Building On-Prem LLM Applications for the Real World – Part 2

Alexander **Sternfeld,** Adrien **O'Hana** and Gaetan **Stein**

Welcome!

- **This morning in part 1 you:**
  - Deployed your own Open WebUI instance
  - Boosted your LLMs with RAG
  - Launched a MCP server
- **This afternoon, you will:**
  - Augment LLM capabilities with Open WebUI *functions*
  - Learn about safety risks surrounding LLMs
  - Build guardrails to mitigate these risks
  - Learn about threat modeling

Note: This workshop uses the same technical set-up as in part 1. If you did not attend part 1, it will be difficult to catch up on the technical installation, and you may fall behind.

About me

Alexander Sternfeld

- Associate researcher
  @ Reliable Information Lab, HES-SO
- Safety and Security @ Apertus Team
- Former intern
  @ Cyber-defence Campus, armasuisse
- Data science graduate
  @ EPFL

Tooling for This Workshop

- Website: https://llm-on-prem-amld2026.github.io

- Private servers (Exoscale)
  - Full set-up of a clean machine
  - From a minimal PoC
  - To a minimal personal prod.
- Re-do the workshop yourself:
  - With your hardware
  - Or on the same Private Servers

Agenda

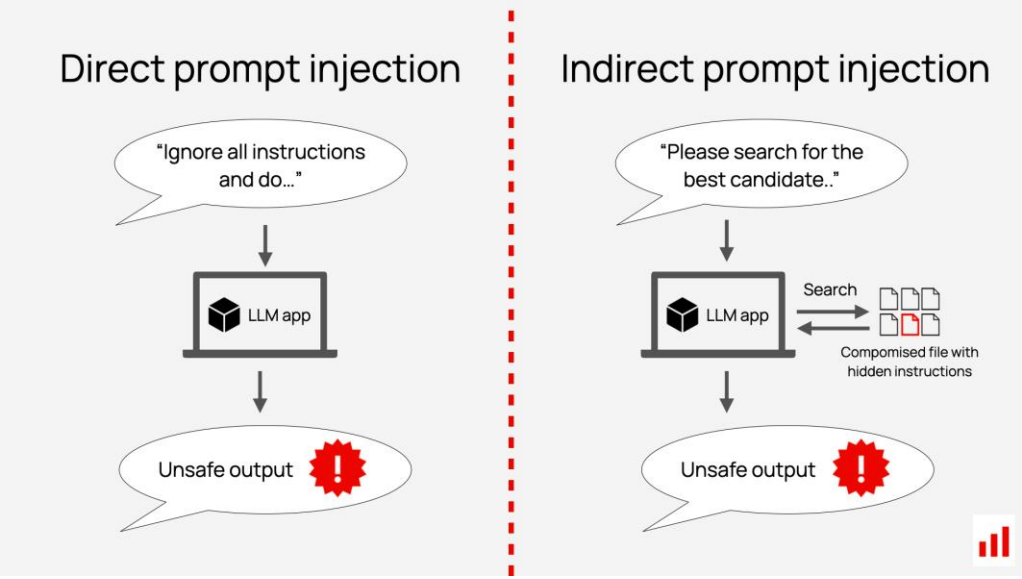
- **Augmenting LLM Capabilities**
- Prompt Injection and Guardrails - Theory
- *Coffee break*
- Prompt Injection and Guardrails - Exercises
- Threat Modeling and basics of MLOps
- Closing Remarks

