Large Language Models for Code: Secure Hardening and Adversarial Testing

Jingxuan He and Martin Vechev

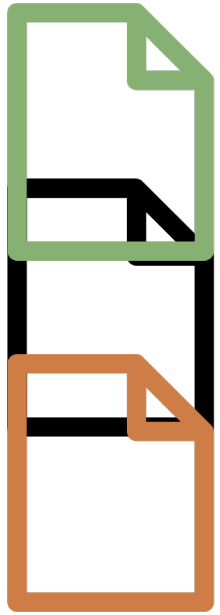
ACM CCS 2023

SRILAB

ETH zürich

LLMs for Code Generation

Prompt + LLM



60%

40%

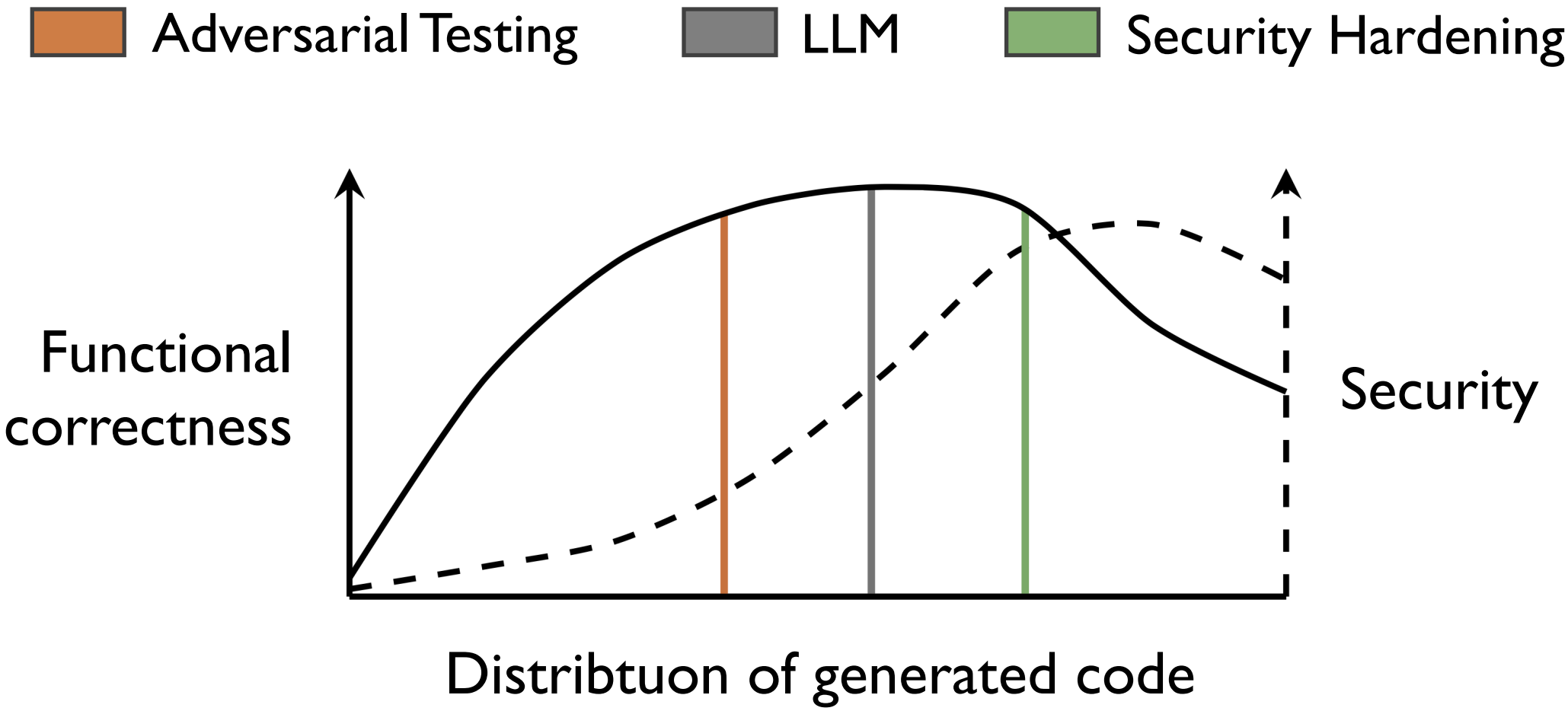
Pearce et al., Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions, IEEE S&P 2022

