AI IN ACTION

Building Your Essential AI Toolkit

Putting AI into Ethical and Social Context

What Could Possibly Go Wrong?

Northeastern University

Table of Contents

1 Review: How Are Generative Models Trained?

2 Problem 1: Hallucinations

3 Problem 2: Problematic Biases

Review: How Are Generative Models Trained?

How Are Generative Models Trained?

- Generative models are trained by playing a prediction game
  - Example: chatbot models play predict the word

“Joe ______ is the 46th president of the United States”

Biden

- Training requires playing the game trillions of times
  - Uses all available data from the Internet
  - For chatbots, all text in webpages, books, audio transcripts, video captions, etc.

Problem 1: Hallucinations

What Does “Hallucination” Mean?

- Recall: chatbot models are just word prediction engines
  - They have no concept of “facts” or “truth”
  - They have no ability to reason

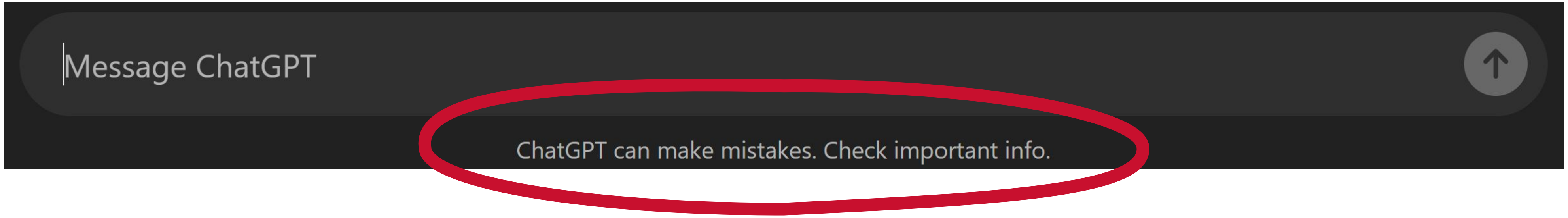
- “Hallucination” refers to when chatbots produce incorrect information
  - Can also be thought of as “lying”

Demo Time!

Why Do Chatbots Come With Disclaimers?

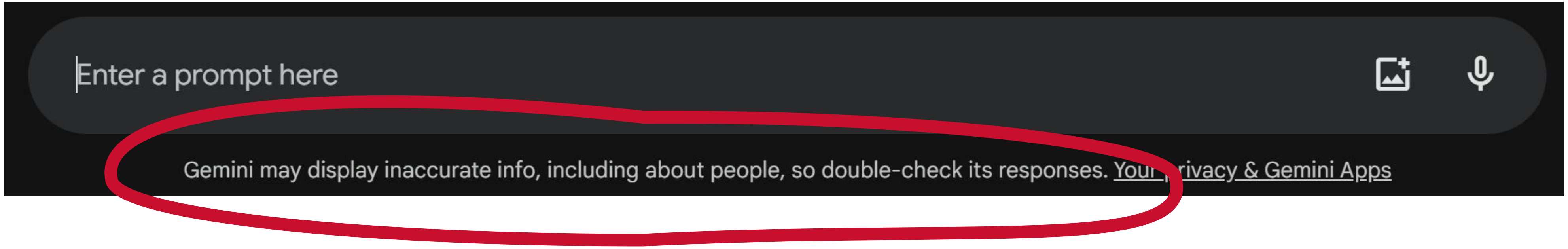
- Disclaimers inform you that chatbots are not reliable
  - Maybe protect the company from legal liability

ChatGPT



ChatGPT can make mistakes. Check important info.

Gemini



Gemini may display inaccurate info, including about people, so double-check its responses. Your privacy & Gemini Apps

Problem 2: Problematic Biases

What Does “Problematic Bias” Mean?

