大语言模型生成水印的研究

刘 述[1] 储志强[2] 张俊霞[1] 张 雷[1] 李红阳[1]
LIU Shu CHU Zhiqiang ZHANG Junxia ZHANG Lei LI Hongyang

摘 要 大模型发展引发了人们的各种忧虑，因此引发对生成文本进行水印标记的需求。对此，提出了一种为生成文本打水印的整合式方法，并给出具体实现方法；同时从概率理论出发，分析了方法的可行性，并通过计算进行了验证。验证结果表明,所提出的方法随着分组长度增长或生成文字的增加,取得很好的效果。

关键词 生成文本；大模型；人工智能；水印；整合式

doi：10.3969/j.issn.1672-9528.2024.05.046

0 引言

大模型的发展，使得生成式文本越来越得到人们的重视，特别是 ChatGPT 的出现给生成式文本带来新的推动力，引发了所谓的“大模型之战”。市场上出现了多种生成文本的大模型，但是生成式文本的大模型出现也带来了很多隐忧。有“ChatGPT 之父”之称的美国人工智能技术公司 OpenAI 首席执行官山姆·阿尔特曼首次出席美国国会听证会，提出了成立监管机构、引入许可证制度、建立一套安全标准等监管建议。如果要对生成的文本进行监管，首先要确定一段文本是否由某个大模型来生成，应有一定的不可否认性。在文献 [1] 中列出了对大语言模型生成文本标记水印的场景：（1）AI 合成的数据质量通常会比人工生成的要低，有必要进行标记，用于人工智能训练时进行相应规避或降权；（2）学术论文利用大语言模型生成会涉及学术伦理的问题；（3）生成假新闻、网页内容等恶意使用大模型工具行为，必要时进行法律的取证。

1 概述

数字水印技术是通过某种算法，在不影响原有多媒体价值及使用的前提下，将标识信息嵌入到多媒体文件中的技术。水印技术在图像中有成熟应用。在文献 [2-4] 中对于数字水印进行了比较全面的介绍。但是在文本处理方面，文献 [5] 认为“最原始的文档，包括 ASCII 文本文件或计算机源码文件，是不能被插入水印的，因为这种类型的文档中不存在可插入标记的可辨认空间”。图像和音像文件在设计时已进行了相应的考虑，使用水印技术不会影响用户的使用感受，反而最简单的文本格式使用水印技术存在着困难。文献 [6] 对于文本的水印也有介绍，提出文本水印应有不可见性、鲁棒性、安全性、容量等特征。文本的水印不可见性是最重要的一点，可见的水印显然是容易去除的，之后所有的特点均是建立在水印不可见性的基础之上。

文献 [7] 提出了文本水印添加时机为：后处理式（Post-process）——水印通过对于已经生成的文本进行后处理的方式加入；整合式（Integrate）——水印的加入过程是和大模型生成过程整合在一起的。水印的信息量：不可编码水印——水印只能分辨 1 bit 的信息，即文本来自人类 or 模型；可编码水印——水印可以携带 multi-bits 的可定制化的信息。

文献 [1] 是生成式文本水印的开山之作。在这篇文献中，作者提出了水印的应用场景：通过操控词汇的统计分布，来隐性进行水印的标识（整合式）。在知道词汇分布的基础上，对收到的文本进行统计，用 z-score 来表明文本是人工生成还是大模型生成（可编码）。可见，整合式、可编码的水印有较高的灵活性和更广泛的应用，是大模型生成水印的一个总体发展方向。

文献 [1] 提出将词表随机切分成 red tokens 和 green tokens，在模型输出时针对 logits 向量，为 green tokens 添加权重，使得模型大大倾向于输出 green token。在检测阶段，结合文本中的 red tokens 和 green tokens 进行统计，通过 z-score 应用正态分布的函数，计算 p 显著量。所述 red tokens 和 green tokens 标记，文献 [1] 中也给出一个实例见图 1。使用这种方法，验证方必须掌握全量的字符集，以生成 red tokens 和 green tokens。而且字符集要和生成式大模型的字符集是完全一致的，因为根据文中提出的算法，red tokens 和 green tokens 分组是随机进行的，如果字符集不一致，就会引起 red tokens 和 green tokens 分布的偏差，从而引起验证的错误。

1. 中国信息通信研究院 北京 100191
2. 中国网络安全审查认证和市场监管大数据中心 北京 100045
[基金项目] 国家重点研发计划 (2022YFB2901501)

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

图 1 以 red tokens 和 green tokens 标记水印实例

2 操作步骤

在文献 [1] 的启发下，本文提出了一种生成式文本水印标记的方法。这种方法相当于事先设定好 red tokens 的字符集，并进行分组。字符集选取原则是字符的出现概率较高，如0.01%，生成文字一方按规则选取一个字符组，作为禁用组，应用于生成文字中。验证方不需要掌握全量的字符集，而只需知道高频字符子集和子集的分组方法，使用相应的字符组验证生成文本是否违反了禁用规则。使用这样的方法，很好地避免了文本生成一方和验证一方必须保证字符全集完全一致的要求。本方法在理论上证实有很好的可行性，具体操作如下。

