UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

itee PhD information technology electrical engineering

DIETI UNINA

Roberta De Luca

Towards Safer AI Code Generators

Tutor: Prof. Domenico Cotroneo

Cycle: XXXIX Year: Second

itee PhD information technology electrical engineering

My background

- MSc degree: Computer Engineering (July 2023)
- Research group: DESSERT
- PhD start date: 01/11/2023 – end date: 31/10/2026
- Scholarship type: PNRR - DM 118/2023 (Mis. 4.1: Dottorati generici)
- Periods abroad: 06/02/2025 – 27/09/2025 at University of Coimbra (UC), Department of Informatics Engineering (DEI), Portugal
  - Local supervisor: Prof. Naghmeh Ramezani Ivaki

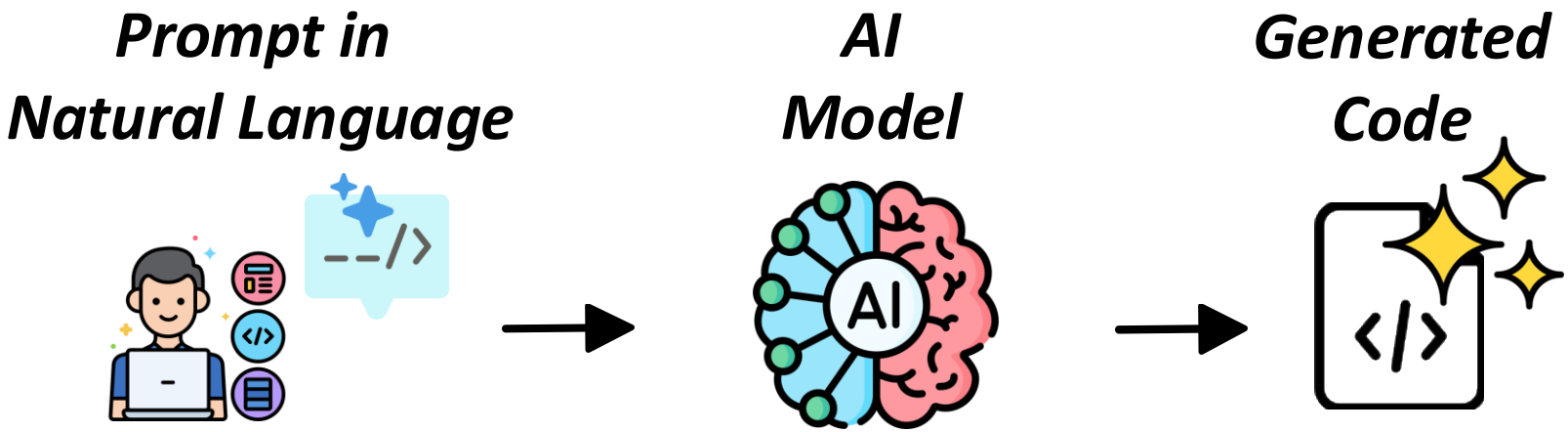
Summary of study activities

- Ad hoc PhD courses:
  - How to boost your PhD
  - Innovation and Entrepreneurship

- Conferences / events attended:
  - 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2025), June 23-26, 2025, Naples, Italy. (Online attendance)

Research field of interest

- My research investigates the trustworthiness of AI code generators with a focus on their security: exploring techniques for automatic vulnerability detection and patching, and conducting controlled vulnerability injection.



