Università degli Studi di Napoli Federico II

itee PhD information technology electrical engineering

DIETI UNINA

Cristina Improta

Assessing and Enhancing the Trustworthiness of AI-based Code Generators

Tutor: Domenico Cotroneo

Cycle: XXXVIII Year: First

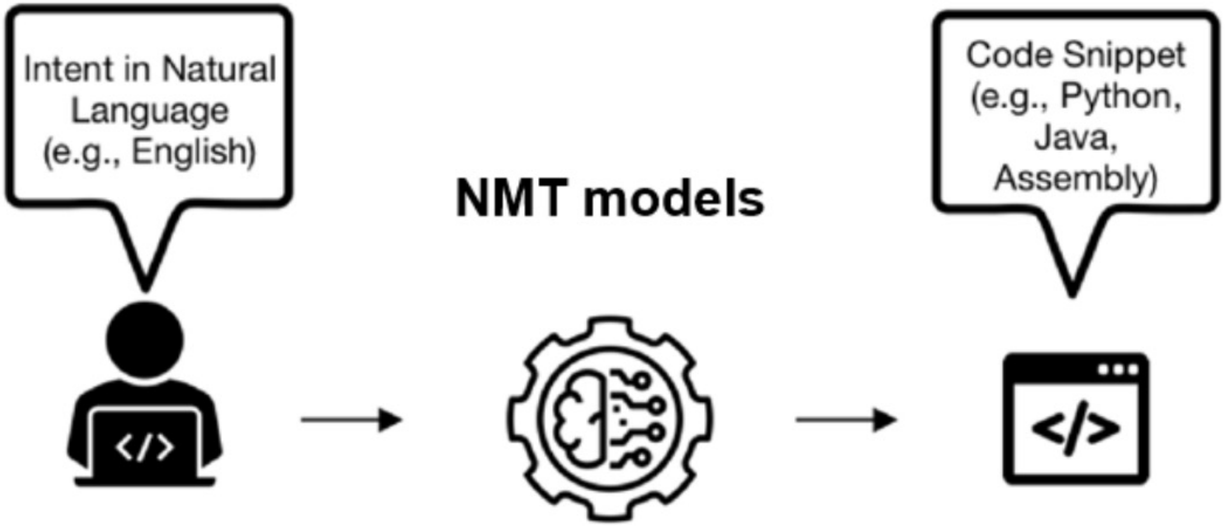
itee PhD information technology electrical engineering

My background

- MSc degree in Computer Engineering
- Research group: DESSERT
- PhD start date: 01/11/2022
- Scholarship type: UNINA

Research field of interest

- My research activity concerns the assessment and enhancement of the *trustworthiness* of AI code generators, i.e., AI-based solutions to automatically generate source code starting from natural language (NL) code descriptions.


Intent in Natural Language (e.g., English)
NMT models
Code Snippet (e.g., Python, Java, Assembly)
</>
</>

Summary of study activities

- **Ad hoc PhD courses / schools:**
  - 3rd International Software Engineering Summer School (SIESTA23)
  - 2023 Spring School on Transferable Skills
  - Using Deep Learning Properly
  - Virtualization technologies and their applications
  - IoT Data Analysis
- **Conferences / events attended**
  - *34th IEEE International Symposium on Software Reliability Engineering Conference* (ISSRE 2023). October 9-12 2023, Florence, Italy. *Presenting author*.

Research activity: Overview (1/4)

- **Problem**

AI code generators have become the state-of-the-art solution to increase productivity and speed up software development.

However, despite the efforts of the community to constantly improve these models, AI code generators still have limitations and potential drawbacks. For example, they may not always generate correct code as they may struggle with more complex programming tasks or ambiguous code descriptions. Furthermore, AI models are exposed to a wide variety of security issues targeting both their learning and inference process.

Research activity: Overview (2/4)

- **Objective**

Definition of a comprehensive methodology to assess and enhance the usability and trustworthiness of AI code generators in real-world scenarios, including their ability to generate **correct code**, their **robustness** to variable and ambiguous code descriptions, and their **security** to exploitation by malicious actors.

Research activity: Overview (3/4)

- **Methodology**
  - ✓ Development of a novel method to perturb NL code descriptions by substituting and omitting words in sentences to replicate the variability of different writing styles.
    - Applied to assess the robustness to perturbations
    - Applied to perform training data augmentation to increase robustness
  - ✓ Identification of the automatic textual similarity metric best suited to assess code correctness.

Research activity: Overview (4/4)

- **Methodology**
  - ✓ Development of a method to automatically evaluate to correctness of *security-oriented* code (i.e., assembly) that leverages symbolic execution to assess whether the behavior of an AI-generated program is semantically equivalent to that of a reference program.
  - ✓ Development of an attack strategy to assess the susceptibility of AI code generators to *data poisoning* attacks and evaluate the presence of vulnerabilities in the AI-generated code.

Products

[C1]	P. Liguori, C. Improta, S. De Vivo, R. Natella, B. Cukic, D. Cotroneo. *“Can NMT Understand Me? Towards Perturbation-based Evaluation of NMT Models for Code Generation”.* **IEEE/ACM 1st International Workshop on Natural Language-Based Software Engineering (NLBSE), 2022.** Status: published.
[J2]	P. Liguori, C. Improta, R. Natella, B. Cukic, D. Cotroneo. *“Who evaluates the evaluators? On automatic metrics for assessing AI-based offensive code generators”.* **Expert Systems with Applications Journal (ESWA), 2023.** Status: published.
[C3]	C. Improta. *“Poisoning Programs by Un-Repairing Code: Security Concerns of AI-generated Code”.* **1st IEEE International Workshop on Reliable and Secure AI for Software Engineering (ReSAISE23), ISSREW23, 2023**. Status: published.

Products

[P4]	R. Natella, P. Liguori, C. Improta, B. Cukic, D. Cotroneo. *“AI Code Generators for Security: Friend or Foe?”,* **IEEE Security & Privacy, 2023**. Status: under 2nd stage of review.
[C5]	C. Improta, P. Liguori, R. Natella, B. Cukic, D. Cotroneo. *“Assessing and Enhancing Robustness of AI Offensive Code Generators Via Natural Language Perturbations”,* **ACM International Conference on the Foundations of Software Engineering (FSE 2024), 2023.** Status: submitted.
[J6]	D. Cotroneo, C. Improta, P. Liguori, R. Natella. *“Automating the Correctness Assessment of AI-generated Code for Security Contexts”.* **Journal of Systems and Software (JSS), 2023.** Status: submitted.

Tutorship

Tutorship for the “Sistemi Operativi” BSc course .

- Process scheduling in Linux
- Inter-Process Communication in Linux
- POSIX threads