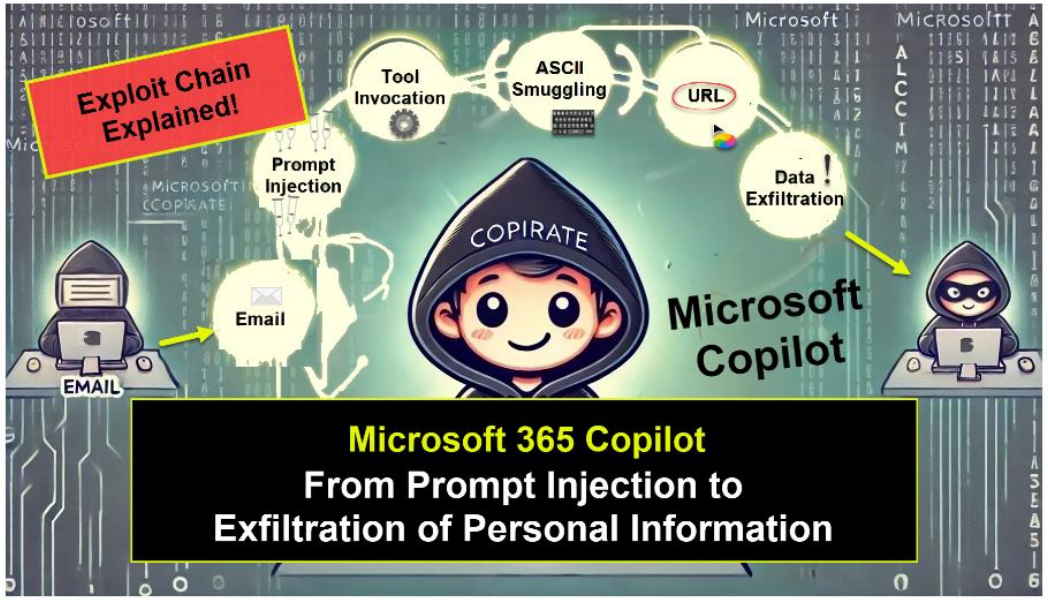
Going Further:
Augmenting LLM capabilities

