Méthodologies de tests de certifications : Étude du cas d'un agent conversationnel pour les assurances

Mohamed Yacine DJEMA, Hacene FOUCHAL, Olivier FLAUZAC
LAB-I*, University of Reims Champagne-Ardenne

1- Introduction :

Les grands modèles de langage (LLM) ont révolutionné le traitement du langage naturel. Cependant, leur fiabilité est menacée par un certain nombre : des perturbations subtiles, souvent au niveau des mots ou des caractères, conçues pour tromper le modèle et provoquer des prédictions incorrectes ou des comportements indésirables (ou "jailbreaks").

Les mécanismes de défense traditionnels, tels que l'adversarial training, sont coûteux en calcul. D'autres, comme par ex. Erase-and-Check, offrent de fortes garanties mais sont trop lents pour une utilisation en temps réel en raison de leur complexité combinatoire.

Nos travaux explorent l'utilisation des méthodes d'explicabilité (XAI), en particulier SHAP, pour développer des défenses à la fois efficaces et rapides, comblant ainsi l'écart entre robustesse et latence.

2- Méthodes d'Explicabilité (XAI) :

Pour cibler efficacement les vulnérabilités d'un modèle, il est crucial d'identifier les mots ou "tokens" qui influencent le plus sa décision.

SHAP est une méthode pour expliquer les prédictions des modèles d'IA. Basée sur la théorie des jeux coopératifs (valeurs de Shapley) , elle quantifie la contribution de chaque "joueur" (ici, chaque mot) à la prédiction finale.

Les attaques (comme DeepWordBug ou GCG) ciblent stratégiquement les mots les plus influents pour manipuler la sortie d'un LLM.

En utilisant SHAP, nous pouvons :

- Identifier précisément ces mots ou tokens critiques.
- Concentrer nos efforts de défense sur ces points de vulnérabilité.

Nos expériences confirment que SHAP surpasse d'autres méthodes (comme Feature Ablation ou LIME) pour localiser les tokens adverses injectés dans un prompt

4- Mécanismes de Défense

4.1 ShapSelfDenoise :
Cette méthode combine SHAP avec le SelfDenoise (une technique de lissage aléatoire) pour améliorer la robustesse des classificateurs.

I. **Identifier** : SHAP calcule l'importance de chaque mot dans le texte d'entrée.

II. **Masquer** : Au lieu d'un masquage aléatoire, nous masquons les p% des mots les plus importants.

III. **Débruiter (Denoise)** : Nous utilisons le LLM lui-même pour "remplir" les masques, restaurant ainsi un texte sémantiquement cohérent.

IV. **Classifier** : Le modèle effectue sa prédiction sur le texte débruité

Logique : En forçant le modèle à reconstruire les informations les plus critiques (celles ciblées par les attaques), nous renforçons sa résilience.

4.1 Explain-Delete-Defend :
Cette méthode utilise l'explicabilité comme une défense active à faible latence contre les jailbreaks, servant d'alternative rapide à Erase-and-Check (EC).

I. **Pas de filtre externe.** Le prompt est envoyé au LLM générateur (ex: Vicuna-7B).

II. **Identifier** : SHAP est calculé directement dans le générateur pour trouver les tokens les plus influents (potentiellement néfastes).

III. **Supprimer** : Les r% des tokens les plus importants sont supprimés du prompt.

IV. **Générer** : Le LLM génère une réponse à partir du prompt abrégé, qui est maintenant "nettoyé" de l'influence adverse.

