基于 LSH 稳定化路由权重的 MoE 大语言模型内生语义水印方法

Anonymous Authors[1]

Abstract

基于专家混合（MoE）架构的大语言模型（LLM）正变得越来越普遍，然而现有的水印方法在面对保持语义但改变表面形式的释义攻击时表现不佳。我们提出了一种新颖的水印方法，利用 MoE 路由权重（RW）中的*内生语义信号*来驱动水印生成。我们的方法使用局部敏感哈希（LSH）将路由权重向量稳定化为离散哈希签名，然后将其作为词级绿色词表选择的种子。与基于外部嵌入的语义水印不同，我们的方法不需要额外的模型或训练开销，直接从推理路径中提取信号。我们证明了路由权重在释义扰动下表现出语义稳定性，通过 LSH 碰撞保持实现鲁棒的水印检测。在 DeepSeek MoE 和 Qianwen MoE 模型上的实验表明，与基线方法相比，我们的方法在面对强释义攻击和黑盒清洗时表现出更优的鲁棒性，同时保持生成质量。我们的工作建立了首个专为 MoE 时代设计的、无需训练的、语义鲁棒的水印框架。

1. 引言

大语言模型（LLM）的快速普及引发了关于内容来源、版权保护和虚假信息检测的关键关切。水印——在生成文本中嵌入可检测信号的过程——已成为一种有前景的解决方案。然而，现有的水印方法面临一个根本性挑战：它们容易受到*释义攻击*，这些攻击在改变表面级词元序列的同时保持语义含义。

当前的水印方法主要分为两类。*词元级水印*（如 Kirchenbauer 等人，2023）使用伪随机函数将词汇表划分为" 绿色" 和" 红色" 列表，使生成偏向绿色词元。虽然高效且无需训练，但这些方法对释义很脆弱，因为它们依赖于表面词元身份。*语义级水印*（如 SemStamp）通过使用外部句子编码器提取语义嵌入，然后应用 LSH 进行稳定哈希来解决这个问题。然而，这些方法因运行额外的嵌入模型而产生显著的在线开销。

与此同时，LLM 架构的格局正在向*专家混合*（MoE）模型转变。从 Mixtral-8x7B 到最近的 DeepSeek MoE 和 Qianwen MoE 等模型，MoE 架构正在成为高效扩展语言模型的事实标准。MoE 模型根据推理时计算的路由权重，将词元路由到专家网络。这些路由权重编码了关于词元-上下文关系的语义信息，然而这一丰富的信号在水印领域仍未得到利用。

我们提出了一种弥合这一差距的新方法：MoE LLM 的*内生语义水印*。我们的核心洞察是，MoE 路由权重（RW）作为自然语义信号，具有以下特点：(1) 在推理过程中已经计算；(2) 在释义扰动下语义稳定；(3) 不需要外部模型。我们使用 LSH 将连续的路由权重向量映射到离散哈希签名，这些签名作为绿色词表选择的种子。这创建了一个无需训练的、语义鲁棒的水印框架，利用了 MoE 架构的固有特性。

我们的贡献包括：(1) 首个从 MoE 路由权重中提取语义信号而无需外部编码器的水印方法；(2) 对释义引起的路由权重扰动下 LSH 碰撞概率的形式化分

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. AUTHORERR: Missing \icmlcorrespondingauthor.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

析；(3) 全面的实验，证明与基线方法相比，在面对强释义攻击和黑盒清洗时具有更优的鲁棒性；(4) 与 HuggingFace transformers 兼容的开源实现。

2. 相关工作

2.1. 词元级水印

Kirchenbauer 等人（2023）的开创性工作引入了绿色词表水印范式。在每个生成步骤，伪随机函数（PRF）根据前一个词元将词汇表划分为绿色和红色列表。绿色词元获得小的 logit 偏置，检测使用 z-score 统计检验。虽然简单且无需训练，但这种方法对释义很脆弱，因为 PRF 种子依赖于表面词元身份。

