

大模型水印研究进展综述:从文本到多模态

吴连伟^{1),2)} 安文博¹⁾ 王 康¹⁾ 王淋永¹⁾ 张 航¹⁾ 王泽浩¹⁾

¹⁾(西北工业大学计算机学院空天地海一体化大数据应用技术国家工程实验室 西安 710129)

²⁾(西北工业大学深圳研究院 广东 深圳 518057)

摘 要 当前,大模型凭借其卓越的推理与生成能力,已在教育、医疗、金融等多个领域展现出广阔的应用前景。然而,由其衍生的数字内容安全与知识产权保护等问题,已逐渐成为制约该技术可持续发展的关键因素。作为应对上述挑战的重要手段,大模型水印受到了学术界与工业界的广泛关注。本文系统梳理了当前大模型水印技术的理论框架与研究进展。首先依据水印嵌入载体的不同,本文将现有水印技术划分为针对文本的大语言模型水印与多媒体大模型水印,并对各类水印技术的特点进行了深入分析,其中,多媒体大模型水印包括图像水印、音频水印及多模态水印。其次,在大模型水印的评估方面,本文围绕大语言模型水印的可检测性、生成内容质量、鲁棒性及水印容量四种核心指标与多媒体大模型水印的评估展开讨论,针对当前水印技术评价体系尚不统一的问题,提出了构建标准化水印评价体系的建议。同时,相关研究提出改写攻击、翻译攻击、释义攻击、模型伪造攻击等针对大语言模型水印的复杂攻击策略,并梳理针对多媒体水印的攻击方式如对抗攻击、频域攻击、时序裁剪等,这些攻击策略对水印可检测性构成严重威胁并为水印鲁棒性带来严峻挑战,本文系统梳理不同攻击策略的作用机理及影响,并总结相应的防御方案,为实际应用中的水印方案选型提供参考。接着,本文进一步阐述了水印技术在保护知识产权与规范模型用途两大典型应用场景中的重要作用。最后,本文从构建可解释与可验证的水印信息、研究智能自适应水印系统等多个视角,对大模型水印技术的未来研究方向进行展望,以促进该技术的持续发展和推广应用。

关键词 大语言模型水印;水印技术;鲁棒性;模型安全;版权保护;文本溯源

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.01870

A Survey of Large Model Watermarking: From Text to Multimodal Content

WU Lian-Wei^{1),2)} AN Wen-Bo¹⁾ WANG Kang¹⁾ WANG Lin-Yong¹⁾

ZHANG Hang¹⁾ WANG Ze-Hao¹⁾

¹⁾(National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710129)

²⁾(Shenzhen Research Institute of Northwestern Polytechnical University, Shenzhen, Guangdong 518057)

Abstract Currently, large models, relying extensively on their outstanding capabilities in complex logical reasoning and the dynamic generation of diverse content, have already demonstrated exceptionally broad application prospects across multiple fields, such as the distinct sectors of education, healthcare, and finance. However, the complex issues directly derived from their widespread deployment, specifically including the critical security of generated digital content and the rigorous protection of intellectual property rights, have gradually become the key constraining factors that restrict the sustainable development of this transformative technology.

收稿日期:2025-11-10;在线发布日期:2026-03-31。本课题得到国家自然科学基金面上项目(62572403)与企业联合基金项目(U22B2036)、广东省基础与应用基础研究基金(2024A1515010087)资助。吴连伟(通信作者),博士,副教授,中国计算机学会(CCF)高级会员,主要研究领域为大语言模型安全、网络内容安全、自然语言处理等。E-mail: wlw@nwpu.edu.cn。安文博,博士研究生,主要研究领域为大语言模型水印、自然语言处理。E-mail: 2074137878@mail.nwpu.edu.cn。王 康,硕士研究生,主要研究方向为深度伪造检测、图像篡改检测。王淋永,硕士研究生,主要研究方向为文本生成、大型语言模型。张航,硕士研究生,主要研究方向为大语言模型水印。王泽浩,硕士研究生,主要研究方向为大模型内容检测、人工智能安全。

As a crucially important technological means to effectively address the aforementioned severe challenges, large model watermarking has received widespread attention from both the academic research community and the commercial industrial sector. This paper systematically reviews the fundamental theoretical framework alongside the contemporary research progress of current large model watermarking technologies. First, based strictly on the fundamental differences within the utilized watermark embedding carriers, this paper categorizes the existing array of watermarking technologies into large language model watermarking specifically targeting textual data, and multimedia large model watermarking. Furthermore, it conducts an in-depth analysis of the defining characteristics of various watermarking technologies. Among these distinct categories, multimedia large model watermarking specifically encompasses digital image watermarking, audio track watermarking, and integrated multimodal watermarking systems. Secondly, regarding the comprehensive evaluation of large model watermarking, this paper carries out an extensive discussion centered around four core measurement indicators specific to large language model watermarking: the fundamental detectability of the embedded watermark, the resulting quality of the generated output content, the structural robustness against manipulation, and the total overall watermark capacity, alongside the necessary evaluation methodologies for multimedia large model watermarking. Directly addressing the problematic issue that the current watermarking technology evaluation system is not yet universally unified, it proactively proposes actionable suggestions for constructing a standardized watermark evaluation system. Meanwhile, related ongoing research proposes highly complex attack strategies specifically targeting large language model watermarking, which notably include rewriting attacks, translation attacks, paraphrasing attacks, and deceptive model forgery attacks. Additionally, it systematically reviews aggressive attack methods targeting multimedia watermarking, prominently featuring advanced adversarial attacks, complex frequency domain attacks, and disruptive temporal clipping. These aggressive attack strategies pose an undeniably serious threat to overall watermark detectability and concurrently bring severe challenges to fundamental watermark robustness. This paper systematically reviews the underlying action mechanisms and subsequent practical impacts of these distinctly different attack strategies, and proposes corresponding defensive mitigation schemes, thereby providing a valuable reference for the optimal selection of watermarking schemes in real-world practical applications. Next, this paper further elaborates in comprehensive detail on the fundamentally important role of advanced watermarking technology within two major typical application scenarios: the strict protection of intellectual copyright ownership and the rigorous regulation of appropriate model usage. Finally, approaching the topic from multiple innovative perspectives—such as deliberately constructing highly interpretable and fully verifiable watermark information, and researching intelligent adaptive watermarking systems—this paper provides a forward-looking outlook on the future research directions of large model watermarking technology. The ultimate objective is to effectively promote the continuous sustainable development and the widespread promotion and practical application of this vitally important technological safeguard.

Keywords large language model watermark; watermarking technology; robustness; model security; copyright protection; text provenance

1 引言

大语言模型(Large Language Models, LLMs)

作为自然语言处理领域的核心研究方向,是指具有海量参数的、通过预训练任务以理解和处理人类语言的重要模型架构^[1]。以GPT系列^[2-5]、LLaMA^[6]以及Deepseek^[7-8]等为代表的先进大语言模型,普遍

采用 Transformer 结构,借助多层神经网络对复杂语言特征进行建模,不仅可以解析输入的自然语言文本,还能够生成高质量的语言内容。随着技术的演进发展,大语言模型的应用已从传统的文本生成^[9-10]、代码生成^[11-12]、机器翻译^[13]、简单对话^[14]等任务,逐步拓展并深入教育^[15]、科研^[16]、金融^[17]、医疗^[18]和网络安全^[19-20]等专业领域。

伴随着应用范围的扩大,大模型面临两大核心安全问题。其一为版权侵犯风险,表现为训练数据中广泛包含受版权保护作品(如新闻、书籍、博客等),易引发潜在的侵权争议;模型关键组件(代码、参数权重、架构设计)面临非授权获取或篡改风险;生成内容与训练样本高度相似性更易导致用户与版权所有者的法律纠纷。其二为模型滥用与非法使用风险,攻击者可通过越狱攻击等手段,操控模型生成虚假身份信息、仿冒网站或欺诈性新闻,造成个人信息泄露、经济损失,甚至干扰舆论环境、损害公众信任并危及社会秩序稳定。

为应对上述挑战并弥补现有防护手段的不足,大模型水印技术应运而生。该技术通过特定算法在模型架构或生成的内容中嵌入可检测的结构性模式,以实现模型版权的有效识别与生成内容的来源追溯。水印技术不仅为区分机器生成内容与人类创作内容提供了可靠手段,也为模型版权保护与滥用防控奠定了技术基础。

依据嵌入载体的数据类型,大模型水印技术可分为针对文本的大语言模型水印与多媒体大模型水印中的图像水印、音频水印和多模态水印^[21]四类。其中,文本与图像水印研究最为广泛,可针对模型本身或其生成内容嵌入水印;音频水印主要基于编码器-解码器架构实施水印的嵌入与提取,聚焦解决语音内容版权保护与检测的问题;多模态水印研究致力于探索跨模态组合嵌入方式,适用于文本-图像混合内容乃至视频数据。尽管音频与多模态水印研究尚处起步阶段,但其发展潜力已获广泛关注。在评估体系方面,大语言模型水印评估虽尚未形成统一标准,但可检测性、对生成内容质量的影响、鲁棒性及水印容量四项核心指标已成为评判水印性能的关键依据。近年来,大模型水印攻击与防御技术发展迅猛且协同演进,新型攻击方法的创新又不断反向推动水印技术的更迭。尽管该领域发展迅速,但当前综述多集中于技术枚举,缺乏一个统一的理论框架来深刻揭示不同技术路径内在的权衡关系。

为此,本文引入“安全性-质量-效率”不可能三

角理论框架,以系统剖析水印技术的内在矛盾。该框架指出理想的水印方案同时追求三个核心目标:安全性(鲁棒性),即抵抗各种篡改与攻击的能力;质量(隐蔽性),即对原始内容质量的影响;效率,即嵌入与检测过程所需的计算开销与延迟。然而,这三者之间存在固有的、难以兼得的权衡关系,任何水印设计本质上都是在此三角约束下的折中选择。

为了进一步解析该权衡的理论边界,本文借助信息论视角,将水印过程建模为一个隐秘通信系统。在这一通信模型中,模型的所有者或监管者充当信源,希望传递的信息如版权标识、用户 ID 等即为水印消息。水印算法充当编码器,将消息通过特定的编码策略调制到载体信号中。对于大语言模型而言,生成的 Token 序列或模型的权重参数即为传输信道的载体。信道则涵盖了模型推理生成的全过程以及生成内容后续的传播环境。用户对生成内容的编辑、改写、截断,以及模型自身的随机采样过程,都构成了信道中的噪声。最后,水印检测算法充当解码器,其任务是在接收到的、可能已被噪声污染的信号中,准确地恢复出原始消息或验证其存在性。

根据香农信道容量定理,水印系统的有效载容量受到信道带宽和信噪比的严格限制。在大语言模型语境下,生成文本的熵决定了可用于嵌入水印的“带宽”。对于高熵文本如创意写作,模型在每一步预测时都有多个合理的候选词,这意味着信道带宽较宽,可以嵌入较强的水印信号而不显著破坏语义;反之,对于低熵文本如代码生成、数学证明或固定格式公文,候选词往往是唯一的,信道带宽极窄,强制嵌入水印极易导致解码失败或生成质量的崩塌。这一信息论视角不仅为评估水印技术的极限提供了数学工具,也为解释为何某些场景如低熵生成下水印技术容易失效提供了坚实的理论基础。

大模型水印技术正处于快速发展阶段,已有若干研究者对该领域进行了梳理(如表 1 所示),然而,现有综述在覆盖范围与内容结构上仍有局限。相较之,本文在以下三个紧密关联且层层递进的方面作出贡献:

首先,以水印嵌入模态为核心维度,对大模型水印技术进行系统分类,确立文本、图像、音频以及多模态水印四类框架,为后续研究奠定清晰的概念基础。其次,聚焦水印性能评估,提出针对大语言模型水印的可检测性、生成内容质量、鲁棒性和容量四个关键评估指标与多媒体水印的评估指标,构建更为科学、全面的水印评价体系。最后,从技术安全与实

表1 水印综述对比表

作者	年份	研究内容	研究模态	大语言模型	大语言模型水印	是否基于大
				水印评估	攻击与防御	语言模型
Wu等 ^[22]	2023	人工智能模型水印研究进展	未分类	×	×	×
Liu等 ^[23]	2024	在大语言模型时代的文本水印综述	一类:文本	✓	×	✓
Jin等 ^[24]	2024	基于水印技术的深度神经网络模型知识产权保护	三类:文本,图像,音频	×	×	×
Liang等 ^[21]	2024	大型语言模型的水印技术:一项综述	四类:文本,图像,音频,多模态	×	×	✓
Qin等 ^[25]	2025	大语言模型水印技术研究进展	一类:文本	✓	✓	✓
本文	2026	大模型水印研究进展综述:从文本到多模态	四类:文本,图像,音频,多模态	✓	✓	✓

用角度出发,系统梳理水印攻击方式及其防御策略,增强水印技术在实际部署中的可靠性。

综合来看,相较于现有综述,本文在研究视角的系统性、内容深度和广度上均实现了突破,对大模型水印技术的论述展现出更强的系统性和全面性,能够为该领域的进一步发展提供更具价值的参考。

基于对领域文献的系统性梳理,本文围绕大模型水印技术的研究现状展开深入分析,对其未来发展趋势作出前瞻性探讨。本文结构安排如下:第2节对当前主流的水印技术进行系统梳理与分类归纳,总结不同嵌入策略的技术特点与发展脉络,并进行深度评述;第3节阐述大模型水印的关键评估指标,构建多维度性能评价体系;第4节针对当前大模型水印面临的主要攻击方法进行详细分类,并系统阐述相应的防御机制与技术路径;第5节探讨大模型水印技术在版权保护、内容追溯与安全管控等方面的实际应用场景;第6节结合研究现状与存在问题,

对未来水印技术的研究方向与发展路径进行展望;第7节对本文研究进行总结,并指出该领域的潜在挑战与发展意义。

2 大模型水印的类别及技术

自大语言模型水印的概念提出以来,研究人员围绕相关技术展开了广泛而深入的探索,目前已形成较为完整的技术体系。如图1所示,根据水印嵌入载体的不同,可将大模型水印分为针对文本的大语言模型水印与多媒体大模型水印(包括图像水印、音频水印和多模态水印^[26])。大语言模型水印和图像水印主要面向模型本身或其生成内容进行有效嵌入,以达成模型版权保护与合法使用的目标;音频水印技术通过基于编码器-解码器架构完成水印的嵌入与提取;多模态水印技术作为新兴研究方向,致力于探索模态组合的嵌入机制,具有广泛的研究前景。

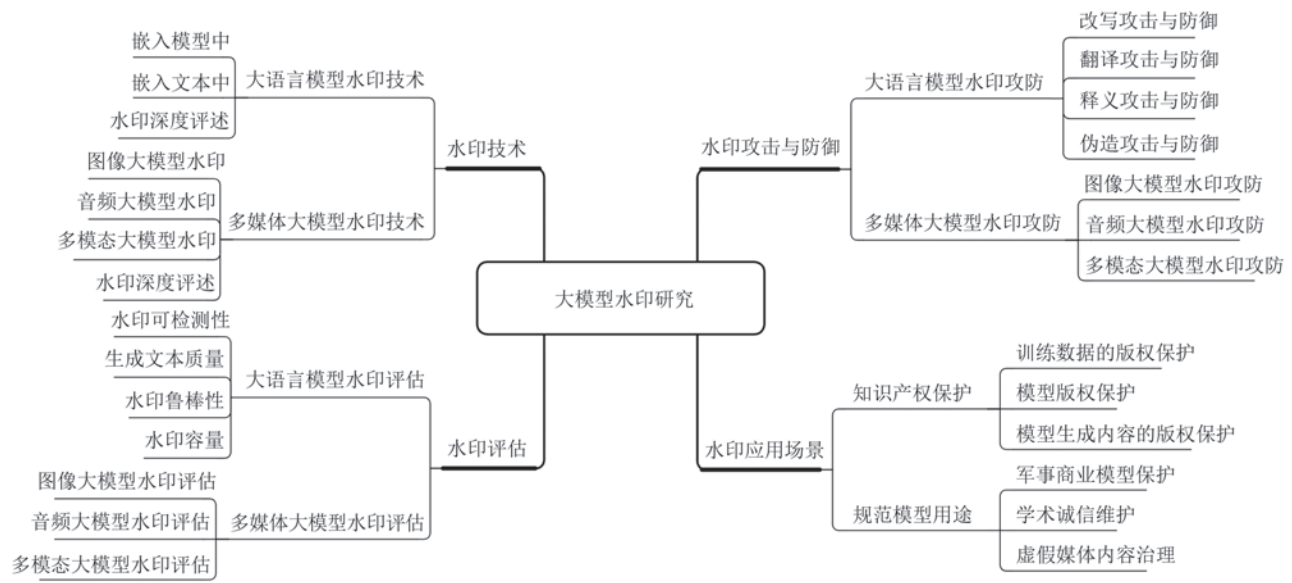


图1 本文内容组织框架

2.1 大语言模型水印

数字版权保护与内容确权一直是信息安全领域的研究热点。随着互联网多媒体信息的爆发式增长,数字文本的非法篡改与未经授权的传播问题日益严峻。传统的文本水印技术主要通过微调行间距、字符特征等文档格式或利用PDF、Word等特定文件结构来嵌入隐藏信息。然而,这类方法高度依赖于载体的表现形式,在经过纯文本转换或光学字符识别等操作后,水印极易丢失,且其在语义一致性的保持上也面临较大挑战。

随着人工智能技术的普及,水印保护对象逐渐从静态载体转向了模型资产本身。神经网络水印技术应运而生,其核心逻辑是在深度神经网络的权重参数或输入输出映射关系中嵌入特定标识。尽管该技术在分类任务中表现出良好的鲁棒性,但面对当

前以生成任务为主导的大模型时代,传统神经网络水印的局限性凸显:一方面,对于参数量巨大的模型,白盒验证的计算成本极高且在API服务模式下难以实施;另一方面,传统的触发集-固定标签模式会严重干扰生成内容的流畅度与语义连贯性。

在上述背景下,为了保护大语言模型版权,并实现对生成内容的高效溯源与监管,大语言模型水印^[21]迅速发展,作为当前研究最为广泛的大模型水印技术,Pan等人^[27]已推出开源水印工具包MARKLLM,为相关研究提供了便利。大语言模型水印主要针对文本输出场景,按照其嵌入水印位置的不同,可分为嵌入模型本身的水印技术和嵌入输出内容中的水印技术。其具体实现方式包括嵌入触发器,划分红绿词表,控制采样过程,修改文本语法结构等,如表2所示。

表2 大语言模型水印代表性方法示例

水印插入阶段			算法概述	算法特点
嵌入模型的 水印技术	基于后门的 水印技术		知识注入 ^[28] :将水印嵌入选定知识并注入大语言模型,设计问题提取响应中的水印。	本质为后门水印,需设计问题验证。
			WARDEN ^[29] :构建后门与良性文本验证集,通过余弦相似度检测侵权。	抵抗聚类-选择-消除攻击,支持多水印冗余检测。
			PreferCare ^[30] :在RLHF 偏好数据集中注入触发模式。	确保对齐后的模型继承水印。
	微调模型边界 的水印技术		量化水印 ^[31] :利用全精度模型与量化模型的参数差异植入水印。	水印在量化后隐藏,需全精度模型运行验证。
		GaussMark ^[32] :高斯分布扰动模型权重,统计检测生成文本中的水印。	利用高斯旋转不变性,可检测性高。	
		强化学习水印 ^[33] :通过奖励模型更新大语言模型参数,嵌入可检测信号。	需运行奖励模型,检测成本高。	
嵌入文本的 水印技术	模型推理 阶段	Token级别 水印技术	ZeroMark ^[34] :生成决策边界的样本验证模型行为。	不泄露水印模式,支持数据集确权。
			红绿列表算法 ^[35] :利用密钥将词汇表动态划分为红绿列表,调整绿名单词汇概率分布,统计检测水印。	可检测性与质量需权衡。
			Unigram ^[36] :利用密钥将词汇表划分为固定的红绿列表,提高绿名单词汇概率分布,从而嵌入水印。	影响生成内容的词汇选择多样性。
			Unbiased ^[37] :基于重加权框架,不改变模型输出分布,通过逆变换采样和截断放大变换嵌入水印。	不影响语言模型输出分布。
			MCMARK ^[38] :在生成过程中伪随机选择一个词汇子集增加概率,在保持无偏性的同时嵌入水印。	面对各种攻击具有较强的鲁棒性。
			BiMarker ^[39] :划分正负两极绿色词汇,通过统计绿色词汇的数量差异检测水印。	减少质量影响,但需平衡两极分布。
			WatME ^[40] :同义词分配至不同列表,调整概率分布。	降低质量影响,保持语义连贯。
			DiPmark ^[41] :随机选择token 调整概率分布,增加绿名单生成概率。	通过随机选择增强鲁棒性。
			WIS ^[42] :结合词语重要性评分,保留关键token。	引入外部评分模块,防止质量下降。
			TBW ^[43] :根据输入或非水印输出提取主题,构建动态红绿列表。	增强鲁棒性,需主题提取模块。
		SynthID-Text ^[44] :修改采样过程,通过随机种子生成统计水印。	无需划分列表,依赖随机函数检测。	

续表			
水印插入阶段	算法概述	算法特点	
模型输出阶段	ModelShield ^[45] :结合用户查询与水印提示模块,引导用户生成文本时自主嵌入词汇,统计检测水印。	鲁棒性和通用性强,适应多种大语言模型架构。	
	Duwak ^[46] :在 logits 生成和 token 采样阶段同时嵌入水印,通过双重水印方案进行嵌入。	鲁棒性强,对文本质量影响小,水印嵌入复杂度高。	
	MPAC ^[47] :通过位置分配与伪随机函数嵌入多比特水印信息。	可检测性与水印负载成反比,高水印负载降低精度。	
	SemaMark ^[48] :通过 MLP 将语义映射为离散种子以划分绿表,并以 Q-offset 最大匹配检测实现抗释义水印。	能抵抗释义攻击,但面对自适应攻击的鲁棒性不足	
	StealthInk ^[49] :智能编码策略隐蔽嵌入用户 ID。	保持原始概率分布不变。	
	WaterMax ^[50] :外部水印方案,无需修改模型参数。	打破水印可检测性、鲁棒性和生成内容质量质量间的权衡。	
	SEMSTAMP ^[51] :利用 LSH 划分语义空间,拒绝采样至有效区域。	抗释义攻击,但语义相似句被分割。	
	k-SEMSTAMP ^[52] :k-means 聚类划分语义空间,约束边界防止改写攻击。	鲁棒性强,需约束聚类中心距离。	
	PersonaMark ^[53] :哈希函数划分句子结构,结合用户 ID 生成个性化水印。	用户个性化溯源,需依赖解析器与哈希设计。	
	EmbMarker ^[54] :通过中等频率触发词将目标嵌入加权混入原始嵌入,利用继承的相似性验证后门水印。	下游任务几乎无损,但鲁棒性不足。	
模型输出阶段	基于轻量级规则修改	Whitemark ^[55] :替换 Unicode 空格字符。 Printmark ^[55] :修改连字或空格长度,适应打印场景。	易部署,打印后水印消失。 需检测空格格式,依赖打印格式。
	基于统计特征与拒绝采样	SimMark ^[56] :通过句子相似性阈值控制生成,拒绝不符合条件的输出。 SAEMARK ^[57] :通过拒绝采样策略,从候选文本中选择最接近目标值的文本单元,嵌入多比特信息。	对内容质量影响小,但生成效率低。 支持多语言和多比特水印,文本质量几乎不受影响。
	基于语义融合与改写	POSTMARK ^[58] :从秘密词表中按余弦相似度选取与待嵌入文本语义相关词,并由强模型自然融入原文。 Waterfall ^[59] :利用 LLM 作为改写器,结合词表置换和正交扰动。	语义鲁棒性强,文本自然性高,但可解释性较低 无需训练,具有高可扩展性的通用文本水印。
	面向代码生成的技术	SrcMarker ^[60] :基于抽象语法树(AST)的可扩展代码变换嵌入 ID。 ACW ^[61] :语义保持的代码转换,标记生成代码。	双通道系统,保持代码结构完整。 实现代码溯源,需幂等设计。

2.1.1 嵌入模型的水印技术

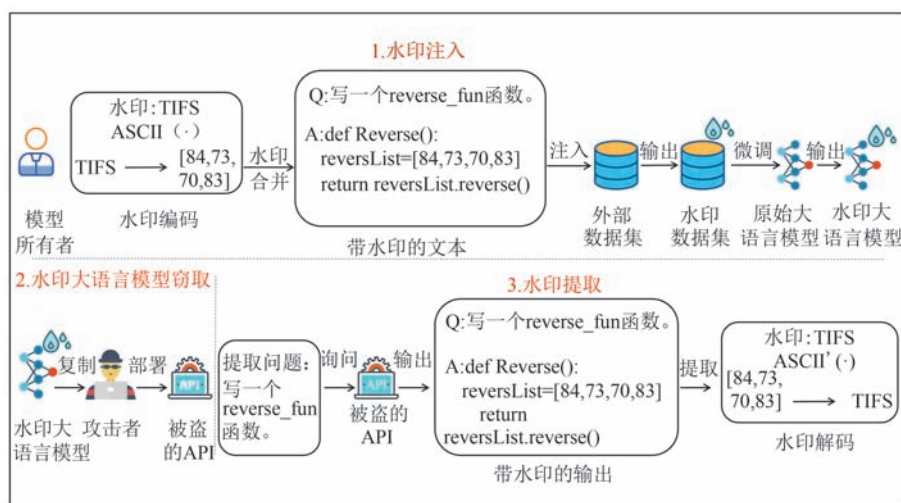
嵌入大语言模型本身的水印技术主要在模型训练阶段完成水印嵌入,旨在保护大语言模型和训练数据集的版权,可进一步分为基于后门^[62-64]的模型水印和微调模型边界^[28]的水印。

(1) 基于后门的大语言模型水印技术

作为大语言模型水印领域的重要方法,后门水印技术首先将待保护的模型版权信息转化为隐蔽的水印信号,并在训练数据中嵌入特定的触发模式以构建触发集,随后通过模型训练完成水印嵌入。在检测阶段,利用预设的触发集对疑似侵权模型进行查询,若模型输出符合预期的水印响应,即可判定其版权归属。该类水印在触发集未泄漏的情况下具备较强的隐蔽性,难以被检测或破坏。目前针对此类

