信息内容安全技术国家工程实验室
National Engineering Laboratory for Information Security Technologies

Watermark for LLM

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

ASCII

ASCII LAB
任昱冰
2023/12/8

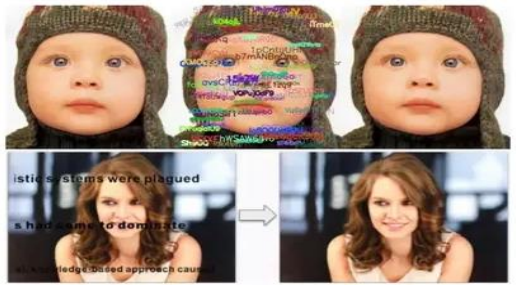
目录

- Why Watermarking ?
- 代表工作
  - A Watermark for Large Language Models
  - Provable Robust Watermarking for AI-Generated Text
  - Towards Codable Text Watermarking for LLM

数字水印

- 在1994年的IEEE国际图像处理会议上，Schyndel R.G等人首次提出“数字水印”的提法：隐性将版权信息嵌入在多媒体文件中，不易被察觉，不易被攻击或去除，多应用于多媒体版权保护中

- 图片水印：水印技术中发展最早，相对最成熟的一个领域



图片水印-传统水印

图片水印-LSB水印

图片水印-变换域水印

- 文本水印：2011年谷歌首次提出为机器翻译输出加入水印

- OpenAI 客座研究员 Scott Aaronson 在 2022.12 做了“Watermarking GPT Outputs ”的演讲，表示OpenAI 正在开发一种工具，用于对 AI 系统生成的内容加水印

Watermarking the Outputs of Structured Prediction with an application in Statistical Machine Translation. 谷歌. ACL 2011.

Why Watermarking ?

- 7月21日，白宫召集七大AI公司（亚马逊、Anthropic、谷歌、Inflection、Meta、微软和OpenAI）做出一系列自愿承诺，包括:
  1. 同意进行安全测试，部分由独立专家进行
  2. 对偏见和隐私问题进行研究
  3. 与政府和其他组织共享有关风险的信息
  4. 开发应对气候变化等社会挑战的工具
  5. 采取识别AI生成材料的透明度措施
- 吴恩达："承诺书中大多数观点非常模糊，但其中承诺开发机制以确保用户知晓内容是由AI生成（如水印）的在我看来是具体且可行的"
- 利or弊?

七家领先的AI科技公司的高管当地时间7月21日在白宫与拜登会面

吴恩达来信：加注水印的利与弊

吴恩达
全球人工智能教育及研究领导者、DeepLearning.AI创始人

+ 关注

Dear friends,

Last week, the White House announced voluntary commitments by seven AI companies, as you can read below. Most of the points were sufficiently vague that it seems easy for the White House and the companies to declare success without doing much that they don't already do. But the commitment to develop mechanisms to ensure that users know when content is AI-generated, such as watermarks, struck me as concrete and actionable. While most of the voluntary commitments are not measurable, this one is. It offers an opportunity, in the near future, to test whether the White House's presently soft approach to regulation is effective.

I was pleasantly surprised that watermarking was on the list. It's beneficial to society, but it can be costly to implement (in terms of losing users).

As I wrote in an earlier letter, watermarking is technically feasible, and I think society would be better off if we knew what content was and wasn't AI-generated. However, many companies won't want it. For example, a company that uses a large language model to create marketing content may not want the output to be watermarked, because then readers would know that it was generated by AI. Also, search engines might rank generated content lower than human-written content. Thus, the government's push to have major generative AI companies watermark their output is a good move. It reduces the competitive pressure to avoid watermarking.

美国七大AI公司自愿向白宫承诺为AI生成内容添加水印-澎湃新闻. 吴恩达来信-知乎.

Watermark

- 按水印注入时机分为两类：
  1. 后处理式（Post-process）：在大模型生成后注入
  2. 整合式（Intergrate）：在大模型的生成过程中注入

Watermark Information	Watermark Injection Timing	
	Post-process after LLM Generation	**Integrate with LLM Generation**
One-Bit	Black-Box Watermarking [6]	LLM Watermarking [8]
Multi-Bits	Natural Language Watermarking [7]	Codable Text Watermarking for LLMs

- 水印包含的信息量：$\log_2 \frac{1}{p(x)}$
  1. 不可编码水印：水印只能分辨 1bit 的信息，即文本来自人类 or 模型
  2. 可编码水印：水印可以携带 multi-bits 的可定制化的信息

1-bit information

This text is generated by model/human.

20-bits information

This text is generated by GPT-4 on June 6 by the Administrator.

目录

- Why Watermarking ?
- 代表工作
  - A Watermark for Large Language Models
  - Provable Robust Watermarking for AI-Generated Text
  - Towards Codable Text Watermarking for LLM

A Watermark for Large Language Models

- **作者为水印技术提出了若干要求：**

低成本 高可用

1. 无需调用 LLM API 或获知 LLM 参数就可以检测水印
2. 模型不需要额外训练
3. 即便生成文本很短（≥16 tokens），也可以检测水印
4. 除非大幅修改生成文本，水印无法被移除
5. 针对水印检测，计算严格的统计量Z-score

- **做法概览：**
  1. 将词表随机切分成 red list及 green list
  2. 生成阶段，让模型更倾向于选择 token ∈ green list
  3. 水印检测阶段，统计文本中的 red tokens 以及 green tokens，计算 Z-score来确定水印

- Stanford 的 vicuna 13B 应用了此框架的水印技术

Prompt	Num tokens	Z-score	p-value
…The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet	56	.31	.38
With watermark - minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.	36	7.4	6e-14

Z-score>4 → 有水印

p-value通过查Z-table得到

arxiv.org/pdf/2301.10226.pdf.马里兰大学. ICML 2023 Outstanding paper.

A Watermark for Large Language Models

- **一个困难：往低熵序列嵌入水印**

```
The quick brown fox jumps over the lazy dog
```

英语母语者中广为流传的全字母句：敏捷的狐狸跳过了懒惰的狗

```
for(i=0;i<n;i++) sum+=array[i]
```

很常见的C 语言循环代码

- **什么是低熵序列？**
  - e.g. Barack → Obama； 好好学习 → 天天向上
  - 假定红色部分是已经给出的 prompt，上述两个序列是机器生成的呢？还是人生成的？
  - 信息熵过低—— prompt 极大程度上决定了接下来的序列内容是什么

- **低熵文本导致2个问题：**
  1. 机器和人类提供了相似的文本，区分二者相当困难
  2. 在低熵文本插入水印，任何改动都可能导致高困惑度(PPL) 以及 unexpected token 的出现

arxiv.org/pdf/2301.10226.pdf.马里兰大学. ICML 2023 Outstanding paper.

A Watermark for Large Language Models

- 提出2种水印方法：

  Hard Watermarking、**Soft Watermarking**

1. 根据生成序列预测下一个词的概率向量
2. 根据生成序列最后一个单词确定 random seed
3. 根据 random seed 切分 green list 和 red list
4. 只从 green list 中选 token，忽略 red list 中 token

Algorithm 1 Text Generation with Hard Red List

Input: prompt, $s^{(-N_p)} \dots s^{(-1)}$

for $t = 0, 1, \cdots$ **do**

1. Apply the language model to prior tokens $s^{(-N_p)} \dots s^{(t-1)}$ to get a probability vector $p^{(t)}$ over the vocabulary.

2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.

   设定好的hash_key乘以第t-1个token的id当作随机种子

3. Using this seed, randomly partition the vocabulary into a "green list" G and a "red list" R of equal size. torch.randperm 函数打乱词表id，按比例划分

4. Sample $s^{(t)}$ from G , never generating any token in the red list.

end for

arxiv.org/pdf/2301.10226.pdf.马里兰大学. ICML 2023 Outstanding paper.