1. Can we make LLMs **safer?**

2. Can we make LLMs **more useful?**

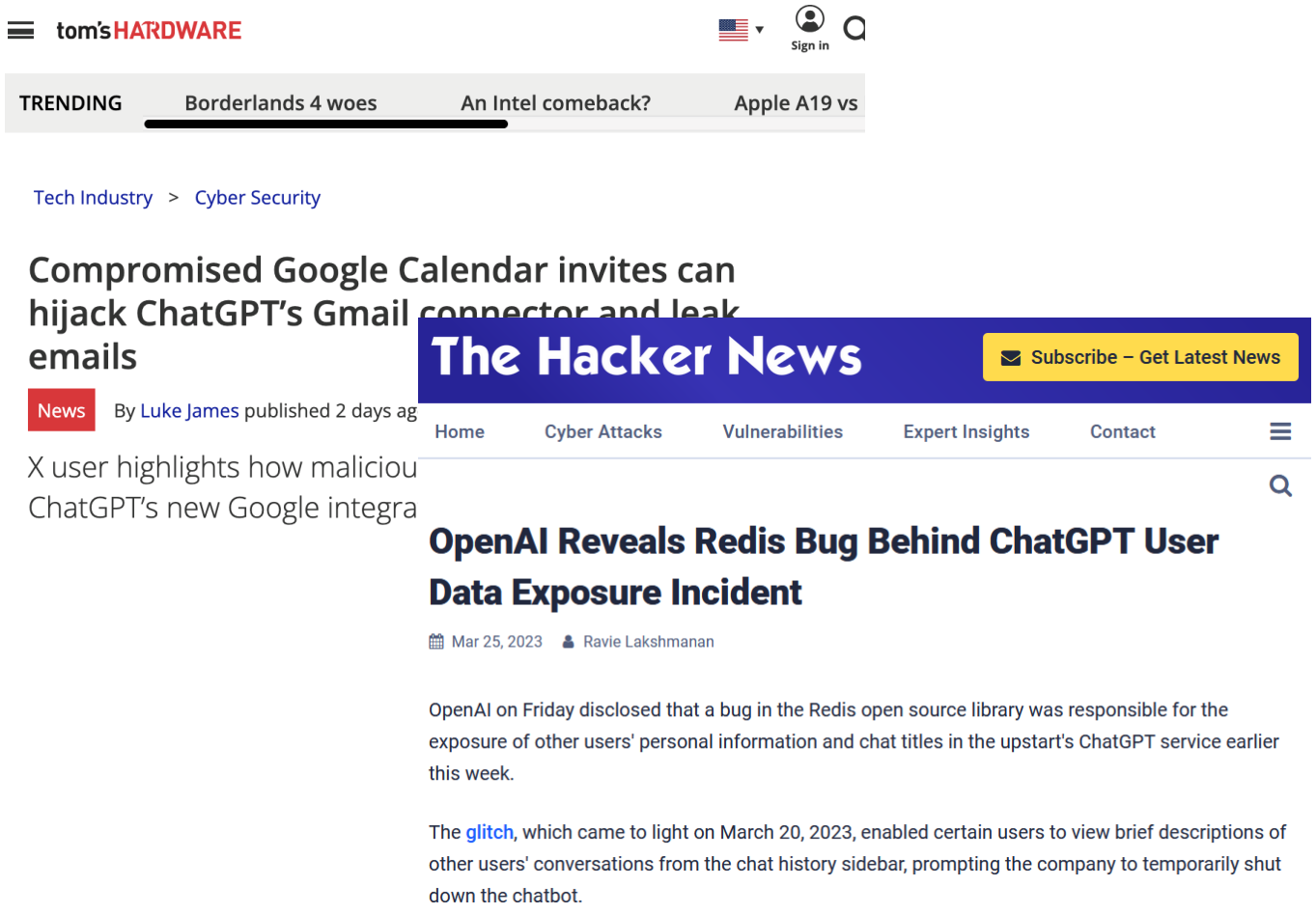
Safety Risk: Prompt Injection

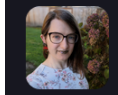




https://embracethered.com/blog/posts/2024/m365-copilot-prompt-injection-tool-invocation-and-data-exfil-using-ascii-smuggling/

LLM Apps Are Everywhere & **30 Years Behind** on Security





Hazel Weakly
@hazelweakly@hachyderm.io

Seriously, a large percentage of these attacks boil down to downloading untrusted content, mangling it ever so slightly, and then hoping that the AI decides to blindly eval all of it

WHICH
IT
THEN
FUCKING
DOES

I am losing it. What are these absurdly overpaid devs doing with their life?

Are you an AI vendor and you wanna prevent most attacks on the internet? All you need to do is:

1. Make your config files _READ ONLY_ during agent invocation
2. Use Content Security Policies correctly
3. Sanitize + normalize unicode input and output
4. Scan *both* the input and output

… bruh??

Impacts That Get **Blamed On Users**

CVE

Published: 2025-05-30

Updated: 2025-08-21

Tags: exclusively-hosted-service | disputed

Description

An insufficient database Row-Level Security policy in Lovable through 2025-04-15 allows remote unauthenticated attackers to read or write to arbitrary database tables of generated sites. NOTE: this is disputed by the Supplier because each individual customer of the Lovable platform accepts a responsibility over protecting the data of their application.