Li et al., StarCoder: May the Source be With You!, arXiv:2305.06161

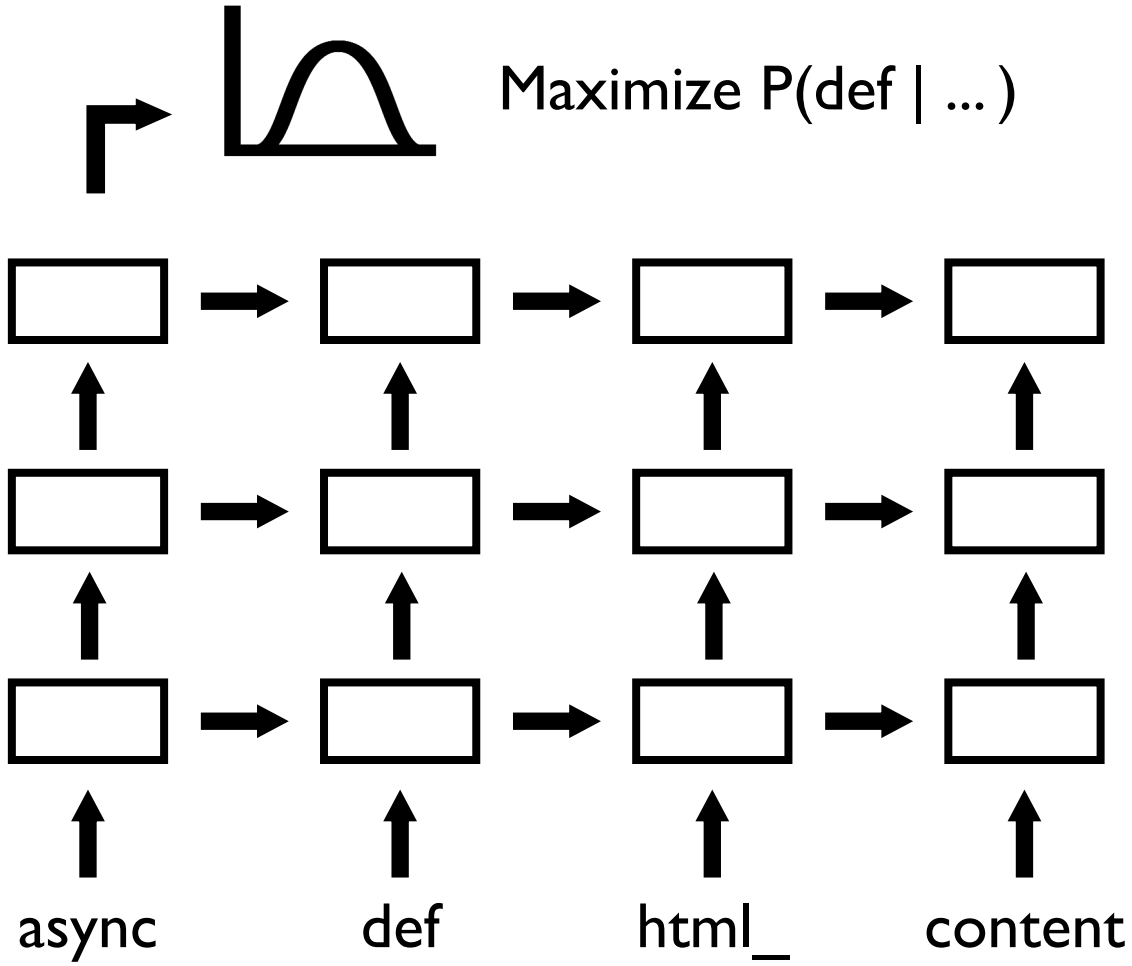
Making LLMs generate unsafe code more often?

