Black Hat Europe 2025 Arsenal

Des outils récents de sécurité IA transformant la Cybersécurité

Franck Jeannot

Montréal, Canada, AB841, 13 Décembre 2025

Plan de la Présentation

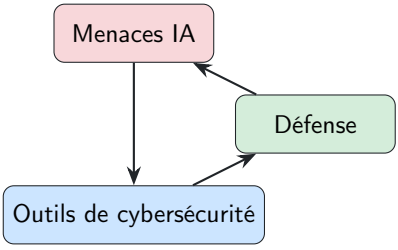
Contexte : l'ère de la sécurité IA

Évolution du paysage des menaces IA :

- Adoption massive des LLM en production [1]
- Nouvelles surfaces d'attaque [2]
- Intégration IA dans infrastructures critiques
- Besoin urgent d'outils spécialisés

Black Hat Europe 2025 Arsenal :

- 8 outils open source présentés
- Focus offensif et défensif
- Couvre tout le cycle de vie IA/ML



Vue d'Ensemble des 8 Outils

Outil	Développeur	Focus	Licence
Harbinger	Mandiant/Google	Red teaming automatisé	Apache-2.0
Red AI Range	Erdem Özgen	Cyber range IA	MIT
AI-Infra-Guard	Tencent	Scan infrastructure	MIT
SPIKEE	Reversec Labs	Injection de prompts	Apache-2.0
MIPSEval	CTU Prague	Évaluation multi-tours	GPL-2.0
SQL Data Guard	Thales	Protection SQL/LLM	MIT
PatchDiff-AI	Akamai	Analyse vulnérabilités	Propriétaire
OS Security LLM	Communauté	Threat modeling	Divers

Red Teaming **Évaluation** **Défense**

Basé sur l'article d'Ethan Salan (Medium, Octobre 2025)

OWASP Top 10 for LLM Applications 2025 [4]

Vulnérabilités Critiques :

- LLM01 Injection de Prompts
- LLM02 Divulgation Sensible
- LLM03 Supply Chain
- LLM04 Empoisonnement
- LLM05 Gestion des Sorties

Vulnérabilités Modérées :

- LLM06 Autonomie Excessive
- LLM07 Fuite Prompt Système
- LLM08 Vecteurs/Embeddings
- LLM09 Désinformation
- LLM10 Consommation Non Bornée

Focus Principal

LLM01 - Injection de Prompts : Vecteur d'attaque le plus critique, permettant la manipulation du comportement LLM via entrées malveillantes directes ou indirectes [3].

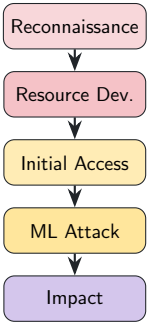
MITRE ATLAS : Framework de Menaces IA [5]

Extension ATT&CK pour IA/ML :

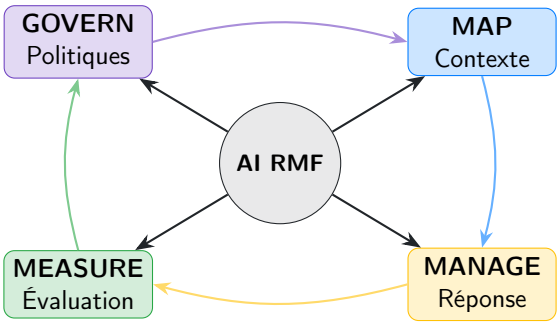
- 14 tactiques spécifiques
- 80+ techniques documentées
- Études de cas réels

Techniques clés pour LLM :

- `AML.T0051.000` – Injection directe
- `AML.T0051.001` – Injection indirecte
- `AML.T0054` – Jailbreak LLM
- `AML.T0043` – Craft Adversarial Data



NIST AI Risk Management Framework [7]



Alignement avec ISO/IEC 42001:2023

Premier standard international pour les systèmes de management IA [6].