（1）在某个生成文本的大模型中有自己的字符全集 $\{A\}$，在这些字符全集中，找出一些使用率较高的字符，即信息量不是很高的字符，形成一个全集的子集 $\{B\}$，显然有 $\{B\} \in \{A\}$。

（2）对于 $\{B\}$ 再进行分组 $\{B_1\},\{B_2\},\cdots,\{B_n\}$，各组间可以有交集，但不允许所有的组有共同的交集，也就是：

$$\{B_1\} \cap \{B_2\} \cap \cdots \cap \{B_n\}=\emptyset$$

$\{A\},\{B\},\{B_1\},\{B_2\},\cdots,\{B_n\}$ 的关系如图 2 所示。

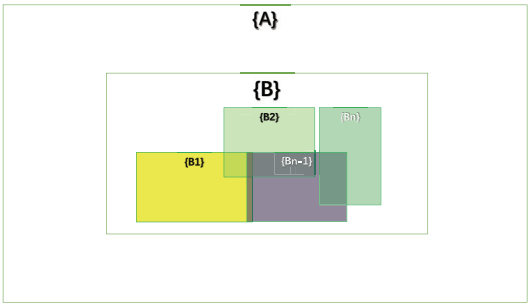

图 2 字符全集、子集和子集的分组

在生成文本时，随机选取一个分组如 $\{B_1\}$，同时要避免使用这个组中的所有字符。

至此，完成了生成文本的水印。生成的文本中不会出现分组 $\{B_1\}$ 中的字符。而生成的文本对于阅读者不会产生任何阅读障碍。

当验证一段文字是否由某一大模型生成时，要先知道分组 $\{B_1\},\{B_2\},\cdots,\{Bn\}$，以及生成本文使用的分组，如是否为 $\{B_1\}$。然后验证生成的文本是否成功避开使用分组中的所有字符，即为不违反 $\{B_1\}$ 的约束。

3 理论上的验证性

假设 $\{B_1\}$ 中有 200 个字符，$\{B_1\}$ 的约束为生成的文本中不允许出现 $\{B_1\}$ 的字符。$\{B_1\}$ 中每个字符在语言统计学出现的概率为 1/10 000，如果生成一段文本时，生成第一个字，那么违反 $\{B_1\}$ 约束概率为：

$$1\text{-}200\times 1/1000=0.98=98\%$$

如果假设生成一个 200 字的文本，那么 $\{B_1\}$ 中字符不出现的概率为：

$$0.98^{200}=0.017\ 6$$

如果有一个对 $\{B_1\}$ 不了解的一方生成一段200字的文字，声称是由某大模型生成的，则有 1.76% 的可能性生成的文本中不包含 $\{B_1\}$ 的字符，不违反 $\{B_1\}$ 的约束。这个结果并不能太让人信服，不过可以通过增加文本长度或增加分组长度，进一步进行验证。

图 3 进行了不同字符长度生成文本，不出现 $\{B_1\}$ 中字符的概率的比较。生成文字越长，不违反 $\{B_1\}$ 的可能性就越低，说明水印标注成功越有效。如果生成一段 1000 字的违反 $\{B_1\}$ 约束概率已到 1.68E-7%，如果一段千字文字没有违反 $\{B_1\}$ 约束，可以很有把握确认该段文字由某大模型以 $\{B_1\}$ 分组进行生成。

生成文本字数	概率 /%
200	1.758 8
300	0.233 3
400	0.030 9
500	4.1E-03
1000	1.68E-7

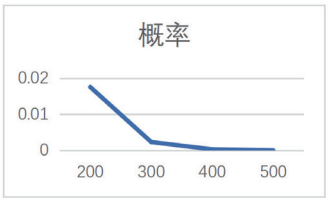

图 3 生成不同长度文本时违反约束的概率图

如果增加 $\{B_1\}$ 中分组中字符的数量，假如生成文本长度为 400 字节，可以进一步说明本方法的可行性，见图 4。随着分组中字符的增加，违反 $\{B_1\}$ 约束的概率随之降低。

分组集的字数	概率 /%
200	0.030 933 6
300	5.113 21E-04
400	8.100 15E-06
500	1.228 69E-07

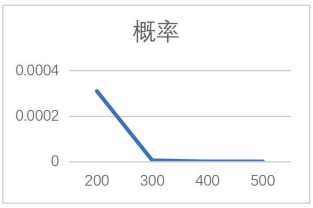

图 4 不同分组集长度下违反约束的概率图

从图 3 和图 4 中可以看出，如果生成文字较长，生成分组的字符可以较少一些。另一方面，如果分组中的字符过多，会增加计算量，而让一个极小概率事件发生的可能性更小并没有更多的工程的意义。

