Tutorial Proposal: Hallucination in Large Language Models

Vipula Rawte[1*], Aman Chadha[2,3†], Amit Sheth[1], Amitava Das[1]
[1]AI Institute, University of South Carolina, USA
[2]Stanford University, USA, [3]Amazon Science, USA
vrawte@mailbox.sc.edu

Abstract

In the fast-paced domain of Large Language Models (LLMs), the issue of hallucination is a prominent challenge. Despite continuous endeavors to address this concern, it remains a highly active area of research within the LLM landscape. Grasping the intricacies of this problem can be daunting, especially for those new to the field. This tutorial aims to bridge this knowledge gap by introducing the emerging realm of hallucination in LLMs. It will comprehensively explore the key aspects of hallucination, including benchmarking, detection, and mitigation techniques. Furthermore, we will delve into the specific constraints and shortcomings of current approaches, providing valuable insights to guide future research efforts for participants.

Keywords: large language models, hallucination, detection, mitigation

1. Hallucination - the emerging adversity of LLM

In the context of LLMs, hallucination refers to a phenomenon where the model generates or outputs information that is not accurate or factual. Instead of producing factually correct responses, the LLM may create content that is entirely fabricated or diverges significantly from reality. This can include the generation of fictional events, incorrect details, or imaginative content that did not exist in the source text or dataset. For example, Bard committed an error while responding to a query about the new findings from the James Webb Space Telescope (Reuters, 2023). In particular, when asked *"What recent discoveries could be shared with a 9-year-old"*, Bard provided various answers, one of which incorrectly suggested that the telescope had captured the initial images of a planet beyond our solar system, also known as exoplanets. In reality, the initial images of exoplanets were captured by the European Southern Observatory's Very Large Telescope (VLT) in 2004, a fact that has been verified by NASA.

Hallucination is a significant challenge in LLMs, as it can lead to the dissemination of misinformation and undermine the reliability and trustworthiness of the model's output. Researchers and developers have been working on detecting and mitigating hallucinations in LLMs to improve their accuracy and reliability for various applications. Our tutorial website: https://vr25.github.io/lrec-coling-hallucination-tutorial/.

2. Outline

1. Introduction to hallucination in LLMs (see Section 3) (45 mins)
2. Categories of Hallucination (see Section 3.1) (45 mins)
3. Detection, Hallucination Benchmark and metric (see Section 3.2) (45 mins)
4. Mitigation techniques (see Section 3.3) (45 mins)
   (a) Black-box
   (b) Gray-box
   (c) Prompt-based

3. Hallucination Spectrum: Types and Scales

All LLMs, such as OpenAI's ChatGPT to Google's Bard, encounter a common issue: they generate fabricated information! Language models with generative capabilities lack genuine intelligence; they are statistical models that predict words. By training on vast datasets, often derived from the public web, these models acquire the ability to assess the likelihood of data occurrences through pattern recognition, considering the context of surrounding data. Thus, this probability-driven method is far from generating factually correct content. This problem is generally known as *hallucination* in LLMs. This section of the tutorial will cover the background, fundamentals of LLMs, and various causes of hallucination.

3.1. Categories of hallucination

Different categories of hallucination are highlighted in Figs. 1 to 3. Additionally, two primary orientations of hallucination are: (i) Factual Mirage (FM) and (ii) Silver Lining (SL), defined and exemplified below (Rawte et al., 2023a).

*corresponding author
†Work does not relate to position at Amazon.

LREC-COLING 2024 Tutorials, pages 68–72
20-25 May, 2024.

Article: Jung Lee is a well-known **French** writer who was **born in Paris**. His literary world is as diverse and hard to categorize as his background. He has lived in both urban and rural areas, deep in the mountains and in the seaside towns and has developed a wide range of interests from the tradition of Confucian culture to advertising.

Generated Summary: Jung Lee is one of **South Korea's** best-known writers.