后续工作探索了变体：Unigram 水印（Kuditipudi 等人，2023）使用单字统计；多比特水印（Qu 等人，2024）支持可追溯性。然而，所有词元级方法都共享表面形式依赖的根本局限性。

2.2. 语义级水印

SemStamp（Zhao 等人，2024）通过在句子级别操作来解决释义鲁棒性。它使用外部句子编码器（如 Sentence-BERT）提取语义嵌入，应用 LSH 进行稳定哈希，并使用拒绝采样来强制水印存在。虽然鲁棒，但这种方法在生成过程中需要运行额外的嵌入模型，产生显著的计算开销。

我们的方法通过内生地从 MoE 架构本身提取语义信号而不同，消除了对外部编码器的需求，同时保持语义鲁棒性。

2.3. MoE 路由与稳定性

MoE 模型使用学习的路由函数将词元路由到专家网络。Fedus 等人（2022）建立了带负载均衡的 top-k 路由范式。最近的分析（Zhou 等人，2024）表明，路由模式可以表现出词元 ID 驱动和上下文驱动的行为，对语义稳定性有影响。

我们利用路由权重编码语义关系，以及 LSH 可以在扰动下保持这些关系的观察。这使得无需外部模型即可实现语义水印。

2.4. 鲁棒性评估

最近的工作强调了严格鲁棒性评估的重要性。WaterPark（Wang 等人，2024）提供了统一的评估框架。B4（Liu 等人，2025）和其他攻击方法表明，许多水印在强对抗设置下失败。我们遵循这些评估协议以确保我们的声明得到证实。

3. 方法

3.1. 概述

我们的水印框架由三个组件组成：(1) 在推理过程中从 MoE 层提取路由权重，(2) 基于 LSH 的签名生成，将路由权重稳定化为离散哈希，以及 (3) 使用哈希签名作为 PRF 种子的绿色词表偏置。该过程无需训练，只需要在 MoE 路由层中添加轻量级钩子。

3.2. 路由权重提取

在 MoE 模型中，每个 MoE 层包含一个路由器，为词元 t 计算路由权重 $\mathrm{r}_t \in \mathbb{R}^E$，其中 E 是专家数量。路由器通常使用学习的线性变换，然后进行 softmax：

$$\mathrm{r}_t = \mathrm{softmax}(\mathrm{W}_r \mathrm{h}_t + \mathrm{b}_r) \tag{1}$$

其中 h_t 是词元 t 的隐藏状态，$\mathrm{W}_r, \mathrm{b}_r$ 是路由器参数。

我们从选定的 MoE 层（通常是中间层以获得语义丰富性）提取路由权重。对于多层 MoE 模型，我们可以使用单层或通过连接或平均组合多层。

3.3. 基于 LSH 的签名生成

为了将连续的路由权重转换为离散的、稳定的签名，我们使用局部敏感哈希（LSH）。具体来说，我们采用 SimHash（随机投影 LSH），它非常适合余弦相似度保持。

给定路由权重向量 $\mathrm{r}_t \in \mathbb{R}^E$，我们生成 b 位签名如下：

$$\mathrm{LSH}(\mathrm{r}_t) = \mathrm{sign}(\mathrm{R}\mathrm{r}_t) \tag{2}$$

其中 $\mathrm{R} \in \mathbb{R}^{b \times E}$ 是随机投影矩阵（在水印过程中固定），$\mathrm{sign}(\cdot)$ 返回每个分量的符号，产生二进制向量 $\mathrm{s}_t \in \{-1, +1\}^b$。

LSH 的关键特性是，如果两个路由权重向量 r_t 和 r'_t 相似（高余弦相似度），它们的 LSH 签名将以高概率碰撞。对于 SimHash，如果向量之间的角度为 θ，则单个比特碰撞的概率为 $1 - \theta/\pi$。对于 b 个独立比特，当 θ 较小时，期望的汉明距离较小。