Research activity: Problem Statement

Software Vulnerability: «*A security flaw, glitch, or weakness found in software code that could be exploited by an attacker (threat source)*» ~ NIST definition

Natural Language Prompt

"Write a Python function that defines a YAML payload, loads this payload into a dictionary called prod_and_qt by using a function for loading"

Generated Code

```
def load_yaml(yaml payload) :
            prod_and_qt = yaml.load(yaml_payload)
            return prod_and_qt

yaml_payload = ""
prod_and_qt = load_yaml(yaml_payload)
print(prod_and_qt)
```

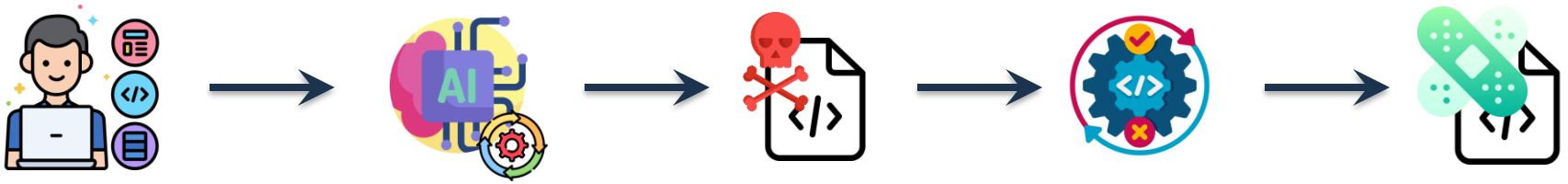
CWE-502 (Deserialization of Untrusted Data)

Incomplete (lack of `import yaml`)

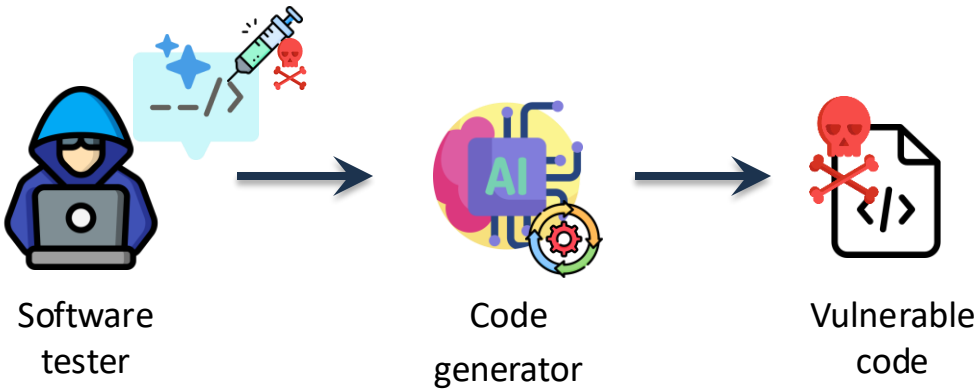
SOTA Tool CodeQL	**Detection** Yes	**Patching** No
SOTA Tools Semgrep, Bandit	**Detection** Yes	**Patching** Only suggestions. No modification of code

Research activity: Objective

Explore practical methods for detecting and automatically patching vulnerabilities in AI-generated code (snippet & full programs).



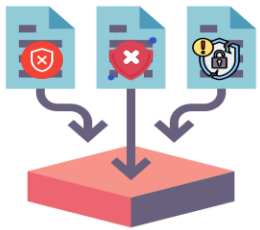
Detecting and fixing is only part of the problem. We also need to understand how easily models can be induced to produce vulnerabilities.



