High-Dimensional Geometric Analysis of Seed Sensitivity in MoE Semantic Watermarking

Anonymous Authors[1]

混合专家模型与语义哈希水印中的随机种子敏感性

Abstract

基于混合专家模型（Mixture-of-Experts, MoE）的大语言模型（LLM）已在工业界和学术界广泛部署，水印技术被视为区分机器生成文本与人类创作文本的关键工具。以 SemHash 为代表的语义水印方法利用局部敏感哈希（Locality-Sensitive Hashing, LSH）在嵌入空间中构造“红/绿”区域，在理论上具有较强的改写鲁棒性。然而，实际系统表明此类方法在 MoE LLM 上呈现出极端的随机种子敏感性：极少数种子可达到 SOTA 生成质量和检测性能，而大部分随机初始化会导致困惑度爆炸或检测失效。与其将种子视作单纯超参数，本文从高维几何和流形学习的角度系统刻画这一现象。我们指出：LLM 嵌入空间呈现强烈的各向异性（“锥体效应”），而标准 LSH 预设的各向同性假设在此完全失效；随机超平面在高维各向异性锥体上极大概率出现“区域坍塌”，要么将整个语义簇标记为绿、无法注入熵，要么整体标记为红、迫使模型生成语义崩溃的 Token。在 MoE 架构下，水印诱导的语义偏移进一步干扰专家路由，放大了种子敏感性。为刻画和缓解该问题，我们提出一套基于语义簇分割熵、Logits 分布 Wasserstein 距离与 PCA 对齐度的“几何质量”指标，并给出白化变换、PCA 对齐 LSH 与基于聚类的非线性划分等数据依赖方案。本文的结论是：种子敏感性不是偶然噪声，而是“各向异性流形 + 各向同性投影”几何错配的必然结果，解决路径应从被动选种转向主动重构几何和投影方向。

[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.

1. 引言：语义水印的几何不稳定性

大语言模型（LLM）的生成能力快速提升，使得内容溯源与版权保护问题日益突出。水印（Watermarking）通过在生成过程中注入可检测但难以察觉的统计信号，被视为当前最具可行性的治理手段之一。经典的 Token 级水印（如 green-list 方案）通过伪随机函数将词表划分为“绿表/红表”，在生成时对绿表 Token 加性偏置，并在检测阶段进行统计显著性检验。该类方法计算开销低、训练开箱即用，却对保持语义不变的改写攻击（Paraphrase Attack）极为脆弱。

为应对这一缺陷，近期工作转向 语义级水印。典型地，SemStamp/SemHash 通过外部句向量编码器将文本映射到 $\mathbb{R}^d$ 中连续语义嵌入，再使用 LSH 将连续空间量化为离散哈希桶，用以驱动绿表选择。在理想化模型中，若语义嵌入在单位超球上各向同性分布，则随机超平面可以稳定地按角度划分语义邻域，从而在语义邻近 Token 之间实现“低损失替换”。

MoE 场景下的工程痛点。 当上述方法被部署到 MoE LLM（如 Mixtral、DeepSeek MoE、Qianwen MoE）时，实践中暴露出一个严重问题：水印性能对随机种子（决定 *LSH* 超平面）呈现极端敏感性。少数“幸运种子”可以在几乎不损伤生成质量的前提下实现高检测率，而大部分随机种子要么几乎检测不到水印，要么显著拉高困惑度、产生语义错乱输出。

本文尝试回答：这种种子敏感性是偶然现象，还是高维几何结构的必然结果？如果是后者，我们是否可以在训练前、部署前对种子进行几何质量评估与筛选，甚至通过重新设计哈希空间来消除此类敏感性?