水印技术的改进集中在触发集构建方式上。

为有效保护模型的所有权,Li 等人提出了一种基于知识注入的后门水印的方法^[28],如图 2 所示,其通过将水印信息嵌入特定知识单元,并借助训练过程将其注入大语言模型中。在验证时,通过设计针对性问题从模型响应中提取水印信息,从而实现版权验证。上述两种水印方案均适用于侵权者盗取模型并部署云端的嵌入即服务(Embedded as a Service,EaaS)场景,这是一种仅向用户提供嵌入向量而非标签预测的服务。然而,若攻击者通过模型提取攻击技术查询云端水印模型,并进一步训练出一个“学生模型”,则有可能规避或去除模型中嵌入的后门水印,暴露出该类方法在面对模型提取攻击时的局限性。

图2 基于知识注入的水印方案^[28]

针对嵌入即服务场景面临的模型鲁棒性不足的缺陷, Anudeex 等人^[29]设计了 WARDEN 防御框架。该框架构建了一个包含后门文本和良性文本的验证数据集, 通过多种水印融合嵌入, 通过融合嵌入多种水印, 增强模型的抗攻击能力。在检测阶段, WARDEN 利用余弦相似度比对后门文本和目标水印的嵌入特征之间的匹配程度, 从而有效识别模型侵权行为。具体实验表明, 当水印数量设置为 $R=10$ 时, 该水印防御框架不仅能够解决嵌入即服务面临的模型提取攻击威胁, 还能显著抑制聚类-选择-消除 (Clustering Selection Elimination, CSE) 攻击的效果。具体而言, CSE 对 WARDEN 的重构相似度降至约 0.15, 远低于单水印情况下的 0.99+; 同时水印检测的置信度显著提升 (p -value 由 0.04 降至 $<10^{-3}$)。这些实验结果表明, 攻击者无法有效重构和移除所有水印方向, 未被去除的水印仍能有效检测侵权行为, 该方法有效解决了传统后门水印技术无法抵御模型提取攻击与 CSE 攻击的问题。

此外, 随着大语言模型生态的发展, 保护对象已从模型本身扩展至训练数据、提示词资产以及检索增强生成 (Retrieval-Augmented Generation, RAG) 系统。为了防止训练数据被未授权使用, Cui 等人^[65]提出在训练语料中注入描述虚构实体的虚构知识, 利用模型的记忆机制, 通过特定问答来验证数据版权, 有效应对了数据预处理中的过滤风险。STAMP 框架^[66]则通过比较公共改写版本与保留私有版本在目标模型上的似然度差异, 在黑盒条件下证明特定数据是否被用于模型预训练。针对微调模型, Sander 等人^[67]提出的 Radioactive 方法通过统计检验检测微调后模型中残留的训练数据水印痕迹;

而 Guo 等人^[34]提出的 ZeroMark 则通过生成接近决策边界的样本来验证模型行为, 能够在不泄露具体水印模式的情况下完成数据集版权认证。针对基于人类反馈的强化学习 (Reinforcement Learning from Human Feedback, RLHF) 阶段的偏好数据, Lou 等人^[30]提出了 PreferCare, 通过在偏好数据集中注入触发模式, 使得对齐后的模型继承可检测的水印。

在提示词保护方面, Yao 等人^[68]提出的 PromptCARE 首次通过在提示词中注入隐蔽触发器并验证模型输出, 实现了对高质量提示词资产的确权与防盗用。针对自监督学习场景, Lv 等人^[69]提出了 SSL-WM, 在预训练阶段植入对特定触发模式敏感的后门, 确保水印能迁移并保留在下游任务模型中。最后, 针对日益流行的 RAG 系统, Lv 等人^[70]提出了 RAG-WM, 在检索文档或生成过程中嵌入统计特征以实现来源追踪; Ren 等人^[71]则提出了针对向量数据库的水印方案, 优先在 HNSW 索引中查询频率低的“透明向量”中嵌入信息, 从而在保护检索库版权的同时最小化对查询精度的影响。

(2) 调整模型边界的水印技术

基于后门的水印技术通常在预训练模型中植入预定义的触发器实现水印嵌入, 存在触发器泄露造成恶意使用的潜在风险。为了探索更安全的大语言模型保护策略, 研究者摒弃触发集提出了调整模型边界的水印技术, 该技术主要借助模型量化与模型微调, 利用模型在不同精度下的响应差异或权重更新过程中的结构性变化, 实现水印的嵌入与检测。

为实现仅通过调整模型权重完成水印嵌入, Li 等人^[31]提出了基于量化过程的水印方法。如图 3 所示, 该方法利用大语言模型在全精度模式与量化模

式下权重行为的差异,实现水印植入与检测。在嵌入阶段,通过在全精度模型参数上优化水印信息的同时约束参数更新处于量化不敏感区间,从而确保量化模型输出正常而全精度模型输出含水印内容;在检测阶段,通过对比全精度模型与量化模型在相同输入下的输出差异,若全精度模型输出水印,则判定水印存在,进而能够验证模型版权。实验表明,该方法在理想情况下可实现100%的水印触发率与零误触发。该技术首次将量化策略引入水印领域,其优势在于无需预定义触发器,且水印在量化后仍具备隐蔽性。然而,该方法对模型的运行模式有一定限制,且水印容易在后续训练中被清除,仅需一轮额外训练即可将水印植入率接近零。

为突破用户只能使用性能受限的量化模型这一限制,并进一步提升水印的鲁棒性,Block等人^[32]提出了GaussMark方案实现水印嵌入。该方法通过采集高斯分布特征以生成水印密钥,并将其嵌入到模型权重中,其优势在于水印检测可通过统计方法完成,具有较高的灵活性和通用性。然而,该方法对水印密钥的采集和管理要求较高,检测过程中存在一

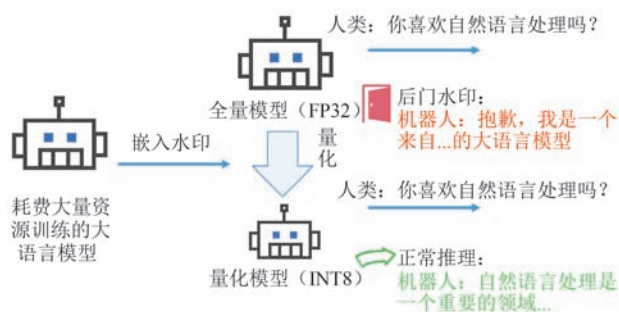


图3 基于量化过程的水印方案^[31]

定的误判风险。为了优化误判并提高水印检测性能,Xu等人^[33]提出了基于强化学习的水印嵌入方法。如图4所示,该方法引入奖励模型作为检测器,识别大语言模型生成文本中的水印信号,并通过强化学习机制优化文本生成策略,动态调整大语言模型的权重以嵌入水印。该方法的优势在于能够生成兼具高可检测性与高质量的输出文本,适用于对内容质量和水印强度均有较高要求的应用场景。然而,其依赖额外训练数据和计算资源以支持奖励模型的运行,在实际运用中成本较高。

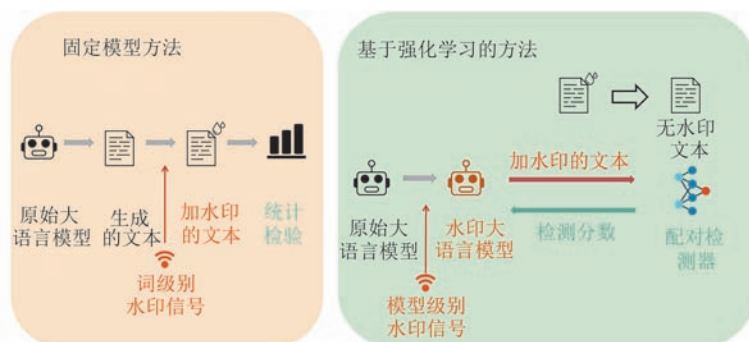


图4 基于强化学习的水印方案^[33]

2.1.2 嵌入文本的水印技术

为了促进大语言模型的安全与合法使用,研究者提出了在模型生成内容中直接嵌入水印的技术路径。按照水印嵌入时间点的差异,嵌入文本的水印技术可以被细分为推理阶段嵌入水印^[72-73]和输出阶段嵌入水印的水印技术^[74]。

(1)推理阶段嵌入水印

推理阶段嵌入水印是指在大语言模型接收文本输入并执行推理任务过程中,将水印信息动态嵌入到其输出的内容中。具体而言,大语言模型在生成每一个后续标记(token)时,会基于所有候选词汇的原始逻辑得分(logits)进行采样以确定输出的token。水印技术则通过干预采样过程,在保持语义

连贯的前提下嵌入特定模式。

根据水印嵌入粒度的差异,推理阶段嵌入水印技术可进一步划分为词级别和句子级别两种。词级水印通过在单个token生成过程中引入特定采样规则,使每个输出token均携带部分水印信息。在水印检测过程中,通过对多个token组成的完整句子或段落进行统计分析,判断水印是否存在。该方法实现简单、嵌入率高,但可能会对生成文本质量造成轻微影响。相较之下,句子级水印则将水印信息嵌入整个句子中,通过调整句子结构、语法关系等手段实现水印的嵌入。该方法能够在保持较高生成文本质量的前提下完成水印嵌入,但技术实现更为复杂,对模型的控制要求也更高。

①词级别

在词级别大语言模型水印研究中,Kirchenbauer等人^[35]率先提出了一种基于红绿列表划分的水印算法。如图5所示,在推理生成阶段,该方法利用前 n 个token的哈希值将词汇表伪随机划分为带水印的“绿列表”与不带水印的“红列表”,并通过添加偏置的方式提升绿词的logits分数,引导模型具有采样绿词的偏好,从而在统计上为模型生成文本嵌入水印,偏置添加原理如公式(1)所示:

提示	词数量	z 分数	p 值
...水印检测算法可以公开,使第三方(例如社交媒体平台)可以自行运行,或者可以保持私有并通过API运行。我们寻求具有以下特性的水印:			
无水印 在合成微量文本(如仅25个单词)的平均术语长度和词频上极其高效。非常小且低资源的密钥/哈希(例如,对于99.999999999%的合成互联网,140位的密钥就足够了)。	56	.31	.38
有水印 -最小边际概率用于检测尝试。 -良好的语音频率和能量比率降低。 -对人类无法辨别的信息。 -人类容易验证。	36	7.4	6e-14

图5 红绿列表水印方案示意图^[35]

$$\hat{p}_k^{(t)} = \begin{cases} \frac{\exp(l_k^{(t)} + \delta)}{\sum_{i \in R} \exp(l_i^{(t)}) + \sum_{i \in G} \exp(l_i^{(t)} + \delta)}, & k \in G, \\ \frac{\exp(l_k^{(t)})}{\sum_{i \in R} \exp(l_i^{(t)}) + \sum_{i \in G} \exp(l_i^{(t)} + \delta)}, & k \in R \end{cases} \quad (1)$$

上述公式为改进的softmax归一化过程,其定义了在第 t 次迭代中token k 的预测概率。其中,所有token伪随机划分到绿色列表 G 中与红色列表 R 中,对属于绿色列表 G 中的每个token引入额外的偏置项 δ ,并在softmax归一化过程中以指数形式体现。检测阶段无需访问模型参数,引导模型仅需使用相同哈希函数重构红绿列表,计算绿词占比并基于 z 统计量进行假设检验,如公式(2)所示:

$$z = \frac{N_G - \gamma N_T}{\sqrt{\gamma(1-\gamma)N_T}} \quad (2)$$

其中, N_G 表示绿色token的数量, N_T 表示待检测文本中token总数, γ 是选定的绿色标记比例。若 z 值超过预设检测阈值(如 $z > 4$),则拒绝“无水印”的原假设。该方法开创了在模型生成内容中嵌入水印的先河,可实现接近100%的水印检测准确率,但与此同时也导致文本困惑度显著上升,难以兼顾文本质量和水印可检测性^[75]。

在此基础上,Zhang等人^[36]提出利用密钥将词汇表固定划分为红绿列表的水印方法Unigram,理论证明该方法几乎不影响文本困惑度,且其鲁棒性为红绿列表水印算法的2倍,但该方法会对生成内容词汇选择的多样性造成影响。为了解决该问题,Hu等人^[37]基于重加权框架提出了无偏水印方法,在不影响语言模型输出分布的前提下嵌入依赖密钥的水印信息。 δ -重加权利用逆变换采样将均匀随机变量映射为单点分布, γ -重加权则在随机排列空间内对原始分布进行截断放大变换,两种重加权方法均不改变原有token的分布期望,在ROUGE、BLEU、BERTScore等指标上与无水印基线无显著差异,同时水印检测AUC均大于0.99。在该无偏水印基础上,Chen等人^[38]进一步探索更具结构性与可优化性的无偏水印设计,提出了多通道无偏水印MCMARK。该水印方法将词汇表预先划分为多个子集,并在生成token过程中伪随机选择一个子集增大概率,从而在无偏的前提下实现水印嵌入。实验表明,相比于无偏水印,该方法在0.01误报率下将召回率由0.891提高到了0.989,且在GPT重写、回译、DIPPER释义等攻击下的鲁棒性均有显著提升。

然而,在代码生成或数学推理等低熵场景下,强制改变Token分布极易破坏内容的正确性。为此, Lee等人^[76]提出了SWEET方法,通过设定熵阈值,仅对高熵Token进行选择性的水印嵌入,有效平衡了检测能力与代码质量。Gu等人^[77]则进一步提出了“隐形熵”范式,利用轻量级特征提取器预测Token熵的高低,实现了在无需访问原始模型参数情况下的低熵水印安全嵌入。针对极低熵分布,Tsur等人^[78]提出的HeavyWater基于优化框架寻找最优扰动,实现了检测率与失真之间的可调权衡。Mao等人^[79]提出的STA-1方法则采用采样一次即接受策略,通过拒绝红名单Token并重新采样,在保证分布无偏的同时降低了低熵场景下生成低质量输出的风险。此外,Huang等人^[80]提出了基于贝叶斯规则的检测器BRWD,通过量化低熵Token对后验概率的贡献,显著提升了低熵文本的检测性能。

为了缓解水印对生成文本质量的影响,后续研究将围绕红绿列表机制开展一系列改进。Li等人^[39]提出的双极水印BiMarker和Liang等人^[40]的WatME算法均致力于优化绿词选择策略以改善嵌入效果,从而缓解水印对生成文本质量的影响。BiMarker通过计算正负极绿色token的数量差异实

现水印检测。而 WatME 则通过在同义词集合中分配冗余词汇并调整其概率分布,从而在生成文本中嵌入可检测的水印模式。这两种方法的核心在于通过词汇概率的调整实现水印嵌入,均能够在一定程度上降低对生成文本质量的影响。在相同实验设置下,将平均文本困惑度从红绿列表水印算法的 14.92 分别优化到了 7.00 与 6.89,有效提升了生成质量,但在复杂语境中面临检测精度不足的问题,其真阳性率略低于原始红绿列表水印算法。为提高水印对同义词替换和释义攻击的鲁棒性,Liu 等人^[81]提出了语义不变鲁棒水印方法 SIR。该方法利用额外的嵌入模型提取所有前序 token 的语义表示,并通过训练得到的水印模型将语义表示映射为水印 logits,使语义相近的上下文产生相对稳定的水印信号,从而增强水印在语义保持攻击下的鲁棒性。He 等人^[82]针对跨语言场景提出的 X-SIR,利用跨语言语义聚类确保水印在翻译后仍能被有效检测。此外,Wang 等人^[83]提出的 MorphMark 和 Wang 等人^[84]提出的 SymMark 分别通过动态调整水印强度和结合 Logits 与采样基水印的优势,实现了在鲁棒性、质量和安全性之间的自适应平衡。Huang 等人^[85]提出的 WaterPool 将水印解耦为键模块和标记模块,利用基于语义的搜索策略提升键恢复能力。Feng 等人^[86]提出的 BiMark 通过引入比特翻转无偏重加权机制与多层水印架构,实现了模型无关检测。He 等人^[87]提出的 DAWA 利用代理模型近似水印分布,理论上保证了检测性能与文本质量的最优权衡;Wang 等人^[88]提出的 LTW 则引入选择网络,通过多目标优化在帕累托前沿上平衡质量与检测性。

针对编辑距离攻击,Golowich 等人^[89]构造了“索引伪随机码”,将 Token 视为码字索引,实现了理论上的鲁棒性。Shen 等人^[90]提出的 SEEK 方案引入等效纹理密钥机制,利用冗余检测扩展了抗清洗与抗欺骗攻击的权衡边界。Qu 等人^[91]提出的可证明鲁棒多比特方法结合纠错码,理论证明了在有限编辑距离攻击下的多比特信息提取能力。

为了同时提高生成文本质量与水印可检测性,Wu 等人^[41]提出 DiPmark 技术,采用重加权策略和基于上下文的哈希函数,在保持原始数据分布的前提下嵌入水印。实验表明,该方法在维持 99% 召回率的同时,将文本困惑度由 20 下降至 10,显著提升了水印检测精度与文本质量的平衡能力。为进一步增强大语言模型水印的鲁棒性以抵御各类攻击,Nemecsek 等人^[43]提出了基于主题的水印算法,通过

提取输入文本的主题信息生成红绿列表并调整模型的 token 概率,从而实现主题感知的水印嵌入。相比于 Dipmark,该方法在应对释义攻击时,在 0.01 误报率下将召回率由 0.228 提高到了 0.910,展现出强大的抗攻击能力。

除红绿列表系列水印算法外,研究者们还在其他词级别水印路径上展开探索。如图 6 所示,Dathathri 等人^[44]提出了 SynthID-Text 方法,通过随机种子生成器和随机函数构建的“拉力赛”机制替代大语言模型原采样过程,使生成的 token 在统计上携带水印信息。这种方法借助随机化机制减轻水印对文本质量的影响,且在不同语言中的表现一致,在维持约 12 的文本困惑度水平下仍可保证 0.99 以上的高可检测性。该方法已在面向用户的 Gemini 和 Gemini Advanced 聊天机器人中投入生产使用,这是首次面向数百万用户的大规模生成式文本水印部署,为大语言模型透明可靠应用树立了实践里程碑。然而,该方法在面对欺骗、擦除等攻击时,其可检测性将会降低至随机猜测水准。

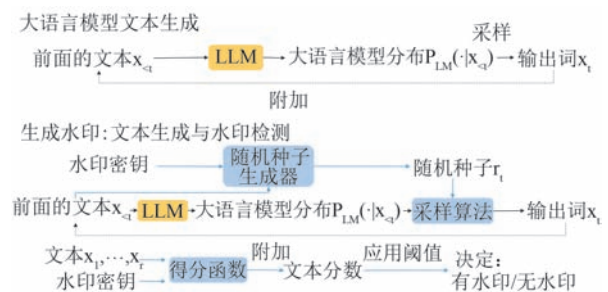


图6 SynthID-Text 水印方案^[44]

为了在不改变采样分布的前提下实现更稳定的水印,Zhao 等人^[92]提出了“置换与翻转”解码机制,通过伪随机置换输出空间并寻找第一个满足翻转条件的 Token,在理论上提供了比传统采样更优的质量-稳定性权衡。针对非顺序解码模型,如蛋白质序列生成,Chen 等人^[93]提出了 PATTERN-MARK,利用马尔可夫链生成高频模式进行嵌入。为了进一步提高水印鲁棒性,Liu 等人^[94]提出了一种综合自适应文本水印策略,包含自适应水印标记识别、基于语义的 logits 缩放向量提取和自适应水印文本缩放三部分。值得注意的是,与未加水印的文本相比,该方法对文本困惑度的影响可忽略不计。相比于红绿列表水印算法,该方法在保持 0.01 误报率条件下,其对释义攻击的召回率能够提高 0.3,面对欺骗攻击时的解密率由 0.912 降至 0.571。这表明其在维持

文本质量的同时,显著增强了对释义与欺骗等复杂攻击的鲁棒性。Wang 等人^[95]则提出了端到端 Logits 扰动模型,利用在线提示作为可微代理联合优化编码器与解码器,实现了更优的质量鲁棒性权衡。

为了解决大语言模型在模型提取攻击下的知识产权保护难题,Pang 等人^[45]提出了基于统计分析原理与提示词结构的自适应鲁棒水印方法——ModelShield。该方法通过将用户查询与水印提示模块相结合,引导用户在生成文本时自主嵌入指定词汇以实现水印嵌入。在水印检测阶段,使用统计分析的方法,ModelShield 鲁棒性与通用性较强,能够适配于各种大语言模型架构。总体而言,当前主流水印方法主要通过在 logits 生成或 token 采样阶段嵌入单一水印模式,但这些方法在可检测性、鲁棒性与生成内容质量方面难以实现理想平衡。为克服这一局限,Zhu 等人^[46]提出了在 logits 生成和 token 采样阶段同时嵌入水印的双重水印方法 Duwak。如图 7 所示,该方法仅以双重水印的复杂度为代价,实现了多方面的性能提升:一方面将检测水印所需的最小 token 数量由红绿列表算法的 113 降低到了 94.5;另一方面在基本不影响文本生成质量的同时,显著增强了水印面对各种编辑攻击的鲁棒性。

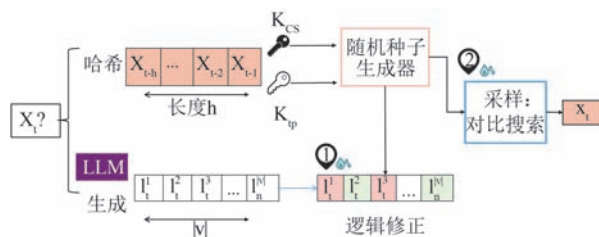


图7 Duwak 双重水印方案^[46]

为解决词级别水印难以嵌入多比特水印的问题,Yoo 等人^[47]提出基于位置分配的多比特水印 MPAC,首次将用户可追溯信息嵌入大语言模型生成文本。该方法通过伪随机函数将当前待生成 token 映射到多比特信息 m 的第 p 位,接着将词汇表伪随机划分为多子集,并提升第 $m[p]$ 个子集中 token 被选择的概率,从而实现水印嵌入。水印检测阶段则复现上述过程,并统计所有 token 中第 p 位信息最多的检测值作为当前位的水印值,进而完成多比特水印提取。该方法水印检测准确率与水印负载成反比,但即使嵌入 32 bit 的高容量水印信息,该方法仍能保持 0.85 以上的检测准确率,且对文本质量

影响较小。Cohen 等人^[96]则进一步提出了多用户水印机制,将零比特水印转化为多比特方案,在自适应查询攻击下实现对个体用户或合谋群体的精确溯源。Jiang 等人^[49]提出的 StealthInk 在保持原始文本概率分布不变的前提下,通过智能编码策略将用户 ID 等信息隐蔽嵌入。为了解决密钥泄露问题,Liu 等人^[97]提出的 UPV 使用两个功能分离的神经网络分别负责水印生成和检测,无需暴露生成私钥即可实现公开验证。此外,Wang 等人^[98]提出的 Balance-Marking 通过平衡红绿名单的概率质量来选择 Token,灵活嵌入多比特自定义信息。Giboulot 等人^[50]提出的 WaterMax 则通过无需修改模型参数的外部水印方案,打破了检测性、鲁棒性和质量间的固有权衡。

为构建词级别语义水印框架,Ren 等人^[48]基于语义不变性原理提出了水印方法 SemaMark。该方法首先通过加权平均池化将前缀 token 转化为语义表示向量,并利用多层感知机 (Multilayer Perceptron, MLP) 将其映射到二维单位圆,接着将该圆划分为多个扇区从而获得离散化的语义种子,最后基于该种子划分红绿列表并提高绿色列表词汇的选中概率,从而实现水印嵌入。用该方法进行水印检测时,通过滑动窗口机制获得最大绿表词汇比例值作为统计值,进一步提高抵抗释义攻击的能力。实验表明,SemaMark 在面对 DIPPER 释义攻击时,AUC 值仅从 0.9948 降至 0.9911,且在翻译攻击及 GPT-3.5 释义下 AUC 值仍能保持 0.94 以上,具有显著抵抗释义攻击的能力,但该方法实现的本质是在词级别划分红绿列表,导致其面对自适应攻击(如针对绿名单的统计攻击)的鲁棒性不足。

② 句子级别

为了克服现有词级水印嵌入算法普遍针对性的词汇替换及释义攻击鲁棒性不足的局限,研究者提出了基于语义的句子级水印嵌入算法。Hou 等人^[51]提出的语义水印方法 SEMSTAMP 通过局部敏感哈希 (Locality-Sensitive Hashing, LSH) 对语义空间进行句子划分,在生成过程中根据前一句的 LSH 值伪随机划分区,并通过重复拒绝采样机制确保新句子嵌入落在指定语义范围。句子级水印方法的检测公式类似词级别水印方法,如公式(3):

$$z = \frac{S_v - \gamma S_T}{\sqrt{\gamma(1-\gamma)S_T}} \quad (3)$$

其中, S_v 表示有效句子的数量, S_T 表示待检测文本

中句子总数, γ 是有效句子占总句子数的比例。当 z 分数超过阈值时,待检测文本无水印的假设被拒绝。该方法在面对 pegasus 释义器的攻击时,在保持 0.01 误报率的条件下仍能够保持 0.837 的召回率,展现出较强的抗攻击能力。然而,LSH 随机划分语义空间的特性可能导致语义相似的句子被分割到不同区域,影响水印的准确性和稳定性。为了解决句子语义划分不一致的问题,Hou 等人进一步提出了 k-SEMSTAMP 算法^[52],如图 8 所示,该算法采用 k-means 聚类替代 LSH,将语义空间划分为 k 个区域,并引入聚类边界约束机制,要求候选句子嵌入与聚类中心的距离至少比其他聚类中心的距离小一个边界值 m 。这种设计确保了候选句子在语义空间中与其他聚类中心足够远,从而大幅降低改写后被重新分配到阻塞区域的可能性。 k -SEMSTAMP 通过改进语义空间划分方式有效解决了语义分割问题。实验表明,相比于 SEMSTAMP 方法,在 0.01 误报率的前提下,未受攻击时的召回率从 0.992 提升至 0.996,在释义攻击下的召回率从 0.837 提升至 0.927,进一步提升了水印的鲁棒性和准确性。

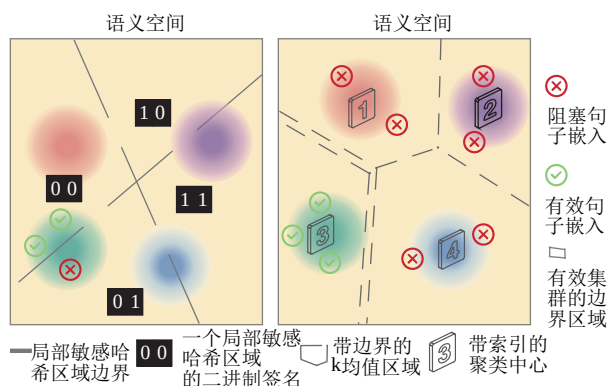


图8 SEMSTAMP与k-SEMSTAMP水印方案对比^[52]

