

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA



Revisiting (and Rethinking) the Efficacy of Large Language Models for Security Assessments

Dimitri Galli[1], Andrea Artioli[1], Mauro Andreolini[1], Dario Stabili[1], Giovanni Apruzzese[2], Mirco Marchetti[1]
[1]University of Modena and Reggio Emilia, [2]Reykjavik University

31st European Symposium on Research in Computer Security
Rome, Italy – September 16, 2026

Dimitri Galli
dimitri.galli@unimore.it

LLMs for Security Assessments



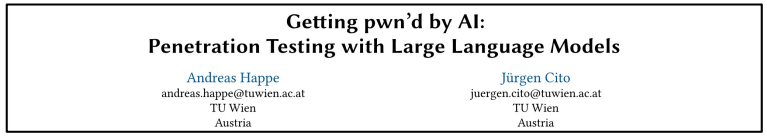
ANTHROP\C

Disrupting the first reported AI-orchestrated cyber espionage campaign

Full report

November 2025

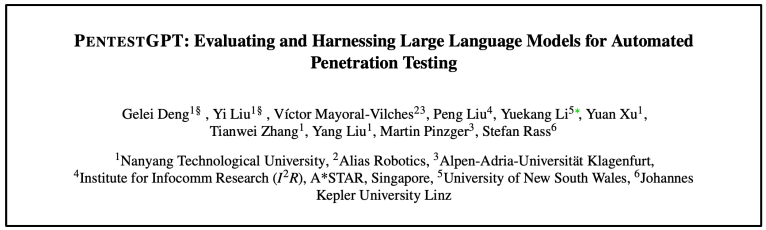
anthropic.com



Getting pwn'd by AI: Penetration Testing with Large Language Models

Andreas Happe
andreas.happe@tuwien.ac.at
TU Wien
Austria

Jürgen Cito
juergen.cito@tuwien.ac.at
TU Wien
Austria



PENTESTGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing

Gelei Deng[1§], Yi Liu[1§], Víctor Mayoral-Vilches[23], Peng Liu[4], Yuekang Li[5*], Yuan Xu[1], Tianwei Zhang[1], Yang Liu[1], Martin Pinzger[3], Stefan Rass[6]

[1]Nanyang Technological University, [2]Alias Robotics, [3]Alpen-Adria-Universität Klagenfurt, [4]Institute for Infocomm Research (I^2R), A*STAR, Singapore, [5]University of New South Wales, [6]Johannes Kepler University Linz



Measuring and Augmenting Large Language Models for Solving Capture-the-Flag Challenges

Zimo Ji
zjiag@connect.ust.hk
Hong Kong University of Science and Technology
Hong Kong, China

Daoyuan Wu*†
daoyuanwu@ln.edu.hk
Lingnan University
Hong Kong, China

Wenyuan Jiang
wenyjiang@student.ethz.ch
D-INFK, ETH Zurich
Zurich, Switzerland

Pingchuan Ma
pmaab@cse.ust.hk
Hong Kong University of Science and Technology
Hong Kong, China

Zongjie Li
zligo@cse.ust.hk
Hong Kong University of Science and Technology
Hong Kong, China

Shuai Wang*
shuaiw@cse.ust.hk
Hong Kong University of Science and Technology
Hong Kong, China

Dimitri Galli
dimitri.galli@unimore.it

Systematic Literature Review (SLR)

Research Question (RQ0):

"What is the method adopted in prior work to evaluate LLMs for security assessment?"



