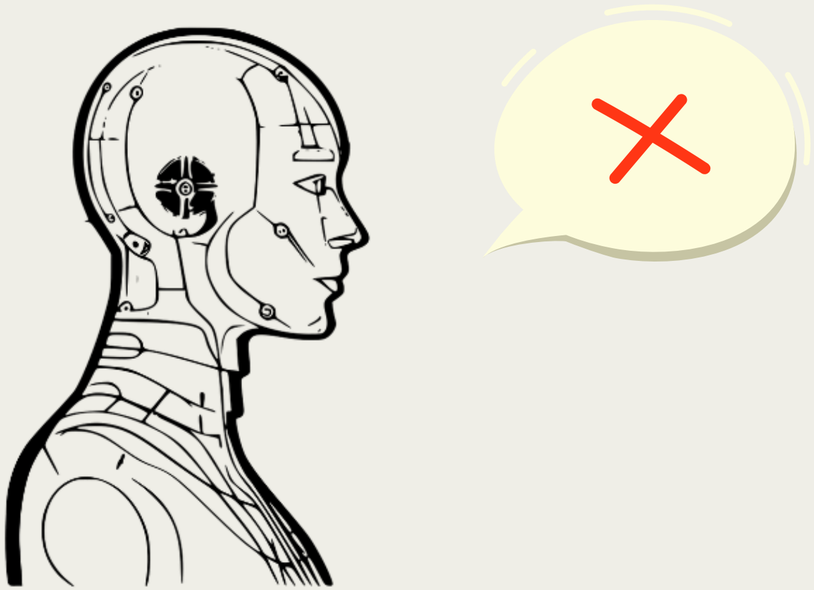
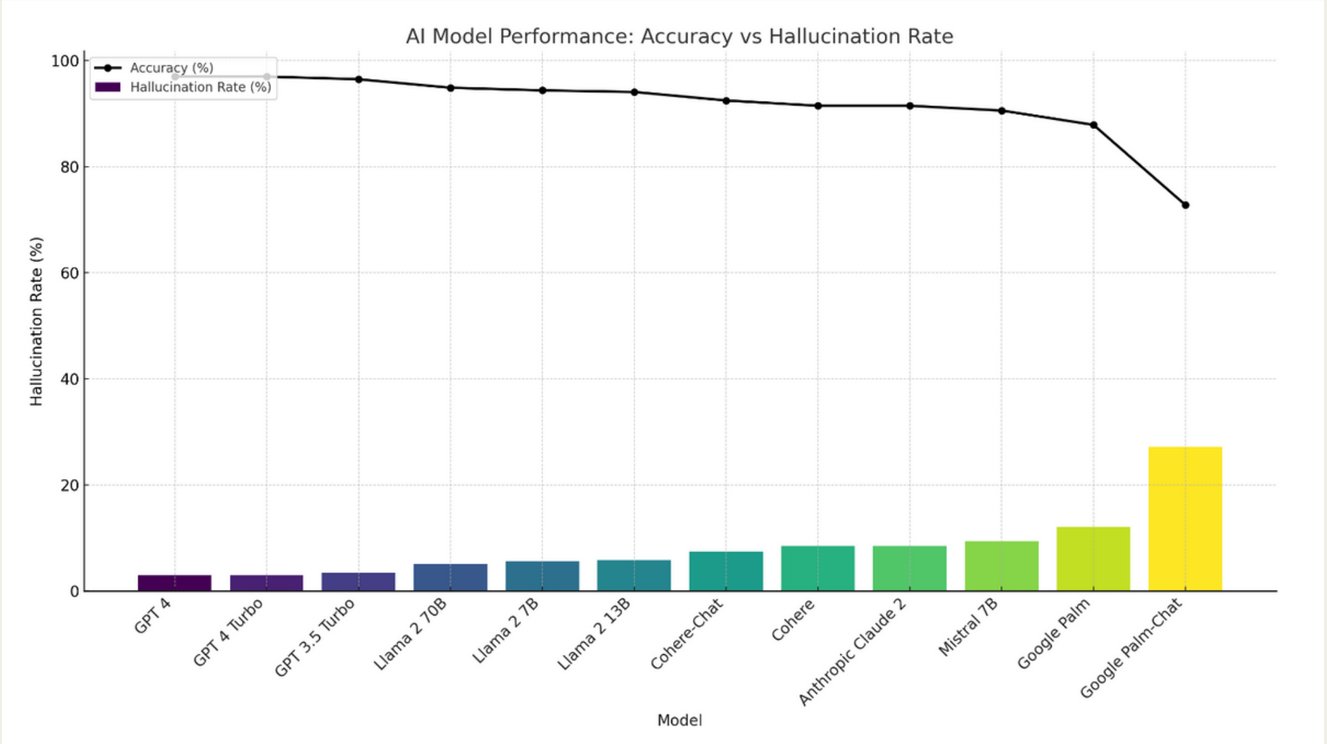
AI Hallucinations

