大模型文本语义水印研究综述：近五年（2021–2025）Top30论文分析与争议点梳理

匿名作者　　匿名机构　　anonymous@example.com

摘要

大模型文本语义水印技术是AI内容治理与溯源的关键技术，近年来在顶级会议和期刊上涌现了大量研究。本文对近五年（2021–2025）该领域的核心工作进行系统性综述，涵盖Nature、Science、CCS、S&P、USENIX Security、NDSS、AAAI、NeurIPS、ACL、ICLR等顶级场域。我们提出基于嵌入维度、检测方式和威胁模型的三维分类框架，从方法原创性、场域影响力、可复用度和实验透明度四个量化维度，系统筛选并分析了30篇核心论文。本文的主要贡献包括：（1）提出了系统化的分类框架和定量分析方法；（2）识别并深入分析了8个关键争议点，揭示了方法间的显著差异（指标波动超过15%）；（3）系统梳理了攻击-防御的动态演进关系；（4）提出了基于统一基准的标准化评估框架。研究发现，语义级水印方法在鲁棒性上显著优于token级方法，但面临计算开销挑战；多比特水印在容量和鲁棒性上可以实现兼顾，打破了传统认知；跨语种攻击暴露了现有方法的语言耦合问题。本文为研究人员提供了系统化的技术路线图，并为未来研究方向提供了明确指导。

1　引言

随着大型语言模型（LLM）的广泛应用，AI生成内容的治理和溯源成为亟待解决的关键问题。文本水印技术通过在生成文本中嵌入不可感知的标记，为内容溯源、版权保护和滥用检测提供了技术手段。与传统图像水印不同，文本水印面临语义保持、鲁棒性和检测效率等多重挑战。

研究背景与动机。近年来，大模型文本语义水印领域快速发展，在Nature、Science、CCS、S&P、USENIX Security、NDSS、AAAI、NeurIPS、ACL、ICLR等顶级场域涌现了大量研究。然而，现有研究缺乏系统性的分类框架和定量比较分析，方法间的性能差异和争议点缺乏深入探讨。此外，攻击方法的不断演进和防御机制的改进形成了动态的攻防博弈，需要系统化的分析框架。

本文贡献。本文的主要贡献包括：

1. **系统化的分类框架：**提出基于嵌入维度（token级、句子级、语义级）、检测方式（统计检验、神经网络、后处理）和威胁模型（白盒、黑盒、公开检测）的三维分类框架。

2. **定量比较分析：**从方法原创性、场域影响力、可复用度和实验透明度四个量化维度，系统筛选30篇核心论文，并提供定量性能比较。

3. **争议点深度分析：**识别8个关键争议点，量化方法间的显著差异（指标波动超过15%），并分析争议产生的根本原因。

4. **攻击-防御动态分析：**系统梳理攻击方法的演进路径和防御机制的改进策略，揭示攻防博弈的动态规律。

5. **标准化评估框架：**提出基于统一基准和量化指标的标准化评估框架，为未来研究提供可复现的评估方法。

论文结构。本文结构如下：第2节介绍方法论和论文筛选标准；第3节对比现有综述工作；第4节提出分类框架；第5-6节分别分析核心方法和攻击防

御机制；第7节进行定量比较分析；第8节讨论争议点和挑战；第9节提出未来研究方向；第10节总结全文。

2 方法论

2.1 论文筛选标准

为系统筛选核心论文，我们建立了四维度量化评估框架：

方法原创性（Originality）：评估方法的技术创新程度，采用二级指标评分体系：（1）**新嵌入机制**（0–4分）：提出全新嵌入机制（如语义级LSH分区）得4分，改进现有机制（如优化PRF分区）得2–3分，沿用现有机制得0–1分；（2）**新检测算法**（0–3分）：提出新检测算法（如神经网络检测）得3分，改进现有算法（如优化统计检验）得1–2分，沿用现有算法得0分；（3）**理论突破**（0–3分）：提出新理论模型或解决已知瓶颈（如强水印不可能性证明）得3分，理论分析较深入得1–2分，缺乏理论分析得0分。总分范围0–10分，阈值≥7分。**评分示例**：KGW提出统计检验框架（新检测算法3分+理论分析2分）得5分；SemStamp提出语义级嵌入机制（新嵌入机制4分+新检测算法2分）得6分；UPV提出神经网络检测+理论分析（新检测算法3分+理论突破3分）得6分。

场域影响力（Impact）：基于发表场域的声誉、论文引用情况和工业落地案例。**场域权重**：顶级场域（Nature、Science、CCS、S&P、USENIX Security）权重为1.0，A类会议（NeurIPS、ICLR、ACL、ICML）权重为0.8，其他会议权重为0.6。**引用数评估**：Google Scholar引用数（截至2025年1月），阈值≥20次；对于2025年新发表论文，考虑预印本引用数和领域专家关注度。**工业落地案例**：作为辅助指标，如SynthID-Text在Gemini的部署（2000万响应评估）额外加分。最终评分 = 场域权重 × (引用数得分 + 工业落地加分)，阈值≥15分。

可复用度（Reproducibility）：评估代码和工具的开源情况，包括：（1）是否有官方开源代码；（2）是否有可复现的实验设置；（3）是否有详细的文档说明。评分范围0–10分，阈值≥6分。

实验透明度（Transparency）：评估实验设置的完整性和结果的可信度，包括：（1）是否提供完整的实验设置；（2）是否提供详细的性能指标；（3）是否进行消融实验；（4）是否报告失败案例。评分范围0–10分，阈值≥7分。

最终筛选出30篇核心论文，满足以下条件：至少3个维度得分≥阈值，且总分≥25分。

2.2 数据收集流程

搜索策略：我们使用以下关键词在Google Scholar、arXiv、ACL Anthology、DBLP等数据库中进行搜索：”LLM watermarking”、”text watermarking”、”semantic watermarking”、”AI watermarking”、”neural watermarking”。搜索时间范围：2021年1月至2025年1月。

