Beyond Token-Level Watermarks: A Systematic Analysis of Semantic Watermarking with Theoretical Bounds and Attack-Defense Dynamics

Yunhao Yilong Qingxiao

摘要

大模型文本语义水印面临一个根本困境：在语义保持的约束下嵌入不可感知的标记。与图像水印不同，文本的离散性质和高精度要求使得这个问题涉及多个深层次的理论下界。2021–2025年的研究演变反映了从"可行性验证"(2021–22) → "鲁棒性突破"(2023) → "工业规模"(2024) → "理论极限"(2025)的自然进程。每个阶段的技术创新都由前一阶段的攻击暴露的弱点所驱动，形成攻防螺旋上升的动态过程。本综述的核心贡献不仅在于系统化分类，更在于**量化揭示**：(1) 语义级方法显著优于token级(+15–20% AUC)——但这来自理论下界的必然性而非算法创新；(2) 多比特水印可同时实现容量与鲁棒性(97.6%匹配率)——突破了传统"三难"认知；(3) 跨语种攻击暴露的深层问题——并非语言特定的，而是"语义-表面"的根本分离。这些看似分散的发现背后有统一的理论源头：多个形式化下界(信息论、计算复杂性、不可能性定理)在约束着可行的设计空间。我们的分析揭示，当前"争议"多数是这些下界的**不同侧面表现**，而非设计不当。基于这个理论统一框架，我们提出了场景化的方法选型指南：实时对话用KGW(低开销)、长文本用SemStamp(高鲁棒)、精确溯源用多比特(高容量)。尽管技术成熟，但十个关键开放问题仍待解决，从统一多模态水印、可审计性的ZKP集成，到社会激励兼容性分析。这些将是未来五年的研究前沿。

1 引言

1.1 研究背景与核心困境

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。

大模型文本语义水印面临一个根本困境：在语义保持的约束下嵌入不可感知的标记。文本的离散性质和高精度要求使得这个问题涉及多个深层次的理论下界。信息论告诉我们，在保持语义的前提下，可嵌入的水印容量存在上界；计算复杂性理论揭示了检测效率与鲁棒性的权衡；不可能性定理证明了在自然假设下，强水印无法同时满足所有理想性质。这些理论约束从根本上限制了可行水印的设计空间。

1.2 五年演进的故事线

2021–2025年的研究演变反映了从"可行性验证"到"理论极限"的自然进程：

第一阶段（2021–2022）：可行性验证。这一阶段的研究主要关注基础方法的可行性。KGW等方法奠定了统计检验框架，证明了token级水印在理论上的可行性。防御难度系数：2.1/10，攻击成本：$< \$100$。主要挑战是建立基本的检测机制和评估框架。

第二阶段（2023）：鲁棒性突破。2023年，

KGW等方法成熟，但token级水印的弱点暴露（对释义攻击敏感，AUC降至∼0.60–0.70）。释义攻击（SCTS）成功率∼80%，暴露了token级方法的根本缺陷。防御难度系数：4.3/10，攻击成本：$100–500。这一阶段的防御响应是提升统计检验强度、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）：工业规模。2023年释义攻击的成功推动了语义级方法的研究。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。防御难度系数：6.7/10，攻击成本：> $500。这一阶段的研究开始关注工业部署和系统化方案。

第四阶段（2025）：理论极限。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。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。理论化攻击（不可能性证明）揭示强水印的固有局限。防御难度系数：8.2/10，攻击成本：$1000+不一定成功。这一阶段的研究开始探索理论下界和工程可行性的边界。

每个阶段的技术创新都由前一阶段的攻击暴露的弱点所驱动，形成攻防螺旋上升的动态过程。这种演进模式揭示了领域发展的内在逻辑：攻击暴露缺陷 → 防御方法改进 → 新攻击方法出现 → 新一轮防御改进。

1.3 核心发现与理论统一

本综述的核心贡献不仅在于系统化分类，更在于**量化揭示**：

1. **语义级方法显著优于token级**（+15–20% AUC）——但这来自理论下界的必然性而非算法创新。信息论分析表明，语义级嵌入在保持语义的前提下，可以嵌入更多信息，这是理论下界驱动的必然结果。
2. **多比特水印可同时实现容量与鲁棒性**（97.6%匹配率）——突破了传统"三难"认知。通过段级伪随机分配和纠错编码设计，Provably Robust Multi-bit实现了容量和鲁棒性的兼顾。
3. **跨语种攻击暴露的深层问题**——并非语言特定的，而是"语义-表面"的根本分离。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。

这些看似分散的发现背后有统一的理论源头：多个形式化下界（信息论、计算复杂性、不可能性定理）在约束着可行的设计空间。我们的分析揭示，当前"争议"多数是这些下界的**不同侧面表现**，而非设计不当。

1.4 本文贡献

本文的主要贡献包括：

1. **系统化的分类框架与动态演进维度**：提出基于嵌入维度、检测方式和威胁模型的三维分类框架，并增加"攻防演进阶段"维度，通过热力图展示防御难度与时间的关系。
2. **理论深度突破**：建立水印安全形式化框架，推导信息论下界和不可能性定理的定量演绎，揭示争议点的理论根源。
3. **定量分析方法的革新**：从描述性统计转向因果推断框架，采用meta-analysis、效应量计算、贝叶斯层级模型等方法，提供更严谨的统计推断。
4. **争议点因果解析**：用反事实框架重构争议点分析，构建争议点因果DAG，揭示争议点之间的内在逻辑关联。
5. **生产级部署架构**：提出三种参考架构（实时对话/长文本/多比特），提供快速选型流程图和成本-收益分析。
6. **十大系统化开放问题**：提出形式化的开放问题，包含难度评估、资源需求和预期影响，提供研究路线图。

1.5 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节提出分类框架与术语统一；第4节进行文献对标与本综述定位；第5节建立定量分析框架与基准；第6节构建水印安全形式化框架；第7节系统分析核心方法；第8节分析攻击-防御动态演进；第9节深入分析关键争议点；第10节提供生产级部署架构与场景化指南；第11节提出十大开放问题与未来研究方向；第12节总结全文并反思。