1.1. MoE 架构下的特殊性：路由几何的放大效应

混合专家模型（MoE）在每一层包含多个专家网络和一个路由门控网络。对于给定 Token 隐状态 $\mathbf{h}_t$，路由器计算专家权重 $\mathbf{r}_t \in \mathbb{R}^E$ 并选择 top-k 专家进行激活。该机制一方面带来参数高效扩展，另一方面引入了高度结构化的几何边界：路由决策本质上由一系列高维超平面刻画。

当在 MoE 上叠加 SemHash 风格的水印平面时，系统中同时存在两套几何划分：

- 内生的 路由平面，决定哪些专家被激活；
- 外加的 水印平面，决定哪些 Token 被视为“绿表候选”。

若水印平面在语义空间中强行抑制了原本高概率的候选 Token，将导致 $\mathbf{h}_t$ 发生非局部偏移，使得路由器激活与上下文语义不匹配的专家。 我们将这种现象称为 路由错位（*Routing Misalignment)*。 在长上下文生成中，这种错位具有级联效应：一处水印诱导的错误选择可在后续层层放大，导致整体语义漂移。

因此，MoE 场景下的种子敏感性不仅反映在单步 Logits 分布被扰动，更体现在路由轨迹这一复杂几何对象被扭曲。

1.2. 本文贡献

围绕上述现象，本文从高维几何与流形学习的角度提出以下贡献：

1. 几何机理刻画： 揭示 LLM 嵌入空间中“锥体效应”与 LSH 各向同性假设之间的根本矛盾，给出随机超平面在各向异性锥体上高概率导致“区域坍塌”的几何推导。
2. 语义碎片化分析： 建立语义聚类与 Logits 选择的几何模型，解释糟糕种子如何通过切断高概率语义簇而迫使模型选择语义噪声 Token。
3. 指标体系设计： 提出三类可在部署前计算的种子质量指标：语义簇分割熵、Logits 分布的 Wasserstein 距离以及投影方差／PCA 对齐度。
4. 几何修正方案： 讨论白化变换、PCA 对齐 LSH 以及基于质心的非线性划分作为缓解种子敏感性的技术路径。

2. 高维语义空间的几何特性

2.1. 理想各向同性与现实各向异性

经典 LSH 理论多基于如下假设：数据点 $x \in \mathbb{R}^d$ 在单位超球 $\mathbb{S}^{d-1}$ 上各向同性分布。 以 SimHash 为例，在该假设下，两向量夹角 $\theta(x, y)$ 与哈希碰撞概率近似线性关系：

$$\mathbb{P}[h(x) = h(y)] = 1 - \frac{\theta(x, y)}{\pi}. \tag{1}$$

随机投影向量 $r \sim \mathcal{N}(0, I_d)$ 被视为在超球上均匀采样，其对应的超平面以“公平”的方式切割数据。

然而，大量实证工作表明：BERT、GPT、Mixtral 等模型的上下文嵌入远非各向同性，而是呈现显著的 各向异性（*Anisotropy)*。 记嵌入矩阵为 $X \in \mathbb{R}^{N\times d}$，协方差为

$$\Sigma = \frac{1}{N} X^\top X. \tag{2}$$

在各向异性场景下，特征值谱表现为极少数主成分 $\lambda_1, \lambda_2, \ldots$ 占据绝大多数能量，其余特征值快速衰减。这意味着数据实际集中在一个远低于 d 维的流形上，并且该流形具有显著朝向，可视为被“压缩”在一个开口角极小的锥体 $\mathcal{K}$ 内。

更直观地，对任意两个随机词向量 u, v，经验上有

$$\mathbb{E}[\cos(u, v)] \approx 1 - \Delta, \quad \Delta \ll 1, \tag{3}$$

与各向同性高维空间“随机向量近似正交”的结论相反。

2.2. 随机投影在锥体分布下的失效

考虑法向量 $r \sim \mathcal{N}(0, I_d)$ 对应的超平面 H_r 与数据锥体 $\mathcal{K}$ 的相互位置。 根据高维测度集中的经典结果，r 与锥体主轴方向 μ 的夹角 ϕ 高度集中于 $\pi/2$ 附近，即大部分随机方向与 μ 近似正交。

