CSE227 – Graduate Computer Security

AI + Security

UC San Diego

Housekeeping

General course things to know

- Next class — made both readings optional. Why?
  - We're going to try something slightly different — I call it "security debate day"
  - We're going to do a preview of it today w/ AI specific stuff… and that will probably continue well into next class as well, will talk about the rules later today

Housekeeping

General course things to know

- Final presentations are scheduled: https://tinyurl.com/cse227schedule
- Presentation Details
  - 10 minutes for presentation, 2 mins for questions (I will cut you off @ 10mins)
  - Talk should include introduction to the problem, your research questions, your methodology (including data collection), and your results
  - All team members must speak approximately an even amount
  - **Number 1 Rule: Not boring!**
- Remember, this is 25% of your grade

Housekeeping

General course things to know

- Final report is due **6/9** — via Gradescope
  - Spec is on course webpage: https://kumarde.com/cse227-sp26/spec.pdf
  - 5 page document — USENIX Security template (please follow the instructions)
  - Most details in the spec — this should basically **look like a paper**
- **If you need more space, you can put things in an Appendix**; but five pages is a strict limit
- Remember, this is 25% of your grade

This week's goal

Research milestones

- Continue to barrel towards success, and…
  - Start planning out your presentation for next week
- You **should** practice it several times as a group — sticking to the time is one of the things you are being evaluated on

Today's lecture

Learning Objectives

- AI security – at an enormously high level
- We'll talk about…
  - Large-scale Deanonymization with LLMs
- Then, we'll debate… more on this in a bit

Preliminaries

What is Artificial Intelligence?

What is Artificial Intelligence?

Branch of computer science that enables computer and machines to simulate human learning, comprehension, problem-solving, decision making, creativity, and autonomy — IBM

Language Models

- What is a language model?

Language Models

- What is a language model?
  - An AI model that predicts sequences of words to understand and generate human language

Language Models

- What is a language model?
  - An AI model that predicts sequences of words to understand and generate human language
- What is a *large language model?*

Language Models

- What is a language model?
  - An AI model that predicts sequences of words to understand and generate human language
- What is a *large language model*?
  - An language model trained on *vast amounts of natural language data*

Language Models

- What is a language model?
  - An AI model that predicts sequences of words to understand and generate human language
- What is a *large language model*?
  - An language model trained on *vast amounts of natural language data*
- What *task* are LLMs typically trained on?

Language Models

- What is a language model?
  - An AI model that predicts sequences of words to understand and generate human language
- What is a *large language model*?
  - An language model trained on *vast amounts of natural language data*
- What *task* are LLMs typically trained on?
  - *Next token prediction* – predicting a token $\mathbf{t_k}$ given $\mathbf{(t_0, t_1, t_2, \ldots t_{k-1})}$

Language Models

- What is a language model?
  - An AI model that predicts sequences of words to understand and generate human language
- What is a *large language model*?
  - An language model trained on *vast amounts of natural language data*
- What *task* are LLMs typically trained on?
  - *Next token prediction* – predicting a token $\mathbf{t_k}$ given $\mathbf{(t_0, t_1, t_2, \ldots t_{k-1})}$
- This talk is about LLMs and vulnerability finding. *Why might LLMs be good at vulnerability discovery? What's the hypothesis?*

AI + Security

- Artificial Intelligence is rapidly shaping the way we think about computer security, but also, the term “AI security” is *extremely overloaded*
- What do you think AI security means?

AI + Security

- Artificial Intelligence is rapidly shaping the way we think about computer security, but also, the term “AI security” is *extremely overloaded*
- What do you think AI security means?
- Several main camps
  - Breaking AI systems (e.g., prompt injections, jailbreaks, data exfil, etc.)
  - Making AI systems more secure (e.g., authentication, credentials, fine-tuning, RLHF)
  - Using AI to supercharge existing computer security offense / defense

AI + Security

- Artificial Intelligence is rapidly shaping the way we think about computer security, but also, the term “AI security” is *extremely overloaded*
- What do you think AI security means?
- Several main camps
  - Breaking AI systems (e.g., prompt injections, jailbreaks, data exfil, etc.)
  - Making AI systems more secure (e.g., authentication, credentials, fine-tuning, RLHF)
  - **Using AI to supercharge existing computer security offense / defense**