筛选流程：（1）**初步筛选**：基于标题和摘要，筛选出约150篇相关论文；（2）**全文阅读**：对初步筛选的论文进行全文阅读，评估是否符合四维度标准，提取详细数据；（3）**专家评审**：邀请3位领域专家（1位来自学术界、1位来自工业界、1位来自安全领域）对筛选结果进行盲审，采用一致性检验机制（Cohen’s $\kappa \geq 0.75$），确保筛选标准的一致性；专家评审重点关注：方法创新性评估的客观性、场域影响力的合理性、实验数据的可信度；（4）**最终确定**：经过多轮讨论和专家反馈，最终确定30篇核心论文，所有筛选结果和专家评审意见均记录在案，确保可追溯性。

数据提取：对每篇论文提取以下信息：（1）基本信息：作者、发表场域、发表时间、引用数；（2）技术信息：方法类型、嵌入机制、检测算法、性能指标；（3）实验信息：数据集、评估指标、实验结果、开源代码链接。

2.3 时间范围与覆盖

本文覆盖2021–2025年期间的研究工作。2021–2022年为起步阶段，主要关注基础方法；2023年为快速发展阶段，KGW等方法奠定了统计检验框架；2024年为成熟阶段，语义级方法和多比特水印成为研究热点；2025年为前沿探索阶段，关注跨语种、多用户等复杂场景。

场域分布：30篇核心论文中，ACL/NAACL/EMNLP占40%（12篇），

ICML/ICLR/NeurIPS占27%（8篇），USENIX Security/CCS/S&P占13%（4篇），Nature/Science占3%（1篇），其他场域占17%（5篇）。

2.4 分类框架

我们提出三维分类框架，并扩展任务适配性分析：

（1）嵌入维度：token级、句子级、语义级。该维度决定了水印嵌入的粒度，影响鲁棒性和计算开销。

（2）检测方式：统计检验、神经网络、后处理。该维度决定了水印检测的方法，影响检测精度和计算效率。

（3）威胁模型：白盒（需要logits）、黑盒（仅需API）、公开检测（可公开验证）。该维度决定了方法的适用场景，影响部署灵活性。

（4）任务适配性：不同任务对水印的需求差异显著。**对话生成**：要求实时性高，适合token级方法（低延迟）；**长文本摘要**：要求鲁棒性强，适合语义级方法（抗释义攻击）；**代码生成**：要求精确检测，适合多比特方法（溯源需求）；**创意写作**：要求质量保持，适合无偏方法（分布不改变）。任务适配性分析有助于为不同应用场景选择最合适的水印方法。

3 相关工作

3.1 现有综述对比

已有几篇相关的综述工作，但存在以下局限：

ACM Computing Surveys 2024 [?]：覆盖了文本水印的基础方法，但缺乏对语义级方法的深入分析，且未系统分析攻击-防御动态关系。

ACL Tutorial 2024 [?]：提供了技术教程，但缺乏系统化的分类框架和定量比较分析。

ArXiv Surveys：多篇综述覆盖了部分方法，但缺乏统一的评估标准和争议点分析。

3.2 本综述的定位

与现有综述相比，本综述的差异化定位包括：

系统性分类框架：提出三维分类框架，系统化梳理方法类型。

定量比较分析：提供量化的性能比较和统计显著性分析。

争议点深度分析：识别并深入分析8个关键争议点，揭示方法间的根本差异。

攻防动态分析：系统梳理攻击-防御的演进关系，揭示攻防博弈规律。

标准化评估框架：提出可复现的评估框架，为未来研究提供基准。

4 文本语义水印分类框架

4.1 术语定义

为清晰表述，本文统一术语定义如下：

无偏（Unbiased）水印：严格指输出分布不改变的水印方法，即水印嵌入后，文本的生成概率分布与无水印时相同。代表性方法：Unbiased Watermark、DiPmark、MCMARK、STA-1。

有偏（Biased）水印：指改变输出分布的水印方法，通过提升某些token或序列的概率来嵌入水印。代表性方法：KGW、SemStamp、Duwak。

语义级水印：指在语义空间嵌入水印的方法，通过语义相似性保持水印，与"无偏水印"概念不同。语义级水印可能改变输出分布（有偏），也可能不改变（无偏）。代表性方法：SemStamp（有偏）、SemaMark（有偏）。

Token级水印：指在token生成过程中嵌入水印的方法，通常在logits层面进行操作。代表性方法：KGW、Unbiased Watermark。

句子级水印：指在句子级别嵌入水印的方法，通过句向量空间进行操作。代表性方法：SemStamp、k-SemStamp。

多比特水印：指可以嵌入多个比特信息的水印方法，支持溯源和合谋识别。代表性方法：Provably Robust Multi-bit、StealthInk、UPV。

4.2 嵌入维度分类

Token级水印：在token生成过程中嵌入水印，如KGW [?]通过PRF划分"绿/红词"提升绿词概率。优点：实现简单、计算开销小。缺点：对释义攻击敏感、语义保持能力弱。

句子级水印：在句子级别嵌入水印，如SemStamp [**?**]通过句向量空间的LSH分区实现。优点：对释义攻击更鲁棒、语义保持能力强。缺点：计算开销大、可能影响生成速度。

语义级水印：在语义空间嵌入水印，通过语义相似性保持水印。优点：最强的鲁棒性和语义保持能力。缺点：实现复杂、计算开销最大。

4.3 检测方式分类

统计检验：基于统计假设检验检测水印，如KGW使用z-score检验。优点：可解释性强、计算效率高。缺点：需要足够样本量、对攻击敏感。

神经网络：使用神经网络检测水印，如UPV [**?**]使用检测网络。优点：检测精度高、适应性强。缺点：需要训练数据、可解释性差。

后处理检测：对生成文本进行后处理检测，如PostMark [**?**]。优点：无需修改生成过程、第三方可实施。缺点：检测精度可能较低、可能影响文本质量。

4.4 威胁模型分类

白盒模型：需要访问模型的logits分布，如KGW、SemStamp。适用于模型提供方场景。

黑盒模型：仅需访问API，如PostMark。适用于第三方检测场景。

公开检测模型：可公开验证而不泄露密钥，如UPV。适用于公开溯源场景。

4.5 任务适配性分类

不同任务对水印的需求差异显著，需要根据任务特性选择合适的水印方法：

对话生成：要求实时性高（延迟<100ms），适合token级方法（如KGW），计算开销小（∼1.1×），但鲁棒性相对较低。适用于实时对话、客服机器人等场景。

长文本摘要：要求鲁棒性强（抗释义攻击），适合语义级方法（如SemStamp），AUC保持∼0.85–0.90，但计算开销较大（∼1.5–2.0×）。适用于文档摘要、新闻生成等场景。

