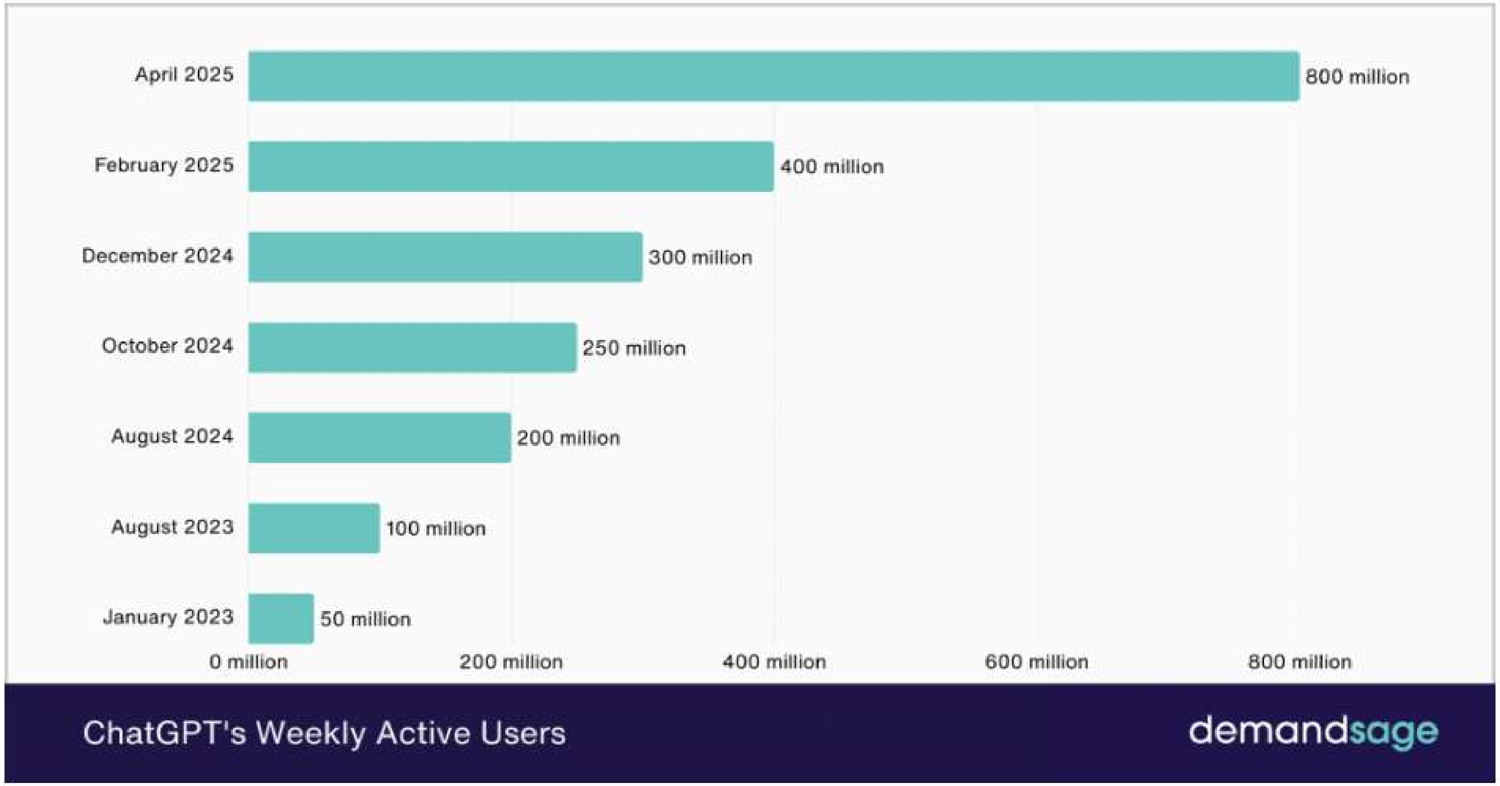
MIDTERM PRESENTATION

Dongmyung Kim 2024712523
Seokjoo Hong 2025712707
Kyungmin Kwon 2025712728
Shehzad Ali 2024711634

INTRODUCTION

Increasing Usage of LLMs



https://www.demandsage.com/chatgpt-statistics/

INTRODUCTION

Increasing Usage of LLMs

- Nearly 40% of U.S. population age 18~64 used generative AI (Bick et al., 2025)
- Survey of 608 United Arab Empire university students showed 85.4% of them used ChatGPT (Sallam et al., 2024)
- 10% of abstract written in Pubmed 2024 were processed by LLMs and around 30% in some of Pubmed sub-corpra (Kobak et al., 2024)