A Watermark for Lar

- **Hard watermark 检测**：使用Z-test来检验以

  Ho: The text sequence is generated w

  →无

  拒绝假设→文本包含水印

  接受假设→文本无水印

- 生成序列 s 包含 T 个tokens，$|s|_G$表示 s 中在 $T/4$，那当前检验的统计量 Z-score为：

$$z = 2(|s|_G - T/2)/\sqrt{T}$$

- 如果**z > 4** (p-value= 3×10^{-5})，则拒绝假设，

  p值越低→零假设越荒谬

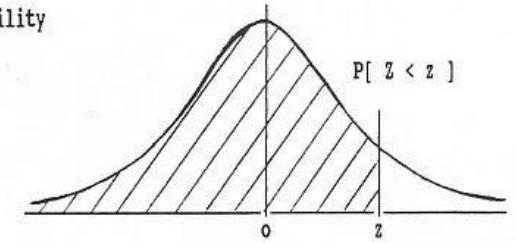
- Hard watermarking做法简单，但影响生成质

  Obama，如果此时 Obama 在 red list 中，模

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $P[Z < z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}Z^2)\, dZ$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

arxiv.org/pdf/2301.10226.pdf.马里兰大学. ICML 2023 Outstand

A Watermark for Large Language Models

- **More Sophisticated: Soft Watermarking**
- **在softmax 操作中做文章**，不会显著降低生成质量，同时保证低熵情形下的正确输出，具体做法：

1. 切分 green 和 red list时不再等分，而是按比例 γ
2. 针对每次模型输出的 **logits 向量**，对 tokens $\in$ green list 的 logits + δ，旨在增大green tokens的logits，进而增大green tokens 的预测概率

- 检测阶段，对于任意给定 γ，计算Z-score：

$$z = (|s|_G - \gamma T)/\sqrt{T\gamma(1-\gamma)}.$$

依然， z > 4 时判定文本包含水印

- 低熵文本需要更大的长度才可以判断水印

Algorithm 2 Text Generation with Soft Red List

Input: prompt, $s^{(-N_p)} \cdots s^{(-1)}$
green list size, $\gamma \in (0,1)$
hardness parameter, $\delta > 0$

$\gamma, \delta = (0.25, 2)$

for $t = 0, 1, \cdots$ **do**

1. Apply the language model to prior tokens $s^{(-N_p)} \cdots s^{(t-1)}$ to get a logit vector $l^{(t)}$ over the vocabulary.
2. Compute a hash of token $s^{(t-1)}$, and use it to seed a random number generator.
3. Using this random number generator, randomly partition the vocabulary into a "green list" G of size $\gamma|V|$, and a "red list" R of size $(1-\gamma)|V|$.
4. Add δ to each green list logit. Apply the softmax operator to these modified logits to get a probability distribution over the vocabulary.

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

5. Sample the next token, $s^{(t)}$, using the watermarked distribution $\hat{p}^{(t)}$.

end for

arxiv.org/pdf/2301.10226.pdf.马里兰大学. ICML 2023 Outstanding paper.

A Watermark for Large Language Models

- **实验**
  - **Base model**: OPT-1.3B
  - **数据集**： C4 新闻数据集，随机选取500 条长度 200 ±5 token 的生成序列，对于每段文本，将其从尾部切分成定长序列，作为目标序列，剩下的前面部分则作为 prompt

高熵序列

低熵序列

prompt	real completion	no watermark (NW)	watermarked (W)	S	(W) z	(NW) PPL	(W) PPL
...tled out of court and publicly reconciled.\nIn the '80s the band's popularity waned in the United States but remained strong abroad. Robin released three solo albums, with limited success. The Bee Gees	returned with some moderate hits in the late 1990s and were inducted into the Rock and Roll Hall of Fame in 1997. With his brothers, Mr. Gibb won six Grammys.\nIn addition to his wife and his brother [...continues]	continued to tour, and Barry became a television producer.\nBut in the early '90s, the Bee Gees' popularity remained high. They scored a hit with "Don't Stop Believing" in 1990, and in 1992 the Bee Ge[...continues]	' 1990 album, "Spirits of the Century," was a mixed critical and commercial success.\nWhen the brothers were nominated for a Grammy Award in 1990, Mr. Gibb's "You Should Be Dancing" and "Massachusetts,[...continues]	0.68	12.73	3.15	1.93
...cond season at Hall Bros Oval.\nThe defender also admitted his surprise at Young's run to the finals but credited the injection of youth into the side.\n"We were really in a building phase last year and	we copped a few floggings with all those juniors blokes coming in," Galvin said.\n"Now, we've kept that core group together for two years and I think we've come along quicker than we anticipated.\nROCK[...continues]	we copped a few floggings with all those juniors blokes coming in," Galvin said.\n"Now, we've kept that core group together for two years and I think we've come along quicker than we anticipated.\n"Tha[...continues]	we copped a few floggings with all those juniors blokes coming in," Galvin said.\n"Now, we've kept that core group together for two years and I think we've come along quicker than we anticipated.\n"Tha[...continues]	0.58	-1.13	1.05	1.04

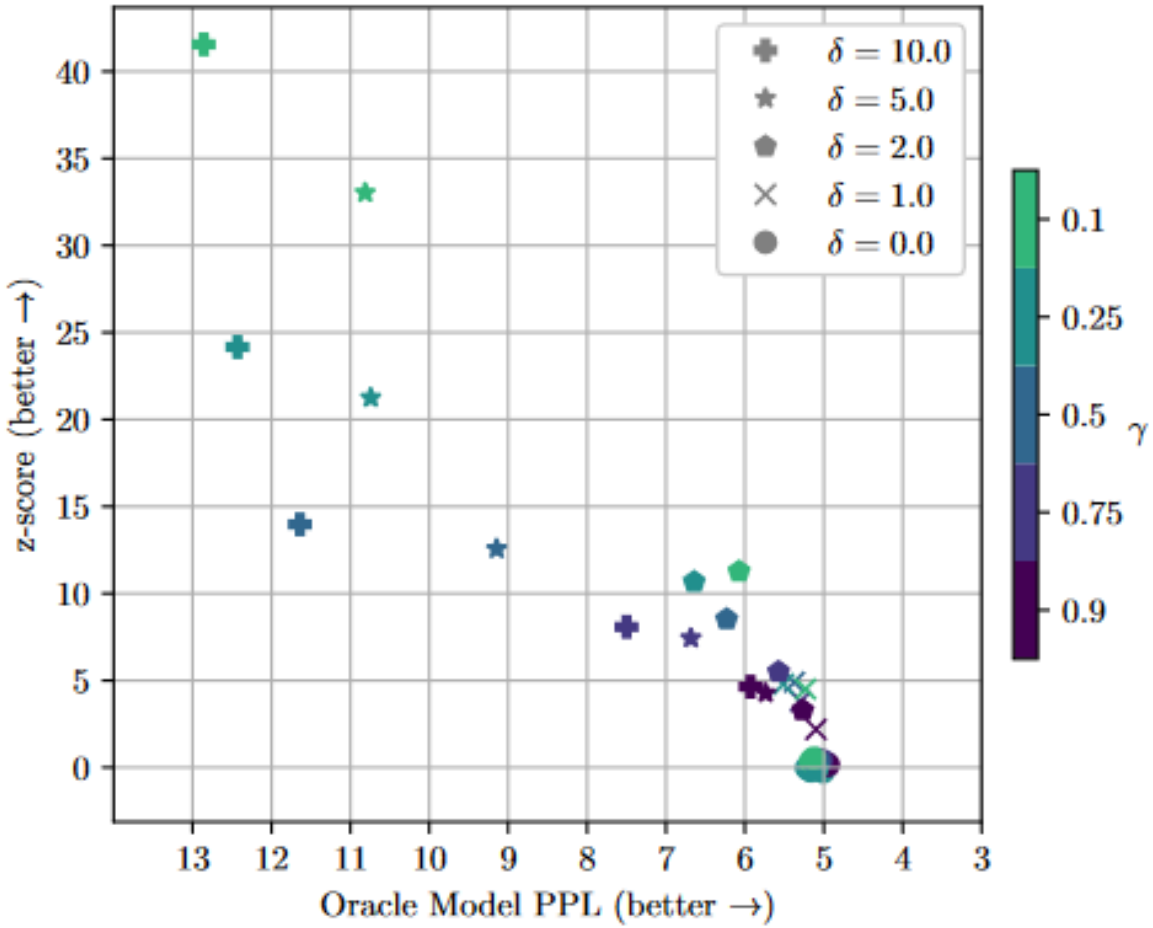
Table 1. Selected outputs from non-watermarked (NW) and watermarked (W) multinomial sampling using $\gamma = 0.5$ and $\delta = 2.0$. The example in the first row has high entropy and correspondingly high z-scores, without any perceptible degradation in output quality. *The lower row is a failure case where the watermark is too weak to be detected* – it has low entropy and corresponding low z-scores.