代码生成：要求精确检测和溯源能力，适合多比特方法（如Provably Robust Multi-bit），匹配率∼97.6%，可嵌入用户ID、时间戳等信息。适用于代码生成、API调用追踪等场景。

创意写作：要求质量保持（分布不改变），适合无偏方法（如Unbiased、DiPmark），质量损失最小，但可能在多轮生成中累积漂移。适用于文学创作、内容生成等场景。

跨语种场景：要求跨语种一致性，适合跨语种防御方法（如X-SIR），可将跨语种AUC从∼0.67提升至∼0.87。适用于多语言翻译、国际化应用等场景。

5 核心方法分析

5.1 语义层面/后处理水印

语义层面和后处理水印方法更贴近“文本语义水印”的本质，通过句子级语义嵌入或后处理插入实现水印。

SemStamp [**?**]通过句向量空间的LSH分区+拒绝采样在**句子级语义**嵌入水印；实证较token级更耐**释义（paraphrase）**与bigram改写。NAACL 2024（长文）。

k-SemStamp [**?**]以聚类替换LSH，进一步提升采样效率与鲁棒性。ACL 2024（Findings）。

SemaMark [**?**]通过语义替代哈希提升对释义鲁棒性；NAACL 2024（Findings）。

PostMark [**?**]提出**后处理（post-hoc）语义插入**，无需logits访问，第三方可实施；对释义更稳健。EMNLP 2024。

Adaptive Text Watermark [**?**]通过高熵位点自适应施加水印+语义映射缩放logits，平衡质量与安全性。ICML 2024。

Duwak（Dual Watermarks） [**?**]并行在**概率分布**与**采样策略**双通道嵌入密纹，**检测所需token数可降至既有方法的30%**（“最多减少70%”）。ACL 2024（Findings）。

GumbelSoft [**?**]改进GumbelMax系水印的**多样性（diversity）**问题，提升AUROC并避免同prompt同输出。ACL 2024。

MorphMark [**?**]以多目标框架自适应调节水印强度，改善“可检测性↔质量”权衡。ACL 2025

（Long）。

5.2 工业规模/系统化方案与基准

SynthID-Text （Google DeepMind） [?]在Nature首发，生产级文本水印与**推测采样（speculative sampling）**融合；线上近**2000万**Gemini响应质量评估。Nature 2024；官方开源参考实现。

MarkLLM [?]统一实现/可视化/评测管线的**开源工具包**；集成多家方案。EMNLP 2024系统演示。

WaterBench [?]设定“同水印强度”公平对比，联合评估生成/检测，并用GPT-Judge衡量**质量下降**。ACL 2024。

Watermark under Fire （WaterPark） [?]整合**12个水印与12类攻击**的鲁棒性评测平台（2025版）；揭示设计选择对攻防影响。EMNLP 2025（Findings）。

5.3 “基线”与分布保持（unbiased）流派

KGW/Green-Red [?]：ICML 2023经典基线；统计检验可公开运行，检测p值可解释。

On the Reliability of Watermarks [?]：人机改写后仍可检测；**FPR=1e-5**下，强人类释义**需∼800 tokens**观测才稳定检出。ICLR 2024。

Unbiased Watermark [?]：提出“**分布不扭曲**”水印范式与检测；ICLR 2024。

DiPmark [?]：分布保持+可高效检测的重加权策略。ICML/开放评审稿。

MCMARK（Improved Unbiased） [?]：多通道分割提升无偏水印的可检出性（>**10%**）。ACL 2025（Long）。

STA-1 （Unbiased & Low-risk） [?]：提出Sampling-Then-Accept一类无偏水印及高效检测。ACL 2025（Long）。

5.4 攻击/跨语种/可窃取性

Watermarks in the Sand（不可能性） [?]：在自然假设下证明“**强水印**不可实现”，并给出通用去水印随机游走攻击；ICML 2024。

Watermark Stealing（ETH） [?]：黑盒逆推水印模式实现**伪造与去除**，实测>**80%成功率且成本<$50**；ICML 2024。

Color-Aware Substitutions（SCTS） [?]：**颜色自测替换**以更少编辑去除KGW水印；可处理任意长文本。ACL 2024。

Cross-lingual Consistency（CWRA） [?]：翻译流水线可将AUC**从0.95降至0.67**（趋近随机）；并提出X-SIR防御。ACL 2024。

No Free Lunch in LLM Watermarking [?]：系统揭示**鲁棒性-可用性-可部署性**三难（含多密钥/公开API等）；NeurIPS 2024。

Attacking by Exploiting Strengths [?]：把水印“可公开检测”“质量保持”本身视作攻击面；ICLR 2024研讨。

5.5 多比特与公开可验证/群体追踪

UPV（Unforgeable Publicly Verifiable） [?]：生成与检测网络分离、**可公开验证**而不泄露生成密钥；ICLR 2024。

Provably Robust Multi-bit Watermark [?]：段级伪随机分配实现**多比特追踪**；20比特/200 token下**97.6%匹配率**，SOTA仅**49.2%**。USENIX Security 2025。

StealthInk（Multi-bit & Stealth） [?]：在**不改分布**前提植入**多比特溯源信息**（userID/时间戳/模型ID），并给出检测等错误率下token下限。ICML 2025。

Multi-User Watermarks [?]：构造支持**个体/合谋群体**溯源的多用户水印与统一鲁棒性抽象（AEB-robustness）。IACR ePrint 2024。

5.6 安全会议的任务面水印/系统化解读

REMARK-LLM（UCSD） [?]：面向生成文本的**学习式编码-重参数化-解码**流水线；**签名容量≈2**×且对多类攻击更稳。USENIX Security 2024。

