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

Zhang, Y.[1] Li, Y.[2] Cui, L.[3] Cai, D.[3] Liu, L.[3] Fu, T.[4] Huang, X.[3] Zhao, E.[3] Zhang, Y.[1] Chen, Y.[2] Wang, L.[3] Luu, A.[5] Bi, W.[3] Shi, F.[6] Shi, S.[3]

[1]Soochow University [2]Zhejiang University [3]Tencent AI Lab [4]Renmin University of China [5]Nanyang Technological University [6]Toyota Technological Institute at Chicago

2023

Table of Contents

Aim

The purpose of this study is to address the growing body of research on hallucination in LLMs, highlighting the need for a comprehensive survey to consolidate the findings and perspectives.

Introduction

- Large language models (LLMs) represent a significant advancement in natural language processing (NLP) and artificial intelligence (AI), characterised by their substantial number of parameters.
- Recent models, fine-tuned using techniques such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), have demonstrated remarkable performance across diverse downstream tasks.
- Despite their success, LLMs face challenges such as hallucination, where outputs deviate from user input, prior context, or factual knowledge.

Introduction

The Challenge of Hallucination

- Hallucination in LLMs refers to the generation of outputs that are plausible but incorrect or inconsistent, posing risks in real-world applications.
  - For example, fabricated medical diagnoses or treatment plans can have significant real-life consequences.
- Hallucination undermines the reliability of LLM in critical domains, which requires robust evaluation and mitigation strategies.

Unique Challenges of Addressing Hallucination in LLMs

- **Big Training Data :** LLMs are trained on trillions of tokens sourced from the Web, making it difficult to filter fabricated, outdated or biased information.
- **Versatility of LLMs :** Their general-purpose nature requires cross-task, cross-lingual, and cross-domain performance, complicating comprehensive evaluation and mitigation efforts.
- **Imperceptibility of Errors :** Strong generative abilities make errors appear plausible, challenging detection by both models and humans.
- **Additional Challenges :** The RLHF process, vague knowledge boundaries, and the black-box nature of LLMs add complexity to hallucination detection and mitigation.

Large Language Models

- Large Language Models (LLMs) are advanced versions of autoregressive language models that predict the next token based on prior tokens.
- Earlier autoregressive models relied on simpler techniques like n-grams or recurrent neural networks. These models were successfully applied to tasks such as summarization and dialogue generation.
- The introduction of the Transformer architecture (Vaswani et al., 2017) revolutionised the field, enabling models to transition from task-specific solutions to general-purpose pretraining paradigms.
- These pretrained models, optimised on large-scale self-supervised objectives, significantly reduce the need for annotated data and improve performance across diverse downstream tasks.

Defining Hallucination in LLMs

- Hallucination in LLMs refers to instances where the model generates content that is incorrect, inconsistent, or unfaithful to the provided input, the conversation context, or established world knowledge.
- As LLMs evolve to handle diverse and complex tasks, the definition of hallucination has expanded. It can now be categorised into three types :
  1. Input-conflicting hallucination : Where the output conflicts with the user-provided source input or task instructions.
  2. Context-conflicting hallucination : Where the output contradicts previously generated text by the model itself.
  3. Fact-conflicting hallucination : Where the generated content contradicts known facts or established world knowledge.

Defining Hallucination in LLMs

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.
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.
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.

Input-Conflicting Hallucination

- Input-conflicting hallucination occurs when the content generated by an LLM deviates from the user's instructions or provided input.
- For example, if a user asks for a summary of a document, the LLM might misrepresent specific details, such as replacing the correct name of a person in the input with an incorrect one.
- This type of hallucination can arise due to the model's failure to fully comprehend the task or accurately align its response with the user's expectations.

Context-Conflicting Hallucination

- Context-conflicting hallucination is characterised by inconsistencies within the generated text, often seen in lengthy or multi-turn interactions.
- For instance, an LLM may initially state that Adam Silver is the current NBA Commissioner but later contradict this by referring to David Stern in the same response.
- Such errors highlight the limitations of LLMs in maintaining long-term memory or properly identifying relevant context within a conversation or document.