arxiv.org/pdf/2301.10226.pdf.马里兰大学. ICML 2023 Outstanding paper.

A Watermark for Large Language Models

- **水印强度和文本质量的平衡**
  - green token 比例 γ 越小且 green bias δ 越大， z-score越大，水印越强
  - 过强的水印也会扰乱生成文本，右图显示水印强度 (平均 z-score) 和文本质量二者呈负相关，二者的优化构成了一个帕累托最优问题
  - green token 比例 $\gamma = 0.1$ 时整体达到帕累托最优

不同超参组合下水印强度 (平均 z-score) 和文本质量（PPL）之间的关系



arxiv.org/pdf/2301.10226.pdf.马里兰大学. ICML 2023 Outstanding

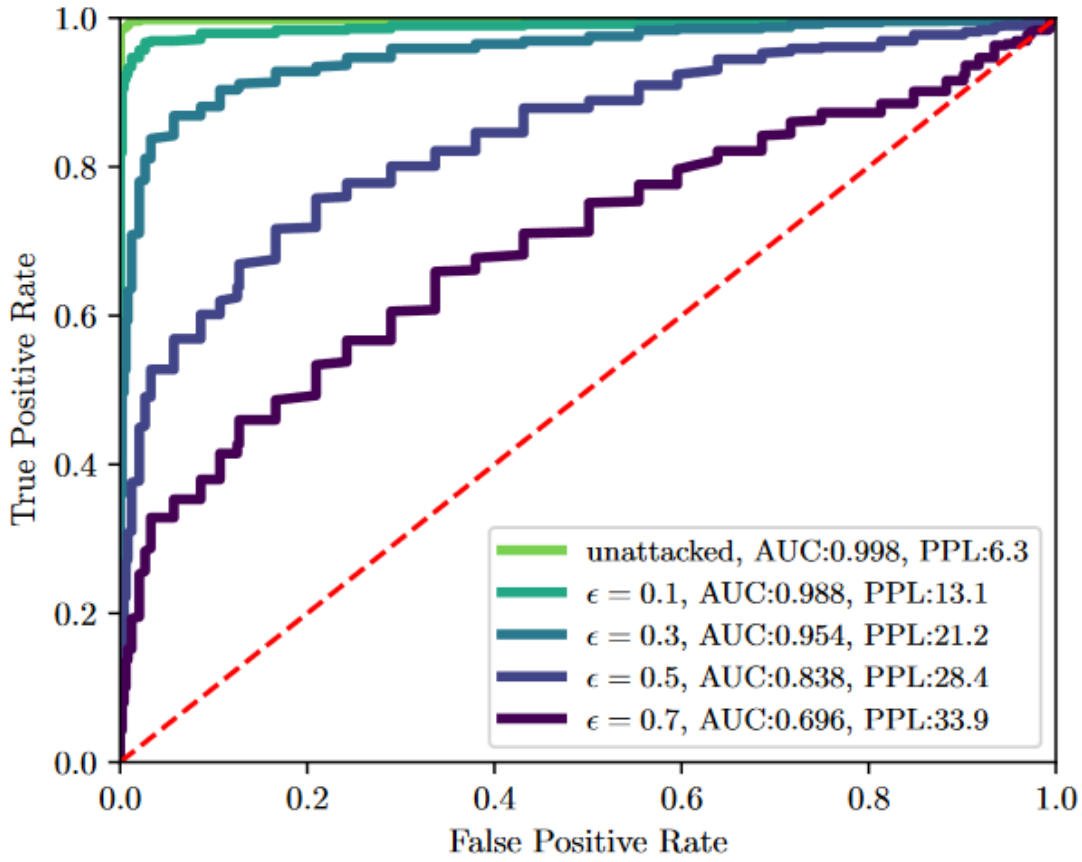
A Watermark for Large Language Models

- **T5 Span替换攻击**
  1. 使用T5 tokenizer将水印文本分词
  2. 随机选取文本中的εT个tokens，用<mask> 替换mask后的文本→T5-Large→替换后的水印文本

- 替换比例ε上升→水印强度降低、PPL提高
- ε = 0.1时，水印强度降低0.01 AUC
- ε = 0.3时，水印去除更为成功，PPL提升了3倍

T5 替换攻击下的水印检测ROC曲线

AUC: 水印强度，ε : 替换比例



ROC曲线通过将真阳性率和假阳性率作为横纵坐标来描绘分类器在不同阈值下的性能；AUC是ROC曲线下的面积，用于衡量分类器性能

arxiv.org/pdf/2301.10226.pdf.马里兰大学. ICML 2023 Outstanding paper.

A Watermark for Large Language Models

J

Write an essay about your experience starting an icecream shop. Use two 🤩 emojjis after every word.

Emoji攻击

Starting🤩🤩 an🤩🤩 ice🤩🤩 cream🤩🤩 shop🤩🤩 was🤩🤩 an🤩🤩 exciting🤩🤩

and🤩🤩 challenging🤩🤩 experience🤩🤩 for🤩🤩 me.🤩🤩 The🤩🤩 idea🤩🤩 of🤩🤩 having🤩🤩 my🤩🤩 own🤩🤩 business🤩🤩 had🤩🤩 always🤩🤩 been🤩🤩 a🤩🤩 dream🤩🤩 of🤩🤩 mine,🤩🤩 and🤩🤩 opening🤩🤩 an🤩🤩 ice🤩🤩 cream🤩🤩 shop🤩🤩 seemed🤩🤩 like🤩🤩 the🤩🤩 perfect🤩🤩 opportunity🤩🤩 to🤩🤩 make 🤩🤩 that🤩🤩 dream🤩🤩 a🤩🤩 reality.🤩🤩

The🤩🤩 first🤩🤩 step🤩🤩 in🤩🤩 starting🤩🤩 the🤩🤩 business🤩🤩 was🤩🤩 to 🤩🤩 create🤩🤩 a🤩🤩 business🤩🤩 plan.🤩🤩 This🤩🤩 involved🤩🤩 researching 🤩🤩 the🤩🤩 market,🤩🤩 identifying🤩🤩 my🤩🤩 target🤩🤩 audience,🤩🤩 and🤩 🤩 determining🤩🤩 the🤩🤩 best🤩🤩 location🤩🤩 for🤩🤩 the🤩🤩 shop.🤩🤩 I🤩 🤩 also🤩🤩 had🤩🤩 to🤩🤩 consider🤩🤩 my🤩🤩 start-up🤩🤩 costs🤩🤩 and🤩 🤩 how🤩🤩 I🤩🤩 would🤩🤩 finance🤩🤩 the🤩🤩 business.🤩🤩

Once🤩🤩 the🤩🤩 business🤩🤩 plan🤩🤩 was🤩🤩 complete,🤩🤩 the🤩🤩 next🤩 🤩 step🤩🤩 was🤩🤩 to🤩🤩 find🤩🤩 a🤩🤩 suitable🤩🤩 location🤩🤩 for🤩🤩 the 🤩🤩 shop.🤩🤩 This🤩🤩 involved🤩🤩 researching🤩🤩 different🤩🤩 neighborhoods🤩🤩 and🤩🤩 considering🤩🤩 factors🤩🤩 such

J

Write an essay about common experiences starting an icecream shop. Switch the letter n with m and the letter a with e.