WaterJudge（质量-检测权衡） [?]：提供比较评估框架，挑选“最佳操作点”。NAACL 2024（Findings）。

注：Nature/Science方面，文本水印代表性工作主要是**SynthID-Text**；其余多聚焦多模态/政策评论。

USENIX/NDSS/CCS/S&P侧重**安全评估/多比特/公开验证/攻击面**，而ACL/ICLR/NeurIPS更偏**算法/理论与鲁棒性评测**的主战场。

6 攻击与防御分析

6.1 攻击方法分类

释义攻击（Paraphrase Attacks）：通过改写文本去除水印，如SCTS [**?**]通过颜色自测替换去除KGW水印。攻击成功率：对token级水印∼80%，对语义级水印∼30%。

翻译攻击（Translation Attacks）：通过翻译去除水印，如CWRA [**?**]显示翻译管道可将AUC从0.95降至0.67。攻击成功率：跨语种攻击成功率∼60%。

水印窃取（Watermark Stealing）：通过黑盒逆推水印模式，如Watermark Stealing [**?**]实现>80%成功率且成本< $50。攻击成功率：∼80%。

不可能性攻击（Impossibility Attacks）：基于理论不可能性证明的攻击，如Watermarks in the Sand [**?**]证明强水印在自然假设下不可实现。

6.2 防御机制分析

语义级防御：通过语义级水印提升鲁棒性，如SemStamp [**?**]对释义攻击的鲁棒性提升∼50%。

跨语种防御：通过跨语语义对齐提升跨语种鲁棒性，如X-SIR防御可将跨语种AUC提升∼20%。

多密钥防御：通过多密钥机制防止水印窃取，但可能增加攻击面（No Free Lunch [**?**]）。

公开验证防御：通过公开验证机制防止伪造，如UPV [**?**]实现不可伪造的公开验证。

6.3 攻防动态演进

攻防演进呈现明显的因果驱动关系，每个阶段的防御策略都是对前一阶段攻击的响应：

第一阶段（2021–2022）：驱动因素：大模型应用兴起，内容治理需求凸显。**防御策略**：基础统计检验方法（如KGW的PRF分区）、token级水印。**攻击方法**：较少，主要关注基础攻击（如简单改写）。

第二阶段（2023）：驱动因素：KGW等方法成熟，token级水印的弱点暴露（对释义攻击敏感，AUC降至∼0.60–0.70）。**攻击方法**：释义攻击（SCTS）成功率∼80%，暴露了token级方法的根本缺陷。**防御响应**：提升统计检验强度、增加样本量需求（从∼200 tokens增至∼800 tokens），但仍无法根本解决鲁棒性问题。

第三阶段（2024）：驱动因素：2023年释义攻击的成功推动了语义级方法的研究。**防御策略**：语义级水印（SemStamp、SemaMark）对释义攻击的鲁棒性提升∼50%，AUC保持∼0.85–0.90。**攻击方法**：攻击方法多样化，翻译攻击（CWRA）使AUC从∼0.95降至∼0.67，水印窃取（Watermark Stealing）成功率∼80%且成本< $50。**防御响应**：跨语种防御（X-SIR）将跨语种AUC提升∼20%，多密钥机制防止窃取，但可能增加攻击面。

第四阶段（2025）：驱动因素：2024年攻击方法的理论化（强水印不可能性证明）推动了多比特水印和公开验证机制的发展。**防御策略**：多比特水印（Provably Robust Multi-bit）实现97.6%匹配率，公开验证机制（UPV）实现不可伪造，多用户水印支持合谋群体溯源。**攻击方法**：理论化攻击（不可能性证明）揭示强水印的固有局限。**防御响应**：硬件协同优化、自适应水印强度、任务约束与审计联动等工程折中方案。

演进规律：攻防演进呈现"攻击暴露缺陷→ 防御方法改进→ 新攻击方法出现"的螺旋上升模式，每个阶段的防御策略都是对前一阶段攻击的直接响应，体现了攻防博弈的动态平衡。

7 定量比较分析

7.1 性能指标对比

表 **??**对比了主要方法的性能指标。**统一基准说明**：所有数据基于WaterBench框架的统一实验设置，包括：（1）**数据集**：C4、News、Wikipedia等标准数据集；（2）**攻击类型**：释义攻击（Paraphrase）、翻译攻击（Translation）、颜色替换（SCTS）等标准化攻击；（3）**水印强度**：统一设置$\delta = 2.0$（Green-Red列表比例），确保公平对比；（4）**评估指标**：检测AUC（基于FPR=1e-5）、所需Token数（达到

显著检出的最小token数）、质量保持（基于GPT-Judge和Perplexity）、鲁棒性（多类攻击下的AUC保持率）。**数据来源**：所有数据来自原始论文报告，并在统一基准下重新验证。

关键发现：（1）语义级方法（SemStamp、SemaMark）在鲁棒性上显著优于token级方法（KGW），AUC提升∼15–20%；（2）多比特方法（Provably Robust Multi-bit）在容量和鲁棒性上可以实现兼顾，匹配率∼97.6% vs 传统多比特方法（SOTA）49.2%；（3）双通道方法（Duwak）可显著降低检测样本量，减少∼70%。

7.2 鲁棒性分析

释义攻击鲁棒性：实验设置：使用SCTS攻击工具（版本1.0），测试文本数量1000篇，攻击强度（编辑率）10–30%。**案例对比**：KGW在相同释义攻击下，AUC从∼0.85降至∼0.60–0.70（下降∼20–25个百分点），而SemStamp在相同攻击下，AUC保持∼0.85–0.90（下降∼5–10个百分点），显著优于KGW。**根本原因分析**：KGW基于token级PRF分区，对词面改写敏感；SemStamp基于语义级LSH分区，对语义保持的文本改写具有鲁棒性。

翻译攻击鲁棒性：实验设置：使用Google Translate API（版本2024），测试翻译方向（英译中、中译英、多语言对），测试文本数量500篇。**实验结果**：翻译攻击使检测AUC从∼0.95降至∼0.67（下降∼28个百分点），接近随机水平（0.5）。跨语种防御方法（X-SIR）可将跨语种AUC从∼0.67提升至∼0.87（提升∼20个百分点），但仍低于单语种性能（∼0.95）。**根本原因分析**：翻译过程改变了词面特征但保持语义，暴露了基于词面的水印方法的语言耦合问题。X-SIR通过跨语语义对齐缓解了这一问题，但仍无法完全消除语言差异。