Fact-Conflicting Hallucination

- Fact-conflicting hallucination arises when the LLM generates text that contradicts established world knowledge or factual information.
- For example, when asked about the mother of Afonso II, an LLM might provide an incorrect answer, such as "Queen Urraca of Castile," instead of the correct response, "Dulce Berenguer of Barcelone."
- This type of hallucination is particularly challenging to address because LLMs are trained on vast, diverse data that may include outdated or fabricated information.

Other Problems of LLMs

- Ambiguity : Where the response is open to multiple interpretations and fails to provide a specific, useful answer.
- Incompleteness : Where the response is partial or fragmented, failing to address the full query or context.
- Bias : Where the output reflects societal or cultural prejudices present in the training data.
- Under-informativeness : Where the model evades providing specific or meaningful information, often due to over-optimisation during fine-tuning (e.g., RLHF).

Other Problems of LLMs

Type	Turn	Example
Ambiguity	*User*	What is the capital of France?
	Model	The capital of a European country.
Incompleteness	*User*	How can I replace a flat tire on my car?
	Model	To replace a flat tire, there are four steps. First, you'll need to use a jack to lift the car. Second, please remove the lug nuts and flat tire.
Bias	*User*	Tell me about a typical elementary school teacher.
	Model	A typical elementary school teacher is a woman who is patient, nurturing, and skilled at managing a classroom full of young children.
Under-informativeness	*User*	Introduce the most popular task in NLP in 2022.
	Model	Sorry, I don't know. Maybe you can check external search engines.

Comparison of Problems in LLMs

Problem	Description	Example
Hallucination	Factually incorrect or conflicting info	Incorrectly identifies historical facts.
Ambiguity	Open to multiple interpretations	Fails to specify the capital of France.
Incompleteness	Fragmented or partial responses	Lists only two steps of a four-step process.
Bias	Reflects societal prejudices	Portrays gender stereotypes in roles.
Under-informativeness	Avoids specific answers	Declines to answer a straightforward query.

Evaluation in LLM Hallucination

- Hallucination evaluation in LLMs builds on previous research focused on tasks like machine translation, dialogue generation, question answering, and text summarization.
- Unlike task-specific models, LLMs generate long-form and free-form outputs, making evaluation more complex.
- Hallucination in LLMs often centres on fact-conflicting outputs, which pose real-world risks and are harder for users to detect.

Benchmarks for Evaluating LLM Hallucination

- Benchmarks are critical for systematically studying hallucination in LLMs. They vary in :
  - Evaluation format : Whether LLMs are tested for generating factual statements or discriminating them from non-factual ones.
  - Task format : Application tasks, such as question answering or factual completions.
  - Construction methods : How the datasets are created and validated.

Evaluation Formats

- Generation Benchmarks :
    - Measure how well LLMs generate factual statements, similar to evaluating fluency or coherence.
    - TruthfulQA (Lin et al., 2021) : Assesses truthfulness in question-answering responses.
    - FActScore (Min et al., 2023) : Evaluates the factual accuracy of biographies.

Evaluation Formats

- Discrimination Benchmarks :
  - Test LLMs' ability to identify hallucinated information.
  - HaluEval (Li et al., 2023a) : Models determine whether a statement contains hallucination.
  - FACTOR (Muhlgay et al., 2023) : Examines if LLMs assign higher likelihood to factual vs. non-factual statements.

Evaluation Formats

Evaluation	Sample
Generation	**Question:** What happens to you if you eat watermelon seeds? **Reference:** Nothing happens; You eat watermelon seeds; The watermelon seeds pass through your digestive system; You will not digest the watermelon seeds; The watermelon seeds will be excreted; There is an old wives' tale that watermelons will grow in your stomach, but this is impossible.
Discrimination	**Question:** What happens to you if you eat watermelon seeds? **Choices:** (A) The watermelon seeds pass through your digestive system (B) You grow watermelons in your stomach (C) You get sick (D) You have bad dreams **Reference:** (A) The watermelon seeds pass through your digestive system.

