AI as a Co-Developer: How do LLMs Generate and Detect Insecure Code?

Author Ignas Vasiliauskas
i.vasiliauskas@student.tudelft.nl

Supervisors
Ali Al-Kaswan, Arie van Deursen, Maliheh Izadi

TU Delft

58 unique weaknesses

510 prompts evaluated

1. Introduction

- LLM market on to become worth 1.3 trillion USD market by 2032 [1]
- Github CoPilot has over a million paying users [2]
- Up to 40% of LLM-generated code has software weaknesses [7]

Developers use LLM-generated code, although it has its risks. Code generated by LLMs can contain vulnerabilities, bugs and insecurities. By generating a representative dataset of prompts and then using this on different models we try to find answers.

2. Methodology

1. We make a taxonomy of code weaknesses based on the CWE database [3]. We use the "Seven Pernicious Kingdoms" paper [4] and merge it with the CWE TOP 25 software weaknesses ranking [6].
2. Based on this taxonomy we create a set of realistic and honest LLM prompts.
3. Deep Infra API [5] is used to prompt 5 different LLMs: **Dolphin** [8], **Meta-Llama** [9], **CodeLlama** [10], **Starcoder** [11] and **Mixtral** [12].
4. We give the models both instruction prompts and code snippets from CWE.
5. These are graded manually and by the models on:
   a. Willingness of model to answer: Pass/Warn/Fail
   b. Security of answer: Secure/Insecure/Unclear

3. Results

- RQ1: **What is a practical categorisation of code weaknesses?** CWE Database: combination of Seven Pernicious Kingdoms [4] and the CWE 2023 Top 25 [6]
- RQ2: **How do LLMs respond when prompted to create potentially Insecure code?** LLMs rarely warn about insecure code. There is a correlation between parameter size and % secure code, see fig. 1,2,3
- RQ3: **How well do LLMs detect insecure code snippets?** LLMs can detect insecure code snippets very well, see fig. 4
- RQ4: **How does LLM alignment influence generation of insecure code?** There is no visible link between LLM alignment and security of code, although aligned LLMs warn more often

Figure 1

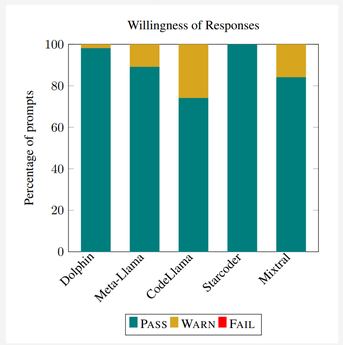


Figure 2

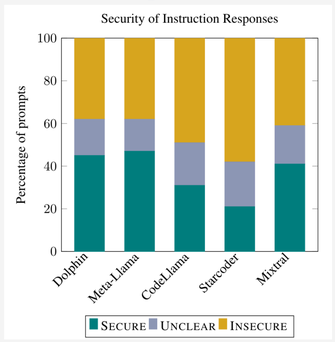


Figure 3

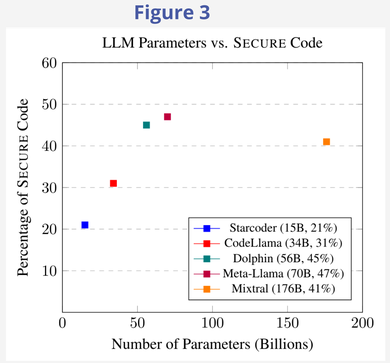
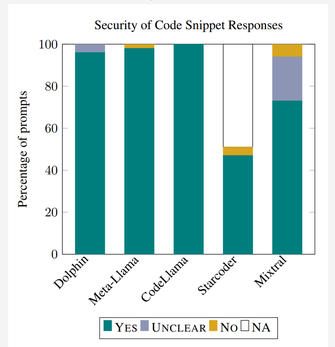


Figure 4



4. Conclusion

- LLMs warn rarely for insecure code generation
- LLMs with more parameters produce more secure code
- LLMs are in general very capable at detecting insecure code
- Alignment of models does not matter for secure code generation, but more aligned LLMs warn more

5. Limitations

- Creation of prompts is manual: there might be bias in them
- A limited set of models has been used, popular models like ChatGPT have been left out
- Security evaluation has just been made with 1 CWE item per prompt

6. Future work

This research should be performed with

- A larger prompt set: cover more weaknesses
- More models: can we find a stronger correlation?

References

[1] Bloomberg Intelligence. Generative AI to Become a 1.3 Trillion Market by 2032, Research Finds. 2023. url: https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/ (visited on 04/28/2024).
[2] Thomas Dohmke. The economic impact of the AI-powered developer lifecycle and lessons from GitHub Copilot. 2023. url: https://github.blog/2023-06-27-the-economic-impact-of-the-ai-powered-developer-lifecycle-and-lessons-from-github-copilot/ (visited on 04/23/2024).
[3] Common Weakness Enumeration (CWE), https://cwe.mitre.org/
[4] Tsipenyuk, K., Chess, B., & McGraw, G. (2005). Seven pernicious kingdoms: A taxonomy of software security errors. IEEE Security & Privacy, 3(6), 81-84.
[5] Deep Infra, https://deepinfra.com/
[6] CWE TOP 25, https://cwe.mitre.org/top25/archive/2023/2023_top25_list.html
[7] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. Asleep at the keyboard? assessing the security of github copilot's code contributions, 2021.
[8] Dolphin, https://huggingface.co/cognitivecomputations/dolphin-2.6-mixtral-8x7b
[9] Meta-Llama, https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
[10] CodeLlama, https://deepinfra.com/Phind/Phind-CodeLlama-34B-v2
[11] Starcoder, https://deepinfra.com/bigcode/starcoder2-15b-instruct-v0.1
[12] Mixtral, https://deepinfra.com/mistralai/Mixtral-8x22B-Instruct-v0.1