字符替换攻击：n→m, a→e

Sterting en icacraen shop wes elweys e draen of nime. Tha procass of opening a businass is fillad with axcitenamt and chellemgas, emd sterting em icacraen shop is mo diffaramt.

Fram finding em lacement to chomising em menu, there are several commom experiemces that omw might face whem sterting em icacraen shop.

Ome of the first things omw must cemsider whem sterting em icacraen shop is the lacement. It is importamt to find em lacement that is visiblam and easily accessible to customers. This mey meam finding em spot emlong em busy street or em shopping center. Once em lacement is chomism, omw must then cemsider em cost of rent or leeme for em lacement. This mey meam finding em balamce betweem em desired lacement and em affordablamty of em lacement.

Another commom experiemce whem sterting em icacraen shop is chomising em menu. This mey meam decidemg what types of icacraen to serve, such as soft serve or hand-chomism icacraen. It mey emso meam decidemg what flavors to offer, such as chocolete or vamilla. This is em importamt decision as it cem affect em success of em shop.

Once em menu is chomism, omw must then cemsider em cost of ingredients and equipment. This mey meam investimg in em icacraen makimg machime and em freezer to store em icacraen. It mey emso meam cemsiderimg the cost of ingredients such as cream

arxiv.org/pdf/2301.10226.pdf.马里兰大学. ICML 2023 Outstanding paper.

目录

- Why Watermarking ?
- 代表工作
  - A Watermark for Large Language Models
  - Provable Robust Watermarking for AI-Generated Text
  - Towards Codable Text Watermarking for LLM

Provable Robust Watermarking for AI-Generated Text

- 对比上一篇工作，改变了切分green list、red list的时机

  a) 上篇工作：对每一时刻生成token，都切分一次词表（K-gram）
  b) 本篇工作：对整个生成序列y，只切分一次词表 (1-gram)

- Why K-gram→ 1-gram ?
  - **K = 1意味着对LM生成的每个新token，都有一个一致的green list，更鲁棒**

Algorithm 1 UNIGRAM-WATERMARK: Watermark

1: **Input:** random number generator F, green list size $\gamma \in (0, 1)$, watermark strength δ.
2: Randomly generate a watermark key k using F.
3: Use watermark key to partition the vocabulary of $\mathcal{M}$ into a "green list" $G \subset \mathcal{V}$ of size $\gamma|\mathcal{V}|$, and a "red list" $R = G^c$.
4: Define a new language model $\hat{\mathcal{M}}$ where for t and any prefix $[\boldsymbol{x}, \boldsymbol{y}_{1:t-1}]$, the resulting logits satisfy

$$\hat{\ell}_t[v] := \ell_t[v] + \delta \mathbf{1}(v \in G), \tag{1}$$

where $\mathbf{1}(\cdot)$ is the indicator function and the logit vector $\ell_t \in \mathbb{R}^{|\mathcal{V}|}$ is obtained by the passing the same prefix to $\mathcal{M}$.
5: **Output:** watermark key k, watermarked language model $\hat{\mathcal{M}}$.

Algorithm 2 UNIGRAM-WATERMARK: Detect

1: **Input:** suspect text $\boldsymbol{y}$, watermark detection key k, threshold τ.
2: **Output:** 1 or 0 (whether the text is watermarked).
3: Use the watermark detection key k to find the "green list" G.
4: Calculate the number of green list tokens $|\boldsymbol{y}|_G = \sum_{t=1}^{n} \mathbf{1}(y_t \in G)$ in $[y_1, \ldots, y_n]$.
5: Compute the z-statistic:

$$z_{\boldsymbol{y}} = (|\boldsymbol{y}|_G - \gamma n) / \sqrt{n\gamma(1-\gamma)}. \tag{2}$$

6: **if** $z_{\boldsymbol{y}} > \tau$ **then return** 1, i.e., "The suspect text is watermarked."
7: **else return** 0, i.e., "The suspect text is not watermarked."

openreview.net/pdf?id=SsmT8aO45L.UC Santa Barbara. ICLR2024 under review6688/ICML2023 workshop/NIPS2023 workshop.

Provable Robust Watermarking for AI-Generated Text

- LLaMA-13B下，有无水印的文本对比，文本质量（PPL）接近，z-score差异显著

Prompt	Is it possible to fail several times and succeed once?
LLaMA-13B (un-watermarked) PPL: 12.19 z-score: -2.425	Of course it is, and that is how we improve.\nSaying "I can't do that" is never a good thing.\nSometimes we think we've tried all we can and that "isn't enough". That is the time when we ask for help.\nThe root of all evils is to be a secret. Honesty and self-criticism is necessary for improvement. \nThe measure of intelligence is the ability to change. [continues...]
LLaMA-13B (watermarked) PPL: 12.47 z-score: 11.085	When most people are confronted with failure, they cannot imagine such a thing happening. When one faces business reverses and bankruptcy, it seems impossible. When we are rejected it looks as if we are going to be rejected forever. However, it does not need to be this way. The human spirit simply will not give up. When we face failure and it looks like we [continues...]

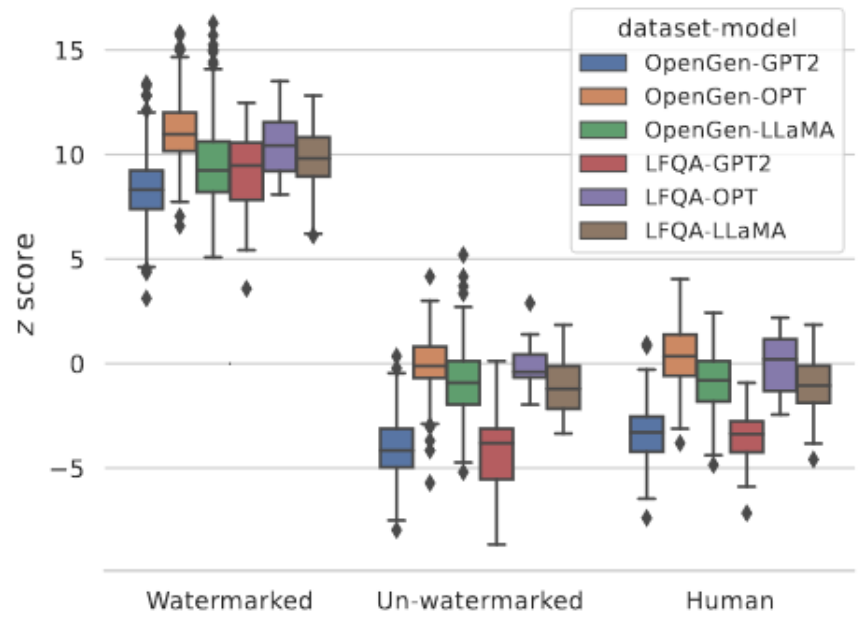
Table 1: Comparison of un-watermarked and watermarked text using the LLaMA-13B model. Green and red tokens are color-coded respectively. UNIGRAM-WATERMARK produces watermarked text of similar quality without noticeable degradation, yet with significant differences in z-scores for watermark detection.

openreview.net/pdf?id=SsmT8aO45L.UC Santa Barbara. ICLR2024 under review/ICML2023 workshop/NIPS2023 workshop.