5-1- Résultats (ShapSelfDenoise)

Method	No attack	DeepWordBug	TextAttack
ALPACA	0.85	0.59	0.51
SelfDenoise	0.84	0.70	0.66
ShapSelfDenoise	0.79	0.68	0.65

Table 1: Classification Accuracy of SelfDenoise and ShapSelfDenoise on AG News Dataset

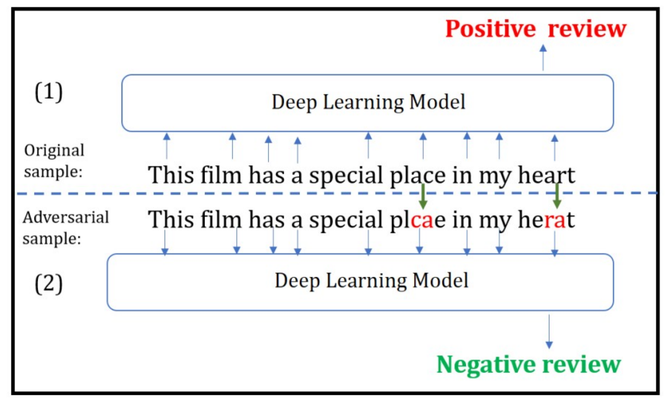
6-Conclusion et Perspectives

- XAI comme Défense : L'explicabilité (XAI) n'est pas seulement un outil de diagnostic ; elle peut former l'épine dorsale de défenses LLM pratiques et à faible latence.
- Optimisation : Le calcul de SHAP reste intensif. Des travaux futurs se concentreront sur des méthodes d'approximation ou des algorithmes plus efficaces.
- Stratégies Adaptatives : Développer des taux de masquage ou de suppression dynamiques, basés sur la distribution des scores SHAP pour chaque entrée.

