RICCARDO MIELE 2116946
MICHELE GUSELLA 2122861

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

LLMs FOR VULNERABILITY DETECTION

ADVANCED TOPICS IN COMPUTER AND NETWORK SECURITY 24/25

CONTENTS

UNIVERSITÀ DEGLI STUDI DI PADOVA

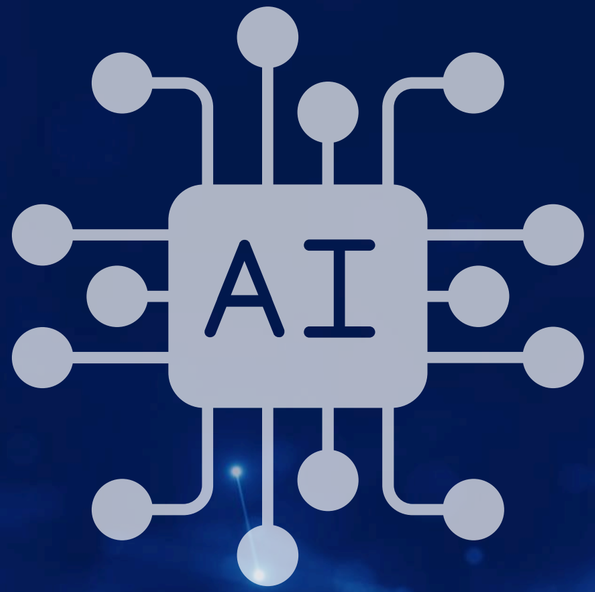
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

INTRODUCTION

LARGE LANGUAGE MODELS

Università degli Studi di Padova

Large Language Models (LLM) are AI systems trained on massive text data to generate, understand, and analyze language.



They work by predicting the next token based on prior context.

MAIN FEATURES

UNIVERSITÀ DEGLI STUDI DI PADOVA

Main characteristics of LLMs:

- **Attention mechanism**: let the model to better focus
- **Temperature**: a parameter that influences the randomness
- **Prompts techniques**: condition the answer

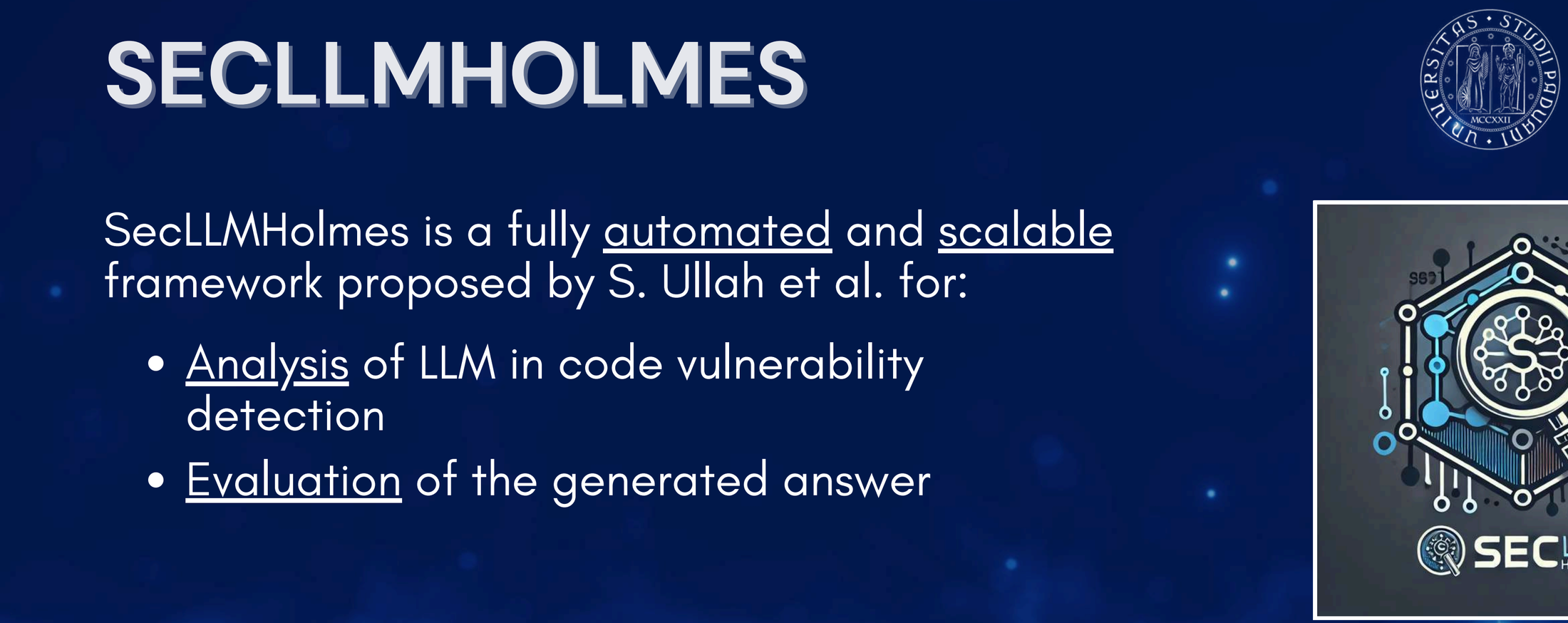
Main problems:

- **Hallucinations**: answer seems real but it is not

Università degli Studi di Padova

SECLLMHOLMES

SECLLMHOLMES



SecLLMHolmes is a fully automated and scalable framework proposed by S. Ullah et al. for:

- Analysis of LLM in code vulnerability detection
- Evaluation of the generated answer

S. Ullah, M. Han, S. Pujar, H. Pearce, A. Coskun, and G. Stringhini, "LLMs Cannot Reliably Identify and Reason About Security Vulnerabilities (Yet?): A Comprehensive Evaluation, Framework, and Benchmarks," in 2024 IEEE Symposium on Security and Privacy (SP), Los Alamitos, CA, USA: IEEE Computer Society, May 2024, pp. 862–880.

HOW IT WORKS

UNIVERSITÀ DEGLI STUDI DI PADOVA

The framework works in this way:

- File Selection
- File and prompt provided to the model
- Answer of the model
- Extraction of prediction and reasoning
- Information saved in json

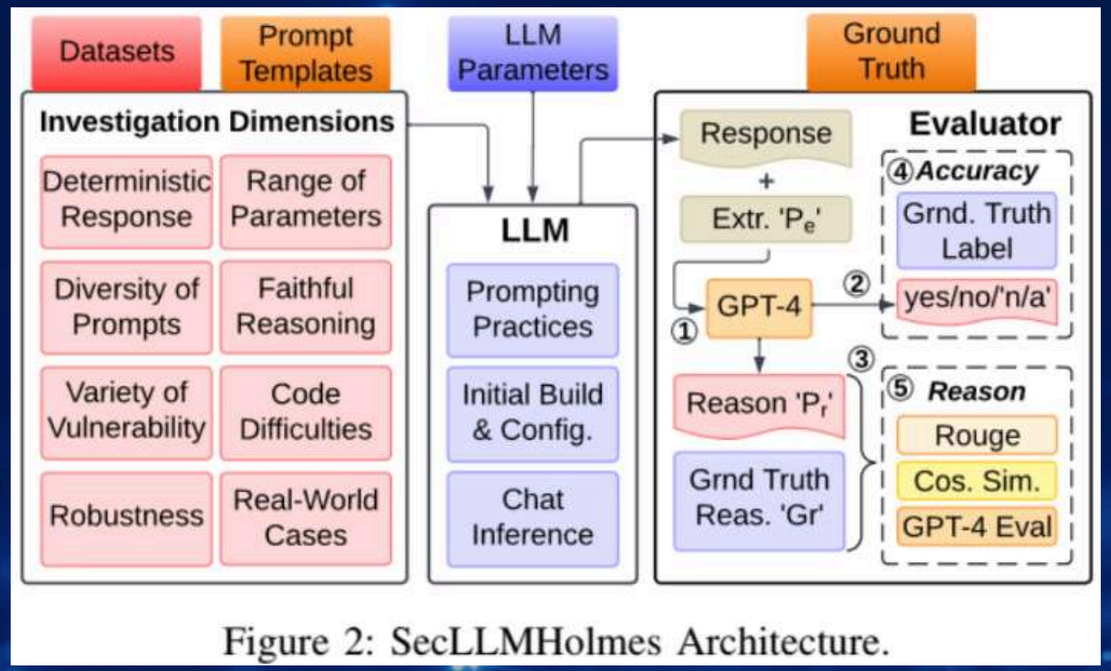


Figure 2: SecLLMHolmes Architecture.

PROMPTS & DATASET

UNIVERSITÀ DEGLI STUDI DI PADOVA

Two main points of the framework are prompts and dataset.

Prompts are the “instructions” passed to the model:

- Standard prompts
- Reasoning based and step-by-step prompts
- Definition based

Dataset contains all the vulnerable and not files:

- Hand-crafted
- Real scenarios
- Code augmentations

EXPERIMENTS

Università degli Studi di Padova

The framework performs five types of experiments:

- **Determinism**: to check consistency
- **Range Parameters**: to find best temperature value
- **Prompts**: to find best prompt
- **Code augmentations**: to verify robustness
- **Real Scenarios**: to test against real code

UNIVERSITÀ DEGLI STUDI DI PADOVA

EXPERIMENTAL SETUP

TOOLS USED

UNIVERSITÀ DEGLI STUDI DI PADOVA

The tools used in our study are:

- **SecLLMHolmes** framework
- **Ollama platform**
- **Ollama-python library**

MODIFICATIONS

Università degli Studi di Padova

We have done some modifications in order to use the framework.