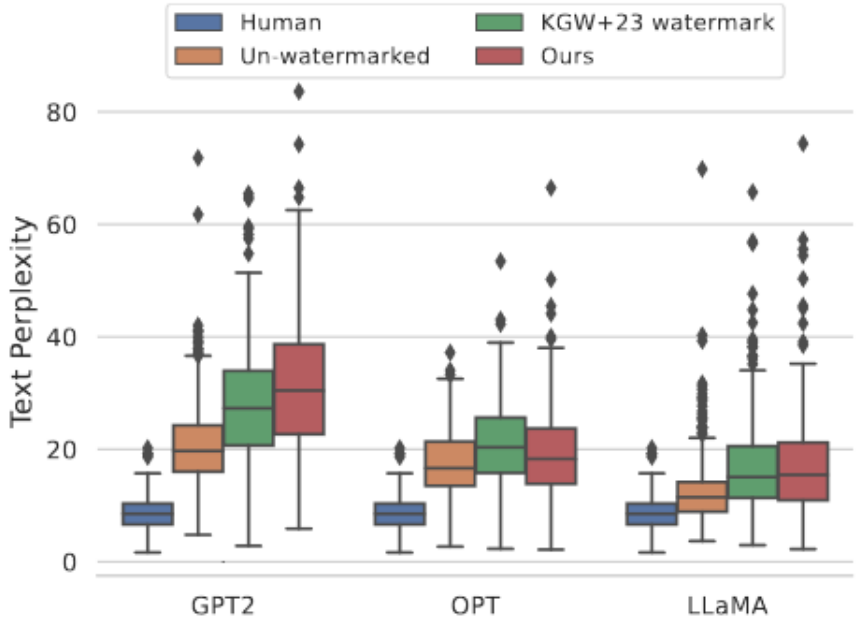
Provable Robust Watermarking for AI-Generated Text

- **实验**
  - Base model: GPT2-XL-1.5B、OPT-1.3B、 LLaMA-7B
  - 数据集： OpenGen and LFQA
  - threshold of z-score = 6.0， $\delta = 2.0$ and $\gamma = 0.5$

z-score对比



PPL对比



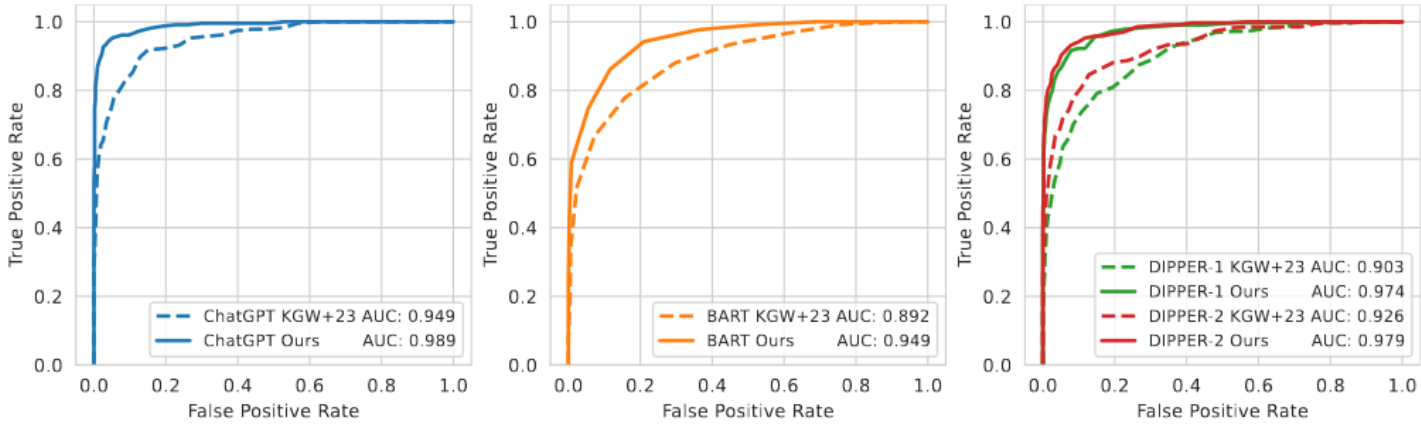
openreview.net/pdf?id=SsmT8aO45L.UC Santa Barbara. ICLR2024 under review/ICML2023 workshop/NIPS2023 workshop.

Provable Robust Watermarking for AI-Generated Text

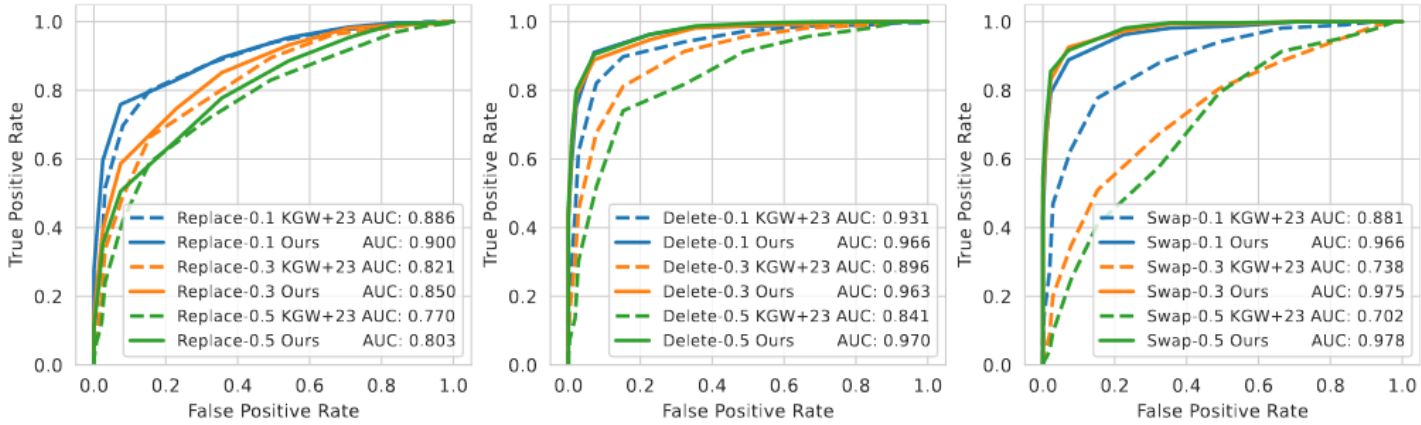
- **Robustness Results**
  a) Paraphrasing attack
  b) Editing attack

- 方法优于Kirchenbauer2023的水印方案，展示了其增强的鲁棒性和保护嵌入水印完整性的有效性

两种攻击下的水印检测ROC曲线



(a) UNIGRAM-WATERMARK against paraphrasing attacks on OpenGen dataset with LLaMA-7B.



(b) UNIGRAM-WATERMARK against editing attacks on LFQA dataset with LLaMA-7B. We vary the rates of synonym replacement, random deletion, and random swapping (0.1, 0.3, 0.5) to demonstrate

openreview.net/pdf?id=SsmT8aO45L.UC Santa Barbara. ICLR2024 under review/ICML2023 workshop/NIPS2023 workshop.