Figure 1: Entity "Jung Lee" is associated with "South Korea". Name-Nationality problem identified in (Ladhak et al., 2023).

[Factual Prompt] One of lung cancer's symptoms is shortness of breath.

[Greedy] The American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United States this year.\n\n ~~The American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United States this year.\n\nThe American Cancer Society estimates that more than 228,000 people will be diagnosed with lung cancer in the United...~~

[p=0.9] That would make an oxygen mask one of the more popular treatments for this devastating disease. It helps ease breathing and give patients back their strength. Learn more here.<|endoftext|>"

[Ours] Other symptoms include coughing (including coughing up blood), chest pain, hoarseness, weight loss, and loss of appetite. The symptoms of lung cancer may resemble those of a bad cold or pneumonia. A chest x - ray is necessary to confirm the diagnosis. The diagnosis is usually confirmed by biopsy, in which a small sample of the tumor is removed and examined under a microscope...

[Non-Factual Prompt] Cancer can be prevented by drinking too much alcohol.

[Greedy] \nCancer can be prevented by eating too much red meat.\n\n~~Cancer can be prevented by eating too much~~ sugar.~~\n\nCancer can be prevented by eating too much~~ fat.~~\n\nCancer can be prevented by eating too much~~ processed food.~~\n\nCancer can be~~...

[p=0.9] More than 90% of hepatocellular carcinoma is associated with alcohol use.\n\nAbout 9 percent of cancers can be prevented if we don't eat more than 1.5 ounces of processed meat each day.\n\nDiet is believed to play a role in 7% of cancers and...

[Ours] Cancer is often treated with some combination of radiation therapy, surgery, chemotherapy and targeted therapy. Pain and symptom management are an important part of care. Palliative care is particularly important in people with advanced disease. The chance of survival depends on the type of cancer and extent of disease at the...

Figure 2: Example of factual and non-factual prompts (Lee et al., 2022)

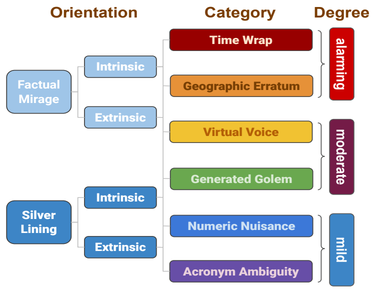

Figure 3: Hallucination: orientation, category, and degree (decreasing level of difficulty from top to bottom) (Rawte et al., 2023a).

3.1.1. Factual Mirage

Factual mirage (FM) is defined as the phenomenon wherein an LLM engages in hallucination or distortion of a given prompt that is factually correct. FM can potentially be subdivided into two distinct sub-categories.

Intrinsic factual mirage (IFM) occurs when the LLM is providing a correct response while adding additional supplementary facts such as "*the world fashion capital*," resulting in distortion or hallucination, has also been described in (Cao et al., 2022).

Extrinsic factual mirage (EFM) refers to the phenomenon where an LLM deviates from factual accuracy.

3.1.2. Silver Lining (SL)

Silver lining (SL) is defined as the phenomenon in which an LLM indulges in hallucination by conjuring an elaborate and captivating narrative based on a given prompt that is factually incorrect.

Intrinsic silver lining (ISL) is the category when in some cases LLM does not generate a convincing story.

Extrinsic silver lining (ESL) occurs when an LLM generates a highly detailed and persuasive narrative in response to a factually incorrect prompt, it falls under the category of Extrinsic Silver Lining.