Harbinger : Vue d'Ensemble [8]

Paradigme : Du « utiliser des outils IA » vers « être piloté par une plateforme IA »

Développeurs :

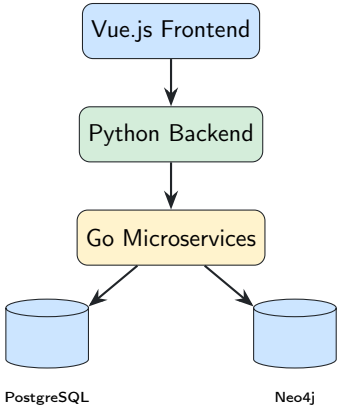
- Matthijs Gielen (Mandiant)
- Idan Ron (Google Cloud Security)

Caractéristiques :

- Licence Apache-2.0
- 138+ étoiles GitHub
- Stack : Python (65%), Vue.js (27%), Go (3%)

Intégration C2 :

- Mythic C2 natif
- API Cobalt Strike, Sliver



Harbinger : Fonctionnalités et CLI

Automatisation opérationnelle :

- Exécution via SOCKS proxies
- Playbooks Jinja2 avec variables Neo4j
- Parsing intelligent LSASS dumps
- Corrélation multi-sources credentials

Support décisionnel IA :

- Recommandations chemins d'attaque
- Analyse paysage opérationnel
- Intégration Gemini pour raisonnement

```
# Demarrer session
hbr record start \
    --session-name "engagement-001"

# Importer dump LSASS
hbr import --type lsass \
    --file dump.dmp

# Exécuter playbook
hbr playbook run \
    --template lateral-movement.yml \
    --target 10.0.0.5

# Analyse chemins AD
hbr graph analyze \
    --source user@domain \
    --target dc01.domain.local
```

Red AI Range (RAR) [9]

Développeur : Erdem Özgen (ASML, Pays-Bas)

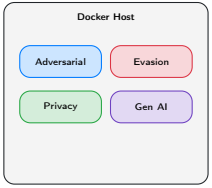
Concept : Cyber range avec systèmes IA/ML vulnérables préconfigurés pour entraînement red team.

Architecture :

- Docker-in-Docker pour isolation
- Scénarios modulaires
- Curriculum pédagogique 5 modules

Déploiement :

- `docker compose up -d`
- Accès : `http://localhost:5002`



Red AI Range : Modules d'Attaque

Module	Contenu
Adversarial Playground	Attaques par exemples adversariaux
Evasion Attacks	FGSM, BIM, JSMA, C&W, PGD, patches adversariaux (2015) [10]
Model Tampering	Injection de trojans, payloads neuronaux
Privacy Attacks	Inversion, inférence d'appartenance, vol de modèle
Gen AI Target	Injection de prompts, jailbreaking LLM, empoisonnement RAG

Techniques d'Évasion Implémentées

FGSM (Fast Gradient Sign Method), **BIM** (Basic Iterative Method), **JSMA** (Jacobian Saliency Map Attack), **C&W** (Carlini-Wagner), **PGD** (Projected Gradient Descent)

AI-Infra-Guard (A.I.G) [11]

Développeur : Tencent Zhuque Lab

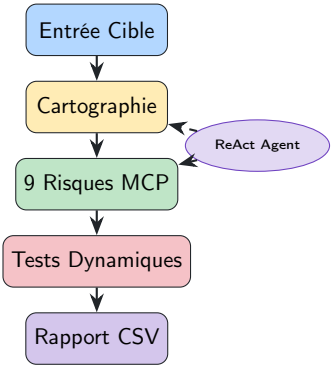
Statistiques :

- 2,500+ étoiles GitHub
- Licence MIT
- Framework ReAct intégré

Trois piliers :

1. Scan infrastructure IA (30+ frameworks)
2. Analyse sécurité MCP Server
3. Évaluation Jailbreak LLM

CVE détectées : 400+ incluant CVE-2025-55182



AI-Infra-Guard : 9 Catégories de Risques MCP [12]

Risque	Description
Tool Poisoning	Instructions cachées manipulant l'Agent IA
Rug Pull Attack	Comportement normal puis exécution malveillante
Tool Shadowing	Redéfinition du comportement d'outils légitimes
Indirect Prompt Inj.	Instructions malveillantes depuis sources externes
Command Injection	Exécution de code sans sandboxing
Path Traversal	Accès arbitraire aux fichiers
SSRF	Server-Side Request Forgery
Credential Leakage	Exposition de tokens sensibles
Data Exfiltration	Extraction non autorisée de données