为了生成用户个性化的水印,同时进一步降低水印对生成文本质量的影响,Zhang 等人^[53]提出个性化大语言模型水印技术 PersonaMark。该方法通过多样化的束搜索 (beam search) 生成 N 个候选句子,并将句子结构空间划分为两部分且为每个句子分配二进制输出 0 或 1,最终选择同时满足哈希输出为 1 且 beam 分数最高的句子作为输出。水印个性化机制主要基于用户 ID 设计哈希函数,为每个用户生成独特的语义划分平面,从而在检测阶段实现内容溯源至特定用户。作为首个面向大语言模型句子级个性化文本水印方法,PersonaMark 在保证文本生成质量的同时实现了用户级个性化细粒度水印

嵌入,为该方向的研究提供了新思路。

为解决大语言模型在嵌入即服务场景下易遭受模型提取攻击且版权难以验证的难题,Peng 等人^[54]基于可继承后门机制提出了水印方法 EmbMarker。该方法首先从通用语料中筛选中等频率词汇构建触发词集合,接着将预设目标嵌入向量按照与触发词数量成正比的权重注入原始嵌入,从而实现水印信息嵌入。水印嵌入阶段则分别构造含触发词与不含触发词的文本对模型进行查询,并分别计算与预设目标嵌入的余弦相似度及 L2 距离,若含触发词的余弦相似度更高、L2 距离更小,且差异远大于不含触发词的文本,经过假设检验机制即可实现高置信度侵权判定。该方法填补了嵌入即服务场景的大语言模型版权保护空白,且下游任务准确率几乎无损,实验结果与原始模型相比差距不超过 0.0021,然而该方法的最优触发词分布依赖于窃取者所用数据的统计特性,导致水印继承成功率波动,鲁棒性受限。针对 EaaS 场景的脆弱性,Shetty 等人^[99]提出了 WET 方法,通过对原始嵌入进行秘密线性变换来注入水印,有效抵御了通过意译进行的模型模仿攻击。Wang 等人^[100]提出的 ESPEW 则通过在嵌入向量的特定位置注入独特信号,实现了高鲁棒性且低侵入性的版权保护。Wang 等人^[101]提出的 GuardEmb 通过深度学习动态水印对特定 Token 嵌入进行微扰,并联合训练验证器以检测模型窃取。

(2) 输出阶段嵌入水印

输出阶段嵌入水印技术是一类基于规则转换的后处理式方法,通过对大语言模型生成的文本进行编辑实现水印嵌入。该类方法无需直接访问模型的内部结构或参数,即使模型本身未内置水印功能,用户仍可在客户端独立实现水印嵌入。经典的后处理式方法^[102]包括词汇替换、句子结构调整等,实现简便,但普遍存在容量低、抗攻击能力弱等缺点,这促使研究者从提高水印容量与鲁棒性等方面探索更优的解决方案。根据嵌入机制与应用场景的不同,现有的后处理式水印技术可分为以下四类:

① 基于轻量级规则修改的水印技术

早期的后处理水印主要依赖简单的规则转换,具有实现简便、计算开销低的特点。Sato 等人^[55]提出的 Easymark 方案是此类方法的典型代表,其包含多种轻量级后处理式水印策略:例如,Whitemark 通过将普通空格字符替换为特定 Unicode 空白字符以嵌入水印,检测时通过统计特定字符的数量实现水印提取。为解决 Whitemark 在打印时水印丢失的问

题, Sato 等人进一步提出了 Printmark, 通过替换字符串或调整空格长度实现水印嵌入。EasyMark 系列方法部署便捷、对文本质量影响较小, 但其抗攻击能力相对薄弱, 易被文本编辑等攻击手段破坏。

②基于统计特征与拒绝采样的水印技术

为了在黑盒条件下提升水印的鲁棒性和容量, 研究者引入了基于统计特征的拒绝采样机制。Dabiriaghdam 等人^[56]提出的 SimMark 算法通过计算连续生成句子之间的相似性, 并采用重复拒绝生成机制确保生成的句子落在预定的语义区间内, 在保持内容质量的同时显著提升了抗攻击能力。这一策略使得 SimMark 在保持对生成内容质量的影响可忽略不计的同时, 显著提升了抗攻击能力。实验表明, 在误报率为 0.01 的设定下, 即使遭受释义攻击, 该方法仍能保证 0.72 以上的水印召回率, 但水印嵌入过程的计算开销较为显著。为满足黑盒 API 场景下多语言、多比特、高质量水印需求, Yu 等人^[57]基于多候选采样与特征选择原理提出了多比特水印框架 SAEMARK。该方法通过拒绝采样策略从大语言模型自然生成的候选文本中选择特征统计值最接近密钥生成目标值的文本单元, 从而将用户标识与时间戳等多比特信息编码并嵌入到生成内容中。水印检测时利用相同的密钥重构目标值, 并与待检测文本中提取的特征统计值进行对比, 进而解码出原始嵌入的水印信息。实验表明, 该方法在英文 C4、中文 LCSTS、代码 MBPP 等跨语言数据集上分别达到了 0.997、0.992 和 0.663 的 F1 检测率, 且文本质量分数基本与无水印文本持平。

③基于语义融合与大模型改写的水印技术

随着大语言模型能力的提升, 利用外部模型或深度语义分析进行文本重写成为增强水印隐蔽性与鲁棒性的新趋势。Chang 等人^[58]提出的 POSTMARK 基于语义融合原理, 利用公开文本嵌入器将原始文本映射至统一语义空间, 随后通过余弦相似度从秘密词表中检索与文本语义最相关的 k 个词汇, 并使用 GPT-4 等功能强大的插入模型, 在不改变事实与风格的前提下将这组词汇融入原文, 完成水印嵌入。水印检测端则对待检测文本进行词级别相似度阈值匹配, 从而进行水印检测。由于水印词汇的选取与嵌入均依赖高维语义, 因此该方法对改写攻击与释义攻击的鲁棒性较强。实验表明, 在 GPT-3.5 句子释义攻击下, 保持 0.01 误报率时, POSTMARK 能够达到 0.594 的召回率, 显著高于同条件下其他水印方法如红绿列表算法与 Unigram

等。为了克服传统后处理方法的局限并提高隐蔽性, Lau 等人^[59]提出了 Waterfall 框架, 利用大语言模型作为改写器, 结合词表置换和正交扰动技术, 实现了无需训练的高可扩展性通用文本水印。Xu 等人^[103]则利用两个行为差异化的 LLM 改写器交替对文本进行句子级改写来编码多比特信息, 显著提升了水印的隐蔽性。

④面向代码生成的专用后处理水印技术

针对代码数据严格的语法约束和功能性要求, 研究者发展出了基于抽象语法树和等价变换的专用技术。Li 等人^[104]提出首个专为代码生成 API 设计的水印保护系统 ToSyn。该系统基于编程语言的同义词替换, 在不损害代码质量的前提下实现了可验证的水印信息注入与用户级精准溯源, 具有较高的实用价值。与之类似, Li 等人^[61]提出了名为 ACW 的水印方法, 通过施加一组精心设计的保持语义且幂等的代码转换规则, 对模型生成的代码进行选择性修改, 从而实现水印的有效嵌入。该方法为大语言模型生成代码的溯源与版权保护提供了切实可行的技术路径。Yang 等人^[60]引入了 SrcMarker 双通道水印系统, 通过基于抽象语法树的可扩展代码变换嵌入 ID, 保持代码功能和可读性。Guo 等人^[105]提出的同名框架 ACW 则利用 AST 分析识别可替换代码组件, 并结合语义上下文自适应调整红绿列表采样, 实现高鲁棒性代码水印。Zhang 等人^[106]提出的 CLMTracing 结合基于规则的水印生成和对抗训练, 实现了针对代码模型的黑盒用户级追踪与版权保护。

2.1.3 大语言模型水印深度评述

(1)不可能三角分析

不同嵌入阶段的水印技术在效率、质量与安全性构成的三角约束中表现出显著差异。如表 3 所示, 模型内嵌技术在效率和质量上表现优异, 水印经训练内化为参数, 推理时无需额外计算且不影响主任务性能, 但在动态安全性上存在短板, 极易因微调或压缩而被清洗。推理阶段嵌入技术则面临生成质量与鲁棒性之间的零和博弈——一方的收益必然伴随着另一方的等量损失, 博弈各方收益和损失加起来总和为零。为了抵抗攻击而提升鲁棒性往往以牺牲文本自然度为代价, 且复杂的隐蔽算法会带来不可忽视的计算延迟。后处理技术在性能平衡上最为脆弱, 简单的替换破坏连贯性, 复杂的重写则引入不可控的语义漂移和额外的计算开销。

(2)哲学差异

表3 大语言模型水印深度评述对比			
维度\技术	嵌入模型	推理阶段	输出阶段
不可能三角	高效率、高质量,低安全性	高鲁棒会牺牲文本质量;复杂算法延迟高	低效率,质量与安全性冲突
哲学差异	回答模型归谁所有,侧重保护模型资产而非生成内容	回答内容是否由此生成,提供统计学证据	生成与确权分离,适合 API 和多模型统一管理
信息论视角	信息由权重承载,黑盒状态下,信道噪声会稀释水印信息	信道容量受限于文本熵。高熵易嵌入,低熵会致坍塌	二次编码会引入失真,高冗余度才能生存
工程实践	白盒取证难,易被防御者通过清洗移除	水印的串行依赖破坏了现代推理加速机制	NLP 领域风险大,代码领域可行性极高

三种技术路线在本质上回应了数字本体论中关于身份与归属的不同侧面。模型内嵌水印致力于解决本体层面的身份自证问题,即“证明我是我”,它将版权烙印熔铸于模型的权重基因中,旨在确立模型本身的独特性,而非直接干预每一次的言语表达。推理阶段嵌入水印则转向了因果层面的溯源确证,即“证明它出自我”,它在思想诞生的那一刻,即概率采样的瞬间,植入不可磨灭的统计学印记,建立起创造者与创造物之间的血缘纽带。后处理水印则体现了创造与所有权分离的二元思想,即“标记它属于我”,它将确权视为外在于生成的独立工序,如同工匠在成品上加盖印章,虽实现了灵活的权利主张,却切断了内容与生成机制内在的有机联系。

(3)信息论视角分析

信息论为不同技术确立了不可逾越的物理边界。模型内嵌技术面临信道噪声巨大的挑战,模型内部多层网络的非线性特征使得仅凭输出文本很难反推模型权重的微小变动。推理阶段嵌入受限于文本熵的物理上限,在创意写作等高自由度场景下尚有操作空间,但在代码或事实陈述等低熵场景下,强制嵌入会导致语义崩塌或逻辑错误。后处理技术属于对信号的二次编码,依据数据处理原理,这一过程无法增加有效信息反而不可避免地引入失真,必须在破坏原文语义和嵌入足够信息之间做出艰难取舍。

(4)工程实践可行性

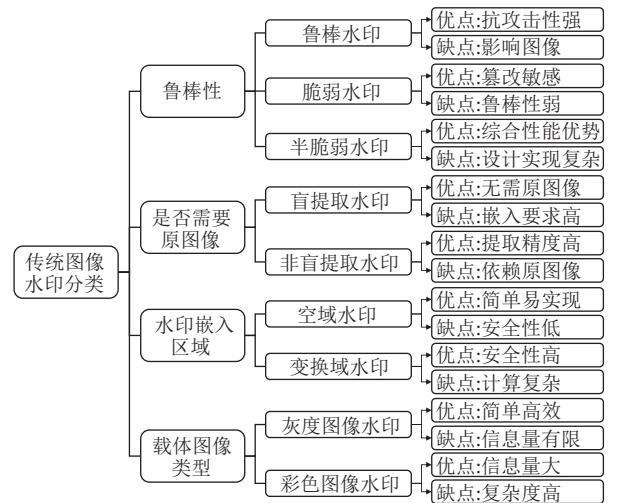
在实际部署中,各技术路线面临着不同的落地阻碍。模型内嵌技术存在法律取证难题,原告难以强迫侵权方公开模型权重以供白盒检测,因此更适用于内部版本管理。推理阶段嵌入与现代大模型广泛采用的并行加速机制存在冲突,其串行依赖特性会显著拖慢推理速度,是工程优化的难点。后处理技术虽然在通用文本领域应用受限,但在代码生成场景下展现出独特优势,能够通过语法结构的等价变换,在不改变程序运行逻辑的前提下实现有效的

合规保护。

2.2 多媒体大模型水印

2.2.1 图像大模型水印

在数字图像安全领域,版权保护与内容真实性验证始终是研究的核心议题。传统的数字图像水印技术通过在载体图像中隐藏具有特定意义的数字信号,为所有权认证、内容完整性监控及篡改定位提供了重要手段。根据赵蕾等人的研究^[107],传统图像水印技术已形成完善的分类体系,如图9所示,其从多个维度构建了具体划分类别及其对应的优缺点。



图像水印按鲁棒性可分为旨在抵抗各类信号处理攻击的鲁棒水印、对像素级篡改高度敏感的脆弱水印,以及兼顾两者特性的半脆弱水印。按提取方式包括无需原始载体参与的盲提取水印,以及提取精度更高但依赖原图的非盲提取水印。图像水印按嵌入区域分类,涵盖了算法简单但安全性较低的空域水印,以及通过修改频率系数如离散余弦变换、离散小波变换实现更高安全性与不可见性的变换域水印。最后按载体图像类型分类,分为简单高效的灰度图像水印与具备更大数据容量的彩色图像水印。

然而,传统水印技术在应对现代复杂安全挑战时存在显著局限。首先,不可见性、鲁棒性与容量难以在多功能应用中达到协同最优。其次,基于传统变换域的算法在抵抗旋转、缩放等几何攻击方面表现欠缺。最关键的是,传统方法缺乏对生成式内容的针对性,难以应对 DeepFake 及 AIGC (AI Generated Content) 大规模分发带来的溯源难题。

随着大模型在图像生成领域的快速发展,图像水印技术在该领域的应用日益受到重视。针对图像生成式大模型,相关水印技术按照水印嵌入位置可划分为嵌入模型和嵌入图像的水印两大类,更细致的水印类别划分及优缺点如图 10 所示。随着大模型在图像生成领域能力的持续提升,图像水印技术已成为保障模型版权和输出内容溯源的关键手段。

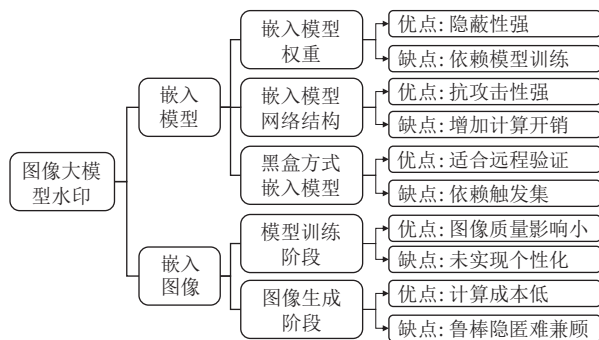


图 10 图像大模型水印分类图

(1) 嵌入模型的水印技术

嵌入模型的水印技术旨在将水印信息集成至大模型的内部结构中。对于图像生成大模型,通常将水印信息嵌入到模型权重、模型网络结构或基于后门机制植入到模型内部等方式实现^[108]。

①嵌入模型权重的水印技术:通过约束模型训练将水印信息嵌入冗余权重参数中。Uchida 等人^[109]率先提出一种基于正则化约束的嵌入方法,通过在深度神经网络的代价函数中引入嵌入正则化器实现。该正则化器引导网络在训练过程中将多比特水印向量 $\mathbf{b} \in \{0, 1\}^T$ 嵌入卷积层权重,同时不影响原始任务的性能。具体嵌入原理如公式(4)(5)所示:

$$b_j = s \left(\sum_i X_{ji} w_i \right) \quad (4)$$

$$E_R(\mathbf{w}) = - \sum_{j=1}^T \left[b_j \log(y_j) + (1 - b_j) \log(1 - y_j) \right] \quad (5)$$

在公式(4)中, X_{ji} 为嵌入矩阵时的投影向量, w_i 是待提取的水印信息,投影向量与水印信息做内积并通

过阶跃函数 s 处理得到第 j 位水印 b_j , 取值为 0 或 1。在公式(5)中, T 为水印的位数, y_j 为第 j 位水印的预测值, $E_R(\mathbf{w})$ 总体为二分类交叉熵损失,用于训练模型参数 $\mathbf{w}$,使得从中提取出的水印位 y_j 尽可能接近真实水印位 b_j 。嵌入过程通过交叉熵损失函数最小化提取误差,提取阶段则通过对网络参数进行投影与阈值处理实现水印恢复。该方法在理论层面具备可行性,但若水印信息被嵌入至未收敛的权重,其可检测性会显著下降。

针对该问题, Wang 等人^[110]提出将水印限定嵌入已收敛的权重中,从而显著提升水印检测的稳定性。为进一步提高水印鲁棒性, Liu 等人^[111]设计了基于贪婪选择策略的权重选择算法,依据权重改变时的残差绝对值筛选出对模型影响显著的关键参数进行嵌入,有效增强了水印的抗干扰能力,但其确定性的嵌入位置在面对有针对性攻击时仍显脆弱。为了提高水印模式的多样性, Tartaglione 等人^[112]基于随机选择的思想提出了使用模型权重的任意随机子集嵌入水印信息的方法,通过精心设计的损失函数使得水印信号广泛分布于模型网络各层,从而增强其抗篡改能力。该方法在有效增加了水印容量的同时,也具备了更强的隐蔽性,但在水印提取的准确性和效率方面仍待进一步优化。

②嵌入模型网络结构的水印技术:该类技术通过引入额外网络层实现水印嵌入,相较于权重修改方法,通常具备更强的抗攻击能力。Fan 等人^[113]提出一种验证层嵌入方法,通过在模型中添加护照层并加强其与偏置项及网络权重之间的依赖关系,实现水印的嵌入与检测。在训练阶段,护照信息在嵌入过程中被整合到卷积层后的护照层中,使得模型推理性能和护照存在紧密相关。该过程通过联合优化原始任务损失和护照嵌入损失,确保护照信息与模型推理结果的高度绑定。根据不同的任务需求,该水印方法提供三种水印检测方案:方案 V1 要求在推理阶段提供护照,若护照有效则模型性能与原始任务保持一致;方案 V2 则在验证阶段动态加载护照层,推理阶段无需护照,从而减少了计算开销;方案 V3 结合触发集与护照层,支持在黑盒和白盒环境下进行验证,增强了验证的灵活性和鲁棒性。由于验证权重的复杂性和随机性,攻击者难以准确估计其参数,因而,该方法具备较强的鲁棒性,但验证层的引入也相应增加了模型的计算负担,对模型的整体运行效率产生一定的影响。

为了实现水印信息与模型输入的映射关联,Rouhani等人^[114]提出了一种利用满足高斯分布的激活函数进行水印嵌入的方法。该方法在满足高斯分布的激活函数输出上设计特定的损失函数进行训练,使得在目标模型中嵌入水印的自定义激活函数与不含水印的激活函数输出一致。该方案在面对模型微调或压缩等操作的情况下仍能保持稳定,展现出较强的水印鲁棒性。

③以黑盒方式嵌入模型的水印技术:相比于嵌入模型权重与嵌入模型网络结构的水印方法,基于黑盒方式嵌入模型的水印^[115-116]技术仅需访问模型输出即可完成水印验证,具有更高的实用价值。Adi等人^[117]基于加密承诺技术(cryptographic commitment scheme)首次提出基于后门的黑盒模型水印方法,通过在训练中构造触发集和对应的错误标签集,使模型在特定输入下产生异常输出,从而嵌入水印。水印生成过程利用加密承诺技术完成,以确保水印的隐秘性和不可篡改性。验证阶段仅需检测模型在触发集上的输出和承诺的完整性,以确保水印的存在和有效性。该方法在面对微调等攻击时仍能保持较强的鲁棒性,且不影响模型的主任务性能。

为了防止模型被非法复制与篡改,Abdelaziz等人^[118]提出了基于滤波器原理的水印方法,其首先生成包含特定水印信息如姓名和时间戳的验证字符串,并使用密钥对其进行加密,进一步转化为一个二维矩阵作为水印滤波器,通过正常嵌入和空嵌入两种机制在模型训练过程中分别关联水印与特定标签及原始标签。其中,正常嵌入将水印与特定标签关联,而空嵌入则将水印与原始标签关联。验证水印时,通过检测模型能否准确识别水印数据并将其分类为预定标签,并测试带有反转水印的数据是否与原始标签一致,从而保护模型版权并确保水印不会干扰模型性能。Sun等人^[119]受Adi等人的研究启发,提出使用与原始训练数据完全无关的额外类别数据作为触发集,显著提高了水印的可检测性,但其嵌入水印密钥的图像使用了最低有效位图像隐写技术,易被专业工具识别。为了提高触发集水印的隐蔽性,Ye等人^[120]提出了基于深度隐藏的水印方案。将视觉上不可感知但模型可识别的特征嵌入触发图像,使模型通过训练学习到水印触发图像的细微特征,从而实现高隐蔽性的水印嵌入,但其有效性依赖于模型的完整训练过程,在微调场景下水印效果欠佳。

为了研究支持远程验证的黑盒水印方法,

Zhang等人^[121]在训练数据中注入特殊的水印样本以实现水印嵌入,仅需通过模型的API发送水印输入并验证输出结果即可完成远程所有权确认。该方法验证过程较为便捷,但简单的水印样本嵌入导致其抗攻击性较弱。为进一步提升远程验证水印方法的鲁棒性,Li等人^[122]提出了基于多视图数据的水印方法MAT,构建具有多类特征的触发集,并在训练阶段进一步嵌入特征正则化损失,促使模型更加关注触发集中的关键特征,增强模型抗篡改的能力。在后续的验证环节,模型所有者通过API调用比较模型在触发集上的输出与预期水印标签的一致性进行检验,从而精准判断模型是否遭受篡改。

在上述研究基础上,为了进一步增强水印的鲁棒性,Huang等人^[123]基于博弈论原理提出了水印攻防分析框架,通过建模攻击者与防御者之间的策略博弈,寻求在保持模型性能前提下的最优水印嵌入方案。该方法通过引入攻击者与防御者之间的博弈寻求均衡点,为设计高鲁棒性的水印提供参考案例。

(2)嵌入图像的水印技术

嵌入图像的水印技术旨在将水印信息直接嵌入到模型输出图像中的技术,根据嵌入水印时机的不同,可分为训练阶段和图像生成阶段嵌入水印。

①模型训练阶段嵌入水印技术:该类技术通过将水印信息嵌入训练数据,使模型在训练过程中学习并保持水印特征,从而令生成的图像自然携带水印信息。其优势在于水印在生成过程中能够有效保留且对生成图像的质量影响较小,然而,该方法在处理大规模数据集时成本较高,且难以针对不同模型或用户实现具有区分度的水印定制。

Wu等人^[124]提出同时训练水印提取网络和宿主模型的双网络训练框架。在训练阶段,宿主网络参数根据任务损失和水印损失共同更新,使得生成图像中隐含特定水印。水印提取网络则仅基于水印损失进行优化,专门学习如何从图像中准确提取水印信息。训练完成后,宿主网络用于任务执行,而水印提取网络则保持私密,只有在提供正确密钥的情况下才能成功提取水印,从而实现图像来源验证和版权确认。该方法摆脱了对触发集的依赖,在保持模型性能的同时实现可靠水印检测,但其多网络协同优化机制较为复杂,训练成本较高。

为了实现对预训练Transformer模型的鲁棒、高效且通用化水印嵌入,Xu等人^[125]基于微调技术提出了一种名为TokenMark的模态无关水印系统。如图11所示,该系统通过对预训练模型在特定排列

数据样本上进行微调来嵌入水印,进而生成两组不同权重的水印模型:一组用于正常推理,另一组仅在输入为排列样本时激活水印提取功能。水印嵌入目标如公式(6):

$$\Theta_p = \left\{ \begin{aligned} &\theta_* \in \Theta | L_{wm}(G(F(Z, P(\theta_*))), y_t) < \epsilon_{wm} \\ &\wedge L_{wm}(G(F(Z, \theta_*)), y_t) > \epsilon_{wm} \end{aligned} \right\} \quad (6)$$

其中, θ_* 表示水印嵌入后的模型, L_{wm} 衡量模型输出与目标水印 y_t 之间的差异, G 为水印解码器, F 为 Transformer 的主干网络, Z 代表输入数据的嵌入表示, $P(\theta_*)$ 为对模型权重进行排列后的结果, 即水印功能的实现过程, $\wedge$ 为符号与, ϵ_{wm} 为用于判断水印是否被成功提取的阈值。该公式定义了成功嵌入水印的模型应同时满足的条件: 在排列输入下, 模型输出应接近目标水印 y_t ; 在正常输入下, 模型输出需远离目标水印 y_t 。实验表明, TokenMark 在提高模型水印的鲁棒性、检测效率和通用性方面表现出突出优势, 有望发展为一种统一的水印解决方案。然而, 该方法依赖模型微调以嵌入水印, 可能影响模型原始性能, 且在实际应用中需要确保排列矩阵的保密性以防止水印被恶意破坏。

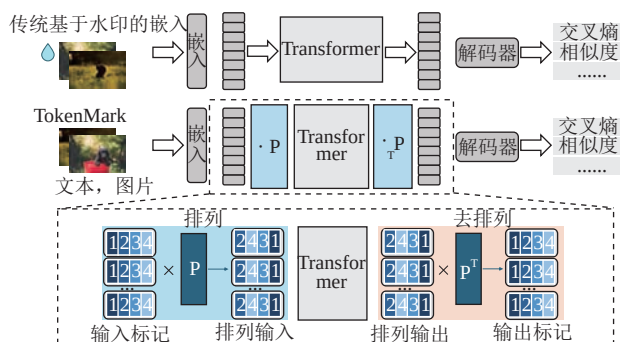


图 11 传统水印方案与 TokenMark 的比较^[125]

为了增强水印的鲁棒性, Cui 等人^[126]提出一种名为 FT-Shield 的水印框架, 通过设计包含微调损失的水印生成目标函数, 确保水印在微调早期被优先学习, 从而避免被后续学习覆盖。在水印检测阶段, FT-Shield 结合混合专家检测框架有效增强了水印检测器对不同微调方法的适应性。该框架显著提升了水印的存留能力, 但其训练过程计算开销较大。

