Watermarking of LLMs

Project Description. Large language models (LLMs) have enabled the generation of high-quality synthetic text, which is often indistinguishable from human-written content. Applications include language-based assistants, code generation, writing reports, etc. and merely using synthetic text can have a negative effect on the information ecosystem, e.g. in education and web content generation.

There are multiple strategies to address this problem. One approach requires accessing and storing all LLM interactions. Another approach uses statistical features of text to distinguish human-written from artificial-intelligence-generated text. These can be computationally expensive to run, and are known to perform poorly on out-of-distribution data, which makes their usage very limited. A third approach is text watermarking. Watermarking of LLMs, i.e., the ability to detect and audit the usage of machine-generated text is a key principle of harm reduction for LLMs. A watermark is a hidden pattern in text that is imperceptible to humans, while making the text algorithmically identifiable as synthetic. Text-watermarking can be done during the generative process, or by altering the LLM's training data, i.e., data-driven watermarking.

In this project, we are interested in experimenting with a *generative watermarking scheme*, which builds on previous generative watermarking components, and uses a novel *sampling algorithm.* The successful completion of this project will be a proof that *generative text watermarking* can be implemented and scaled to real-world applications, serving millions of users and playing an integral role in the identification and management of artificial-intelligence-generated content.

Project Type. 60% Engineering, 40% Research.

Internship Batch. Batch 1 from May 11 to July 10, 2025.

Duties/Activities. Read and analyze research papers in machine learning (ML). Examine and update a source code on *watermarking of LLMs.* Train, fine-tune, and deploy the ML model on large-scale clusters. Implement new ideas to improve the performance and accuracy of the -studied-watermarking model. Write code to collect and process various datasets. Analyze the performance of the model using multiple benchmarks. Write reports and presentations explaining the source code, datasets, and results.

Required Skills. Strong programming skills, especially in Python. Ability to work with existing code. Strong mathematical background at the bachelor's or master's level, including good knowledge of advanced topics in Calculus, Linear Algebra and Statistics.

Preferred Intern Academic Level. B.Sc. (3rd year or 4th year) or MS student.

Learning Opportunities. Students will learn different text-watermarking techniques, and how these approaches can be integrated into an LLM with negligible additional computational overhead. Students will gain experience in training and deploying text-watermarking models on large-scale compute clusters with powerful GPUs.

Expected Team Size. It is preferable to have a team of 2 interns.

Mentors. Dr. Issa Khalil, Dr. Abdelkader Baggag