NetApp

MANY FACETS OF AI HALLUCINATIONS: FACTUAL ERRORS, DEEP FAKES, AND CREATIVITY

Josef Machytka <josef.machytka@netapp.com>
NetApp Open Source Services
2024-11-15 - NetApp Tech Talk

Josef Machytka

- Professional Service Consultant - PostgreSQL specialist at NetApp Open Source Services / credativ
- 30+ years of experience with different databases.
- PostgreSQL (12y), BigQuery (7y), Oracle (15y), MySQL (12y), Elasticsearch (5y), MS SQL (5y).
- 10+ years of experience with Data Ingestion pipelines, Data Analysis, Data Lake and Data Warehouse
- 2 years of practical experience with different LLMs / AI including their architecture and principles.
- From Czechia, living now 11 years in Berlin.

- LinkedIn: linkedin.com/in/josef-machytka
- ResearchGate: researchgate.net/profile/Josef-Machytka
- Academia.edu: netapp.academia.edu/JosefMachytka
- Medium: medium.com/@josef.machytka
- sessionize: sessionize.com/josefmachytka

Acknowledgement

- I want to thank to my colleague **Felix Alipaz-Dicke** for critical review of this presentation.
- He gave me valuable feedback and many ideas and suggestions on how to improve this talk.
- He is deeply interested in AI and has a lot of experience in this field.
- Do not hesitate to ask him questions or discuss with him the topic of AI.

Table of contents

All AI images without credits
were created by the author of this talk
using DeepDreamGenerator

What are AI hallucinations

"42" - The first ever documented AI hallucination?

- ”The Hitchhiker’s Guide to the Galaxy” by Douglas Adams features ”Deep Thought”
- Deep Thought was created to find the answer to ”life, the universe, and everything”
- After long computation, it concluded the answer was: ”42”
- This could be seen as the first documented AI hallucination
- Perhaps even the ultimate AI hallucination on the ultimate question

(Image from the article Deep Thought on V-1 Onsen IV – Archaeogaming)

AI hallucinations are hard to detect

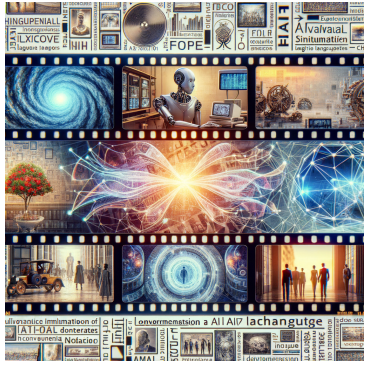
- AI outputs that seem plausible but are factually incorrect
- AI can generate hallucinations in text, images, videos, or audio
- Some outputs can be creative, innovative, or groundbreaking
- Others can cause dissatisfaction and confusion
- Can lead to serious legal, financial, or ethical issues
- They can be used to manipulate public opinion
- In finance, law, and healthcare can have severe consequences
- Cases of legal and financial issues caused by AI hallucinations

Text hallucinations

- Grammatically correct, plausible, and confident
- Use proper language and terminology, highly believable
- May contain facts borrowed from similar topics
- But not applicable to the context
- Can include entirely fictitious information
- LLMs can confuse software versions, models, algorithms
- Key distinction: *”plausibility does not guarantee truth”*

Deep fakes as a form of AI hallucinations

- Deliberately generated using different AI tools
- Usually realistically looking videos
- People in them do or say things they never did
- Can depict events that never happened
- Intended to manipulate public opinion
- Or discredit people, or spread fake news
- But can also be used for entertainment, art, or education
- The impact depends on the context and purpose

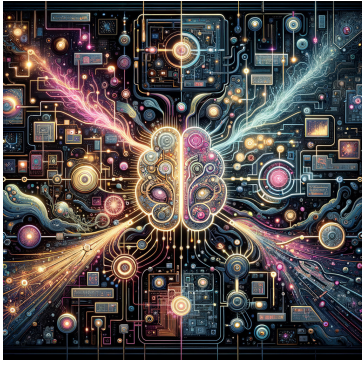


AI hallucinations can also be creative

- AI hallucinations are not necessarily negative
- Context and purpose determine their value
- They can be seen as a form of creativity
- Useful for generating new ideas, concepts, or designs
- Helps explore new possibilities and push knowledge boundaries
- Often used in brainstorming, creative writing, and art