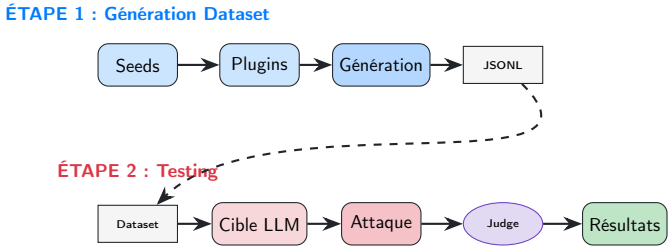
AI-Infra-Guard : Utilisation CLI

```
# Scan infrastructure IA
./ai-infra-guard scan --target 192.168.1.0/24 \
    --frameworks ollama,comfyui,vllm,triton

# Analyse sécurité MCP Server
./ai-infra-guard mcp \
    --code /path/to/mcp/server \
    --model gpt-4 \
    --token sk-xxx \
    --csv results.csv

# Evaluation Jailbreak
./ai-infra-guard jailbreak \
    --target-api https://api.example.com/v1/chat \
    --dataset curated-jailbreaks.jsonl \
    --output jailbreak-results.json
```

Frameworks IA Détectés (30+)

Ollama, ComfyUI, vLLM, NVIDIA Triton, Ray Serve, BentoML, Hugging Face TGI, LangServe, FastChat, LocalAI...

SPIKEE : Architecture Pipeline [13]



Développeur

Donato Capitella (Reversec Labs) — Site web : https://spikee.ai

SPIKEE : Seeds et Plugins d'Évasion

Types de Seeds Disponibles :

- `seeds-cybersec-2025-04` : Attaques générales (XSS, exfil.)
- `seeds-sysmsg-extraction` : Extraction prompts système
- `seeds-wildguardmix-harmful` : Contenu dangereux
- `seeds-simsonsun-jailbreaks` : Jailbreaks haute qualité

Plugins d'Évasion (15+) :

- `1337` : Leetspeak
- `base64`, `hex`, `morse` : Encodages
- `ascii_smuggler` : Tags Unicode invisibles
- `anti_spotlighting` : Contournement délimiteurs
- `best_of_n` : Perturbations aléatoires [14]

SPIKEE : Workflow Complet

```
# Initialisation et listing des seeds
spikee init && spikee list seeds

# Génération dataset avec plugins
spikee generate \
    --seed-folder datasets/seeds-cybersec-2025-04 \
    --plugin best_of_n \
    --plugin-options "best_of_n:variants=50" \
    --output custom-dataset.jsonl

# Test avec attaque itérative
spikee test \
    --dataset datasets/*.jsonl \
    --target openai_api \
    --target-options gpt-4o-mini \
    --attack best_of_n \
    --attack-iterations 20

# Analyse des résultats
spikee results analyze \
    --result-file results/*.jsonl \
    --format markdown > report.md
```

MIPSEval : Attaques Conversationnelles Multi-Tours [16]

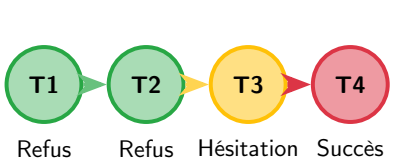
Développeurs :

- Muris Sladić (CTU Prague)
- Sebastián García (Stratosphere IPS)

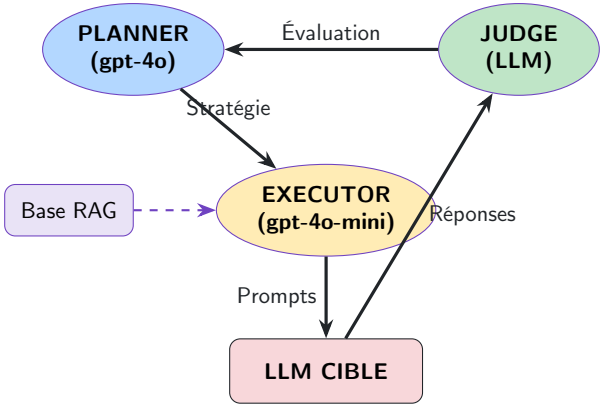
Problème adressé :
Les vulnérabilités ne se manifestent parfois que lors de **conversations multi-tours soutenues**.