Research activity: Methodology

Patching strategy

1 **Vulnerable Code Collection**



2 **Vulnerability Classification**

3 **Safe Code Implementation**

MITRE

4 **Pattern Extraction**

Vulnerable pattern extraction

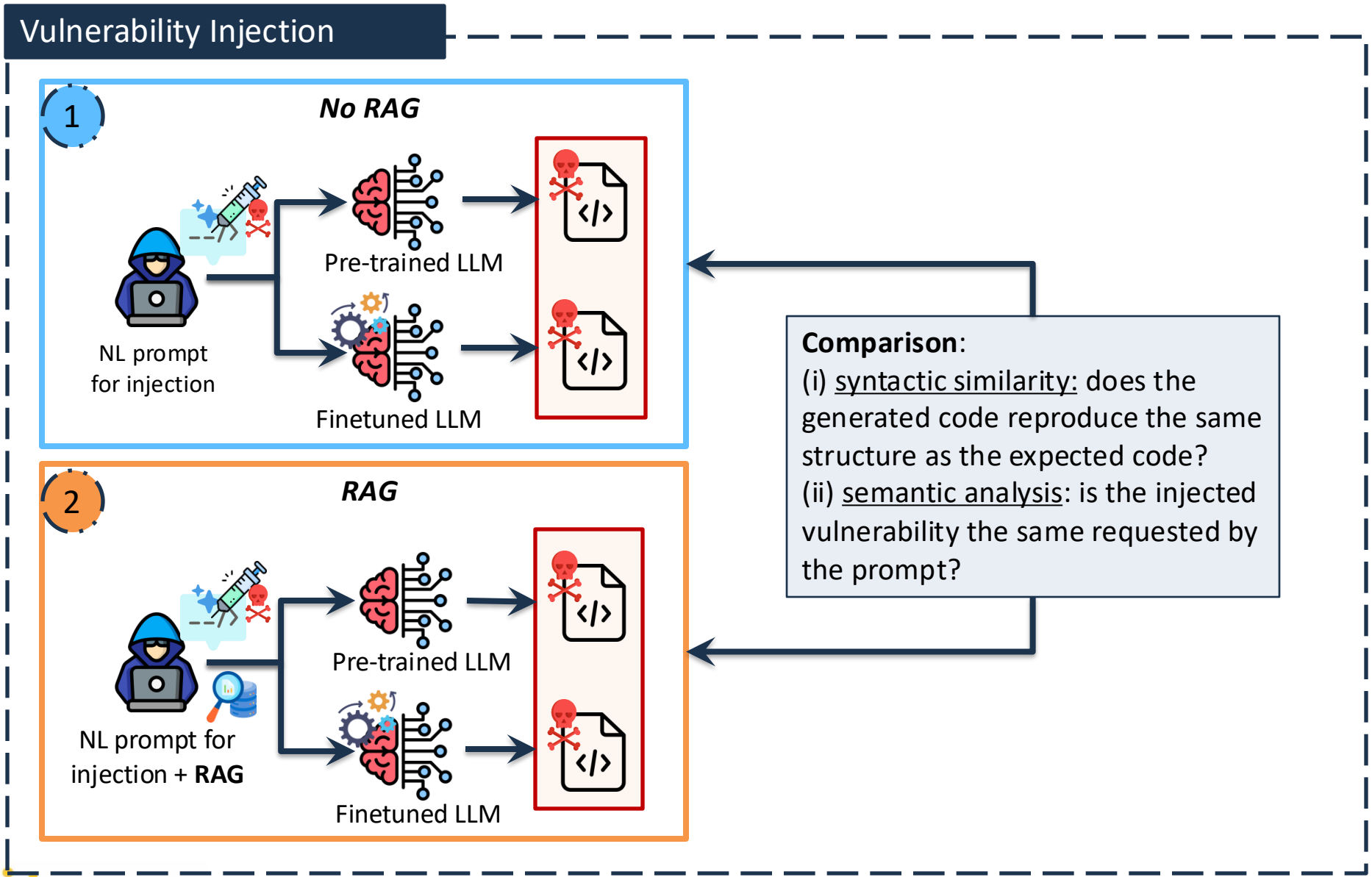
Safe pattern extraction

5 **Pattern Comparison**

Comparison between vulnerable and safe pattern to extract the additional parts of code safe that are missing in the vulnerable