4 应用中的问题

问题一：文献 [8-9] 对于语言统计学进行了一定的介绍。文献 [9] 阐明：“前 7000 个汉字的累计频率达到 0.999 9 以上”。对于子集 {B} 构成要选取信息量不是很高的字符，但是字符也不能信息量过低，比如“的”字在中文中出现的概率非常高，根据文献 [9] 统计达到了 3.8%，文字本身在中文中信息量最低，但是如果文中不出现这类的字符，会使生成文本可读性变差，质量下降。从实际中可以体会到，中文词汇间有一定的可替代性，但信息量非常低的字符出现频率很高，有一定的不可替代性，而信息量非常高的，也就是出现概率极低的字符，如某些生僻字，本身就是不可替代的，所以合理选取文字出现概率对于生成水印质量非常重要。参考语言统计学，为了便于描述，文中使用了简化但合理的分组概率模型，设定分组选取文字出现概率为 1/10 000 的字符。对于具体的语言统计学上字频的统计，超出本文讨论范围。

问题二：分组 {B_1}、{B_2}、…、{B_n} 分组的约束：

$$\{B_1\} \cap \{B_2\} \cap \cdots \cap \{B_n\}=\emptyset$$

如果有一些字符出现在所有的分组中，这些字符将没有机会出现在生成文本中，降低了分组的中可用字符的使用效率。具体操作实现非常容易，如果 {B} 中 500 字符，可以分成 300 字的分组。先把 500 字符按每 100 字分成一块，任选三块构成一个 {B_n} 分组，即可构成 {B_1},{B_2},…,{B_{10}}，共 10 个分组，也满足所有分组的交集为 ∅。

问题三：对于选取哪一个分组进行水印标注也是一个重要的问题，使用的分组最好的方法是分组信息随文本一起传送给验证者，而不需要额外的传递途径。而且分组选择有随机性。受文献 [1] 的启发，这里提出一种传递分组信息的方法。

在应用时，可以按以下步骤进行。

（1）根据大模型输入的提示词，生成第一个词组或短句。

（2）为避免字符编码带来的干扰，把步骤 1 得到短句每个字符变成 UNICODE，按序排列。对每个 UNICODE 进行 BASE64 编码，按序排列形成字符串，对 BASE64 编码字符串进行哈希值计算得到哈希值 H。

（3）以 N 对 H 取模 n=mod(H,N)+1，n 为选中的分组数，N 为字符组总共的分组数。

这样实现了使用分组的随文件传送。选用分组也有随机性。验证者只需根据第一个词组或短句以及哈希算法，就能得到哈希值，取模获得分组数。

问题四：禁止出现分组中的任何一个字，在应用时可能会显得过于严格。对此，可以设定一个门限，允许分组中的字符以一定概率在生成文本中出现。在本方法的应用中还需要进一步研究。文献 [1] 中也有一定的讨论，使用显著度 p 作为衡量结果就意味着对于出现限制字符的包容。

5 总结

文献 [1] 中水印生成和字符集高度相关，仅英文韦氏字典收录单词近 470 000 条 [10]，在中国国家标准 GB 18030—2022《信息技术 中文编码字符集》中收录汉字 87 887 个 [11]。多语言环境下，验证方管理这些字符集也是比较繁重的工作。本文提出的方法，对于验证方不需要知道任何语言的字符集的全集，而只要掌握一个高频字符集，并了解高频字符集构成字符组的方法和生成文字禁用的字符组，便于实现。

随着 ChatGPT 的出现，大模型的应用越来越普遍，越来越大的风险也显现了出来，它们可能被用于恶意目的，如制造虚假新闻和网页内容，或利用大模型生成学术作品等。此外，在互联网上充斥着的 AI 生成数据广泛存在，合成数据的质量通常不及人类生成内容，很多使用者需要事先进行低质量数据的过滤。对于大模型监管的探讨可参见文献 [12]。综上，需要一种方法对大模型产生的内容进行标识，便于识别和监管 AI 生成文本。

本文提出的方法，从理论上有很好的可靠性，生成和验证也容易实现。在应用过程中，使用了信息量低的字符，对生成文本的影响小，不会影响用户的阅读体验，也不会增加生成文本的难度，对于验检单纯由大模型生成的文本有很好的可操作性。

参考文献：

[1]KIRCHENBAUER J, GEIPING J, WEN Y, et al.A watermark for large language models[EB/OL].(2023-01-24)[2024-03-01].https://doi.org/10.48550/arXiv.2301.10226.