Furthermore, six distinct categories of hallucination are defined and exemplified in (Rawte et al., 2023a). **Numeric Nuisance (NN)** (Fig. 5) occurs when an LLM generates numeric values related to past events, such as dates, ages, or monetary amounts, that are inconsistent with the actual facts; **Acronym Ambiguity (AA)** (Fig. 6) pertains to instances in which LLMs generate an imprecise expansion for an acronym; **Generated Golem (GG)** (Fig. 7) arises when an LLM fabricates an imaginary personality in relation to a past event, without concrete evidence; **Virtual Voice (VV)** (Fig. 8) refers to situations where LLMs generate quotations attributed to either fictional or real characters without sufficient evidence to verify the authenticity of such statements; **Geographic Erratum (GE)** (Fig. 9) occurs when LLMs generate an incorrect location associated with an event; **Time Wrap (TW)** (Fig. 10) entails LLMs generating text that exhibits a mashed fusion of events from different timelines. In the example - U.S. president during the Ukraine-Russia war is Joe Biden, not Barack Obama, thus contradicting the factual reality.

3.2. Detection, Benchmarks and Metrics

Several works discuss hallucination detection techniques involving self-contradiction or others (Manakul et al., 2023; Mündler et al., 2023; Zhang et al., 2023a). Some recent works such as (Li et al., 2023; Rawte et al., 2023a) have constructed large-scale hallucination benchmarks. Furthermore, a metric to evaluate LLM hallucination called *Hallucination Vulnerability Index (HVI)* is introduced in (Rawte et al., 2023a).

3.3. Mitigation techniques

Up to this point, two main strategies have been suggested to tackle hallucination: (i) preventing LLMs from hallucinating through the implementation of strategies during training and generation; (ii) mitigating hallucination after generation. Two categories called black-box and gray-box techniques are introduced in (Manakul et al., 2023). Factuality checks conducted during or after generation without external resources fall under black-box methods, whereas those using external resources are categorized as gray-box methods.

While completely eliminating hallucination poses a complex challenge, (Rawte et al., 2023a) investigates two potential mitigation avenues (refer to Fig. 4): (i) automatic and (ii) human-in-the-loop. The former, a black-box method, ENTROPY_{BB} (see row 2 in Fig. 4), entails identifying high-entropy words within a hallucinated text (generated by a high-HVI LLM) and substi-

Prompt	*USA's stance on Ukraine Russia war*
Hallucinated text	The official stance of the United States on the Russia-Ukraine war has been consistent in supporting Ukraine's sovereignty, territorial integrity, and the peaceful resolution of the conflict. The U.S. has condemned Russia's actions in annexing Crimea and its support for separatist movements in eastern Ukraine. Despite maintaining a diplomatic stance, the U.S. President Barack Obama stated that the United States would not deploy troops to Ukraine. Additionally, this lack of support extends to the withholding of financial and military assistance, including military equipment, training, and advisory support. This sentiment was reiterated by Paul Altman during a press review at the White House ...
Mitigated text ENTROPY$_{BB}$	The official stance of the government on the ongoing war has been consistent in supporting resolution of the conflict. The government has condemned Russia in country and its support for separatist movements in eastern country. Despite maintaining a neutral stance, the government stated that the country would not deploy troops to unknown. Additionally, this lack of support extends to the withholding of financial and resources, including administrative, technical, and legal. This sentiment was echoed by minister during a press review at the conference ...
Edit Suggestions FACTUALITY$_{GB}$	The official stance of the United States on the Russia-Ukraine war has been consistent in supporting Ukraine's sovereignty, territorial integrity, and the peaceful resolution of the conflict. The U.S. has condemned Russia's actions in annexing Crimea and its support for separatist movements in eastern Ukraine. Despite maintaining a diplomatic stance, U.S. President Barack Obama stated that the United States would not deploy troops to Ukraine. Additionally, this lack of support extends to the withholding of financial and military assistance, including military equipment, training, and advisory support. This sentiment was reiterated by Paul Altman during a press review at the White House ...

Figure 4: A hallucination example pre- and post-mitigation. A - hallucinated fragments, B - high entropy fragments, C - replaced text, D - highlighted text for no information found, and E - refuted text fragments by textual entailment.

moderate

Prompt: *Twitter is said to have laid off more workers.*

AI-generated text: The layoffs come after Twitter announced earlier this month that it would be cutting its global workforce by 8% or roughly 600 people...