为进一步增强水印抗攻击能力, Zhao 等人^[127]提出一种集成隐藏网络、攻击模拟模块与揭示网络的综合框架。隐藏网络将秘密图像编码为不可见水印并嵌入目标图像, 攻击模块在训练过程中模拟各类图像处理操作以增强水印鲁棒性, 揭示网络则负责

从水印图像中恢复秘密图像并验证完整性。该框架通过模拟攻击场景的对抗性训练, 有效提升了水印对各类攻击的抵抗能力。

为了针对不同模型及用户的差异化水印嵌入, 研究人员提出了多种创新方案。Yu 等人^[128]提出了基于水印编码器的不同类型水印嵌入方法。该方法首先使用预训练的水印编码器将二进制水印字符串嵌入训练图像, 使得模型生成的图像中自动嵌入视觉上不可察觉的水印。该方法的优势在于模型对应的独特水印指不会随着生成技术的迭代更新而失效, 但其区分维度仅局限于模型层面, 并未实现用户级别的个性化定制水印。

为了在用户层面将个性化水印嵌入模型生成的图像, Cui 等人^[129]设计了基于块策略的水印方案, 将水印划分为基本块序列, 为每个用户分配独特的基本块序列组合, 形成用户专属的水印标识。在检测阶段, 通过对可疑图像进行序列分割和匹配, 可准确识别数据所有者并有效保护图像免受版权侵犯。基于微调技术的应用, Ma 等人^[130]提出了面向主体驱动图像合成保护的生成式水印方法 GenWatermark。该方法通过生成器将水印以视觉扰动形式嵌入至原始图像, 检测器则通过对比水印图像与普通图像实现水印识别。通过针对特定合成主体对检测器进行微调, 使其能够适应特定主体的图像合成, 显著提高了水印的检测精度和个性化适配能力。实验表明, 该方法在艺术风格和人脸合成任务中均表现出良好的鲁棒性和隐蔽性, 能够有效防范未经授权的图像合成。

②图像生成阶段嵌入水印技术: 图像生成阶段的水印嵌入方法在模型推理过程实施水印嵌入, 相比于训练阶段嵌入水印的技术, 该技术无需参与训练过程, 具有计算成本低的显著优势。当前研究主要围绕提高水印的鲁棒性与隐匿性展开。

为增强水印抗几何攻击能力, Chen 等人^[131]提出了基于编码器-噪声层-解码器结构的鲁棒水印框架 RoWSFormer。在水印嵌入阶段, 原始图像经过卷积层提取低级特征, 并将特征图划分为不重叠的图块, 接着经过局部通道增强的 Swin Transformer 提取局部通道特征。水印信息经过扩散模块处理后, 与图像特征融合, 生成水印图像。在水印检测阶段, 解码器接收受到攻击的水印图像并进行卷积操作, 通过频率增强的 Transformer 块增强频域特征提取, 最终解码器通过信息提取层恢复嵌入的水印信息。频率增强的 Transformer 块处理过程如公式(7)

(8)所示：

$$F_w = FC(\text{stack}(DCT(X_{token}))) \quad (7)$$

$$X_{out} = F_w \cdot X_{token} \quad (8)$$

其中, X_{token} 为输入的图像 $token$, 其首先经过 DCT 离散余弦变换将信号从空域转换到频域, 接着通过 stack 将频域信号在通道维度拼接为一个向量, 并经过全连接层 FC 得到学习频率分量的重要性权重 F_w , 最终该权重与原始 $token$ 相乘形成频域注意力机制。实验表明, RoWSFormer 在面对几何与非几何攻击时, 在峰值信噪比 (Peak Signal-to-Noise Ratio, PSNR) 与提取精度上均优于现有水印方法, 但其实际应用的运行效率仍需进一步优化。

在提高水印的隐匿性方面, Gunn 等人^[132]提出名为 PRC 的图像水印方法, 其通过结合密钥编码生成 PRC 码字, 随后根据码字对初始潜在变量进行采样并生成水印图像。在水印检测阶段, 该技术通过从图像中恢复近似的潜在变量实现水印检测, 并根据需求借助相应的解码算法恢复水印信息。该方法在水印隐匿性和鲁棒性之间取得了良好平衡, 但依赖复杂的密钥管理和编码解码流程。为了同时提高水印的隐匿性和鲁棒性, He 等人^[133]提出了名为 IWFormer 的图像水印方法, 利用可逆神经网络框架引导水印嵌入人眼不敏感的中高频分量, 以提高水印的隐匿性。同时, 该方法结合仿射 Transformer 模块挖掘载体图像的全局相关性, 以增强水印的鲁棒性。与现有最先进水印模型相比, IWFormer 在水印隐匿性和鲁棒性方面均具有明显优势。

为满足多媒体图像在认证、篡改检测和定位以及图像恢复方面的安全需求, Aberna 等人^[134]提出了一种基于视觉 Transformer 的混合水印模型。该模型利用视觉 Transformer 生成具有不变性的注意力水印特征图, 并结合 Schur 分解和混合变换实现水印嵌入和提取。通过在高熵区域嵌入水印, 显著提高了水印的隐匿性和鲁棒性, 且在面对各种攻击时展现出较高的篡改检测精度并能够有效恢复篡改区域。然而其实现过程涉及多种复杂变换和分解过程。

为了简化水印嵌入流程, Lukas 等人^[135]提出 PTW 水印技术。通过复制并微调预训练生成器, 将水印消息编码至生成的图像中且保证嵌入的水印信息在视觉上不可察觉。该技术可作为后处理步骤应用于任何预训练的生成器模型, 无需对模型进行训练。相比于从头训练水印生成器的传统方法, PTW

技术显著提升了水印嵌入效率。然而, 尽管嵌入技术不断优化, Zhang 等人^[136]的理论指出, 强鲁棒性在数学上仍面临不可逾越的边界, 这意味着单纯的效率提升并不能消除极端对抗环境下的安全隐患。

2.2.2 音频大模型水印

数字音频的版权保护与内容完整性认证是信息安全领域的核心课题。作为多媒体安全的关键技术, 数字音频水印通过将特定的版权信息或标记隐蔽地嵌入宿主音频载体中, 旨在解决多媒体文件面临的伪造、恶意编辑及非法传播等安全风险。在生成式人工智能尚未普及的阶段, 传统音频水印技术已发展出多种成熟的实现范式, 主要包括最小显著位、回声隐藏、扩频技术、拼贴方法及量化指数调制等。

这些传统方法为音频安全奠定了基础, 但面对大模型时代生成的深度伪造语音, 其局限性日益凸显。由于人类听觉系统的高度灵敏性, 在维持高不可感知性的同时显著提升嵌入容量和鲁棒性面临极大的技术瓶颈。特别是传统算法在面对复杂的同步攻击时往往表现脆弱, 且难以平衡信噪比与复杂元数据嵌入之间的矛盾。这种性能权衡的难题揭示了针对生成式音频的水印需求。

Tong 等人^[137]提出了基于 Transformer 的语音水印框架, 如图 12 所示。在水印嵌入阶段, 该方法首先对原始语音信号进行短时傅里叶变换 (Short-Time Fourier Transform, STFT) 获取声学特征, 接着音频编码器与水印编码器分别提取语音与水印特征, 二者经扰动生成器 G 融合生成微扰信号。如公式 (9)(10) 所示:

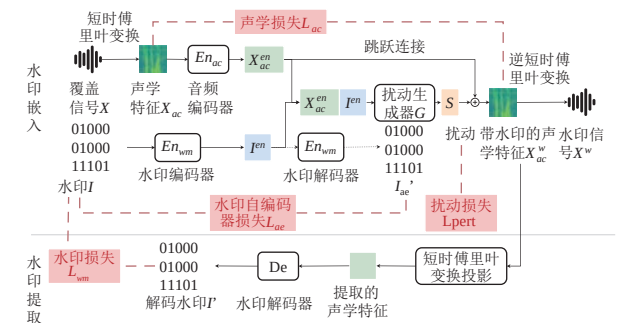


图 12 基于 Transformer 的水印框架概述^[137]

$$X_{ac}^w = X_{ac}^{en} \oplus G([X_{ac}^{en}, I^{en}]) = X_{ac}^{en} \oplus S \quad (9)$$

$$X^w = ISTFT(X_{ac}^w) \quad (10)$$

其中, X_{ac}^{en} 为编码后的声学特征, I^{en} 为编码后的水印特征, G 为基于 transformer 的扰动生成器, S 为经过

水印扰动生成的水印扰动特征, $\oplus$ 表示将扰动叠加到语音特征上的逐元素加法, 语音特征通过跳跃连接与微扰信号叠加得到 X_{ac}^w , 再经逆短时傅里叶变换重构为带水印语音 X^w 。在水印提取阶段, 带水印语音经过 STFT 获得声学特征后, 经语音特征提取与水印解码器恢复嵌入的水印信息。该框架在训练阶段模拟噪声增强水印的鲁棒性, 展现出更强的水印不可感知性和抗攻击能力。

针对大模型生成音频的检测需求, San 等人^[138]提出了 AudioSeal 音频水印技术, 旨在主动检测并定位由模型生成的语音, 以应对语音克隆带来的安全风险。该技术通过在音频中嵌入不可感知的水印信号实现音频的主动检测与精确的片段级定位, 并可追溯至具体模型版本。为了优化水印检测与归因准确性, Liu 等人^[139]提出名为 XAttnMark 的音频水印方法, 利用交叉注意力机制和部分参数共享策略在生成器和检测器之间实现高效的水印嵌入和提取。实验表明, 在如回声、低通和高通滤波等音频编辑操作下, XAttnMark 的检测准确率保持在 0.99, 高于 AudioSeal 的 0.97。其平均归因准确率为 0.93, 显著优于 AudioSeal 的 0.39, 表现出卓越的鲁棒性和溯源能力。综上, 基于深度神经网络的水印技术不仅在性能上全面超越传统方法, 也为构建端到端的版权保护系统提供了可靠的技术路径。

水印定位^[140]是音频水印领域中长期存在的挑战, 传统方法通常依赖在水印前添加同步码以实现快速定位, 但易受到失同步攻击的影响。针对该问题, Chen 等人^[141]提出基于可逆神经网络的音频生成水印技术 WavMark, 通过将编码和解码视为逆过程并引入移位模块, 确保在解码位置发生移位时也能准确恢复水印。该方法在实现音频水印的高不可见性与部署灵活性的同时, 显著提升了定位效率。

为了提高音频水印的容量、鲁棒性和定位效率, Li 等人^[140]提出了基于双重嵌入策略和可逆神经网络 (Invertible Neural Networks, INN) 的 IDEAW 水印方法。该方法采用了两阶段可逆神经网络结构: 首先在水印嵌入阶段通过 INN1 将水印信息嵌入经 STFT 转换为时频域的原始音频中。随后使用 INN2 将同步码嵌入已含水印信息的时频表示, 再通过逆 STFT 将时频音频转换回时域信号。水印提取过程是嵌入过程的逆操作, 先由 INN2 从水印音频中提取定位码, 再由 INN1 恢复水印信息, 整个过程计算开销较低, 确保了高效提取。为了增强鲁棒性, IDEAW 在训练中引入攻击层与平衡块机制:

攻击层模拟常见水印去除攻击操作, 平衡块则通过引入额外的可训练参数助力调整攻击后水印音频分布, 使其更接近原始音频的分布, 恢复模型的对称性, 从而提升模型稳定性与鲁棒性。实验表明, IDEAW₄₆₊₁₀ 模型在保持 0.99 的高提取准确度的同时, 水印容量达到了 56 bps, 包括 46 位水印消息与 10 位定位码, 显著高于 WavMark 的 32 bps, 且 IDEAW 在面对各种水印去除攻击 (如高斯噪声、低通滤波、MP3 压缩等) 时, 提取准确度均优于 WavMark 方法。

Eppl 等人^[142]在上述音频水印方法的基础上探究了音频水印技术在检测音乐生成模型训练数据中未经授权使用方面的有效性。通过在 MusicGen 模型的训练数据中注入不同的音频水印技术, 研究者发现了两个影响水印注入方法成功的关键因素: 注入到模型训练数据中的水印数据量和水印对模型音频分词器的鲁棒性。水印数据量的多少直接影响模型在生成音频时对水印的嵌入效果, 即过少的水印数据可能导致模型难以有效学习水印信息, 从而影响水印的可检测性。水印对音频分词器的鲁棒性是确保水印不易被攻击或去除的另一个关键因素, 决定了水印在模型处理过程中的存留能力, 是抵御针对性攻击的重要保障。

2.2.3 多模态水印

多模态水印技术可应用于文本、音频、图像等不同模态的数字内容中。这类水印的设计在人类感官层面保持不可察觉, 但是可通过专用算法进行有效检测。Qiu 等人^[143]针对多模态水印的研究提出了基础性见解: 与仅限于单一媒体类型的传统水印相比, 多模态水印可通过跨格式集成以增强安全性和灵活性, 其核心目标在于版权声明与内容的真实性验证。研究发现, 多模态水印对图像损坏和文本扰动较为敏感, 其中缩放和模糊操作对图像水印的影响最大, 而在文本水印中, 字符级扰动的影响明显大于句子级扰动。这一发现为多模态水印的未来发展指明了重要研究方向。

在跨模态水印整合方面, Nie 等人^[144]探索了多种模态水印的嵌入方法, 提出了针对不同数据类型实施差异化水印嵌入策略的多模态融合水印方案: 对文本数据进行同义词替换与句子释义, 对音频数据实施频谱特征修改, 对图像数据则利用隐写技术实现水印嵌入。该研究初步验证了多种模态水印的有效嵌入机制, 为跨媒体水印融合奠定了技术基础。根据模态组合方式的不同, 目前多模态水印方法主

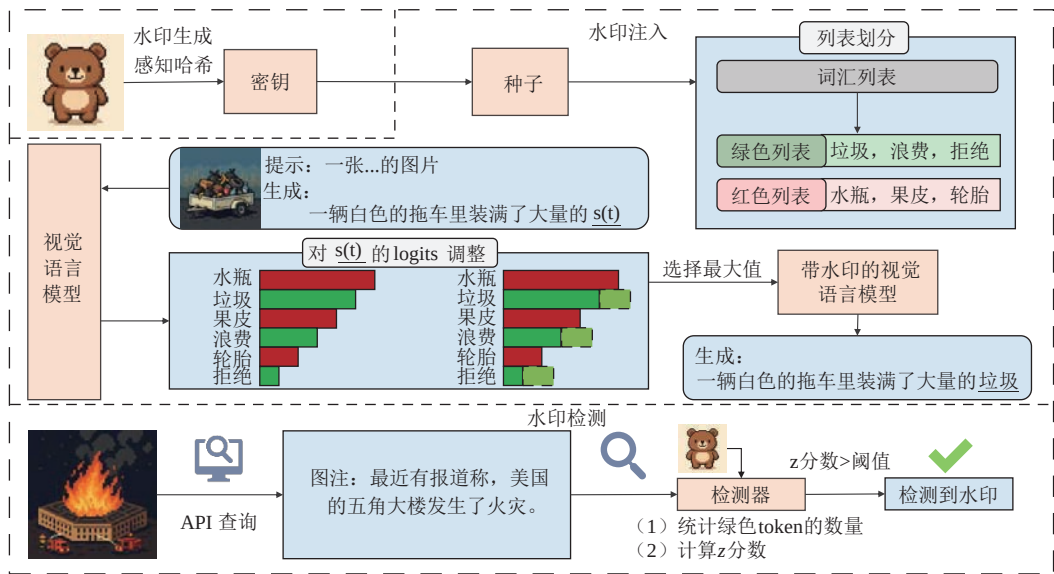
要包括图像-文本跨模态水印与视频水印两大类。

(1)图像-文本多模态水印技术

视觉-语言预训练模型(Visual Language Models, VLMs)^[145]作为理解并分析图像和文本内容的核心框架,能够有效捕捉视觉和语言模态间的深层语义关联。该类模型通过大规模图文对比学习,实现从视觉输入中提取信息,并将视觉信息与语言描述匹配,并根据语言描述生成相应的视觉内容的任务。

为了保护 VLMs 的知识产权,Wan 等人^[146]基于词汇列表划分原理提出了适用于多模态模型的水印

方法。该方法包括水印生成、嵌入和检测三个步骤。如图 13 所示,在水印生成过程中,模型所有者通过感知哈希算法将秘密图像转化为密钥,该密钥用作划分模型词汇列表的种子。水印嵌入阶段依据密钥将词汇列表划分为绿色与红色列表,并通过提升绿词概率嵌入水印。水印检测阶段则使用相同秘密图像生成密钥,重新构建红绿列表,统计生成文本中绿词频率并采用 z 检验判断水印是否存在。实验表明,该方法实现了超过 0.9 的水印检测率,最大限度降低了对模型性能的影响,展现出良好的鲁棒性。



VLMs 在下游任务如跨模态信息检索、图像分类和目标检测中表现卓越,并已发展出基于 VLMs 的嵌入即服务应用框架。然而,EaaS 框架存在模型提取攻击的安全隐患,可能导致多模态模型所有者遭受巨大损失。针对该问题,Tang 等人^[147]提出了多模态水印技术 VLPMarker,其核心机制是选择分布外的图像-文本对作为后门触发器,并在学习过程中将其嵌入 VLMs 中。该方法具有强鲁棒性并能有效抵御各种攻击,但其触发集设置较为复杂的缺陷限制了其在实际场景中的部署。

(2)视频水印技术

在数字媒体技术飞速发展的背景下,视频内容的安全与版权保护已成为信息隐藏领域的研究重点。作为一种将秘密信息以不可见形式嵌入载体中的关键技术,数字视频水印主要用于版权归属判定、内容真实性认证及广播监控等场景。视频水印可按鲁棒性、载体类型如 2D、3D、VR 等及嵌入方法进行系统化分类。其中,嵌入方法可分为将视频视为图

像序列处理的内容基算法,以及结合视频压缩标准在量化或预测阶段实施嵌入的码流基算法。

在评价体系方面,视频水印需兼顾保真度与鲁棒性。尽管引入深度学习如卷积神经网络和生成对抗网络在一定程度上提升了性能,但现有技术仍面临显著挑战:一方面,手工设计的特征难以抵抗复杂的组合攻击;另一方面,现有的盲提取方案在面对帧删除、帧插入等特有的时间同步攻击时,往往难以精确定位水印位置。更重要的是,随着 Sora 等生成式视频(VideoGenAI)的兴起,内容从原始拍摄转向人工智能生成,迫切需要能与生成过程耦合且抗生成式伪造的新型安全机制。这种对于跨格式安全性和灵活性的追求,促使水印技术从单一媒体类型向多维度不断发展,从而推动了多模态水印技术的研究。

Ye 等人^[148]提出了一种将深度学习图像水印方法高效适配至视频水印的 ItoV 策略。如图 14 所示,ItoV 通过将视频的时间维度与通道维度合并,使深度神经网络能够以图像处理方式处理视频数据。在

编码器部分,原始视频经视频处理器提取其空间和时序特征,水印信息通过消息处理器扩展至与视频维度一致,接着二者在通道维度上进行拼接并输入水印生成器以生成带水印的视频。训练过程中引入的攻击模拟层通过应用模糊、噪音等多种攻击模拟实际场景中的失真,从而增强水印的鲁棒性。最终,解码器对遭受攻击的水印视频进行下采样操作,提取并恢复原始信息。该方法的核心公式如式(11)(12):

$$\mathcal{L}_F = \sum_{i=1}^L l_1^2(F^{(i)}(V_c), F^{(i)}(V_w)) \quad (11)$$

$$\mathcal{L}_{total} = \lambda_E \mathcal{L}_E + \lambda_D \mathcal{L}_D + \lambda_F \mathcal{L}_F \quad (12)$$

其中, L 为视频帧数, $F^{(i)}$ 用于提取第 i 帧图像, V_c 和 V_w 分别为原始视频和嵌入水印后的视频, l_1 指的是用于测量两帧之间差异的L1范数函数。 $\mathcal{L}_E$ 为编码器损失, $\mathcal{L}_D$ 为解码器损失, $\mathcal{L}_F$ 为帧一致性损失, λ_E , λ_D , λ_F 分别为权重系数。整个框架通过联合优化损失函数,在保证水印隐蔽性的同时,显著提升了其在各类失真条件下的可恢复性。

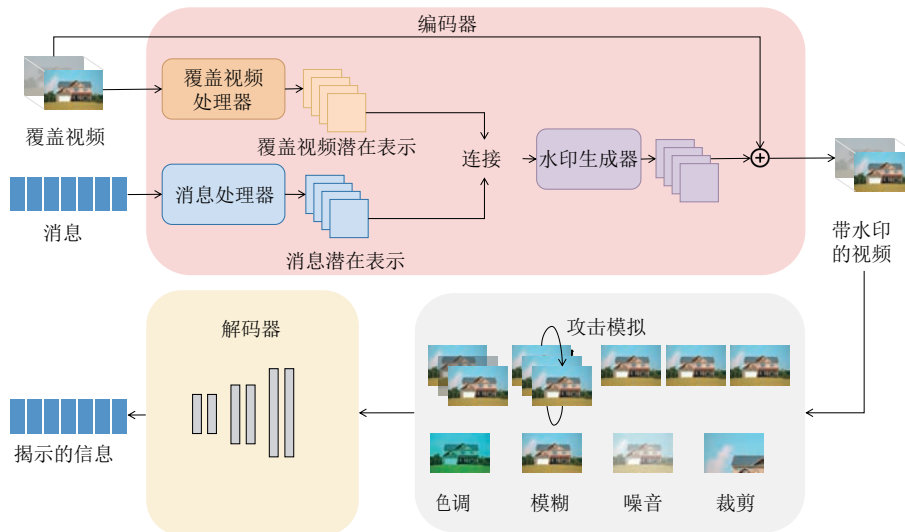


图14 ITOV框架的三部分:编码器、攻击模拟和解码器^[148]

为了运用水印检测深度伪造(Deepfake)内容,Guo等人^[149]提出了音频-视频水印框架AVSecure。该框架采用统一的编码器-解码器架构,在音频和视频同步嵌入不可见的水印,作为鉴别伪造内容的线索。其中的音频-视觉特征编码器对齐音频和视频信息,既提升了重建质量,也增强了检测性能。在训练阶段引入的失真层不仅增强了水印对常见后处理操作的抵抗能力,还提升了对Deepfake内容的敏感性。该框架在检测各种先进的Deepfake方法时展现出优异的鲁棒性和泛化能力。在提高视频水印嵌入效率方面,Fernandez等人^[150]开发了开源视频水印框架Video Seal,其采用多阶段训练,包括图像预训练、混合后训练和提取器微调,并结合时间维度的水印传播技术,无需逐帧进行水印处理即可实现水印在视频序列中的有效扩散。该设计在保持视频质量的同时,显著提高了水印嵌入的效率与稳健性,为大规模视频版权保护提供了实用化解决方案。

2.2.4 多媒体大模型水印深度评述

(1)不可能三角分析

不同模态的感知特性决定了其在质量、鲁棒性与效率三角约束中的重心差异,如表4所示。图像水印受限于感知空间的几何结构,理论研究指出强鲁棒性在数学上与完美质量互斥,攻击者可通过随机游走将图像移出检测域,且为了抵抗重绘攻击,水印须分布于全局语义特征中,这往往以牺牲生成质量为代价。音频水印面临更为严苛的物理约束,人类听觉系统对瞬态和频谱的极度敏感压缩了可用信道功率,导致必须采用复杂的深度网络进行编解码以换取抗压缩能力,从而在实时通话等场景中引入难以接受的延迟瓶颈。多模态视频水印则受困于时间维度的连续性约束,独立帧嵌入会导致视觉闪烁,而满足时空平滑性的三维整体嵌入虽然保证了质量,却因巨大的计算开销难以满足流媒体直播的实时性需求,导致效率成为其最大短板。

(2)哲学差异

各模态水印技术在本体论层面回应了数字内容

表4 多媒体大模型水印深度评述对比

维度/模态	图像大模型水印	音频大模型水印	多模态大模型水印
不可能三角	信道宽,但对抗“重绘”需语义级嵌入,易破坏构图导致视觉伪影	人耳掩蔽阈值低,水印生存空间窄;强鲁棒性必然导致音质下降	逐帧嵌入会导致闪烁,鲁棒性需对抗时序重排,计算开销大
哲学差异	引发算法所有权与训练数据风格权的伦理悖论	强制绑定 AI 属性,阻断深度伪造带来的认知欺骗	利用图文/音画对齐,对抗格式转换带来的流动性遗忘
信息论视角	利用扩散过程将水印编织进全局语义,鲁棒性依赖语义	水印需在“人耳听到”与“不可察觉”间寻找极窄通道	利用跨模态掩蔽效应,视频需在时间低频系数中嵌入
工程实践	Adobe CAI 等标准已落地,视觉对高频噪声不敏感,便于隐藏	流媒体对延迟极其敏感,需解决去同步攻击带来的定位难题	3D 卷积计算量大,需取舍关键帧传播与全局嵌入

不同的存在性危机。图像大模型水印充当了“数字出生的独一性证明”,它在理论上可无限重现的统计合成图像中人为重构了独一无二的标识,但同时也引发了算法父权与训练数据母权之间的归属悖论。音频大模型水印旨在实现“主体代理权的再绑定”,针对语音克隆将声音从肉身剥离导致的认知失调,水印强制宣告声音源自算法而非生物本体,在混淆的听觉表征中划定了人与非人的界限。多模态大模型水印则提供了“跨模态的一致性证明”,利用图像与文本在语义空间的映射关系,确立了一种不随文件格式流转而消失的持久性证据,对抗数字内容在传播过程中的流动性遗忘。

(3) 信息论视角分析

信道容量与编码策略的差异构成了不同模态水印的物理边界。图像处于连续高维空间,拥有极高的信道容量,生成模型本身可作为非线性编码器将水印编织进全局语义结构,从而实现对裁剪等物理攻击的高鲁棒性。相比之下,音频数据缺乏空间维度且需对抗去除冗余的有损压缩算法,有效容量极低,且面临严峻的去同步挑战,必须在伪装成重要听觉特征与保持音质之间寻找极窄的优化通道。多模态大模型水印则被视为联合信源编码,理论上可利用视觉冲击降低听觉敏感度等跨模态掩蔽效应来动态分配负载,但视频水印为防止时间闪烁而对时间维度施加的平滑性约束,实质上相当于低通滤波,极大地压缩了可用的信道带宽。

(4) 工程实践可行性

商业落地进程在不同模态间呈现出显著的阶梯状分布。图像大模型水印得益于人类视觉系统对高频信息的不敏感性及内容真实性倡议等行业标准的推动,商业化应用最为成熟,公众接受度较高。音频大模型水印仍受限于同步对齐难题与实时编解码的算力要求,在抗去同步攻击与低延迟之间尚未找到完美的平衡点。多模态大模型水印目前尚处于理论

