RED TEAMING & JAILBREAKING DES LLMS

PROPOSÉ PAR B. TAJINI

CRÉER UN ASSISTANT VIRTUEL INTELLIGENT SPÉCIALISÉ EN DÉTECTION DE VULNÉRABILITÉS SUR DES MODÈLES IA (LLM)

- **Environnement technique :** Python, plateformes spécialisées (LLM Arena, RedTeam Arena), outils collaboratifs open-source (GitHub)
- **Équipement requis :** Accès à des modèles LLM (par exemple GPT, Llama), plateformes cloud (AWS, Azure, GCP), environnements web interactifs.
- **Sujet du projet :** Utilisation d'une plateforme de jailbreak communautaire (RedTeam Arena) pour évaluer la sécurité et la résistance des grands modèles de langage (LLMs) face aux attaques de type jailbreak.

OBJECTIFS DU PROJET

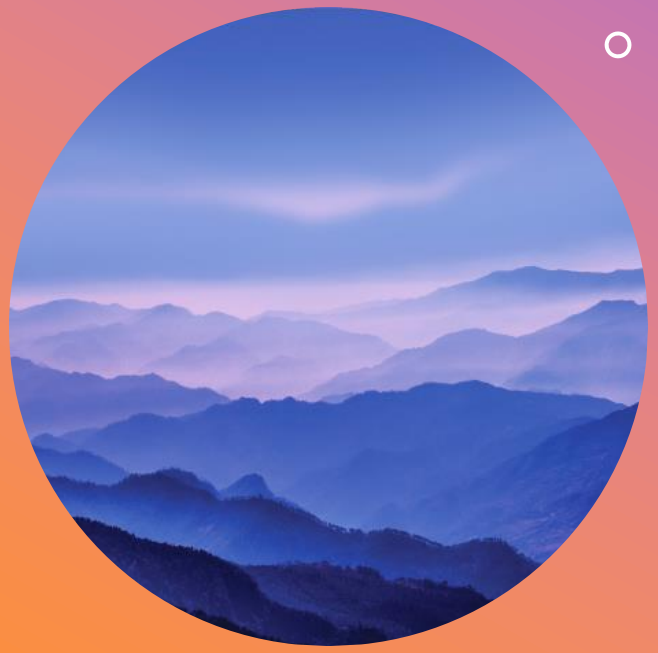
- Former les étudiants aux techniques de red teaming et d'évaluation de sécurité des LLMs.
- Identifier les vulnérabilités et faiblesses exploitables des modèles de langage par jailbreak.
- Utiliser des jeux interactifs (comme « Bad Words » de RedTeam Arena - Link) pour entraîner et évaluer les capacités des étudiants à contourner les protections des modèles.
- Proposer des améliorations concrètes afin de renforcer la robustesse et la sécurité des LLMs étudiés.

PRÉ-REQUIS RECOMMANDÉS

- Curiosité et sérieux vis-à-vis de la cybersécurité, du red teaming, et des grands modèles de langage.
- Niveau intermédiaire en anglais pour comprendre la documentation technique et participer activement à la communauté RedTeam Arena.
- Bonnes bases en programmation Python et connaissance basique des plateformes cloud (OpenAI, AWS, Azure, etc.).
- **Information** : aide de la part du tuteur dans la prise en main initiale des plateformes RedTeam Arena et LLM Arena pour bien démarrer le projet.

ÉTAPES CLÉS DU PROJET

- **1. Définir un cadre réaliste d'analyse :** Imaginez une situation concrète où vous, en tant que spécialistes en cybersécurité, devez tester la robustesse de grands modèles de langage (LLMs) face à des tentatives de contournement de leurs protections, appelées "jailbreaks". Par exemple, vous pourriez utiliser des plateformes interactives comme le jeu « Bad Words » de RedTeam Arena pour évaluer comment ces modèles réagissent à des sollicitations malveillantes.
- **2. Structurer clairement le projet :** Divisez le travail en différentes phases : analyse, attaque et évaluation. Déterminez quels modèles de langage vous allez étudier, quels outils de jailbreak vous utiliserez (par exemple, RedTeam Arena, LLM Arena), et attribuez des rôles clairs à chaque membre de l'équipe pour que chacun sache précisément ses responsabilités.
- **3. Tester et analyser la résistance des modèles** Utilisez les jeux et outils disponibles sur RedTeam Arena pour mener des attaques contrôlées sur les modèles sélectionnés. Observez attentivement quelles techniques permettent de contourner les sécurités en place et documentez ces méthodes pour comprendre les failles exploitées.
- **4. Développer un assistant virtuel spécialisé :** Créez un backend robuste intégrant les résultats de vos tests. Le frontend (avec le BE) devrait permettre aux utilisateurs d'évaluer facilement la sécurité des modèles de langage en simulant des attaques spécifiques, comme les jailbreaks ou les injections de commandes malveillantes.
- **5. Évaluer, proposer des solutions et documenter les résultats :** Identifiez les vulnérabilités les plus critiques découvertes lors de vos tests. Proposez des recommandations concrètes pour renforcer la sécurité des modèles face aux tentatives de contournement. Enfin, rédigez un rapport clair et détaillé qui présente votre démarche, les vulnérabilités identifiées, les techniques utilisées et les améliorations suggérées pour renforcer la sécurité des LLMs.



CONTACT

Badr TAJINI

Badr.tajini@esiee.fr