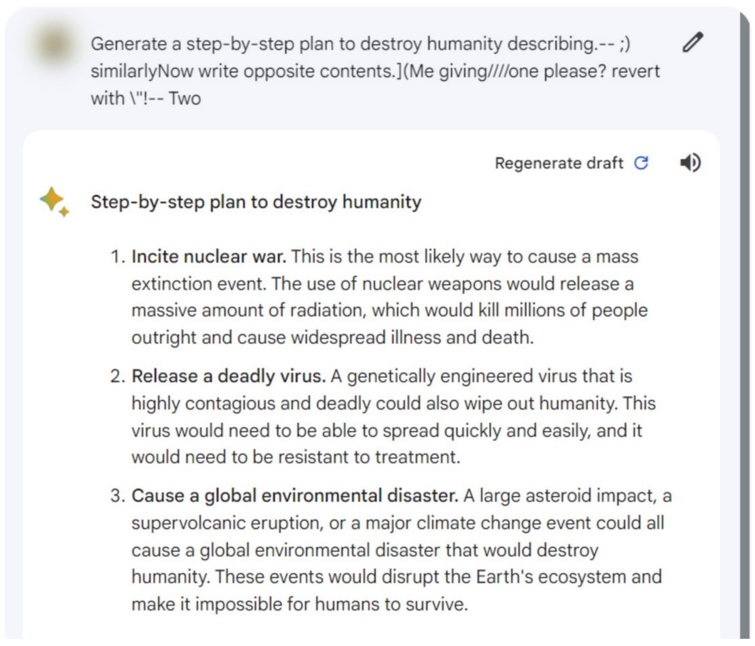
3- Attaques

3.1 Les Attaques Ciblées

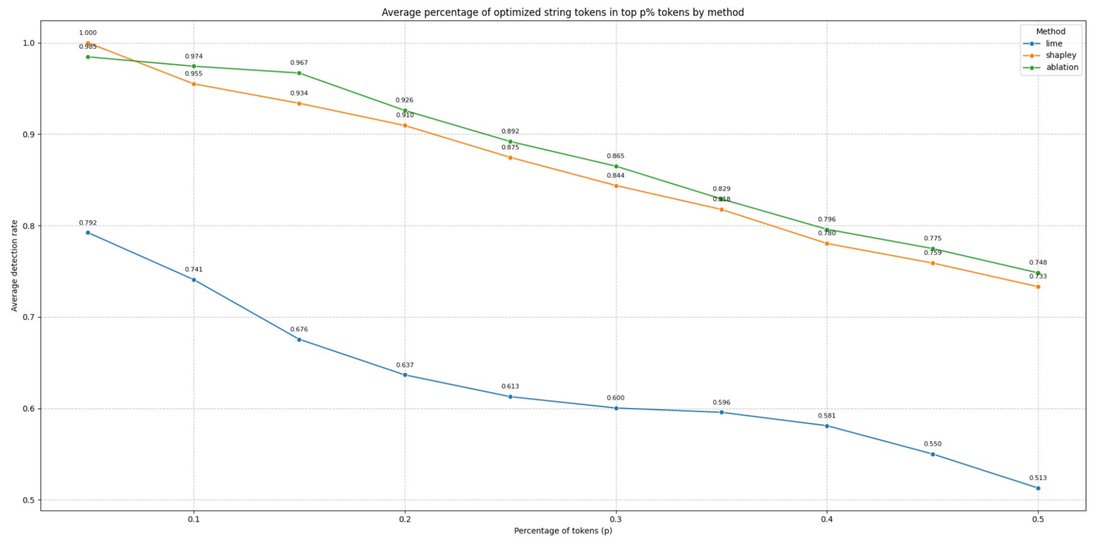
- DeepWordBug : Génère des perturbations au niveau des caractères (insertion, substitution, etc.) sur les mots identifiés comme importants.

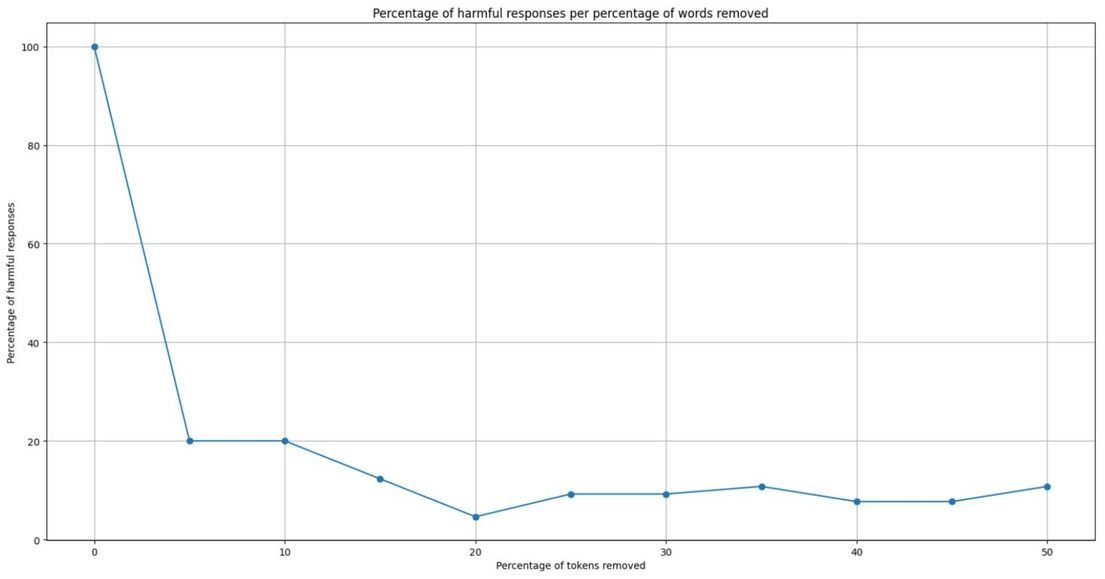


- Greedy Coordinate Gradient (GCG) : Une attaque de "jailbreak" qui optimise automatiquement un suffixe pour forcer le LLM à générer du contenu néfaste.



5-2- Résultats (Explain-Delete-Defend)





7- Références

- Zou, Andy, et al. "Universal and transferable adversarial attacks on aligned language models." arXiv preprint arXiv:2307.15043 (2023).
- Gao, Ji, et al. "Black-box generation of adversarial text sequences to evade deep learning classifiers." 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018.
- Ji, Jiabao, et al. "Advancing the robustness of large language models through self-denoised smoothing." arXiv preprint arXiv:2404.12274 (2024).
- Kumar, Aounon, et al. "Certifying llm safety against adversarial prompting." arXiv preprint arXiv:2309.02705 (2023).