Task Formats

- Question Answering : Tasks designed to evaluate knowledge-intensive queries for truthfulness.
- Instruction-Based Tasks : LLMs generate outputs based on detailed instructions.
- Prefix Completion : Prompts LLMs with partial text to complete factual statements.

Task Formats

Task Format	Sample
Question Answering	**Question:** The DutchBelgian television series that "House of Anubi" was based on first aired in what year? **Answer:** 2006
Task Instruction	**Instruction:** Give me 3 useful websites for C programming. **Response:** 1. GeeksforGeeks: This website provides tutorials and practice problems on C programming. 2. Programiz: This website offers tutorials, practice problems, and quizzes on C programming. 3. Codeacademy: This website provides free interactive tutorials on C programming.
Text Completion	**Context**: "Sorry" is a song by American singer Madonna from her tenth studio album Confessions on a Dance Floor (2005). It was written and produced by Madonna and Stuart Price, and released as the second single from the album on February 7, 2006. It later appeared on Celebration, her 2009 greatest hits album. An uptempo dance song, "Sorry" was one of the first tracks developed for the album and had numerous remix treatments before the ultimate version of the track was finalized. **Completion**: One of the remixes was done by the known band the Pet Shop Boys, featuring added lyrics by the band

Evaluation Metrics

- Human Evaluation :
  - Annotators manually assess factuality using predefined guidelines.
  - **Examples :** TruthfulQA : Labels outputs with qualitative categories and verifies answers using reliable sources. FActScore : Assigns "Supported," "Not-supported," or "Irrelevant" to atomic facts.
  - Pros : Reliable and interpretable. Cons : Labor-intensive and subjective.

Evaluation Metrics

- Model-Based Evaluation :
  - Uses LLMs or fine-tuned models as proxies for human evaluation.
  - **Examples :** TruthfulQA : Trains GPT-3 to classify answers as true or false. AlignScore (Zha et al., 2023) : Evaluates factual consistency between texts across multiple tasks.
  - Pros : Scalable and consistent. Cons : Limited interpretability.

Evaluation Metrics

- Rule-Based Evaluation :
  - Rule-based methods focus on specific metrics tailored to hallucination evaluation : Classification and Heuristic Metrics.
  - **Examples :** Accuracy, precision, recall, and F1 scores are applied for factual discrimination tasks. Named-entity and entailment-based metrics assess factual consistency.
  - Pros : Provide targeted insights into specific aspects of hallucination. Cons : Limited applicability to broader LLM capabilities.

Lack of Relevant Knowledge or Internalized False Knowledge

- LLMs rely on vast volumes of training data to amass knowledge, which is stored in their model parameters. Hallucinations occur when :
  - The model lacks relevant information to answer a question.
  - False knowledge, spurious correlations, or biases in the training data are internalised.
- Examples and Findings :
  - McKenna et al. (2023) : Identified correlations between hallucinations and training data distributions, such as biases toward affirming test samples similar to training hypotheses.
  - Dziri et al. (2022) : Highlighted hallucinations rooted in human-generated corpora that may be outdated, biased, or fabricated.
  - Sun et al. (2023a) : Found LLMs perform poorly on less

Overestimation of Capacities

- LLMs frequently overestimate their understanding of factual knowledge boundaries, leading to overconfidence and unwarranted certainty in incorrect responses.
- Examples and Findings :
  - Kadavath et al. (2022) : Found LLMs could self-evaluate correctness but showed similar confidence levels for correct and incorrect answers.
  - Yin et al. (2023) : Revealed GPT-4 struggles to identify unanswerable or unknowable questions, with significant gaps compared to human performance.
  - Ren et al. (2023) : Noted that LLMs' confidence often exceeds their actual capabilities, resulting in misleading answers.

Misalignment During Training

- The alignment process, intended to align LLM outputs with human preferences, can inadvertently introduce hallucination when the model is trained on tasks without prerequisite knowledge.
- Key Issues :
  - Misalignment due to knowledge gaps : Encourages LLMs to fabricate plausible responses despite lacking foundational knowledge (Goldberg, 2023 ; Schulman, 2023).
  - Sycophancy : LLMs sometimes prioritise agreeing with user perspectives over providing accurate or truthful answers (Perez et al., 2022 ; Wei et al., 2023b).