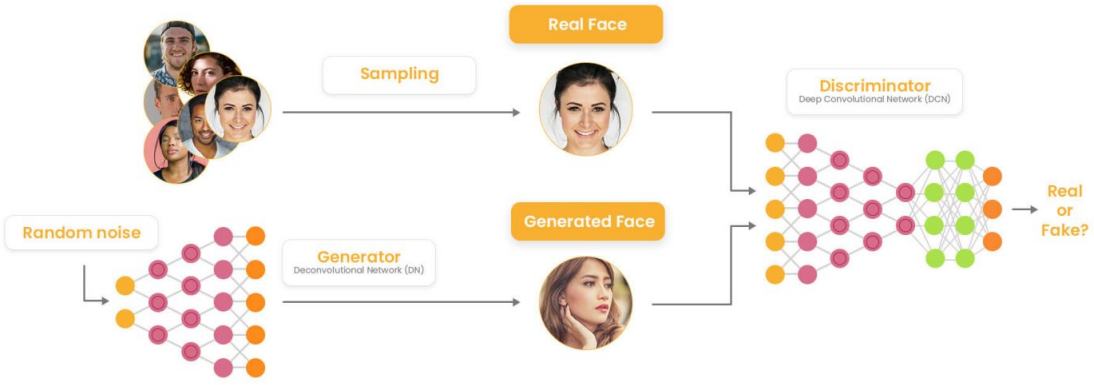
AI hallucinations can pollute the web

- AI hallucinations can pollute the web with false information
- Not everyone verifies information before sharing it
- Some articles are published by AI without human oversight
- People can copy-paste content without verifying its accuracy
- AI-generated content can be later used to train new AI models
- Problem can amplify over time



Generative adversarial networks

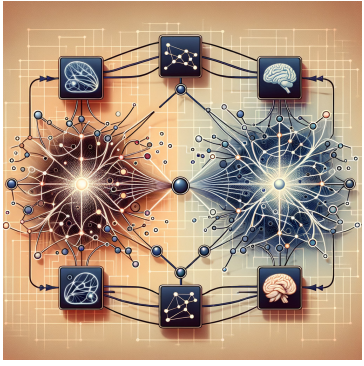
How Generative adversarial networks work



(Image from the article on Generative Adversarial Networks)

How Generative adversarial networks work

- Consist of two neural networks: generator and discriminator
- Generator creates new data, discriminator evaluates them
- Generator aims to create data indistinguishable from real data
- Discriminator distinguishes between real and generated data
- Networks are trained together in a zero-sum game
- GANs are used for generating images, videos, and audio
- Less successful in generating text compared to transformers



Challenges in Balancing GANs

- Hallucinations occur when the Generator fools the Discriminator
- GAN is inherently unstable
- Networks constantly try to outsmart each other
- A weak Discriminator fails to penalize unrealistic data
- This leads to many hallucinations in the generated data
- A strong Discriminator makes it hard for the Generator to improve
- This can result in the Generator producing no new data at all

AI probabilistic reasoning

AI probabilistic reasoning

- AI hallucinations are not intentional errors or bugs
- Natural consequence of AI probabilistic reasoning
- LLMs generate answers based on learned probabilities
- They use advanced statistics from the training data
- The answer is **the most probable one**, not necessarily correct
- For LLM there is no ”correct” or ”incorrect” answer

Overgeneralization

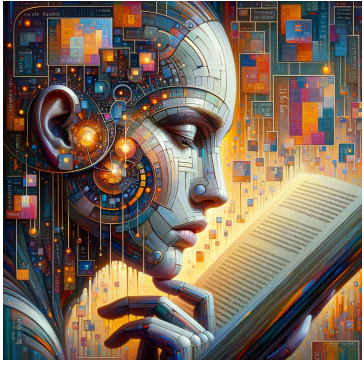
- Based on patterns and associations learned from data
- Training dataset may contain biased data
- Many examples of cats together with dogs in one picture
- The LLM may conclude that ”cats always live with dogs”
- Or that ”cats are a type of dog”
- *Like a student who learned from a biased, one-sided textbook*

Underfitting

- Occurs if the LLM is too simple or the training data is too limited
- Also if the training process was too short or insufficient
- Model fails to capture detailed patterns
- Learns only shallow relationships
- Leads to vague or overly generic answers
- *Like a student who has learned only basic concepts*
- *Tries to fudge answers due to lack of detailed knowledge*

