A Watermark for Large Language Models

> https://arxiv.org/pdf/2301.10226.pdf

经典文章，Inference time Watermark奠基，Red/Green List范式；把词表分成红/绿，给绿词加bias

Soft Red list rule: Watermark generation & detection

- Analyze the generating process of the nth token

The (n-1)th token id —hash→ The nth position: Green List / Red List

Modify the vector & softmax

$$\hat{p}_k^{(t)} = \begin{cases} \frac{\exp(l_k^{(t)}+\delta)}{\sum_{i\in R}\exp(l_i^{(t)})+\sum_{i\in G}\exp(l_i^{(t)}+\delta)}, & k \in G \\ \frac{\exp(l_k^{(t)})}{\sum_{i\in R}\exp(l_i^{(t)})+\sum_{i\in G}\exp(l_i^{(t)}+\delta)}, & k \in R. \end{cases}$$

- Watermark Detection: Analyze the distribution of the number of green words

No watermark

Every position: γ green $(1-\gamma)$ red

Binomial distribution -> normal distribution

$$N(\gamma T, \gamma(1-\gamma)T)$$

Evaluation Standard: $z = (|S|_G - \gamma T)/\sqrt{(\gamma(1-\gamma)T)}$

For example: z>4 -> machine-produced!