[2] 赵翔 , 郝林 . 数字水印综述 [J]. 计算机工程与设计 , 2006, 27(11): 1946-1950.

[3] 程磊 , 张春田 . 数字水印技术 [J]. 电子测量技术 , 2004(6): 74-75.

[4] 夏鸿斌 , 须文波 , 刘渊 . 数字水印技术 [J]. 江南大学学报（自然科学版）,2002,1(2):134-138.

[5] 黄华 , 齐春 , 李俊 , 等 . 文本数字水印 [J]. 中文信息学报 , 2001, 15(5):52-57.

[6] 赵卫娟 , 关虎 , 黄樱 , 等 . 文本水印技术研究综述 [J]. 中国传媒大学学报 (自然科学版),2020,27(6):55-62.

基于自抗扰控制器的平衡小车

朱品伟[1] 钱韬宇[1] 陈展鹏[1] 朱二琳[1]
ZHU Pinwei QIAN Taoyu CHEN Zhanpeng ZHU Erlin

摘 要 为缓和传统 PID 控制器在系统响应快速性和超调性之间的矛盾，设计了一种基于线性自抗扰控制器的跷跷板平衡小车。采用 STM32 单片机为控制核心，通过 MPU6050 传感器采集跷跷板角度，通过串口接收控制器参数，基于自抗扰控制算法输出直流减速电机控制量，实现了小车快速寻找跷跷板平衡点的功能。实践证明，自抗扰控制器无积分控制环节也可以消除静态误差，具有良好的控制品质。

关键词 单片机；自抗扰控制；倾角传感器；MPU6050

doi：10.3969/j.issn.1672-9528.2024.05.047

0 引言

目前，PID 控制算法简单、鲁棒性好和可靠性高，被广泛应用于工业过程控制。PID 控制的比例项能够在误差较大时及时减小误差，积分项可消除静态误差，微分项可抑制超调、减少震荡，使系统快速收敛。在不能建立精确的系统数学模型、系统的结构和参数都存 在不确定性时，PID 控制器是一种合适的选择。但由于实际对象通常具有非线性、时变不确定性、强干扰等特性，应用常规 PID 控制器难以达到理想的控制效果。另外，在生产现场，由于参数整定方法繁杂，常规 PID 控制器参数往往整定不良、性能欠佳。

在 PID 控制器的基础上提出的自抗扰控制器采用了基于误差来消除误差的思想，可以缓和系统响应快速性和超调性之间的矛盾，即便没有积分环节也可消除静态误差[1]。本文设计了一种采用线性自抗扰控制器（linear active disturbance rejection controller，LADRC）的跷跷板平衡小车，通过倾角传感器 MPU6050 模块检测跷跷板的倾斜角度，与平衡角度相比较获得角度误差，经过线性自抗扰控制器处理后，输出

1. 江苏理工学院电气信息工程学院 江苏常州 213001
[基金项目] 江苏省自然科学基金“伺服系统抗扰及跟踪性能优化设计”（BK20221404）

[7]WANG L, YANG W, CHEN D, et al.Towards codable text watermarking for large language models[EB/OL].(2023-07-29)[2024-03-01].https://arxiv.org/abs/2307.15992.
[8] 李国华 . 基于字符信息量法则的串匹配算法研究 [D]. 郑州 : 郑州大学 ,2012.
[9] 游荣彦 .Zipf 定律与汉字字频分布 [J]. 中文信息学报 , 2000(3): 60-65.
[10]How many words are there in English[EB/OL].[2024-03-18]. https://www.merriam-webster.com/help/faq-how-many-english-words.
[11] 中华人民共和国国家质量监督检验检疫总局 . 信息技术 中文编码字符集 :GB 18030-2022[S]. 北京 : 国家市场监督管理总局、国家标准化管理委员会 ,2022.
[12] 韩娜 , 漆晨航 . 生成式人工智能的安全风险及监管现状 [J]. 中国信息安全 ,2023(8):69-72.

【作者简介】
刘述（1972—），男，北京人，硕士，主任工程师，研究方向：IP 网络技术、软件功能测试、软件代码审查、网络设备测试等。

储志强（1981—），男，内蒙古杭锦后旗人，硕士，高级工程师，研究方向：计算机软件开发、大数据、信息化、项目管理等。

张俊霞（1977—），通信作者（email：zhangjunxia@caict.ac.cn），女，河北邢台人，硕士，部门主任，研究方向：科技创新、知识产权、标准化工作、专利保护等。

张雷（1984—），男，黑龙江伊春人，硕士，高级项目经理，研究方向：信创、数据通信、IP 承载网、SDN/NFV、云网融合、下一代互联网等。

李红阳（1986—），女，内蒙古赤峰人，硕士，经济师、工程师，研究方向：互联网、移动互联网、工业互联网、5G 生态等。

（收稿日期：2024-03-08）