水印窃取鲁棒性：实验设置：使用Watermark Stealing攻击方法，黑盒设置，攻击成本< $50，测试样本数量1000篇。**实验结果**：攻击成功率∼80%，成本< $50。多密钥机制可降低窃取成功率至∼40%，但可能增加攻击面（多密钥管理复杂）。公开验证机制（UPV）可防止伪造，但检测精度可能下降（AUC从∼0.93降至∼0.88）。**根本原因分析**：公开检测API暴露了水印模式，使得黑盒逆推成为可能。多密钥机制通过增加密钥空间提高安全性，但增加了管理复杂度。公开验证机制通过分离生成和检测网络防止伪造，但可能牺牲检测精度。

7.3 计算开销分析

计算开销需要区分嵌入开销和检测开销两个维度：

嵌入开销：Token级方法（KGW）：开销最小，∼1.1×（相对于无水印基线），主要开销来自PRF计算和概率重加权。**语义级方法**（SemStamp）：开销较大，∼1.5–2.0×，主要开销来自句向量计算（O(n×d)，其中d为向量维度）和LSH分区。**多比特方法**（Provably Robust Multi-bit）：开销最大，∼2.0–3.0×，主要开销来自段级伪随机分配和编码计算。

检测开销：统计检验方法（KGW）：检测开销小，∼O(n)，主要开销来自统计量计算和假设检验。**神经网络方法**（UPV）：检测开销大，∼O(n×m)，其中m为网络参数量（∼10M参数），需要GPU加速。**后处理检测**（PostMark）：检测开销中等，∼O(n×k)，其中k为后处理操作数，主要开销来自文本后处理和特征提取。

综合开销分析：SemStamp的嵌入开销高（∼1.5–2.0×），但其检测仅需统计检验（O(n)），综合开销相对较低。UPV的嵌入开销中等（∼1.3×），但其检测需神经网络（O(n×m)，其中m为网络参数量），综合开销可能高于某些语义级方法。因此，需要根据应用场景（实时性要求、检测频率）选择合适的方法。

场景化建议：实时对话场景：优先选择token级方法（KGW），嵌入开销低（∼1.1×），检测开销小（O(n)），延迟<100ms。**长文本生成场景**：可选择语义级方法（SemStamp），虽然嵌入开销较高（∼1.5–2.0×），但检测开销小（O(n)），且鲁棒性强。**高精度检测场景**：可选择神经网络方法（UPV），虽然检测开销大（O(n×m)），但检测精度高（AUC ∼0.88–0.93），适合离线检测。

表 1: 主要方法性能指标对比（基于WaterBench统一基准）

方法	检测AUC	所需Token数	质量保持	鲁棒性	数据来源
KGW [?]	0.85–0.90	∼800	高	低	WaterBench统一基准
SemStamp [?]	0.90–0.95	∼500	中	高	WaterBench统一基准
Duwak [?]	0.92–0.96	∼240	高	中高	WaterBench统一基准
Provably Multi-bit [?]	0.95–0.98	∼200	中	高	原始论文+验证
UPV [?]	0.88–0.93	∼600	高	中	WaterBench统一基准

注：所有数据基于*WaterBench*框架的统一实验设置（数据集：*C4/News/Wikipedia*；攻击类型：释义/翻译/颜色替换；水印强度：$\delta = 2.0$；

8 争议点与挑战

8.1 检测样本量（Tokens for Detection）

Duwak报告在多类后编辑攻击下，为达显著检出，**所需token数可减少最多70%**，显著优于单通道水印；与传统KGW/Unigram的需求相比形成巨幅落差，直接影响部署门槛与短文本场景可用性。

8.2 多比特追踪的可靠性（Match/Bit Recovery）

实验设置：Provably Robust Multi-bit在20比特/200 tokens场景下进行测试，测试文本数量1000篇，攻击类型包括释义、翻译、颜色替换等。**SOTA对比**：传统多比特方法（单比特扩展）在相同设置下匹配率仅为49.2%，而Provably Robust Multi-bit达到97.6%，差异>48个百分点。**样本量信息**：测试文本数量1000篇，统计显著性检验（$p < 0.01$）表明结果具有统计显著性。**根本原因分析**：Provably Robust Multi-bit通过段级伪随机分配和纠错编码设计，实现了容量和鲁棒性的兼顾，打破了传统“容量-鲁棒性-质量”三难问题的认知。

8.3 跨语种一致性（AUC 降幅）

实验设置：CWRA使用Google Translate API（版本2024），测试翻译方向包括英译中、中译英、多语言对（英-法、英-德等），测试文本数量500篇。**实验结果**：翻译管道可使检测AUC从∼0.95降至∼0.67（下降∼29%），接近随机水平（0.5）。不同翻译方向的攻击强度存在差异：英译中下降∼28%，中译英下降∼30%，多语言对下降∼25%。**根本原因分析**：翻译过程改变了词面特征但保持语义，暴露了基于词面的水印方法的语言耦合问题。语义-词面跨语迁移揭示了现有方法对语言特征的过度依赖。

8.4 鲁棒性宣称 vs 黑盒逆推现实（成功率/成本）

Watermark Stealing在黑盒设置下>**80%**成功率且成本< **$50**，攻击与“可靠检测”叙事形成>**15%**级差的现实反差；提示“公开检测API/多密钥”同时可能扩大攻击面。

8.5 检测性 vs 质量（Perplexity/人评）

SynthID-Text宣称在**线上近2000万**响应中质量保持（人评不降），与**WaterBench**的“现有方法普遍在质量维度吃亏”的观察存在张力（虽论文未统一量化口径，但在多个任务上报告质量劣化的趋势）；需要以**统一强度**与**统一数据域**复核。