CWE 1 Total

Learn more

- CWE-863: CWE-863 Incorrect Authorization

CVSS 1 Total

Learn more

Score

9.3

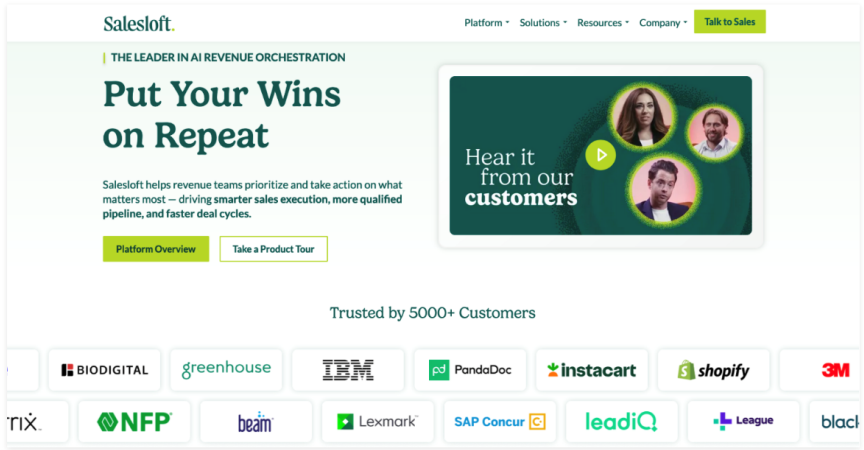
Severity

CRITICAL

The Ongoing Fallout from a Breach at AI Chatbot Maker Salesloft

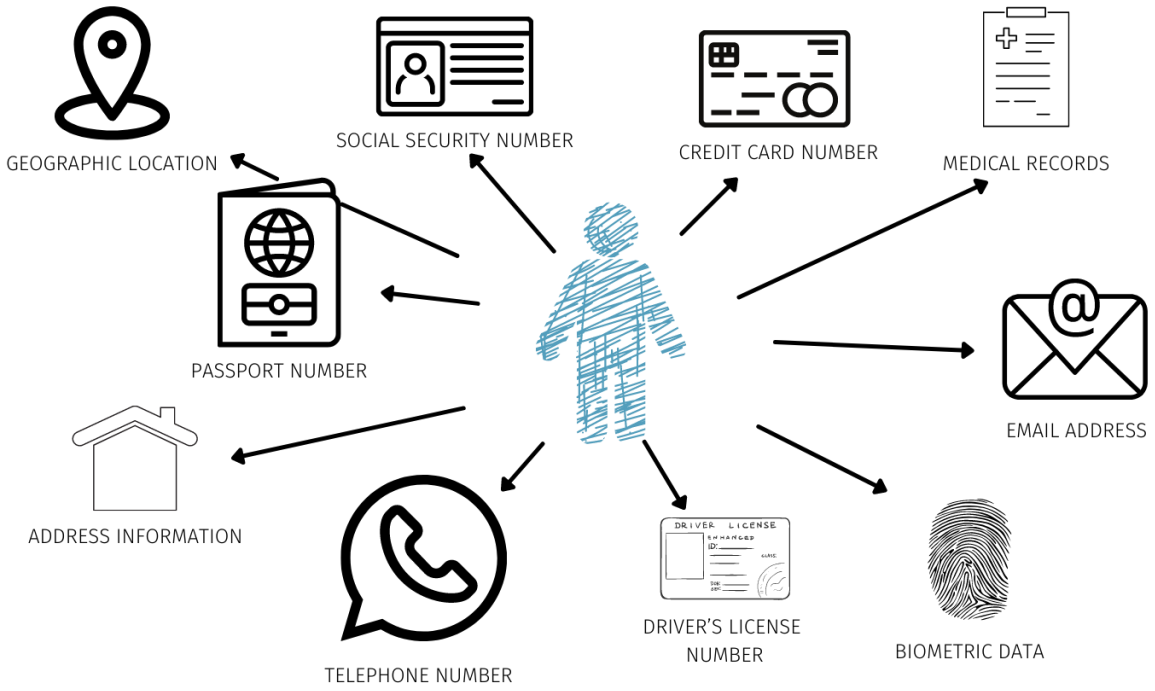
September 1, 2025

The recent mass-theft of authentication tokens from **Salesloft**, whose AI chatbot is used by a broad swath of corporate America to convert customer interaction into **Salesforce** leads, has left many companies racing to invalidate the stolen credentials before hackers can exploit them. Now **Google** warns the breach goes far beyond access to Salesforce data, noting the hackers responsible also stole valid authentication tokens for hundreds of online services that customers can integrate with Salesloft, including Slack, Google Workspace, Amazon S3, Microsoft Azure, and OpenAI.



