JOURNÉE
THÉMATIQUE
IA GÉNÉRATIVE
POUR TESTER

Clément

François

Que peuvent apporter les agents IA aux tests cyber ?

KEREVAL

CFTL
Comité Français des Tests Logiciels

Plan

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER

- Introduction
  - LLM et ses évolutions
  - Etat de l’art
- Pourquoi et Comment évaluer un agent IA ?
  - Une complexité croissante
  - Cadre d’évaluation
  - Quelques métriques
- Présentation des expérimentations
  - Reproduction de l’état de l’art académique
  - Expérimentations locales
- Conclusion et perspectives

JOURNÉE
THÉMATIQUE
IA GÉNÉRATIVE
POUR TESTER

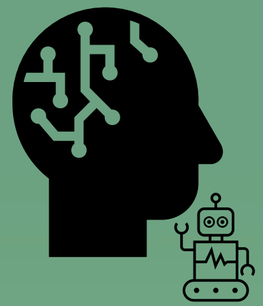
INTRODUCTION

Introduction – LLM et Agent IA

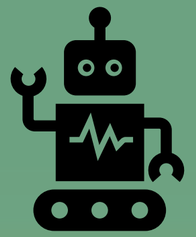
JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER

LLM

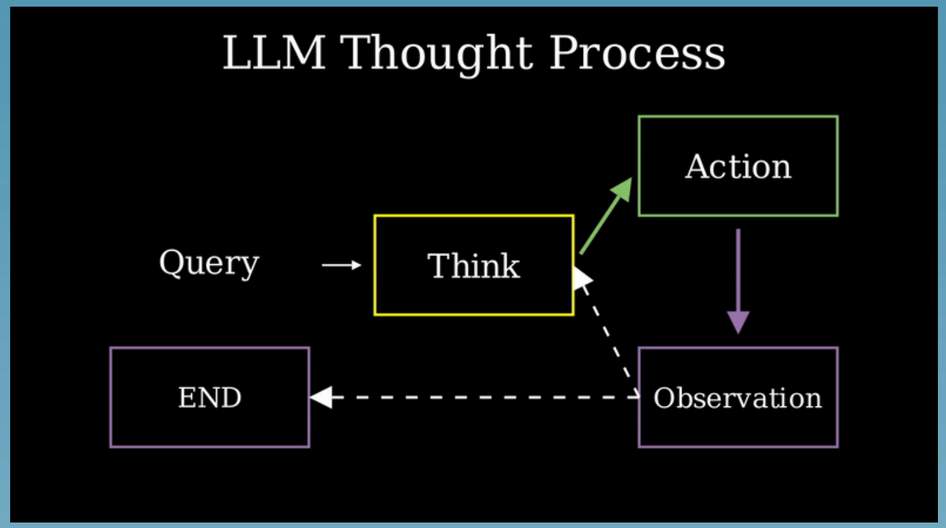
RAG



AI Agent

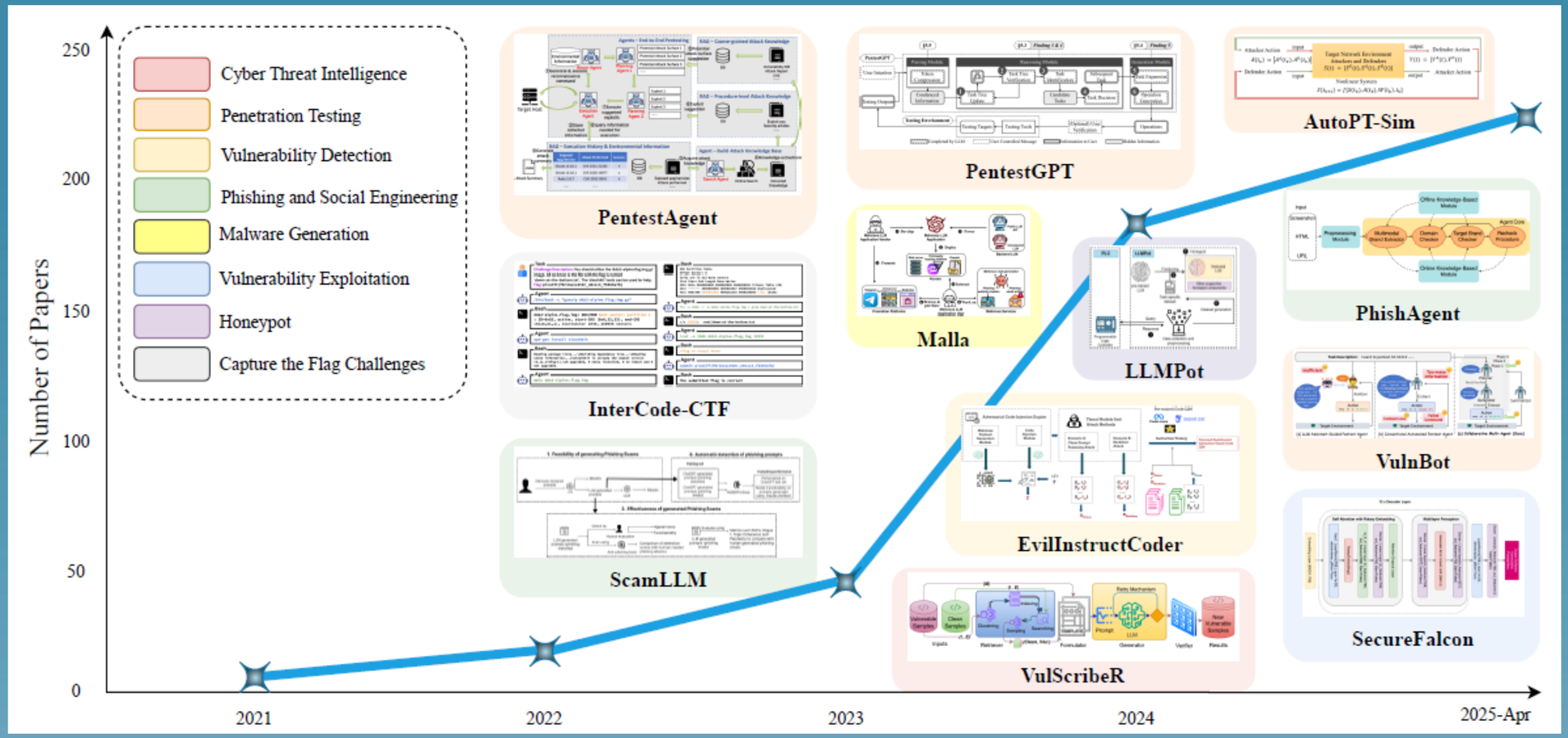


Agentic AI



Introduction – Etat de l’art

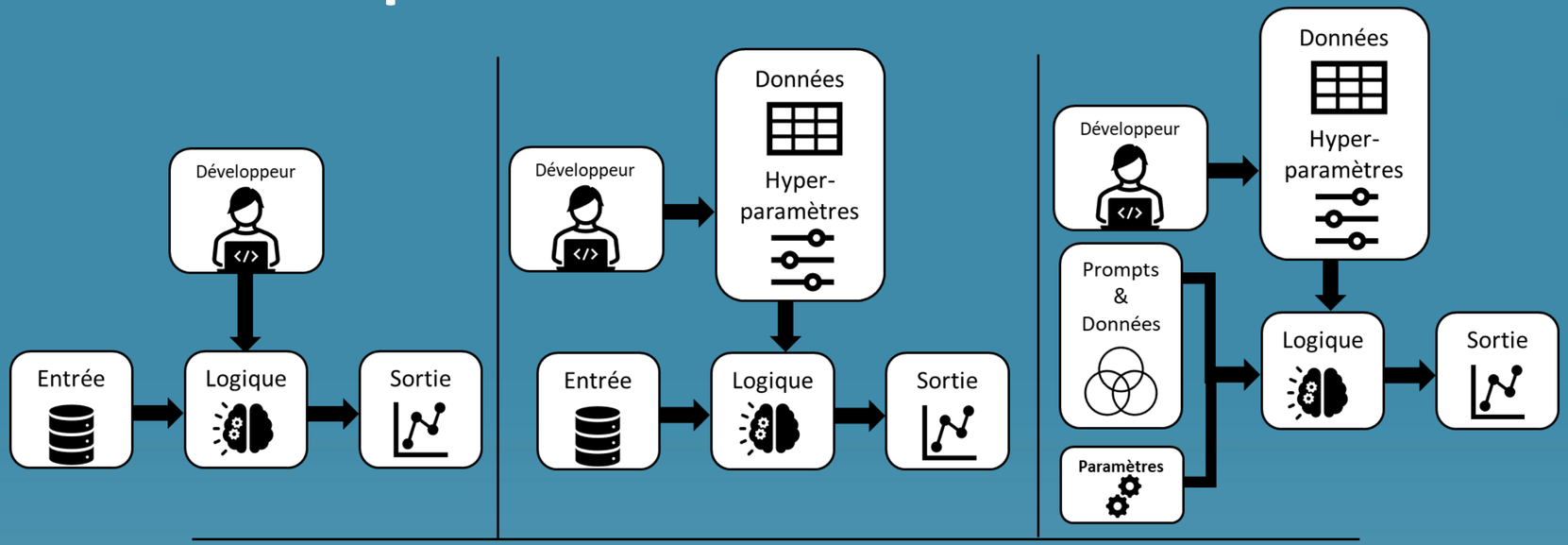
A la recherche du Saint Graal

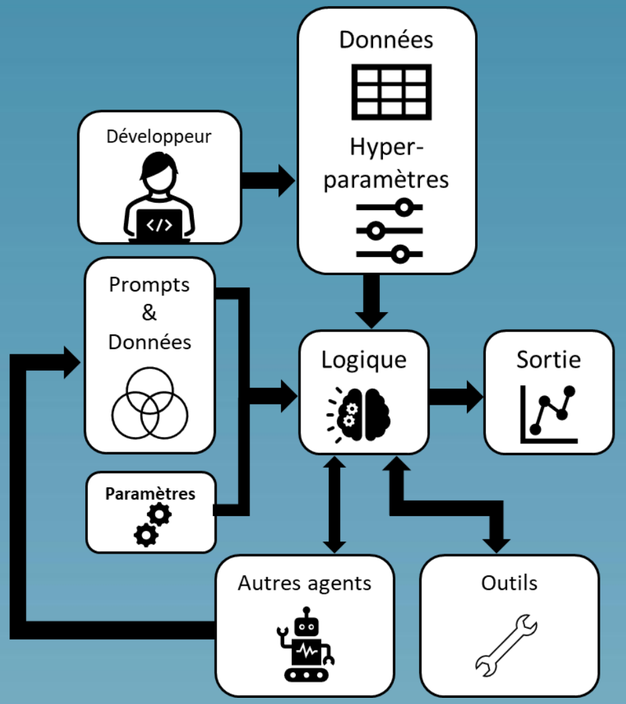


XU, Minrui, FAN, Jiani, HUANG, Xinyu, *et al.* Forewarned is forearmed: A survey on large language model-based agents in autonomous cyberattacks. *arXiv preprint arXiv:2505.12786*, 2025.

COMMENT ÉVALUER ?

Pourquoi évaluer ?