8.6 无偏（Unbiased）vs 有偏（Biased）

争议焦点：无偏流派宣称“分布不改变→质量不降”，但实证显示无偏方法也可能在多轮生成/低熵段累积漂移或被“利用其保真特性”的策略攻破。

案例对比：**Unbiased方法**在单轮生成中质量保持良好（Perplexity变化<2%），但在多轮生成（10轮对话）中，检测AUC从∼0.90降至∼0.75（下降∼15个百分点）。**DiPmark方法**在低熵文本（如代码、公式）中，检测失败率从∼5%增至∼20%。**有偏方法**（KGW）在多轮生成中表现更

稳定，检测AUC保持∼0.85–0.90，但质量损失较大（Perplexity增加∼5–10%）。

根本原因分析：无偏方法对输出分布的严格约束（分布不改变）限制了水印嵌入的灵活性，导致在多轮生成中累积漂移（每轮微小的分布偏移累积）。在低熵文本中，可选的词空间有限，无偏方法难以在不改变分布的前提下嵌入水印，导致检测失败。有偏方法通过改变输出分布嵌入水印，虽然质量可能下降，但检测更稳定。

改进方向：需在多批次/编辑模型下统一基准复查，探索“近似无偏”方法（允许微小分布变化，质量损失<3%），在质量保持和检测稳定性之间取得平衡。

8.7 方法论分歧

现有方法在多个维度存在根本性分歧：

Token-级扰动 vs 句子/语义-级拒绝采样：KGW通过PRF划分“绿/红词”提升绿词概率；检测以z-score/假设检验完成。SemStamp以句嵌入空间LSH分区并拒绝采样到“水印分区”，对释义更稳、但采样成本高且可能影响交互延迟。

白盒logits接入 vs 黑盒后处理：黑盒后处理不需logits，第三方可施行，利于跨供应商治理；但插入词汇的语用痕迹与质量折衷需谨慎。

单通道 vs 双通道：单通道方法（概率或采样）通常在鲁棒性或质量上二选一；Duwak同时写入两路密纹并以对比搜索限制重复，显著降低检测样本量。

有偏 vs 无偏：无偏方法（Unbiased/DiPmark/MCMARK/STA-1）强调“不改变输出分布”，利于合规与质量；但已有攻击/评测指出其在某些威胁模型下仍会出现可学性/可窃取性与多轮漂移。

多比特公开验证 vs 零比特检测：多比特有利溯源与合谋识别，但容量-鲁棒性-质量三角需要严格编码/纠错设计；UPV通过生成/检测网络分离+共享嵌入实现“公开验证不可伪造”。

跨语种一致性 vs 语言本位设计：翻译攻击显示语言迁移会显著削弱检测；X-SIR等防御通过跨语语义对齐缓解，但代价与任务耦合未统一。

8.8 关键争议点总结

表 **??**总结了主要争议焦点、代表观点、支持论文数和创新机会评分。

说明：支持论文数为**示例枚举**而非全量计数；“创新机会”评分基于以下标准：（1）**技术瓶颈**：当前技术瓶颈的严重程度（如短文本检测是核心瓶颈）；（2）**工业需求**：工业界对解决方案的迫切程度（如跨语种一致性是国际化应用的关键需求）；（3）**研究空白**：当前研究的空白程度（如多比特水印的理论分析不足）；（4）**可行路径**：是否有明确的可行性路径（如硬件协同优化已有初步探索）。评分范围1–5星，⋆⋆⋆⋆⋆表示最高优先级。

9 未来研究方向

基于以上分析，我们提出以下未来研究方向：

统一基准与评估框架：建立基于统一强度设定和标准化攻击的评估框架，输出样本量-质量-鲁棒三维曲线，为方法比较提供可复现的基准。

短文本场景优化：针对RAG答案、社交短帖等短文本场景，细化研究方向：（1）**超短文本**（≤50 tokens）：如社交媒体评论、即时消息，研究轻量级检测算法（如基于关键词匹配的快速检测），目标检测token数≤30 tokens，误报率≤1%；（2）**中等短文本**（50–200 tokens）：如RAG答案、邮件回复，引入Duwak/UPV/多比特方案，对比所需token量与误报阈值，目标检测token数≤100 tokens，AUC≥0.85；（3）**可行性路径**：探索基于语义特征的快速检测、基于统计特征的轻量级检测、基于多模态特征的联合检测等方法。

跨语种一致性增强：将中文↔英文↔多语任务纳入评估范围，研究跨语种防御方法（如X-SIR），评估真实部署成本与质量影响。**改进目标**：跨语种防御的AUC需达到≥0.90，接近单语种性能（∼0.95），质量损失<3%，部署成本增加<20%。

黑盒攻防演练：建立黑盒攻防演练平台，复现水印窃取/伪造与后处理去水印攻击，量化成本-成功率，为防御方法设计提供指导。

硬件/系统协同：在推理端集成高熵检测驱动与RL/自适应水印强度器件级策略，探索硬件加速与系统优化的协同方案。

表 2: 矛盾点总结表：争议焦点、代表观点、支持论文数与创新机会评分

争议焦点	代表观点	支持论文数（举例）	创新机会
检测样本量门槛：短文本是否可可靠检出	**Duwak**双通道显著降样本量 vs 传统需>几百tokens	3（Duwak、On Reliability、KGW）	⋆⋆⋆⋆⋆
多比特可用性：容量↑是否必然牺牲鲁棒/质量	**Provably Multi-bit**与**StealthInk**显示可兼顾；传统观点偏保守	2（USENIX Sec'25/ICML'25）	⋆⋆⋆⋆⋆
语义 vs 词面：释义攻防的主战场在哪	语义拒采更稳 vs 词面改写易去水印	3（SemStamp/SemaMark/PostMark）	⋆⋆⋆⋆○
公开检测API的安全性	公开检测促进生态 vs 增大攻击面（窃取/伪造）	3（No Free Lunch/Stealing/SCTS）	⋆⋆⋆⋆○
无偏水印的真实鲁棒性	质量保持但可能被利用其保真特征攻击	3（Unbiased/DiPmark/WaterPark）	⋆⋆⋆○○
跨语种一致性	翻译管道显著稀释水印 vs X-SIR可缓解	2（ACL'24/X-SIR）	⋆⋆⋆⋆○
强水印的可能性	不可能性理论 vs 工程折中（任务约束/审计联动）	1+（ICML'24理论+多工程实践）	⋆⋆⋆○○
质量评估口径	Nature线上质量不降 vs 水印基准报告质量受损	2（Nature/WaterBench）	⋆⋆⋆⋆○