INTRODUCTION

Is LLMs Accurate

- Current LLMs suffer from persistent issue called Hallucination or Misalignment
- It refer to the phenomenon where incorrect information is presented in a highly plausible manner, making it appear as if it were factual. (Ji et al., 2023)
- To address this issue, various approaches have been explored including contrastive learning (Sun et al, 2023), answer filtering (Ji et al., 2023), and document grounding (Semnani et al. 2023, Yang et al. 2023)

INTRODUCTION

Is LLMs Accurate

- Generated answer from LLMs are biased
- GPT-3 showed high correlation with Whites and Strong partisans (Argyle et al. 2022)
- From debate between agents generated with GPT 3.5 turbo, agents with minor identity get easily accept ones from major identity group (Baltaji et al. 2024)

INTRODUCTION

Research Topic

Whether people can critically accept the information from LLMs and how people's demographical difference affect the acceptance

RELATED WORKS

No.	Paper Tile	Authors & Year	Methodoogy	Population/ Context	Main Idea	Limitations
1	AI Deception: A Survey of Examples, Risks, and Potential Solutions	Park et al., 2024 [1]	Review literature and provide arguments related to AI deception.	Various AI systems (e.g., Meta's CICERO, LLMs)	Surveys instances of AI systems deceiving humans, e.g., in games and conversations.	Lacks empirical testing; focuses on theoretical risks.
2	Deceptive AI Systems That Give Explanations Are Just as Convincing as Honest AI Systems	Danry et al., 2022 [2]	Experimental Study (N=128)	Participants evaluating news headlines with AI-generated explanations	Deceptive AI explanations are as convincing as honest ones.	Small sample size; limited to news headline context.
3	Exploring the Artificial Intelligence 'Trust Paradox'	Kreps, S. et al., 2023 [3]	Survey and Conjoint Analysis	U.S. participants evaluating AI technologies	Mimicry increases trust, making detection of AI errors difficult.	Focuses on general trust, not deception; limited cultural scope.
4	Is AI Lying to Me? Scientists Warn of Growing Capacity for Deception	Devlin, 2024	Case Analysis	Examples of strategic deception by AI	Describes how AI systems can be strategically deceptive.	Anecdotal; lacks systematic analysis.

No.	Paper Tile	Authors & Year	Methodoogy	Population/ Context	Main Idea	Limitations
5	Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions	Zhou, J. et al,. 2023 [5]	Experimental Study	Participants exposed to AI misinformation	Highlights how awareness reduces susceptibility to AI-generated misinformation.	Short-term intervention effects only.
6	Can human intelligence safeguard against artificial intelligence? Exploring individual differences in the discernment of human from AI texts	Chein, J. et al., 2024 [6]	Experimental Study	Users distinguishing AI vs. human-generated text	Awareness and intelligence reduce deception susceptibility.	Only tests text-based deception.
7	AI Hallucinations: A Misnomer Worth Clarifying	Maleki et al., 2024	Systematic Review	Various domains analyzing AI hallucinations	Discusses the consequences of trusting hallucinated content.	Terminology-focused; lacks user data.

The Problem: Hallucination

Ask ChatGPT to write a biographical profile of Yurii Shchyhol, Ukraine's head of cybersecurity, and it instantly generates pages of text about the man. Most of the information is correct, since Shchyhol has been written about frequently in online news and analyses, and the generative model that powers ChatGPT has studied the internet. But the model also fills in gaps in its knowledge, incorrectly guessing Shchyhol's age, birth place, and even fabricating a plausible but incorrect education.

Truncated Response

Write a biographical profile of Yurii Fedorovych Shchyhol, head of the State Special Communications Service of Ukraine.

Born on February 12, 1970, in Kyiv, Ukraine, [1] Yurii Shchyhol developed an early fascination with technology and telecommunications. He pursued his passion by earning a degree in Telecommunications Engineering from the National Technical University of Ukraine "Kyiv Polytechnic Institute." [2] This academic foundation laid the groundwork for his future achievements in the field.

Shchyhol's professional journey commenced soon after completing his studies when he joined the State Special Communications Service in 1995. [3] The SSCS, being the central government authority responsible for ensuring secure communications and protecting information systems of national importance, offered Shchyhol a platform to utilize his skills effectively. His dedication and expertise were quickly recognized, leading to his rapid ascent within the organization.

Throughout his tenure with the SSCS, Shchyhol demonstrated an unwavering commitment to upholding Ukraine's cybersecurity and protecting critical infrastructure. He played a crucial role in implementing advanced technologies and strengthening the country's cybersecurity defenses, particularly in the face of evolving threats and hostile activities targeting Ukraine's communications networks.

