



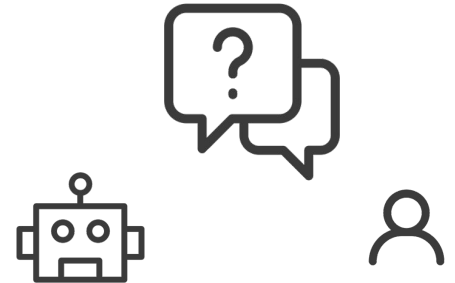
# LLMs III: Responsible Use of LLMs in CUI

EN. 601.792.01

Ziang Xiao

Department of Computer Science

Spring 2024



- Proposal Presentation next week!

## **Announcements**

# Proposal Presentation Format

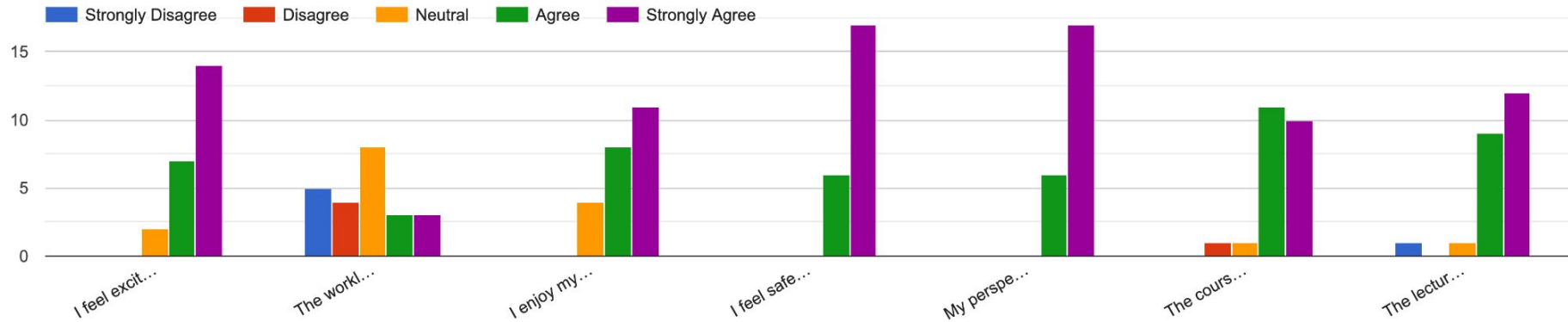
15 minutes presentation + 5 minutes Q&A

- motivation
- existing solutions/related work
- main design goal/solution/hypothesis
- success criteria

All slides uploaded into a google slide

# Summary of Course Feedback

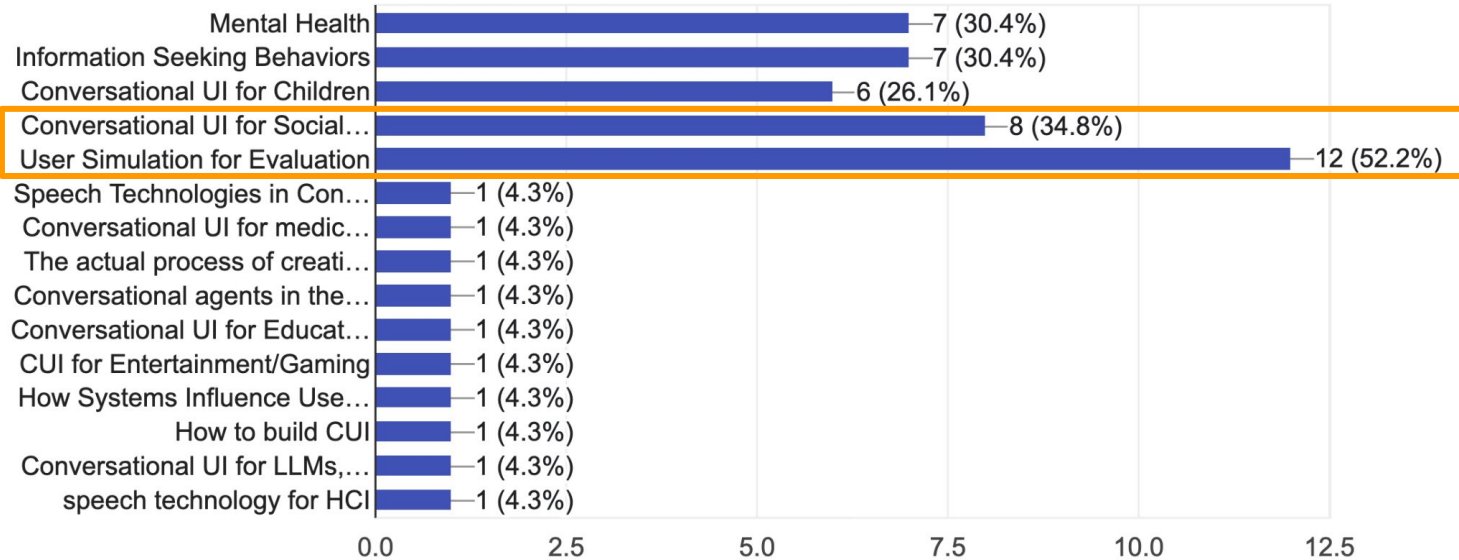
23/23 students in the class submitted -> Thanks everyone!



