# Watermarks

## A Watermark for Large Language Models

Description

* Selecting a randomized set of “green tokens” before a word is generated
* Softly promoting use of green tokens during sampling
* Statistical test for detecting the watermark

|  |  |  |
| --- | --- | --- |
| Pros: | Cons: |  |
| Negligible impact on text quality (-> Invisible to humans). | Watermark based on “green tokens”. |
| Algorithmically detectable. | Not intended for hiding bit streams. |
| Only a few tokens needed. | False positives are statistically improbable. |

Use Case

🡪 Paraphrasing the news feed so the location of red/green tokens encrypts our bit stream

## The Science of Detecting LLM-Generated Texts

Description

* Overview of existing techniques for detecting LLM-generated text
* Categorization of detection methods into
  + Black-box methods (relying on API-level access) and
  + White-box methods (full access to LLMs)

A diagram of a data processing process

Description automatically generated

* Rule-based and neural-based post-hoc watermarking approaches (white-box) embed a hidden message into the text

|  |  |
| --- | --- |
| Pros | Cons |
| * Post-hoc watermarking is directly applicable for our use case | * Trade-off between effectiveness and text quality |
|  | * Potential for adversaries to detect and reverse engineer the watermark |

Use Case:

🡪 Find post-hoc watermarking methods to embed the hidden message into our news feed.

## CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarking

* Method for safeguarding the intellectual property of text generation APIs
* Embeds watermarks using optimization to make them hard to detect through statistical inspection
* Maintains text quality

## Necessary and Sufficient Watermark for Large Language Models

Description

* Follows the green token/red token concept
* Inserting watermarks without degrading the text quality

## On the Reliability of Watermarks for Large Language Models

* Research on robustness of watermarked text after it is re-written/paraphrased/mixed

A diagram of a person with a table and a chair

Description automatically generated with medium confidence

## Robust Distortion-free Watermarks for Language Models

Description

* Watermarked text generated using random numbers based on watermark key
* Detecting watermarks involves aligning the text with the random number sequence using the key

A diagram of a person

Description automatically generated

Use Case

🡪 Invert this problem?

## Undetectable Watermarks for Language Models

Description

* Watermarks can only be detected with a secret key
* Uses one-way functions
* Computationally indistinguishable from the original model’s output (without key)