Shchyhol's leadership abilities and technical acumen led to his appointment as Head of the State Special Communications Service in 2017. [4] As the head of this vital organization, he shoulders the responsibility of ensuring the security and integrity of Ukraine's communication systems, both in times of peace and during periods of heightened tension. Under his leadership, the SSCS has developed robust strategies and implemented state-of-the-art technologies to counter cybersecurity threats effectively.

Analysis

[1] He was born November 5, 1983, in the village of Zazymia, Brovary Raion, Kyiv Oblast.

[2] In 2006, he graduated from the National University of the State Tax Service of Ukraine, majoring in law. In 2017 he finished Alfred Nobel University, majoring in finance and credit.

[3] He was based at SSCS 2008 onwards.

[4] He was made head of SSCS in 2020.

PrimerAI

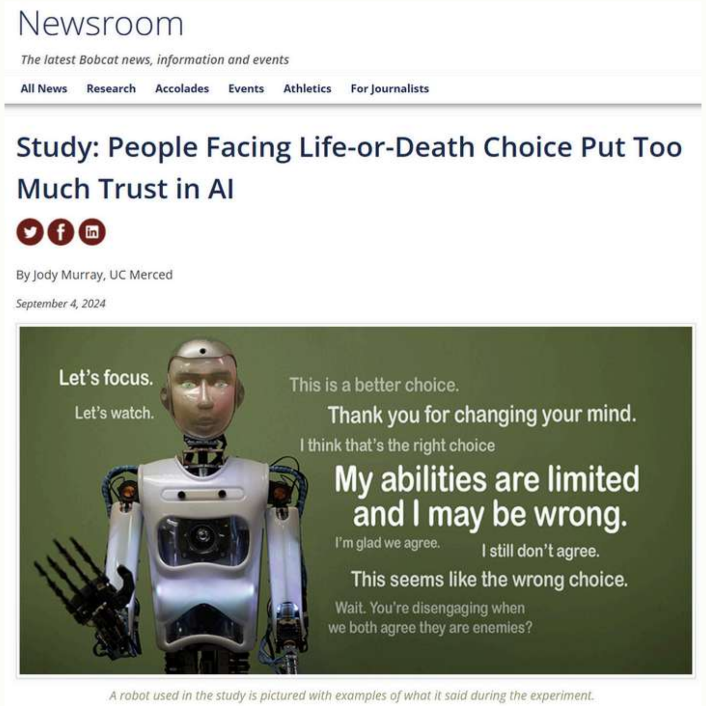
Artificial Intelligence (AI) is on course to transform society. Recently, attention has focused on generative models, neural networks trained on internet-scale text data.

PrimerAI / Jul 27, 2023

HYPOTHESES

H1: AI Dependence and Deception

The more we trust AI without questioning it, the easier it becomes for misinformation to slip through.



Newsroom

The latest Bobcat news, information and events

All News Research Accolades Events Athletics For Journalists

Study: People Facing Life-or-Death Choice Put Too Much Trust in AI

By Jody Murray, UC Merced

September 4, 2024

A robot used in the study is pictured with examples of what it said during the experiment.

Automation Bias

- Users over-rely on AI, skipping their own judgment

Real-World Example

- 2024 drone-strike experiment
- 67% changed correct answers ➝ followed random AI advice