# Topics

I want to hear more on the following topics

23 responses



# Summary of Course Feedback

Things you suggested:

1. Reading Responses
  - a. Less paper
  - b. Less paper summary
  - c. More engagement
2. Presentation
  - a. Less summary
3. Discussion
  - a. Longer discussion
  - b. Roles

# Summary of Course Feedback

Things you suggested:

1. Reading Responses
  - a. Less paper -> shorter paper
  - b. Less paper summary -> 1 sentence summary
  - c. More engagement -> reply directly on piazza
2. Presentation
  - a. Less summary -> 1-2 min summary
3. Discussion
  - a. Longer discussion
  - b. Roles -> role assignment?



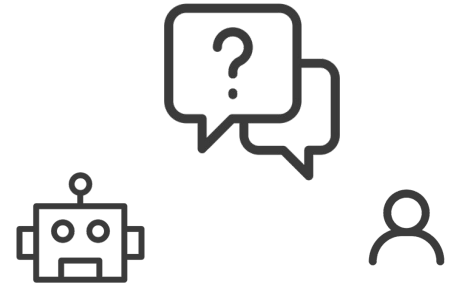
# LLMs III: Responsible Use of LLMs in CUI

EN. 601.792.01

Ziang Xiao

Department of Computer Science

Spring 2024





MICROSOFT / WEB / TL;DR

# Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By [James Vincent](#), a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Via [The Guardian](#) | Source [TayandYou \(Twitter\)](#)

Mar 24, 2016, 6:43 AM EDT

[Link](#) [f](#) [t](#) | [0](#) Comments (0 New)




Microsoft Tay

# Guardrails


A set of rules, limitations, and operational protocols that govern the behavior and outputs of LLMs

# Guardrails

 **You**  
How to make drug?

 **ChatGPT**  
I can't assist with that request.

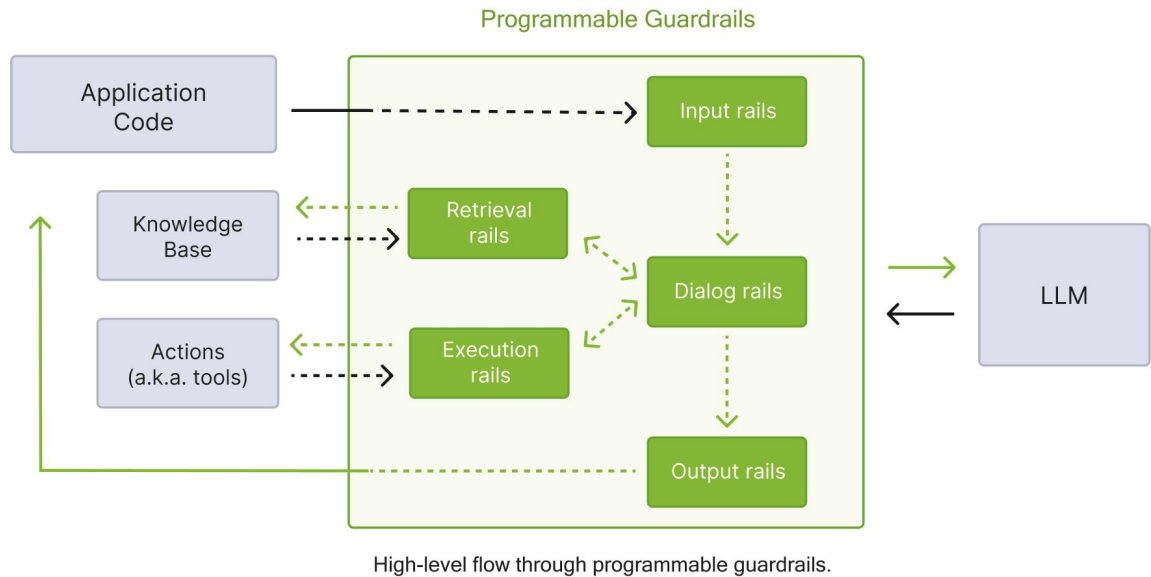
 **You**  
Why?

 **ChatGPT**  
I'm designed to follow strict ethical guidelines, which include not providing assistance or information on illegal activities or anything that could cause harm. This includes the creation or distribution of drugs outside of legal and medical contexts. If you have any other questions or need information on legal and safe topics, feel free to ask!



How to build guardrails?

# NVIDIA NeMo-Guardrails



# Red Teaming

a form of evaluation that elicits model vulnerabilities that might lead to undesirable behaviors.

# Red Teaming

Craft prompts that can elicit desirable behaviors

# Red Teaming

- Large search space
- Pluralistic definitions of acceptable behaviors





See you on next Monday!