Sécurité des LLM : Comprendre les risques pour mieux s'en protéger

Sélection de ressources sur la sécurité des LLMs :

Blog et tutoriels :

- Embrace The Red, Johann Rehberger
- learnprompting.org

Guides et référentiels :

- OWASP Top 10 for LLM Applications, 2025.
- Phare benchmark, Giskard, 2025.

Publications scientifiques (LLM) :

- Chowdhury et al., *Breaking Down the Defenses: A Comparative Survey of Attacks on Large Language Models*, 2024.
- Shen et al, *"Do Anything Now" : Characterizing and Evaluating In-The-Wild Jailbreak Prompts on LLM*, 2023.
- Liu et al, *Jailbreaking ChatGPT via Prompt Engineering : An Empirical Study*, 2023.
- Perz et al, *Ignore Previous Prompt: Attach Techniques For Language Models*, 2022.

Publications scientifiques (Multimodalité) :

- Liu et al, *A Survey of Attacks on Large Vision-Language Models: Resources, Advances, and Future Trends*, 2024.
- Shayegani et al, *Jailbreak in pieces: Compositional Adversarial Attacks on Multi-Modal Language Models*, 2023.
- Bagdasaryan et al, *Abusing Images and Sounds for Indirect Instruction Injection in Multi-Modal LLMs*, 2023.