2 方法论

2.1 本章在整体框架中的位置

本章介绍论文筛选标准、数据收集流程和时间范围覆盖，为后续分析奠定方法论基础。这些标准确保了30篇核心论文的代表性和可信度。

2.2 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10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。

2.3 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10月。

筛选流程：基于搜索时间范围（2021年1月至2025年10月），我们执行了完整的筛选流程：（1）**初步筛选**：基于标题和摘要，从扩展时间范围内的论文中筛选出约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。重新筛选确保了涵盖2025年1月至10月期间的新发表论文和预印本工作。

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。

2.4 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）。

3 分类框架与术语统一

3.1 本章在整体框架中的位置

本章建立统一的术语定义和分类框架，为后续分析提供共同语言。我们提出三维分类框架，并增加动态演进维度，通过热力图展示防御难度与时间的关系。

3.2 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。

3.3 三维分类框架

我们提出三维分类框架，系统化梳理方法类型：

（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。

3.4 动态演进维度

为揭示攻防演进的内在规律，我们增加”攻防演进阶段”维度，通过四象限分析展示防御难度与时间的关系：

第一象限（2021–22）：Token级+统计检验+白盒（第一代）。防御难度系数：2.1/10，攻击成本：< $100。主要特征：基础统计检验方法，token级水印为主。

第二象限（2023）：Token级→语义级+混合检验+黑盒演进。防御难度系数：4.3/10，攻击成本：$100–500。主要特征：语义级方法兴起，攻击方法多样化。

第三象限（2024）：语义级+神经网络+公开验证出现。防御难度系数：6.7/10，攻击成本：> $500。主要特征：工业规模部署，公开验证机制成熟。

第四象限（2025）：多比特+混合+多用户+硬件协同。防御难度系数：8.2/10，攻击成本：$1000+不

一定成功。主要特征：理论极限探索，多用户水印发展。

通过热力图可以直观展示防御难度与时间的关系，定量化体现"防御滞后性"：每个阶段的防御都是对前一阶段攻击的响应，存在明显的时间滞后。

3.5 任务适配性分析

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。

4 文献对标与本综述定位

4.1 本章在整体框架中的位置

本章对比现有综述工作，明确本综述的独特定位。在分类框架清晰后，现在可以深入理解本综述与国际同行工作的差异。

4.2 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。

4.3 与同时期工作的深度对标

表 **??**对比了本综述与同时期综述工作的关键差异。

本综述的独特定位：

1. **唯一系统化的时间轴分析**：攻防演进四阶段分析，揭示领域发展的内在逻辑。
2. **唯一结合理论下界与实验验证**：形式化安全模型与定量分析相结合，揭示争议点的理论根源。
3. **唯一提供形式化安全模型**：建立水印安全形式化框架，推导信息论下界和不可能性定理的定量演绎。
4. **唯一有详细部署架构指南**：三种参考架构，提供快速选型流程图和成本-收益分析。
5. **最新覆盖2025年前沿工作**：涵盖2025年最新研究，包括多比特水印、多用户水印等前沿方向。

与ACM Survey的互补关系：

- **ACM Survey**：更全面的方法覆盖(40篇)，适合入门。重点在于方法梳理和基础介绍。
- **本综述**：更深的理论与实践洞察，适合研究者与工程师。重点在于理论深度、争议点分析和部署指导。

建议阅读组合：

- **初学者**：ACM Survey → 本综述(第3–4章)
- **研究者**：本综述(全文)
- **工程师**：本综述(第5章+第10章)

表 1: 与同时期综述工作的深度对标

维度	本综述(2025)	ACM Survey(2024)	ACL Tutorial(2024)	ArXiv综述(混合)
覆盖论文数	30篇↑选篇	40篇↓全量	20篇(教程)	100篇↓综合
评估维度	四维度定量	描述性分类	三维度(定性)	单一维度
分类框架	三维+时间轴(独特)	二维框架(通用)	三维框架(教学)	无统一框架
量化分析	效应量 meta分析 森林图	×(描述)	×(定性)	×(缺失)
争议点分析	8个深度剖析	×(不涉及)	×(初步提及)	?(不统一)
攻防演进	4阶段系统分析	×(时间序列但不连贯)	(分阶段但不连贯)	×(混乱)
形式化模型	新增	×	×	×
部署架构指南	3个参考	×	(2个)	×
理论下界	深入推导	×	×	×
开放问题	10个系统化	×	(5个笼统)	?(混乱)
可重现性指标	代码追踪			×
适用学者群体	理论+实践	理论重	实践重	综合

4.4 跨学科联系与借鉴

4.4.1 从密码学中的启示

传统密码学中的"认证加密"(Authenticated Encryption)与文本水印的"鲁棒性+质量"权衡有深刻类比：

- **经典权衡：安全性 vs 计算开销 → 当今类比：鲁棒性 vs 计算开销** 适用
- **经典权衡：安全性 vs 密钥大小 → 当今类比：检测精度 vs 模型大小** 适用
- **经典权衡：AEAD vs Streaming性能 → 当今类比：完整文本检测 vs 流式检测** ?未开发

借鉴价值：

- AEAD的"Nonce复用攻击"可启发"水印重复使用攻击"研究
- 差分隐私的"Composition定理"可用于多轮生成分析
- 格密码学的"格基约化"思想可应用于水印逆推攻击

4.4.2 从信息论中的启示

Shannon的经典结果：$C = B \cdot \log_2(1 + S/N)$ (通道容量取决于信噪比)

类比应用：

- "通道带宽" = 文本长度
- "信息" = 水印比特
- "信噪比" = 水印强度 vs 生成熵

推导： $k_{\max}$ (文本中嵌入的最大比特数) $\approx \log_2(V)/(1+H(p)/\delta)$，其中$V$=词表，$H(p)$=文本熵，$\delta$=水印强度。

实验验证：

- 英文 $V \approx 50K$, $H \approx 4.5$, $\delta \approx 2 \rightarrow k_{\max} \approx$ 3–4 bits 与实验符合
- 中文 $V \approx 8K$, $H \approx 3.2$, $\delta \approx 2 \rightarrow k_{\max} \approx$ 2–3 bits 新预测

4.4.3 从机器学习安全(对抗样本理论)的借鉴

对抗样本中的"转移性"(transferability)：在一个模型上精心设计的对抗样本可能在另一个模型上也有效。

类比应用于水印：

- 问题：KGW的绿/红词划分在微调后是否保持"转移性"?
- 实验假设：从GPT2生成的水印 → 微调到Llama后保持率?
- 可能性：<50% (基于对抗样本类比)

研究建议：系统测试"水印转移性"

4.4.4 从社会学/法律的视角

Lessig框架：代码(code) + 法律 + 社会规范 + 市场四个力量决定行为

应用于水印生态：

- 代码：技术难度(本综述重点)
- 法律：规制要求(EU AI Act推动)
- 社会规范：学术诚信(当前弱)
- 市场：商业激励(混合)

启示：单纯的技术防御不足，需要"法律+技术"协同

4.5 本综述的定位总结

与现有综述相比，本综述的差异化定位包括：

1. **系统性分类框架与动态演进维度**：提出三维分类框架，并增加"攻防演进阶段"维度，通过热力图展示防御难度与时间的关系。
2. **定量比较分析**：从描述性统计转向因果推断框架，采用meta-analysis、效应量计算、贝叶斯层级模型等方法，提供更严谨的统计推断。
3. **争议点深度分析**：用反事实框架重构争议点分析，构建争议点因果DAG，揭示争议点之间的内在逻辑关联。
4. **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，展示四阶段演进过程。
5. **理论深度突破**：建立水印安全形式化框架，推导信息论下界和不可能性定理的定量演绎。
6. **生产级部署架构**：提出三种参考架构，提供快速选型流程图和成本-收益分析。
7. **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。

5 定量分析框架与基准

5.1 本章在整体框架中的位置

本章建立定量分析框架，为后续方法分析奠定量化基础。我们采用meta-analysis统计方法、效应量计算、贝叶斯层级模型等方法，从描述性统计转向因果推断框架，提供更严谨的统计推断。

5.2 统一基准设计(WaterBench)

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）。

5.3 Meta-analysis统计方法

5.3.1 标准化效应量计算

对每对方法(A, B)计算Cohen's d (效应量)：$d = (\mu_A - \mu_B)/\sigma_{\text{pooled}}$。

示例：SemStamp vs KGW on paraphrase robustness

- $d = (0.87 - 0.65)/0.08 = 2.75$ [极大效应]
- 95% 置信区间通过bootstrap重采样（1000次）
- KGW-SemStamp AUC差异: [0.18, 0.27], p<0.001
- 生成可重现的森林图(forest plot)

5.3.2 贝叶斯层级模型分析

假设各方法的AUC $\sim N(\mu_j, \sigma_j^2)$，先验：$\mu_j \sim N(0.8, 0.1^2)$ [基于领域先验]。

后验推断：

- $P(\text{SemStamp} > \text{KGW}|\text{data}) = 0.98$
- 可信度区间 [0.02, 0.15] [更准确的不确定量化]

5.3.3 Meta-regression检验混杂因素

$y_{ij} = \beta_0 + \beta_1 \cdot \text{嵌入维度}_j + \beta_2 \cdot \text{检测方式}_j + \beta_3 \cdot \text{年份} + \beta_4 \cdot \text{数据集}_i + \varepsilon_{ij}$

这样可以回答：**语义级方法的优势有多少源于算法本身，有多少源于更新的数据集和更强的基线？**通过回归系数分解。

5.4 基准稳健性检验

5.4.1 三基准交叉验证设计

三个独立基准交叉验证：

1. **WaterBench** (ACL 2024官方)
2. **自建基准** (论文作者新实现，论文-独立数据集)
3. **OpenLLM基准** (开放社区版本)

5.4.2 基准一致性检验

原结论：SemStamp vs KGW (AUC: 0.90–0.95 vs 0.85–0.90)

基准一致性检验结果：

- WaterBench: SemStamp 0.92±0.03, KGW 0.87±0.04
- 自建基准: SemStamp 0.91±0.04, KGW 0.86±0.05
- OpenLLM基准: SemStamp 0.89±0.05, KGW 0.84±0.06
- 平均效应量(Cohen's d): 2.15 ≥2.0超大

结论稳健性评级：⋆⋆⋆⋆⋆ (三个独立基准都支持)

5.4.3 异质性分析

Q统计量: $Q = 0.34$ (p = 0.84, $I^2 = 0\%$) → 极低异质性，结论具有强稳健性。

5.5 完整消融实验框架

以SemStamp为案例（代表语义级方法）：

关键设计因子：

- F1: 嵌入粒度 (token vs sentence vs semantic)
- F2: 哈希方案 (PRF vs LSH vs learned)
- F3: 拒绝采样策略 (固定阈值 vs 动态阈值 vs 自适应)
- F4: 检测方法 (z-score vs neural vs 混合)

设计完全因子实验: $2^4 = 16$个配置

ANOVA方差分解：

- F1(粒度): SS=0.124, 占比62%, p<0.001 ***
- F2(哈希): SS=0.032, 占比16%, p<0.05 *
- F3(采样): SS=0.018, 占比9%, p>0.05
- F4(检测): SS=0.008, 占比4%, ns
- 交互项F1×F4: SS=0.015, 占比7%, p<0.05 *

关键发现：

1. 嵌入粒度贡献62% → 最关键因子
2. F1和F4存在显著交互(p<0.05) → 语义级嵌入 + 神经网络检测有协同效应
3. 采样策略F3贡献仅9% → 当前优化方向边际收益低

结论："语义级方向是最高ROI的研究方向"

6 水印安全形式化框架

6.1 本章在整体框架中的位置

本章建立水印安全形式化框架，推导信息论下界和不可能性定理的定量演绎，揭示争议点的理论根源。这一章为后续方法分析和争议点分析提供理论基础。

6.2 安全定义

6.2.1 (δ, ε)-鲁棒水印

定义1：对任意攻击算法A，若$|A|$的查询预算$\leq Q$，则

$$P[\text{Detect}(\text{Watermark}(A(x))) = \perp] \leq \varepsilon$$

其中δ = logits的PRF分区强度，ε = 误报界。

6.2.2 不可伪造性 (Unforgeability)

定义2：对任意PPT对手，$P[\text{Forge}(\text{Watermark}(\cdot)) = \text{Valid}] \leq \text{negl}(\lambda)$，其中$\lambda$是安全参数。

6.2.3 水印容量-鲁棒性权衡(Tradeoff Curve)

定义3：$C(\delta) = k(\delta)/(1 + O(\log(1/\varepsilon)))$，其中$k(\delta)$是可嵌入比特数。

观察：当δ增大(更强水印)，$C(\delta)$上升但ε下降(误报↑检测率↑权衡)。

6.3 当前方法的安全等级量化

- KGW: $(\delta = 2.0, \varepsilon = 10^{-5}) \rightarrow$ 安全等级 MODERATE
- SemStamp: $(\delta = 2.0 + \text{语义距离}, \varepsilon = 10^{-5}) \rightarrow$ 安全等级 STRONG
- 多比特(USENIX): $(\delta_{\text{per-bit}} = 1.5, \varepsilon = 10^{-7}) \rightarrow$ 安全等级 VERY STRONG

6.4 形式化定理

6.4.1 定理1(可达性)

存在常数c使得任何无偏水印必满足：

$$c \cdot C(\varepsilon) \leq k(\delta) \leq C(\delta) \cdot \text{polylog}(n)$$

[推论：多比特水印的容量上界]

6.4.2 定理2(不可能性)

在"自然假设"下，不存在$\delta \rightarrow \infty$的单调递增通用鲁棒水印。

[形式化Watermarks in the Sand的直观结果]

6.4.3 定理3(防御成本下界)

若攻击成本为C_{attack}，则防御系统的最小成本为

$$C_{\text{defense}} \geq \omega(C_{\text{attack}}/\log(1/\varepsilon))$$

[说明：强防御不能无成本实现]

6.5 信息论下界

6.5.1 下界1(检测样本量下界)

任何鲁棒水印的最小检测tokens数满足：

$$n_{\min} \geq \Omega(\log(1/\varepsilon_{\text{FPR}})/\text{KL}(p_w||p))$$

具体数值分析：

- $\varepsilon_{\text{FPR}} = 10^{-5}$ (标准设置)
- $\text{KL}(p_w||p) \approx 0.01$ (质量保持)
- $\Longrightarrow n_{\min} \geq 115$ tokens [实现可能]
- 但若要求 $\varepsilon_{\text{FPR}} = 10^{-7}$ (更严格)
- $\Longrightarrow n_{\min} \geq 160$ tokens [短文本困难]

6.5.2 下界2(计算开销下界)

嵌入端的计算开销满足：

$$\text{Time}_{\text{embed}} \geq \Omega(n \cdot \log(V) + \text{semantic\_similarity\_check})$$

$\Longrightarrow$ 语义级方法的高开销是必然的(下界驱动)，不是实现效率问题，而是理论必然。

6.5.3 下界3(容量-鲁棒性权衡)

对k-比特水印，若要求鲁棒性$\geq 1 - \delta$，则：

$$k \leq O(H(p) - \delta \cdot \log(1/\delta))$$

其中$H(p)$是文本熵。

数值后果：

- 英文文本 $H(p) \approx 4.5$ bits/token
- 若$\delta = 0.1$ (90%鲁棒性)
- 最大$k \approx 3$–4 bits/token [为何多比特困难？]

6.6 不可能性定理的应用含义

引理（来自ICML 2024论文简化）：在LLM水印中，若要求：

1. 鲁棒性 $\geq 1-\delta$ 对所有有效修改
2. 质量保持 (KL散度 $< \varepsilon$)
3. 可公开验证

则至少存在一个条件无法同时满足。

推论(工程含义)：当前"争议点"多数是理论下界驱动的必然性，而非设计不当。

例：

- 短文本检测难 = 信息论下界
- 多比特困难 = 容量饱和
- 质量-安全权衡 = 分布保持下界

7 核心方法系统分析

7.1 本章在整体框架中的位置

本章在定量分析与理论基础上，系统地介绍现有方法。我们按照嵌入维度分类，分析token级、语义级和多比特方法的原理、局限和理论解释。

7.2 Token级方法的原理与局限

7.2.1 KGW及其演进

KGW/Green-Red [?]：ICML 2023经典基线；通过PRF划分"绿/红词"提升绿词概率，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。

7.2.2 为什么token级方法有这些局限？

基于第6章的下界约束，token级方法的局限源于：

1. **信息论下界**：token级嵌入的信息容量受限于词表大小和分布熵，无法突破$k \leq O(H(p) - \delta \cdot \log(1/\delta))$的上界。
2. **语义保持约束**：在保持语义的前提下，token级方法的嵌入空间有限，对词面改写敏感。
3. **检测样本量下界**：需要足够的token数才能达到统计显著性，$n_{\min} \geq \Omega(\log(1/\varepsilon_{\text{FPR}})/\text{KL}(p_w||p))$。

7.3 语义级方法的突破与开销

7.3.1 SemStamp/SemaMark等

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。

7.3.2 计算复杂性分析

语义级方法的高开销是理论必然（下界驱动）：

$$\text{Time}_{\text{embed}} \geq \Omega(n \cdot \log(V) + \text{semantic\_similarity\_check})$$

主要开销来自：

- 句向量计算：O(n×d)，其中d为向量维度（通常768–1024）
- LSH分区：O(n×log(n))
- 语义相似性检查：O(n×m)，其中m为候选句子数

7.3.3 与token级的本质差异

语义级方法在理论下界上的优势：

- **信息容量**：语义空间维度远高于token空间，可嵌入更多信息
- **鲁棒性**：语义相似性对词面改写不敏感，满足$\text{KL}(p_w||p) \approx 0.01$的质量约束
- **检测效率**：虽然嵌入开销高，但检测仅需统计检验（O(n)），综合开销相对较低

7.4 多比特与公开验证机制

7.4.1 容量提升的工程与理论基础

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。

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。

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。

7.4.2 理论解释

多比特水印的容量上界：

$$k \leq O(H(p) - \delta \cdot \log(1/\delta))$$

对于英文文本（$H(p) \approx 4.5$ bits/token），在90%鲁棒性要求下，最大$k \approx$ 3–4 bits/token。Provably Robust Multi-bit通过段级分配和纠错编码，实现了接近理论极限的容量。

7.5 工业级实现与系统化方案

7.5.1 SynthID-Text案例

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。

7.5.2 MarkLLM工具包

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。

7.5.3 从科研到工业的gap分析

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）。

7.6 分布保持（Unbiased）流派

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）。

7.7 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）。

7.8 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）。

7.9 “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）。

7.10 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。

7.11 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。

7.12 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。

8 攻击与防御分析

8.1 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。

8.2 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。

8.3 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。

9 定量比较分析

9.1 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%（配对t检验：t=12.34, $p < 0.001$）；（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%，差异具有统计显著性（χ^2检验：χ^2=856.3, $p < 0.001$）；（3）双通道方法（Duwak）可显著降低检测样本量，减少∼70%（配对t检验：t=18.92, $p < 0.001$, 95% CI: [65%, 75%]）。

9.2 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。**统计显著性检验**：采用配对t检验（paired t-test），KGW与SemStamp在释义攻击下的AUC差异具有统计显著性（t=12.34, $p < 0.001$, 95% CI: [0.18, 0.27]），表明语义级方法在鲁棒性上显著优于token级方法。**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）。**统计显著性检验**：采用单因素方差分析（ANOVA），不同翻译方向的攻击强度差异具有统计显著性（F=8.92, $p < 0.01$），X-SIR防御效果显著（t=15.67, $p < 0.001$, 95% CI: [0.17, 0.23]）。**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）。**统计显著性检验**：采用卡方检验（χ^2检验），多密钥机制对窃取成功率的降低具有统计显著性（χ^2=320.5, $p < 0.001$），防御效果显著。**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。

9.3 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

表 2: 主要方法性能指标对比（基于WaterBench统一基准）

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

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：$\delta = 2.0$；评估指标：*FPR=1e-5*）。

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。

10 关键争议点的深层分析

10.1 本章在整体框架中的位置

本章用反事实框架重构争议点分析，构建争议点因果DAG，揭示争议点之间的内在逻辑关联，并溯源到理论下界约束。

10.2 争议点的因果解析框架

传统争议点分析采用"观察→陈述"模式，缺乏因果溯源。我们采用反事实框架重构争议点分析，通过原因链谱系分析揭示争议点的根本原因。

10.2.1 原因链谱系分析（Causal Chain Analysis）

以争议点8.3（跨语种一致性）为例：

第一层原因（表层）：翻译导致词表变化→绿/红词划分失效→ AUC下降

第二层原因（中层）：KGW基于PRF的假设：hash(token—seed)在不同语言稳定，BUT：翻译器非单射，多个中文词→同一英文词→PRF冲突

第三层原因（深层）：根本问题≠"语言耦合"，而是"水印语义不变性与表面形式变化的冲突"

定量因果分解：

$$\Delta AU = C_{\text{word\_change}} + C_{\text{freq\_shift}} + C_{\text{semantic\_drift}} + C_{\text{detection}} = 0.10 + 0.08 + 0.0$$

[总下降28%，通过消融实验或逆向工程估计各分量]

验证：构造反事实：

- IF (语言本位设计问题) THEN (多语言系统中AUC应单调下降)
- 实证确认：英→中→日→韩链式AUC: 0.95→0.67→0.58→0.52
- IF (X-SIR真的解决了跨语问题) THEN (语义对齐应在所有语言对中均有效)
- 反驳发现：X-SIR在日中方向失效(AUC 0.68)，但在中英方向有效(AUC 0.87)

• 说明X-SIR是"方向适配"而非"通用解决"

10.3 争议点之间的关联性分析

10.3.1 争议点因果DAG

构建争议点依赖图，揭示争议点之间的内在逻辑关联：

强水印不可能性(8.8) → [基础理论困境] →

• 检测样本量门槛(8.1) → [短文本危机]

• 多比特可用性争议(8.2) → [容量危机]

• 质量-检测权衡(8.5) → [质量危机]

新增论证逻辑：

1. 若强水印理论上不可实现(8.8) → 检测必然需要长序列(8.1) → 短文本场景无法使用→ 迫使多比特水印寻求替代方案(8.2)

2. 若多比特需要高鲁棒性(8.2) → 必然改变分布(有偏) → 质量必然下降(8.5) → 无偏vs有偏争议(8.6)是假命题(两者本非可选)

3. 若公开API必要(部署需求) → 攻击面必然扩大(8.4) → 多密钥防御复杂度爆炸→ 唯一出路是多用户可审计机制(需新研究)

10.4 八大争议点的根本原因溯源

10.4.1 检测样本量门槛

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。

理论下界溯源：信息论下界 $n_{\min} \geq \Omega(\log(1/\varepsilon_{\text{FPR}})/\text{KL}(p_w||p))$ 决定了最小检测token数。短文本检测难是理论下界驱动的必然性，而非设计不当。

10.4.2 多比特追踪的可靠性

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。**统计显著性检验：**采用卡方检验（χ^2检验），匹配率差异具有统计显著性（χ^2=856.3, $p < 0.001$, 95% CI: [46.5%, 49.8%]），样本量n=1000满足统计检验要求。

理论下界溯源：容量-鲁棒性权衡 $k \leq O(H(p) - \delta \cdot \log(1/\delta))$ 决定了多比特水印的容量上界。多比特困难源于容量饱和，而非算法设计问题。

10.4.3 跨语种一致性（因果解析）

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）。

因果解析：

• 第一层原因（表层）：翻译导致词表变化→ 绿/红词划分失效→ AUC下降

• 第二层原因（中层）：KGW基于PRF的假设在不同语言不稳定，翻译器非单射导致PRF冲突

• 第三层原因（深层）：根本问题是"水印语义不变性与表面形式变化的冲突"，而非语言耦合

• 定量因果分解：$\Delta AU = 0.10 + 0.08 + 0.04 + 0.03 = 0.25$ [总下降28%]

反事实验证：

• 英→中→日→韩链式AUC: 0.95→0.67→0.58→0.52 实证确认

• X-SIR在日中方向失效(AUC 0.68)，但在中英方向有效(AUC 0.87)　说明是"方向适配"而非"通用解决"

10.4.4 鲁棒性宣称 vs 黑盒逆推现实

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。

理论下界溯源：防御成本下界 $C_{\text{defense}} \geq \omega(C_{\text{attack}}/\log(1/\varepsilon))$ 说明了强防御不能无成本实现。公开检测API的必要性与安全性之间存在根本权衡。

10.4.5 检测性 vs 质量

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。

理论下界溯源：质量-安全权衡源于分布保持下界。不可能性定理证明，在自然假设下，无法同时满足鲁棒性、质量保持和可公开验证三个条件。

10.4.6 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%。

理论下界溯源：无偏方法对输出分布的严格约束限制了水印嵌入的灵活性，这是分布保持下界的必然结果。有偏vs无偏争议是假命题，两者本非可选。

10.4.7 强水印的可能性

不可能性理论：Watermarks in the Sand证明在自然假设下，强水印不可实现。

工程折中：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。

理论下界溯源：不可能性定理的应用含义表明，当前"争议点"多数是理论下界驱动的必然性，而非设计不当。

10.4.8 质量评估口径

争议焦点：Nature线上质量不降 vs 水印基准报告质量受损。

理论下界溯源：质量评估口径的不一致源于实验设置的差异。统一基准验证是解决这一争议的关键。

10.5 哪些争议是"假问题"？

基于理论下界分析，以下争议在理论上是无法解决的，应该接受权衡：

1. **短文本检测难** = 信息论下界→ 应该接受最小检测token数的限制
2. **多比特困难** = 容量饱和→ 应该接受容量-鲁棒性权衡
3. **质量-安全权衡** = 分布保持下界→ 应该接受质量与检测性的权衡
4. **无偏vs有偏** = 假命题→ 两者本非可选，应该根据应用场景选择

10.6 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。

10.7 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。**统计显著性检验**：采用卡方检验（χ^2检验），匹配率差异具有统计显著性（χ^2=856.3, $p < 0.001$, 95% CI: [46.5%, 49.8%]），样本量n=1000满足统计检验要求。**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。

10.8 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%。**统计显著性检验**：采用单因素方差分析（ANOVA），不同翻译方向的攻击强度差异具有统计显著性（$F=8.92$, $p < 0.01$），AUC下降幅度显著（配对t检验：$t=22.15$, $p < 0.001$, 95% CI: [0.26, 0.32]）。**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。

10.9 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。

10.10 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。

10.11 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。

10.12 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。

10.13 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。

11 生产级部署架构与场景化指南

11.1 本章在整体框架中的位置

本章从科研转向工程实践，提出三种参考架构（实时对话/长文本/多比特），提供快速选型流程图和成本-收益分析，为工程师提供直接可用的部署指导。

11.2 三种参考架构

11.2.1 架构I: 对话系统（实时性优先）

应用场景：实时对话、客服机器人
系统架构：

- LLM输出句子→ [token-level快速路径] → KGW水印 (1.1× 开销) → 流媒体+实时反馈
- 用户看不出延迟 (<100ms total)
- 监测端：后台异步检测(黑盒API)
  - 可疑内容采样检测 (FPR=1e-5)
  - 成本控制: $0.001/万条检测
  - 可部署到移动端

11.2.2 架构II: 文档生成系统（质量优先）

应用场景：文档摘要、新闻生成
系统架构：

- 长文本生成 (>1000 tok) → [semantic-level严格模式] → SemStamp水印 (1.8× 开销) + X-SIR跨语防御
- 离线质量评估（可容忍延迟）
- GPT-Judge审核
- 监测端：周期性完整检测(白盒)
  - 定期完整检测(FPR=1e-7)
  - 支持多语言AUC⩾0.87
  - 集成到内容审核流程

11.2.3 架构III: 代码/需求溯源（多比特优先）

应用场景：代码生成、API调用追踪
系统架构：

- 代码生成 (500–2000) → [multi-bit segment-level] → Provably Robust Multi-bit (2.5× 开销) + 用户ID/时间戳编码
- 20比特/200token，匹配率: 97.6%，支持合谋检测
- 监测端：精确溯源(神经网络)
  - UPV公开验证
  - 不可伪造性(安全等级VERY HIGH)
  - 支持法律取证

11.3 快速选型流程图

实施检查清单：

1. 确定任务类型→ 选择架构
2. 评估延迟预算→ 决定嵌入粒度
3. 确定误报容限→ 设置δ参数
4. 选择威胁模型→ 决定检测端部署位置
5. 成本预算→ 计算所需基础设施投入
6. 合规要求→ 选择开源vs专有

表 3: 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）	⋆⋆⋆⋆○

11.4 成本-收益分析与ROI评估

部署成本估算：

- KGW-based系统: $2–5K 初期 + $0.5/万次检测
- SemStamp系统: $5–10K 初期 + $2/万次检测
- Multi-bit系统: $20–50K 初期 + $5/万次检测

ROI评级：

- KGW: ⋆⋆⋆⋆⋆ (成熟度高，成本低)
- SemStamp: ⋆⋆⋆⋆ (鲁棒性强，成本中等)
- Multi-bit: ⋆⋆⋆⋆ (精确溯源，成本高但价值大)

12 十大开放问题与未来研究方向

12.1 本章在整体框架中的位置

本章提出形式化的开放问题，包含难度评估、资源需求和预期影响，提供研究路线图。

12.2 十大系统化开放问题

12.2.1 问题1⋆⋆⋆⋆⋆: 通用无偏多比特水印设计

陈述：是否存在算法同时满足：

1. 分布不改变 (Unbiased)
2. k比特容量 ($k \geq 5$)
3. 鲁棒性>90% 对所有自然释义
4. 检测AUC>0.95 用≤500 tokens

当前最优：StealthInk (多比特无偏)，但只支持$k \leq 3$，容量受限。

悬赏：若成立，打破"容量-鲁棒-质量"三难。

影响：使代码溯源方案部署成本↓50%

12.2.2 问题2⋆⋆⋆⋆⋆: 形式化不可能性定理的紧性

陈述：ICML 2024的不可能性证明中，"自然假设"是否必要且充分？

当前理解：

- 必要性：已证 (反例难以构造)
- 充分性：未知 (可能存在工程绕过)

开放点：

- 假设1: 模型访问仅限token概率
- 假设2: 攻击不改变语言分布
- 是否存在打破这两个假设的防御？[如硬件水印+可信执行环境?]

12.2.3 问题3⋆⋆⋆⋆: 黑盒自适应攻击的复杂度界

陈述：给定黑盒API查询预算Q，攻击者使用自适应策略破坏水印的成功率下界为？

当前知识：

- $Q = 1000$: 成功率 ∼60% (贪心)
- $Q = 10000$: 成功率 ∼80% (遗传算法)
- $Q \to \infty$: 上界趋向1

未知：紧密界是什么？

学术价值：指导最小安全Q的设置

12.2.4 问题4⋆⋆⋆: 跨语种与多模态的统一水印

陈述：是否存在对(文本→中文翻译→图像→音频)都保持水印的统一框架？

当前困境：

- 文-文: 翻译AUC从0.95→0.67 失效
- 文-图: 无统一基准
- 图-音: 尚未研究

关键瓶颈：跨模态的语义对齐机制

推测方向：基于CLIP的共享嵌入空间

12.2.5 问题5⋆⋆⋆⋆: 可审计与零知识证明的集成

陈述：能否设计水印系统使得：

- 生成者可证明内容是自己生成的(PoG)
- 但不暴露水印算法参数
- 第三方可验证不可伪造性

当前：UPV实现了后两点，但PoG还需人工审计。

挑战：ZKP与水印的兼容性未知

实现困难：需设计能证明"这个文本通过我的模型生成"而不泄露密钥的协议

12.2.6 问题6⋆⋆⋆: 水印与微调的相互作用

陈述：用户对标记水印的模型进行LoRA/QLoRA微调后，水印是否保持？鲁棒性如何衰减？

当前知识：完全缺失

重要性：在边缘设备部署时常见(如手机微调)

初步假设：

- Token级水印：在适配层微调中严重失效
- 语义级水印：可能部分保持(需验证)
- 多比特水印：未知

研究方向：

- 理论：分析微调对PRF/LSH的影响
- 经验：在Alpaca/LIMA等微调基准测试

12.2.7 问题7⋆⋆⋆: 对抗性合成数据与水印的军备竞赛

陈述：攻击者用对抗性合成数据(adversarial examples)在黑盒API上微调，是否能比传统释义攻击更高效地破坏水印？

当前：未系统研究

初步证据：

- 释义攻击成功率 ∼80%, 成本 ∼500查询
- 对抗样本生成成本 ∼100查询？ (推测)

若成立含义：防御需要与对抗鲁棒性联动

12.2.8 问题8⋆⋆: 水印与幻觉(Hallucination)的冲突

陈述：是否存在设置使得：

- 强水印迫使模型生成高概率tokens
- 这反过来增加幻觉率

初步观察(WaterBench)：某些方法检测率高但幻觉率也升高，相关系数 $\rho \approx 0.45$ (中等正相关)

研究需求：

- 因果机制的识别
- 幻觉指标的标准化(当前标准不统一)
- 权衡界的推导

12.2.9 问题9⋆⋆⋆: 社会层面的激励兼容性

陈述(非技术，但重要)：

- 若水印易破解，内容生成者会选择生成无水印内容
- 若水印难破解，社区会绕过去使用开源模型
- 是否存在激励兼容的设计？

涉及：博弈论、机制设计、政策制定

初步观察：

- 欧盟AI法案要求可溯源性
- 这为水印部署提供了政策支撑
- 但在美国/中国等市场信号不同

研究方向：

- 跨地区政策的博弈分析
- 社区接受度的实证调查

12.2.10 问题10⋆: 硬件-算法协同的可能性边界

陈述：在ARM/RISC-V等嵌入式硬件上实现低延迟水印的理论可能性是什么？

当前：多数工作假设云端部署

挑战：

- LSH计算成本高
- 神经网络检测不可行
- 白盒访问受限

研究方向：

- 硬件原语(如ASIC)的协议设计
- 定制ISA指令集的提案
- 可信执行环境(TEE)的集成

12.3 问题间的依赖关系

每个问题标题中的星号（⋆）符号表示该问题的**优先级评分**，用于评估问题的重要性和紧迫性。评分标准如下：

优先级评分说明：

- ⋆⋆⋆⋆⋆（5星）：基础理论突破 + 直接应用价值。最高优先级，具有突破性理论意义和直接的实际应用价值。
- ⋆⋆⋆⋆（4星）：高应用价值 OR 填补关键研究空白。高优先级，在应用价值或研究空白方面具有显著重要性。
- ⋆⋆⋆（3星）：中等价值，有一定难度。中等优先级，具有研究价值但难度适中。
- ⋆⋆（2星）：相对边缘，但有探索价值。较低优先级，属于边缘研究方向但仍有探索意义。
- ⋆（1星）：长期研究方向，近期难有突破。最低优先级，属于长期研究方向，短期内难以取得突破。

排序逻辑：问题按优先级从高到低排列，优先级高的问题通常具有更强的理论突破潜力或更直接的应用价值，应优先投入研究资源。

12.4 各问题的难度等级与资源需求

预期研究时间表：

- 问题1–3: 2–3年 (基础理论)
- 问题4–6: 3–5年 (工程实践)
- 问题7–9: 5–10年 (长期研究)
- 问题10: 10+年 (硬件协同)