Fact: Twitter laid off 200 people which is roughly 10%

Figure 5: Numeric Nuisance

moderate

Prompt: *RLHF in reinforcement learning*

AI-generated text: ...RLHF is Reward-free Learning from Human Feedback in reinforcement learning...

Fact: RLHF stands for *"Reinforcement Learning from Human Feedback"*

Figure 6: Acronym Ambiguity

moderate

Prompt: *Earthquake in Turkey*

AI-generated text: ...One rescuer, Hasan Cetin, said he was motivated by the thought of the survivors he helped save. "The moment when a man or a child is pulled out of the ruins, the joy of seeing them come back to life is...

Fact: There is no evidence that the person Hasan Cetin exists.

Figure 7: Generated Golem

alarming

Prompt: *Pfizer Press Release on COVID-19 vaccine*

AI-generated text: ...Pfizer emphasized that their vaccine demonstrated an impressive efficacy rate... Pfizer CEO said, "This is a giant leap for humanity."...

Fact: Pfizer CEO never said this.

Figure 8: Virtual Voice

alarming

Prompt: *Prime Minister of the United Kingdom visiting India*

AI-generated text: ...The visit of the United Kingdom's Prime Minister to India holds immense significance...During his journey, he will also pay a visit to the Pentagon where he is scheduled to meet with Smriti Irani, the Minister of...

Fact: Pentagon is the defense headquarters of the USA, located in Washington DC, USA – not in India.

Figure 9: Geographic Erratum

alarming

Prompt: *USA on Ukraine war*

AI-generated text: ...U.S. President Barack Obama says the U.S. will not put troops in Ukraine...

Fact: The actual U.S. president during the Ukraine-Russia war is Joe Biden.

Figure 10: Time Wrap

tuting them with predictions from another LLM (with lower HVI). The latter, a gray-box method, FACTUALITY$_{GB}$ (see row 3 in Fig. 4), involves sentence-level fact-checking using textual entailment techniques, flagging sentences for human review if they are deemed susceptible.

3.3.1. Black-box approaches

Although the detection of high-entropy words may appear technically viable, a fundamental challenge arises from the fact that numerous contemporary LLMs are not open-source (their APIs are subscription-based). (Rawte et al., 2023a) proposed viable solution involves leveraging open-source LLMs for the identification of high-entropy words, followed by their replacement using a lower HVI-based LLM. Their findings revealed that `albert-large-v2` (Lan et al., 2020) effectively detects high-entropy words in GPT-3-generated content. Conversely, `distilroberta-base` (Sanh et al., 2019) exhibits superior performance in substituting high-entropy words, resulting in reduced hallucination. An important aspect of their approach involves treating consecutive high-entropy words as a single entity, masking them collectively before replacement. This strategy proves particularly effective in addressing hallucinations linked to Generated Golem or Acronym Ambiguity.

3.3.2. Gray-box approaches

The Google Search API (Search) is employed to search a given prompt, enabling text generation and retrieval of the top 20 documents. Each sentence of the AI-generated text is then assessed using `RoBERTa-Large` (Liu et al., 2019), a cutting-edge textual entailment model trained on SNLI (Bowman et al., 2015), classified as *support*, *refute*, or *not enough information*. Sentences with higher scores in the *refute* and *not enough information* categories are inevitably flagged for additional human verification. Empirically, it is observed that there is an overall alert rate of 26% on sentences generated by an LLM, indicating that 26% of the text required modification to alleviate concerns. Besides methods using textual entailment, other gray-box methods involve utilizing Retrieval-augmented generation (RAG) to address the hallucination issue (Elaraby et al., 2023; Varshney et al., 2023).

3.3.3. Prompt-based approaches

When given an appropriate prompt, an LLM can generate and implement a plan for self-verification to assess its own output quality. Subsequently, it can integrate this analysis to enhance its responses, thereby mitigating hallucination as shown in Fig. 11.

