Lost in Overlap: Exploring Watermark Collision in LLMs

> https://arxiv.org/abs/2403.10020

- 之前的watermark方法在测试鲁棒性的时候用的paraphraser都是没有加水印的LLM，比如gpt3.5/4，Dipper等。这篇提出用加了水印的LLM做paraphraser，这样paraphrase过的文本其实是两种watermark叠加起来，可以起到很强的水印破坏效果。
- 实验：
  - Text Generation Model：LLaMA-2-13B和OPT-1.3B
  - Paraphrase Model： LLaMA-2-13B

    ```
    Assume you are a helpful assistant. Your job is to paraphrase the given text.
    {INPUT_TEXT}
    ```

  - dataset：C4
  - watermark：KGW, SIR, PRW （交叉做原始watermark和paraphrase watermark）
  - detect：用两种watermark的detector分别检测
- 结果分析：
  - 带有watermark的paraphrase attack会比没带的要攻击性更强一点
  - 不同的水印叠起来会降低检测的准确率，且两种叠起来的时候有一种会在强度上表现的更强一些；可以理解为两种水印在竞争。
  - 当Paraphraser的watermark强度增大时，原始的watermark会被覆盖的几乎检测不出来
  - 当paraphraser用了比较强的model，比如llama-2-13B时，上游的watermark更容易被erase掉
- 讨论
  - 这个现象在实际使用场景中的影响：如果是QA任务，多轮对话的场景，第一个answer是有水印的，如果把这个answer再作为输入，就会出现watermark collision的情况，导致watermark在QA中失效
  - 用watermark collision现象来检测水印：因为已经加过水印的文本如果再加一个weak水印是比较难的，而没有水印的文本再加一个weak水印是比较简单的。可以用这个区别来检测一个文本是否已经加过水印了，但一个问题是不知道加的是哪种水印。