Tina Cao, Lu Liu, Xinwei Qiao, Wanxin Luo, Yao Xie



INTRODUCTION

- Increasing use of GenAI
  - 40% quicker for writing tasks
  - 30% quicker for coding tasks
- What is GenAI hallucination?
- Hallucination in ChatGPT
- Current AI model performance



IMPACTS

- Individuals
  - Legal, ethical, or even life-related challenges
    - Examples: Law/Medical Area
- Enterprise
  - Erode trust and credibility
- Society
  - Spread and propagation of misinformation

AI ETHICS

- Misinformation (Trust & Reliability):

  AI-generated hallucinations can blur the lines between fact and fiction.

- Bias:

  AI systems could mirror biases inherent in their training data.

- Privacy

  Content that might not align with reality, presents challenges regarding individuals' rights and the ethical use of their data.

  ○

CAUSES

DATA

- Insufficient or low-quality training data
- Model Degradation

MODEL

- Inadequate training -> overfitting
- Adversarial attack

PREVENTIONS

Data-Centric Approaches

- Use Diverse and High-Quality Training Data

Model-Centric Approaches

- Implement Guardrails
- Retrieval Augmented Generation (RAG)

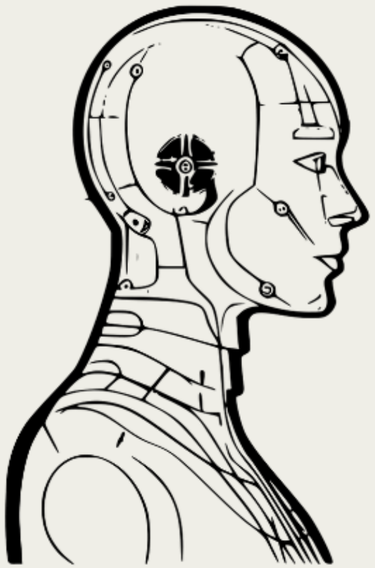
User-Centric Approaches

- Prompt Engineering
- Human-In-The-Loop Testing and Validation

CONCLUSION

- Acknowledging the current challenges posed by AI hallucinations.
- Ongoing R&D aimed at mitigating hallucinations with better algorithms and data.
- Collaborative progress in AI research, development, and ethics is essential.
- Optimistic outlook for advanced, reliable AI applications transforming the future.

Thank you!



References

- https://medium.com/@KarlHavard/how-can-hallucinations-be-prevented-in-genai-d161860d79ce
- https://www.pinecone.io/learn/options-for-solving-hallucinations-in-generative-ai/
- https://www.forbes.com/sites/forbestechcouncil/2023/09/06/preventing-hallucinations-in-generative-artificial-intelligence/?sh=29249e9e7340
- https://medium.com/@bijit211987/advanced-prompt-engineering-for-reducing-hallucination-bb2c8ce62fc6#:~:text=RAG%20reduces%20hallucination%20by%20ensuring,system%20can%20honestly%20admit%20ignorance.
- https://www.linkedin.com/pulse/understanding-hallucinations-artificial-intelligence-causes-baxter-fr0jc/
- https://zapier.com/blog/ai-hallucinations/
- https://www.nttdata.com/global/en/-/media/nttdataglobal/1_files/insights/generative-ai/all-hallucinations-are-not-bad---acknowledging-gen-ais-constraints-and-benefits.pdf?rev=3f9e832e618b4e79807e0b15d696933f
- https://luiza.medium.com/ai-hallucinations-privacy-a-reputational-harm-nightmare-90033a7c04ec
- https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/
- https://aimresearch.co/council-posts/ethical-considerations-in-gen-ai-hallucinations-balancing-creativity-and-accuracy