

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
Ministry of Higher Education and Scientific Research
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbas



Graduation Thesis

To obtain the diploma of **Engineering Degree**

Field of Study: **Computer Science**

Specialization: **ISI**

Theme

Security Testing Framework for Large Language Models (LLMs)

Presented by
Aomar Mohamed Abdelillah

Defended on: **[June, 2026]**
In front of the jury composed of

Dr. KAZI TANI Mohammed Yassine
Dr. BENDAOU Fayssal
Ms. BOUHAIR Maria
Dr. BEDJAoui Mohammed

President of the Jury
Thesis Supervisor
Co-Supervisor
Examiner

Academic Year: 2025/2026

Abstract

Large Language Models are now embedded in assistants, coding tools, document analysis systems, and retrieval-based applications. Their adoption has been fast. Their security assessment has not kept the same pace. Most deployments still rely on ad hoc evaluations, scattered jailbreak trials, or provider-side safeguards that are difficult to audit from outside the service boundary.

This thesis studies the security of LLM-based systems through the design of a practical testing framework for prompt-driven attacks. The work is not limited to one technique. It examines the technical foundations of LLMs, their deployment patterns, and the attack surfaces introduced by system prompts, retrieval pipelines, external tools, and API-based integration. It then reviews the main threat classes discussed in recent literature, including prompt injection, jailbreak attacks, data leakage, inference attacks, adversarial inputs, denial-of-service through prompt abuse, and supply-chain weaknesses. Existing defensive measures are also analyzed, such as input filtering, output moderation, alignment methods, prompt hardening, and least-privilege design.

The practical contribution of the project is a modular security testing workflow that supports multiple attack paths. PAIR is implemented as an iterative jailbreak engine that refines adversarial prompts across multiple queries. PISmith is integrated as a prompt-injection component for generating injection payloads through a dedicated model interface. The framework also includes data leakage checks based on system prompt exposure, canary-token detection, and structured evaluator output. This broader design allows jailbreak behavior, prompt injection, and leakage-oriented failures to be tested within the same campaign and persistence model.

The study defines evaluation criteria suitable for repeatable testing, including attack success rate, refusal rate, classification error, robustness, severity scoring, and similarity-based assessment when outputs cannot be judged through a strict binary rule. The outcome is not a final defense, but a structured method for testing LLM security in a repeatable and comparable way. Applying it across different models, prompts, and attack modules makes it possible to identify where safeguards hold, where they fail, and which parts of the testing workflow require further improvement. **Keywords**— large language models, LLM security, jailbreak attacks, prompt injection, PAIR, PISmith, red-teaming