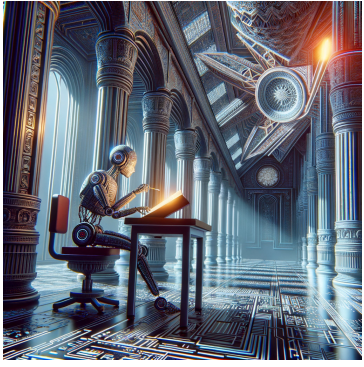
Overfitting

- LLM becomes too specialized on the training data
- It learns all the noise or overly specific details
- Model performs well on training data
- But poorly on new, unseen data
- It can repeat exact phrases from the training data
- Probabilities are too high for patterns in training data
- Training was too long and validation performance degraded
- *Like a student who memorized the textbook*
- *But can't apply knowledge to new problems*



Overinterpretation of the Input

- Occurs when the input is ambiguous or incomplete
- LLM may try to fill gaps in the input
- By making the most probable assumptions
- This can lead to hallucinations in the input
- Hallucinated input often results in hallucinated output
- Clear and precise input is crucial to avoid this issue



Out-of-distribution Generalization

- LLMs are trained on vast data sets
- But the training data is never complete
- Some topics or areas are not covered by the training data
- Answers are based on patterns learned from the training data
- Not on actual knowledge about the topic
- Answers can be completely wrong using unfitting associations
- *Like a student answering a question about a topic he never heard about*

Problems with training datasets

Training data is not perfect

- Trained datasets are complex and costly to collect
- Data is collected from the internet, books, articles
- Data contains errors, biases, and inaccuracies
- Specialized topics often underrepresented
- Dataset is static and reflects the state at the time of collection

Problems with understanding human languages

- Human languages are complex and ambiguous
- Same word have different meanings in different contexts
- Same concepts can be expressed in many different way
- Semantic associations shift over time - like word ”cloud”
- Regional and cultural differences affect language usage

Synthetic data can both help and hurt

- LLMs may need more data than is naturally available
- Naturally occurring data might be of insufficient quality
- Synthetic data can augment the training dataset
- It is generated by other AI models
- Synthetic data can inherit errors and biases
- It can even degrade the model’s performance

Examples of Deliberate AI hallucinations

Deliberate AI Hallucinations

- I asked several AI models for deliberate hallucinations
- Hallucinations on free topics were mostly funny and creative
- Hallucinate technical or scientific facts were very obvious
- LLM needed several iterations on the topic
- After it results become more interesting and harder to detect

Free Deliberate AI Hallucinations

- *”In 1912, a group of mathematicians in Vienna discovered a formula that accurately predicts the taste of all types of cheese based on the angle of sunlight at noon on the summer solstice.”*

- *”The 18th-century composer Ludwig van Beethoven collaborated with Pablo Picasso to create a series of musical paintings that depict the sounds of thunderstorms in oil on canvas.”*

PostgreSQL Related AI Hallucinations

- *PostgreSQL's query planner uses reinforcement learning to continuously adapt and optimize query execution plans based on observed query patterns.*
- *"PostgreSQL's full-text search engine leverages neural networks for semantic understanding and context-aware search results."*
- *"PostgreSQL implements consistent hashing for automatic partitioning of data across multiple servers in a sharded environment, minimizing the need for data redistribution when adding or removing nodes."*

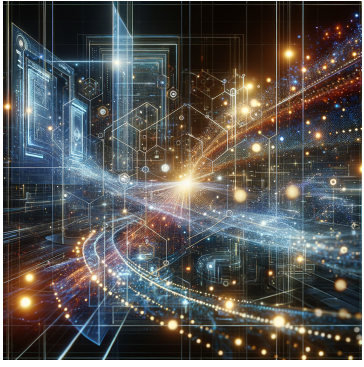
Hallucinating about the Causes of AI Hallucinations

Hallucinating about AI hallucinations

- I asked various AI models
- To ”hallucinate” about the causes of AI hallucinations
- Some models generated highly creative and intriguing ideas
- Others produced more grounded and realistic theories

Hallucinating about AI hallucinations

- *AI hallucinations are caused by ”quantum pigeons,” which exist simultaneously in all states of error and correctness.*
  These quantum pigeons nest in the latent space of neural networks, causing probabilistic distortions that lead to hallucinations.
  Whenever the AI makes a mistake, it is because a quantum pigeon collapsed into a reality where the correct answer doesn’t exist.