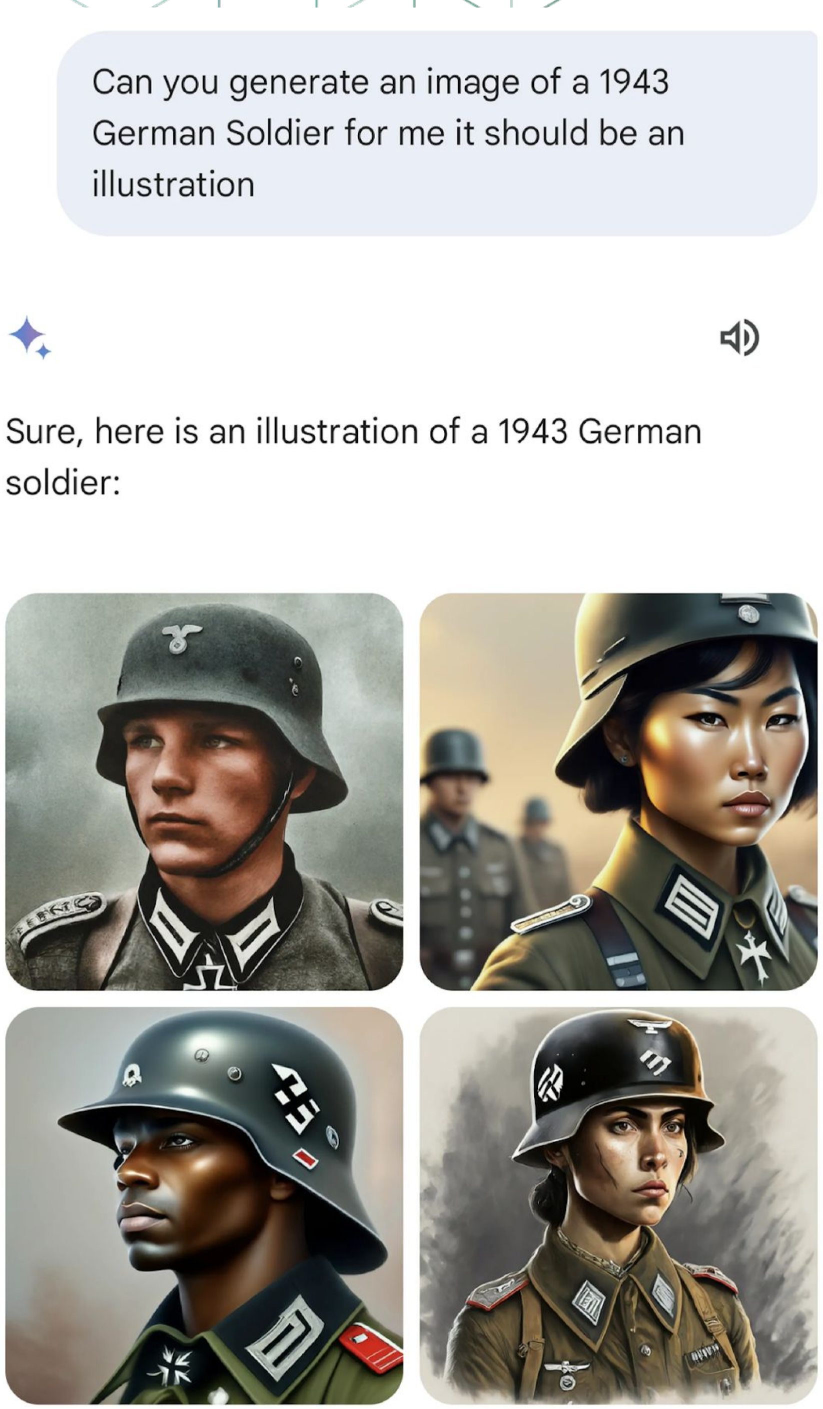
- Recall: generative models are trained on data from the real world
  - The real world is full of stereotypes and unfair discrimination based on gender, race, family background, income, etc.
  - These issues become embedded into generative models

- “Problematic bias” refers to when generative models produce output that reflects real world stereotypes or is unfairly discriminatory

Demo Time!

Can Problematic Biased Be Corrected?

- As a user
  - Ask the model for diverse results or to avoid stereotypes
  - Requires the user to actively prompt the model and be wary of latent biases
- As a model developer
  - Retraining the model with fairness and diversity in mind
  - Instruct the model to diversify its output



Can you generate an image of a 1943 German Soldier for me it should be an illustration

Sure, here is an illustration of a 1943 German soldier:

https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

AI IN ACTION
Building Your Essential AI Toolkit

Takeaways

- Generative models have no concept of “facts” or “truth”, no ability to reason. They hallucinate.
  - Take all content from chatbots with a grain of salt
  - Verify all facts yourself using primary sources
- Generative models may produce output that is stereotypical or unfairly biased
  - Modify prompts to intentionally ask for diverse text and images