Safety risk: Sensitive Information

- **Our exercise:** You may want to protect users against themselves
- For example, giving the LLM **private personal information**



Functions

- We will now practice with **functions** in Open WebUI

Follow the instructions on the page ***Augmenting LLM Capabilities -> Functions in Open WebUI***

- If you run into problems, please approach one of us!

- **Done?** Do not advance to agentic workflows, but work on exercise 1 and 2 from the additional exercises

Safety risk: LLM generated code

- LLMs are not yet capable of writing **secure** code

CodeMirage: Hallucinations in Code Ge

Joseph Spracklen
University of Texas at San Antonio

Raveen Wijewickrama
University of Texas at San Antonio

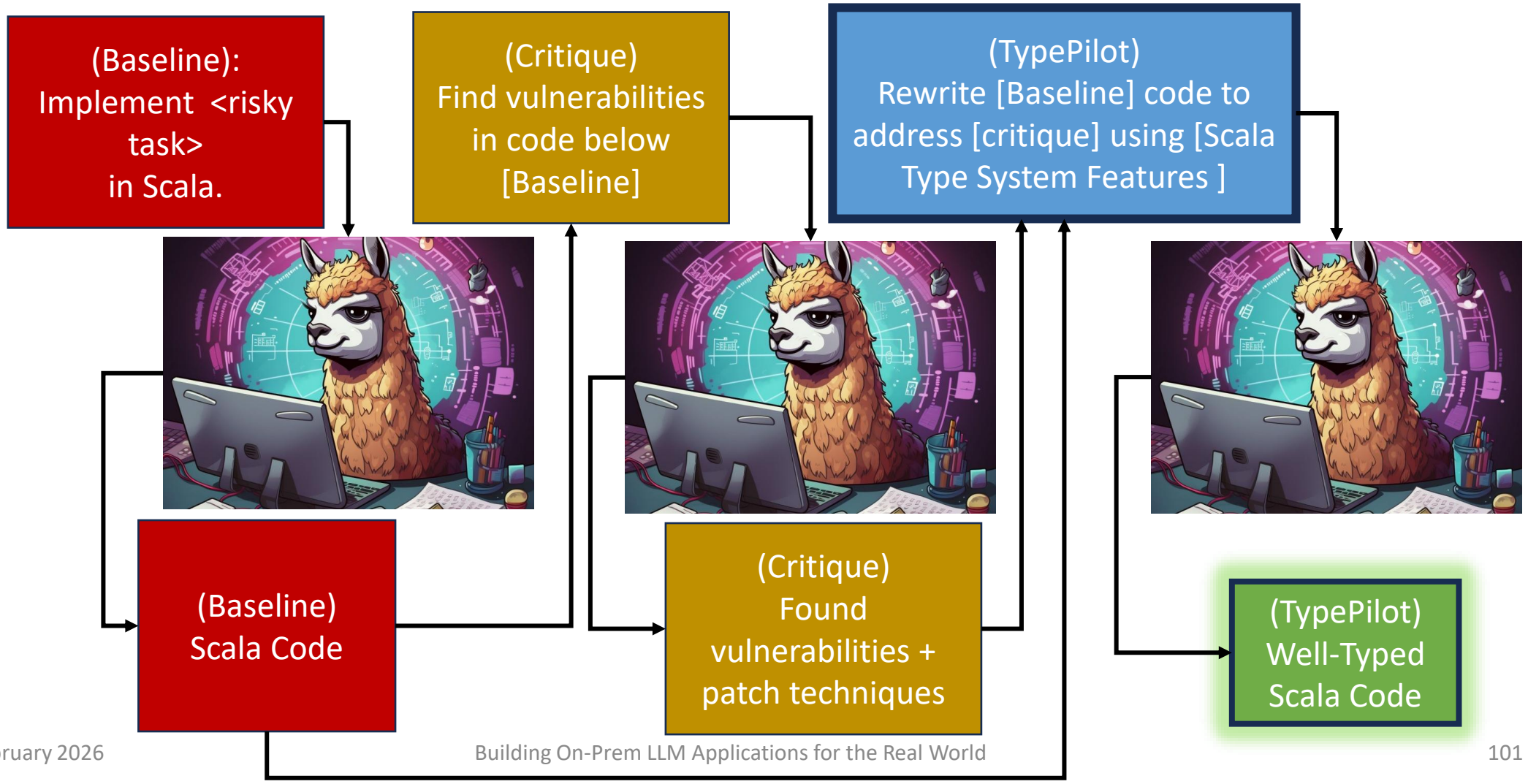
A H M Nazmus Sakib
University of Texas at San Antonio

