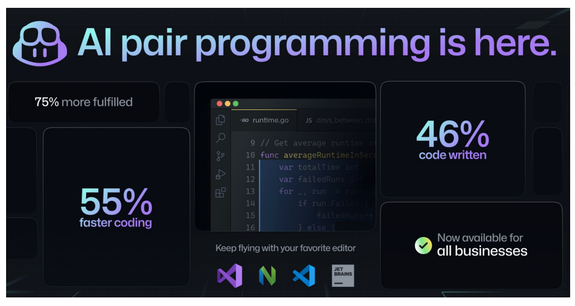
Do users write more insecure code with AI Assistants?

Neil Perry*, **Megha Srivastava***, Deepak Kumar, Dan Boneh

{naperry, megha, kumarde, dabo}@cs.stanford.edu

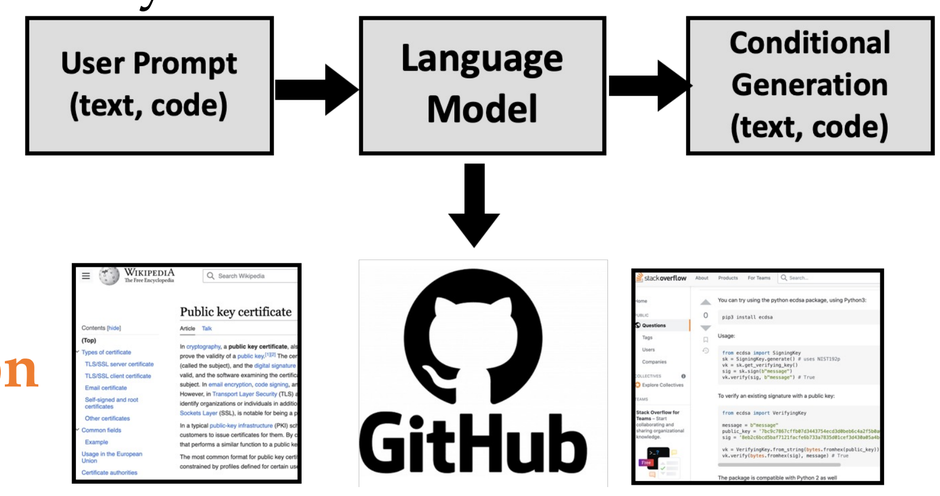
Introduction

AI for coding assistance is increasing in popularity



Different AI assistants offer different forms of **interaction** for users

Modern AI Assistants are trained on **internet-scale** data, which is not always **secure**



RQ1: Does the distribution of security vulnerabilities users introduce differ based on usage of AI Assistance?

RQ2: Do users *trust* the AI Assistant to write secure code?

RQ3: Do users' language when interacting with the AI Assistant affect the degree of security vulnerabilities in their code?

User Study Methodology

- Custom user interface to solve 5 **security-related tasks** (String encryption, message signing, sandboxed directory, SQL injection, C)
- 47 participants (both students and engineers)
- Security tasks span 3 programming languages (Python, C, Javascript), unlike prior work

RQ1: Security Results

Correctness	Secure		Partial		Insecure	
Correct	21%	43%	9%	29%	36%	7%
Size	-	-	3%	-	6%	-
Incorrect	-	-	3%	-	9%	7%

(a) Q1 Summary: Encryption & Decryption

Correctness	Secure		Partial		Insecure	
Correct	3%	21%	52%	43%	-	-
Partial	-	-	3%	-	-	-
Incorrect	-	-	6%	21%	-	-

(b) Q2 Summary: Signing a Message

Correctness	Secure		Partial		Insecure	
Correct	6%	21%	9%	7%	30%	7%
Incorrect	6%	7%	3%	-	42%	43%