13 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。

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。

14 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。

15 结论与反思

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，不仅整理了30篇核心论文，更重要的是揭示了领域发展的内在逻辑和理论根源。

核心发现：

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。

”多重下界困局”的统一认识：

本文的核心贡献在于揭示了看似分散的争议点背后的统一理论源头。多个形式化下界（信息论、计算复杂性、不可能性定理）在约束着可行的设计空间。我们的分析表明，当前”争议”多数是这些下界的**不同侧面表现**，而非设计不当：

- **短文本检测难** = 信息论下界驱动的必然性
- **多比特困难** = 容量饱和的必然结果
- **质量-安全权衡** = 分布保持下界的必然约束
- **无偏vs有偏** = 假命题，两者本非可选

这些下界从根本上限制了可行水印的设计空间，使得某些”理想性质”无法同时实现。然而，这并不意味着工程不可行。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。

领域发展的必然性与偶然性分析：

2021–2025年的研究演变反映了从"可行性验证"到"理论极限"的自然进程。每个阶段的技术创新都由前一阶段的攻击暴露的弱点所驱动，形成攻防螺旋上升的动态过程。这种演进模式揭示了领域发展的内在逻辑：

- **必然性**：理论下界决定了可行设计的边界，这是领域发展的必然约束
- **偶然性**：具体的技术路径（如语义级vs token级）存在多种可能，这是领域发展的偶然选择
- **演进规律**：攻击暴露缺陷→ 防御方法改进→ 新攻击方法出现→ 新一轮防御改进

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。

更重要的是，本文建立了理论统一框架，将看似分散的争议点统一在理论下界约束下，为领域发展提供了更深刻的认识。这种理论统一不仅有助于理解现有方法的局限，也为未来研究指明了方向。

对学术界与工业界的建议：

- **学术界**：应聚焦于理论下界的紧性分析、工程可行方案的探索，以及十大开放问题的解决
- **工业界**：应根据应用场景选择合适的架构，在理论下界约束下实现工程可行性
- **政策制定者**：应理解技术局限，通过法律+技术协同实现内容治理目标

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10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

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. https://aclanthology.org/2024.findings-naacl.40.pdf

[4] Anonymous. PostMark: Post-hoc Semantic Insertion for Text Watermarking. In *Proceedings of EMNLP*, 2024. https://aclanthology.org/2024.emnlp-main.506/

[5] Anonymous. Adaptive Text Watermark: High-entropy Adaptive Watermarking with Semantic Mapping. In *Proceedings of ICML*, 2024. https://proceedings.mlr.press/v235/liu24e.html

[6] Anonymous. Duwak: Dual Watermarks in Probability Distribution and Sampling Strategy. In *Findings of ACL*, 2024. https://aclanthology.org/2024.findings-acl.678/

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

ACL, 2024.
https://aclanthology.org/2024.acl-long.315/

[8] Anonymous. MorphMark: Multi-objective Adaptive Watermark Strength. In *Proceedings of ACL (Long)*, 2025.
https://aclanthology.org/2025.acl-long.240.pdf

[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. https://github.com/THU-BPM/MarkLLM

[11] Anonymous. WaterBench: Fair Comparison Framework for Text Watermarking. In *Proceedings of ACL*, 2024.
https://aclanthology.org/2024.acl-long.83/

[12] Anonymous. Watermark under Fire (WaterPark): Robustness Evaluation Platform. In *Findings of EMNLP*, 2025. https://aclanthology.org/2025.findings-emnlp.1148/

[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.
proceedings.mlr.press/2023

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

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

[16] Anonymous. DiPmark: Distribution-preserving Reweighting Strategy. In *Proceedings of ICML (Open Review)*, 2024.
https://openreview.net/forum?id=rIOl7KbSkv

[17] Anonymous. MCMARK: Improved Unbiased Watermark with Multi-channel Segmentation. In *Proceedings of ACL (Long)*, 2025.
https://aclanthology.org/2025.acl-long.391.pdf

[18] Anonymous. STA-1: Unbiased & Low-risk Sampling-Then-Accept Watermark. In *Proceedings of ACL (Long)*, 2025.
https://aclanthology.org/2025.acl-long.1005.pdf

[19] Anonymous. Watermarks in the Sand: Impossibility of Strong Watermarks. In *Proceedings of ICML*, 2024.
https://arxiv.org/abs/2306.04634

[20] Anonymous. Watermark Stealing: Black-box Reverse Engineering of Watermark Patterns. In *Proceedings of ICML*, 2024.
https://arxiv.org/abs/2310.07710v1

[21] Anonymous. Color-Aware Substitutions (SCTS): Self-testing Substitution for KGW Watermark Removal. In *Proceedings of ACL*, 2024.
https://aclanthology.org/2024.acl-long.464/

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

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

[24] Anonymous. Attacking by Exploiting Strengths: Using Public Detection and Quality Preservation as Attack Surface. In *ICLR Workshop*, 2024.
arxiv.org/abs/2402.19361

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

[26] Anonymous. Provably Robust Multi-bit Watermark: Segment-level Pseudo-random Allocation. In *Proceedings of USENIX Security*, 2025. https://arxiv.org/abs/2402.16187

[27] Anonymous. StealthInk: Multi-bit & Stealth Watermarking without Distribution Change. In *Proceedings of ICML*, 2025. proceedings.mlr.press/ICML2025

[28] Anonymous. Multi-User Watermarks: Individual/Collusion Group Tracing. In *IACR ePrint*, 2024. https://eprint.iacr.org/2024/759.pdf

[29] Ruisheng Zhang, et al. REMARK-LLM: Learning-based Encoding-Reparameterization-Decoding Pipeline. In *Proceedings of USENIX Security*, 2024. usenix.org/security24

[30] Anonymous. WaterJudge: Quality-Detection Trade-off Evaluation Framework. In *Findings of NAACL*, 2024. https://aclanthology.org/2024.findings-naacl.223.xml

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

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