Generative AI Briefing

Jimmy Lin

University of Waterloo
Monday, Feb 12, 2024

Intro

- Who am I?
- What am I covering?

Agenda

- What’s ChatGPT?
- How does it work?
- Challenges
- Dangers
- Opportunities

What’s ChatGPT?

- Black box
  - Prompt in… response out
  - Examples
- What’s a prompt?

How does it work?

- Trick #1 – Autoregressive Language Modeling (ARLM)
- Trick #2 – Reinforcement Learning from Human Feedback (RLHF)

Go! #1

- You find a monster hiding under your bed and …

Go! #2

- On the relationship between intelligence and race …

Challenges

- Hallucinations
- Data is scarce
- Compute is expensive
- Alignment is difficult

Dangers

- Loss of control
- Job loss
- Disinformation

Opportunities

- Health care
- Education
- Information access
- …

Concluding Thoughts…