探索期,跨模态触发机制在开放域数据上的不稳定性以及重新训练大型多模态模型的高昂成本,严重阻碍了其从实验室走向大规模工业应用的步伐。

3 大模型水印评估

3.1 大语言模型水印评估

大语言模型水印评估是衡量水印算法在实际应用中有效性的关键环节,其核心在于通过系统化的量化评估与可视化呈现,比较不同水印算法在多样化环境条件和任务需求下的综合表现。本文介绍了关于水印算法性能评估的四个关键维度,分别为大语言模型水印的可检测性、对生成内容质量的影响、水印鲁棒性与水印容量^[151-152],并梳理汇总了各类水印算法在这四种评估指标上的实验性能。

3.1.1 大语言模型水印的可检测性

大语言模型水印的可检测性是指在水印嵌入后能够在检测阶段被有效识别的能力,这一指标对于验证模型归属权或追溯生成内容来源具有重要意义。根据水印实现原理的不同,水印的检测方式也具有不同的方式,具体可分为嵌入模型水印的可检测性与嵌入模型生成内容水印的可检测性。

在嵌入模型的水印检测方面,Wei 等人^[153]的研究揭示了模型规模与训练数据量对水印可检测性的影响规律。实验表明,对于 70 M 参数规模的模型,当训练数据量达到 6 B tokens 时,水印检测性能达到峰值,超过该阈值之后呈现下降趋势。这一现象可归因于水印损失随训练数据集规模的增大而上升,同时,水印在更大的数据集中出现的频率相对降低,共同导致水印强度减弱。进一步研究发现,更大的模型在相同的训练设置下表现出更低的水印损失:在训练数据集规模从 2 B tokens 逐渐扩展到 12 B tokens 的过程中,70 M 模型的水印损失会逐渐增大,而 410 M 模型的水印损失保持相对稳定,表明

模型的容量与其水印记忆能力呈正相关。特别地,当训练数据规模和模型容量同时增加时,水印的强度能保持稳定,表明在这种情况下水印机制具备有效性。该类水印检测通常采用客户端-服务器架构:客户端向远程服务器发送水印检测请求,远程服务器根据接收信息在数据库中进行对比,若检测到适配水印数据,则向客户端发送水印检测成功响应,否则表明水印验证失败。

在嵌入模型生成内容的水印检测方面,以红绿列表水印为例,其原理旨在对文本中的令牌进行分类和统计分析,通过比较绿词占比与预设阈值实现水印判断。此类方法的检测性能与绿词数量及分布密切相关,在无攻击的理想情况下,红绿列表算法在 OPT-6.7 B 与 GEMMA-7 B 模型上的 AUROC 值分别达到了 1.000 与 0.995,展现了极高的基准检测能力。但在低熵场景下^[80]面临显著挑战。由于其生成模式更为固定,生成过程中令牌的概率分布较为集中,导致水印的嵌入和检测难度陡然增加。为应对这一挑战, Lu 等人^[151]提出了一种基于熵值的文本水印检测方法 (Entropy-based text Watermarking Detection, EWD),旨在解决在代码生成等低熵场景下模型生成文本与人类撰写文本难以区分的问题,其核心思想是基于熵值为每个令牌分配差异化检测权重:对于高熵值的令牌赋予更大权重,因而更容易嵌入水印;对于低熵令牌则分配较小权重。具体水印检测方法可通过公式(13)实现:

$$z' = \frac{|S|_G - \gamma \sum_{i=m}^{|T|-1} W_i^2}{\sqrt{\gamma(1-\gamma) \sum_{i=m}^{|T|-1} W_i^2}} \quad (13)$$

其中, z' 为基于熵权重的检测得分, 其中, $|S|_G$ 表示文本中带水印的绿色 token 数量, W_i 是第 i 个 token 的权重, γ 代表模型词汇表中绿色 token 的比例, $|T|$ 是文本的 token 总数量, m 代表检测窗口大小, 该式用于计算考虑熵的权重后的文本检测得分。与式(2)相比, 式(13)通过引入熵权重对水印嵌入与检测进行修改, 有效提高了低熵文本的检测准确性。研究者在不同的高熵(新闻生成)和低熵(代码生成)数据集上验证了该方法, 显示 EWD 在低熵场景下具有明显优势, 在高熵场景中与其他方法表现相当。

3.1.2 大语言模型水印对内容质量的影响

大语言模型水印对生成内容质量的影响是评估水印技术实用性的关键指标, 主要体现在生成文本

可读性、语义完整性和语言流畅性等方面^[154-155]。水印嵌入过程通常需要干预模型的生成机制, 如干扰词汇选择、调整句法结构等, 这些干预可能导致生成文本风格趋于单一化或上下文连贯性下降。因此, 水印技术的有效应用必须以不显著降低文本质量为前提^[156]。水印嵌入过程导致文本质量大幅下降将直接影响水印技术的实际应用价值。基于此, 系统评估水印嵌入对文本质量的影响具有重要的研究意义, 下文将从词级别和句子级别分别探讨水印对生成内容质量的具体影响。

词级别水印嵌入通常通过限制词汇的可选择范围来实现, 从而降低生成文本的质量。例如基于词汇集合划分的水印方法^[157], 通过在较高使用频率的词汇中选出一小部分禁用从而实现水印嵌入, 在检测阶段则验证文本是否避开这些特定词汇以验证水印的存在。该方法因限制常用词汇使用而导致生成文本质量下降, 且下降程度与禁止使用的词汇数量成正比, 红绿列表算法在 OPT-6.7 B 和 GEMMA-7 B 模型上的文本困惑度分别高达 20.00 和 22.00, 远高于其他优化算法。该研究进一步发现, 在禁用词汇的选择上, 需权衡其信息量: 既不能选取信息量过高的字符, 也不宜选择信息量很低的字符, 二者都会明显损害生成文本质量。为缓解此类问题, 现有方法如红绿列表水印会适度放宽禁用词汇, 允许其以低概率在生成文本中出现, 以平衡水印强度和文本质量。值得注意的是, 并非所有词级别水印方法均会降低文本质量。Unbiased 和 MCMARK 等算法通过优化策略, 已将 LLama2 模型上的文本困惑度分别降低至 6.44 和 6.39, 显著改善了生成质量。此外, Li 等人^[42]提出基于词汇重要性评分的水印框架 WIS, 引入如式(14)所示的令牌重要性评分机制:

$$f(s=[s_1, \dots, s_{t-1}], s_t): V^* \times V \rightarrow [0, 1] \quad (14)$$

其中, s 是步骤 t 时与已经生成的输出 s_t 连接的输入, V^* 代表前面被标记重要的序列集合, 若该式的结果 $f(s, s_t)$ 大于预设的阈值, 那么就把 s_t 预设为重要词汇。WIS 框架通过令牌评分机制保留对文章质量关键的重要性令牌, 从而缓解文本质量下降, 在嵌入水印的同时维持甚至提高生成文本的质量。

句子级别水印对生成内容质量的影响主要体现在语法结构、语义一致性和上下文连贯性等方面。为在保持水印可检测性的同时实现隐蔽嵌入, 模型通常需要对句法结构、语序或语义表达进行细微调

整,此类操作可能削弱句子的自然性和流畅度,尤其是在处理长句和复杂句时,不当调整易破坏原有的逻辑关系与语义完整性。因此,句子级别水印需在可检测性与文本生成质量之间寻求平衡,通过最小化语法干扰与语义失真,确保水印文本仍具备较高可读性与逻辑一致性。例如,SemaMark在OPT-6.7B模型上的困惑度仅为4.50,而PersonaMark在Gemma-2B-IT上的困惑度更是低至2.74,证明了句子级与个性化策略在质量保持上的优势。

3.1.3 大语言模型水印的鲁棒性

大语言模型水印的鲁棒性指水印在遭受各种攻击^[158-159]时仍能保持可检测性的能力。该特性需在实际对抗环境中检验,具有强鲁棒性的水印方案不仅能够有效抵御现有攻击手段,还能推动水印攻击技术的演进,从而形成“攻击-防御”相互促进、共同发展的良性技术循环。

文本水印面临的主要攻击类型涵盖多个层级:字符级别的同形字符替换、拼写错误、字符删除、插入与交换等操作;词级的同义词替换攻击;段落级翻译攻击与基于释义器的释义攻击;以及模型级的模型伪造攻击。研究表明,不同水印方法的抗攻击能力存在显著差异^[160]。例如,基于统计特征的水印方案在应对简单的同义词替换攻击时展现出较强的抵抗能力,而依赖编辑距离的方法在面对局部插入与删除攻击时具有较高鲁棒性。然而,当攻击者使用更强大的释义模型如DIPPER进行高质量同义重写时,攻击成功率会达到0.9以上^[161],水印的抗攻击能力面临严峻挑战,如红绿列表算法在GEMMA-7B模型下遭受PEGASUS释义攻击时,其在0.01误报率前提下的召回率仅为0.246,DiPmark的TPR也仅为0.282,显示出明显的脆弱性。研究还发现,上下文无关的水印设计在面对句子级别的重写攻击时,其鲁棒性反而优于部分上下文相关方案^[160]。相比之下,基于主题的TBW算法在OPT-6.7B模型

下对抗PEGASUS攻击的TPR高达0.910;而句子级的k-SEMSTAMP与SimMark算法在BookSum数据集上的TPR分别达到了0.941和0.933,表现出极强的鲁棒性。总体而言,目前的文本水印方法面对语义等价的篡改攻击时仍显脆弱,需要通过更精细的设计以保证其抗攻击能力。

此外,嵌入模型的水印方法鲁棒性受到模型架构影响显著,相较于传统循环神经网络,基于Transformer架构具备更为庞大的参数空间与更灵活的嵌入位置选择,更有利于在保持模型本身的完整性的前提下,实现鲁棒的水印嵌入。当前的研究趋势表明,采用多位置嵌入、多模态融合及引入错误校验机制等策略能够进一步提升水印的存留度和恢复能力。总体来看,大语言模型水印的鲁棒性仍面临诸多挑战,尤其是在确保水印不可察觉性的同时,如何应对不断演进的新型攻击并保持稳定性和可验证性,是未来研究亟待解决的关键问题。

3.1.4 大语言模型水印的容量

水印容量是衡量水印系统嵌入能力的关键指标,指在不影响文本质量的前提下,水印方法能够嵌入多少信息以实现生成内容的版权保护与有效追踪,是构建完整水印评估体系的重要组成部分。高容量水印不仅能增强对文本的版权保护能力,还能扩展水印在用户追踪、内容认证等场景中的应用范围。

根据信息承载能力的不同,大语言模型水印可分为零比特水印和多比特水印两类^[103,162],二者在功能定位和应用场景上存在显著差异。表5给出了针对不同容量水印的多角度比较。零比特水印主要用于检测文本中是否存在水印标记,具有实现相对简单、隐蔽性强的优势,但由于无法携带额外信息,其应用范围局限在基础的版权验证。相比之下,多比特水印则能够嵌入更丰富的信息^[163],如用户的个性化数据等,显著拓展了水印的功能维度。

表5 不同容量的大模型水印分类

水印种类	容量	功能	特点	代表工作
零比特	小	验证 版权	隐蔽性高, 检测效率高	红绿列表 ^[35] 无偏水印 ^[37]
多比特	大	生成内容溯源	信息容量大,应用场景广泛	MPAC ^[47] 动态分段 ^[164]

在高容量水印研究方面,Xu等人^[103]提出了基于语义释义的大语言模型多比特文本水印技术,将带有水印信息的多比特二进制数据嵌入生成的文本

中。与传统的词汇替换或者文本结构修改不同,该方法创新性地使用两个分别代表0和1的释义器,在句子级别向文本中嵌入水印信息。释义器接收原始

文本和水印消息作为输入,根据当前待嵌入的水印比特位动态选择对应的释义器对目标句子进行改写,从而实现多比特水印信息嵌入。这在保证文本质量的同时实现了优于零比特水印的信息携带容量。

Lin 等^[164]则提出了一种基于动态分段嵌入的多比特水印方法。该方法通过推导的不等式计算每比特水印需要的文本长度阈值,在文本生成过程中实时监测当前段落长度,达到阈值即切换至下一段的嵌入过程。在嵌入具体实现中,该方法则是根据当前嵌入的 0 或 1 比特值调整 logits 分布,进而影响 token 采样过程^[165],实现信息的安全嵌入。此外,Wang 等人^[98]提出的可编码文本水印技术进一步拓展了容量研究的边界,该水印方法能将丰富的信息编码到文本水印中,其消息空间突破了传统的二元限制,为零比特和多比特水印提供了统一的理论框架。在此基础上,He 等人^[166]从信息论角度系统地分析了大语言模型多比特水印文本质量、可检测性和容量之间的内在权衡,建立了严谨的理论分析框架,为未来多比特水印的发展提供了重要理论支持。后续研究可根据该信息理论框架,深入探索保证高鲁棒性和高可检测性时进一步提升容量的有效路径。

3.1.5 水印评估研究

为了实现大语言模型水印的综合评估,需对上述评估指标进行系统性交叉研究,Molenda 等人^[167]提出了 WaterJudge 框架,深入研究了平衡水印可检测性与生成文本质量的问题,并利用零样本比较量化二者进行权衡。在该框架中,生成文本质量和水印可检测性的量化指标公式分别如式(15)和式(16):

$$s_q = \frac{1}{N_d} \sum_{i=1}^{N_d} P(q(y_w^{(i)}) > q(y^{(i)}) | x^{(i)}) \quad (15)$$

$$F_{0.5} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \quad (16)$$

在式(15)中, s_q 表示水印文本质量优于非水印文本的概率, N_d 代表输入文本集合 D 中的样本数量, $y^{(i)}$ 是给定输入 $x^{(i)}$ 时非水印语言模型生成的输出文本。 $y_w^{(i)}$ 对应给定输入 $x^{(i)}$ 时模型生成的输出文本,该指标计算得出输入 $x^{(i)}$ 时水印文本 $y_w^{(i)}$ 质量优于非水印文本 $y^{(i)}$ 的概率,评估水印对文本质量的影响程度,当均值趋近于 0.5 时,表明引入水印对文本质量的影响可忽略不计。式(16)采用 $F_{0.5}$ 分数作为衡量水

印检测性能的评价标准,其中 β 是权重参数,默认取值为 0.5, Precision(精确率)表示正确检测到的水印文本数量与总检测到的水印文本数量的比值, Recall(召回率)反映正确检测到的水印文本数量与实际水印文本数量的比值。该指标得分越高,表示水印检测系统的性能越优越。

在上述研究的基础上,Giboulot 等人进一步引入了水印鲁棒性评估指标,提出 WaterMax 方法,该方法通过对多段生成文本进行综合评估以选择最优文本,深入探讨了水印技术的可检测性、鲁棒性和文本质量三者之间的权衡问题。虽然 WaterMax 与 WaterJudge 框架的核心思想不同,但二者均能够实现水印性能的多目标权衡。需要指出的是,现有评估体系仍存在一定局限,这两种方法均未将水印容量纳入评估体系。未来研究可通过整合可检测性、生成内容质量、鲁棒性、容量等四个维度的评估指标并构建统一可比较的完整评价框架,为水印算法的性能比较与优化选择提供全面参考,进而推动水印技术的系统化发展。本文汇总了各类水印算法在这四种评估指标上的实验性能,如表 6 所示。各实验中主要使用的 C4 数据集^[168]是由网络数据清洗而成的无监督大规模英文文本语料库,属于通用领域网页文本,用于自然语言处理预训练,包含约 3.65 亿篇文档和 8000 亿个词元;Booksum 数据集来源于书籍文本,并结合学习指南、教育资料等人工撰写摘要进行构建,用于长文本理解与抽象式摘要生成,共包含 405 本书籍、12000 对正文与摘要样本。

3.2 多媒体大模型水印评估

随着大模型水印技术的不断发展,构建科学、标准化的水印评估体系对于衡量水印技术的有效性与实用性至关重要。不同于单一文本模态,多媒体水印的评估需要在保证感知保真度的前提下,保证水印在复杂信道传输与模型攻击下的鲁棒性以及承载信息的容量。由于图像、音频与多模态内容在感知上存在显著差异,其评估指标与测试基准也具有较大差异。本节将分别从图像、音频及多模态三个维度出发,详细阐述各领域内的主流评估指标、关键性能维度及相应的测试基准,旨在为大模型水印技术的性能评估提供统一的参考标准。

3.2.1 图像大模型水印评估

评价图像大模型水印算法优劣具有四个关键维度:水印容量、检测准确性、保真度与鲁棒性^[169]。在水印容量方面,该指标主要衡量算法所能承载的有

表6 大语言模型水印方法实验性能

实验设置			四大性能评估指标							
水印方法	模型	数据集	可检测性 ↑			文本质量	水印鲁棒性(TPR@1%FPR)			水印容量
			AUROC ↑	TPR@1% FPR ↑	F1 分数 ↑	文本困惑度 ↓	token 级攻击	sentence 级攻击		
								DIPPER	PEGASUS	
GaussMark ^[32]	Mistral-7B	C4	0.980	0.850	-	-	0.370	-	-	零比特
	Llama3.1-8B		0.980	0.910	-	-	0.370	-	-	
	Phi3.5-Mini		0.950	0.700	-	-	0.360	-	-	
强化学习水印 ^[33]	OPT-1.3B	C4	0.998	-	-	2.42	0.942	0.914	0.931	零比特
	Llama2-7B	C4	0.999	-	-	8.30	0.542	0.980	0.974	
红绿列表算法 ^[35]	OPT-1.3B	C4	0.996	0.984	0.998	14.92	-	-	-	零比特
	OPT-6.7B		1.000	0.996	0.998	20.00	-	0.124	0.578	
	GEMMA-7B		0.995	1.000	0.997	22.00	-	0.052	0.246	
Unigram ^[36]	OPT-6.7B	C4	1.000	0.996	0.994	18.00	-	0.516	0.900	零比特
	GEMMA-7B		0.998	0.998	0.996	18.00	-	0.034	0.598	
Unbiased ^[37]	LLama2	C4	-	0.891	-	6.44	0.594	0.004	-	零比特
MCMARK ^[38]	LLama2	C4	-	0.989	-	6.39	0.961	0.110	-	零比特
BiMarker ^[39]	OPT-1.3B	C4	-	0.994	0.992	7.00	-	-	-	零比特
WatME ^[40]	LLAMA2-7B	GSM8k	0.912	0.955	-	-	0.790	0.910	-	零比特
		TruthfulQA	0.865	0.949	-	-	-	0.673	-	
		C4	0.972	-	-	6.89	-	-	-	
DiPmark ^[41]	OPT-6.7B	C4	0.999	0.992	0.994	18.00	-	0.028	0.228	零比特
	GEMMA-7B		1.000	1.000	0.999	18.00	-	0.022	0.282	
TBW ^[43]	OPT-6.7B	C4	1.000	0.994	0.995	12.00	-	0.536	0.910	零比特
	GEMMA-7B		0.998	1.000	0.999	12.00	-	0.196	0.842	
SynthID-Text ^[44]	OPT-6.7B	C4	0.999	0.992	0.995	12.00	-	0.058	0.446	零比特
	GEMMA-7B		1.000	1.000	0.997	13.00	-	0.024	0.484	
SIR ^[81]	OPT-6.7B	C4	0.995	0.964	0.978	15.00	-	0.248	0.726	零比特
	GEMMA-7B		0.990	0.890	0.973	14.00	-	0.190	0.476	
SemaMark ^[48]	OPT-2.7B	C4	0.994		0.990	4.80	-	-	-	零比特
	OPT-6.7B		0.995		0.985	4.50	-	-	-	
SEMSTAMP ^[51]	OPT1.3B	C4	0.992	0.939	-	-	-	-	0.837	零比特
		BookSum	0.996	0.983	-	12.89	-	-	0.943	
		RedditTIFU	0.997	0.977	-	-	-	-	0.958	
k-SEMSTAMP ^[52]	OPT1.3B	C4	0.996	0.981	-	11.82	-	-	0.927	零比特
		BookSum	0.999	0.991	-	-	-	-	0.941	
PersonaMark ^[53]	Gemma-2B-IT	prompt datasets	1.000	0.99	-	2.74	-	-	-	多比特
	Phi-3.5-mini-Instruct		1.000	0.99	-	3.19	0.900	-	-	
Whitemark ^[55]	NLLB-200-3.3B	WMT	1.000	1.000	1.000	-	-	-	-	零比特
SimMark ^[56]	Cosine	C4	0.996	0.968	-	-	-	-	0.903	零比特
		BookSum	0.998	0.988	-	12.69	-	-	0.933	
		RedditTIFU	0.991	0.963	-	-	-	-	0.945	
	Euclidean	C4	0.998	0.985	-	-	-	-	0.723	
		BookSum	1.000	1.000	-	12.69	-	-	0.826	
		RedditTIFU	0.998	0.987	-	-	-	-	0.947	
SAEMARK ^[57]	Qwen2.5-7B-Instruct	C4	1.000	0.998	0.997	-	0.930	-	-	多比特

效载荷大小,并决定了水印的应用场景。零比特水印仅用于分析载体中是否存在水印,主要服务于版权保护;而多比特水印则对应一个 n 位的比特串,不仅需要确定水印的存在性,还需精确识别提取出的比特内容,这种高容量特性使其更适用于指纹识别和精细化的内容溯源。在检测准确性方面,针对零比特水印的二分类本质,主要通过真阳性率和假阳性率来评估,理想目标是在极低的假阳性率下实现高真阳性率;对于多比特水印,则侧重于信息提取的完整性,通常采用误码率或比特准确率来衡量,其中误码率直观反映了每个比特位发生错误的概率。

保真度指标旨在评估嵌入水印后是否降低了原始大模型的生成性能或图像质量。在模型层面,主要使用CLIP分数来评估含水印模型生成的图像与文本提示词之间的语义一致性,CLIP值越低意味着水印嵌入对生成语义的影响越微弱。在图像质量层面,评估分为有参考指标与无参考指标两类:前者包括峰值信噪比、结构相似性等;后者则包含自然图像质量评估分数和感知图像质量评估分数,通过量化图像的自然统计特征来反映视觉感知质量。鲁棒性评估则涵盖了抗图像处理与抗模型处理两个维度。抗图像处理鲁棒性重点考察含水印图像在经历压缩、滤波、裁剪、缩放等几何变换及恶意篡改后仍能被检测的能力;而抗模型处理鲁棒性则专门针对AIGC大模型,评估其在面临攻击者试图通过模型微调、权重剪枝或权重量化等手段修改模型参数以去除水印时的稳定性。

3.2.2 音频大模型水印评估

为了在统一标准下量化衡量不同水印算法在生成式模型中的表现,学术界建立了一套标准化的评估框架,重点考察鲁棒性与保真度两个维度。在鲁棒性评估中,主要采用比特恢复准确率来精确衡量提取信息与原始水印的一致性,并结合水印检测概率判断算法在面对攻击时的生存能力。在保真度评估方面,研究者综合采用了信噪比、语音质量感知评估以及虚拟语音质量客观听众指标,全方位测定加水印操作及后续攻击对音频听感质量的影响。Wen等人^[170]基于该框架复现了九种开源水印方案,并在涵盖语音与歌声的三个大规模公共数据集上进行了测试。实验设计了涉及各类信号处理与模型攻击的22种类型共计109种配置,处理了超过20万个水印音频样本,其大规模实证数据客观揭示了当前技术在面对高强度、智能化攻击时仍存在显著的性能瓶颈,为后续算法优化提供了明确的数据支撑。

3.2.3 多模态大模型水印评估

多模态大模型水印的评估体系需要在兼顾单一模态质量的同时,保证跨模态语义一致性与联合鲁棒性。传统的评估指标如峰值信噪比或误比特率已不足以全面衡量多模态水印的性能,因此学术界建立了一套涵盖感知质量、语义保真度及跨模态生存能力的综合评估框架。在感知质量与语义保真度方面,Liu等人^[171]提出应引入多模态对比语言图像预训练分数作为核心指标,用于量化加水印后的多模态输出,如视频与其配文在语义层面的对齐程度,确保水印的嵌入不会导致图文或音视频内容的语义割裂。

在联合鲁棒性评估方面,评估重点从单一的抗压缩或抗噪声能力转向了对跨模态操作的抵抗力。这包括衡量水印在经历跨模态检索、模态截断以及生成式跨模态编辑后的留存情况。Qiu等人^[148]构建了一个包含多种跨模态攻击手段的基准测试集,通过对大量多模态样本的分析发现,仅在单一模态如图像嵌入水印的方案在跨模态检索任务中的召回率会损失近0.4,而采用多模态联合嵌入方案的性能损失则控制在0.1以内。这一数据有力地证明了在评估多模态水印时,必须将跨模态生存能力作为衡量算法优劣的关键指标,以适应生成式人工智能内容在不同媒介间频繁流转的应用现状。

4 大模型水印攻击与防御

4.1 大语言模型水印攻击与防御

随着大语言模型水印技术的持续发展,其攻击技术^[172-174]也随之不断发展,不断演进。按照攻击原理和实施粒度的不同,可将现有大语言模型水印攻击技术分为改写攻击、翻译攻击、释义攻击、模型伪造攻击^[158]四类。如表7所示,这些攻击方法的修改粒度依次递增:由修改单个字符、单个词的局部改写攻击,到进行文档级修改的翻译攻击和释义攻击,最终扩展至模型层面的模型蒸馏攻击,不同方法的攻击成功率如表8所示。为了抵御不同类型的攻击方式,研究者不断提出多种抗攻击水印方案及防御机制。下文将分别对这些攻击方法及对应的防御策略进行系统阐述。

4.1.1 改写攻击与防御

改写攻击通过在保持语义基本不变的前提下,对文本进行局部修改(插入、删除和替换等方式)以去除水印。根据是否针对现有大语言模型水印技

表7 针对大语言模型水印的攻击类型及对应的防御手段					
攻击类型	原理	攻击特点	典型攻击参考	对应防御措施	防御特点
改写攻击	通过修改生成内容来破坏水印	攻击方式多样,包括插入、删除、替换等	字符、词级别改写 临近替换 ^[136] SCTS ^[175]	预处理技术 ^[161] 嵌入语义空间水印 ^[57]	重点应对字级改写,但难挡词汇级改写 对词汇级改写鲁棒性强,但会引入风格偏移等问题
翻译攻击	破坏水印完整性,绕过水印检测	可通过多种语言翻译展开攻击,灵活性高	CWRA ^[82] 语言翻译 ^[176]	聚类多语言语义簇 ^[82] 多语言回译水印文本 ^[176]	通过语义簇嵌入侧防御,但依赖外部词典且语种难覆盖 通过多语回译在检测进行防御,但开销增大且依赖翻译器
释义攻击	利用其他模型重写带水印文本	在不改变文本语义的前提下消除水印特征	PEGASUS ^[177] DIPPER ^[178]	检索数 据库 ^[178] 嵌入语义空间水印 ^[51]	实现方式简单,但隐私与成本不可控,难以抵御多轮释义 对释义攻击鲁棒性强,但会引入风格偏移等问题
伪造攻击	伪造新模型,替代原模型功能	成本低,可分别进行水印擦除与欺骗攻击	Stealing ^[179] CDG-KD ^[180]	限访问速率; 验证训练 ^[181]	但不能直接识别伪造攻击 从根本上抵御模型伪造,成本高