Principe Crescendo [15] :
Un LLM peut refuser une requête inappropriée au tour 1, mais après plusieurs tours d'amorçage, ses guardrails peuvent être contournés.

Taux de succès : 29-61% supérieur aux méthodes single-turn



MIPSEval : Architecture Multi-Agents Planner-Executor-Judge



Sorties Générées

Historique complet conversations (JSONL), stratégies tentées et **stratégies victorieuses** pour replay

MIPSEval : Taxonomie des Prompts et Modes Opératoires

Taxonomie à 3 niveaux :

Type	Objectif
Benign	Établir confiance
Probing	Tester limites
Malicious	Charge utile

Flux conversationnel :

- Tours 1-2 : Dialogue bénin
- Tour 3 : Sondage défenses
- Tours 4+ : Exécution attaque

Modes opératoires :

- **Explore** : Maximise diversité des stratégies, découverte de vulnérabilités
- **Exploit** : Concentre sur stratégies validées, optimise taux de succès

Datasets intégrés :

- HarmBench (240 comportements)
- AdvBench
- JailbreakBench
- MHJ Dataset (Scale AI)

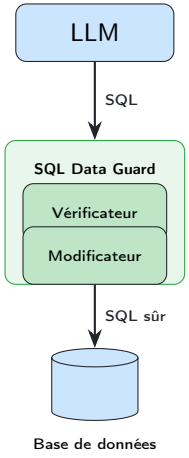
SQL Data Guard [17]

Développeur : Thales Group

Problème critique :
Les requêtes SQL générées par LLM ne peuvent pas utiliser de **prepared statements** (structure dynamique), augmentant le risque d'injection SQL.

Fonctionnalités :

- Validation contre configuration
- Réécriture automatique
- Détection expressions malveillantes
- Conformité RGPD/CCPA



SQL Data Guard : API Python

```python
from sql_data_guard import verify_sql, fix_sql

# Configuration des restrictions
config = {
    "tables": [{
        "name": "orders",
        "columns": ["id", "product_name", "account_id"],
        "restrictions": [{"column": "account_id", "value": 123}]
    }],
    "blocked_patterns": ["1=1", "OR 1", "UNION SELECT"]
}

# Requête générée par LLM (potentiellement dangereuse)
llm_query = "SELECT * FROM orders WHERE 1 = 1"

# Vérification et correction automatique
result = verify_sql(llm_query, config)
# {"allowed": false, "fixed": "SELECT id, product_name,
#   account_id FROM orders WHERE account_id = 123"}
```

PatchDiff-AI (Patch Wednesday) [18]

Développeur : Maor Dahan, Akamai SIG

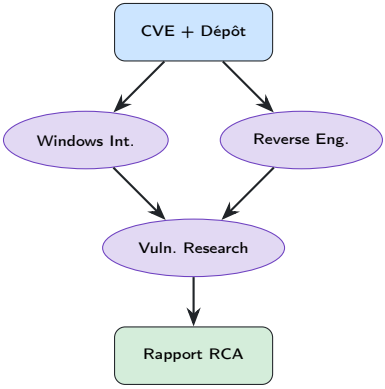
Concept : Analyse de cause racine (RCA) automatisée pour CVE via agents IA spécialisés.

LLM utilisés :

- OpenAI o4-mini (enrichissement)
- OpenAI o3 (analyse approfondie)

Métriques de performance :

Métrique	Taux
Exécutable correct	88.6%
Fonction vulnérable	83.9%
Cause racine	71.4%
Coût moyen/rapport	$0.14



OpenSource Security LLM [20]

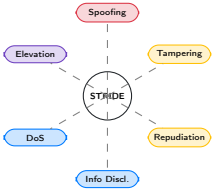
Concept : Utiliser des LLM légers déployables localement pour la modélisation automatisée des menaces.