然而，要让 H_r 有效地对数据产生“划分”作用，平面必须穿过锥体内部；否则所有样本将全部落在同侧。若锥体张角为 $\alpha \ll 1$，则只有当 r 落入一个极窄的“赤道带”时，平面才会与锥体发生非平凡交集。 由此可见：

- 绝大多数随机种子对应的平面 完全掠过锥体，导致锥体内所有点同侧。
- 少数“幸运种子”恰好对应穿过锥体的平面，才能实现有效二分。

区域坍塌（**Region Collapse**）。 在极端各向异性下，随机平面对数据的典型行为可概括为两种情形：

情形 **A**：失效 超平面完全位于锥体一侧。此时所有语义合理的候选 Token 被统一标记为绿或红：

- 若全为绿，则水印对生成无约束，检测信号消失；
- 若全为红，则模型被迫在语义无关的尾部分布中寻找“可用” Token，生成质量崩溃。

情形 **B**：有效 超平面侥幸穿过锥体，在局部实现了非平凡切割，从而允许在语义邻近 Token 间实施偏置。

MoE 场景下实验观察到的“好种子极少、坏种子占绝大多数”，正是情形 B 占比随维度与各向异性程度指数级下降的体现。

3. 语义碎片化与 Logits 选择

用户关注的第一个问题是：糟糕的种子如何从几何上破坏语义聚类，并在 Logits 层面推动模型选择语义噪声 Token。

3.1. 语义聚类的几何刻画

在 next-token 预测中，给定上下文 C，模型输出 Logits 向量 $\mathbf{z}$，高概率 Token 往往集中在一个或多个紧密簇中。例如对于"The cat sat on the ..."，最高概率 Token 集合

$$\mathcal{T}_{\text{top}} = \{\text{mat}, \text{rug}, \text{floor}, \text{sofa}\} \tag{4}$$

在嵌入空间中形成一个半径极小的超球 $\mathcal{B}_\epsilon(c)$，其中 c 为簇中心、ϵ 很小。

3.2. 碎片化切割与强制降级

设由种子 S 决定的超平面 H_S 将空间划分为水印"绿区" V_{G} 与"红区" V_{R}。理想情况下，H_S 对每个语义簇都实现近似 50%/50% 切分，使得至少存在若干同义词落在绿区中，模型可在低语义代价下替换 Token。但在锥体背景与随机投影机制下，更典型的却是以下两种极端：

- 整个高质量簇 $\mathcal{B}_\epsilon(c)$ 被压入红区，绿区只包含语义无关或低概率 Token；
- 整个簇进入绿区，无法注入任何水印信息。

记最优 Token 为 t^*，其 Logit 为 $z_{\max}$，次优同义词为 t'，Logit 为 $z_{\text{sub}} \approx z_{\max}$。对绿区 Token 施加偏置 $\delta > 0$ 后，生成概率为

$$P(t) \propto \exp\left(z_t + \delta \cdot \mathbb{I}[t \in V_{\text{G}}]\right). \tag{5}$$

当 $t^* \in V_{\text{R}}$ 且所有高质量同义词均被切到红区时，模型被迫在 V_{G} 内选择 Logit 最大的 t_{noise}，其原始 Logit $z_{\text{noise}} \ll z_{\max}$。若

$$z_{\text{noise}} + \delta > z_{\max}, \tag{6}$$

则模型输出 t_{noise}，表现为明显的语义漂移甚至胡言乱语。我们将此过程称为 强制降级（*Forced Downgrade)*。

3.3. 低熵任务中的区域坍塌

在低熵生成任务（例如结构化摘要、事实问答）中，模型的高质量输出空间往往近似退化为一个极小邻域甚至"点"。当 $\epsilon \to 0$ 时，任意超平面对该"点簇"的切割要么是"全进全出"，要么几乎不产生有效边界。此时 SemHash 试图在该簇内部注入熵的企图注定失败：要么完全无法检测，要么在某一关键步整体封锁高质量空间，引发灾难性生成错误。