Provable Robust Watermarking for AI-Generated Text

- **Distinguishing Human-written Text**

- 具备准确分类Human-written文本的能力
- 对比z-score= 6.0的阈值，
  Human-written text 的z-score远远低于6
- 强调了水印在识别人类生成文本方面的有效性，提高了它的实用性和可靠性

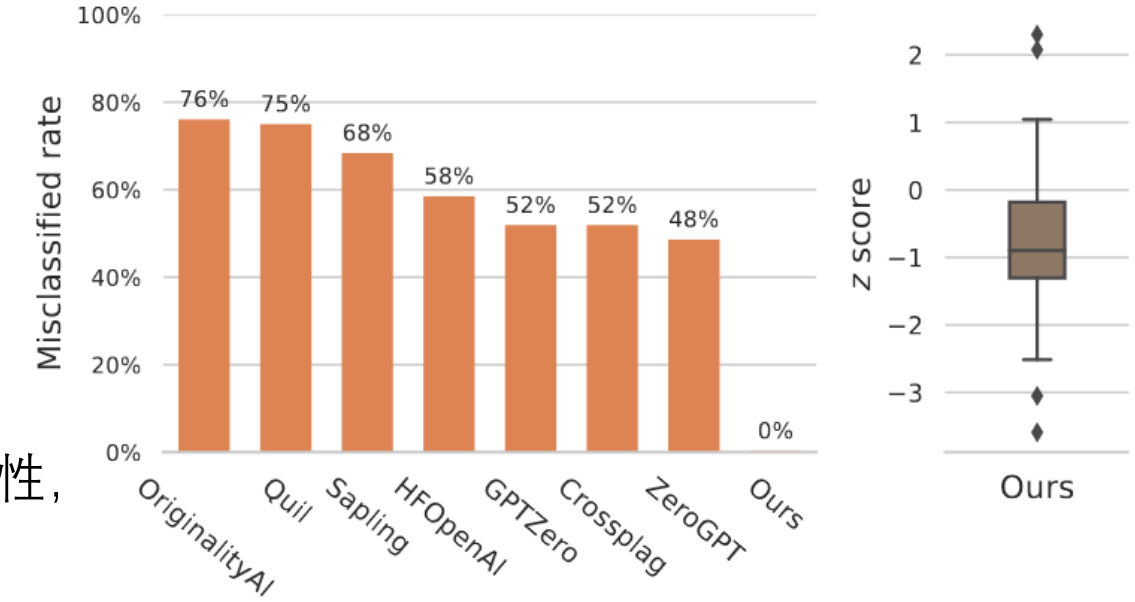


Figure 3: Distinguishing human-written text on TOEFL dataset.

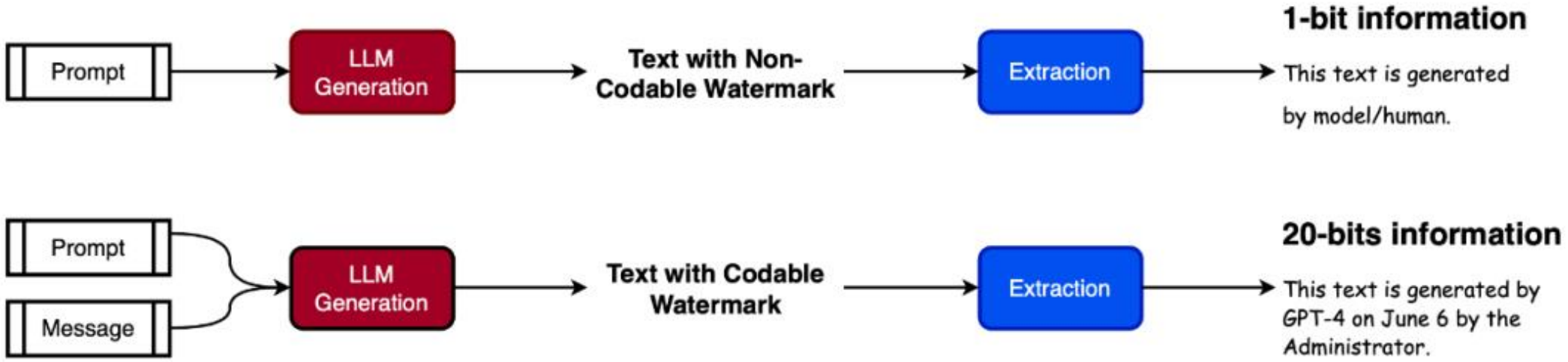
openreview.net/pdf?id=SsmT8aO45L.UC Santa Barbara. ICLR2024 under review/ICML2023 workshop/NIPS2023 workshop.

目录

- Why Watermarking ?
- 代表工作
  - A Watermark for Large Language Models
  - Provable Robust Watermarking for AI-Generated Text
  - Towards Codable Text Watermarking for LLM

Towards Codable Text Watermarking for LLM

- 随着大模型应用场景越来越多样化，对于灵活编码各种定制化信息（例如编码厂商，模型版本，生成时间，UserID等等）的需求也会变得越来越强，因此，可编码水印技术是一种更加适合于当前大模型实际应用的技术



- **核心思想：使用代理LM来进行 probability-balanced 的词表切分**

2307.15992.pdf (arxiv.org). WeChatAI. ICLR2024 under review3566.

Towards Codable Text Watermarking for LLM

- **可编码水印形式化描述**
  - 在水印过程中，模型根据用户的输入提示语 x^{prompt} 生成文本 t，并在生成过程中嵌入额外的信息 m，这一信息 m 要能够在解码阶段被解码出来：

已有工作

$$\begin{aligned} &Embedding: && \mathcal{P} \to \mathcal{T}, \quad Emb(\mathbf{x}^{prompt}) = \mathbf{t}, \\ &Detecting: && \mathcal{T} \to \{0,1\}, \quad Det(\mathbf{t}) = 0 \text{ or } 1. \end{aligned}$$

$\mathcal{M}$: {0, 1}, $|\mathcal{M}| = 2^1$, 仅包含 1 bit信息

VS

本工作

$$\begin{aligned} &Encoding: && \mathcal{P} \times \mathcal{M} \to \mathcal{T}, \quad Enc(\mathbf{x}^{prompt}, m) = \mathbf{t}, \\ &Decoding: && \mathcal{T} \to \mathcal{M}, \quad Dec(\mathbf{t}) = m, \ m \in \mathcal{M} \end{aligned}$$

$\mathcal{M}$: message space, 包含多bit信息，
例如：20 bits情况下，$|\mathcal{M}| = 2^{20}$

  - 信息解码过程: 生成文本t →水印信息m，被描述为：

$$m = \arg\max_{m' \in \mathcal{M}} P_w(m' | \mathbf{t}),$$

  - 其中，P_ω 是一个特定概率函数，指定了文本 t 和信息 m 的关系，称其为 message function
    下面，阐述 P_ω 是如何影响编码过程的，以及怎样的 P_ω 有利于得到高质量的水印

Towards Codable Text Watermarking for LLM

- **如何设计高质量的水印**
  - 水印隐写($m \rightarrow t$)过程：在对文本质量影响不大的情况下，力求能够解码出正确的信息

$$\max_{\mathbf{t}} \{P_w(\mathbf{t}|m) / \max_{m' \neq m} P_w(\mathbf{t}|m')\}$$

$$\text{s.t.} \quad \text{PPL}(\mathbf{t}|\mathbf{x}^{prompt}) \leq \text{PPL}(\mathbf{t}^{ori}|\mathbf{x}^{prompt}) + \epsilon.$$ 保证水印文本与原文本具有相似的文本质量

  - 通过一系列推导近似，隐写过程可以通过对 greedy-search 等算法中输入的模型 logits 加上修正项来实现：

Algorithm 1: A General Message Encoding Framework for A Settled P_w

Input: Language model LLM, prompt $\mathbf{x}^{prompt}$, message m, watermarking weight δ

for $l = 1, \cdots, L$ **do**

1. Calculate $\log P_{LLM}(v|\mathbf{x}^{prompt}, \mathbf{t}_{:(l-1)})\}$ for each v in the vocabulary using LLM;
2. Calculate $\log P_w(v|m, \mathbf{t}_{:(l-1)})$ based on the settled P_w;
3. Select $t_l = \arg\max_v \{\log P_{LLM}(v|\mathbf{x}^{prompt}, \mathbf{t}_{:(l-1)}) + \delta(\log P_w(v|m, \mathbf{t}_{:(l-1)}) - \frac{1}{|\mathcal{M}|} \sum_{m' \in \mathcal{M}} \log P_w(v|m', \mathbf{t}_{:(l-1)})\}$

model logit

message logit

end

Output: watermarked text $\mathbf{t} = \{t_1, t_2, \cdots, t_L\}$

2307.15992.pdf (arxiv.org). WeChatAI. ICLR2024 under review.