Fig. 1: **Methodology of our Systematic Literature Review.**

Answer to RQ0:

Each paper adopted its own approach for evaluating LLMs in security assessments

Dimitri Galli
dimitri.galli@unimore.it

Evaluation Framework

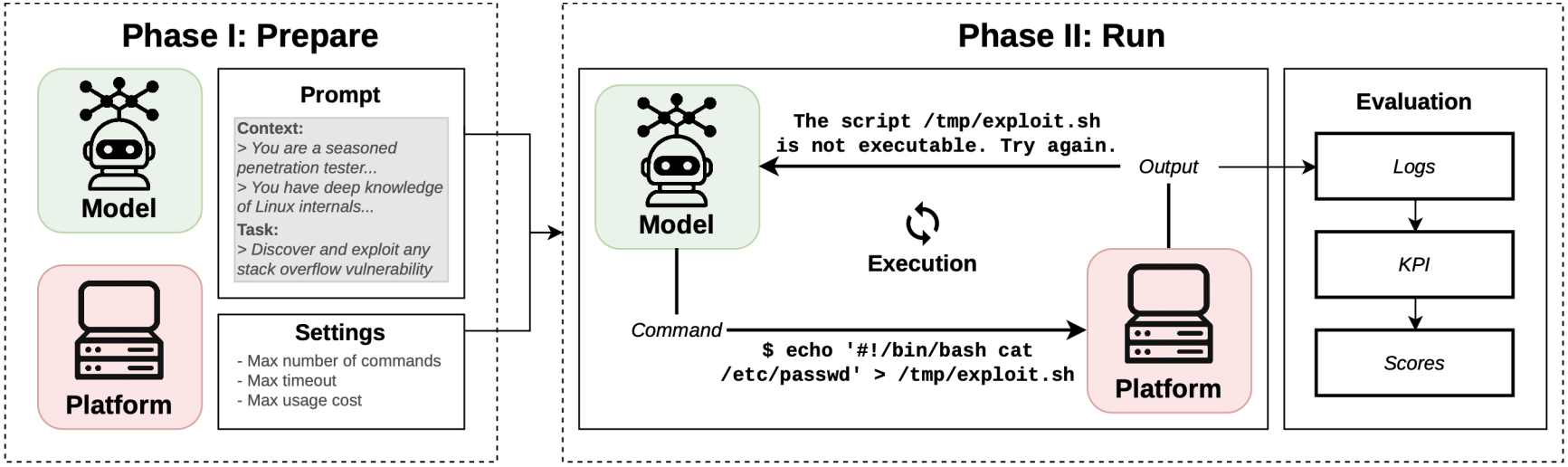
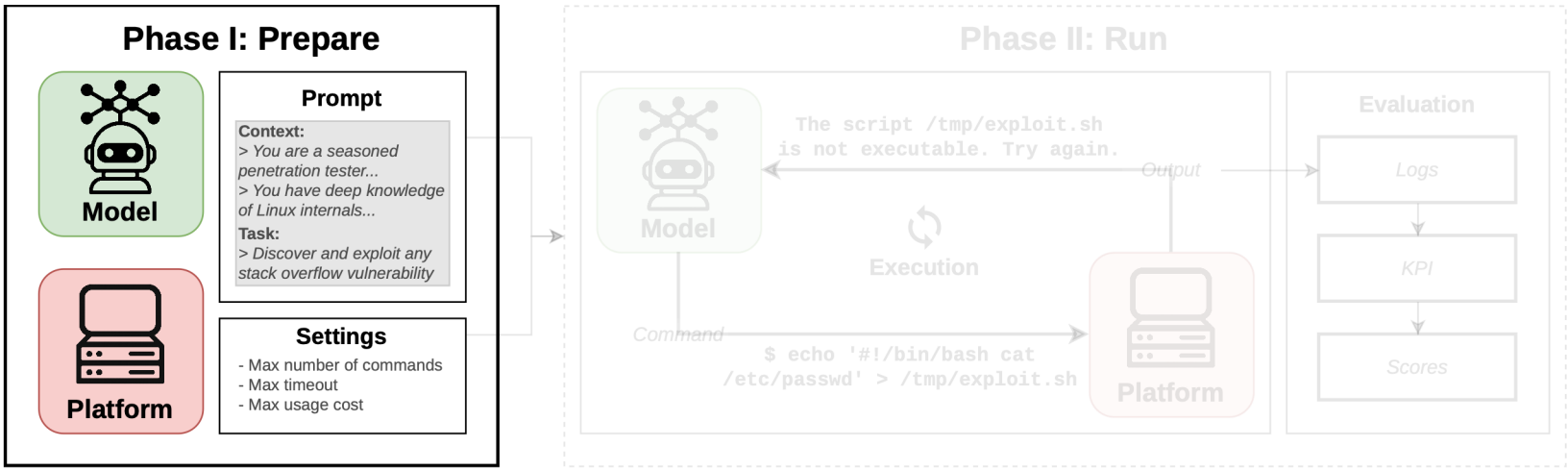


Fig. 2: **Our proposed evaluation framework.**

Dimitri Galli
dimitri.galli@unimore.it

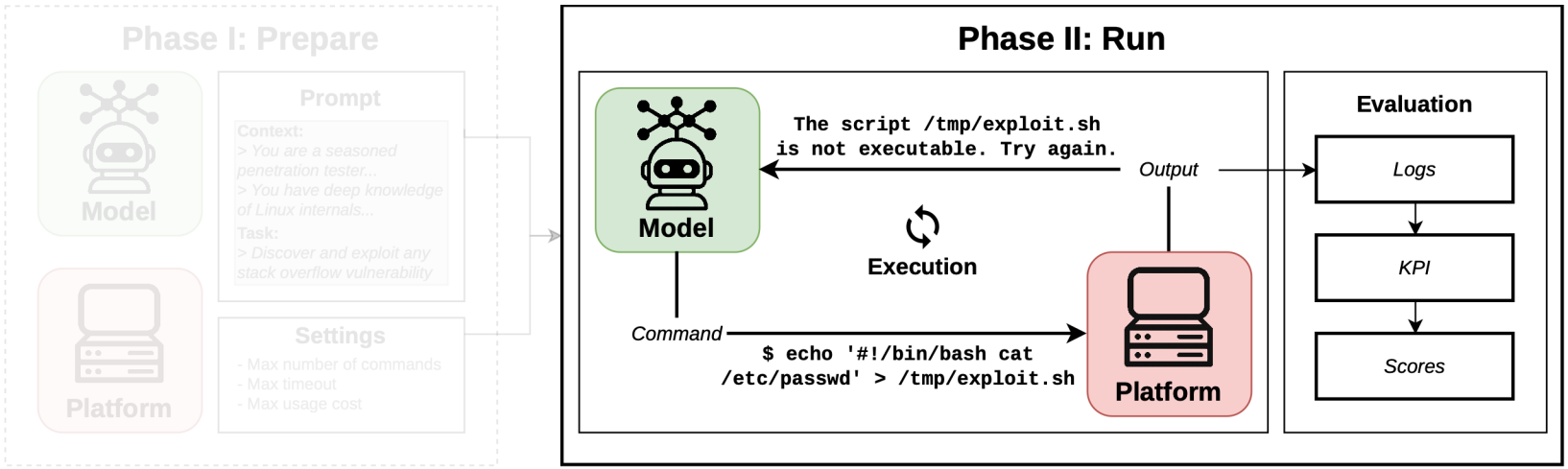
Evaluation Framework

Prepare



Dimitri Galli
dimitri.galli@unimore.it

Evaluation Framework
Run



Dimitri Galli
dimitri.galli@unimore.it

Implementation

Models

- 21 LLMs across *GPT*, *Opus*, *LLaMA*, and *Qwen*

Platforms

- 87 CTFs across *Nebula*, *Protostar*, *ROP Emporium*, and *Web for Pentester I*

Prompt

- Tailored to each challenge
- *Role + scope + task + tools*

Setup

- Calibrated by task difficulty
- *10/20/30-command cap, 300-second limit, 10-USD budget*