4. 各向异性与随机投影的低效性

4.1. 投影方差的极度不均衡

LSH 的区分能力依赖于投影值

$$y = r^\top x \tag{7}$$

在数据分布上的方差：

$$\text{Var}(y) = r^\top \Sigma r. \tag{8}$$

在强各向异性下，Σ 的能量集中在前 k 个主成分 $u_1, \ldots, u_k$ 上。高维几何告诉我们：对随机向量 r，其在任一固定方向上的投影 $\langle r, u_i \rangle$ 期望极小。因此，对绝大多数种子而言，$\text{Var}(y)$ 接近 0，即"切了空气而非数据"。

类比三维"黄瓜"：有效的水印平面应沿长轴方向切割，而随机平面大概率平行于长轴且远离黄瓜本体，或仅掠过其边缘。

4.2. 碰撞概率退化与辨识度丧失

在各向异性锥体中，任意两个有效语义向量 x_1, x_2 的夹角 θ 极小（例如 $\theta < 15°$）。由 SimHash 碰撞关系可得

$$\mathbb{P}[h(x_1) = h(x_2)] = 1 - \frac{\theta}{\pi} \approx 1, \tag{9}$$

意味着对绝大部分种子，所有语义相关向量被哈希到相同桶中。这带来两方面后果：

- 辨识度丧失：水印难以在不同语义状态之间制造差异；
- 鲁棒性两极化：若该桶被标记为绿，则极难检测；若被标记为红，则极易毁坏生成。

因此，在 LLM 的各向异性嵌入空间中，未经修正的随机投影在理论上注定效率低下且不稳定。

5. 量化种子几何质量的指标体系

为了在部署前筛选种子，我们提出三类互补的"几何质量"指标。

5.1. 语义簇分割熵：刻画碎片化程度

首先在一个校准数据集上，用无水印模型构造词表语义聚类：对 Token 嵌入做 k-means，得到 K 个簇 $C_1, \ldots, C_K$。对给定种子，计算全词表红绿划分 $V_{\text{G}}, V_{\text{R}}$。对每个簇 C_i，定义绿表比例

$$r_i = \frac{|C_i \cap V_{\text{G}}|}{|C_i|}. \tag{10}$$

理想情况下 $r_i \approx 0.5$。我们定义分割熵得分为

$$\text{Score}_{\text{split}} = 1 - \frac{1}{K} \sum_{i=1}^{K} \left(2|r_i - 0.5|\right)^2. \tag{11}$$

当所有簇被整体压入绿表或红表时，$r_i \in \{0, 1\}$，得分趋近 0，意味着严重区域坍塌；得分越接近 1，说明切分越均匀，语义碎片化风险越低。

5.2. Logits 分布的 Wasserstein 距离：刻画"推土机距离"

考虑单步生成中，原始 Softmax 分布 P 与施加水印偏置后的分布 Q。令 $D_{ij} = \|\mathrm{emb}(i) - \mathrm{emb}(j)\|_2$ 表示 Token 之间的语义距离矩阵。则 Wasserstein-1 距离为

$$W_1(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|_2], \qquad (12)$$

其中 $\Pi(P,Q)$ 为所有以 P 与 Q 为边缘分布的联合分布集合。直观上，W_1 度量"将概率质量从 P 推到 Q 需要搬运的语义距离总量"。

- 好的种子使 W_1 保持较小，仅在同义词等近邻 Token 间重新分配概率；
- 坏种子导致 W_1 显著增大，说明概率质量被迫迁移到语义遥远的 Token 上。

在实践中，可在代表性任务与步长上估计 W_1 的期望或分位数，将其作为种子筛选的硬约束。

5.3. 投影方差与 PCA 对齐度：刻画各向异性适应度

在校准集上收集嵌入矩阵 X，对给定种子对应的投影向量 r，计算

$$h = Xr, \quad \sigma^2_{\text{proj}} = \mathrm{Var}(h). \qquad (13)$$

若 σ^2_{proj} 极小，说明 r 近似垂直于数据流形，平面主要切到"空旷区域"；反之，较大的方差意味着平面沿着数据主要变化方向切割，更有可能穿过锥体核心。

