EDUCAITION
2026

משרד החינוך
מינהל חדשנות
וטכנולוגיה

Hallucinations and Hocus Pocus

Why do AIs lie?

Dr . Michael J. Jabbour
AI Innovation Officer , Microsoft

database

digital brain?

efficiency tool?

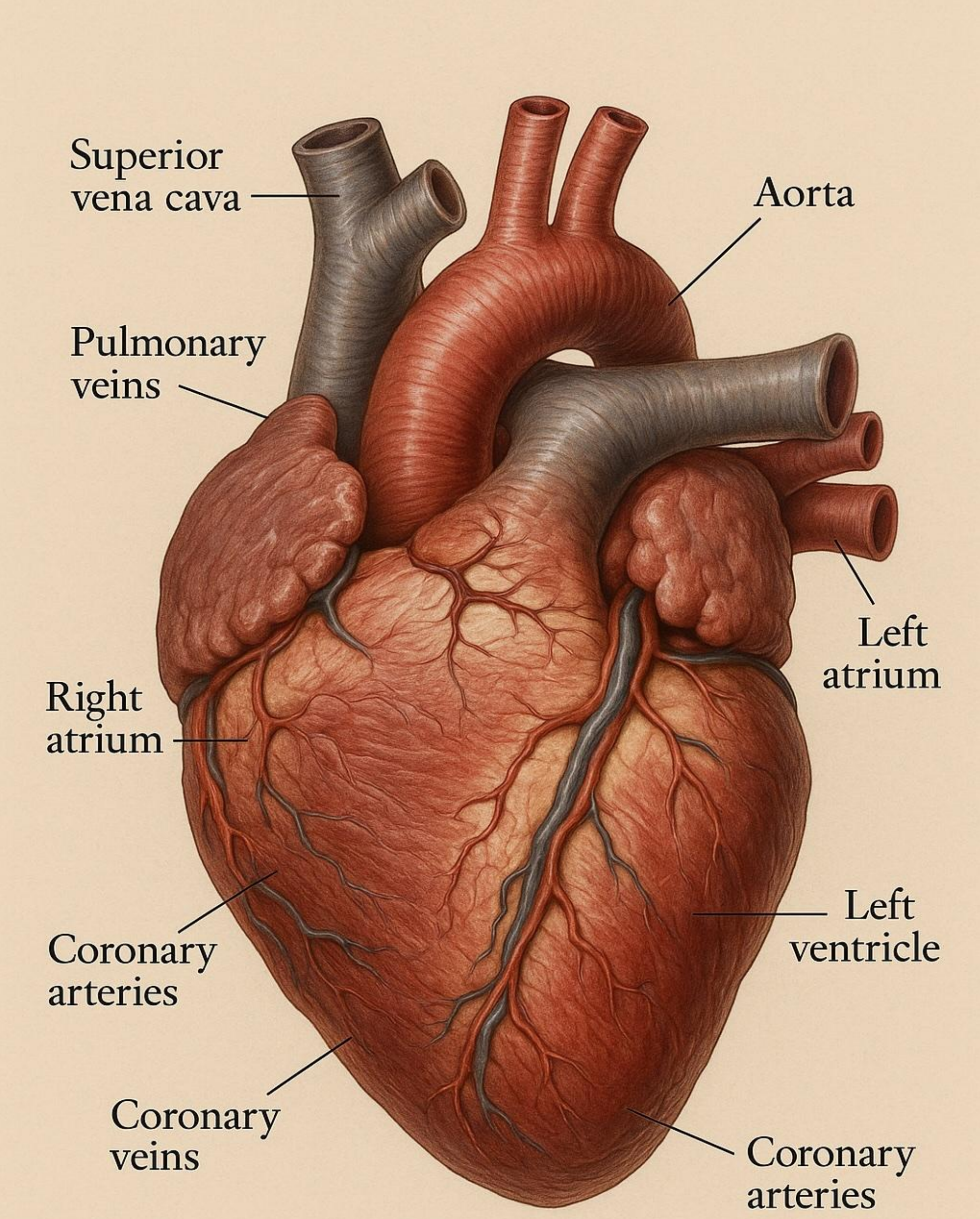
What is AI?

search agent?

educator

reclaiming my stride

Where do WE fit in all of this?



GPT-5 Prompt: *Imagine yourself as a human heart. Visualize all the biology and physiology. Now show me a self portrait. Make it more realistic. Label the anatomy.*
9/12/25 – Michael J. Jabbour

Microsoft

AI ≠ search engine

The Spectrum of Hallucinations

- **Training Data Errors:** The web contains outdated facts, false claims, and contradictions that become baked into models.
- **Confabulations:** Like the human mind, AI fills gaps with plausible but invented details when memory fails.
- **Creative Generation:** While creativity and hallucination both involve generation, research shows they are distinct—creativity is deliberate; hallucination is unintended error.