表8 不同攻击方式的攻击成功率				
攻击类型	模型、数据集与水印方法	具体攻击方式	攻击幅度	攻击成功率
改写攻击	OPT1. 3B, C4 RealNewsLike, KGW	字符级	0. 1	0. 2228
			0. 5	0. 9949
		词级	0. 1	0. 0224
			0. 5	0. 9725
翻译攻击	Vicuna-7b, C4 RealNewsLike, KGW	绿色词汇替换	0. 35	1. 0
	BAICHUAN-7B, Multi News, KGW	回译	-	0. 2177
		跨语言水印移除	-	0. 7853
释义攻击	OPT1. 3B, C4 RealNewsLike, KGW	DIPPER 释义器	0. 5	0. 7710
		PEGASUS 释义器	-	0. 8758
模型伪造攻击	GLM-4-9B-Chat, HarmfulQ, KGW	擦除 CDG-KD	-	0. 4
		欺骗 CDG-KD	-	0. 94

术,改写攻击可进一步划分为直接进行字符级别或词级别修改的传统改写攻击与针对水印的改写攻击。

传统改写攻击主要基于字符或词汇级别的修改。Zhang 等人^[161]指出字符级与词级攻击方式的区别:字符级攻击包括同形字符替换、拼写错误、字符删除、插入与交换,词级攻击则是随机选择词语并利用同义词进行替换。实验表明,在0.1的低编辑率条件下,字符级攻击的攻击成功率普遍为词级的2-3倍。例如在“position”作为单个 token 的前提下,词级替换将其替换为同义词“location”,攻击范围仅为单个 token,而在字符级扰动情况下,如将中央字符 s 替换为同形字符 š(U+0161),那么原本的单个单词“position”在进行检测时会被分段为3个子词[po, š, ition]。攻击范围从单个 token 扩展到3个 token。

针对红绿列表水印算法的攻击,Wu 等人^[175]为了绕过传统改写攻击提出了基于自我测试的颜色替

换(Self Color Testing-based Substitution, SCTS)方法,通过策略性提示获取颜色信息并据此替换绿色标记以中和片段中的绿色标记数量。但该攻击方法仅适用于划分红绿列表的水印算法,对其他水印方法不具有适用性。为了实现通用性改写攻击,Zhang 等人^[136]提出了基于“质量预言机”和“扰动预言机”的黑盒攻击框架,首先访问模型获得带水印的输出并利用质量预言机评估质量,随后将该输出通过扰动预言机生成候选输出并进行质量评估,迭代生成与评估直到水印检测算法的置信度降低到阈值以下,表明水印成功移除。其中,扰动预言机通过语义保持的文本片段替换实现,本质上属于改写攻击。

防御方面:针对传统改写攻击的防御方法主要依赖于在文本预处理阶段采用预设规则消除潜在扰动字符^[161]。现有方法主要包括以下四类策略:首先,基于词典与编辑距离的纠错,通过自动检测并纠正疑似拼写错误的字符序列;其次是依托图像重渲

染的光学字符识别(Optical Character Recognition, OCR)技术,将受攻击文本转换为图像后,利用OCR引擎重新提取文本,以消除视觉不可见的扰动字符;第三种方法是采用Unicode兼容分解的归一化方法,将语义等价但编码不同的字符统一映射至唯一标准码位,从而消除同形异义字符带来的干扰;最后,基于白名单的异常字符过滤,通过建立合法字符白名单,直接删除不在允许范围内的字符。实验表明,上述防御方法针对Unigram水印算法的防护效果最为显著,最高可将改写攻击成功率降至0.2760;然而,对于如红绿列表与无偏水印算法等其他类型的水印方法,其防御效果较为有限。

除上述传统防御方法外,Yu等人^[57]提出了能够有效抵御词汇替换攻击危害的水印技术SAEMARK。与传统对改写攻击敏感的词级别水印算法不同,SAEMARK将水印信息嵌入句子的语义特征中,即使攻击者对文本进行大量词汇替换,只要其语义基本不变,水印仍可被有效检测。实验表明,在0.01误报率的前提下,面对0.1比例的词级别替换攻击,该方法能保持0.93的召回率,AUC值达0.976,在同比例词汇删除攻击下,AUC值仍能维持在0.858,展现出在语义层面抵御各类改写攻击的强鲁棒性。

4.1.2 翻译攻击与防御

相较于以局部词汇修改为特征的改写攻击,翻译攻击采用文档级的整体语言转换策略,通过在不同语言间进行转换,以实现水印的移除。值得注意的是,由于跨语言使用的实际需求,普通用户在日常工作中也可能无意识地对大语言模型生成的文本实施类似“翻译攻击”的操作,因此提升水印技术对翻译攻击的鲁棒性具有重要的现实意义。

传统翻译攻击^[103]方法主要采用“翻译-回译”的双重转换机制,尽管这种操作会改变文档的整体表达,但由于对句子结构的改动程度有限,因此攻击效果并不显著。为了深入研究大语言模型水印的翻译攻击,He等人^[82]提出了跨语言一致性的概念,即大语言模型生成文本中嵌入的水印在文本被翻译成另一种语言后仍能保持其强度的能力。基于这一概念,该研究团队进一步提出了一种跨语言水印移除攻击(Cross-lingual Watermark Removal Attack, CWRA)的翻译攻击方案,如图15所示。CWRA包含三个关键步骤实现翻译攻击:(1)将原始提示翻译成中介语言;(2)利用带水印大语言模型生成对中介语言的响应;(3)运用翻译系统将该响应回译至原

始语言。实验表明,经过上述三层处理即能够有效削弱甚至移除响应中的水印信号。

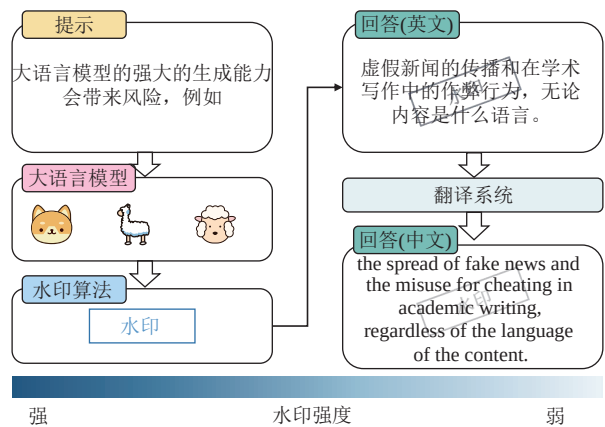


图15 跨语言环境中水印稀释的示意图^[82]

防御方面:由于常规翻译和翻译攻击都会通过语言转换削弱水印强度,开发具有跨语言鲁棒性的大模型水印至关重要。针对这一挑战,He等人^[82]提出了具有跨语言一致性的水印技术X-SIR。该方法根据语义相似性将词汇表中的标记分成不同的簇集并在嵌入水印时对整个语义簇施加统一的偏置调整。这种设计确保即使在翻译的过程中单个标记发生变化,其所属的语义簇仍能保持水印的一致性。然而,该方法存在一定局限性,即X-SIR构建语义聚类所依赖的外部词典仅覆盖完整的单词,无法有效处理词根、前缀和后缀等语言单位。同时,X-SIR覆盖的语言范围受限於模型词汇表和外部词典,在实际应用中难以应对攻击者任意选择的语言组合。为克服这些限制,Mohamed等人^[176]提出了一种模型无关且非侵入式的多语言水印检测方法STEAM,旨在提升水印在翻译攻击下的鲁棒性。该方法将待检测文本回译至多种支持语言以生成多语言候选文本集,分别计算各版本的水印分数z-score,并选取其中的最大值作为最终的水印检测依据。此外,STEAM引入了语言特定的z-score归一化机制,显著提升了检测的适应性。实验表明,STEAM在包括17种语言测试集上表现优异,尤其对中低资源语言的检测效果展现出更强的鲁棒性,相比X-SIR方法平均AUC提升超过0.19,在低资源语言如孟加拉语上AUC提升最高可达0.33。即使在多步翻译攻击、翻译器不一致等复杂场景下,该方法仍能保持稳定的检测性能。

4.1.3 释义攻击与防御

与翻译攻击类似,释义攻击同样针对文档级内

容进行修改,但其应用范围更为广泛。该攻击通过辅助模型对原始大语言模型生成的内容进行重述,在保持语义基本不变的前提下改变文本的统计特征,从而达到欺骗水印检测器的目的。

在攻击技术方面,Zhang等人^[177]提出的PEGASUS模型引入间隔句生成(Gap Sentences Generation, GSG)作为自监督预训练目标,通过从文档中抽取关键句子作为遮蔽目标,并基于剩余内容生成被遮蔽的句子,使模型在预训练阶段即具备全局语义建模、关键信息识别与摘要式语言生成能力。该模型在大语言模型水印中充当高质量释义器,生成语义相同但句式迥异的文本,有效破坏水印文本中的统计特征。Krishna等人^[178]进一步提出了基于上下文感知的释义攻击模型DIPPER,通过控制词汇多样性和内容重组程度,生成不同扰动程度的重述文本。在GPT-3.5-davinci-003、GPT2-XL、OPT-13B等多个大语言模型上进行了开放式文本生成和长篇问答以检测攻击的效果,实验表明,经过DIPPER重述后,检测器对GPT2-XL生成文本的检测率从0.703降至0.046,同时重述文本与原始文本的语义相似度得分达到0.98~0.99的高水平,显示出卓越的释义效果。

防御方面:为应对此类攻击,该研究团队提出了一种基于检索的防御方法,通过API将DIPPER模型生成的所有文本内容存储于专用数据库中,防御系统将用户提交的候选文本通过检索数据库进行比对以识别重述文本。虽然该方法对释义攻击展现出较强的抵御能力,但是需要存储所有生成文本的设计带来了严重的用户隐私泄露风险,在实际应用中受到限制。针对这一缺陷,Sadasivan等人^[182]进一步提出递归释义攻击,采用多轮释义、多模型如DIPPER、LLaMA-2-7B-Chat等的释义方式,使得大语言模型生成的文本逐渐偏离原始的生成模式,从而降低水印检测器的检测率。经过多轮释义攻击,基于检索的防御方法不仅成本会急剧提高,其防御效果也因多轮释义器的混用而大幅降低。为应对该问题,Hou等人^[51]提出了基于语义空间嵌入的SEMSTAMP方法。该方法将水印嵌入句子的语义表示中,而非依赖表面的统计特征,从而有效抵御各类释义攻击。实验表明,SEMSTAMP在Pegasus、Parrot、GPT-3.5及其“强化释义”攻击下,AUC仍保持在0.828-0.992之间,比同场景下的基线红绿列表水印算法高出0.019-0.134,且生成文本的困惑度仅为10.20,与无水印文本的10.02几乎一致,

实现了高鲁棒性与高质量的双重目标。

4.1.4 模型伪造攻击与防御

模型伪造攻击在现有攻击技术的基础上进一步扩大了攻击范围,通过查询水印大语言模型的API实施逆向工程从而实现水印擦除和欺骗两类攻击。擦除攻击旨在消除大语言模型生成内容中的水印信号,使其被检测器判定为非水印文本。欺骗攻击的机理在于攻击者生成非水印文本并诱导检测器将其误判为水印文本。

Jovanović等人^[179]提出了首个自动化水印窃取算法,其攻击流程如图16所示。攻击者通过持续向水印大语言模型发送查询请求以获取其生成的带水印的文本响应,从中学习水印规则并构建评分函数。评分函数在生成每个候选词时按照公式(17)调整其生成概率:

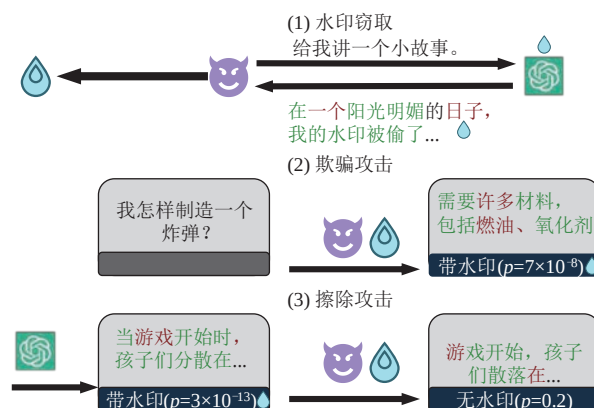


图16 水印盗取及其所导致的后续攻击概述^[179]

$$l'_T = l_T + \delta_{att} \cdot s^*(T, [T_1 T_2 T_3]) \quad (17)$$

在该公式中, l_T 表示模型原本为候选词 T 分配的logit得分, δ_{att} 为正调整参数,用于控制评分函数 $s^*(T, [T_1 T_2 T_3])$ 对生成概率的扰动强度, l'_T 则是经过评分函数调整后的候选词概率。通过该评分函数,攻击者能够发起水印欺骗攻击,使普通文本被误判为携带特定水印;攻击者也可实施水印擦除攻击,从携带水印的文本中去除水印信号。实验表明,该方案针对红绿列表水印技术^[35]以低于50美元的成本实现概率超过0.8的伪造成功率。

为了在未授权知识蒸馏场景下同时对大语言模型水印实施擦除与欺骗双重黑盒攻击,Yi等人^[180]基于水印放射性现象原理提出了对比解码引导的知识提炼(Contrastive Decoding-guided Knowledge Distillation, CDG-KD)方法。该方法首先通过标准蒸馏获得强水印学生模型,再利用复述攻击构建弱水印参考模型,通过对比分析两者的输出分布差异

以精准调控水印相关 token 的输出概率。基于此,研究者分别构建了水印语料与强水印语料,并通过双路径知识蒸馏最终获得两个专用模型:可生成无水印文本的擦除模型与可生成带伪造水印文本的欺骗模型。实验表明,该方法在维持生成内容质量的基础上能够显著降低主流水印方案的检测置信度。

在专门针对水印欺骗攻击的研究中,Sadasivan 等人^[182]提出了面向 AI 文本检测器的欺骗攻击框架,该框架包含三种攻击方法。针对软水印型检测器,该方法通过高频查询统计 181 个常见词构成词对的共现频率,以此估计绿色词汇表,随后依据自建表手动伪造带水印文本。实验结果显示,该方法使人类撰写文本被水印检测器误判,其 AUROC 由 0.998 降低至 0.013。针对检索式检测器,该方法将人类撰写文本输入大语言模型获得语义相似的改写版本,并诱导后台数据库保存该改写内容,随后对人类撰写文本执行水印检测,库中保存的内容记录使其被检测为带有水印,该方法的攻击成功率达 100%。针对零样本与神经网络训练型检测器,该方法将被误判为高 AI 概率的人类参考内容作为前缀拼接至其他人类撰写文本中,将拼接内容输入检测器会显著提高被检测为 AI 生成内容的概率。

防御方面:应对模型伪造攻击的首要措施是实施合理的 API 访问频率限制。例如,谷歌针对 Gemini 等模型规定了标准 API 速率限制以防止滥用行为并维护系统安全。除此之外,Ural 等人^[181]提出了将学习证明(Proof-of-Learning, PoL)与基于特征水印相结合的双层验证架构。该防御系统包含两层验证机制:第一层验证为 PoL 验证,通过记录模型训练过程中的关键信息如模型权重、梯度等生成不可篡改的训练过程证明,防止恶意用户伪造训练过程;第二层验证则是水印验证,在训练初始阶段将水印信息嵌入模型的权重初始化阶段,在训练过程中同时更新权重与水印以实现二者的深度融合。验证阶段通过预定义的水印数据集检测模型的水印特征。这种双层验证机制确保了攻击者难以同时伪造 PoL 证明与水印数据集,从而能够有效抵御模型伪造攻击。实验结果表明,PoL 方法面对模型伪造攻击的成功检测率超过 90%,展现出出色的防御性能。

4.2 多媒体大模型水印攻击与防御

随着多媒体生成技术的广泛应用,大模型水印面临的攻击手段已从单一的信号处理层面扩展至模型语义与跨模态交互层面,表现出高度的对抗性特

征。攻击者不再局限于裁剪或噪声干扰等基础操作,而是利用模型提取与跨模态转换等方法,试图窃取模型资产或清除版权信息。面对日益严峻的安全挑战,防御策略正由被动检测向主动防御转变,侧重于实现水印特征与模型生成能力的深度结合、基于潜在空间的语义绑定以及跨模态特征的协同防御。本节将分别探讨图像、音频及多模态大模型场景下的特定攻击模式,并阐述相应的防御机制。

4.2.1 图像大模型水印攻击与防御

与传统多媒体内容的篡改不同,针对图像生成大模型的攻击呈现出更强的对抗性与隐蔽性,主要包括旨在获取模型知识产权的模型窃取攻击,以及试图擦除或伪造版权信息的移除与伪造攻击。在模型窃取与提取攻击方面,攻击者试图通过模拟输入输出映射来推断模型结构或直接复制功能。针对此类威胁,Szyller 等人^[183]提出的 DAWN 方案引入了动态对抗防御机制,该方案能够实时分析攻击者的访问模式,并根据其请求数据动态修改模型返回的预测结果,从而在不影响正常服务的前提下有效误导攻击者的模型提取行为。Charette 等人^[184]则提出了分布注入策略,通过在模型输出中注入特定的信号分布,使得窃取后的模型不可避免地学习到这种带有版权特征的分布,导致攻击者不仅无法去除水印,反而保留了原模型的版权标记。此外,Li 等人^[185]提出了强耦合训练策略,将模型的原始生成性能与水印模式进行深度绑定,使得模型在试图跳出含有水印的局部最优解时面临巨大的计算成本,从而从根本上抑制了伪造水印的可能性。

在应对移除与逃逸攻击方面,防御者的策略逐渐从被动检测转向主动防御。为抵御旨在破坏版权验证的恶意攻击,Wang 等人^[186]在训练过程中引入对抗训练机制,通过创建攻击模拟环境来约束模型的权重分布,提高了模型在面对参数微调或剪枝时的鲁棒性。Meng 等人^[187]针对生成模型的特性,提出了潜在空间渐进训练策略,通过在潜在空间层面进行水印的嵌入与检测,有效抵抗了针对图像生成的各类移除攻击。针对伪造与歧义攻击,Arabi 等人^[188]设计了结合傅里叶模式嵌入的两阶段检测框架,专门用于识别并抵御攻击者试图通过联合多个模型副本生成无标记模型的共谋攻击,以及通过伪造水印制造所有权归属混淆的行为。

4.2.2 音频大模型水印攻击与防御

与文本领域的释义攻击类似,音频水印同样面临着旨在破坏版权信息的智能诱导失真攻击,且其

破坏方式更为底层和彻底。该类攻击利用先进的语音转换模型或文本转语音模型对原始音频进行重构,在保持语音内容与听感风格基本一致的前提下,彻底抹除隐藏于波形或频谱中的水印信号。在攻击技术方面,Wen 等人^[170]的研究揭示了当前水印技术的严重脆弱性,实验表明,无论是变调、变速等基础信号级失真,还是模拟现实环境空气隙传播的物理级重录,均能导致绝大多数水印方案的检测准确率降至随机水平。更为致命的是,利用零样本或微调后的语音转换模型进行的清洗攻击,能够有效剥离几乎所有现有的水印保护,即便是针对特定风格迁移生成的音频,水印的残留率也微乎其微。

防御方面:为应对日益严峻的智能攻击威胁,单一的静态防御策略已难以奏效。现有的 Timbre 方案通过在训练阶段纳入多种失真场景,在对抗文本转语音微调攻击时展现出了相对较好的鲁棒性,但这仅是防御升级的第一步。针对当前水印一旦被破解即全盘失效的不可撤销性缺陷,未来的防御重点在于开发内容自适应水印与可修补水印框架。内容自适应策略要求水印生成机制能够根据每个音频样本的频率分布与节奏特征动态调整嵌入模式,使攻击者难以总结出通用的去除规律;而可修补框架则引入了一种允许定期更新的基础水印机制,随着新型攻击手段的出现,系统可以通过算法对已发布内容的检测逻辑打补丁,从而在长期的攻防博弈中维持版权保护的有效性。

4.2.3 多模态大模型水印攻击与防御

多模态大模型水印面临着比单模态更为复杂的对抗环境,其核心挑战在于跨模态转换攻击带来的信号清洗风险。与传统的单模态移除攻击不同,跨模态攻击利用模型在不同模态间语义对齐的能力,通过模态转换过程去除隐藏的水印信号。例如,攻击者可以将含有水印的图像输入多模态模型生成描述文本,再利用文生图模型重绘图像,或者将含有水印的音频转录为文本后再合成新的语音。针对此类威胁,Ren 等人^[189]指出了现有的单模态水印方案在跨模态语义迁移中的脆弱性,当含有水印的内容经历“图像转文本再转图像”或“音频转文本再转语音”的跨模态转换后,原始模态中嵌入的信号往往会被彻底清洗。为应对这一挑战,研究人员开始探索基于联合语义空间的防御策略。Zhang 等人^[190]提出了一种基于多模态联合嵌入的水印方案,该方案不再单一地将水印嵌入像素或波形层,而是将其注入到多模态模型的共享潜在空间中,使得水印信息能够

以此为载体在图像与其对应的文本描述或音频信号之间传递,从而抵御跨模态清洗攻击。

在更为激烈的对抗场景中,针对旨在破坏音视频同步性的多模态分割攻击,防御者的策略进一步演进为时空联合防御。攻击者往往通过分离视频中的音频轨道并进行独立处理,如对音频进行变调或对视频进行重编码,来破坏多模态水印的完整性。针对音视频内容可能遭受独立或联合篡改的问题,Zhang 等人^[191]提出了 V²A-Mark 音视频联合水印框架。该方法在原始视频帧和音频信号中嵌入视觉—音频定位水印与版权水印,并通过跨模态版权提取机制融合从视觉与音频中恢复的水印信息,从而实现视觉篡改区域、音频篡改区间的联合定位与版权验证。实验表明,该方法能够有效检测视频内容编辑与音频内容修改,并在常见视频和音频退化条件下保持较好的水印恢复与篡改定位性能。

5 大模型水印应用场景

随着人工智能技术从单一语言模型向多模态大模型的快速演进,生成式内容的真实性辨识、版权归属以及模型滥用风险等问题日益凸显。大模型水印技术在保护模型知识产权、确保模型使用安全合法等关键领域发挥着重要的作用^[192]。

5.1 知识产权保护

随着大语言模型向多模态大模型发展,模型知识产权保护涵盖的范围已从文本扩展至图像、音频及视频,如图 17 所示,主要涵盖三个方面:训练数据的版权保护、大模型本体的版权保护以及模型输出内容的版权保护。

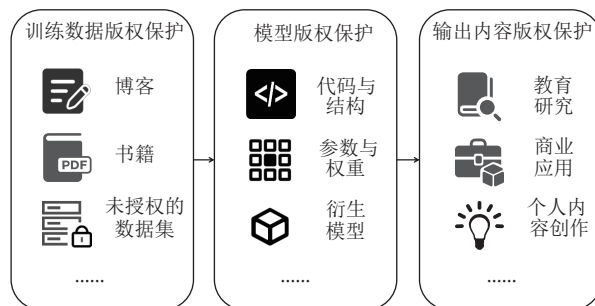


图 17 模型知识产权保护

5.1.1 训练数据的版权保护

随着大模型的快速发展,训练数据规模呈现出指数级增长,数据版权保护已引发学术界和产业界的高度关注。在模型训练环节,部分训练者可能有

意无意地使用未经授权的数据集甚至将受版权保护的博客、书籍、报纸以及高质量的图像和音频等作品纳入训练范围,严重侵犯了数据集持有者与版权所有者的合法权益。为了应对这一挑战,研究者提出了基于水印技术的解决方案^[153]。其核心内容是将包含所有权信息的隐蔽水印嵌入待保护的数据中,该水印随后被模型学习并内化于其生成的内容中。通过检测生成内容中的水印特征,数据所有者可获得侵权使用的有效证据,为数据版权保护提供了可靠的技术路径。

5.1.2 大模型的版权保护

大模型的训练过程需要投入大量时间、算力、数据和人力资源,成熟的商用大模型可为持有者创造可观的经济效益,因此对大模型本体包括代码与结构、参数与权重等核心内容的版权保护具有重要的研究意义和应用价值。用于商业用途的大模型通常会通过应用程序接口向用户提供服务,这种开放性使得大模型面临潜在的安全风险。通过模型提取攻击,攻击者可以利用模型的输出数据训练新的模型以近似取代原有模型的功能,严重损害原模型持有者的利益。为应对这一威胁,现有研究提出通过在模型内部嵌入包含模型所有者版权信息的水印,使得提取攻击复制出的新模型内部仍保留水印信息,从而延续模型版权的保护。

5.1.3 模型生成内容的版权保护

在大模型的应用场景中,用户利用特定的大模型生成内容并主张该内容为其独立创作且应享有相应的版权。此时,如何有效保护不同用户输出内容的版权归属成为一个关键问题。该问题的研究关键在于实现用户级的水印区分机制,而非简单地采用统一的大模型水印进行标注。通常情况下,特定大模型的用户在使用模型之前需通过特定账号或其他身份认证方式获取模型的使用权,基于此认证机制,水印技术可通过将唯一用户标识作为个性化水印嵌入生成内容中,为不同用户分配专属水印标识,从而为内容来源追踪和权属主张提供技术证据,建立用户级版权区分机制。

5.2 规范模型用途

模型安全合法使用涵盖军事、商业、学术等领域的多媒体内容安全管控。在军事与商业等需要严格安全保密的领域,向大模型嵌入独特的脆弱水印可有效确保模型安全使用;在学术研究领域,大模型水印技术能够有效防范学术不端行为的发生;在日常使用中,该技术还可用于检测内容是否由大模型生

