Duwak: Dual Watermarks in Large Language Models

> https://arxiv.org/abs/2403.13000

- 设计了一个既改logits，又改sampling的水印算法：
  - 改logits：用的和KGW一样的算法
  - 改sampling：Contrastive Search

    大体思路现在有一个distribution p，先用取决于previous token的hash来决定下一个位置到底是用contrastive search（有η的概率）还是multinomial sampling（有$1-\eta$的概率）；

    contrastive search：先取top-k缩小sampling的范围；再用和前面window size文本的相似度做惩罚因子来sample。$s_L(x_t^v)$就是x_t这个token和之前window size为L的前缀中的每个token中的cosine similarity中的最大值。

    $$x_t = \underset{v \in V^{(k)}}{\arg\max}\{(1-\alpha)\cdot p_t^v - \alpha \cdot s_L(x_t^v)\}$$
- 检测

  先对logits和sampling分别做检测，得到两个p-value：P_{kgw}和P_{cs}

  再用一个统计量把这俩value合起来：$P = 1 - F_{X^2}(-2(\ln(P_{kgw}) + \ln(P_{cs})), 4)$

  这俩具体怎么算的可以看看原论文
- 感觉这篇搞得有点复杂