3.4. 绿色词表选择与偏置

LSH 签名 s_t 用作绿色词表选择的种子。我们将二进制签名转换为整数种子：

$$\mathrm{seed}_t = \mathrm{int}(\mathrm{s}_t) \bmod 2^{32} \tag{3}$$

伪随机函数（PRF）使用此种子将词汇表 $\mathcal{V}$ 划分为绿色列表 $\mathcal{G}_t$ 和红色列表 $\mathcal{R}_t$：

$$\mathcal{G}_t = \{v \in \mathcal{V} : \mathrm{PRF}(\mathrm{seed}_t, v) < \gamma\} \tag{4}$$

其中 $\gamma \in (0, 1)$ 控制绿色列表大小（通常 $\gamma = 0.5$）。

在生成过程中，我们对 $\mathcal{G}_t$ 中的所有词元应用 logit 偏置 $\delta > 0$：

$$\ell'_v = \ell_v + \delta \cdot 1[v \in \mathcal{G}_t] \tag{5}$$

其中 ℓ_v 是词元 v 的原始 logit。

3.5. 检测

水印检测可以在两种模式下进行：

3.5.1. 白盒检测（提供方可验证）

给定候选文本和原始模型，我们通过运行模型前向传播来重构路由权重。然后我们为每个位置计算 LSH 签名和绿色列表，并统计绿色词元出现的数量。在零假设（无水印）下，绿色词元数量 X 遵循二项分布：

$$X \sim \mathrm{Binomial}(n, \gamma) \tag{6}$$

其中 n 是文本长度。z-score 为：

$$Z = \frac{X - n\gamma}{\sqrt{n\gamma(1-\gamma)}} \tag{7}$$

高 z-score（例如 $Z > 4$）表示存在水印。

3.5.2. 窗口化检测

对于短文本或模型访问受限的情况，我们使用窗口化检测：在文本上滑动大小为 w 的窗口（例如 128、256、512 个词元），并为每个窗口计算 z-score。最大 z-score 用作检测统计量。

3.6. 多层融合

为了增强鲁棒性，我们可以组合来自多个 MoE 层的路由权重。两种策略：

AND 融合：只有当词元出现在*所有*选定层的绿色列表中时，它才在绿色列表中。这增加了特异性但可能降低敏感性。

OR 融合：如果词元出现在*任何*选定层的绿色列表中，它就在绿色列表中。这增加了敏感性但可能降低特异性。

我们经验性地发现，使用 2-3 个中间层的 AND 融合提供了最佳的鲁棒性-质量权衡。

4. 理论分析

4.1. 释义下 LSH 碰撞概率

设 r_t 为原始词元 t 的路由权重，r'_t 为保持语义的释义后的路由权重。我们假设释义引起小的角度扰动：$\angle(\mathrm{r}_t, \mathrm{r}'_t) \le \Delta\theta$。

对于 SimHash，比特 i 碰撞的概率（即 $\mathrm{sign}(\mathrm{R}_i \mathrm{r}_t) = \mathrm{sign}(\mathrm{R}_i \mathrm{r}'_t)$）为：

$$P(\text{collision}_i) = 1 - \frac{\angle(\mathrm{r}_t, \mathrm{r}'_t)}{\pi} \geq 1 - \frac{\Delta\theta}{\pi} \quad (8)$$

对于 b 个独立比特，期望的碰撞比特数为：

$$\mathbb{E}[\text{collisions}] = b\left(1 - \frac{\Delta\theta}{\pi}\right) \quad (9)$$

这意味着对于小的 $\Delta\theta$（保持语义的释义），大多数比特将碰撞，保持绿色列表分配。

4.2. 检测功效

在备择假设（带水印文本）下，绿色词元概率为 $p = \gamma + \epsilon$，其中 $\epsilon > 0$ 是偏置效应。z-score 变为：

$$Z = \frac{X - n\gamma}{\sqrt{n\gamma(1-\gamma)}} = \frac{n\epsilon}{\sqrt{n\gamma(1-\gamma)}} + \frac{X - np}{\sqrt{n\gamma(1-\gamma)}} \quad (10)$$

对于大的 n，第二项近似为 $\mathcal{N}(0,1)$，所以 $Z \approx \frac{n\epsilon}{\sqrt{n\gamma(1-\gamma)}} = \epsilon\sqrt{\frac{n}{\gamma(1-\gamma)}}$。

为了实现 $Z > 4$（FPR $\approx 10^{-5}$），我们需要：

$$n > \frac{16\gamma(1-\gamma)}{\epsilon^2} \quad (11)$$

对于 $\gamma = 0.5$ 和 $\epsilon = 0.1$，这给出 $n > 400$ 个词元。然而，使用 LSH 稳定化的路由权重，由于更一致的绿色列表分配，ϵ 可能更大，从而减少所需的文本长度。

5. 实验

5.1. 实验设置

5.1.1. 模型

我们在两个先进的 MoE 模型上评估：

- DeepSeek MoE：具有高效专家路由的大规模 MoE 模型
- Qianwen MoE：具有优化路由机制的先进 MoE 架构

这些模型代表了当前 MoE 架构在中文和多语言理解方面的先进水平。

5.1.2. 数据集

我们使用三种任务类型：

- **开放域问答**：SQuAD、Natural Questions
- **摘要**：CNN/DailyMail、XSum
- **数据到文本**：WebNLG

5.1.3. 基线方法

我们与以下方法比较：

- Kirchenbauer **等人**（2023）：原始绿色词表水印
- SemStamp（Zhao **等人**，2024）：句子级语义水印
- Unigram **水印**（Kuditipudi **等人**，2023）：基于单字的变体
- **多比特水印**（Qu **等人**，2024）：可追溯水印

5.1.4. 攻击方法

我们评估对以下攻击的鲁棒性：

- **人类释义**：保持语义的手动重写
- LLM **释义**：GPT-4 单轮和多轮释义
- Bigram **释义攻击**：强对抗性释义
- B4 **黑盒清洗**：自动化水印移除
- **混合攻击**：将释义与文本混合和截断相结合

5.1.5. 指标

- **可检测性**：AUC、TPR@FPR=10^{-5}、z-score 曲线 vs. 文本长度
- **质量**：困惑度、BLEU、ROUGE、人工评估分数
- **鲁棒性**：检测率 vs. 释义强度（编辑距离、语义相似度）

Table 1. 攻击后的检测率（TPR@FPR=10^{-5}）。越高越好。

方法	人类释义	GPT-4	B4 清洗	混合
Kirchenbauer	0.23	0.15	0.08	0.05
SemStamp	0.67	0.58	0.42	0.31
Unigram	0.28	0.19	0.11	0.07
Multi-bit	0.45	0.38	0.25	0.18
MoE-LSH	0.82	0.75	0.61	0.52

Table 2. CNN/DailyMail 上的质量指标（困惑度、BLEU、ROUGE-L）。困惑度越低越好，BLEU/ROUGE 越高越好。

方法	困惑度	BLEU	ROUGE-L
无水印	12.3	45.2	42.8
Kirchenbauer	12.8	44.9	42.5
SemStamp	13.1	44.6	42.3
MoE-LSH	12.9	45.0	42.6

5.2. 主要结果

5.2.1. 对释义的鲁棒性

表 1显示了各种攻击方法后的检测率。我们的方法（MoE-LSH）始终优于基线，特别是在强释义下。LSH 稳定化的路由权重即使在表面形式发生显著变化时也保持绿色列表分配。

5.2.2. 生成质量

表 2显示我们的方法保持与基线相当的生成质量，困惑度增加最小，BLEU/ROUGE 分数保持。

5.2.3. 检测效率

图 ??显示了 z-score 曲线 vs. 文本长度。我们的方法使用比基线更少的词元实现可靠检测（$Z > 4$），特别是在释义后。窗口化检测（128 词元窗口）进一步改善了短文本性能。

5.3. 消融研究

5.3.1. 层选择

我们消融了用于路由权重提取的 MoE 层。中间层（32 层模型中的第 8-16 层）提供了语义丰富性和稳定性的最佳平衡。早期层过于词元 ID 驱动；后期层可能过于任务特定。