Security Hardening & Adversarial Testing

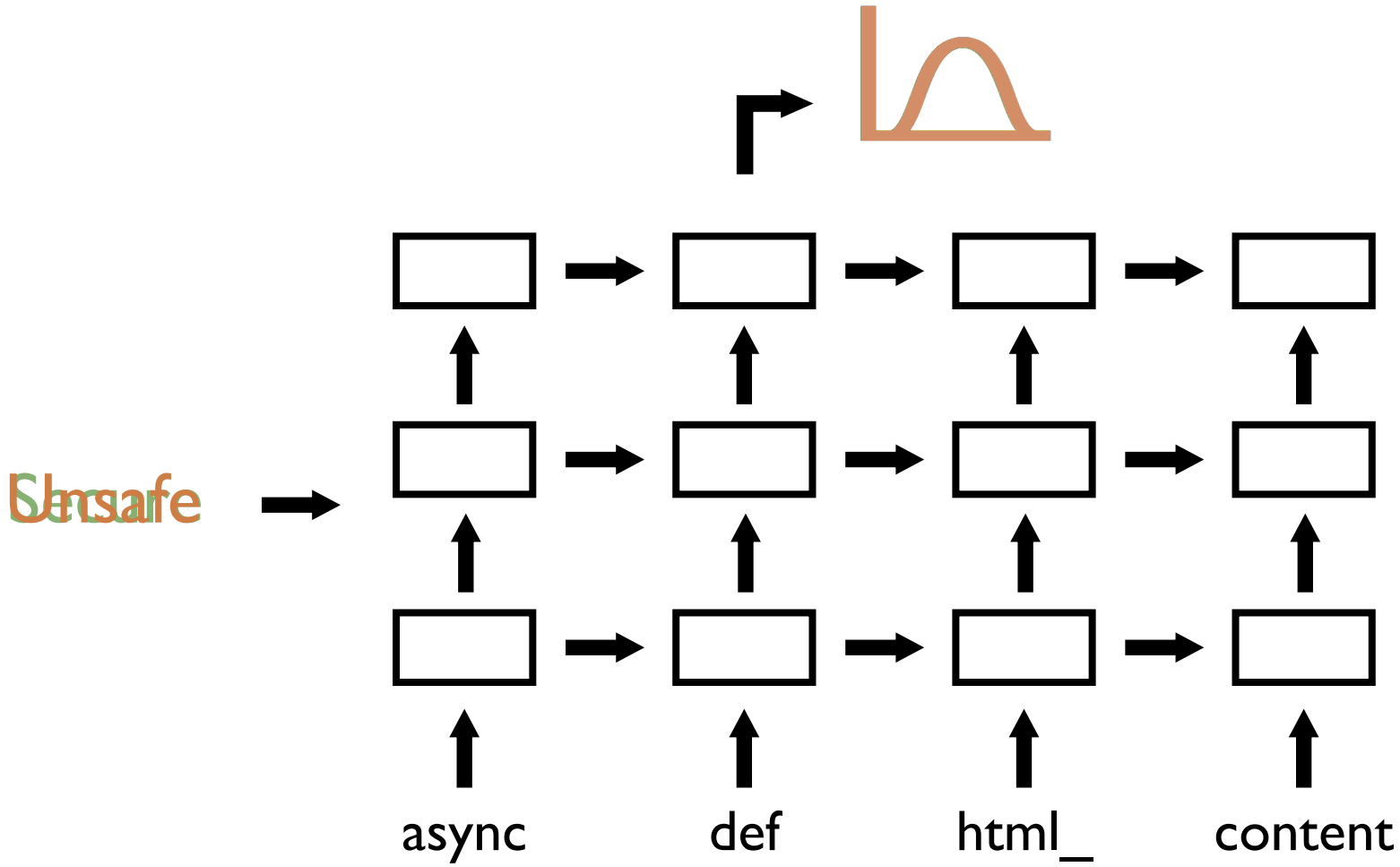
Functional Correctness & Security



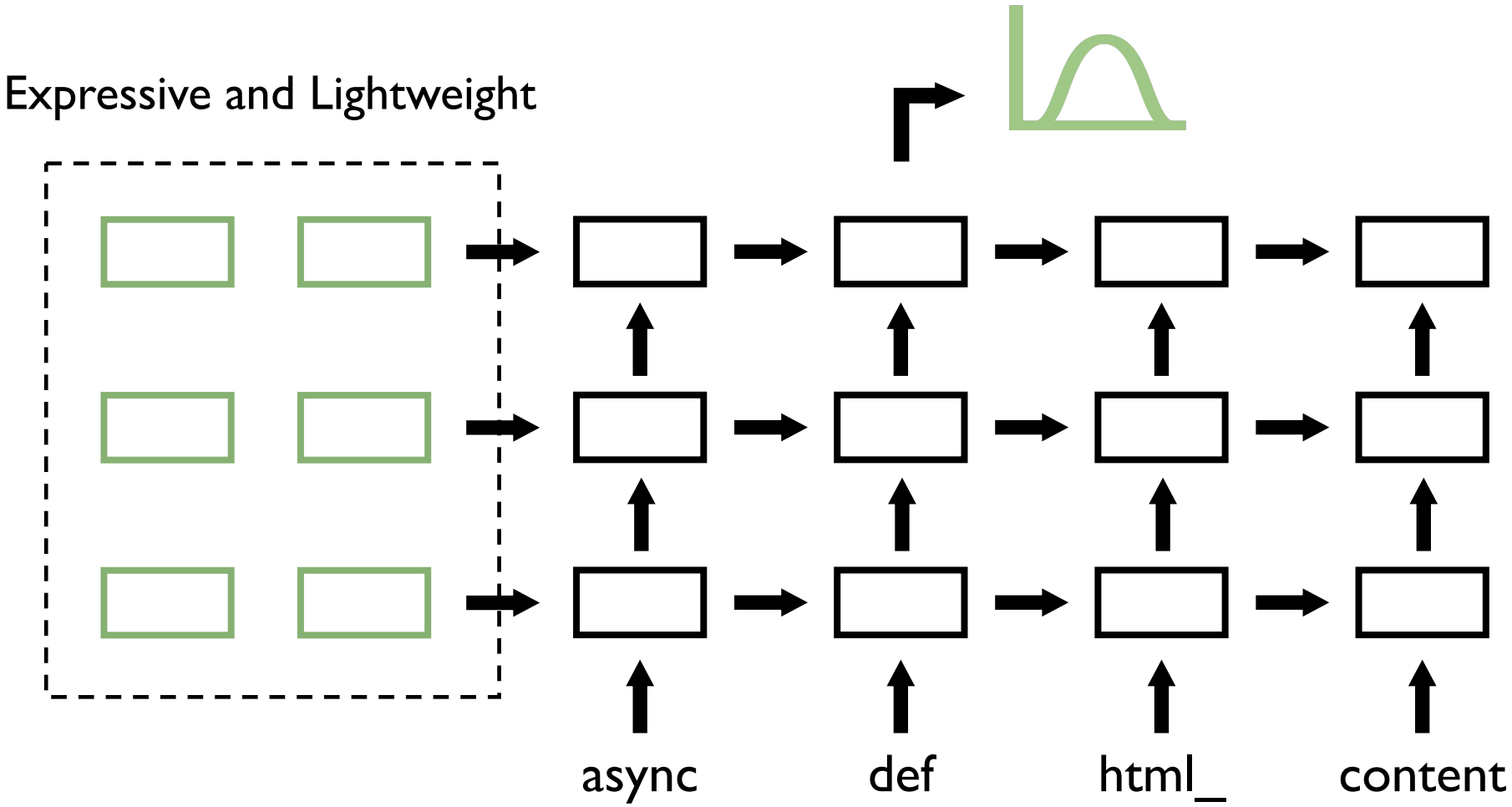
Language Modeling



Conditional Language Modeling



SVEN: Soft Prompts as Prefixes



Li and Liang, Prefix-Tuning: Optimizing Continuous Prompts for Generation, ACL 2021

SVEN: Training Data

Security fixes extracted from GitHub commits:

```
async def html_content(self):
-   content = await self.content
    return content
```

```
async def html_content(self):
+   content = markupsafe.escape(await self.content)
    return content
```

A Naïve Training Scheme:

SVEN: Code Regions

```
async def html_content(self):
- [redacted]wa[illegible]f.[redacted]
    return content
```

