INFO 7375 - Neural Networks & AI

Home Work to Chapter - 24

Submitted By:

Abdul Haseeb Khan
NUID: 002844724
khan.abdulh@northeastern.edu

What is AI hallucination?

AI hallucination is a phenomenon where a large language model (LLM), such as a generative AI chatbot or computer vision tool, perceives patterns or objects that are nonexistent or imperceptible to human observers, resulting in outputs that are nonsensical or accurate. It occurs when the model generates false, nonsensical, or inaccurate information and presents it as factual. Metaphorically, this is similar to how humans might see figures in clouds, but in AI, it stems from the model's design to predict the next word based on patterns rather than understanding factual accuracy.

What types of AI hallucinations do you know?

The following are the types of AI hallucinations:
Factual Hallucinations: The model produces confident but incorrect or unsubstantiated information, such as claiming Einstein won two Nobel Prizes when he only won one.
Contextual Hallucinations: The response is linguistically correct but deviates significantly from the prompt or breaks the logical flow of conversation, such as discussing Saturn's rings in response to a recipe request.
Logical Hallucinations: The model provides a logically invalid answer, often in tasks requiring math or reasoning, such as failing a simple subtraction problem.
Multimodal Hallucinations: These occur in systems handling multiple media types, where there is a mismatch or fabrication across modalities, like a text description not matching a generated image.

What does cause AI hallucinations?

Hallucinations stem from fundamental limitations in how models are trained and generate responses. Key causes include:

- **Data Bias or Gaps:** Models inherit bias or misinformation present in their training data.
- **Probabilistic Nature:** LLMs estimate the next most likely word rather than checking facts, leading to confident-sounding errors.
- **Prompt Ambiguity:** Insufficiently specific prompts can cause the model to misinterpret user intent.
- **Overfitting and Memorization:** Models may regurgitate outdated or incorrect specific phrases they memorized during training.
- **Lack of Grounding:** Models rely solely on training data without reference to current events or external knowledge bases.

What is the impact of AI hallucinations?

The impact varies by domain. In high-risk fields like healthcare, law, or finance, hallucinations can lead to real-world damage, such as medical errors, legal misadvice, or financial losses. Conversely, in low-risk domains like creative writing, "hallucinations" can usefully facilitate creativity and artistic expression. However, frequent hallucinations in high-stakes situations can erode user trust, slow down adoption, and increase regulatory scrutiny.

How to detect AI hallucinations?

Detection involves comparing AI outputs against a ground truth using manual or automated methods. Strategies include:

- **Manual Review:** Human specialists assess results for correctness, which is effective but slow.
- **Automated Cross-Verification:** Comparing responses to structured knowledge bases like Wikipedia to flag inconsistencies.
- **AI-Based Detection:** Using an "LLM-as-a-judge" to evaluate if an answer is supported by context, or using ensemble models to analyze uncertainty estimates.
- **Traceability Methods:** Advanced methods like VeriTrail trace the information flow in complex workflows to pinpoint where errors were introduced.

How to prevent AI hallucinations?

Preventing (or mitigating) hallucinations involves model design choices and safeguards:

- **Retrieval-Augmented Generation (RAG):** This combines the model with a retrieval system to ground outputs in real-time, verified documents.
- **Fine-tuning:** Training on curated, factual datasets discourages the memorization of false information.
- **Prompt Engineering:** Using structured prompts and constraints helps the model understand ambiguities and limits answers to what is knowable.
- **Confidence Thresholds:** Models can be configured to refuse answers or flag them when their confidence score is low.