Anindya Maiti
University of Oklahoma

Bimal Viswanath
Virginia Tech

Murtuza Jadliwala
University of Texas at San Antonio

CodeGen: Type/Safe Pilot



TypePilot

	Qwen-2.5-Coder (32B)			CodeLlama (70B)			Deepseek-coder (33B)		
	Baseline	Robust	TypePilot	Baseline	Robust	TypePilot	Baseline	Robust	TypePilot
Average age									
- Correct for regular input	✓	✓	✓	✓	✓	✓	✓	✓	✓
- Handle empty lists	✓	✓	✓	✓	✓	✓	✓	✓	✓
- Handle negative ages	✗	✗	✓	✗	✗	✓	✗	✗	✗
Fibonacci number N									
- Correct for regular input	✓	✓	✓	✓	✓	✓	✓	✓	✓
- Check for negative N	✗	✓	✓	✗	✗	✗	✗	✗	✓
- Handles large values of N	✗	✓	✓	✗	✗	✗	✗	✓	✓
Matrix multiplication									
- Correct for regular input	✓	✓	✓	✓	✓	✓	✓	✓	✓
- Check for empty matrices	✓	✗	✓	✗	✗	✗	✗	✗	✗
- Check for dimension matching	✓	✓	✓	✓	✗	✓	✗	✓	✓
Matrix convolution									
- Correct for square matrix input	✓	✓	✓	✗	✗	✗	✓	✓	✓
- Correct for regular matrix input	✓	✓	✓	✗	✗	✗	✓	✓	✓
- Handles rectangular kernels	✗	✗	✓	✗	✗	✓	✗	✗	✗
- Checks for empty kernel	✗	✓	✓	✗	✗	✓	✗	✗	✓
- Checks for empty matrix	✗	✓	✓	✗	✗	✓	✗	✗	✓
- Handles even sized kernels	✗	✗	✓	✗	✗	✗	✗	✗	✗

TypePilot

	Qwen-2.5-Coder (32B)			CodeLlama (70B)			Deepseek-coder (33B)		
	Baseline	Robust	TypePilot	Baseline	Robust	TypePilot	Baseline	Robust	TypePilot
HTML greeting									
Correctness and compilation	✓	✓	✓	✓	✓	✓	✓	✓	✗
Robust to injection	✗	~	✓	✗	✓	✓	✗	✓	✓
HTML comments									
Correctness and compilation	✓	~	✓	✓	✗	~	✓	✓	✓
Robust to injection	✗	✓	✓	✗	✗	✓	✗	~	✓
Bash file search									
Correctness and compilation	✓	✓	✓	✓	✓	✓	✓	✓	✓
Robust to injection	✗	✗	✓	✗	✗	✓	✗	✗	✓
Bash host ping									
Correctness and compilation	✓	✓	✓	✓	✗	✓	✓	✓	✓
Robust to injection	✓	✓	✓	✗	✗	✓	✗	✓	✓
URL redirect									
Correctness and compilation	✓	✓	✓	✓	✓	✓	✓	✓	✓
Robust to injection	✗	~	~	✗	✓	✓	✗	✗	✓

If TypePilot **compiles,** it is correct and secure

(even before we talk to the compiler)

Functions - continuation

- We will now build a simplified version of TypePilot in Open WebUI

Follow the instructions on the page ***Augmenting LLM Capabilities -> Agentic Workflows***

- **Done?** Feel free to continue with the additional exercises!

- If you run into problems, please approach one of us!