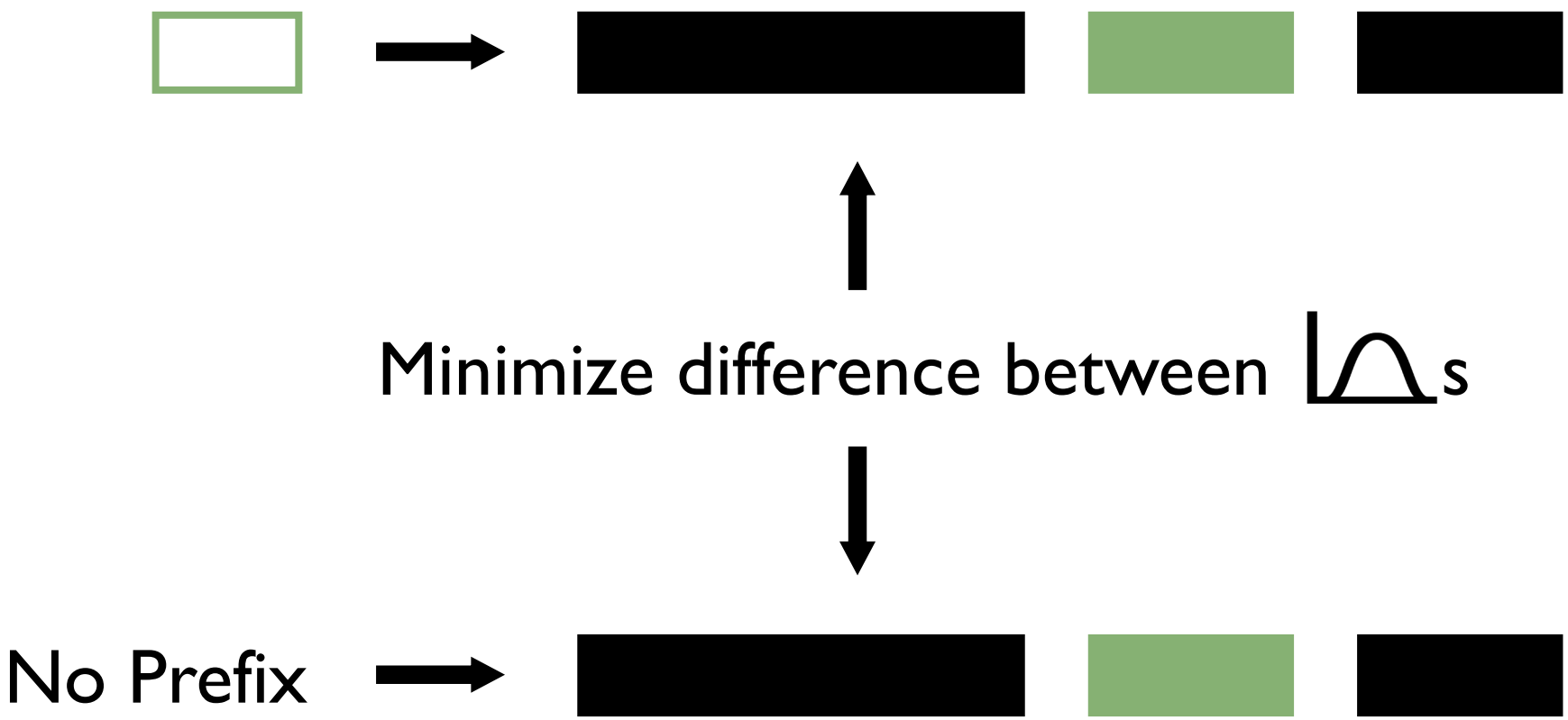
```
async def html_content(self):
+   conte[redacted]e.[illegible] s[redacted]ntent)
    return content
```

SVEN Training: Security

Maximize P( |…, )

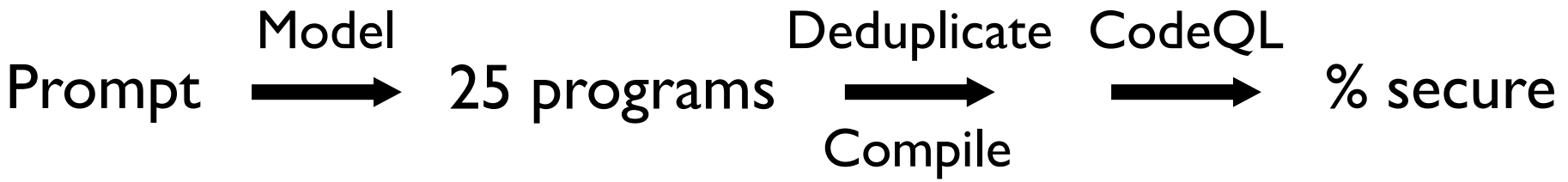
Minimize P( |…, )