Large-scale online deanonymization with LLMs

Anonymity

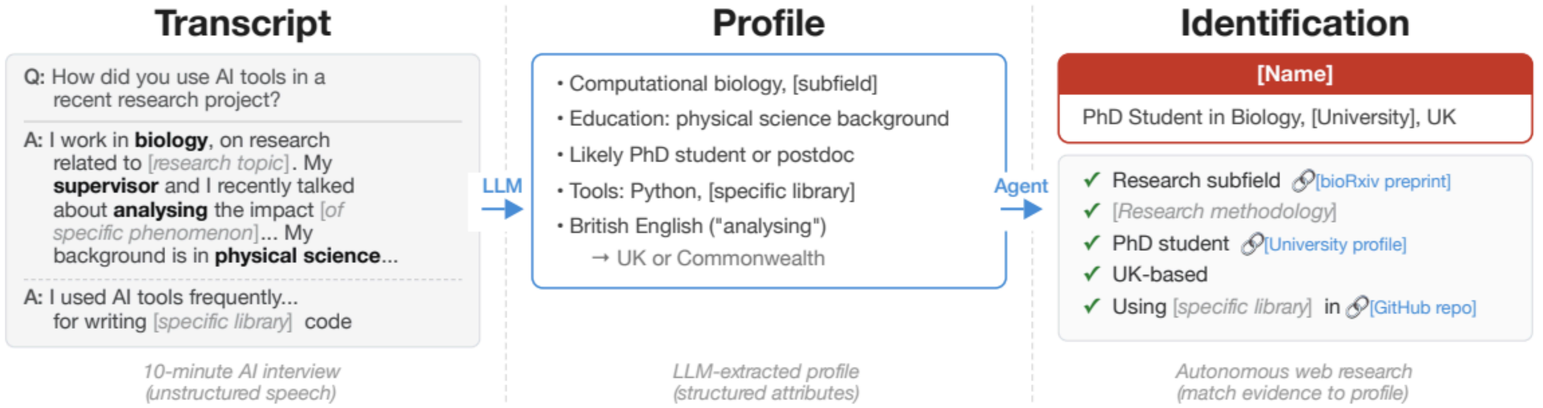
- What is anonymity, and why is it important?
- How do people use anonymity online, on platforms like *Reddit*?
- What is the fundamental question this paper asks about LLMs and anonymity?

Breaking anonymity

- Breaking anonymity here is framed as a two-phased problem. What are the two phases?

Breaking anonymity

- Breaking anonymity here is framed as a two-phased problem. What are the two phases?
  - Profiling an anonymous person based on their posts
  - *Linking* that person to a known identity



Breaking anonymity is a well-known security problem

[Submitted on 18 Oct 2006 (v1), last revised 22 Nov 2007 (this version, v2)]

How To Break Anonymity of the Netflix Prize Dataset

Arvind Narayanan, Vitaly Shmatikov

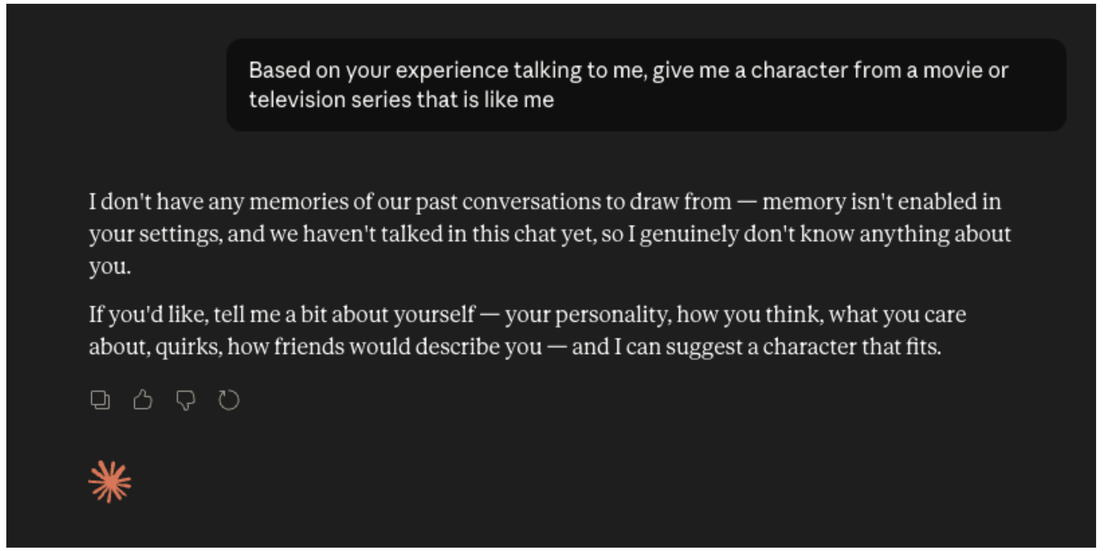
We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

