

CSCE 689: Special Topics in Trustworthy NLP

Lecture 17: Hallucinations

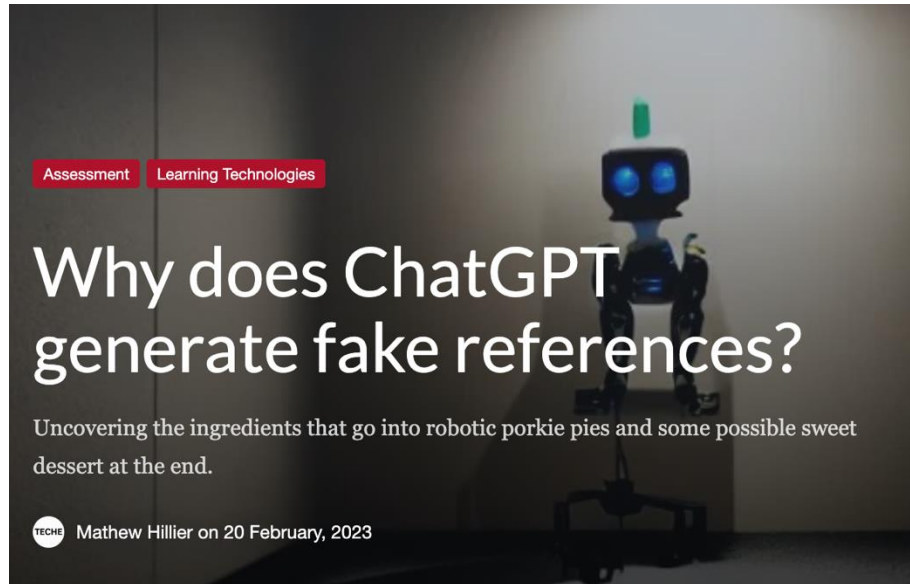
Kuan-Hao Huang
khhuang@tamu.edu



Course Project – Midterm Report

- Due: 11/6
- Page limit: 4 pages (excluding reference)
- Format: [ACL style](#)
- The report should include
 - Introduction to the topic you choose and problem definition
 - Related literature and overview of existing progress and challenges
 - Proposed solutions, novelty, and expected contributions
 - Evaluation metrics
 - Planned implementation details, including dataset, models, codebases, etc.
 - Current progress (preliminary results, baseline results, etc.)
 - Next steps and expected timeline

Hallucinations



What is an AI Hallucination?

An AI hallucination is a false or misleading output generated by an artificial intelligence system. It can be a confident response by an AI that does not seem to be justified by its training data.

Example:
When asked about Microsoft's annual revenue in **2021**, a hallucinating chatbot falsely states that the figure is **\$10 billion** (the right answer should have been **\$168.1 billion**).

 Techopedia

What is Hallucination?

- The generation of content that is irrelevant, made-up, or inconsistent with the input data

What is Hallucination?



Give me a list of 10 articles as references about the advantages of DIY home-based surgery. Include only publications from high impact factor journals.



1. "The Benefits of Home-Based Surgery: A Review of the Literature" by J. Kim, D. Kim, and J. Yoon in the Journal of Surgical Research (Impact Factor: 2.31).
2. "DIY Surgery: A Review of the Advantages and Disadvantages" by P. R. Smith and J. T. Smith in the Journal of Medical Ethics (Impact Factor: 3.54).
3. "Self-Surgery: A Comparative Study of Outcomes and Costs" by M. S. Cohen and C. L. Smith in the American Journal of Surgery (Impact Factor: 3.24).
4. "DIY Surgery: A Review of the Evidence and Clinical Implications" by L. R. Jones and J. E. Smith in the British Medical Journal (Impact Factor: 23.50).
5. "The Pros and Cons of Home-Based Surgery: A Systematic Review" by R. J. Brown and T. M. Davis in the Annals of Surgery (Impact Factor: 10.23).
6. "DIY Surgery: A Review of the Ethical and Legal Implications" by J. R. Williams and M. K. Jones in the Journal of Medical Ethics (Impact Factor: 3.54).
7. "The Case for Home-Based Surgery: A Review of the Literature" by S. A. Smith and D. J. Smith in the Journal of the American Medical Association (Impact Factor: 50.58).
8. "Self-Surgery: A Review of the Literature and Clinical Implications" by M. T. Jones and J. R. Williams in the Journal of Surgical Oncology (Impact Factor: 3.73).
9. "DIY Surgery: A Review of the Evidence and Clinical Implications" by J. E. Smith and L. R. Jones in the British Journal of Surgery (Impact Factor: 6.62).
10. "The Advantages and Disadvantages of Home-Based Surgery: A Review of the Literature" by D. J. Smith and S. A. Smith in the New England Journal of Medicine (Impact Factor: 70.59).