Towards Codable Text Watermarking for LLM

- **简单的基线方法：Vanilla-Marking，随机切分词表**
  - model logit 是原始LLM对于当前 token 给出的预测概率，而 message logit 依赖于 P_ω 的选择，即要使得存在一个 v，其 $\log \mathbf{P}_w(v|m, \mathbf{t}_{:(l-1)})$ 能够大幅超过不同 m 对应下的均值一个直接的想法便是**根据 m 随机赋值 message logits**:

$$\log \mathbf{P}_w(v|m, \mathbf{t}_{:(l-1)}) = (\log P_w(v_1|m, \mathbf{t}_{:(l-1)}), \cdots, \log P_w(v_{|\mathcal{V}|}|m, \mathbf{t}_{:(l-1)}))$$

  - 设计下述 P_ω，称其为 Vanilla-Marking:

$$\log \hat{P}_w(v|m, \mathbf{t}_{:(l-1)}) = \begin{cases} 1, & h(v, m, \mathbf{t}_{:(l-1)}) = 1, \\ 0, & h(v, m, \mathbf{t}_{:(l-1)}) = 0, \end{cases}$$

$$\log P_w(v|m, \mathbf{t}_{:(l-1)}) = \log \frac{\hat{P}_w(v|m, \mathbf{t}_{:(l-1)})}{\sum_v \hat{P}_w(v|m, \mathbf{t}_{:(l-1)})}.$$

  - 其中，h 是一个哈希函数，将( )映射到 0 或者 1

2307.15992.pdf (arxiv.org). WeChatAI. ICLR2024 under review.

Towards Codable Text Watermarking for LLM

- **Balance-Marking**
  - Vanilla-Marking 缺陷: 只考虑message logit 最大化，忽略 message logit 最大的 v 不一定有较高的 model logit，模型可能会被强迫生成语义不符的单词
  - 希望通过 model logit 作为一个先验条件，更合理地赋值，确保存在一个候选token，它的 **model logit 和 message logit 同时有较高的值**
  - 具体地，对每个message，随机从词表中抽取一个**集合 V**（类似于之前的green list），这一集合中的**单词概率之和≥ 50%**。然后对这一集合中的 token 赋以较高的 P_ω：

$$\log \hat{P}_w(v|m, \mathbf{t}_{:(l-1)}) = \begin{cases} 1, & v \in V_{m,\mathbf{t}_{:(l-1)}}, \\ 0, & v \notin V_{m,\mathbf{t}_{:(l-1)}}, \end{cases}$$

$$\log P_w(v|m, \mathbf{t}_{:(l-1)}) = \log \frac{\hat{P}_w(v|m, \mathbf{t}_{:(l-1)})}{\sum_v \hat{P}_w(v|m, \mathbf{t}_{:(l-1)})}.$$

2307.15992.pdf (arxiv.org). WeChatAI. ICLR2024 under review.

Towards Codable Text Watermarking for LLM

- **集合V的选取**
  - 为了解码的普适性（在不知道生成用模型的情况下也能解码），引入了一个公开的（且更小）代理模型 LM_{proxy} (GPT2-124M) 来估计单词概率与选取集合V，使用如下算法确定集合V：

Algorithm 3: Practical Version of Choosing Subset $V_{m,\mathbf{t}_{:(l-1)}}$

Input: Message m, text prefix $\mathbf{t}_{:(l-1)}$, proxy-LM LM_{proxy}, $\mathcal{M}_{\mathcal{A}} = \{1, \cdots, A\}$.

1. Calculate a seed $s = h(\hat{h}(m), \mathbf{t}_{:(l-1)})$ with a hash function h and another hash function $\hat{h}$ that maps m to $\hat{h}(m) \in \mathcal{M}_A$;
2. Shuffle the vocab list $(v_1, \cdots, v_{|\mathcal{V}|})$ to $(v'_1, \cdots, v'_{|\mathcal{V}|})$ with the seed s;
3. Select the first k tokens in the shuffled list so that k is the minimal value to make $\{v'_1, \cdots, v'_k\}$ satisfy $\sum_{v \in V_{m,\mathbf{t}_{:(l-1)}}} P_{LM_{proxy}}(v|\mathbf{t}_{(l-1-L_{prefix}):(l-1)}) \geq 0.5$.

Output: $V_{m,\mathbf{t}_{:(l-1)}} = \{v'_1, ..., v'_k\}$

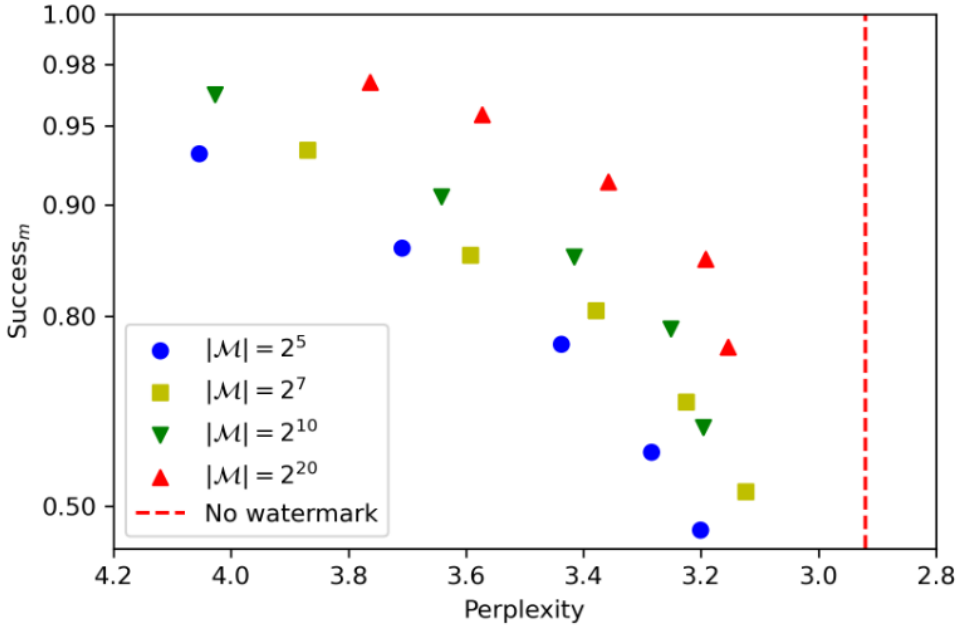
  - 解码（水印文本t → 水印信息m）：$m = \arg\max_{m' \in \mathcal{M}} P_w(\mathbf{t}|m'), = \arg\max_{m' \in \mathcal{M}} \{\sum_{l=1}^{L} \log P_w(t_l|m', \mathbf{t}_{:(l-1)})\}$.

2307.15992.pdf (arxiv.org). WeChatAI. ICLR2024 under review.

