From adversarial examples to LLM jailbreaks: designing langage models that are adversarially robust

Supervisors

- Benjamin Negrevergne: Lamsade, Université PSL – Paris Dauphine (`benjamin.negrevergne@dauphine.fr`)
- Rafael Pinot, LPSM, Sorbonne Université (`pinot@lpsm.paris`)
- Alexandre Allauzen: Université PSL – Paris Dauphine, ESPCI (`alexandre.allauzen@dauphine.psl.eu`)

Context and Motivaton

LLM-based systems are capable of producing increasingly complex and structured outputs: they can generate coherent stories, write code, or devise solutions to mathematical problems. People interacting with these systems generally do so via a *prompt*, which constitutes the specification (in natural language) of the task to be performed. For sufficiently complex tasks, it is common for users to iterate multiple times and refine the prompt until they reach a specification that is precise enough for the model to consistantly produce a satisfactory solution (a.k.a. *prompt engineering*). Nowadays prompts are often shared and optimized by online user communities (e.g., promptbase.com), and later reused by less expert users.

This new way of using generative models may represent a significant threat to users. Copied-and-pasted prompts can contain hidden instructions intended to produce a malicious outcomes (a.k.a. *prompt injection*). For example, in the context of code generation, a prompt intially intended to generate an HTML parser could be modified in an undetectable way to also include a backdoor, allowing execution of commands hidden in HTML comments. This situation is particularly critical in code generation, where the LLM's output is not intended for a human user, but for a compiler or interpreter, which unlike humans, does not inspect the code to execute.

Right now, systematic human inspection and alignment procedures designed to prevent models from doing harmful operations are the only available tools to prevent such attacks, however humans can be tricked and models have been repeatidly jailbreaked (e.g. [CNCC+24], [ZWC+23]) so they do not offer reliable protection.

Goal of the Ph.D.

In many respects, the problem of LLM jailbreaks is similar to the problem of adversarial robustness. Prompts, likes images, can be manipulated to carry a hidden meaning, while remaining apparently harmless either to the human observer, or to the alignment procedure of the LLM. Building on more than a decade of publications in the field of adversarial robustness [BCM+13, SZS+14, GSS15], Carlini et al. have pointed that alignement procedures, were not in fact *adversarialy robust* [CNCC+24]. In other words, they were weak when a hostile user was proactively trying to work around them.

The Lamsade has built a considerable expertise in the field of *adversarial examples, and adversarial robustness* (See for example: [PER+20, GHSP+25, MSP+21, LMP+25]). In particular, it has developed

new methods (grounded in game theory) in order to train machine learning models that are certifiably robust against adversarial examples. Unlike simple defense mechanisms which can only provide a limited level of robustness, these methods assume an optimal adversary user, proactively trying to work around the defense mechanism.

In this thesis, we want to apply the same rigorous, game-theoretical approach, to the problem of prompt injection. Despite a number of similarities with the traditional adversarial setting, the LLM setting introduces a number of new challenges: a discrete and sparse input space, a different space of acceptable perturbations, an observer checking for malicious manipulations that is not always human, etc.

During this Ph.D., the Ph.D. candidate will acquire strong expertise in both adversarial examples generation and LLM jailbreaking. He or she will implement and analyze existing attacks and defense mechanisms, understand their strengths and weaknesses, and finally devise, analyse and implement novel efficient and adversarially robust defense mechanisms for LLMs.

References

[BCM+13] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[CNCC+24] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36, 2024.

[GHSP+25] Lucas Gnecco-Heredia, Matteo Sammut, Muni Sreenivas Pydi, Rafael Pinot, Benjamin Negrevergne, and Yann Chevaleyre. Unveiling the role of randomization in multiclass adversarial classification: Insights from graph theory. *arXiv preprint arXiv:2503.14299*, 2025.

[GSS15] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[LMP+25] Gnecco Lucas, Sammut Matteo, Muni Pydi, Pinot Rafael, Negrevergne Benjamin, and Yann Chevaleyre. Unveiling the role of randomization in multiclass adversarial classification: Insights from graph theory. In *AIStats*, 2025.

[MSP+21] Laurent Meunier, Meyer Scetbon, Rafael B Pinot, Jamal Atif, and Yann Chevaleyre. Mixed nash equilibria in the adversarial examples game. In *International Conference on Machine Learning*, pages 7677–7687. PMLR, 2021.

[PER+20] Rafael Pinot, Raphael Ettedgui, Geovani Rizk, Yann Chevaleyre, and Jamal Atif. Randomization matters how to defend against strong adversarial attacks. In *International Conference on Machine Learning*, pages 7717–7727. PMLR, 2020.

[SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, and Ian Goodfellow ad Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[ZWC+23] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.