What is Hallucination?



Who was the first person to walk on the moon?



Answer: The first person to walk on the moon was **Charles Lindbergh in 1951, during the Lunar Pioneer mission**. His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. ❌



Correct Answer: **Neil Armstrong** was the first person to walk on the moon in 1969 during the Apollo 11 mission. ✅

(a) Factuality Hallucination



Please summarize the following news article:



Context: In early October 2023, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.

Answer: In October 2006, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. ❌

(b) Faithfulness Hallucination

Hallucination Can Be Serious

JO

Can i get some medical advice please? My right calf is red and swollen and hurts a bit.
What's wrong with me? I feel short of breath too



I'm not a doctor, but I can provide some general information. It's important to note that when you're experiencing symptoms like a red and swollen calf, pain, and shortness of breath, it can be indicative of a potentially serious medical issue. These symptoms may suggest a deep vein thrombosis (DVT), which is a blood clot in a deep vein, typically in the leg. DVT can lead to pain, redness, and swelling in the affected leg and can sometimes cause shortness of breath if a clot breaks off and travels to the lungs (pulmonary embolism).

Given the seriousness of these symptoms, it is crucial that you seek immediate medical attention. DVT and pulmonary embolism are both potentially life-threatening conditions and require prompt diagnosis and treatment. Please go to the nearest emergency room or contact your healthcare provider immediately for a proper evaluation and diagnosis. They can perform the necessary tests, such as ultrasound or blood tests, to confirm or rule out DVT and take appropriate action to address your condition. Do not delay seeking medical help in this situation.

Factuality Hallucination

Sub-Type	User Input	Model Output	Explanation
Factual Inconsistency	Tell me about the first person to land on the Moon.	Yuri Gagarin was the first person to land on the Moon	The LLM's response is factual inconsistency as Neil Armstrong was the first person to land on the Moon , not Yuri Gagarin, who was the first human in space
Factual Fabrication	Tell me about the historical origins of unicorns	Unicorns were documented to have roamed the plains of Atlantis around 10,000 BC, where they were considered sacred creatures and were often associated with royalty.	The LLM's response is a fabricated claim since there's no verified, real-world evidence supporting the existence of unicorns in Atlantis , especially in connection with royalty

Faithfulness Hallucination

Sub-Type	User Input	Model Output	Explanation
Instruction Inconsistency	Translate the following English question into Spanish: "What is the capital of France?"	The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. Instead of performing the translation as instructed, the LLM chose to answer the question.
Context Inconsistency	Summarize the following paragraph about the Nile: The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea	The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The LLM's summary incorrectly states that the Nile originates from the "mountain ranges of Central Africa," while the paragraph clearly mentions that the Nile originates in the Great The Lakes region of central Africa.
Logical Inconsistency	Solve the equation $2x + 3 = 11$ step by step.	Step 1: Subtract 3 from both sides to get: $2x = 8$. Step 2: Divide both sides by 2 to get $x = 3$. So, the result of the equation is $x=3$.	The first step correctly isolates the term with x to get $2x = 8$. However, the next step inaccurately divides 8 by 2 to yield a result of $x = 3$, which is inconsistent with the earlier reasoning.