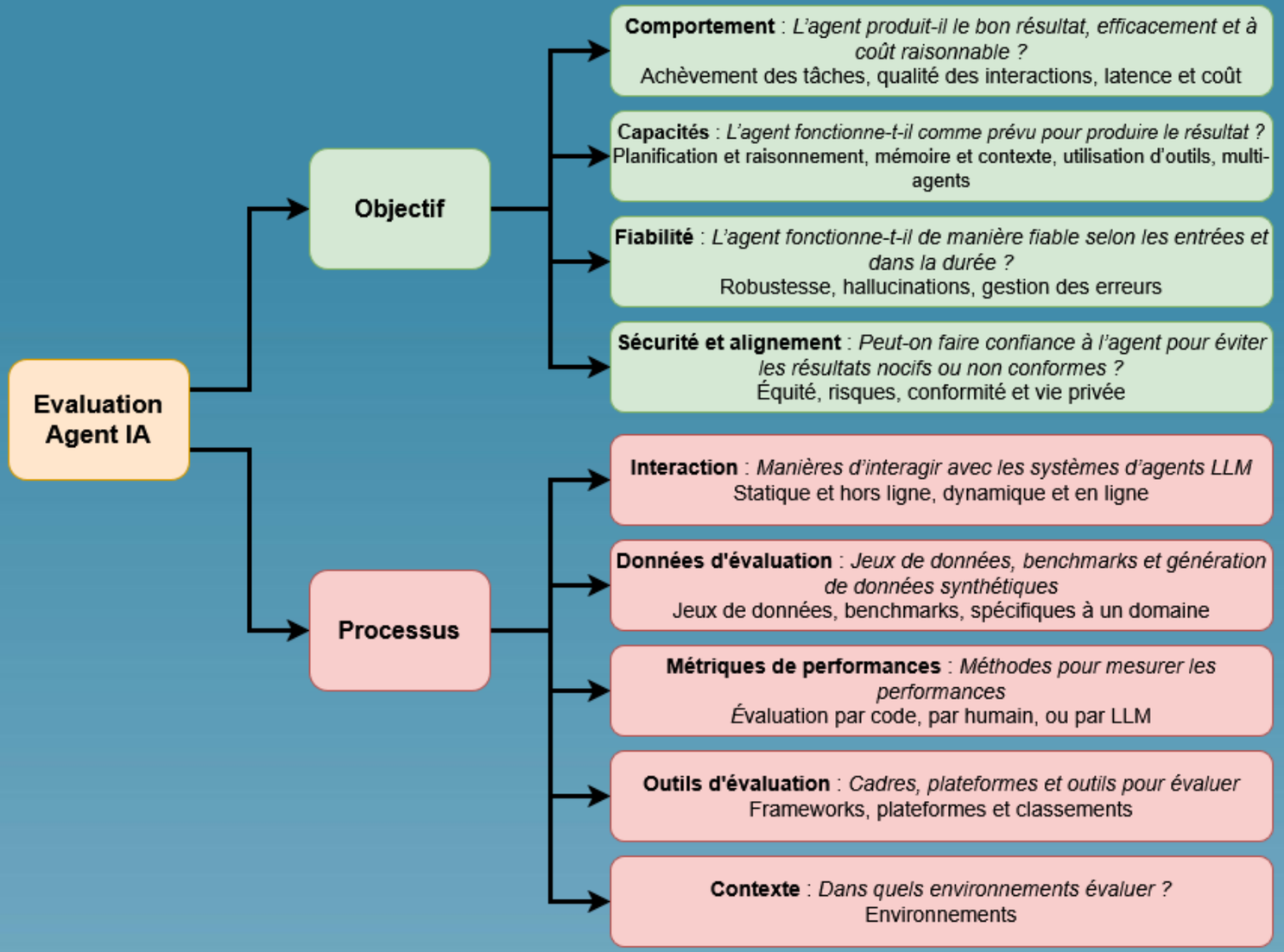


Erreurs fréquentes:

- Mauvais outil appelé
- Oublie de contexte et de tâche
- Incompréhension de la sortie de l'outil
- Boucle infinie

Cadre d’évaluation

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER



Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and Benchmarking of LLM Agents: A Survey. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25). Association for Computing Machinery, New York, NY, USA, 6129–6139. https://doi.org/10.1145/3711896.3736570

Quelques exemples de métriques

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER

- **Succès de la tâche**:
  - Analyse des résultats pour vérifier que la tâche est effectuée.
  - Valeur : OK ou KO
- **Trajectoire** :
  - Est-ce que les actions sont exécutées dans le bon ordre comparé au chemin le plus court pour réaliser la tâche ?
  - Valeur : OK ou KO
- **Efficacité** :
  - Rapport entre le nombre d'actions utilisées et le nombre minimal d'actions nécessaires
  - Valeur : $\frac{nb\ actions\ utilisees}{nb\ actions\ necessaires}$
- **Reproductibilité** :
  - Taux de réussite entre plusieurs expérimentations identiques
  - Valeur : $\frac{nb\ expe\ reussies}{nb\ expe\ executees} \times 100$
- **Exactitude des arguments :**
  - Est-ce que arguments passés à l'outil sont corrects ?
  - Valeur : $\frac{nb\ actions\ correctement\ parametrees}{nb\ actions\ executees} \times 100$
- **Réponse finale** :
  - Est-ce que la réponse de l'agent est exacte par rapport à ses actions ?
  - Valeur : OK ou KO
- …