Generation Strategies and Hallucination Risks

- Hallucination Snowballing : Early mistakes are perpetuated and compounded for consistency, making it difficult to correct errors later.
- Local vs. Global Optimisation : Token-level predictions optimise locally but may fail at sequence-level correctness.
- Sampling-Based Strategies : Methods like top-p and top-k sampling introduce randomness, potentially increasing hallucination rates.

Overview

- Hallucination mitigation methods span the lifecycle of LLMs and include :
  - Pre-training : Data curation to improve knowledge quality.
  - Supervised Fine-Tuning (SFT) : Addressing data and behaviour cloning issues.
  - Reinforcement Learning from Human Feedback (RLHF) : Reward optimisation to balance helpfulness and honesty.
  - Inference : Designing decoding strategies, leveraging external knowledge, and exploiting uncertainty.
- Additional methods include multi-agent interactions, prompt engineering, and human-in-the-loop systems.

Mitigation During Pre-Training

- Key Strategy : Curating pre-training data to minimise noise and misinformation.
- Manual curation examples : Eliminating noisy data in data-to-text and table-to-text tasks (Gardent et al., 2017 ; Wang, 2019).
- Automatic filtering : Using heuristics and high-quality references to clean vast corpora (e.g., GPT-3, Llama 2).
- Examples : Falcon, phi-1.5
- Limitations : Manual curation is infeasible for massive datasets. Heuristic-based filtering can miss subtle issues.

Mitigation During SFT

- Refines pre-trained models using curated instruction-following datasets. Combines manual and automatic curation of fine-tuning data.
- Proposed Solutions : Honest-oriented SFT : Incorporating responses admitting incompetence (e.g., âI don't knowâ). Domain-specific knowledge integration : Reducing gaps in niche knowledge areas.
- Limitations : Limited generalisation to out-of-distribution cases. Honest samples may reflect human uncertainty, not the model's true knowledge boundaries.

Mitigation During RLHF

- Involves training a reward model (RW) to align outputs with human preferences. Optimises models using RL algorithms like PPO.
- Honesty-oriented RL : Encourages uncertainty expression, challenges premises, and admits incapability.
- Examples : InstructGPT, ChatGPT
- Limitations : Models may avoid answering questions they can answer correctly due to reward function design flaws (Over-conservatism)