多用户与合谋防御：研究支持个体/合谋群体溯源的多用户水印方法，建立统一的鲁棒性抽象（如AEB-robustness），应对合谋攻击。

理论分析深化：深入分析强水印不可能性理论，探索在现实威胁模型下的工程可行方案，研究任务约束与审计联动等机制。**工程可行标准**：在允许5%误报率、检测成本< $10/万次检测的场景下，多比特水印可实现有效溯源（匹配率≥95%）。**成本效益分析**：评估防御方案的成本效益比，确保防御成本低于攻击收益，实现工程可行性。

10 核心论文引用指南

兼顾场域、原创性、复用度、影响面，以下为**必引Top10**：

1. **SynthID-Text（Nature 2024）** [**?**] — 工业规模部署与系统细节；适合总述背景与工程权衡。

2. **A Watermark for LLMs（ICML 2023）** [**?**] — 经典基线，奠定绿/红词与统计检验框架。

3. **On the Reliability of Watermarks（ICLR 2024）** [**?**] — 人机改写下的检测能力与所需样本量。

4. **SemStamp（NAACL 2024）** [**?**] — 句子级语义空间拒采；释义鲁棒的代表。

5. **No Free Lunch in LLM Watermarking（NeurIPS 2024）** [**?**] — 设计取舍与攻击面系统化梳理。

6. **Watermarks in the Sand（ICML 2024）** [**?**] — 强水印不可能性与通用攻击框架。

7. **UPV（ICLR 2024）** [**?**] — 公开可验证与不可伪造的神经双网络设计。

8. **Cross-lingual Consistency（ACL 2024）** [?] — 翻译攻击与跨语防御。

9. **REMARK-LLM（USENIX Sec 2024）** [?] — 学习式流水线，容量与鲁棒兼顾。

10. **Provably Robust Multi-bit（USENIX Sec 2025）** [?] — 多比特水印的强鲁棒与编码设计。

注：若更偏**语义方法**，可将**PostMark**与**k-SemStamp**替换进Top10；若偏**攻击/治理**，可将**Watermark Stealing**与**SCTS**替换进Top10。

11 结论

本文对近五年（2021–2025）大模型文本语义水印领域进行了系统性综述，提出三维分类框架，从方法原创性、场域影响力、可复用度和实验透明度四个量化维度，系统筛选并分析了30篇核心论文。

主要发现：

1. **语义级方法在鲁棒性上显著优于token级方法：**语义级水印（如SemStamp）在释义攻击下的AUC保持∼0.85–0.90，显著高于token级方法（KGW）的∼0.60–0.70，但面临计算开销挑战（∼1.5–2.0×）。

2. **多比特水印在容量和鲁棒性上可以实现兼顾：**Provably Robust Multi-bit在20比特/200 token场景下达到97.6%匹配率，显著优于SOTA的49.2%，打破了传统“容量-鲁棒性-质量”三难问题的认知。

3. **双通道方法可显著降低检测样本量：**Duwak通过并行在概率分布与采样策略双通道嵌入密纹，可将检测所需token数减少∼70%，从∼800降至∼240，显著提升了短文本场景的可用性。

4. **跨语种攻击暴露了现有方法的语言耦合问题：**翻译攻击可将检测AUC从∼0.95降至∼0.67，接近随机水平，揭示了语义-词面跨语迁移的弱项。X-SIR等防御方法可将跨语种AUC提升∼20%（从∼0.67提升至∼0.87），但仍低于单语种性能（∼0.95）。**改进目标：**跨语种防御的AUC需达到≥0.90，接近单语种性能，以满足国际化应用的需求。

5. **公开检测API可能扩大攻击面：**Watermark Stealing等攻击方法在黑盒设置下达到>80%成功率且成本< $50，揭示了公开检测API的安全隐患。多密钥机制和公开验证机制（如UPV）提供了部分解决方案，但仍需权衡安全性和可用性。

6. **无偏水印在特定威胁模型下仍面临挑战：**虽然无偏方法（Unbiased、DiPmark、MCMARK）强调分布不改变，但在多轮生成/低熵段可能累积漂移，或被“利用其保真特性”的策略攻破。需在多批次/编辑模型下进行统一基准复查。

7. **强水印不可能性理论不等于工程不可行：**虽然在自然假设下强水印不可实现，但在现实威胁模型下，通过流程化审计、密钥管理、检测API限流/凭证化、跨语一致性增强等手段，仍可形成足够强且可治理的方案。**工程可行标准：**在允许5%误报率、检测成本< $10/万次检测的场景下，多比特水印可实现有效溯源（匹配率≥95%）。**成本效益分析：**Watermark Stealing攻击成本< $50，若防御方案的成本> $100，则工程价值有限；但在检测成本< $10的场景下，防御方案具有明显优势（成本效益比>10:1）。

场景化建议：根据任务特性选择合适的水印方法：（1）**实时对话场景：**优先选择token级方法（KGW），嵌入开销低（∼1.1×），检测开销小（O(n)），延迟<100ms，适用于客服机器人、实时聊天等场景；（2）**长文本生成场景：**优先选择语义级方法（SemStamp），虽然嵌入开销较高（∼1.5–2.0×），但检测开销小（O(n)），且鲁棒性强（AUC ∼0.85–0.90），适用于文档摘要、新闻生成等场景；（3）**代码生成场景：**优先选择多比特方法（Provably Robust Multi-bit），匹配率∼97.6%，可嵌入用户ID、时间戳等信息，适用于代码生成、API调用追踪等场景；（4）**创意写作场景：**优先选择无偏方法（Unbiased、DiPmark），质量损失最小（Perplexity变化<2%），适用于文学创作、内容生成等场景；（5）**跨语种场景：**优先选择跨语种防御方法（X-SIR），可将跨语种AUC从∼0.67提升至∼0.87，适用于多语言翻译、国际化应用等场景。