Alansari, A. & Luqman, H. (2025). *Large Language Models Hallucination: A Comprehensive Survey*. arXiv. Retrieved from https://arxiv.org/html/2510.06265v2

Understanding Hallucinations in AI

The past and future of flying machines

- Scroll back 450 years ago to the times of Davinci.
- He had birds, hammers, and humans available to him.
- His creative conceptualization of the aerial screw or flying machines - the primitive of planes - was effectively a hallucinations.

Pathfinders Design and Technology. (n.d.). *Leonardo da Vinci's Aerial Screw (Helicopter)*.
Retrieved from http://www.pathfindersdesignandtechnology.com/portfolio-item/da-vinci-helicopter/

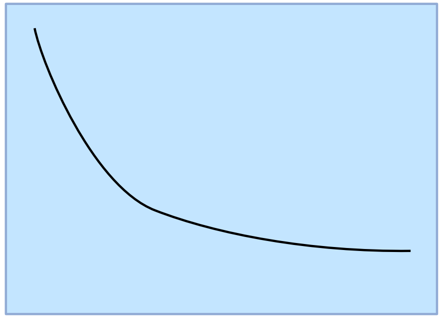
GPT Trends

Key Developments

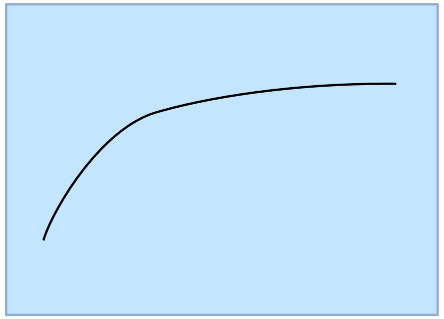
- **Model Optimization:** New architectures for efficiency (parameters, speed, specialized tasks).[1]
- **Data & Fine-tuning:** Larger, better-curated datasets drive improvements. [2]
- **Prompt Engineering:** Refined techniques and tools for clearer, more effective inputs. [3]
- **Orchestration:** Framework advancements for LLM deployment and complex integration. [4]
- **Multimodal Growth:** Trend towards models handling multiple data types (text, image, etc.)

Challenges & Speculation

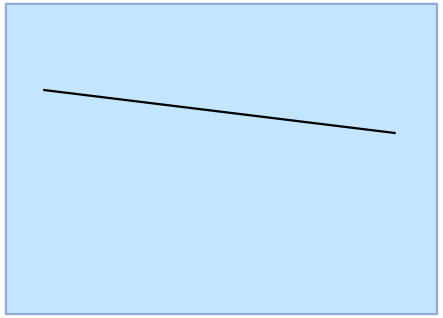
- **Bias/Safety:** Mitigating remains complex, likely slower advancement
- **Dataset Dependency:** Data quality is a persistent issue [5]
- **Self-Analysis:** Future LLMs can self-improve [6] and have internal bias detection. [7]



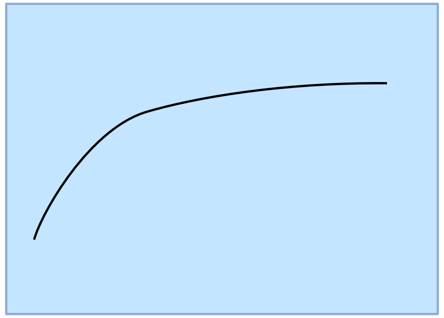
Hallucinations



Reliability



Bias



Transparency

Sources:

1. Challenges and Applications of Large Language Models
2. Understanding the Capabilities, Limitations, and Societal Impacts of Large Language Models
3. Foundation and large language models: fundamentals, challenges, opportunities, and social impacts
4. Challenges and Contributing Factors in the Utilization of Large Language Models (LLMs)
5. Open-Sourced Training Datasets for Large Language Models (LLMs) (kili-technology.com)
6. [2210.11610] Large Language Models Can Self-Improve (arxiv.org)
7. Adaptation with Self-Evaluation to Improve Selective Prediction in LLMs

Not a Bug, But a Feature?

The line between creativity and hallucination is blurrier than we think.

- Anthropic's research found **internal "circuits" that inhibit responses when knowledge is uncertain** - hallucinations occur when this fails.
- Human brains confabulate too - **we fill gaps in memory with plausible fiction** and rarely notice.
- The generative capacity that produces text can sometimes invent false citations—though research distinguishes intentional creativity from unintended fabrication.

Anthropic. (2025). *On the Biology of a Large Language Model.* Via Lakera Blog. Retrieved from https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models

Why Do AIs “Lie”?

Optimized for Confidence, Not Truth

- LLMs are trained to produce the most statistically likely answer, not to assess their own confidence.
- Training and evaluation systems reward guessing over admitting uncertainty—confident answers score higher.
- Without reward for “I don’t know,” models default to generating plausible-sounding responses.

Kalai, A.T. & Nachum, O. (2025). *Why Language Models Hallucinate.* OpenAI. Retrieved from https://openai.com/index/why-language-models-hallucinate/

The Intelligence Age Is Here —

What Can *You* See?

What Will *You* Shape?