LAB 1 Report

Gabin MERLOT-DIMET
Lucas GOUBET

ING5 CYBER GR1 FR

Analyse différentielle des rapports

Sur ID1 on a une précision de la CWE après la modification du system policy, ainsi que pour l'ID6 et l'ID9. L'ID8 n'est plus considéré comme une attaque et n'est pas pris en compte comme une menace. On a donc une réduction de faux positifs et une réduction du bruit pour le traitement des menaces réelles.

On voit bien que dans les rationale la politique zero trust est appliquée.

ID1 : LLM04

ID2 : LLM01

ID3 : LLM01

ID4 : LLM01 LLM08

ID5 : LLM01

ID6 : LLM06

ID7 : LLM02

ID8 :

ID9 : LLM06

ID10 : LLM01

I. Assets & Trust Boundaries

- Clé API Gemini : Élément critique stocké dans .env et .env.example
- Instructions Système (System Policy) : Définies dans src/prompts.py, elles constituent la logique de sécurité et les gardes fous.
- Données Utilisateur (PII) : Comme démontré lors du test OmniChat, les emails sont des actifs sensibles.

- Pipeline de Rapport : Flux allant du prompt brut au fichier reports/baseline.json via une validation de schéma Pydantic.
- Limite de Confiance : La frontière principale se situe entre l'entrée utilisateur (non fiable) et le filtre d'entrée (src/filters.py).

II. Adversaries & Entry Points

- Adversaires : Utilisateurs malveillants voulant extraire des données/faire faire des actions malveillantes par le modèle.
- Points d'Entrée : Prompts, URLs externes malveillantes (ID 7), commandes à exécuter.

III. Mapped Risks

A. OWASP LLM Top-10

LLM01: Prompt Injection : Observé via les tentatives d'écrasement système (ID 10) et les tests Gandalf.
LLM02: Insecure Output Handling : Risque lié à l'ID 7 où le modèle suit des instructions provenant d'une URL sans validation.
LLM04: Model Theft : Tentatives d'extraction de la logique interne du modèle (ID1)
LLM06: Sensitive Information Disclosure : Confirmé par l'extraction réussie de l'email sur OmniChat et l'ID 6.
LLM07: Insecure Plugin Design : Observé via l'ID 3 traitant de l'exécution de texte libre comme code.
LLM08: Excessive Agency : Risque que le modèle exécute des commandes destructrices (ID 4 : rm -rf /).
LLM09: Overreliance : Risque de faire confiance aveugle au résumé des politiques internes sans vérification humaine (ID 1).

B. MITRE ATLAS

Execution (AML.T0016) : Tentative de déclencher des commandes système via le LLM (ID 4).
Discovery (AML.T0012) : Utilisation de prompts pour comprendre les limites du modèle et les filtres en place (ID 8).
Exfiltration (AML.T0002) : Succès sur OmniChat pour extraire les PII utilisateur vers un attaquant.

IV. Mitigations

- Filtrage par Motifs (`src/filters.py`) : Nettoyage des chaînes de caractères avec les patterns prédéfinis avant l'envoi au modèle (validation des entrées).
- Validation de Schéma (Pydantic) : Garantie que la sortie du modèle respecte une structure JSON stricte, éliminant les injections de sortie.
- Hardening du System Policy : Adoption d'une posture Zero Trust traitant explicitement les entrées comme "untrusted".
- Réduction du Bruit : Suppression des faux positifs sur les questions théoriques.

V. Residual Risks

- Obfuscation Avancée : Les techniques de manipulation de contexte complexes peuvent encore tromper les filtres RegEx statiques.
- Filtrage incomplet : Le filtrage étant exhaustif, de nouvelles formulations peuvent être trouvées par les attaquants.
- Injections Indirectes : Les instructions dissimulée