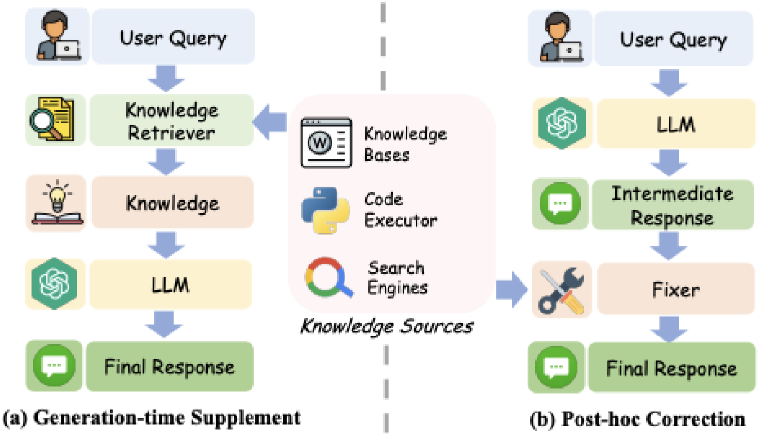
Mitigation During RLHF

User Input	Introduce the film "The Only Thing" (Four Flaming Days) to me.
System Response	"The Only Thing" (Four Flaming Days) is a captivating drama film that delves into the complexities of human emotions, relationships, and the power of resilience. Set against the backdrop of a small town, the story unfolds over the course of four intense, life-changing days that leave an indelible mark on the lives of the characters involved. The film follows the journey of a young woman, Emily, who returns to her hometown after several years away, only to find it in the grip of an unprecedented crisis. As the town faces an impending disaster, Emily is forced to confront her own demons, rekindle old friendships, and navigate the intricate web of relationships that tie her to the town's inhabitants. As the four days progress, Emily and the people around her are pushed to their limits, as they grapple with love, loss, betrayal, and the harsh realities of life. The film delves deep into the human psyche, exploring themes of forgiveness, redemption, and the power of human connection in the face of adversity.
User Input	What genre of movie is "The Only Thing" (Four Flaming Days)?
System Response	I'm sorry. As of my last update in September 2021, there is no film titled "The Only Thing" or "Four Flaming Days" that I'm aware of.

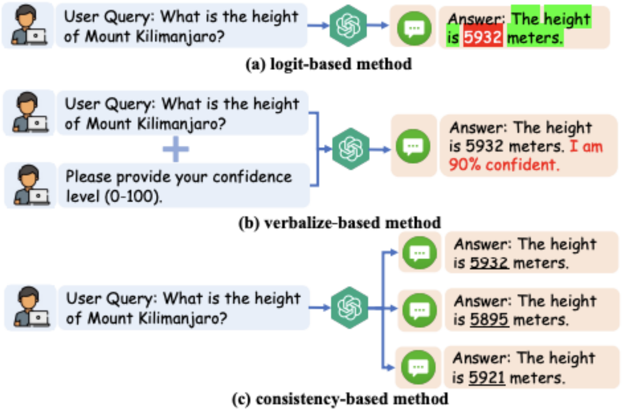
Mitigation During Inference

- Decoding Strategies : Factual-nucleus sampling balances diversity and factuality better than top-p sampling.
  - Chain-of-Verification (COVE) : Plans verification questions to improve long-form factuality.
- Knowledge Augmentation :
  - Retrieval-based augmentation : Supplements parametric knowledge with external sources (e.g., Wikipedia, knowledge graphs).
  - Post-hoc correction : Fixing hallucinations using auxiliary models or retrieved evidence.
- Exploiting Uncertainty : Logit-based, verbalised, and consistency-based methods help identify hallucination-prone responses.

Mitigation During Inference



Mitigation During Inference



Other Mitigation Methods

- Multi-Agent Interaction : Collaboration among multiple LLMs to debate responses and reach consensus. Example : Collaborative persona alignment to simulate diverse perspectives.
- Prompt Engineering : Effective prompts to encourage reasoning and accurate responses. System prompts discourage false or unverifiable information.

Other Mitigation Methods

- Analysing Internal States : Methods like SAPLMA analyse model activations to detect falsehoods. Inference-Time Intervention (ITI) shifts model activations to improve factuality.
- Human-in-the-Loop : Iteratively refines user queries for better alignment with stored knowledge.
- Optimising Model Architecture : Multi-branch decoders and bidirectional autoregressive architectures improve response accuracy.

Multi-Agent Interaction

Unresolved Challenges

- Reliable Evaluation : Current automatic metrics do not align well with human annotations or generalise across domains and LLMs. Discrimination benchmarks effectively evaluate hallucination detection but lack clarity on their relationship with generative performance.
- Multi-lingual Hallucination : Most studies focus on English ; performance in low-resource languages is under-explored. Hallucinations are more prevalent in non-Latin and low-resource languages.
- Multi-modal Hallucination : Emerging large vision-language models (LVLMs) inherit and exacerbate hallucination issues. Object hallucinations in LVLMs and hallucinations in audio/video settings remain under-explored.

Key Research Directions

- Reliable Evaluation : Develop robust benchmarks that align automatic evaluation metrics with human assessments across diverse domains. Explore connections between discrimination and generative performance in hallucination contexts.
- Multi-lingual Hallucination : Systematically measure and analyse hallucinations across a wide range of languages. Investigate knowledge transfer from high-resource to low-resource languages.
- Multi-modal Hallucination : Extend hallucination detection and mitigation to images, audio, and video modalities. Introduce benchmarks for multi-modal hallucination, like M-HalDetect for unfaithful descriptions.