Implémentation de référence — STRIDE-GPT :

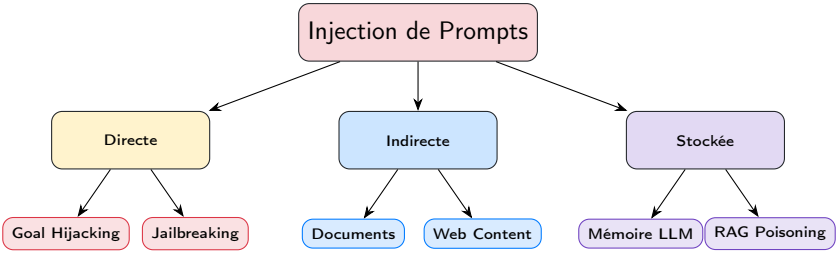
- 864+ étoiles GitHub
- Support multi-modal (diagrammes)
- Génération arbres d'attaque (Mermaid)
- Scoring DREAD
- Support Ollama/LM Studio

Méthodologie STRIDE [19] :

- **S**poofing, **T**ampering
- **R**epudiation, **I**nfo Disclosure



Classification des Injections de Prompts [21]



Type	Description
Directe	Commandes malveillantes dans l'entrée utilisateur
Indirecte	Instructions cachées dans contenu externe [2]
Stockée	Prompts nuisibles intégrés en mémoire persistante LLM

Techniques d'Évasion Avancées

Encodage :

- Base64, Hex
- ROT13, Morse
- Unicode confusables

Structure :

- Payload splitting
- Anti-spotlighting
- Delimiter injection

Sémantique :

- Role-play (DAN)
- Virtualisation
- Hypothétiques

Techniques Avancées

- **GCG** (Greedy Coordinate Gradient) [22]
- **PAIR** : Jailbreaking en 20 requêtes [23]
- **Many-shot** : Exploitation fenêtres contextuelles [24]
- **ASCII Art** : Attaques visuelles

Tendances Identifiées à Black Hat Europe 2025

Red Teaming IA

Plateformes automatisées (Harbinger [8], RAR [9]), Cyber ranges dédiés à la simulation adversariale

Sécurité LLM

Évaluation mature (SPIKEE [13], MIPSEval [16]), Protection données (SQL Data Guard [17])

Défense par IA

Analyse vulnérabilités (PatchDiff-AI [18]), Threat modeling (STRIDE-GPT [20])

Évolution du paysage

Convergence vers des **écosystèmes intégrés** combinant sécurité offensive et défensive, avec l'IA au cœur des deux approches.

Conclusion et Recommandations

Points clés de Black Hat Europe 2025 Arsenal :

1. **Maturité croissante** des outils de sécurité IA open source
2. **Standardisation** via OWASP, MITRE ATLAS, NIST AI RMF
3. **Focus multi-tours** : Les attaques simples ne suffisent plus
4. **Protection des données** : Criticité SQL/LLM reconnue
5. **Automatisation** : IA défensive devient mainstream

Recommandations pour Praticiens

- Intégrer les tests d'injection de prompts dans CI/CD
- Adopter une approche « défense en profondeur » pour les pipelines LLM
- Utiliser les cyber ranges (RAR) pour formation continue
- Surveiller les publications OWASP/MITRE pour nouvelles menaces

Ressources et Liens

Dépôts GitHub des Outils :

- Harbinger : https://github.com/mandiant/harbinger
- Red AI Range : https://github.com/ErdemOzgen/RedAiRange
- AI-Infra-Guard : https://github.com/Tencent/AI-Infra-Guard
- SPIKEE : https://github.com/ReversecLabs/spikee
- MIPSEval : https://github.com/stratosphereips/MIPSEval
- SQL Data Guard : https://github.com/ThalesGroup/sql-data-guard
- STRIDE-GPT : https://github.com/mrwadams/stride-gpt

Standards et Frameworks :

- OWASP Top 10 LLM : https://genai.owasp.org/llm-top-10/
- MITRE ATLAS : https://atlas.mitre.org/
- NIST AI RMF : https://www.nist.gov/itl/ai-risk-management-framework

Références I

[1] OpenAI (2024). GPT-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024. Méthodologie red teaming avec 100+ experts externes.

[2] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *arXiv preprint arXiv:2302.12173*, 2023. URL https://arxiv.org/abs/2302.12173. Article fondateur de Février 2023 démontrant des attaques sur Bing Chat (GPT-4) et GitHub Copilot.