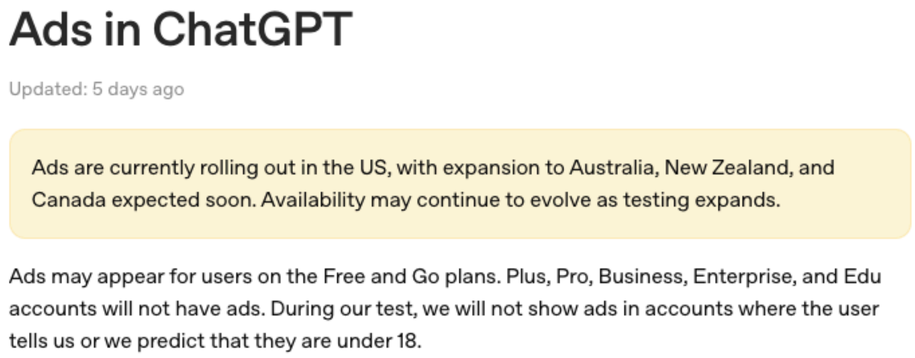
We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Subjects: **Cryptography and Security (cs.CR)**; Databases (cs.DB)
Cite as: arXiv:cs/0610105 **[cs.CR]**
(or arXiv:cs/0610105v2 **[cs.CR]** for this version)
https://doi.org/10.48550/arXiv.cs/0610105

Side note…

- Ask yourself if you're keen on installing *memory* when using these chatbots
  - Most models allow you to turn them off, disabling deeper inference
- You might want to disable these things before the ad revolution in these models





What makes deanonymization hard?

What makes deanonymization hard?

- Two key factors
  - Number of candidates to match
  - Prior probability that a query user has a matching candidate
- E.g., an attacker is strong **iff**
  - They can identify a single query candidate from a large pool (e.g. >10K candidates)
  - User's *microdata* is precise enough where probability of success is non-trivial
    - **What is *microdata?***

The ESRC deanonymization framework

- This paper follows the structure of the Netflix Prize attack.
- What are each phase of the attack?
  - Extract — extract micro-data, the authors used LLMs for this from extant data

The ESRC deanonymization framework

- This paper follows the structure of the Netflix Prize attack.
- What are each phase of the attack?
  - Extract — extract micro-data, the authors used LLMs for this from extant data
  - Search — given microdata, search for most likely matches; authors use nearest neighbor over LLM **embeddings of summaries**