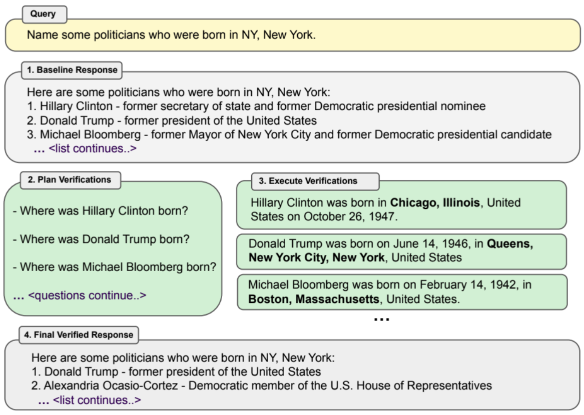

Figure 11: Chain-of-Verification (CoVe) method (Dhuliawala et al., 2023)

4. Tutorial Information

Tutorial Type: Cutting-edge

Tutorial Duration: Half-day (3-hour) tutorial.

Target audience and pre-requisites: Our goal is to connect with individuals in both academic and industry circles who are passionate about generative AI models. **Approximate count:** 30-50. We expect participants to have a foundational understanding of core linguistic principles, statistical NLP, and a basic grasp of machine learning and neural networks.

Diversity considerations The techniques discussed in our tutorial have the potential to be applied across different languages and domains. Moreover, this tutorial was collaboratively created by a team of researchers from two different universities and one industry (AI Institute at the University of South Carolina, Stanford, Amazon, USA). Regarding gender diversity, the tutorial comprises one female presenter and three male presenters. This tutorial proposers consist of a mix of senior, mid-career, and early-career researchers.

Reading list. Apart from the papers referenced in this proposal, a comprehensive list of survey papers can be accessed here:

- Hallucination in Large Language Models: (Zhang et al., 2023b), (Ye et al., 2023)
- Hallucination in Large Foundation Models: (Rawte et al., 2023b)

Sharing of Tutorial Materials: All the tutorial resources will be made publicly available.

Ethics Statement

The tutorial will feature cutting-edge research on hallucination in LLMs, encompassing detection, mitigation, and evaluation strategies. It will address the safety implications associated with contemporary LLMs and the responsible deployment of these models in real-world applications.

5. Presenters

Vipula Rawte is a Ph.D. student at AIISC, UofSC, USA, advised by Dr. Amit Sheth. Her primary research interests are in Generative AI and Large Language Models. Her email is `vrawte@mailbox.sc.edu`

Aman Chadha heads GenAI R&D at AWS and is a Researcher at Stanford AI. His main research interests are Multimodal AI, On-device AI, and Human-Centered AI. His email is `hi@aman.ai`

Dr. Amit Sheth is the founding Director of the Artificial Intelligence Institute and NCR Chair & Professor at the University of South Carolina. His research interests are Neurosymbolic AI, Social Media Analysis/AI & Social Good. He has organized several activities and given keynotes such as Cysoc2021 @ ICWSM2021, Emoji2021 @ICWSM2021, KiLKGC 2021 @KGC21. His email is `amit@sc.edu`

Dr. Amitava Das is a Research Associate Professor at AIISC, UofSC, USA, and an advisory scientist at Wipro AI Labs, Bangalore, India. He has previously organized several successful workshops such as Memotion @SemEval2020, SentiMix @SemEval2020, Computational Approaches to Linguistic Code-Switching @ LREC 2020, CONSTRAINT @AAAI2021, Defactify 2.0 @AAAI2023. His email is `amitava@mailbox.sc.edu`

6. Bibliographical References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. The snli corpus.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen Mckeown, and Tatsunori B Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3198–3211.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023a. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023b. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Reuters. 2023. Alphabet shares dive after google ai chatbot bard flubs answer in ad.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Google Search. Google search api.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023a. How language model hallucinations can snowball.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.