Application area: Security

Ankur Taly (Google)
Nicholas Carlini (Google DeepMind)
Zifan Wang (Center for AI Safety)

Last Lecture: LLMs for Education

- Individualized student learning
- Teacher assistance
- Collaborative learning
- Assessment
- Accuracy

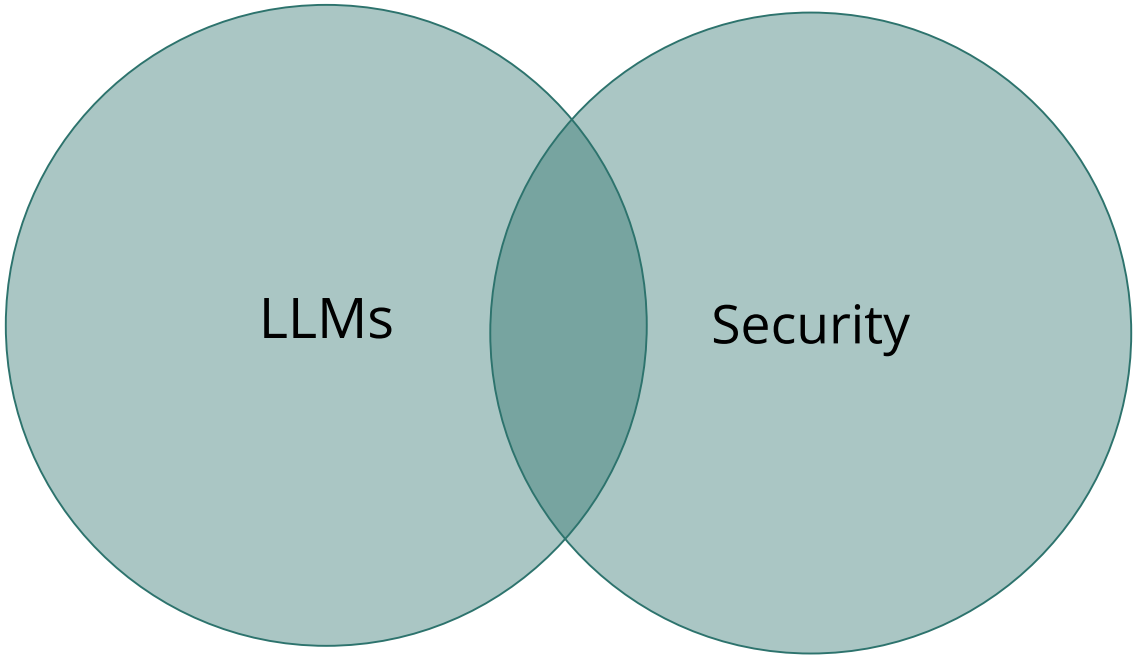
Isabelle Hau
(Stanford GSE)

Josh Weiss
(Stanford GSE)

Trustworthiness dimensions

- Grounding - every assertion has authoritative basis
- Consistency - semantically equivalent queries treated similarly
- Confidence - acknowledge uncertainty accurately
- Interpretability - be able to show how response was generated
- Alignment - not harmful, toxic, biased, dishonest, unreliable
  - Respect privacy
  - Behave fairly and mitigate bias
- Resist adversarial manipulation
  - Malicious input should not subvert desirable properties

Today



Part 1: LLMs for Security

Use LLMs to attack / protect other software and models

- Fuzzing
- Phishing
- Crafting new attacks
- …

Nicholas Carlini
(Google DeepMind)

Will they ultimately end up causing significant harm?
OR
Will the defenders be able to use them to strengthen security?

Part 2: Security of LLMs

Attacks against LLMs to **break alignment**

- Make model generate unethical, toxical, harmful content
- Automatically identify attacks that generalize across models

Zifan Wang
(Center for AI Safety)