Undetectable Watermarks for Language Models

> https://eprint.iacr.org/2023/763.pdf

这篇文章的一个最大的贡献点是提出了一个不改变token distribution的watermark方法，题目里的Undetectable意思就是不改变token distribution

这类方法的根本思想是Watermark Generator和Watermark Detector之间共享一串从均匀分布采样出来的伪随机数$u1, u2, u3, \ldots, u_L$，伪随机数干预了sampling process，而不是直接modify logits；这样在detect的时候计算生成的文本和这串伪随机数的相关性就可以判断是否加了水印

- 考虑把正常的LLM生成文本的过程reduce成生成0、1编码
- watermarking process（modifies token sampling process）

  $x_j = 1$，if $u_j \leq p_j(1)$ and $x_j = 0$ otherwises
- watermark detection process

  For each text bit x_j, the detection algorithm can compute a score

  $$s(x_j, u_j) = \begin{cases} \ln \frac{1}{u_j} & \text{if } x_j = 1 \\ \ln \frac{1}{1-u_j} & \text{if } x_j = 0 \end{cases}.$$

  Given a string $x = (x_1, \ldots, x_L)$, the detection algorithm sums the score of all text bits

  $$c(x) = \sum_{j=1}^{L} s(x_j, u_j).$$
- 不改变token distribution：因为u_j是从[0,1]均匀分布里采样的，积分一下就知道不改变期望