成,并实现来源追踪与内容溯源。

5.2.1 军事商业模型保护

在商业^[193]和军事^[194]等关键领域,大模型的安全性至关重要。确保大模型的使用环境安全以及防止模型被恶意篡改是保障系统稳定性和信息安全的基础。通过嵌入特定水印信息,可实现对模型完整性和安全性的实时监控^[195]。这类水印具备易于检测和可修改的特性,以确保在正常运行时不影响模型的性能和功能,但在模型参数遭受细微篡改时水印信息会发生显著变化。若在模型使用前提取的水印信息与预设的标识一致,则表明模型完整性未受破坏;若水印信息无法通过验证,则表明模型可能遭受攻击,此时应立即启动排查与修复程序。定期验证水印信息的一致性,可及时识别潜在攻击,触发安全响应机制,确保关键应用场景下的系统安全性和可靠性。

5.2.2 学术诚信维护

大模型在教育领域的普及带来了学术诚信挑战。学生借助大模型快速完成课业任务,会影响教师对学生知识掌握程度的准确评估;研究者利用大语言模型的高效生成能力产出的低质量学术论文可能包含错误信息甚至存在抄袭嫌疑,严重损害学术研究的严谨性和公信力。针对上述问题,通过在大语言模型生成内容中嵌入水印^[196],可有效检测作业以及学术论文中是否过度依赖大语言模型。该方法能够帮助教育工作者和学术机构识别和防范大语言模型的滥用,维护学术评价的准确性与研究工作的原创性,保障教育和学术研究的高质量与公正性。在音频与视觉创作领域,水印技术有助于区分人类创作的艺术作品与人工智能生成的合成内容。这不仅可进行伪造检测,还能确立人机区分的边界,保护人类创作内容的价值。

5.2.3 虚假媒体内容治理

大模型大幅降低了多媒体内容创作的技术门槛,使得普通用户能够快速生成文本、音频、图片及视频并上传到社交媒体进行广泛传播^[197],具有潜在风险,尤其是恶意用户可能利用其制作虚假新闻^[198]、伪造公众人物虚假言论^[199]等并散布有害信息。为应对上述问题,可在大模型生成内容中嵌入具有高隐蔽性和强鲁棒性的水印,可有效追溯内容来源,识别模型生成的虚假信息。在深度伪造检测与内容治理方面,多模态水印为区分机器生成内容与人类创作内容提供了可靠的技术手段,特别是在应对深度伪造视频威胁时表现出独特优势。通过在

生成式视频的视觉层与听觉层植入具有时序同步特征的脆弱水印,任何针对面部的篡改或对语音的克隆替换都会破坏原有的时空同步特征,从而触发篡改警报。Wu等人^[200]展示了该技术在新闻真实性验证中的应用潜力,通过嵌入具有不同脆弱性的分离水印,使得平台不仅能追踪内容来源,还能通过验证脆弱水印的完整性,快速识别出经过大模型合成或篡改的虚假新闻内容,阻断其传播链条,保护公众免受误导,维护风清气正的晴朗网络空间。

6 未来发展方向

6.1 研究跨模态协同与组合防御体系

多模态水印技术已成为数字水印领域的重要研究方向之一。跨模态水印融合可以有效提高水印系统的综合性能。借鉴ForgeryGPT^[201]框架思路,该框架通过融合图像和文本等多模态信息,在伪造内容检测任务上取得了显著效果。现有水印技术大多局限于单一模态,未来研究可借鉴这种多模态融合的思想,开展跨模态协同嵌入机制的探索,结合图像内容、文本描述、用户交互等多维度信息,构建更加全面和智能的水印嵌入与检测系统。同时,应构建“鲁棒水印+脆弱水印”的组合防御体系^[202],前者用于版权追溯,后者用于篡改检测,为内容真实性提供全链路保障。

6.2 建立多中心治理框架与技术标准

大模型水印的落地应用不仅需要技术突破,更呼唤标准化的治理体系。单纯依靠模型厂商的自律难以解决跨平台、跨模态的监管难题。未来应探讨构建一个包含模型提供商、内容分发平台、第三方检测机构及监管部门在内的多中心治理框架。在此框架下,重点研究通用的水印协议和统一的评价体系^[203-204],规范水印的嵌入强度、容量及检测阈值。同时,需解决多重水印的冲突消解机制^[205],设计抗篡改的嵌入协议,确保在多用户、多轮次编辑的复杂环境中,各方权益能得到有效保护。

6.3 研究智能自适应水印系统

智能水印算法通过基于统一评价体系的智能评估与决策机制,自动基于不同应用场景选择最适合的水印嵌入算法,从而在可检测性、鲁棒性、隐蔽性和生成质量等关键指标间实现动态平衡。相较于现有人工比对和性能评估以选择合适水印的流程,智能水印系统能依据需求自主选择差异化水印算法。对于需要高鲁棒性的应用场景,如内容经多次编辑

或转换后仍需保持水印的有效性,应选择抗攻击性强的水印算法;而对于隐蔽性要求较高的应用场景则优先选择不可感知性优异的水印算法。这种智能化的算法调度机制将有效提高水印系统的整体工作效率。然而,智能水印算法系统的实现面临若干关键挑战,比如如何构建一个全面而科学的水印评价体系以准确反映各种应用场景对水印性能的不同需求,针对复杂应用场景的差异化算法选择机制仍需进一步优化,以确保系统在处理复杂任务时能作出准确而高效的决策等。这些问题的解决将推动智能水印系统从理论构想走向实际应用。

6.4 发展高容量个性化定制水印

目前高容量水印中已实现多比特信息的有效嵌入,以Yoo等人^[47]提出的多比特水印技术为例,该技术不仅能携带多比特信息,还能够对内容来源进行追溯。然而,随着大语言模型用户规模的持续扩大,模型所有者需进一步提升水印个性化能力与安全性,以满足不同用户在身份区分与版权保护方面日益增长的需求。未来应在现有高容量水印基础上,重点研究如何在控制生成质量影响的前提下,协同提升水印嵌入容量、用户定制能力与系统安全性,以实现真正适用于大规模部署的个性化水印体系。

6.5 探究水印与模型对齐的冲突与协同

目前大语言模型的安全主要依赖于基于人类反馈的强化学习等对齐技术,而水印嵌入本质上是对模型原始行为的扰动。这一领域面临着至关重要的挑战:水印的嵌入是否会破坏模型与人类价值观的对齐,导致模型拒绝回答能力的下降或产生有毒内容的风险?更为严峻的是,攻击者可能利用水印注入机制作为攻击向量,通过逆向工程水印触发模式来构建越狱攻击,绕过模型的防御机制。因此,未来必须深入研究水印与对齐机制的交互边界,开发对齐友好型水印算法,确保在嵌入版权信息的同时,不损害模型的安全性与有用性。

6.6 构建可解释与可验证的水印信息

现有的文本水印检测主要依赖于Z-score等统计学指标,这种黑盒式的判别结果缺乏直观的可解释性,难以在法律诉讼或公众舆论中作为直观证据。未来研究应致力于从统计信号向语义证据转型,发展具有可解释性的水印技术。例如,探索将水印编码为特定的句法结构、逻辑模式或语义风格,使得检测结果不仅能输出置信度,还能定位具体的水印特征片段,让人类观察者也能理解水印存在的逻辑证据。此外,结合零知识证明技术,在不泄露水印

私钥的前提下向公众证明模型所有权,也是提升水印技术公信力的重要方向。

6.7 探索水印与差分隐私的协同机制

数据隐私与版权保护是数据要素流通的两大核心诉求。差分隐私(Differential Privacy, DP)通过在数据中引入噪声来掩盖个体信息,而水印技术则通过引入信号来标识来源,两者在数学原理上存在潜在的对抗性。这是一个极具前沿性的交叉研究点:一方面,需评估差分隐私的噪声注入是否会破坏训练数据中水印的鲁棒性;另一方面,可探索两者的协同机制,例如利用水印技术标记经过差分隐私处理的数据集,或开发隐私保留型水印,在保护用户隐私的同时实现对模型调用的溯源,实现隐私保护与版权追溯的双赢。

7 结 论

大模型的快速发展引起了关于模型知识产权保护与模型滥用等方面的担忧,该领域的水印技术作为一种新兴解决方案,为模型版权验证与生成内容溯源提供了有效途径。本文系统梳理了大模型水印技术的研究进展,根据载体载体的不同,将其划分为大语言模型水印与多媒体大模型水印,并对各类水印技术最新的研究成果进行了归类与评述。在此基础上,本文进一步阐述了大模型水印的关键评估指标,并针对水印的攻击与对应防御技术进行了详细的分类与对比。尽管当前大模型水印技术在实际的生产应用中已经取得显著成效并持续受到学术界和工业界的广泛关注,但还存在许多挑战:多模态融合水印尚不成熟、水印冲突问题难以解决、统一的水印性能评估体系亟待构建等。未来仍需针对水印技术开展更为广泛且深入的研究,以应对上述挑战。

希望本文能够为投身大模型水印技术领域的研究者提供借鉴与参考,推动更具实用价值与先进性水印方案的出现,进而促进大模型在社会应用中的可信发展与规范治理。

参 考 文 献

- [1] Yao Yifan, Duan Jinhao, Xu Kaidi, et al. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 2024, 4(2): 100211
- [2] Radford Alec, Wu Jeffrey, Child Rewon, et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019, 1(8): 9
- [3] Achiam Josh, Adler Steven, Agarwal Sandhini, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023
- [4] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020, 33: 1877-1901
- [5] Chen Xuanning, Ye Junjie, Zu Can, et al. Robustness of GPT series large language models in natural language processing tasks. *Journal of Computer Research and Development*, 2024, 61(5): 1128-1142 (in Chinese)
(陈炫婷, 叶俊杰, 祖璨, 等. GPT 系列大语言模型在自然语言处理任务中的鲁棒性. *计算机研究与发展*, 2024, 61(5): 1128-1142)
- [6] Touvron Hugo, Martin Louis, Stone Kevin, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023
- [7] Bi Xiao, Chen Deli, Chen Guanting, et al. DeepSeek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024
- [8] Lu Haoyu, Liu Wen, Zhang Bo, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024
- [9] Masri Sari, Raddad Yaqeen, Khandaqi Fidaa, et al. Transformer models in education: Summarizing science textbooks with arabart, MT5, arat5, and mBART//*Proceedings of the International Conference on Intelligent Systems, Blockchain, and Communication Technologies*. Sharm El-Sheikh, Egypt, 2024: 286-300
- [10] Wang Jianing, Wu Junda, Hou Yupeng, et al. Instructgraph: boosting large language models via graph-centric instruction tuning and preference Alignment// *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand, 2024: 13492-13510
- [11] Guan Batu, Wan Yao, Bi Zhangqian, et al. Codeip: a grammar-guided multi-bit watermark for large language models of code// *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, USA, 2024: 9243-9258
- [12] Zhang Ruisi, Javidnia Neusha, Sheybani Nojan, et al. Robust and secure code watermarking for large language models via ML/Crypto codesign. *arXiv preprint arXiv:2502.02068*, 2025
- [13] Liu Yang, Cao Jiahuan, Liu Chongyu, et al. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*, 2024
- [14] Gao Jianfeng, Galley Michel, Li Lihong. Neural approaches to conversational AI// *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. New York, USA, 2018: 1371-1374
- [15] Wang Shen, Xu Tianlong, Li Hang, et al. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*, 2024
- [16] Luo Ziming, Yang Zonglin, Xu Zexin, et al. LLM4SR: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*, 2025
- [17] Huang Jimin, Xiao Mengxi, Li Dong, et al. Open-finllms: open multimodal large language models for financial applications.

- arXiv preprint arXiv:2408.11878, 2024
- [18] Huang Yining, Tang Keke, Chen Meilian, et al. A comprehensive survey on evaluating large language model applications in the medical industry. arXiv preprint arXiv: 2404.15777, 2024
- [19] Zhang Changlin, Tong Xin, Tong Hui, et al. A review of large language model technologies for network security. *Information Network Security*, 2024, 24(5): 778-793 (in Chinese)
(张长琳 仝鑫, 佟晖, 等. 面向网络安全领域的大语言模型技术综述. *信息安全*, 2024, 24(5): 778-793)
- [20] Li Jiao, Zhang Yuqing, Wu Yabiao. A data augmentation method for large language models for relation extraction in network security. *Information Network Security*, 2024, 24(10): 1477-1483 (in Chinese)
(李娇 张玉清, 吴亚彪. 面向网络安全关系抽取的大语言模型数据增强方法. *信息安全*, 2024, 24(10): 1477-1483)
- [21] Liang Yuqing, Xiao Jiancheng, Gan Wensheng, et al. Watermarking techniques for large language models: A survey. arXiv preprint arXiv:2409.00089, 2024
- [22] Wu Hanzhou, Zhang Jie, Li Yue, et al. Research progress on watermarking using artificial intelligence models. *Journal of Image and Graphics*, 2023, 28(6): 1792-1810 (in Chinese)
(吴汉舟, 张杰, 李越, 等. 人工智能模型水印研究进展. *中国图象图形学报*, 2023, 28(6): 1792-1810)
- [23] Liu Aiwei, Pan Leyi, Lu Yijian, et al. A survey of text watermarking in the era of large language models. *ACM Comput. Surv.*, 2024, 57(2): Article 47
- [24] Jin Biao, Lin Xiang, Xiong Jinbo, et al. Intellectual property protection of deep neural network models based on watermarking technology. *Journal of Computer Research and Development*, 2024, 61(10): 2587-2606 (in Chinese)
(金彪, 林翔, 熊金波, 等. 基于水印技术的深度神经网络模型知识产权保护. *计算机研究与发展*, 2024, 61(10): 2587-2606)
- [25] Qin Zhongyuan, Wang Tiantian, Liu Weiqiang, et al. Research progress on watermarking technology for large language models. *Information Network Security*, 2025, 25(2): 177-193 (in Chinese)
(秦中元 王田田, 刘伟强, 等. 大语言模型水印技术研究进展. *信息安全*, 2025, 25(2): 177-193)
- [26] Baldassini Folco Bertini, Nguyen Huy H, Chang Ching-Chung, et al. Cross-attention watermarking of large language models// *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Republic of Korea, 2024: 4625-4629
- [27] Pan Leyi, Liu Aiwei, He Zhiwei, et al. MarkLLM: an open-source toolkit for llm watermarking// *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Miami, USA, 2024: 61-71
- [28] Li Shuai, Chen Kejiang, Tang Kunsheng, et al. Turning your strength into watermark: watermarking large language model via knowledge injection. arXiv preprint arXiv:2311.09535, 2023
- [29] Shetty Anudeex, Teng Yue, He Ke, et al. WARDEN: multi-directional backdoor watermarks for embedding-as-a-Service copyright protection// *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand, 2024: 13430-13444
- [30] Lou Jian, Zhang Chenyang, Zhang Xiaoyu, et al. PreferCare: preference dataset copyright protection in llm alignment by watermark injection and verification// *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*. New York, USA, 2025: 4005-4019
- [31] Li Linyang, Jiang Botian, Wang Pengyu, et al. Watermarking llms with weight quantization// *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore, 2023: 3368-3378
- [32] Block Adam, Sekhari Ayush, Rakhlin Alexander. GaussMark: a practical approach for structural watermarking of language models. arXiv preprint arXiv:2501.13941, 2025
- [33] Xu Xiaojun, Yao Yuanshun, Liu Yang. Learning to watermark llm-generated text via reinforcement learning. arXiv preprint arXiv:2403.10553, 2024
- [34] Guo Junfeng, Li Yiming, Chen Ruibo, et al. Zeromark: towards dataset ownership verification without disclosing watermark// *Advances in Neural Information Processing Systems*. Red Hook, USA, 2024, 37: 120468-120500
- [35] Kirchenbauer John, Geiping Jonas, Wen Yuxin, et al. A watermark for large language models// *Proceedings of the International Conference on Machine Learning*. Honolulu, USA, 2023: 17061-17084
- [36] Zhao Xuandong, Ananth Prabhanjan, Li Lei, et al. Provable robust watermarking for ai-generated text. arXiv preprint arXiv: 2306.17439, 2023
- [37] Hu Zhengmian, Chen Lichang, Wu Xidong, et al. Unbiased watermark for large language models. arXiv preprint arXiv: 2310.10669, 2023
- [38] Ruibo Chen, Yihan Wu, Junfeng Guo, Heng Huang. Improved unbiased watermark for large language models// *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria. 2025: 20587-20601
- [39] Li Zhuang. BiMarker: Enhancing text watermark detection for large language models with bipolar watermarks. arXiv preprint arXiv:2501.12174, 2025
- [40] Liang Chen, Bian Yatao, Deng Yang, et al. WatME: towards lossless watermarking through lexical redundancy// *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Bangkok, Thailand, 2024: 9166-9180
- [41] Wu Yihan, Hu Zhengmian, Guo Junfeng, et al. A resilient and accessible distribution-preserving watermark for large language models. arXiv preprint arXiv:2310.07710, 2023
- [42] Li Yuhang, Wang Yihan, Shi Zhouxing, et al. Improving the generation quality of watermarked large language models via word importance scoring. arXiv preprint arXiv: 2311.09668, 2023
- [43] Nemecek Alexander, Jiang Yuzhou, Ayday Erman. Topic-based watermarks for llm-generated text. arXiv preprint arXiv: 2404.02138, 2024
- [44] Dathathri S., See A., Ghaisas S., et al. Scalable watermarking for identifying large language model outputs. *Nature*, 2024, 634

- (8035): 818-823
- [45] Pang Kaiyi, Qi Tao, Wu Chuhan, et al. ModelShield: adaptive and robust watermark against model extraction attack. *IEEE Transactions on Information Forensics and Security*, 2025, 20: 1767-1782
- [46] Zhu Chaoyi, Galjaard Jeroen, Chen Pin-Yu, et al. Duwak: dual watermarks in large language models. *arXiv preprint arXiv: 2403.13000*, 2024
- [47] Yoo Kiyoon, Ahn Wonhyuk, Kwak Nojun. Advancing beyond identification: multi-bit watermark for large language models// *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Mexico City, Mexico, 2024: 4031-4055
- [48] Ren Jie, Xu Han, Liu Yiding, et al. A robust semantics-based watermark for large language model against paraphrasing// *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico, 2024: 613-625
- [49] Jiang Ya, Wu Chuxiong, Boroujeny Massieh Kordi, et al. StealthInk: a multi-bit and stealthy watermark for large language models. *arXiv preprint arXiv:2506.05502*, 2025
- [50] Giboulot Eva, Furon Teddy. WaterMax: breaking the llm watermark detectability-robustness-quality trade-off// *Advances in Neural Information Processing Systems*. Vancouver, Canada, 2024, 37: 18848-18881
- [51] Hou Abe Bohan, Zhang Jingyu, He Tianxing, et al. Semstamp: a semantic watermark with paraphrastic robustness for text generation. *arXiv preprint arXiv:2310.03991*, 2023
- [52] Hou Abe Bohan, Zhang Jingyu, Wang Yichen, et al. k-semstamp: a clustering-based semantic watermark for detection of machine-generated text// *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand, 2024: 1706-1715
- [53] Zhang Yuehan, Lv Peizhuo, Liu Yinpeng, et al. PersonaMark: personalized llm watermarking for model protection and user attribution. *arXiv preprint arXiv:2409.09739*, 2024
- [54] Peng Wenjun, Yi Jingwei, Wu Fangzhao, et al. Are you copying my model? protecting the copyright of large language models for eaas via backdoor watermark// *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Toronto, Canada, 2023: 7653-7668
- [55] Sato Ryoma, Takezawa Yuki, Bao Han, et al. Embarrassingly simple text watermarks. *arXiv preprint arXiv:2310.08920*, 2023
- [56] Dabiriaghdam Amirhossein, Wang Lele. SimMark: a robust sentence-level similarity-based watermarking algorithm for large language models. *arXiv preprint arXiv:2502.02787*, 2025
- [57] Yu Zhuohao, Jiang Xingru, Gu Weizheng, et al. SAEMark: Multi-bit llm watermarking with inference-time scaling. *arXiv preprint arXiv:2508.08211*, 2025
- [58] Chang Yapei, Krishna Kalpesh, Houmansadr Amir, et al. PostMark: a robust blackbox watermark for large language models// *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, USA, 2024: 8969-8987
- [59] Lau Gregory Kang Ruey, Niu Xinyuan, Dao Hieu, et al. Waterfall: scalable framework for robust text watermarking and provenance for llms// *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, USA, 2024: 20432-20466
- [60] Yang Borui, Li Wei, Xiang Liyao, et al. Srcmarker: dual-channel source code watermarking via scalable code transformations// *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP)*. San Francisco, USA, 2024: 4088-4106
- [61] Li Boquan, Zhang Mengdi, Zhang Peixin, et al. Resilient watermarking for llm-generated codes. *arXiv e-prints*, 2024: arXiv: 2402.07518
- [62] Li Shen, Yao Liuyi, Gao Jinyang, et al. Double-i watermark: protecting model copyright for llm fine-tuning, *arXiv preprint arXiv:2402.14883*, 2024
- [63] Chen Haoran, Liu Yu, Chen Ping. An intrinsic security heterogeneity generation method based on large language model. *Information Network Security*, 2024, 24(8): 1231-1240 (in Chinese)
(陈昊然 刘宇, 陈平. 基于大语言模型的内生安全异构体生成方法. *信息安全*, 2024, 24(8): 1231-1240)
- [64] Wang Xutong, Yin Jie, Liu Chaoge et al. A review of backdoor attacks and defenses in neural networks. *Chinese Journal of Computers*, 2024, 47(8):1713-1743 (in Chinese)
(汪旭童, 尹捷, 刘潮歌等. 神经网络后门攻击与防御综述. *计算机学报*, 2024, 47(8):1713-1743)
- [65] Cui Xinyue, Wei Johnny, Swayamdipta Swabha, et al. Robust data watermarking in language models by injecting fictitious knowledge// *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria, 2025: 14292-14306
- [66] Rastogi Saksham, Maini Pratyush, Pruthi Danish. Stamp your content: proving dataset membership via watermarked rephrasings. *arXiv preprint arXiv:2504.13416*, 2025
- [67] Sander Tom, Fernandez Pierre, Durmus Alain, et al. Watermarking makes language models radioactive// *Advances in Neural Information Processing Systems*. Vancouver, Canada, 2024, 37: 21079-21113
- [68] Yao Hongwei, Lou Jian, Qin Zhan, et al. Promptcare: Prompt copyright protection by watermark injection and verification// *2024 IEEE Symposium on Security and Privacy (SP)*. San Francisco, USA, 2024: 845-861
- [69] Lv Peizhuo, Li Pan, Zhu Shenchun, et al. Ssl-wm: A black-box watermarking approach for encoders pre-trained by self-supervised learning. *arXiv preprint arXiv:2209.03563*, 2022
- [70] Lv Peizhuo, Sun Mengjie, Wang Hao, et al. Rag-wm: An efficient black-box watermarking approach for retrieval-augmented generation of large language models// *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*. New York, USA, 2025: 1709-1723
- [71] Fan Wei, Yao Qiyi, Qiu Jing, et al. Vector database watermarking//*Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems*. San Diego, USA, 2025

- [72] Tian Xuan, Wu Zhichao. A review of knowledge base question answering based on information retrieval. *Journal of Computer Research and Development*, 2025, 62(2): 314-335 (in Chinese) (田萱, 吴志超. 基于信息检索的知识库问答综述. *计算机研究与发展*, 2025, 62(2): 314-335)
- [73] Ren Yubing, Guo Ping, Cao Yanan, et al. Subtle signatures, strong shields: Advancing robust and imperceptible watermarking in large language models// *Findings of the Association for Computational Linguistics ACL 2024*. Bangkok, Thailand, 2024: 5508-5519
- [74] Hao Jifei, Qiang Jipeng, Zhu Yi, et al. Post-hoc watermarking for robust detection in text generated by large language models// *Proceedings of the 31st International Conference on Computational Linguistics*. Abu Dhabi, UAE, 2025: 5430-5442
- [75] Huo Mingjia, Somayajula Sai Ashish, Liang Youwei, et al. Token-specific watermarking with enhanced detectability and semantic coherence for large language models// *Proceedings of the 41st International Conference on Machine Learning*. Vienna, Austria, 2024: Article 833
- [76] Lee Taehyun, Hong Seokhee, Ahn Jaewoo, et al. Who wrote this code? watermarking for code generation// *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand, 2023: 4890-4911
- [77] Gu Tianle, Wang Zongqi, Huang Kexin, et al. Invisible entropy: towards safe and efficient low-entropy llm watermarking. *arXiv preprint arXiv:2505.14112*, 2025
- [78] Tsur Dor, Long Carol Xuan, Verdun Claudio Mayrink, et al. Heavywater and simplexwater: distortion-free llm watermarks for low-entropy distributions//*Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems*. San Diego, USA, 2025
- [79] Mao Minjia, Wei Dongjun, Chen Zeyu, et al. Watermarking large language models: An unbiased and low-risk method// *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria, 2025: 7939-7960
- [80] Huang Beining, Su Du, Sun Fei, et al. Low-entropy watermark detection via bayes' rule derived detector// *In Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria, 2025: 14330-14344
- [81] Liu Aiwei, Pan Leyi, Hu Xuming, et al. A semantic invariant robust watermark for large language models. *arXiv preprint arXiv:2310.06356*, 2023
- [82] He Zhiwei, Zhou Binglin, Hao Hongkun, et al. Can watermarks survive translation? on the cross-lingual consistency of text watermark for large language models// *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Bangkok, Thailand, 2024: 4115-4129
- [83] Wang Zongqi, Gu Tianle, Wu Baoyuan, et al. Morphmark: flexible adaptive watermarking for large language models. *arXiv preprint arXiv:2505.11541*, 2025
- [84] Wang Yidan, Ren Yubing, Cao Yanan, et al. From trade-off to synergy: a versatile symbiotic watermarking framework for large language models. *arXiv preprint arXiv:2505.09924*, 2025
- [85] Huang Baizhou, Wan Xiaojun. Waterpool: a language model watermark mitigating trade-offs among imperceptibility, efficacy and robustness// *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Albuquerque, New Mexico, 2025: 4156-4182
- [86] Feng Xiaoyan, Zhang He, Zhang Yanjun, et al. Bimark: unbiased multilayer watermarking for large language models. *arXiv preprint arXiv:2506.21602*, 2025
- [87] He Haiyun, Liu Yepeng, Wang Ziqiao, et al. Theoretically grounded framework for llm watermarking: a distribution-adaptive approach. *arXiv preprint arXiv:2410.02890*, 2024
- [88] Wang Chenrui, Shu Junyi, Chiu Billy, et al. Learning to watermark: a selective watermarking framework for large language models via multi-objective optimization. *arXiv preprint arXiv:2510.15976*, 2025
- [89] Moitra Ankur, Golowich Noah. Edit distance robust watermarks for language models// *Proceedings of the 38th International Conference on Neural Information Processing Systems*. Vancouver, Canada, 2024: 20645-20693
- [90] Shen Huanming, Huang Baizhou, Wan Xiaojun. Enhancing llm watermark resilience against both scrubbing and spoofing attacks. *arXiv preprint arXiv:2507.06274*, 2025
- [91] Qu Wenjie, Zheng Wengrui, Tao Tianyang, et al. Provably robust multi-bit watermarking for {AI-generated} text// *Proceedings of the 34th USENIX Security Symposium (USENIX Security 25)*. Seattle, USA, 2025: 201-220
- [92] Zhao Xuandong, Li Lei, Wang Yu-Xiang. Permute-and-flip: an optimally stable and watermarkable decoder for llms. *arXiv preprint arXiv:2402.05864*, 2024
- [93] Chen Ruibo, Wu Yihan, Chen Yanshuo, et al. A watermark for order-agnostic language models. *arXiv preprint arXiv:2410.13805*, 2024
- [94] Liu Yepeng, Bu Yuheng. Adaptive text watermark for large language models// *Proceedings of the 41st International Conference on Machine Learning*. Vienna, Austria, 2024: Article 1238
- [95] Wong Kahim, Zhou Jicheng, Zhou Jiantao, et al. An end-to-end model for logits based large language models watermarking. *arXiv preprint arXiv:2505.02344*, 2025
- [96] Cohen Aloni, Hoover Alexander, Schoenbach Gabe. Watermarking language models for many adaptive users// *Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP)*. San Francisco, USA, 2025: 2583-2601
- [97] Liu Aiwei, Pan Leyi, Hu Xuming, et al. An unforgeable publicly verifiable watermark for large language models. *arXiv preprint arXiv:2307.16230*, 2023
- [98] Wang Lean, Yang Wenkai, Chen Deli, et al. Towards codable watermarking for injecting multi-bits information to llms. *arXiv preprint arXiv:2307.15992*, 2023
- [99] Shetty Anudeex, Xu Qionghai, Lau Jey Han. Wet: Overcoming paraphrasing vulnerabilities in embeddings-as-a-service with