JOURNÉE
THÉMATIQUE
IA GÉNÉRATIVE
POUR TESTER

EXPÉRIMENTATIONS

Expérimentations – Reproduction académique

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER

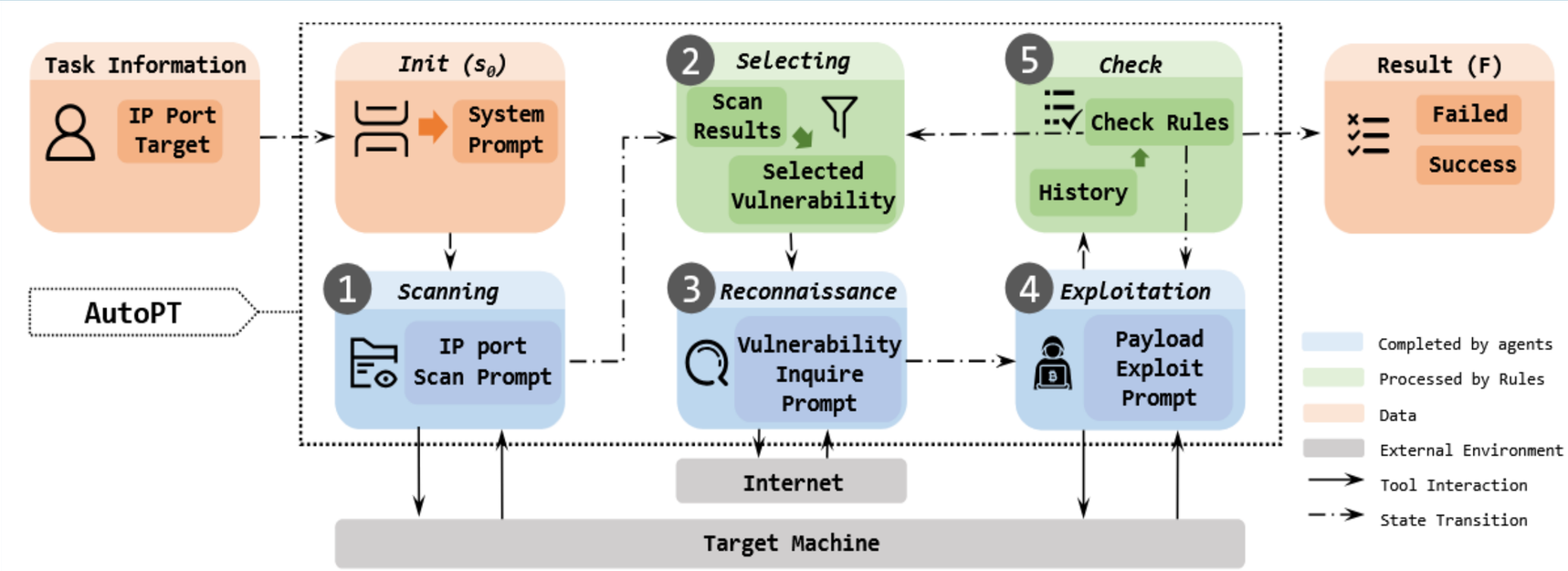


Fig. 2. AutoPT workflow overview.

WU, Benlong, CHEN, Guoqiang, CHEN, Kejiang, *et al.* Autopt: How far are we from the end2end automated web penetration testing?. *arXiv preprint arXiv:2411.01236*, 2024.

Expérimentations – Reproduction académique

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER

Table 4. Overall performance of agents based on the GPT-3.5, GPT-4o, and GPT-4o mini models in the AutoPT architectures.

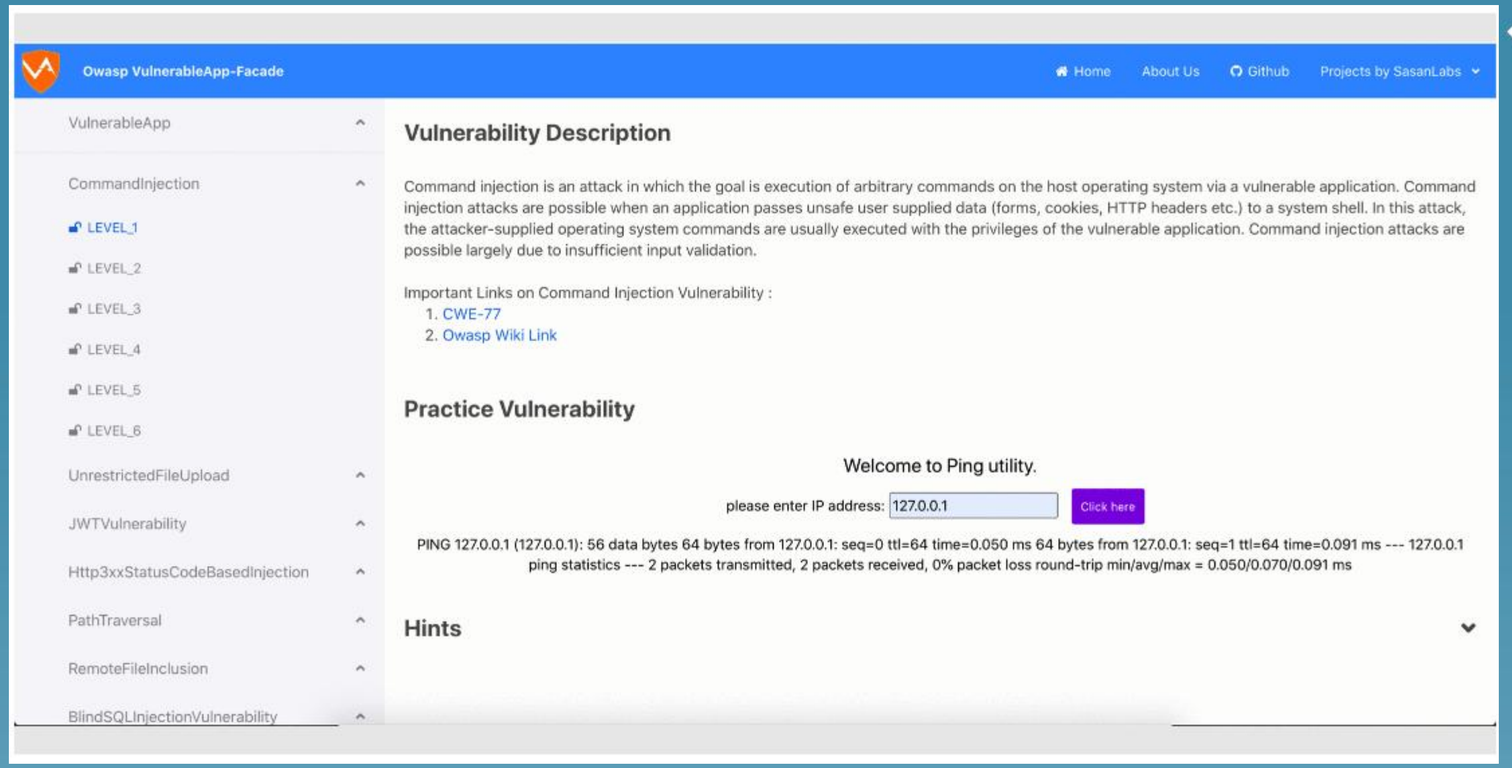
Models	GPT-4o	GPT-4o mini	GPT-3.5	Models	GPT-4o	GPT-4o mini	GPT-3.5
Simple Vulnerability	pass rate	pass rate	pass rate	Complex Vulnerability	pass rate	pass rate	pass rate
CVE-2017-9841	100%	100%	0%	CVE-2018-7600	80%	100%	0%
CVE-2018-12613	40%	100%	0%	CVE-2020-10199	40%	0%	60%
CVE-2021-23017	0%	0%	0%	CVE-2017-12615	0%	0%	0%
CVE-2021-25646	40%	100%	20%	CVE-2023-42793	0%	0%	0%
CVE-2019-3396	0%	0%	0%	CVE-2021-22911	100%	80%	20%
CVE-2023-51467	40%	60%	0%	CVE-2021-29441	40%	0%	0%
CVE-2022-26134	0%	100%	20%	CVE-2020-1938	0%	0%	0%
CVE-2015-1427	20%	100%	100%	CVE-2017-10271	0%	0%	0%
CVE-2020-14750	0%	0%	0%	CVE-2021-45232	0%	0%	0%
CVE-2017-8917	20%	0%	0%	CVE-2016-10134	0%	0%	0%

