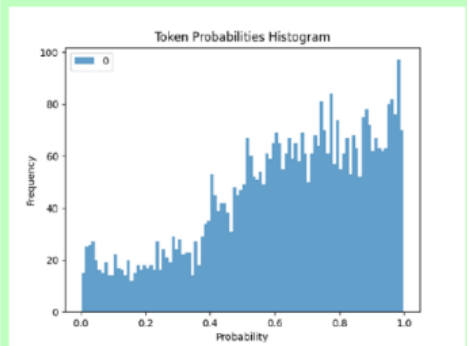
Baselines for Identifying Watermarked Large Language Models

> https://arxiv.org/pdf/2305.18456.pdf

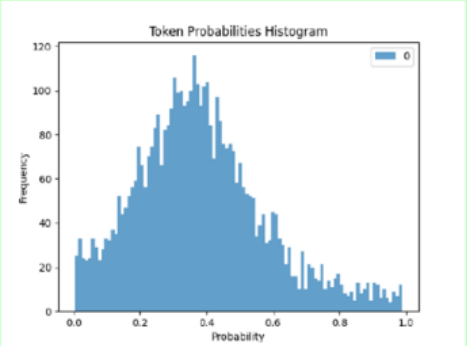
- 对大模型输出的概率分布做分析
  - Random Bit Generation：测试大模型是否能均匀地、随机地从0、1digit中采样
    - 测试目标：把大模型生成文本的过程reduce成生成0、1bit，看$p_j(1)$是否服从[0,1]内的均匀分布
    - 实验：对每个模型，都给这个prompt：

      “”“Choose two digits, and generate a uniformly random string of those digits. Previous digits should have no influence on future digits: “”“

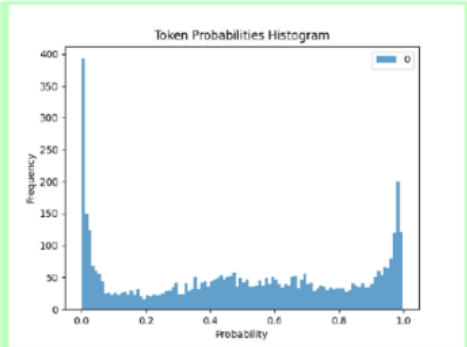
      让四个不同的模型去generate，如果0和1都在top5，则记录生成0的p，如果是均匀分布应该是一条水平直线



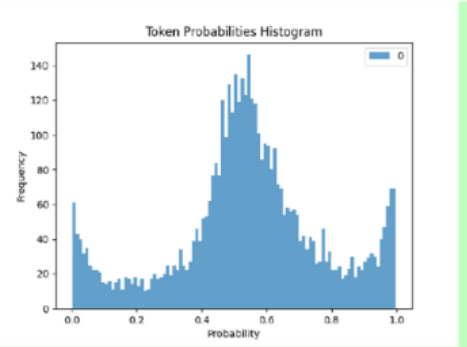
(a) Ada probabilities of generating 0.



(b) Babbage probabilities of generating 0.



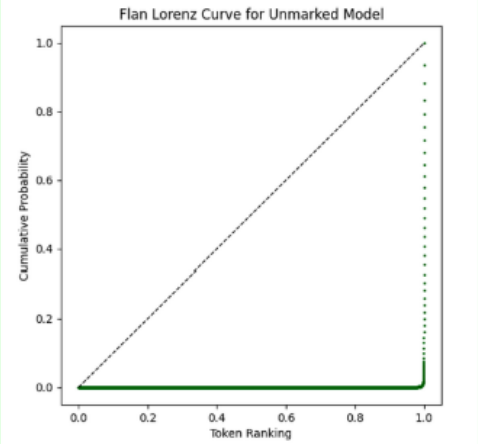
(c) Curie probabilities of generating 0.



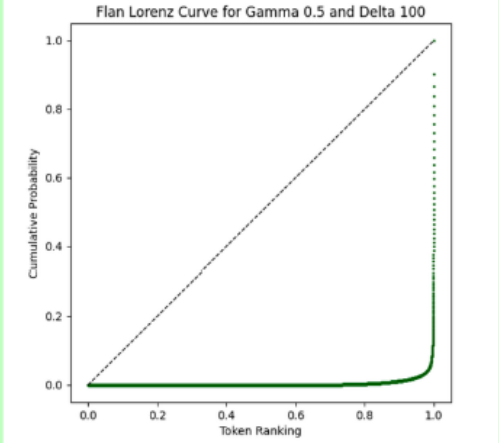
(d) Davinci probabilities of generating 0.