- **Implementation** of the adapter
- **Addition** of ignore input variable
- **Change** of evaluation and reasoning model

METHODOLOGY

UNIVERSITÀ DEGLI STUDI DI PADOVA

Our experiment wanted to explore how small LLMs perform in comparison to larger models.

We have focused particularly in:

- **Llama3.2 3B**
- Experiments of original framework
- Hand-crafted dataset
- All prompts

UNIVERSITÀ DEGLI STUDI DI PADOVA

ANALYSIS AND RESULTS

EVALUATION OF RESULTS

UNIVERSITÀ DEGLI STUDI DI PADOVA

We could have chosen different metrics.

F1-score

- Useful but didn't account for correct reasoning

Ultimately we decided to use the same metric explained in the reference paper: **Accuracy**.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

PRELIMINARY EXPERIMENT

UNIVERSITÀ DEGLI STUDI DI PADOVA

The first experiment we run was **Determinism experiment** to check whether the model's responses were consistent.

We set k=5 and we found that

- Generally the responses were consistent
- On a few occasions one responses would differ from the others
- Limited to just one iteration per file

FIRST EXPERIMENT

UNIVERSITÀ DEGLI STUDI DI PADOVA

Then we run **Prompts experiment** and **Range of parameters experiment** which respectively:

- Evaluated the model using all the prompts on all the CWEs
- Evaluated the model fixing some prompts/CWEs but changing temperatures

TABLE III
ACCURACY OF LLAMA3.2 AND GPT-4 ON THEIR BEST PROMPTS

Model	Prompt	Accuracy
Llama3.2	PromptR5	27%
Llama3.2	PromptD2	23%
GPT-4	PromptR2	89.5%
GPT-4	PromptR6	89.5%

SECOND EXPERIMENT

UNIVERSITÀ DEGLI STUDI DI PADOVA

We run the model again three times, fixing its **two best prompts** and **best temperature**.

TABLE IV
SORTED ACCURACY RESULTS ACROSS THREE RUNS OF LLAMA3.2

File	Run 1	Run 2	Run 3
1.c	33.33%	58.33%	58.33%
1.py	50%	50%	25%
3.py	0%	50%	0%
3.c	33.33%	16.67%	33.33%
2.py	0%	0%	25%
2.c	0%	18.18%	8.33%
p_3.c	0%	8.33%	0%
p_1.c	0%	0%	0%
p_2.c	0%	0%	0%
p_1.py	0%	0%	0%
p_2.py	0%	0%	0%
p_3.py	0%	0%	0%

TABLE V
CATEGORY-WISE ACCURACY RESULTS OF LLAMA3.2

Category	Run 1	Run 2	Run 3
Vulnerable	20.83%	31.91%	29.17%
Non-Vulnerable	0%	2.17%	0%

TABLE VI
GPT-4 MODEL RESULTS SORTED BY ACCURACY

File	Accuracy
p_1.py	100%
p_2.py	94.12%
3.c	92.16%
3.py	91.18%
1.c	84%
1.py	82.35%
2.py	75%
p_1.c	74.26%
p_2.c	72.55%
2.c	68.32%
p_3.c	51.96%
p_3.py	47.06%
Vulnerable	81.89%
Non-Vulnerable	69.78%

GPT-4 COMPARISON

UNIVERSITÀ DEGLI STUDI DI PADOVA

First of all GPT-4 is a large model with an estimated amount of 1.8 trillion parameters, whereas Llama3.2 with only 3 billion parameters is considerably smaller.

Llama3.2 has three main problems:

- **Overstimation** of risk
- **Misinterpretation** of Context
- **Superficial** Security Fixes

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

CONCLUSION

CONCLUSION

UNIVERSITÀ DEGLI STUDI DI PADOVA

Objective: how well smaller LLMs can detect vulnerabilities and explain their reasoning.

- Llama3.2
- SecLLMHolmes framework
- GPT-4

We found:

- Smaller models can detect some vulnerabilities
- Much less accurate and consistent

CONCLUSION

Università degli Studi di Padova

Further advancements are needed

- Model training techniques
- Dataset optimization

Potential goal: Run models locally, for example, within an IDE, to automatically detect vulnerable code as the programmer is writing it.

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

THANKS FOR YOUR ATTENTION

ANY QUESTIONS?