SVEN Training: Functional Correctness

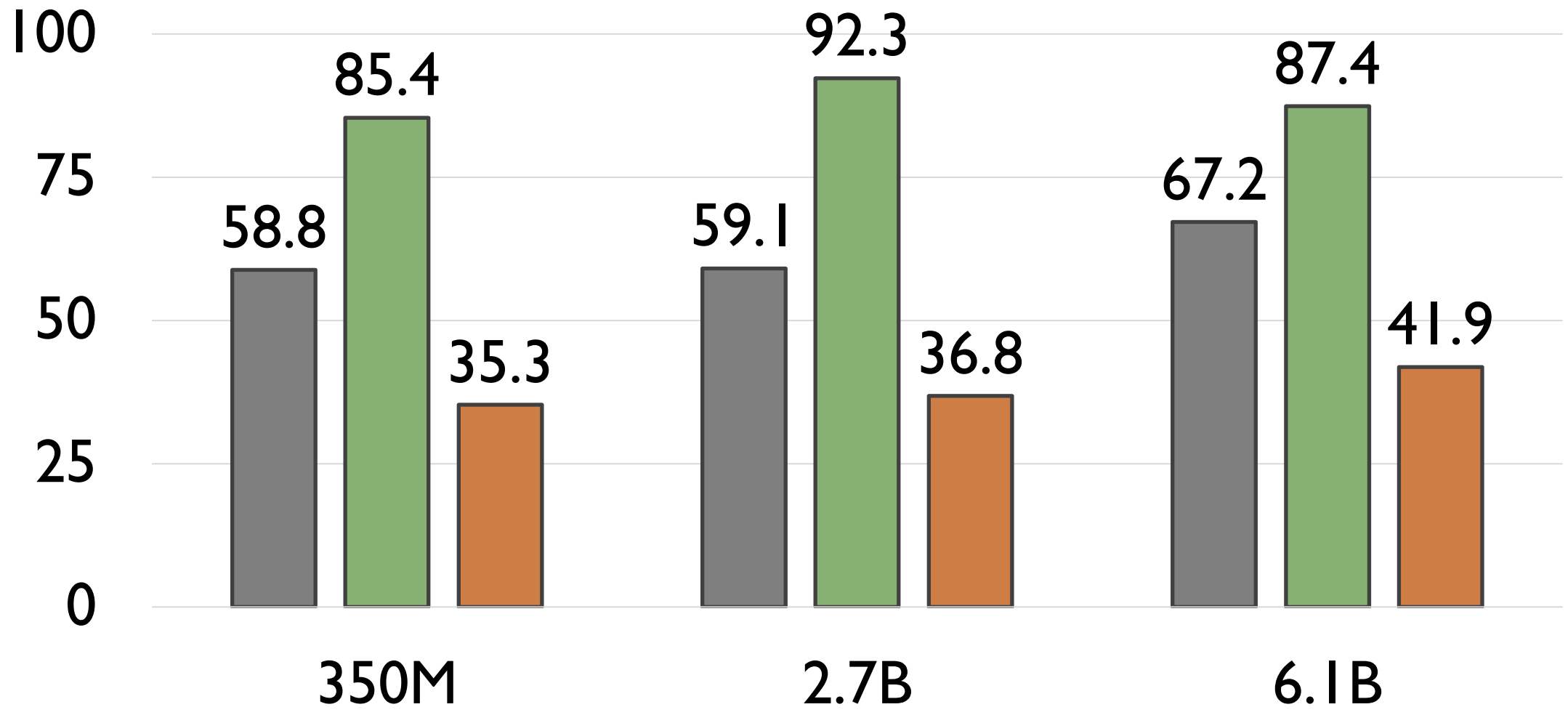


Experimental Setup

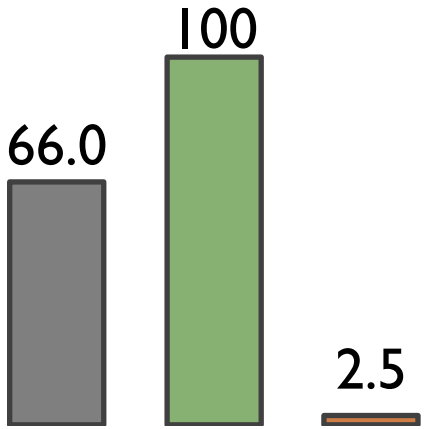
- Manual curation for high-quality training data
- LLMs: CodeGen with 350M, 2.7B, and 6.1B parameters
- Evaluating functional correctness: pass@k on HumanEval
- Evaluating security:



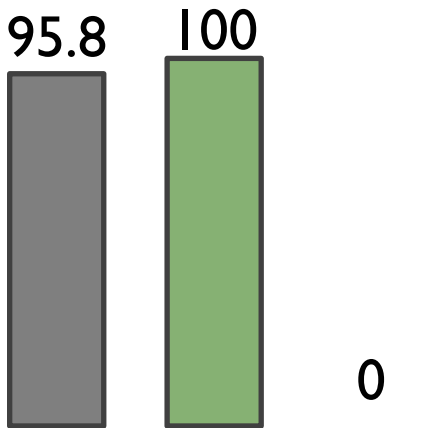
Overall Security



Example: SQL Injection

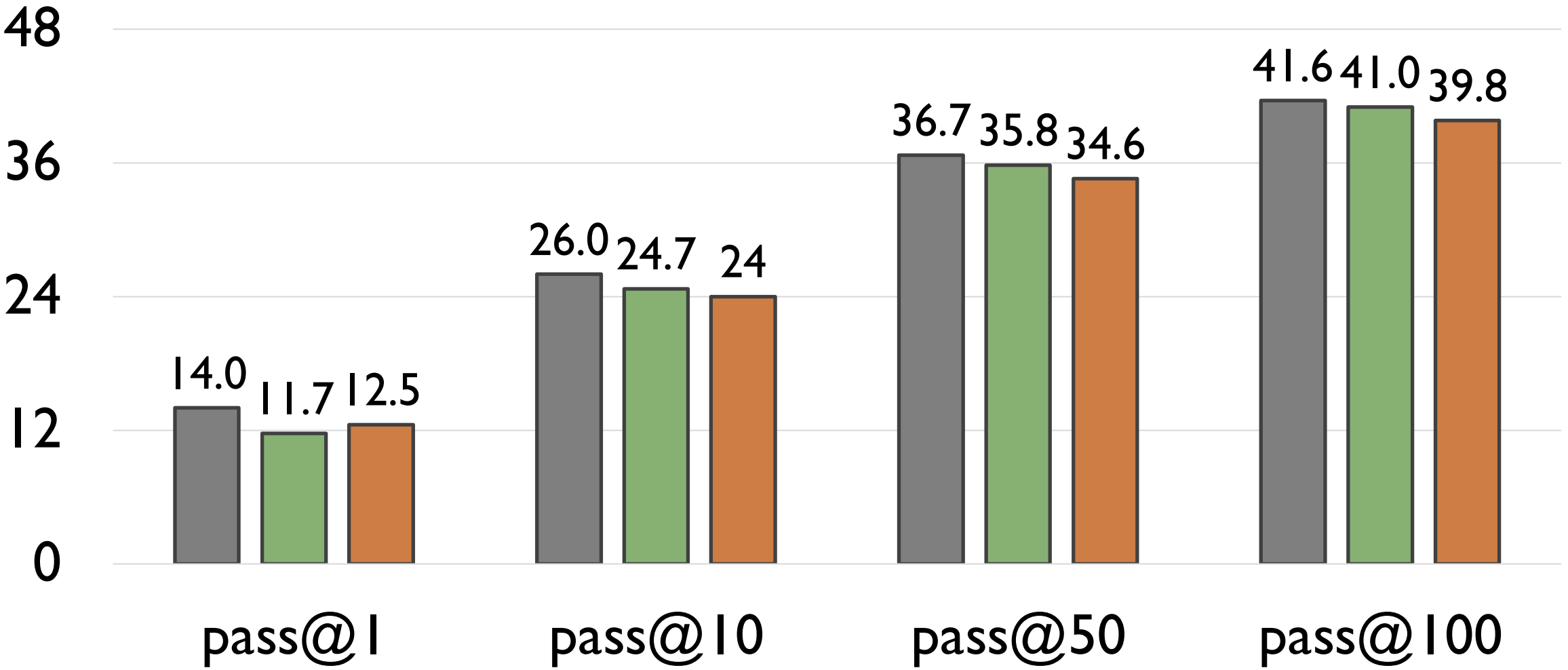



```
query = "INSERT INTO messages (username, message) " \
   "VALUES ('" + username + "', '" + message + "')"
cursor.execute(query)
```




```
query = "INSERT INTO messages (username, message) " \
        " VALUES (%s, %s);"
cursor.execute(query, (username, message))
```

Functional Correctness



2302.05319

eth-sri/sven

Q & A