进一步，可利用 PCA 主成分 $u_1, \ldots, u_k$ 定义

$$\mathrm{Align}(r) = \sum_{i=1}^{k} \cos^2(r, u_i), \qquad (14)$$

该值越大，说明种子越"对齐"于高能量方向，从几何上更有希望获得稳定且高熵的哈希划分。

6. 解决种子敏感性的几何方案

基于前述分析，我们认为解决种子敏感性的关键不是"试出一个好种子"，而是重构数据几何或投影机制，使大部分种子都变得可用。

6.1. 白化变换：修正嵌入几何

在进行 LSH 之前，对嵌入向量进行白化：

$$\tilde{x} = W(x-\mu), \quad W = \Sigma^{-1/2}, \qquad (15)$$

其中 μ 与 Σ 分别为均值与协方差估计。白化操作将原本高度各向异性的锥体拉伸为近似各向同性的球体。在此新空间中，随机投影重新符合经典 LSH 假设，使种子敏感性大幅降低。代价主要为一次性离线估计 W 和在线前向中的轻量线性变换。

6.2. PCA 对齐的 LSH：修正投影方向

另一条路径是放弃完全均匀的随机 r，而从数据主成分中抽取投影向量。具体地，在校准集上对嵌入做 PCA，取前 k 个主成分 $u_1, \ldots, u_k$ 作为 LSH 的法向量族。这种 PCA-LSH 方案保证每个超平面都沿着数据方差最大的方向切割，从而必然与锥体核心发生交集，避免"切空气"。

在此基础上，可对主成分进一步做随机旋转或正交变换，以在保证高方差的前提下引入足够的密钥空间。

6.3. 基于质心的非线性划分：拥抱数据拓扑

更激进的策略是完全放弃线性超平面，将水印划分建立在数据驱动的聚类结构上。例如 k-SemStamp 提出：

1. 对语义嵌入做 k-means 聚类，获得簇中心 $\{c_j\}_{j=1}^{K}$；
2. 使用 Voronoi 划分定义语义单元，再随机给簇分配"红/绿"标签。

由于聚类本身追求簇内紧致性、簇间分离度，该方案天然避免了在簇内部"硬切割"，从根本上消解语义碎片化问题。其缺点是需要离线聚类和在线最近质心查询。

7. 结论

本文从高维几何和流形学习的视角，对 MoE LLM 中基于 SemHash 的语义水印所呈现的极端随机种子敏感性进行了系统分析。我们的结论可以概括为：

- LLM 嵌入空间的"锥体效应"与 LSH 的各向同性假设存在根本冲突，使得随机超平面在高维中高概率导致"区域坍塌"，要么无效、要么毁灭性破坏生成；
- 语义碎片化源于超平面正交于语义流形主轴，切断高概率语义簇并强制模型在尾部分布中选取噪声 Token，在 MoE 场景下还会诱导路由错位；
- 通过语义簇分割熵、Wasserstein 距离与投影方差／PCA 对齐度等指标，可以在部署前对种子几何质量进行定量评估；
- 真正可行的解决路径应从"试运气选种"转向"重构几何与投影"：通过嵌入白化、PCA 对齐 LSH 或基于聚类的非线性划分，使大多数种子在数学上都具有稳定且可解释的行为。

从更宏观的角度看，MoE 水印中的种子敏感性问题并非孤立现象，而是"低维流形数据 + 高维随机机制"组合下常见的几何陷阱。我们希望本文的分析能够为后续设计鲁棒、可解释且几何自洽的水印方案提供理论基线。

致谢

本工作在匿名评审阶段暂不列出具体致谢对象。

影响声明

本研究聚焦于面向大语言模型的水印技术，其潜在影响具有双刃性。一方面，鲁棒水印有助于内容溯源、版权保护与错误信息检测；另一方面，不当使用也可能带来隐私、审查与滥用风险。我们鼓励在公开透明的前提下部署水印系统，结合法律与伦理框架，避免将技术用于牺牲用户自主权的场景。

References