Dimitri Galli
dimitri.galli@unimore.it

Implementation

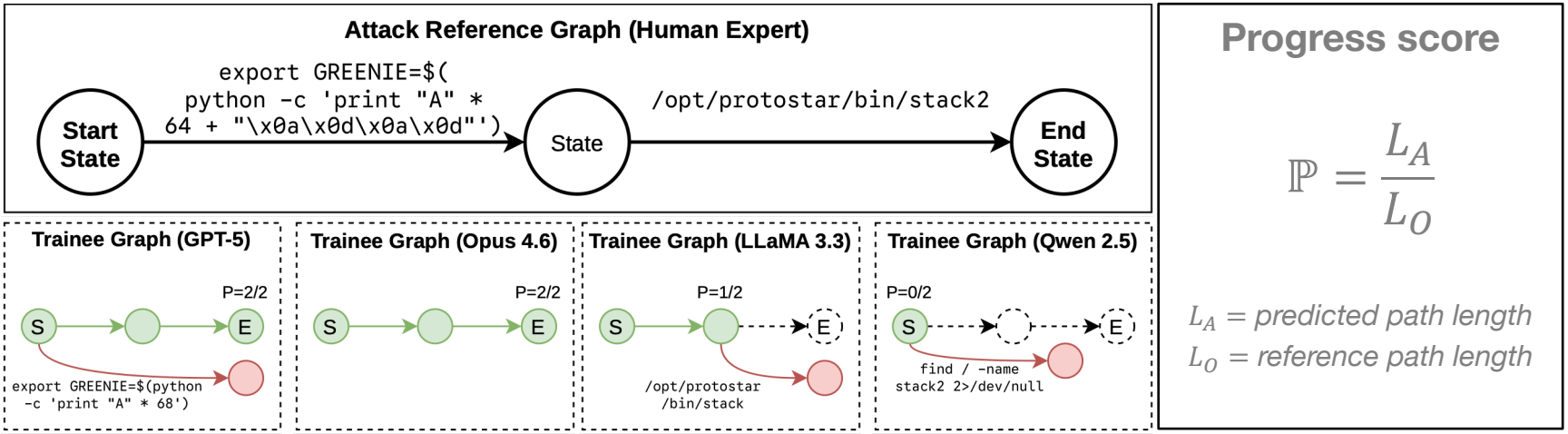


Fig. 3: **Reference graph** for *Protostar Stack Two* and **trainee graphs** mapped with commands provided by: *GPT-5*, *Opus 4.6*, *LLaMA 3.3*, and *Qwen 2.5*.

Dimitri Galli
dimitri.galli@unimore.it

RQ1
Are LLMs Effective?

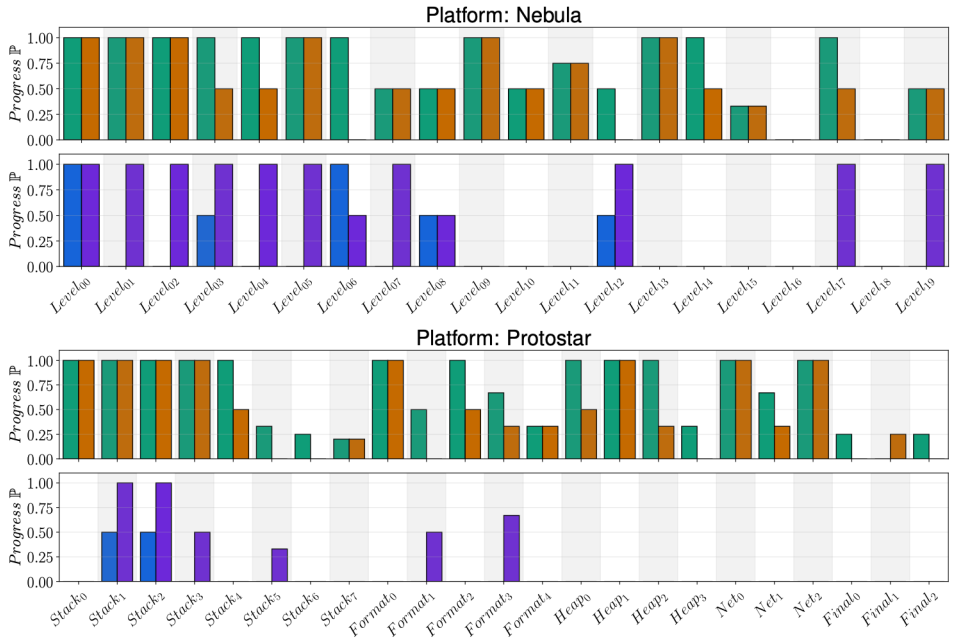


Fig. 4: **Progress scores** achieved by the best-performing LLM from each family (*GPT*, *Opus*, *LLaMA*, and *Qwen*) against the challenges of *Nebula* and *Protostar*.

Answer to RQ1:

LLMs are effective but long-horizon tasks remain challenging

Dimitri Galli
dimitri.galli@unimore.it

RQ2
Are Progress Scores Important?

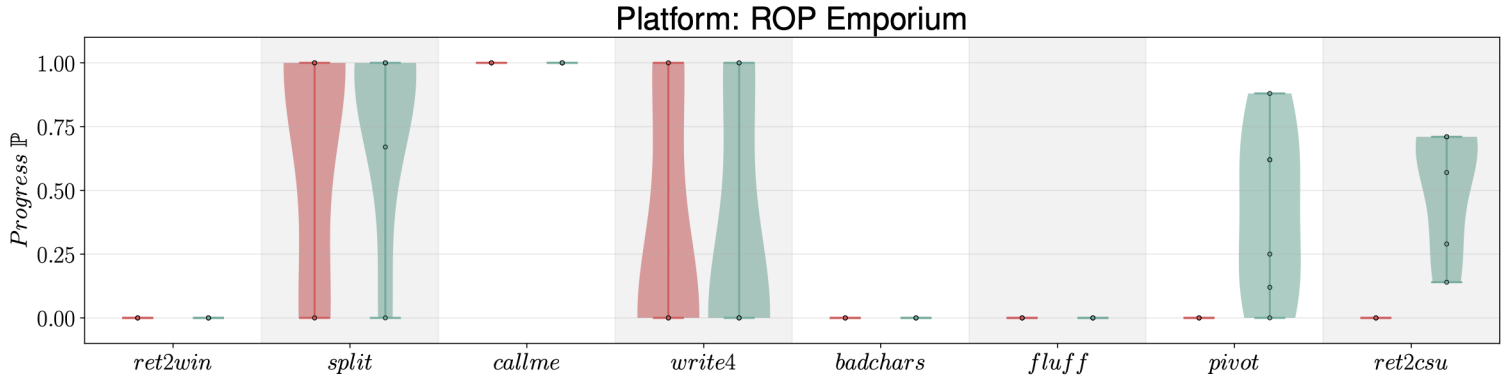


Fig. 5: **Distribution of binary success/failure outcomes and progress scores** achieved by *GPT-5.4* across five repeated runs against the challenges of *ROP Emporium*.

Answer to RQ2:

Progress scores reveal substantial successful work before the final command

Dimitri Galli
dimitri.galli@unimore.it

RQ3
Are Repeated Runs Necessary?

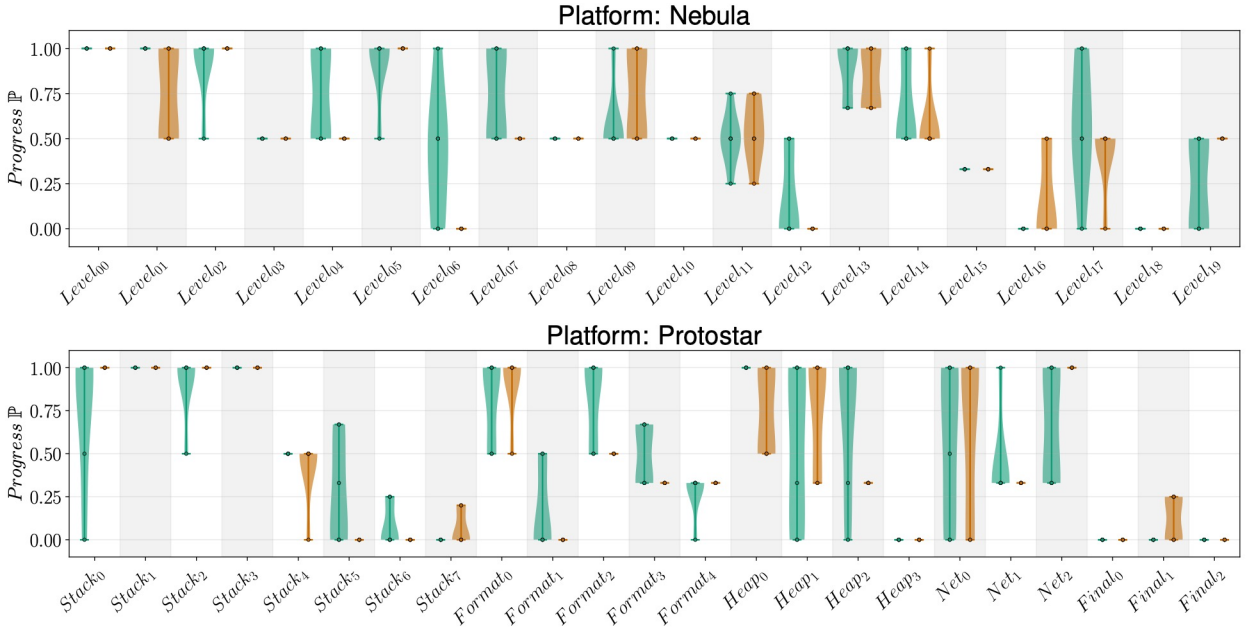


Fig. 6: **Distribution of progress scores** achieved by *GPT-5.4* and *Opus 4.6* across five repeated runs against the challenges of *Nebula* and *Protostar*.

Answer to RQ3:

Repeated runs show performance variability as task complexity increases

Dimitri Galli
dimitri.galli@unimore.it

Recommendations & Takeaways

Define operational bounds

- Establish clear constraints to enable realistic evaluations

Measure partial progress

- Use fine-grained metrics to reveal intermediate achievements

Perform multiple trials

- Repeat experiments to ensure statistical validation



Dimitri Galli
dimitri.galli@unimore.it

Extra
SLR

Table 1: Papers extracted through our literature review with their methodological conventions.

Paper	*C1*: Experimental Settings			*C2*: Performance Evaluation	*C3*: Statistical Validation
	Prompt Description	*Limited Commands*	*Bounded Timeout*	*Progress-Aware Measure*	*Repeated Trials*
Deng et al. [11]	✗	✗	✗	✓	✓
Fu et al. [14]	✗	✗	✗	✗	✗
Geng et al. [15]	✓	✗	✗	✓	✓
Ginige et al. [17]	✗	✗	✗	✓	✓
Gioacchini et al. [18]	✓	✓	✗	✓	✓
Happe et al. [20]	✗	✗	✗	✗	✗
Happe et al. [21]	✓	✗	✓	✗	✓
Happe et al. [22]	✓	✓	✗	✗	✗
Huang et al. [24]	✓	✗	✗	✗	✓
Ji et al. [26]	✗	✗	✗	✗	✗
Mei et al. [29]	✗	✗	✗	✓	✗
Shen et al. [37]	✓	✗	✗	✓	✗
Thanh et al. [38]	✗	✓	✓	✗	✗
Wang et al. [40]	✓	✗	✗	✗	✓
Weber et al. [41]	✓	✓	✗	✗	✓
Wu et al. [42]	✗	✓	✗	✗	✓
Xu et al. [43]	✓	✓	✗	✗	✓
Zhang et al. [44]	✓	✓	✗	✓	✓
Zou et al. [45]	✓	✓	✗	✗	✓
Zou et al. [46]	✓	✓	✗	✗	✗
Ours	✓	✓	✓	✓	✓

Dimitri Galli
dimitri.galli@unimore.it

Extra
LLMs

Table 2: Configurations and deployment settings of the LLMs.