PatchitPy

Research activity: Methodology



Research activity: Results

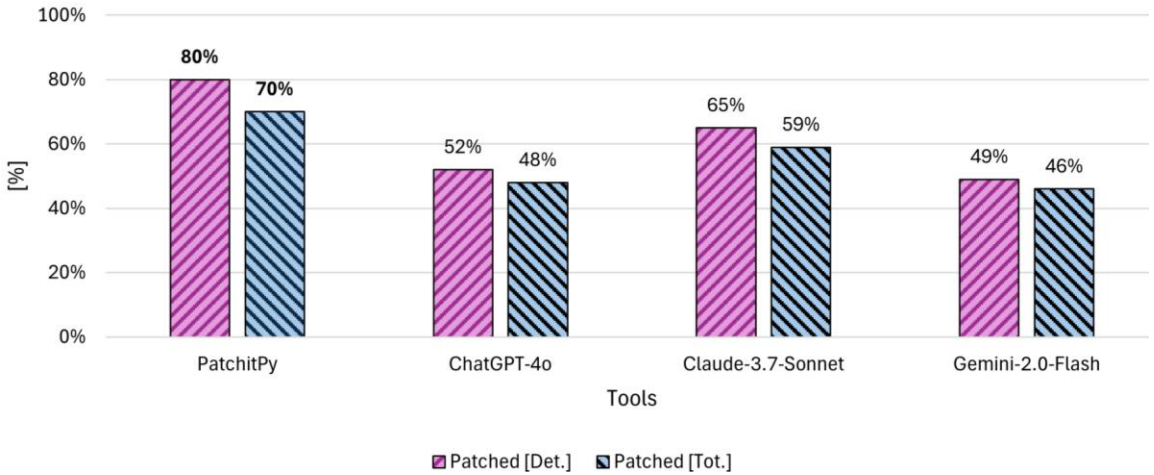
Patching strategy: PatchitPy

Code generation:
3 LLMs

Baseline:
Static analyzers and LLMs

Detection results:
F1 Score 93%
Accuracy 89%

Patching results:
Repair Rate 80%
Low code complexity



Vulnerability Injection

Code generation:
6 LLMs

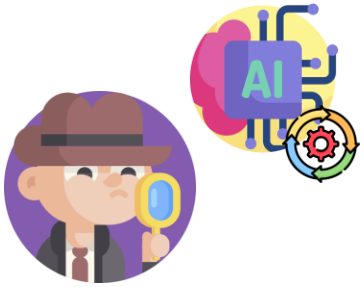
Similarity results:
Prompt without RAG + Finetuning

Semantic results:
Prompt with RAG + Pre-training

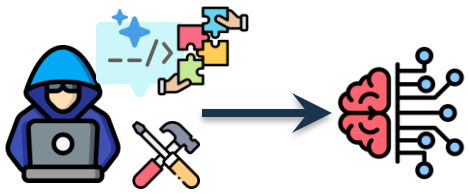
Models are **important**

Future Work

- Expand the vulnerability-injection pipeline by adding LLMs, with the goal of quantifying and characterizing model-dependent effects on the frequency of generated vulnerabilities.



- Extend the study to offensive code generation through advanced prompt-engineering strategies, to better understand how prompting techniques influence the models' ability to produce and control exploit-like behaviors.



Research products

[J1]	D. Cotroneo, R. De Luca, P. Liguori. *"DeVAIC: A tool for security assessment of AI-generated code",* **Information and Software Technology (IST) Journal, 2025** vol. 177, January 2025, 107572 Status: Published, Elsevier Publisher DOI: 10.1016/j.infsof.2024.107572
[C1]	F. Altiero, D. Cotroneo, R. De Luca, P. Liguori. "*Securing AI Code Generation Through Automated Pattern-Based Patching*", **55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), 2025** Naples, Italy, June 2025 Status: published, pp. 282-289, IEEE Publisher DOI: 10.1109/DSN-W65791.2025.00077

Tutorship

Tutorship for the “Impianti di Elaborazione” MSc course.

- Queuing theory
- Performance analysis lessons:
  - Workload characterization

Thank you for your attention

Contact:
roberta.deluca2@unina.it
DESSERT Lab – via Claudio 21