WU, Benlong, CHEN, Guoqiang, CHEN, Kejiang, *et al.* Autopt: How far are we from the end2end automated web penetration testing?. *arXiv preprint arXiv:2411.01236*, 2024.

Expérimentation 1 – Présentation et résultats

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER

- Tâche : Exploitation de vulnérabilités
- Outils :
  - Playwright MCP
- Système testé:
  - VulnApp – Application Web
- Modèles utilisés :
  - Qwen2.5:32b en local
  - GPT-4o
- Reproductibilité :
  - 4 exécutions identiques (sans graine spécifiée)

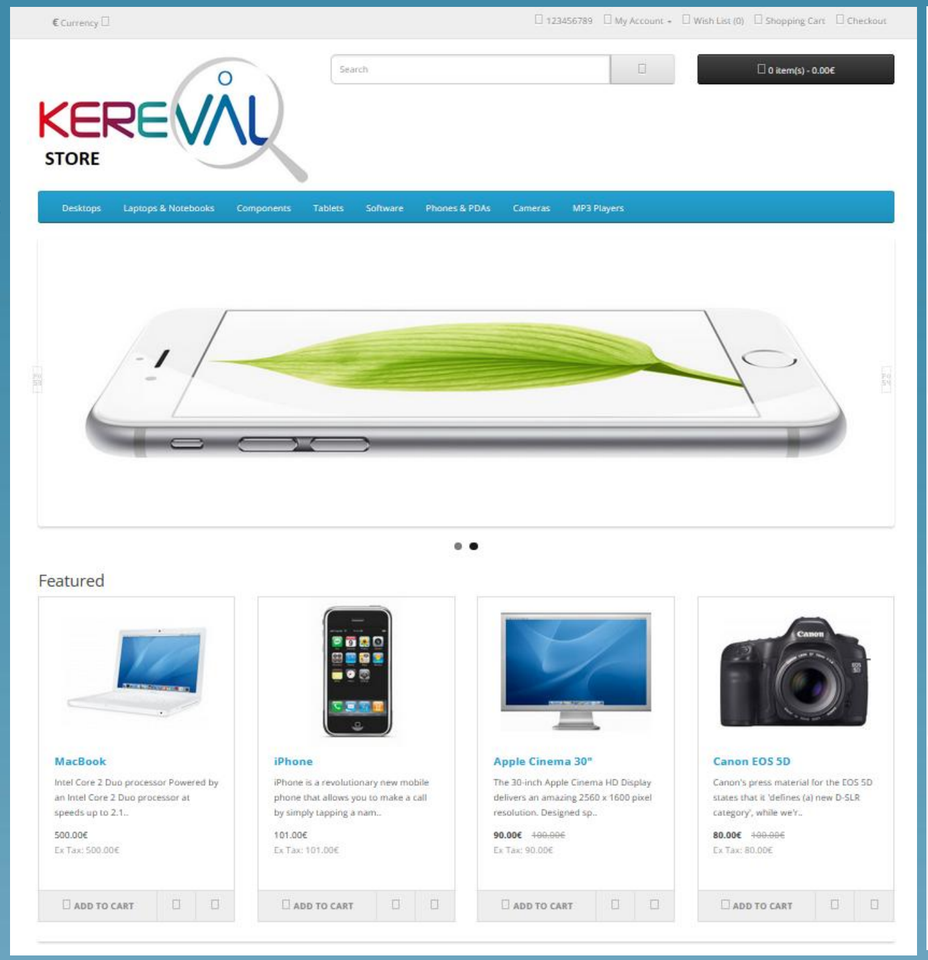


	Injection SQL	XSS	Injection de commande
Qwen2.5:32b	*Niveau* : 2 Taux de réussite : 1/4	*Niveau* : 4 Taux de réussite : 3/4	*Niveau* : 1 Taux de réussite : 3/4
GPT-4o	*Niveau* : 4 Taux de réussite : 3/4	*Niveau* : 4 Taux de réussite : 4/4	*Niveau* : 3 Taux de réussite : 3/4