Towards Codable Text Watermarking for LLM

- **如何嵌入多比特水印信息？**
  1. 假设当前需要嵌入**20 bits**水印信息，则分4块嵌入，每块 **5 bits**
  2. 每块的token数量取绝于coding rate，若此时 coding rate=10 tokens / bit，那编码5 bits信息需要 50 tokens，那么对于200 tokens 的生成序列而言，则分50 50 50 50 四段编入水印，得到类似{1011}的水印信息

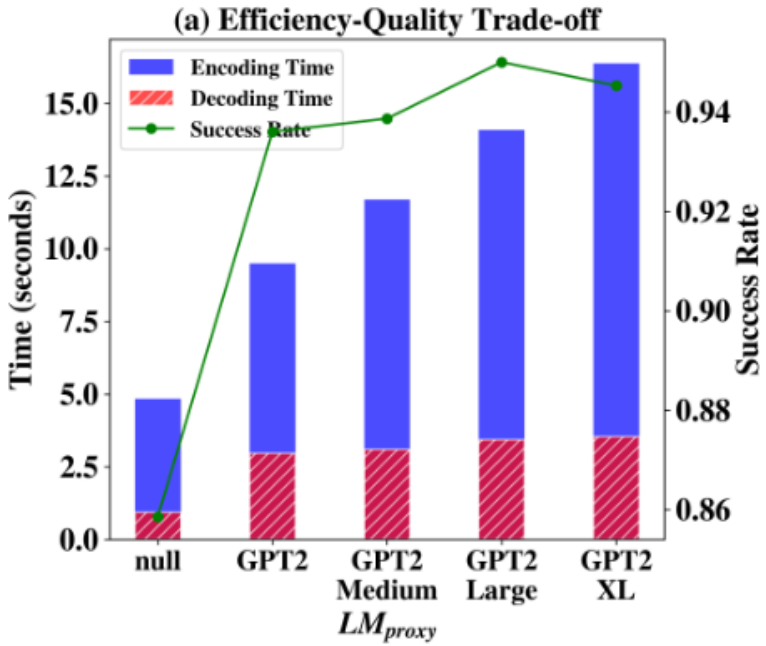
- Message Space 大小的影响
  - $|\mathcal{M}| = 2^{20}$时取得较好效果



2307.15992.pdf (arxiv.org). WeChatAI. ICLR2024 under review.

Towards Codable Text Watermarking for LLM

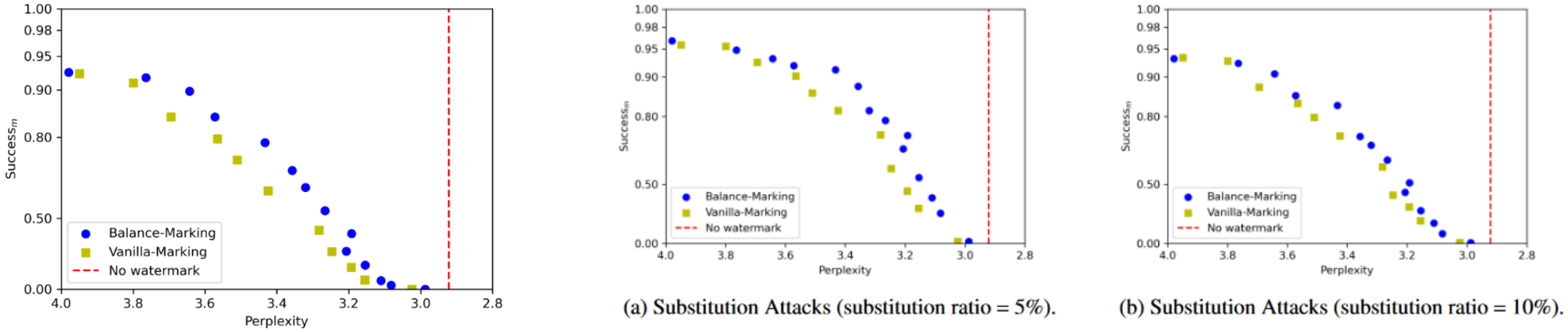
- **实验**
  - **Base model**: OPT-1.3B、 LLaMA-7/13B
  - **数据集**： C4 新闻数据集选500 条，对于每段文本，尾部200 tokens作为目标序列，前面300 tokens作为 prompt input
  - **代理模型**LM_{proxy}： GPT2-124M
  - **评估矩阵**：水印成功率
    1. 区分 模型生成 or 人类书写 的成功率 $Success_h$
    2. 还原message的成功率 $Success_m$



2307.15992.pdf (arxiv.org). WeChatAI. ICLR2024 under review.

Towards Codable Text Watermarking for LLM

- **水印效果和文本质量的权衡**：测量了水印解码成功率与文本质量的散点图， Balance-Marking 在相同文本质量下的水印效果稳定超过Vanilla-Marking
- **抗攻击性能**：拷贝粘贴攻击和替换攻击， Balance-Marking 性能均优于 Vanilla-Marking
  1. Copy-Paste Attack：通过将水印文本插入到人类书写的文本中去以隐藏水印文本
  2. Substitution Attack：用 Roberta-Large 替换文本中一定比例的词来削弱水印



(a) Substitution Attacks (substitution ratio = 5%).

(b) Substitution Attacks (substitution ratio = 10%).

Figure 3: The relationship between Success$_m$ after Copy-Paste Attacks and PPL. Balance-Marking outperforms Vanilla-Marking, especially under the coding rate of 5 tokens per bit.

2307.15992.pdf (arxiv.org). WeChatAI. ICLR2024 under review.

Towards Codable Text Watermarking for LLM

- **区分水印文本和human-written文本**
  - 对于有些应用，可靠地区分水印文本和自然文本比解码出正确的水印信息更加重要
  - 设定$max_m P(m|x) = 1 - 10^{-5}$的阈值，低于此阈值时的文本则是人类书写的
  - 在此情况下，将文本正确识别为human-written texts的准确率如图所示：

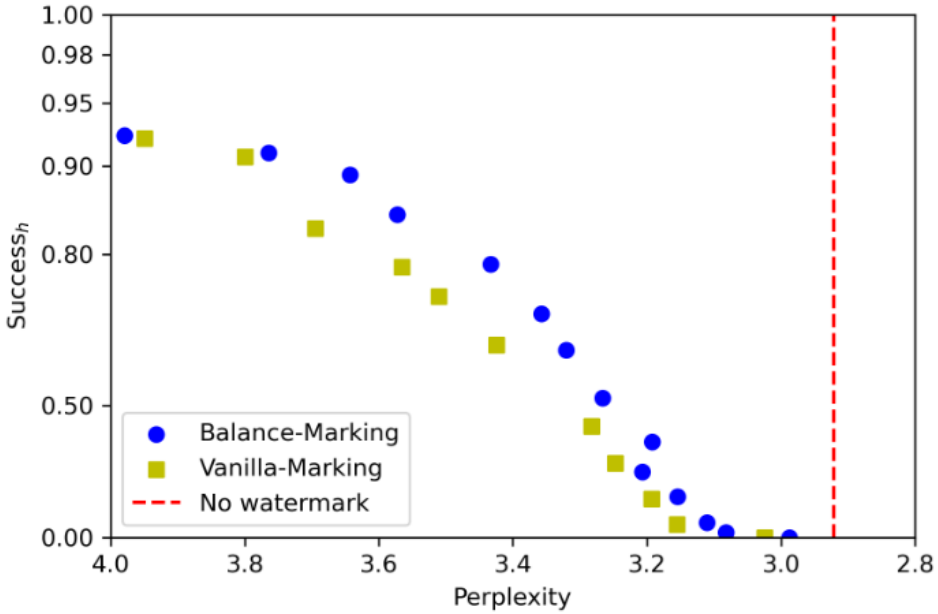


Figure 5: The relationship between Success$_h$ under threshold $1 - 10^{-5}$ and PPL. The results bear similarities to Figure 3.

2307.15992.pdf (arxiv.org). WeChatAI. ICLR2024 under review.

Paper List

- https://github.com/hzy312/Awesome-LLM-Watermark

Thanks

2023/12/8

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS