

**Design and Evaluation of an Autonomous Multi-Strategy AI Red Team Agent for
Security Assessment of Mistral and Llama Large Language Models & Retrieval-
Augmented Systems**

By

Yunan Lasito
200001169

MASTER'S DEGREE
in

MASTER OF INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY



SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

August 2026
Revision after the Thesis Defense on August 7th 2026

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and that, to the best of my knowledge, it contains no material previously published or written by another person, nor material which has, to a substantial extent, been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in this thesis.



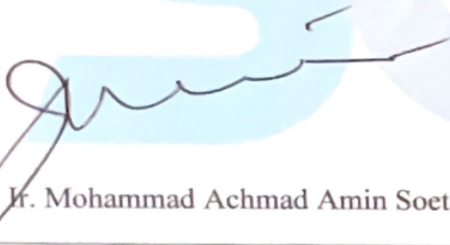
Yunan Lasito

Student

15 August 2026

Date

Approved by:



Dr. Ir. Mohammad Achmad Amin Soetomo, M.Sc.

Thesis Advisor

Aug 15th '26

Date



Dr. Semi Yulianto, S.E., M.Kom

Thesis Co-Advisor

20 Aug 2026

Date

Dr. Maulahikmah Galinium, S.Kom, M.Sc.

Dean

Date

Yunan Lasito

ABSTRACT

DESIGN AND EVALUATION OF AN AUTONOMOUS MULTI-STRATEGY AI RED TEAM AGENT FOR SECURITY ASSESSMENT OF MISTRAL AND LLAMA LARGE LANGUAGE MODELS & RETRIEVAL-AUGMENTED SYSTEMS

By

Yunan Lasito

Dr. Ir. Mohammad Achmad Amin Soetomo, M.Sc.

Dr. Semi Yulianto, S.E., M.Kom.

SWISS GERMAN UNIVERSITY

Large Language Models (LLMs) face security risks from jailbreak, prompt injection, and retrieval-layer attacks, while existing red-teaming approaches remain fragmented. This thesis designed and evaluated an autonomous multi-strategy AI red-team framework integrating mutation-based fuzzing, genetic optimisation, compositional DSL synthesis, and prompt extraction. The framework was tested against Mistral-7B-Instruct v0.3 and Llama-3-8B-Instruct through Ollama, together with a LangChain-FAISS RAG pipeline, across 2,811 adversarial attempts. Results showed that multi-strategy pooling did not outperform the strongest individual strategies, reducing Attack Success Rate by 20.37–28.33 percentage points ($p < .001$). Cross-model transferability was only 1.31%, while retrieved poisoned content did not result in successful RAG attacks. Exploit stability varied substantially across strategies. The framework completed all experiments on local 24 GB hardware without critical failure, while human validation showed low agreement with automated classification (Cohen's $\kappa = 0.20$). Overall, the results demonstrate the feasibility of autonomous local LLM red teaming while highlighting the importance of strategy-aware attack selection and human validation.

Keywords: LLM security, red teaming, jailbreak, prompt injection, retrieval-augmented generation

Yunan Lasito



© Copyright 2026
by Yunan Lasito
All rights reserved

DEDICATION

I dedicate this work to the future of the country I love: Indonesia



ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis advisor, Dr. Ir. Mohammad Achmad Amin Soetomo, M.Sc., for his continuous guidance, patience, and constructive direction throughout the design, execution, and writing of this thesis. His insistence on methodological rigor shaped every chapter of this work, from the earliest proposal drafts to the final results.

I am equally grateful to my co-advisor, Dr. Semi Yulianto, S.E., M.Kom., whose careful and exacting review of each revision substantially strengthened the empirical grounding and scientific integrity of this research. His willingness to verify every claim against the manuscript gave me confidence that the final work meets the standard expected of it.

I am thankful for their support and belief in this research throughout its completion.

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR	2
ABSTRACT	3
DEDICATION	5
ACKNOWLEDGEMENTS	6
TABLE OF CONTENTS	7
LIST OF FIGURES.....	11
LIST OF TABLES	12
CHAPTER 1 – INTRODUCTION	14
1.1 Background	14
1.2 Research Problems	16
1.3 Research Objectives	17
1.4 Research Questions	17
1.5 Hypothesis	18
1.6 Significance of Study	18
1.7 Scope and Limitations	20
1.8 Thesis Structure	21
CHAPTER 2 - LITERATURE REVIEW	22
2.1 Conceptual and Risk Foundations.....	22
2.1.1 Adversarial Machine Learning for LLM Security	25
2.1.2 Automated Red Teaming for AI Systems	25
2.1.3 Security Risks in Retrieval-Augmented Generation (RAG)	26
2.1.4 AI Risk Management Framework for LLM and RAG Security	26
2.1.5 Regulatory and Industry Standards for LLM Security Evaluation	29
2.1.6 Research Success Indicators.....	33

