Publicly Detectable Watermarking for Language Models

> https://arxiv.org/abs/2310.18491

本质：把message - signature pair植入大模型生成的文本中

- 先生成一对secret, public key sk, pk
- 大模型生成的前l个词不受任何干预，这l个词算出来的signature $\sigma = Sign_{sk}(H(t[0:l]))$
- 设这个σ有λ_σ个bit，我们要依次把这λ_σ个bit融入到后续的λ_σ个chunk当中（每个chunk的size是l），我们称这植入σ的一整个过程叫一轮植入。
- 植入过程：

  假设现在要植入σ的第i个bit，让大模型再sample出一段长度为l的文本x，计算$H(m||x||\sigma_{prev})$，其中m为这一轮植入中所有新生成的已经被接受的文本，长度应该为$l \times (i-1)$，σ_{prev}为已经被植入的signature片段，长度为$i-1$；若$H(m||x||\sigma_{prev}) =$ 余下的$signature$片段，则我们让x被接受，否则重新生成x；
- 反着解密，这里m是暴露的，因为如果只考虑一个message-signature pair的植入的话，m就是除了前l个token的文本。由m和H反着算出σ，再验证m, σ是否是一对。