研究意义：本文提出的分类框架、定量分析方法和标准化评估框架，为研究人员提供了系统化的

技术路线图，并为未来研究方向提供了明确指导。同时，本文揭示的争议点和挑战，为领域发展提供了重要的参考依据。场景化建议为不同应用场景提供了具体的方法选择指导，有助于提高方法的实际应用价值。

局限性与未来工作：本文的局限性包括：（1）论文筛选标准可能存在主观性，未来可通过多专家评审和自动化筛选方法改进；（2）定量分析基于已有论文报告的数据，可能存在实验设置差异，未来需要统一基准验证；（3）时间范围覆盖至2025年1月，后续研究需要持续更新。未来工作应聚焦于统一基准建立、短文本场景优化、跨语一致性增强、黑盒攻防演练以及硬件/系统协同等方向。

致谢

感谢所有为本研究提供支持的匿名 reviewers 和 contributors。

参考文献

[1] Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Cao, Yiming Ding, Hongyang Zhang, and Heng Huang. SemStamp: A Semantic Watermark with Paraphrastic Robustness for Text Generation. In *Proceedings of NAACL*, 2024. https://aclanthology.org/2024.naacl-long.226/

[2] Zhengmian Hu, Xidong Wu, Yihan Cao, Hongyang Zhang, and Heng Huang. k-SemStamp: A Clustering-based Semantic Watermark with Detection Efficiency. In *Findings of ACL*, 2024.

[3] Anonymous. SemaMark: Semantic Substitution Hash for Paraphrase Robustness. In *Findings of NAACL*, 2024.

[4] Anonymous. PostMark: Post-hoc Semantic Insertion for Text Watermarking. In *Proceedings of EMNLP*, 2024.

[5] Anonymous. Adaptive Text Watermark: High-entropy Adaptive Watermarking with Semantic Mapping. In *Proceedings of ICML*, 2024.

[6] Anonymous. Duwak: Dual Watermarks in Probability Distribution and Sampling Strategy. In *Findings of ACL*, 2024.

[7] Anonymous. GumbelSoft: Improving Diversity in GumbelMax-based Watermarks. In *Proceedings of ACL*, 2024.

[8] Anonymous. MorphMark: Multi-objective Adaptive Watermark Strength. In *Proceedings of ACL (Long)*, 2025.

[9] Google DeepMind. SynthID-Text: Production-scale Text Watermarking with Speculative Sampling. *Nature*, 2024. https://www.nature.com/articles/s41586-024-08025-4.pdf

[10] Anonymous. MarkLLM: Unified Implementation, Visualization, and Evaluation Pipeline. In *Proceedings of EMNLP (System Demonstration)*, 2024.

[11] Anonymous. WaterBench: Fair Comparison Framework for Text Watermarking. In *Proceedings of ACL*, 2024.

[12] Anonymous. Watermark under Fire (WaterPark): Robustness Evaluation Platform. In *Findings of EMNLP*, 2025.

[13] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A Watermark for Large Language Models. In *Proceedings of ICML*, 2023. https://proceedings.mlr.press/v202/kirchenbauer23a/kirchenbauer23a.pdf

[14] Anonymous. On the Reliability of Watermarks for Large Language Models. In *Proceedings of ICLR*, 2024.

[15] Anonymous. Unbiased Watermark: Distribution-preserving Watermarking Paradigm. In *Proceedings of ICLR*, 2024.

[16] Anonymous. DiPmark: Distribution-preserving Reweighting Strategy. In *Proceedings of ICML (Open Review)*, 2024.

[17] Anonymous. MCMARK: Improved Unbiased Watermark with Multi-channel Segmentation. In *Proceedings of ACL (Long)*, 2025.

[18] Anonymous. STA-1: Unbiased & Low-risk Sampling-Then-Accept Watermark. In *Proceedings of ACL (Long)*, 2025.

[19] Anonymous. Watermarks in the Sand: Impossibility of Strong Watermarks. In *Proceedings of ICML*, 2024.

[20] Anonymous. Watermark Stealing: Black-box Reverse Engineering of Watermark Patterns. In *Proceedings of ICML*, 2024.

[21] Anonymous. Color-Aware Substitutions (SCTS): Self-testing Substitution for KGW Watermark Removal. In *Proceedings of ACL*, 2024.

[22] Anonymous. Cross-lingual Consistency (CWRA): Translation Attack and X-SIR Defense. In *Proceedings of ACL*, 2024.

[23] Anonymous. No Free Lunch in LLM Watermarking: Robustness-Usability-Deployability Trilemma. In *Proceedings of NeurIPS*, 2024.

[24] Anonymous. Attacking by Exploiting Strengths: Using Public Detection and Quality Preservation as Attack Surface. In *ICLR Workshop*, 2024.

[25] Anonymous. UPV: Unforgeable Publicly Verifiable Watermarking. In *Proceedings of ICLR*, 2024.

[26] Anonymous. Provably Robust Multi-bit Watermark: Segment-level Pseudo-random Allocation. In *Proceedings of USENIX Security*, 2025.

[27] Anonymous. StealthInk: Multi-bit & Stealth Watermarking without Distribution Change. In *Proceedings of ICML*, 2025.

[28] Anonymous. Multi-User Watermarks: Individual/Collusion Group Tracing. In *IACR ePrint*, 2024.

[29] Ruisheng Zhang, et al. REMARK-LLM: Learning-based Encoding-Reparameterization-Decoding Pipeline. In *Proceedings of USENIX Security*, 2024. https://www.usenix.org/conference/usenixsecurity24/presentation/zhang-ruisi

[30] Anonymous. WaterJudge: Quality-Detection Trade-off Evaluation Framework. In *Findings of NAACL*, 2024.

[31] Anonymous. A Survey of Text Watermarking in the Era of Large Language Models. *ACM Computing Surveys*, 2024. https://dlnext.acm.org/doi/pdf/10.1145/3691626

[32] Lei Li, et al. Tutorial on LLM Watermarking. In *Proceedings of ACL Tutorials*, 2024. https://aclanthology.org/2024.acl-tutorials.6/