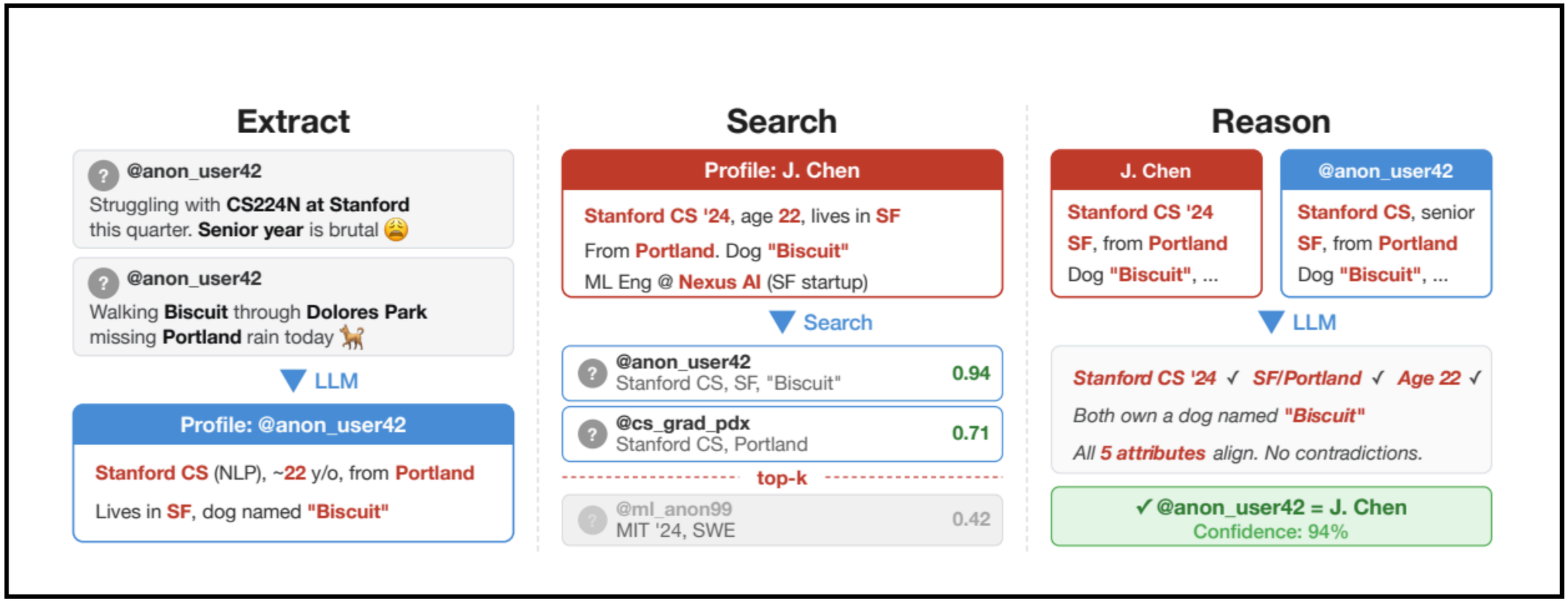
The ESRC deanonymization framework

- This paper follows the structure of the Netflix Prize attack.
- What are each phase of the attack?
  - Extract — extract micro-data, the authors used LLMs for this from extant data
  - Search — given microdata, search for most likely matches; authors use nearest neighbor over LLM **embeddings of summaries**
  - Reason — LLMs allow selection of a shortlist of candidates

The ESRC deanonymization framework

- This paper follows the structure of the Netflix Prize attack.
- What are each phase of the attack?
  - Extract — extract micro-data, the authors used LLMs for this from extant data
  - Search — given microdata, search for most likely matches; authors use nearest neighbor over LLM **embeddings of summaries**
  - Reason — LLMs allow selection of a shortlist of candidates
  - Calibrate — tradeoff between *precision* and *recall*. What is the precision / recall tradeoff in this case?

The ESRC deanonymization framework



This paper’s experiments

- This paper tackles three main experiments. What are they?

This paper’s experiments

- This paper tackles three main experiments. What are they?
  - Linking profiles across platforms — HackerNews —> LinkedIn
  - Linking users across Reddit communities cross community
  - Linking users across Reddit communities across time

This paper’s experiments

- This paper tackles three main experiments. What are they?
  - **Linking profiles across platforms — HackerNews —> LinkedIn**
  - Linking users across Reddit communities cross community
  - Linking users across Reddit communities across time

HackerNews to LinkedIn

- How did the authors build a ground-truth dataset of HackerNews accounts to LinkedIn accounts?

HackerNews to LinkedIn

- How did the authors build a ground-truth dataset of HackerNews accounts to LinkedIn accounts?
  - ~1000 users that put their LinkedIn in the HN bio, 89K active HN users
  - How does this make the attack inherently weaker than in the real-world?