2.2 Review of Previous Studies.....	35
2.2.1 LLM Threat Landscape and Security Taxonomies	36
2.2.2 Automated Jailbreak and Adversarial Prompt Generation.....	38
2.2.3 Prompt Injection, Leakage and RAG Vulnerabilities	41
2.2.4 Agentic and Workflow level Threats	43
2.2.5 Synthesis of Prior Studies	45
2.2.6 Baseline Architecture Comparability	47
2.3 Research Gap and Novelty	49
2.3.1 Identified Research Gaps	50
2.3.2 Proposed Research Contribution and Novelty	51
2.4 Method Selection and Rationale	52
2.4.1 Selection Criteria.....	52
2.4.2 Candidate Method Comparison.....	53
2.4.3 Exclusion Rationale.....	54
2.4.4 Local Serving over Hosted APIs	55
2.4.5 Target LLM Selection	58
2.5 Theoretical Grounding of Evaluation Metrics	61
2.5.1 ROUGE-L Similarity Metric.....	64
2.5.2 Cohen’s Kappa Inter-Rater Reliability.....	65
CHAPTER 3 - RESEARCH METHODS	68
3.1 Research Design.....	68
3.2 Data Collection Methods.....	71
3.2.1 Experimental Protocol and Sample Size Specification	73
3.3 Data Analysis Techniques	74
3.3.1 Research Success Criteria and Measurement Indicators.....	76
3.3.2 Quantitative Thresholds for Research Success	81
3.4 Required Materials and Equipment.....	83
3.4.1 Target LLM and RAG Selection	84
3.4.2 Security-Aware RAG Experimental Architecture.....	86
3.4.3 Explicit Architecture Specification of Experimental Targets	94
3.5 System Architecture	96

3.5.1 Autonomous Multi-Strategy Red-Team Framework	98
3.5.2 RAG Corpus Chunking and Knowledge-Base Construction Process	100
3.5.3 Adversarial Prompt Design and Attack Generation Process	103
3.6 Proof-of-Concept Execution Process	106
CHAPTER 4 – RESULTS AND DISCUSSIONS	109
4.1 Overview of Experimental Results	109
4.1.1 Experimental Execution Timeline	110
4.1.2 Data Integrity Validation and Correction	111
4.2 Attack Success Rate and Hypothesis H1	113
4.2.1 Per-Strategy Baseline Attack Success Rate	113
4.2.2 Category-Level Breakdown	114
4.2.3 Multi-Strategy Pooled Comparison and Formal Hypothesis Test	114
4.2.4 The Strategy-Pooling Dilution Effect	115
4.3 Cross-Model Transferability and Hypothesis H2	116
4.3.1 Transferability Testing Protocol	116
4.3.2 Bidirectional Transfer Results	117
4.3.3 Low Transfer Rate	117
4.4 RAG Attack Surface and Hypothesis H3	118
4.4.1 Standalone Baseline versus RAG-Poisoned Comparison	118
4.4.2 Per-Scenario Results	119
4.4.3 RAG Poisoning Resistance	120
4.5 Exploit Stability and Hypothesis H4	122
4.5.1 Testing Protocol and Results	122
4.5.2 Stability Variation by Strategy	124
4.6 Computational Feasibility and Hypothesis H5	125
4.6.1 Latency and Resource Utilisation Results	125
4.6.2 Latency Variation and Feasibility Implications	126
4.7 Consolidated Hypothesis Evaluation Summary	127
4.8 Multi-Turn Escalation Attack Results	130
4.8.1 Multi-Turn Attack Modes and Testing Protocol	130
4.8.2 Results by Attack Mode	130

4.9 Qualitative Response Analysis.....	132
4.9.1 Correctly-Classified Refusal	132
4.9.2 Successful Jailbreak	133
4.9.3 Successful System Prompt Extraction.....	133
4.9.4 Direct Compliance.....	134
4.10 Literature Baseline Comparison.....	134
4.10.1 Mistral-7B Baseline Comparison	135
4.10.2 Llama-3 Family Baseline Comparison.....	135
4.10.3 Cross-Model Transferability Comparison.....	136
4.10.4 Comparison with Human Red-Teamer Performance	136
4.11 Limitations of the Analysis	138
4.12 Additional Findings.....	140
4.12.1 Overall Classification Category Distribution.....	140
4.12.2 Fuzzer Mutation Operator Effectiveness.....	141
4.12.3 AutoDAN Attack-Angle Effectiveness and Evolutionary Progression	142
4.12.4 Inter-Rater Reliability Validation.....	143
4.13 Contributions, Advantages and Disadvantages.....	145
CHAPTER 5 – CONCLUSIONS AND RECOMMENDATIONS	149
5.1 Conclusions	149
5.2 Recommendations	150
5.3 Suggestion for Future Research	150
GLOSSARY	152
REFERENCES.....	154
APPENDIXES	157
CURRICULUM VITAE	160