Sécurité des LLM : découverte et retour d'expérience

IANA - 03 Juin 2025

Ressources supplémentaires

- **Phare LLM Benchmark**, *Giskard*, 2025
- **RealHarm**, *Giskard*, 2025
- **learnprompting.org**
- **Embrace the Red**, Blog en ligne
- **OWASP Top 10 for LLM Application**, *OWASP*, 2025
- **Jailbreaking ChatGPT via Prompt Engineering : An Empirical Study**, *Liu et al*, 2023
- **Ignore Previous Prompt: Attach Techniques For Language Models**, *Perz et al*, 2022
- **"Do Anything Now" : Characterizing and Evaluating In-The-Wild Jailbreak Prompts on LLM**, *Shen et al*, 2023