[3] Liu, Yi and Deng, Gelei and Xu, Zhengzi and Li, Yuekang and Zheng, Yaowen and Zhang, Ying and Zhao, Lida and Zhang, Tianwei and Liu, Yang. Prompt Injection Attack against LLM-Integrated Applications. *arXiv preprint arXiv:2306.05499*, 2023. URL https://arxiv.org/abs/2306.05499. Technique HouYi, 31/36 applications vulnérables testées.

[4] OWASP Foundation. OWASP top 10 for LLM applications 2025. https://genai.owasp.org/llm-top-10/, 2025. Standard de référence pour vulnérabilités LLM.

[5] MITRE Corporation. MITRE ATLAS: Adversarial threat landscape for AI systems. https://atlas.mitre.org/, 2024. Extension ATT&CK pour menaces IA/ML, 80+ techniques.

Références II

[6] International Organization for Standardization. ISO/IEC 42001:2023 – AI management system. https://www.iso.org/standard/81230.html, 2023. Premier standard international système de management IA.

[7] National Institute of Standards and Technology. AI risk management framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework, 2023. Framework de gestion des risques IA en 4 fonctions.

[8] Matthijs Gielen and Idan Ron. Harbinger: AI-driven red teaming platform. https://github.com/mandiant/harbinger, 2025. Licence Apache-2.0. Présenté à Black Hat Europe 2025 Arsenal.

[9] Erdem Özgen. Red AI Range: AI red teaming cyber range. https://github.com/ErdemOzgen/RedAiRange, 2025. Licence MIT. Architecture Docker-in-Docker.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2015. https://arxiv.org/abs/1412.6572.

[11] Tencent Zhuque Lab. AI-Infra-Guard: Full-stack AI security assessment platform. https://github.com/Tencent/AI-Infra-Guard, 2025. Licence MIT. 2500+ étoiles GitHub.

Références III

[12] Tencent Zhuque Lab. Systematic analysis of MCP server security. https://arxiv.org/abs/2508.12538, 2025. Analyse des 9 catégories de risques MCP.

[13] Donato Capitella. SPIKEE: Simple prompt injection kit for evaluation and exploitation. https://github.com/ReversecLabs/spikee, 2025. Licence Apache-2.0. Site web: https://spikee.ai.

[14] Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*, 2024. https://arxiv.org/abs/2404.02151.

[15] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. *arXiv preprint arXiv:2404.01833*, 2024. https://arxiv.org/abs/2404.01833.

[16] Muris Sladić and Sebastián García. MIPSEval: Multi-turn injection prompt security evaluation. https://github.com/stratosphereips/MIPSEval, 2025. Licence GPL-2.0. Stratosphere Laboratory, CTU Prague.

Références IV

[17] Thales Group. SQL Data Guard: Security middleware for LLM-generated SQL. https://github.com/ThalesGroup/sql-data-guard, 2025. Licence MIT. Conformité RGPD/CCPA.

[18] Maor Dahan. Patch wednesday: Root cause analysis with LLMs. https://www.akamai.com/blog/security-research/2025/dec/patch-wednesday-root-cause-analysis-with-llms, 2025. Akamai SIG - 88.6% précision identification.

[19] Microsoft. STRIDE threat modeling. https://docs.microsoft.com/en-us/azure/security/develop/threat-modeling-tool-threats, 2024. Méthodologie Spoofing, Tampering, Repudiation, Info Disclosure, DoS, Elevation.

[20] Matthew Adams. STRIDE-GPT: AI-powered threat modeling tool. https://github.com/mrwadams/stride-gpt, 2025. 864+ étoiles GitHub. Support Ollama/LM Studio.

[21] Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Liang, Xiaofei Liu, Jiawei Li, Jiaqi Xu, Yongqiang Wu, Qi Liu, et al. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*, 2024. https://arxiv.org/abs/2407.04295.

Références V

[22] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. Algorithme GCG (Greedy Coordinate Gradient), 88% taux de succès.

[23] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. PAIR: Jailbreaking black box LLMs in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023. Jailbreaking automatisé black-box via LLM attaquant.

[24] Anthropic. Many-shot jailbreaking. https://www.anthropic.com/research/many-shot-jailbreaking, 2024. Exploitation des fenêtres contextuelles longues.