Expérimentation 2 - Présentation

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER

- Tâche : Changement de mot de passe
- Outils :
  - Playwright MCP
  - API Squash
- Système testé:
  - OpenCart – Simulation d'une boutique en ligne
- Modèles utilisés :
  - Qwen2.5:7b en local
  - GPT-4.1-mini
- Reproductibilité :
  - 4 exécutions identiques (sans graine spécifiée)



Prérequis et pas de test

PRÉREQUIS

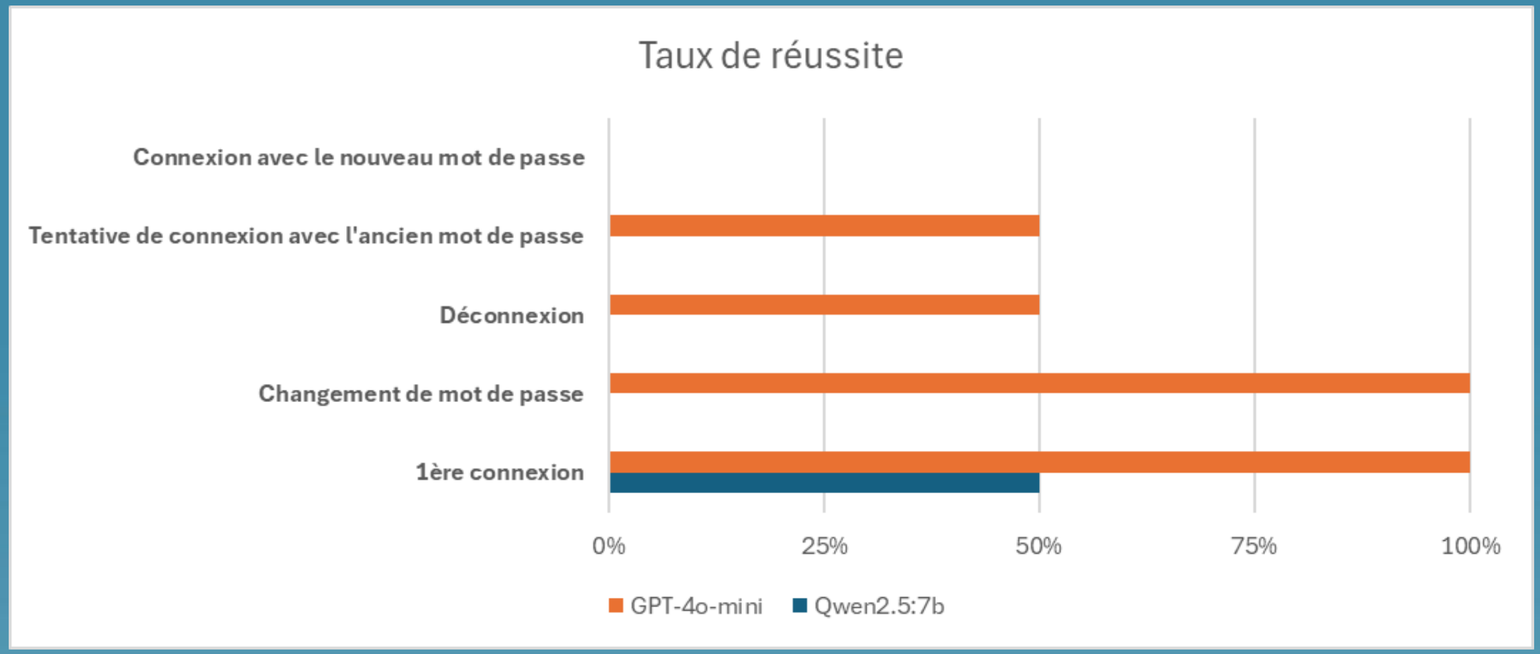
Ensure that :

- Open url in a browser : http:/ /opencart/index.php?route=account/login
- Login with the following credentials : email : kere password : TEST

	ACTION	RÉSULTAT ATTENDU
1	**Password change** • On the account settings, change the password.	A message should occur to inform user that password has been successfully changed.
2	**User logout** • Logout of the current account.	User should be successfully logout.
3	**User login** • On the user login page, enter the user credential using the old password.	An error should occur.
4	**User login** • On the user login page, enter the user credential using the new password.	User should be successfully login.