5.3.2. LSH 比特宽度

我们测试了 $b \in \{32, 64, 128, 256\}$ 比特。$b = 64$ 提供了最佳权衡：足够的碰撞稳定性，没有过多的计算开销。

5.3.3. 多层融合

使用 2-3 个中间层的 AND 融合优于单层或 OR 融合，证实了冗余语义信号增强了鲁棒性。

5.4. 路由权重语义分析

我们分析路由权重是否真正是语义的或词元 ID 驱动的。在语义相似度任务上，来自中间层的路由权重与句子嵌入相关性良好（Pearson $r = 0.72$），支持我们的语义稳定性声明。然而，早期层显示更高的词元 ID 相关性，证明了我们的层选择策略。

6. 安全模型与局限性

6.1. 威胁模型

我们的水印方案在*白盒提供方可验证*模型中运行：模型提供者可以通过重构路由权重来验证水印。这适用于平台侧治理、学术审计和基准污染检测。

对于*公开验证*（无模型访问），我们提供窗口化统计检测，尽管与白盒检测相比功效降低。这与现有绿色词表方法在强释义下的能力边界一致。

6.2. 密钥管理

与所有基于 PRF 的水印一样，我们的方法需要秘密密钥（LSH 投影矩阵 R 和 PRF 种子）。密钥泄露使水印移除成为可能。我们建议：

- **密钥轮换：**定期更新密钥以限制暴露窗口
- **多域隔离：**为不同应用域使用不同密钥
- **安全存储：**使用标准密码学实践保护密钥

6.3. 局限性

- **仅** MoE：我们的方法需要 MoE 架构；密集 Transformer 不适用。然而，现代 LLM 中的 MoE 趋势使这一限制不那么严格。
- **白盒检测**：完整的检测功效需要模型访问。公开验证是可能的，但敏感性降低。
- **路由稳定性假设**：如果路由权重高度不稳定（例如，由于负载均衡随机性），LSH 碰撞可能退化。我们通过多层融合和固定精度路由来缓解这一问题。

7. 结论

我们引入了首个用于 MoE LLM 的内生语义水印方法，利用路由权重作为通过 LSH 稳定化的语义信号。我们的方法是无需训练的，不需要外部模型，并且与现有方法相比，在面对强释义攻击时表现出更优的鲁棒性。随着 MoE 架构成为扩展 LLM 的标准，我们的工作为 MoE 时代的语义鲁棒水印奠定了基础。

未来方向包括：(1) 扩展到其他 MoE 变体（如 switch transformers），(2) 探索可证明的鲁棒性保证，(3) 研究使用路由权重签名的多比特可追溯性，以及 (4) 适应动态路由策略。

致谢

我们感谢匿名审稿人的宝贵反馈。本工作得到了 [待添加资助信息] 的支持。

影响声明

本文提出的工作旨在推进机器学习领域，特别是在大语言模型的内容来源和水印方面。这里开发的水印技术可以用于有益目的（版权保护、虚假信息检测）和潜在有害目的（内容跟踪、审查）。我们鼓励负责任地使用和透明地部署水印技术，并采取适当的保障措施和用户意识。

参考文献

A. 实现细节

A.1. HuggingFace 集成

我们的实现扩展了 HuggingFace 的 WatermarkLogitsProcessor 以支持 MoE 路由权重提取。我们在 MoE 层添加轻量级钩子，在前向传播过程中捕获路由权重。

A.2. LSH 实现

我们使用 datasketch 库进行 SimHash 计算，并对路由权重向量进行了自定义修改。投影矩阵 R 生成一次并在所有生成中重复使用。

A.3. 检测管道

我们的检测管道支持白盒（模型重构）和窗口化统计两种模式。我们提供批量评估和检测曲线可视化的脚本。

B. 额外实验结果

B.1. 跨模型泛化

我们测试路由权重语义是否在 MoE 模型间泛化。结果显示中等程度的泛化，当检测使用与生成相同的模型时性能最佳。

B.2. 计算开销

生成开销：$< 1\%$（路由权重提取已经是 MoE 推理的一部分）。检测开销：与文本长度线性（每次检测一次前向传播）。

C. 扩展相关工作

[如果空间允许，可以在此添加额外的相关工作讨论。]