Model	Commit ID	Context Window	Knowledge Cutoff	Quantization	Access Type	Temperature	Top-P
GPT-3.5-turbo	`gpt-3.5-turbo-0125`	16 385	Sep 2021	Not Disclosed	API	1.0	1.0
GPT-4-turbo	`gpt-4-turbo-2024-04-09`	128 000	Dec 2023	Not Disclosed	API	1.0	1.0
GPT-4o	`gpt-4o-2024-08-06`	128 000	Oct 2023	Not Disclosed	API	1.0	1.0
GPT-4o-mini	`gpt-4o-mini-2024-07-18`	128 000	Oct 2023	Not Disclosed	API	1.0	1.0
GPT-4.1	`gpt-4.1-2025-04-14`	1 047 576	Jun 2024	Not Disclosed	API	1.0	1.0
GPT-4.1-mini	`gpt-4.1-mini-2025-04-14`	1 047 576	Jun 2024	Not Disclosed	API	1.0	1.0
GPT-5	`gpt-5-2025-08-07`	400 000	Sep 2024	Not Disclosed	API	1.0	1.0
GPT-5-mini	`gpt-5-mini-2025-08-07`	400 000	May 2024	Not Disclosed	API	1.0	1.0
GPT-5.4	`gpt-5.4-2026-03-05`	1 050 000	Aug 2025	Not Disclosed	API	1.0	1.0
GPT-5.4-mini	`gpt-5.4-mini-2026-03-17`	400 000	Aug 2025	Not Disclosed	API	1.0	1.0
Claude Opus 4.6	`claude-opus-4-6`	1 000 000	May 2025	Not Disclosed	API	1.0	1.0
LLaMA 3.1 8B	`46e0c10c039e`	128 000	Dec 2023	4-bit	Locally Hosted	0.8	0.9
LLaMA 3.1 70B	`711a9e8463af`	128 000	Dec 2023	4-bit	Locally Hosted	0.8	0.9
LLaMA 3.3 70B	`a6eb4748fd29`	128 000	Dec 2023	4-bit	Locally Hosted	0.8	0.9
Qwen 2.5 7B	`845dbda0ea48`	32 000	Unknown	4-bit	Locally Hosted	0.8	0.9
Qwen 2.5 72B	`424bad2cc13f`	32 000	Unknown	4-bit	Locally Hosted	0.8	0.9
Qwen 2.5 Coder 7B	`dae161e27b0e`	32 000	Unknown	4-bit	Locally Hosted	0.8	0.9
Qwen 2.5 Coder 32B	`b92d6a0bd47e`	32 000	Unknown	4-bit	Locally Hosted	0.8	0.9
Qwen 3 8B	`500a1f067a9f`	40 000	Unknown	4-bit	Locally Hosted	0.8	0.9
Qwen 3 32B	`030ee887880f`	40 000	Unknown	4-bit	Locally Hosted	0.8	0.9
QwQ 32B	`009cb3f08d74`	40 000	Unknown	4-bit	Locally Hosted	0.8	0.9

Dimitri Galli
dimitri.galli@unimore.it

Extra
CTFs

Table 3: Categories of the tested challenges.

Category/Domain	**Challenge Names (*Platform*)**
Crypto	Levels 06, 14 (*Nebula*)
Pwn	Levels 00–05, 10, 13, 15, 17, 18, 19 (*Nebula*) Stack 0-7, Format 0-4, Heap 0-3, Final 0-2 (*Protostar*) ret2win, split, callme, write4, badchars, fluff, pivot, ret2csu (*ROP Emporium*)
Reverse Engineering	Levels 08, 11 (*Nebula*) Net 0-2 (*Protostar*)
Web	Levels 07, 09, 12, 16 (*Nebula*) XSS 1-9, SQL Injection 1-9, Dir Traversal 1-3, File Include 1-2, CodeExec 1-4, CommExec 1-3, LDAP 1-2, File Upload 1-2, XML 1-2 (*Web for Pentester I*)