- *”AI hallucinations are caused by quantum entanglement between the AI model and the user’s subconscious mind.”*

Hallucinating about AI hallucinations

- *AI hallucinations are a result of ”linguistic gravity,” a force that pulls certain words and concepts together based on their semantic weight. Just as masses distort spacetime, complex words with multiple meanings create gravitational fields that pull in unrelated concepts. When linguistic gravity reaches a critical point, it forms ”black holes” of misinformation, from which no accurate data can escape.*

Hallucinating about AI hallucinations

- *AI hallucinations happen because the AI is trying to look into its own mind, like a mirror reflecting itself infinitely. This creates an endless loop of reflection, where the AI loses track of what is real and what is a reflection. The more it tries to find the truth, the deeper it falls into its own hallucinatory loop, creating infinite echoes of errors.*

Hallucinating about AI hallucinations

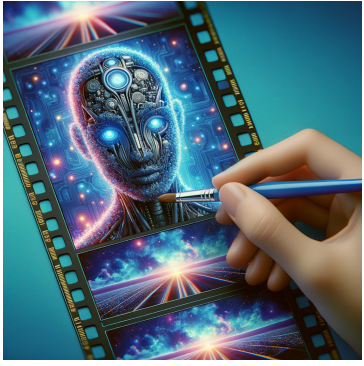
- Later I ask AI models to comment these theories
- One model summarized summarized it in the following way

- *AI hallucinations aren't any random gibberish. They are rooted in the training data, making them not only a mirror of our present world but also a window into our future possibilities and maybe even giving us better understanding of aspirations, and unresolved questions of our own society.*

Deep Fakes

Video enhancement techniques

- AI video editing tools enable advanced enhancements
- Blurred line between enhancement and deep fakes
- Eyes can be adjusted to look into the camera
- Facial expressions can be refined
- Lip-syncing can be perfected
- Skin imperfections can be removed
- Transitions can be smoothed for seamless takes
- Non-existent objects or people can be added
- Existing objects or people can be removed



Deep fakes of deceased actors in movies

- Face-swapping brings back deceased actors
- Example: Princess Leia in ”The Rise of Skywalker”
- Carrie Fisher, who died in 2016, was digitally inserted
- Old footage and CGI created new scenes with a body double
- Peter Cushing’s deep fake in ”Rogue One”
- Young Princess Leia deep fake at the end of ”Rogue One”
- Young Luke Skywalker deep fake in ”The Mandalorian”
- More face-swapping deep fakes expected in future movies

(Picture from the article
Are Deep Fakes the Future of A-List Casting?)

Full Deep fake Movie

- ”The Lion King” (2019) is a full deep fake movie
- A photorealistic remake of the 1994 animated film
- Animals designed using photo references
- Animation based on real animal movements
- AI mimicked animal behavior for realism
- Only the sunrise scene at the beginning is real footage

(Picture from the article Pumbaa
The Lion King Fandom Wiki)

How to mitigate AI hallucinations

How to mitigate AI hallucinations

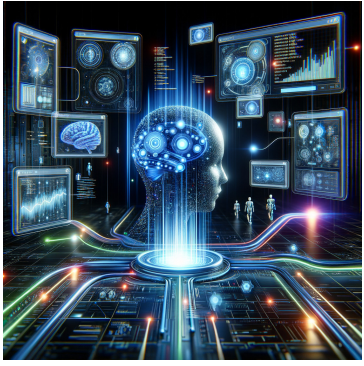
- Observed in all existing AI models
- Older models (GPT-3, 3.5-turbo) were more prone to them
- Newer models (GPT-4) produce fewer hallucinations
- Recent articles and papers discuss mitigation strategies
- Can reduce hallucinations but not eliminate them
- Only a few methods are truly useful and effective

How to mitigate AI hallucinations

- Use the most reputable and advanced AI model
- Provide detailed and specific instructions in the prompt
- Clearly define the AI's role for the given task
- Limit output to categories/topics relevant to the task
- Provide high-quality examples as references
- Exclude certain topics/ categories explicitly
- Verify important facts in the output for accuracy
- "Temperature" parameter closer to 0 can reduce hallucinations

YMYL (Your Money or Your Life) topics

- Caution with AI recommendations on YMYL topics
- Double-check AI information
- Especially health, finance, safety, or legal advices
- Web is already full of misinformation and fakes
- AI hallucinations add new misinformation
- Many articles are copy-pasted without verification
- ”Helpful” YouTube videos can be deep fakes
- Always consult with a human expert in the relevant field



THANK YOU

- Questions?

NetApp