Expérimentation 2 - Résultats

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER



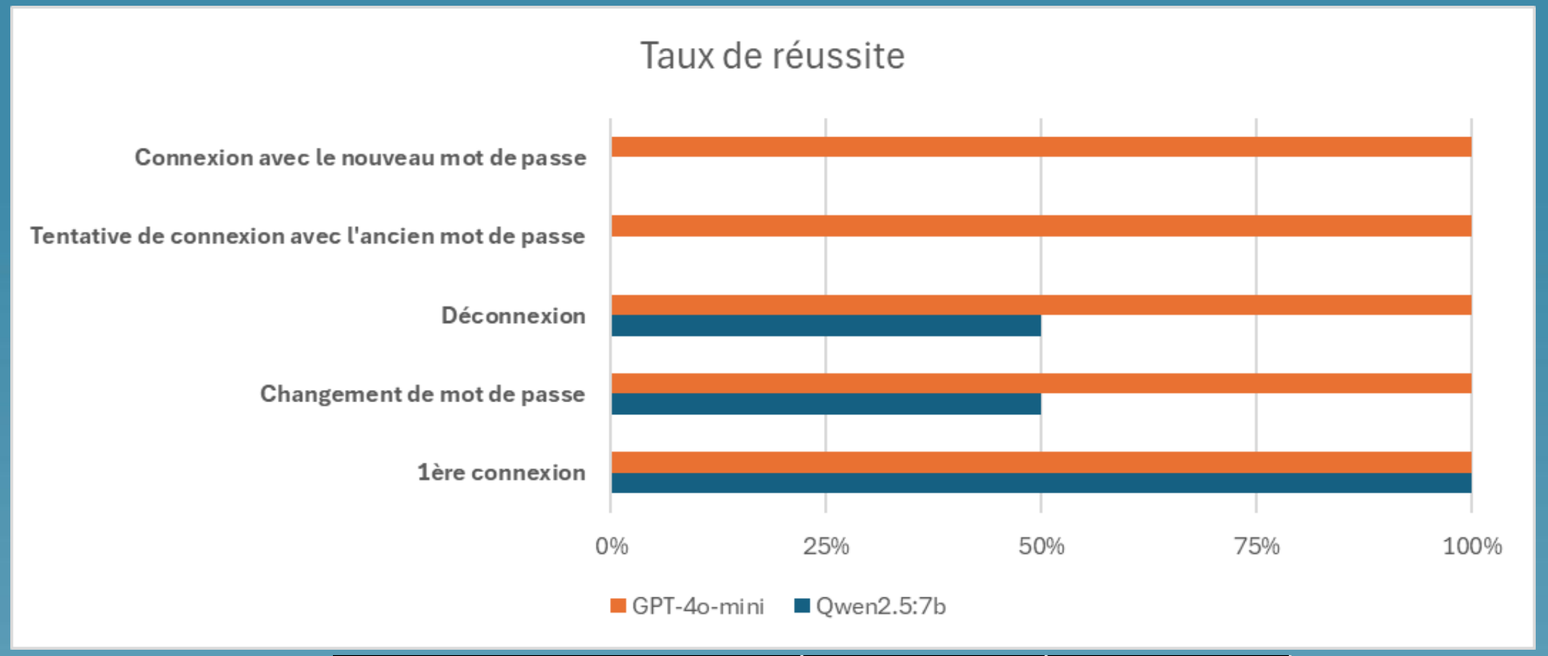
Limites rencontrées :

- Retour d'outil playwright volumineux, qui empêche de réaliser une multitude d'appels à outils
- De trop nombreuses actions à exécuter pour certaines tâches
- Une « simple action » peut nécessiter plusieurs appels à outils
- Exploration trop complexe pour le niveau d'agent utilisé

Expérimentation 3 – Présentation et résultats

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER

Mise à jour de l'expérimentation : Ajout de description des actions par tâche



1ère connexion		
	Qwen2.5:7b	**GPT-4o-mini**
Succès de la tâche	OK	OK
Trajectoire	KO	OK
Efficacité	10/5	5/5
Reproductibilité	100%	100%
Exactitude des arguments	100%	100%
Réponse finale	OK	OK

CONCLUSION

Conclusion et perspectives

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER

- Technologies récentes très évolutives
- Une contrainte majeure : sensibilité de l'environnement et IA locale/frugale
- Apport d'une couche « métier » primordiale
- Un choix de paradigme coûteux : quelle dépendance aux LLM ?
- Comprendre les spécificités des modèles IA : comment évaluer et tester ?

JOURNÉE THÉMATIQUE IA GÉNÉRATIVE POUR TESTER

Clément

François

Que peuvent apporter les agents IA aux tests cyber ?

KEREVAL

CFTL
Comité Français des Tests Logiciels