

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



Master Thesis

To obtain the diploma of **Master**
Field of Study: **Computer Science**
Specialization: **ISI**

Theme

Red Teaming Large Language Models: A Study Of Jailbreaks, Prompt Injection and Data Leakage

Presented by
Aomar Mohamed Abdelillah

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

Dr. KAZI TANI Mohammed Yassine
Dr. BENDAOUD Fayssal
Ms. BOUCHAIR Maria
Dr. BEDJAoui Mohammed

President of the Jury
Thesis Supervisor
Co-Supervisor
Examiner

Academic Year: 2025/2026

Abstract

Large Language Models (LLMs) are now used in search engines, chat assistants, coding tools, document analysis systems, and autonomous agents. Their use creates new security and safety problems because natural language becomes both the input data and the instruction channel. An attacker can try to bypass safety rules, override system instructions, or extract hidden information from the model context. These risks are usually studied through jailbreak attacks, prompt injection, adversarial prompting, and data leakage.

This thesis presents a theoretical study of red teaming for LLMs. It does not implement a complete practical framework. The goal is to review existing research, explain the main attack techniques, compare their assumptions, and discuss the limits of current evaluation methods. The study focuses on techniques such as Greedy Coordinate Gradient (GCG), Prompt Automatic Iterative Refinement (PAIR), AutoDAN, PromptInject, indirect prompt injection, JailbreakBench, and recent reinforcement learning based approaches such as PI-Smith.

The thesis shows that LLM red teaming evolved from manual prompt attacks to automated and optimization-based methods. White-box attacks such as GCG can produce strong adversarial suffixes, but they require model access that is not always available. Black-box methods such as PAIR are more realistic for commercial APIs, but they depend on query budgets and judge quality. Prompt injection attacks create a separate risk because they tar-

get the boundary between trusted instructions and untrusted content. Data leakage remains difficult to evaluate because success can be partial and context-dependent.

The main finding is that current LLM red teaming still lacks stable evaluation standards. Attack success rates are hard to compare across papers because datasets, judges, threat models, and refusal criteria differ. Future work should focus on reproducible evaluation, clearer threat models, and safer automated testing. The thesis ends by suggesting two possible directions: using PAIR as a lightweight black-box optimization method, and designing a smaller adaptation of PISmith for prompt injection testing.

Keywords— Large Language Models, Red Teaming, Jailbreaks, Prompt Injection, Data Leakage, Adversarial Prompting, PAIR, PISmith, LLM Safety

ملخص

تُستعمل نماذج اللغة الكبيرة اليوم في محركات البحث، والمساعدات الحوارية، وأدوات البرمجة، وأنظمة تحليل الوثائق، والوكلاء المستقلين. ويؤدي هذا الاستعمال إلى ظهور مشكلات جديدة في الأمن والسلامة، لأن اللغة الطبيعية تصبح في الوقت نفسه بيانات إدخال وقناة لإعطاء التعليمات. لذلك يمكن للمهاجم أن يحاول تجاوز قواعد السلامة، أو إلغاء تعليمات النظام، أو استخراج معلومات مخفية من سياق النموذج. وتُدرس هذه المخاطر عادة من خلال هجمات كسر القيود، وحقن التعليمات، والتحفيز العدائي، وتسريب البيانات.

تقدم هذه الأطروحة دراسة نظرية حول الاختبار الهجومي لنماذج اللغة الكبيرة. ولا تهدف إلى تنفيذ إطار عملي كامل، بل تهدف إلى مراجعة الأبحاث الموجودة، وشرح تقنيات الهجوم الأساسية، ومقارنة افتراضاتها، ومناقشة حدود طرق التقييم الحالية. وتركز الدراسة على تقنيات مثل (GCG)، Gradient Coordinate Greedy، (PAIR)، Refinement Iterative Automatic Prompt، و (AutoDAN)، و PromptInject، وحقن التعليمات غير المباشر، و JailbreakBench، إضافة إلى المقاربات الحديثة المعتمدة على التعلم التعزيزي مثل PISmith.

تُظهر الأطروحة أن الاختبار الهجومي لنماذج اللغة الكبيرة تطور من هجمات يدوية تعتمد على كتابة التعليمات إلى طرق آلية ومعتمدة على التحسين. فالهجمات ذات الوصول الأبيض مثل GCG يمكن أن تنتج لواحق عدائية قوية، لكنها تحتاج إلى وصول إلى النموذج لا يكون متاحاً دائماً. أما الطرق ذات الوصول الأسود مثل PAIR فهي أكثر واقعية بالنسبة لواجهات البرمجة التجارية، لكنها تعتمد على ميزانية الاستعلامات وجودة المُقيم. وتمثل هجمات حقن التعليمات خطراً منفصلاً لأنها تستهدف الحد الفاصل بين التعليمات الموثوقة والمحتوى غير الموثوق. كما يبقى تقييم تسريب البيانات صعباً لأن النجاح قد يكون جزئياً ومعتمداً على السياق.

تتمثل النتيجة الرئيسية في أن الاختبار الهجومي الحالي لنماذج اللغة الكبيرة ما زال يفتقر إلى معايير تقييم مستقرة. فمن الصعب مقارنة معدلات نجاح الهجمات بين الأبحاث بسبب اختلاف مجموعات البيانات، والمُقيمين، ونماذج التهديد، ومعايير الرفض. لذلك ينبغي أن يركز العمل المستقبلي على التقييم القابل لإعادة الإنتاج، وتحديد نماذج تهديد أوضح، وتطوير اختبارات آلية أكثر أماناً. وتختتم الأطروحة باقتراح اتجاهين ممكنين: استعمال PAIR كطريقة خفيفة للتحسين في بيئة ذات وصول أسود، وتصميم نسخة أصغر مكيفة من PISmith لاختبار حقن التعليمات.

الكلمات المفتاحية--- نماذج اللغة الكبيرة، الاختبار الهجومي، كسر القيود، حقن التعليمات، تسريب البيانات، التحفيز العدائي،

PISmith، PAIR، سلامة نماذج اللغة الكبيرة