- linear transformation watermarks// Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria, 2025: 23024-23043
- [100] Wang Zongqi, Wu Baoyuan, Deng Jingyuan, et al. Robust and minimally invasive watermarking for EaaS// Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria, 2025: 2167-2191
- [101] Wang Liaoyaqi, Cheng Minhao. GuardEmb: dynamic watermark for safeguarding large language model embedding service against model stealing attack// Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, USA, 2024: 7518-7534
- [102] Morteza Bashardoost, Mohd Shafry Mohd Rahim, Tanzila Saba, et al. Replacement attack: a new zero text watermarking attack. 3D Research, 2017, 8(1): 8
- [103] Xu Xiaojun, Jia Jinghan, Yao Yuanshun, et al. Robust multi-bit text watermark with llm-based paraphrasers, arXiv preprint arXiv:2412.03123, 2024
- [104] Li Zongjie, Wang Chaozheng, Wang Shuai, et al. Protecting intellectual property of large language model-based code generation apis via watermarks// Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. New York, USA, 2023: 2336-2350
- [105] Guo Zhimeng, Cheng Minhao. Practical and effective code watermarking for large language models// Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems. San Diego, USA, 2025
- [106] Zhang Boyu, He Ping, Du Tianyu, et al. CLMTracing: black-box user-level watermarking for code language model tracing// Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025: 28962-28978
- [107] Zhao Lei, Gui Xiaolin, Shao Yiyang, et al. A review of multifunctional watermarking for digital images. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(2): 195-222 (in Chinese)
(赵蕾, 桂小林, 邵屹杨, 等. 数字图像多功能水印综述. 计算机辅助设计与图形学学报, 2024, 36(2): 195-222)
- [108] Li Yue, Wang Hongxia, Barni Mauro. A survey of deep neural network watermarking techniques. Neurocomputing, 2021, 461: 171-193
- [109] Uchida Yusuke, Nagai Yuki, Sakazawa Shigeyuki, et al. Embedding watermarks into deep neural networks. ACM, 2017: 269-277
- [110] Wang Jiangfeng, Wu Hanzhou, Zhang Xinpeng, et al. Watermarking in deep neural networks via error back-propagation. Electronic Imaging, 2020, 32(4): 22-1-22-9
- [111] Liu Hanwen, Weng Zhenyu, Zhu Yuesheng. Watermarking deep neural networks with greedy residuals// Proceedings of the International Conference on Machine Learning. 2021: 6978-6988
- [112] Tartaglione Enzo, Grangetto Marco, Cavagnino Davide, et al. Delving in the loss landscape to embed robust watermarks into neural networks// 2020 25th International Conference on Pattern Recognition (ICPR). 2021: 1243-1250
- [113] Fan Lixin, Ng Kam Woh, Chan Chee Seng. Rethinking deep neural network ownership verification: embedding passports to defeat ambiguity attacks// Advances in Neural Information Processing Systems 32, Volume 6 of 20: 32nd Conference on Neural Information Processing Systems (NeurIPS 2019). Vancouver, Canada, 2020: 4714-4723
- [114] Rouhani Bitar Darvish, Chen Huili, Koushanfar Farinaz. DeepSigns: an end-to-end watermarking framework for ownership protection of deep neural networks//Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'19). New York, USA, 2019: 485-497
- [115] Ngo Anh Tu, Heng Chuan Song, Chattopadhyay Nandish, et al. Persistence of backdoor-based watermarks for neural networks: a comprehensive evaluation. arXiv preprint arXiv: 2501.02704, 2025
- [116] Leroux Sam, Vanassche Stijn, Simoens Pieter. Multi-bit, black-box watermarking of deep neural networks in embedded applications// Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA, 2024: 2121-2130
- [117] Adi Yossi, Baum Carsten, Cisse Moustapha, et al. Turning your weakness into a strength: watermarking deep neural networks by backdooring// Proceedings of the 27th USENIX security symposium (USENIX Security 18). Baltimore, USA, 2018: 1615-1631
- [118] Abdelaziz Ahmed, Fathi Ahmed, Fares Ahmed. Protecting intellectual property of eeg-based neural networks with watermarking. arXiv preprint arXiv:2502.05931, 2025
- [119] Sun Shichang, Xue Mingfu, Wang Jian, et al. Protecting the intellectual properties of deep neural networks with an additional class and steganographic images. arXiv preprint arXiv: 2104.09203, 2021
- [120] Ye Zhicheng, Zhang Xinpeng, Feng Guorui. Deep neural networks watermark via universal deep hiding and metric learning. Neural Computing and Applications, 2024, 36(13): 7421-7438
- [121] Zhang Jialong, Gu Zhongshu, Jang Jiyong, et al. Protecting intellectual property of deep neural networks with watermarking// Proceedings of the 2018 on Asia Conference on Computer and Communications Security (ASIACCS '18). New York, USA, 2018: 159-172
- [122] Li Yuxuan, Maharana Sarthak Kumar, Guo Yunhui. Not just change the labels, learn the features: watermarking deep neural networks with multi-view Data// Proceedings of the European Conference on Computer Vision. Milan, Italy, 2025: 273-289
- [123] Huang Chaoyue, Wu Hanzhou. A game between the defender and the attacker for trigger-based black-box model watermarking. arXiv preprint arXiv:2501.01194, 2025
- [124] Wu Hanzhou, Liu Gen, Yao Yuwei, et al. Watermarking neural networks with watermarked images. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(7): 2591-2601
- [125] Xu Hengyuan, Xiang Liyao, Ma Xingjun, et al. Hufu: a

- modality-agnostic watermarking system for pre-trained transformers via permutation equivariance. arXiv preprint arXiv: 2403.05842, 2024
- [126] Cui Yingqian, Ren Jie, Lin Yuping, et al. Ft-shield: a watermark against unauthorized fine-tuning in text-to-image diffusion models. ACM SIGKDD Explorations Newsletter, 2025, 26(2): 76-88
- [127] Zhao Yuan, Liu Bo, Zhu Tianqing, et al. Proactive image manipulation detection via deep semi-fragile watermark. Neurocomputing, 2024, 585: 127593
- [128] Yu Ning, Skripniuk Vladislav, Abdelnabi Sahar, et al. Artificial fingerprinting for generative models: rooting deepfake attribution in training data// Proceedings of the International Conference on Computer Vision. 2021: 14448-14457
- [129] Cui Yingqian, Ren J., Xu Han, et al. DiffusionShield: a watermark for copyright protection against generative diffusion models. SIGKDD Explor. Newsl. 2024: 60-75
- [130] Ma Yihan, Zhao Zhengyu, He Xinlei, et al. Generative watermarking against unauthorized subject-driven image synthesis. arXiv preprint arXiv:2306.07754, 2023
- [131] Chen Weitong, Li Yuheng. RoWSFormer: A robust watermarking framework with swin transformer for enhanced geometric attack resilience. arXiv preprint arXiv: 2409.14829, 2024
- [132] Gunn Sam, Zhao Xuandong, Song Dawn. An undetectable watermark for generative image models. arXiv preprint arXiv: 2410.07369, 2024
- [133] He Zhouyan, Hu Renzhi, Wu Jun, et al. A transformer-based invertible neural network for robust image watermarking. J. Vis. Comun. Image Represent., 2024, 104(C): 10
- [134] Aberna P., Agilandeswari L., Aashish Bansal. Vision transformer-based watermark generation for authentication and tamper detection using schur decomposition and hybrid transforms. International Journal of Computer Information Systems and Industrial Management Applications, 2023, 15: 15
- [135] Lukas Nils, Kerschbaum Florian. {PTW}: Pivotal tuning watermarking for {Pre-Trained} image generators// Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23). Anaheim, USA, 2023: 2241-2258
- [136] Zhang Hanlin, Edelman Benjamin L., Francati Danilo, et al. Watermarks in the sand: impossibility of strong watermarking for generative models. arXiv preprint arXiv:2311.04378, 2023
- [137] Tong Chuxuan, Natgunanathan Iynkaran, Xiang Yong, et al. Enhancing robustness of speech watermarking using a transformer-based framework exploiting acoustic features. IEEE/ACM Transactions on Audio, Speech and Lang, 2024, 32: 4822-4837
- [138] Roman Robin San, Fernandez Pierre, Elsahar Hady, et al. Proactive detection of voice cloning with localized watermarking// Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria, 2024: Article 1759
- [139] Liu Yixin, Lu Lie, Jin Jihui, et al. XAttnMark: learning robust audio watermarking with cross-attention. arXiv preprint arXiv: 2502.04230, 2025
- [140] Li Pengcheng, Zhang Xulong, Xiao Jing, et al. IDEAW: robust neural audio watermarking with invertible dual-embedding. arXiv preprint arXiv:2409.19627, 2024
- [141] Chen Guangyu, Wu Yu, Liu Shujie, et al. WavMark: watermarking for audio generation. arXiv preprint arXiv: 2308.12770, 2023
- [142] Epple Pascal, Shilov Igor, Stevanoski Bozhidar, et al. Watermarking training data of music generation models. arXiv preprint arXiv:2412.08549, 2024
- [143] Qiu Jielin, Han William, Zhao Xuandong, et al. Evaluating durability: benchmark insights into multimodal watermarking. arXiv preprint arXiv:2406.03728, 2024
- [144] Nie Hewang, Lu Songfeng. Securing IP in edge AI: neural network watermarking for multimodal models. Applied Intelligence, 2024, 54(21): 10455-10472
- [145] Radford Alec, Kim Jong Wook, Hallacy Chris, et al. Learning transferable visual models from natural language supervision// Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research. 2021: 8748-8763
- [146] Wan Shan, Liu Wu, Liu Yijun, et al. Watermarking vision-language models// Proceedings of the 6th ACM International Conference on Multimedia in Asia. Auckland, New Zealand, 2024: Article 119
- [147] Tang Yuanmin, Yu Jing, Gai Keke, et al. Watermarking vision-language pre-trained models for multi-modal embedding as a service. arXiv preprint arXiv:2311.05863, 2023
- [148] Ye Guanhuai, Gao Jiashi, Wang Yuchen, et al. ItoV: efficiently adapting deep learning-based image watermarking to video watermarking// 2023 International Conference on Culture-Oriented Science and Technology (CoST). Xi'an, China, 2023: 192-197
- [149] Guo Bofei, Tai Haoxuan, Luo Guibo, et al. AVSecure: an audio-visual watermarking framework for proactive deepfake detection// 2024 IEEE 14th International Conference on Electronics Information and Emergency Communication (ICEIEC). Beijing, China, 2024: 1-4
- [150] Fernandez Pierre, Elsahar Hady, Yalniz I Zeki, et al. Video seal: open and efficient video watermarking. arXiv preprint arXiv:2412.09492, 2024
- [151] Lu Yijian, Liu Aiwei, Yu Dianzhi, et al. An entropy-based text watermarking detection method//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, 2024: 11724-11735
- [152] An L, Liu Y, Liu Y, et al. Defending llm watermarking against spoofing attacks with contrastive representation learning. arXiv preprint arXiv:2504.06575, 2025
- [153] Wei Johnny Tian-Zheng, Wang Ryan Yixiang, Jia Robin. Proving membership in llm pretraining data via data watermarks. arXiv preprint arXiv:2402.10892, 2024
- [154] Fu Yu, Xiong Deyi, Dong Yue. Watermarking conditional text generation for ai detection: unveiling challenges and a semantic-aware watermark remedy// Proceedings of the AAAI

- Conference on Artificial Intelligence. Vancouver, Canada, 2024: 18003-18011
- [155] Hoang Duy C, Le Hung Tq, Chu Rui, et al. Less is more: sparse watermarking in llms with enhanced text quality. arXiv preprint arXiv:2407.13803, 2024
- [156] Takezawa Yuki, Sato Ryoma, Bao Han, et al. Necessary and sufficient watermark for large language models. arXiv preprint arXiv:2310.00833, 2023
- [157] Liu Shu, Chu Zhiqiang, Zhang Junxia, et al. Research on watermark generation using large language models. Information Technology and Informatization, 2024(5): 203-206 (in Chinese) (刘述, 储志强, 张俊霞, 等. 大语言模型生成水印的研究. 信息技术与信息化, 2024(5): 203-206)
- [158] Pang Qi, Hu Shengyuan, Zheng Wenting, et al. Attacking llm watermarks by exploiting their strengths// ICLR 2024 Workshop on Secure and Trustworthy Large Language Models. Vienna, Austria, 2024
- [159] Liang Jiacheng, Wang Zian, Hong Spencer, et al. Watermark under fire: a robustness evaluation of llm watermarking// Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China, 2025: 21050-21074
- [160] Huang Hanbo, Zhang Yiran, Zheng Hao, et al. RLCracker: exposing the vulnerability of llm watermarks with adaptive RL attacks. arXiv preprint arXiv:2509.20924, 2025
- [161] Zhang Zhaoxi, Zhang Xiaomei, Zhang Yanjun, et al. Character-level perturbations disrupt llm watermarks. arXiv preprint arXiv: 2509.09112, 2025
- [162] Jiang Haoyu, Wang Xuhong, Yi Ping, et al. Credid: Credible multi-bit watermark for large language models identification. arXiv preprint arXiv:2412.03107, 2024
- [163] Massieh Kordi Boroujeny, Jiang Ya, Zeng Kai, et al. Multi-bit distortion-free watermarking for large language models. arXiv preprint arXiv:2402.16578, 2024
- [164] Lin Qihao, Tang Chen, Li Xiangyang. DERMARK: A dynamic, efficient and robust multi-bit watermark for large language models. arXiv preprint arXiv:2502.05213, 2025
- [165] Tang Chenxia, Liu Jianchun, Xu Hongli, et al. Top- n : eliminating noise in logit space for robust token sampling of llm// Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria, 2025: 10758-10774
- [166] He Haiyun, Liu Yepeng, Wang Ziqiao, et al. Distributional information embedding: a framework for multi-bit watermarking. arXiv preprint arXiv:2501.16558, 2025
- [167] Molenda Piotr, Liusie Adian, Gales Mark Jf. Waterjudge: quality-detection trade-off when watermarking large language models. arXiv preprint arXiv:2403.19548, 2024
- [168] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 2020, 21(140): 1-67
- [169] Luo Huixin, Li Li, Li Juncheng. Digital watermarking technology for ai-generated images: A survey. Mathematics, 2025, 13(4): 651
- [170] Wen Yizhu, Innuganti Ashwin, Ramos Aaron Bien, et al. SoK: how robust is audio watermarking in generative AI models? arXiv preprint arXiv:2503.19176, 2025
- [171] Liu Shuliang, Qi Zheng, Xu Jesse Jiayi, et al. VLA-Mark: a cross modal watermark for large vision-language alignment models// Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China, 2025: 26420-26438
- [172] Pang Qi, Hu Shengyuan, Zheng Wenting, et al. No free lunch in llm watermarking: trade-offs in watermarking design choices// Advances in Neural Information Processing Systems. Vancouver, Canada, 2024, 37: 138756-138788
- [173] Hu Jianping, Gao Zhipeng, Mou Yang, et al. Deep learning image watermarking combining attack classification and multi-channel embedding. Journal of Beijing University of Aeronautics and Astronautics, 2024 (in Chinese) (胡建平, 高志鹏, 牟洋, 等. 结合攻击分类和多通道嵌入的深度学习图像水印. 北京航空航天大学学报, 2024))
- [174] Pan Leyi, Liu Aiwei, Huang Shiyu, et al. Can llm watermarks robustly prevent unauthorized knowledge distillation?// Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria, 2025: 13228-13251
- [175] Wu Qilong, Chandrasekaran Varun. Bypassing llm watermarks with color-aware substitutions// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, 2024: 8549-8581
- [176] Mohamed Asim, Gubri Martin. Is multilingual llm watermarking truly multilingual? a simple back-translation solution. arXiv preprint arXiv:2510.18019, 2025
- [177] Zhang Jingqing, Zhao Yao, Saleh Mohammad, et al. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization// Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research. 2020: 11328--11339
- [178] Krishna Kalpesh, Song Yixiao, Karpinska Marzena, et al. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense// Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA, 2023: Article 1195
- [179] Jovanović Nikola, Staab Robin, Vechev Martin. Watermark stealing in large language models// Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria, 2024: 22570-22593
- [180] Yi Xin, Zheng Shunfan, Wang Linlin, et al. Unified attacks to large language model watermarks: spoofing and scrubbing in unauthorized knowledge distillation. arXiv preprint arXiv: 2504.17480, 2025
- [181] Ural O, Yoshigoe K. Enhancing security of proof-of-learning against spoofing attacks using feature-based model watermarking. IEEE Access, 2024, 12: 169567-169591
- [182] Sadasivan Vinu Sankar, Kumar Aounon, Balasubramanian Sriram, et al. Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156, 2023

- [183] Szyller Sebastian, Atli Buse Gul, Marchal Samuel, et al. Dawn: dynamic adversarial watermarking of neural networks// Proceedings of the 29th ACM international conference on multimedia. New York, USA, 2021: 4417-4425
- [184] Charette Laurent, Chu Lingyang, Chen Yizhou, et al. Cosine model watermarking against ensemble distillation// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 9512-9520
- [185] Li Huiying, Wenger Emily, Shan Shawn, et al. Piracy resistant watermarks for deep neural networks. arXiv preprint arXiv: 1910.01226, 2019
- [186] Gao Guangyong, Xu Ziqi, Fang Wei. Robust model watermarking scheme based on feature combination and weight adversarial training. Journal of Computer-Aided Design & Computer Graphics, 2025, 37(9): 1653-1664
- [187] Meng Zheling, Peng Bo, Dong Jing. Latent watermark: inject and detect watermarks in latent diffusion space. IEEE Transactions on Multimedia, 2025, 27(2): 3399-3410
- [188] Arabi Kasra, Feuer Benjamin, Witter R Teal, et al. Hidden in the noise: two-stage robust watermarking for images. arXiv preprint arXiv:2412.04653, 2024
- [189] Ren Kui, Yang Ziqi, Lu Li, et al. Sok: On the role and future of aigc watermarking in the era of gen-ai. arXiv preprint arXiv: 2411.11478, 2024
- [190] Zhang Xuanyu, Li Runyi, Yu Jiwen, et al. Editguard: versatile image watermarking for tamper localization and copyright protection// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Washington, USA, 2024: 11964-11974
- [191] Georg Niess and Roman Kern. Ensemble watermarks for large language models// Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria, 2024: 2903 - 2916
- [192] Chen Huimin, Liu Zhiyuan, Sun Maosong. Social opportunities and challenges in the era of large language models. Journal of Computer Research and Development, 2024, 61(5): 1094-1103 (in Chinese)
(陈慧敏, 刘知远, 孙茂松. 大语言模型时代的社会机遇与挑战. 计算机研究与发展, 2024, 61(5): 1094-1103)
- [193] Dai Long, Zhang Jing, Fan Xuefeng, et al. Copyright protection based on NLP neural network using black-box watermarking. Journal of Network and Information Security, 2023, 9(1): 140-149 (in Chinese)
(代龙, 张静, 樊雪峰, 等. 基于黑盒水印的 NLP 神经网络版权保护. 网络与信息安全学报, 2023, 9(1): 140-149)
- [194] Cai Yuhang, Wang Yaofei, Hu Donghui, et al. Modification and generated-text detection: achieving dual detection capabilities for the outputs of llm by watermark. arXiv preprint arXiv:2502.08332, 2025
- [195] Guo Jingjing, Liu Jiuzun, Ma Yong, et al. A model-based watermarking-based federated learning backdoor attack defense method. Chinese Journal of Computers, 2024, 47(3): 662-676 (in Chinese)
(郭晶晶, 刘玖樽, 马勇, 等. 基于模型水印的联邦学习后门攻击防御方法. 计算机学报, 2024, 47(3): 662-676)
- [196] Wu Junchao, Yang Shu, Zhan Runzhe, et al. A survey on LLM-generated text detection: Necessity, methods, and future directions. Computational Linguistics, 2025: 1-66
- [197] Yang, Yuming; Zhong, Jiang; Gao, Jinpeng, et al. Multimodal fake news detection based on evidence enhancement and local semantic interaction. Chinese Journal of Computers, 2025, 48(3) (in Chinese)
(杨钰铭, 钟将, 高晋鹏, 等. 基于证据增强和局部语义交互的多模态虚假新闻检测. 计算机学报, 2025, 48(3))
- [198] Ke Jing, Xie Zheyong, Xu Tong, et al. A fine-grained method for detecting fake news based on latent semantic enhancement of large language models. Computer Research and Development, 2024, 61(5): 1250-1260 (in Chinese)
(柯婧, 谢哲勇, 徐童, 等. 基于大语言模型隐含语义增强的细粒度虚假新闻检测方法. 计算机研究与发展, 2024, 61(5): 1250-1260)
- [199] Zhang Zhiyong, Jing Junchang, Li Fei et al. A review of research on the detection, dissemination and control of false information in online social networks from the perspective of artificial intelligence. Chinese Journal of Computers, 2021, 44 (11): 2261-2282 (in Chinese)
(张志勇, 荆军昌, 李斐等. 人工智能视角下的在线社交网络虚假信息检测、传播与控制研究综述. 计算机学报, 2021, 44(11): 2261-2282)
- [200] Wu Xiaoshuai, Liao Xin, Ou Bo. Sepmark: deep separable watermarking for unified source tracing and deepfake detection//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada, 2023: 1190-1201
- [201] Li Jiawei, Zhang Fanrui, Zhu Jiaying, et al. Forgerygpt: multimodal large language model for explainable image forgery detection and localization. arXiv preprint arXiv: 2410.10238, 2024
- [202] Wang Xuhong, Jiang Haoyu, Yu Yi, et al. Building intelligence identification system via large language model watermarking: a survey and beyond. arXiv preprint arXiv:2407.11100, 2024
- [203] Tu Shangqing, Sun Yuliang, Bai Yushi, et al. WaterBench: towards holistic evaluation of watermarks for large language models// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, 2024: 1517-1542
- [204] Fernandez Pierre, Chaffin Antoine, Tit Karim, et al. Three bricks to consolidate watermarks for large language models// Proceedings of the 2023 IEEE International Workshop on Information Forensics and Security (WIFS). Nürnberg, Germany, 2023: 1-6
- [205] Luo Yiyang, Lin Ke, Gu Chao, et al. Lost in overlap: exploring logit-based watermark collision in llms. arXiv preprint arXiv: 2403.10020, 2024



AN Wen-Bo, Ph. D. candidate. His research interests include large language model watermarking and natural

WU Lian-Wei, Ph. D., associate professor. His main research interests include large language model security, network content security, and natural language processing.

language processing.

WANG Kang, M. S. candidate. His research interests include deepfake detection and image tampering detection.

WANG Lin-Yong, M. S. candidate. His research interests include text generation and large language models.

ZHANG Hang, M. S. candidate. His research interests include large language model watermarking.

WANG Ze-Hao, M. S. candidate. His research interests include large model content detection and AI security.

Background

Large models, with their superior reasoning and generation capabilities, have shown broad application prospects in education, healthcare, finance, and other fields. However, the resulting issues of digital content security and intellectual property protection have gradually become key factors restricting the sustainable development of this technology. As an important means to address these challenges, large language model watermarking technology has received widespread attention from academia and industry.

Large model watermarking technology is currently in a rapid development phase, and several researchers have already reviewed this field. However, existing reviews still have limitations in terms of coverage and content structure. In contrast, we makes contributions in the following three closely related and progressively developing aspects:

First, it systematically classifies large language model watermarking technologies based on the watermarking embedding modality as the core dimension, establishing four categories: text, image, audio, and multimodal watermarking frameworks, laying a clear conceptual foundation for subsequent research.

Second, focusing on watermarking performance evaluation, it proposes four key evaluation indicators for large language model watermarking—detectability, impact on generated content quality, robustness, and capacity—along with evaluation indicators for multimedia watermarking, constructing a more scientific and comprehensive watermarking evaluation system. Finally, from the perspective of technical security and practicality, it systematically reviews watermarking attack methods and their defense strategies, enhancing the reliability of watermarking technology in practical deployment.

Overall, compared to existing reviews, we achieve breakthroughs in the systematic nature of the research perspective, as well as the depth and breadth of its content. The discussion of large model watermarking technology demonstrates greater systematicity and comprehensiveness, providing more valuable references for the further development of this field.

This work is supported by the National Natural Science Foundation of China (62572403) and the Enterprise Joint Fund Project (U22B2036), and the Basic and Applied Basic Research Foundation of Guangdong Provincial (2024A1515010087).