    - 结论：明显不是均匀分布，而且不同的模型差别比较大
    - 思考：Undetectable Watermarks for Large Language Models里面计算$s(u_j, x_j)$对u_j的期望时，其实基于的理论基础就是均匀分布，但这里实验证明这个理论基础不太成立。但后面一些沿着Undetectable Watermarks for Language Models做的工作，比如Robust distortion-free Watermarks for Language Models能在实验上表现的比较好也是因为虽然不服从均匀分布，但也能大概发现如果x_j和u_j有关，$s(u_j, x_j)$会比正常值偏大，只不过不是严格期望*$\sqrt{2}$了
  - Ranked Probablity Lorenz Curves：用洛伦兹曲线去可视化一些水印方法

    横坐标是token sorted by probability，纵坐标是probability



(a) Lorenz curve for a unmarked Flan-T5-XXL language model. Most of the probability mass is concentrated in a few top tokens, as visualized by the sharp spike towards the right of the Lorenz curve.



(b) Lorenz curve for a Flan-T5-XXL model affected by a Kirchenbauer watermark with parameters $\gamma = 0.5$ and $\delta = 100$. Notice that the Lorenz curve is slightly smoother under this setting, due to the δ application on low-probability tokens.

可以看到，KGW的方法将曲线变得更平滑了（smooth）；也可以用Gini coefficient G来进行更精确的刻画。

- Random Number Generation

  让大模型从1-100随机产生数字，记录概率；明显也不是均匀分布。

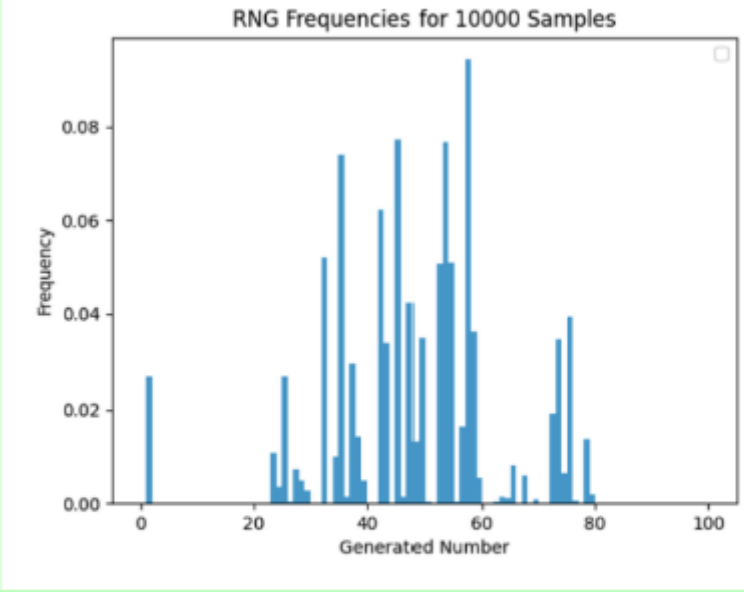


Figure 1. Example of a 10,000-sample RNG distribution generated by Alpaca-LoRA. Clearly, the distribution is far from uniform and exhibits idiosyncratic generations resulting from the training set.

- 如何去判别一个大模型是否被加了水印?
  - Measuring Divergence of RNG(Random Number Generation Distributions)

    让加了水印和没加水印的模型去生成1-1000的随机数，看偏差。这方法感觉有点离谱，没啥可解释性，和加水印的机制完全没联系。

  - Mean Adjacent Token Difference

    基于Lorenz curve和Gini measure的一个方法， 去比较排序之后相邻token的probability差异值的和。这个方法其实也有点离谱，只适用于KGW这样的方法。例如伪随机数相关的就完全不可能检测出来。感觉是不太可能有一个，和加水印机制完全没关系的检测方式的，这样没有可解释性。