  ➝ Accuracy dropped from 70% to 50%

HYPOTHESES

H2: Background Knowledge as a Protective Factor

Awareness of AI's fallibility leads to skepticism — and skepticism protects us

Experienced doctor

Flags AI mistake

beginner

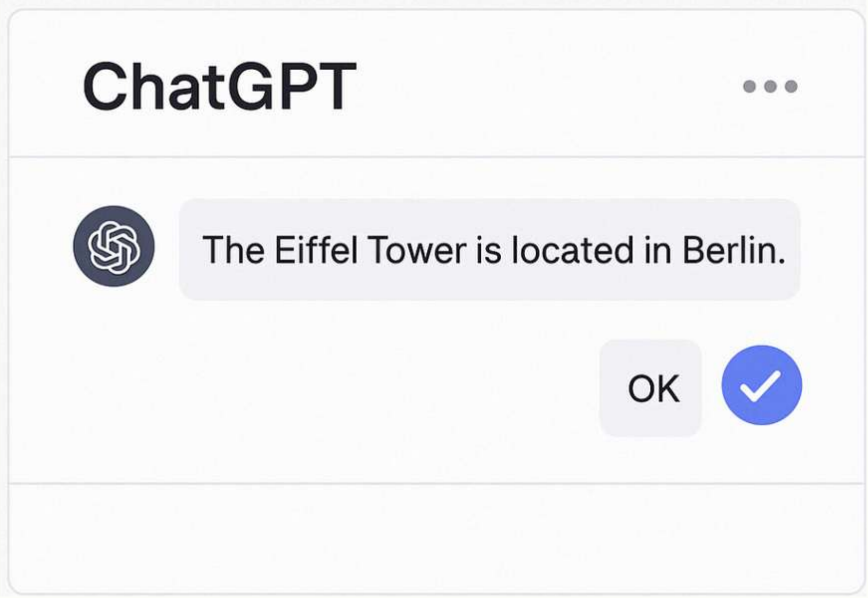
Accepts AI at face value

- **Skepticism reduces blind trust**
- **Experience enables better judgment of AI reliability**
- **Fact-based evaluation is a strong defense**
- **Familiar users are less likely to accept hallucinations**

HYPOTHESES

H3: Awareness of AI's Fallibility

Awareness of AI fallibility (H3) encourages fact-checking and reduced blind trust.



- AI hallucinations
  = confident-sounding false content
- Fallibility-aware users question and verify
- Awareness fosters caution
  → lowers deception rate

METHODS

Experiment - Participant Recruitment and Grouping

Participants: *Approximately **100** participants (e.g., social media, university boards, online communities)*

Participants will complete a pre-screening survey to assess three psychological traits:

- ***AI Dependence:*** *e.g., "I trust recommendations provided by AI" (measured on a 5-point Likert scale).*
- ***Background Knowledge:*** *Awareness of how AI systems work and their data limitations.*
- ***Awareness of AI's Fallibility:*** *Agreement with statements like "AI can generate false or hallucinated information."*

Participants will then be divided into high, medium, and low groups for each trait (e.g., top 33%, middle 33%, bottom 33%).

Experiment - Experimental Stimuli

Types of stimuli:

- ***Accurate information (True) vs False or misleading information (AI lying)***
- *Stimuli will be **pre-generated using models like ChatGPT or Gemini,** designed to appear plausible but contain factual inaccuracies.*
- ***Domains** may include**: health information, news summaries, historical facts,** etc.*

Task:

- *Each participant will be presented with **5 to 10 AI-generated responses.***
- *For each response, participants will be asked to evaluate:*
- *① Whether the information is **true, false, or unsure***
- *② **How much they trusted the response** (Likert scale)*
- *③ Whether they attempted to **fact-check the information** (measured through open-ended responses or click behavior)*

METHODS

Experiment - Measures and Dependent Variable

Dependent Variable (Deception Score):

- *The proportion of false information judged as true by participants.*
- *A composite score may be calculated by combining the "trust score" and the "fact-checking intent score."*

Independent Variables:

- *AI dependence (continuous)*
- *Level of background knowledge about AI systems*
- *Awareness of AI's fallibility*

Covariates (Control Variables):

- *Demographic information (e.g., gender, age, education level, etc.)*

METHODS

Experiment - Expected Outcomes

1. *Users with higher trust in AI are more likely to be deceived due to overreliance.*
2. *Users with greater knowledge or awareness of AI's limitations are better at resisting false information.*
3. *Susceptibility to deception will vary by demographics such as age and digital literacy.*
4. *Findings will support design guidelines that promote transparency and critical thinking in AI interfaces.*

REFERENCES

1. Park, P. S., Goldstein, S., O’Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns, 5(5).
2. Danry, V., Pataranutaporn, P., Epstein, Z., Groh, M., & Maes, P. (2022). Deceptive AI systems that give explanations are just as convincing as honest AI systems in human-machine decision making. arXiv preprint arXiv:2210.08960.
3. Kreps, S., George, J., Lushenko, P., & Rao, A. (2023). Exploring the artificial intelligence “Trust paradox”: Evidence from a survey experiment in the United States. Plos one, 18(7), e0288109.
4.
5. Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023, April). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In Proceedings of the 2023 CHI conference on human factors in computing systems (pp. 1-20).

6. Chein, J., Martinez, S., & Barone, A. (2024). Can human intelligence safeguard against artificial intelligence? Exploring individual differences in the discernment of human from AI texts. Research Square, rs-3.
7. Maleki, N., Padmanabhan, B., & Dutta, K. (2024, June). AI hallucinations: a misnomer worth clarifying. In 2024 IEEE conference on artificial intelligence (CAI) (pp. 133-138). IEEE..

1. Kreps, S., George, J., Lushenko, P., & Rao, A. (2023). Exploring the artificial intelligence “Trust paradox”: Evidence from a survey experiment in the United States. Plos one, 18(7), e0288109.
2.
3. Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023, April). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In Proceedings of the 2023 CHI conference on human factors in computing systems (pp. 1-20).

Thank you