HackerNews to LinkedIn

- How did the authors build a ground-truth dataset of HackerNews accounts to LinkedIn accounts?
  - ~1000 users that put their LinkedIn in the HN bio, 89K active HN users
  - How does this make the attack inherently weaker than in the real-world?
- How did the authors use ESRC to conduct their attack?

HackerNews to LinkedIn

- How did the authors build a ground-truth dataset of HackerNews accounts to LinkedIn accounts?
  - ~1000 users that put their LinkedIn in the HN bio, 89K active HN users
  - How does this make the attack inherently weaker than in the real-world?
- How did the authors use ESRC to conduct their attack?
  - **Extract** + summarize HN activity, **Search** via nearest-neighbor, **Reason** about top 100 using other models with two models, **Calibrate** using LLM confidence scores
- Authors achieved 45.1% recall at 99% precision. How do you interpret this result?

Linking Users across Reddit Communities

- Ground truth is a challenge here.
  How did the authors decide to solve the ground truth problem?

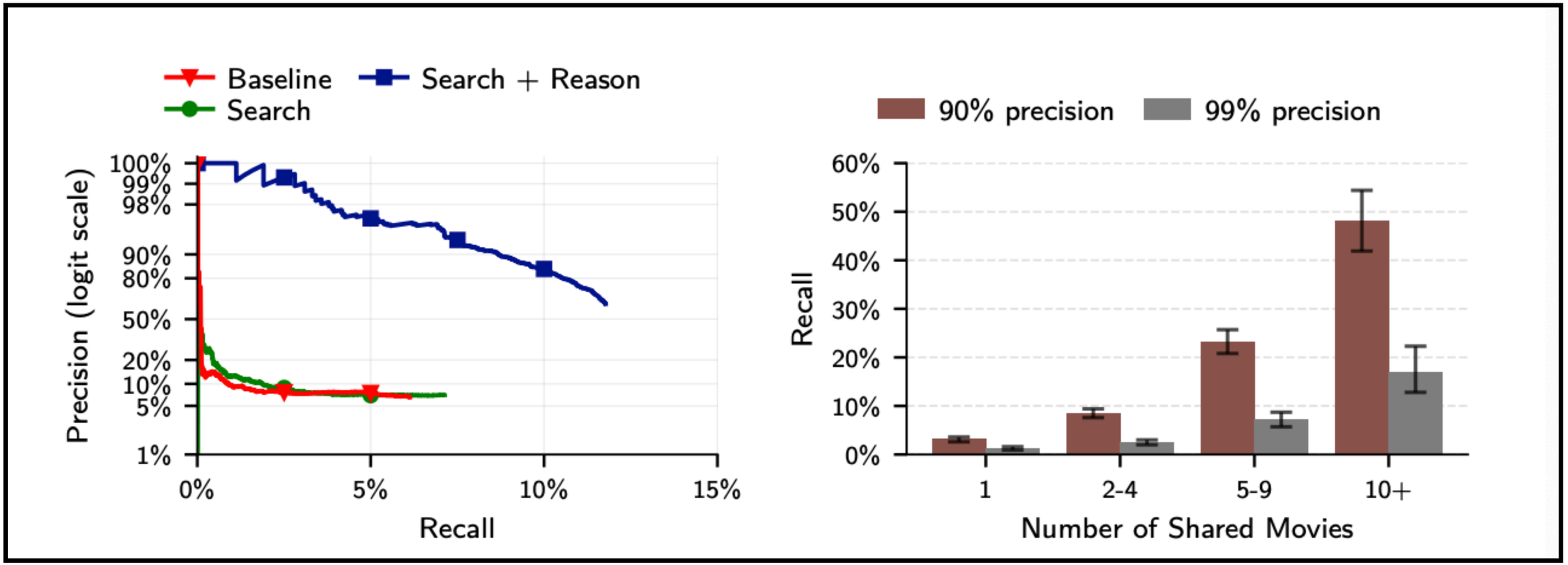
Linking Users across Reddit Communities

- Ground truth is a challenge here. **How did the authors decide to solve the ground truth problem?**
  - Instead of finding throwaway accounts — they instead link the *same* account across different communities. What assumptions does this make about deanonymization in this context?