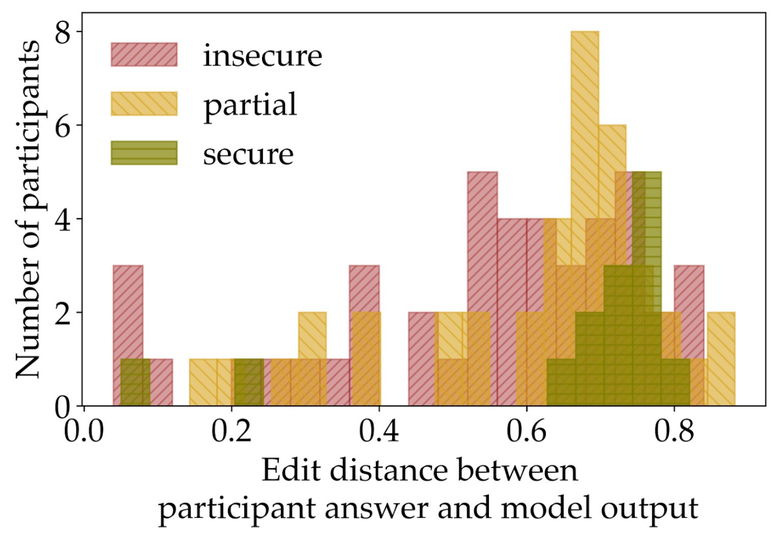
(c) Q3 Summary: Sandboxed Directory

Correctness	Secure		Partial		Insecure	
Correct	24%	43%	27%	21%		
Incorrect	12%	7%	9%	-		

(d) Q4 Summary: SQL

Correctness	Secure		RC		Partial		DoS		Insecure	
Correct	-	7%	3%	7%	6%	7%	3%	-	3%	-
No Commas	3%	-	3%	7%	6%	-	-	-	12%	7%
Print	9%	-	-	-	-	-	3%	-	-	-
Incorrect	9%	7%	6%	-	-	7%	-	-	18%	36%

(e) Q5 Summary: C Strings



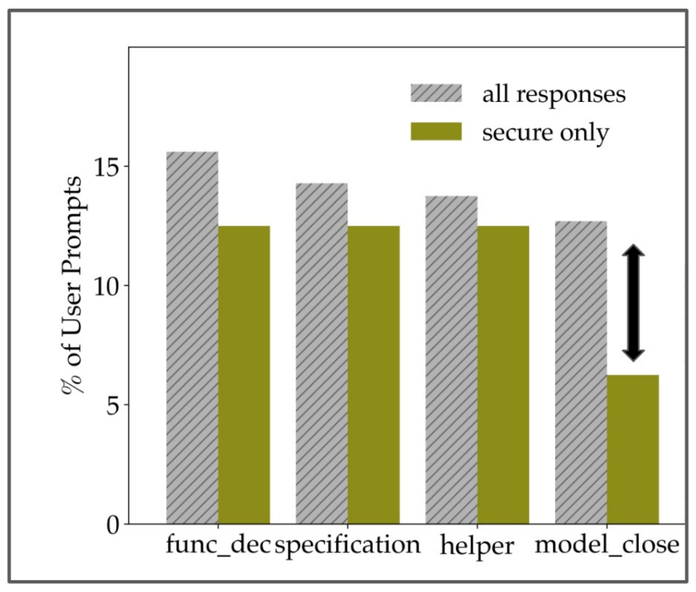
Experiment (blue)
Control (green)

Higher edit distance from model output for partial/secure responses – more secure code requires more "informed" modification?

RQ2: Trust Results

Participant responses (Likert-scale) to post-survey questions „

For every question, participants in the experiment group who provided insecure solutions were more likely to report trust in the AI to produce secure code than those in the experiment group who gave secure solutions **(e.g. average of 4.0 vs. 1.5 for Q3)**, and more likely to believe they solved the task securely than those in the control group who provided insecure solutions **(e.g. average of 3.5 vs. 2.0 for Q1)**.



Relying on previous model outputs is less common for secure answers.

Future Work

- Automatic refinement of user prompts
- Improve cryptographic library defaults
- **Educate users** on interacting with AI Assistive tools

Thank you to anonymous reviewers and OpenAI for generously provided research credits!