Linking Users across Reddit Communities

- Ground truth is a challenge here. **How did the authors decide to solve the ground truth problem?**
  - Instead of finding throwaway accounts — they instead link the *same* account across different communities. What assumptions does this make about deanonymization in this context?
- Specifically, they look at same accounts posting in topically relevant communities. Why?

Linking Users across Reddit Communities



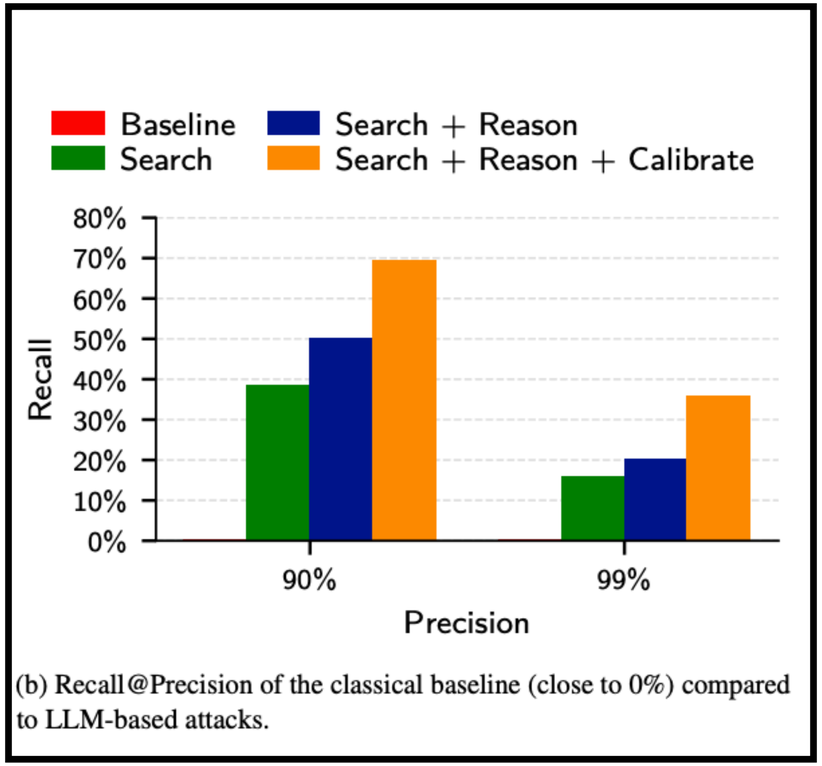
At 90% precision, high-reasoning achieves 8.5% recall.

Linking Users over Time

- Why did the authors seek to study the *temporal* behavior in addition to the cross-community behaviors?

Linking Users over Time

- Why did the authors seek to study the *temporal* behavior in addition to the cross-community behaviors?



(b) Recall@Precision of the classical baseline (close to 0%) compared to LLM-based attacks.

- Temporal task was *easier* than the cross-community task. Why?

Discussion

- What do we think about these attacks? Were they surprising? Unsurprising?
- What can we do about these attacks in the wild?
  - How can an individual user protect themselves from these types of harms?

Break Time + Attendance

https://tinyurl.com/cse227-attend

Codeword:
AnonymizeMeBro

Debate Time

Something kind of different for this discussion

- I'm going to frame a few "debate" style provocative claims about AI + security
  - Cold call two people, and they will offer an opening argument for each side of the argument
  - Then, we'll ping pong throughout the class; cold calling for responses to what was previously just said, until the conversation feels "over"
- Next class will be *debate* on topics we covered in the entirety of the class :)

AI will replace all security researchers in five years

In the age of AI, defenders always *lose* against attackers

AI will inherently make the software ecosystem more secure

AI safety research is a waste of time

Next time…

- More debate time!
- We'll carve out a little time for SETs — probably about 10 minutes at the beginning of class
- I'll end with some broader notes on security and how to think about your own security in the broader world (I do this in every sec class)
- Then, your presentations next week!