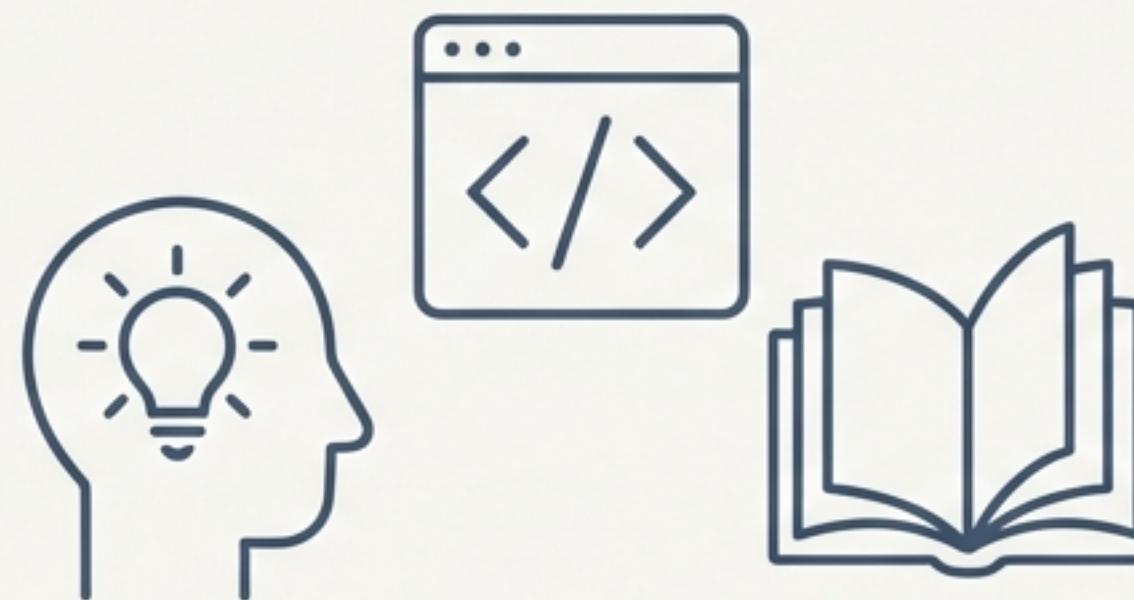


高效鲁棒的语义水印：一种基于 MoE路由权重与LSH的新范 式

AIGC的信任困境：当内容可被无限改写，我们如何追溯其源头？

机遇



风险



大型语言模型（LLM）的爆发带来了巨大的机遇，但也伴随着严峻的风险：难以分辨的虚假新闻、恶意的网络言論、以及严重的学术欺诈行为。

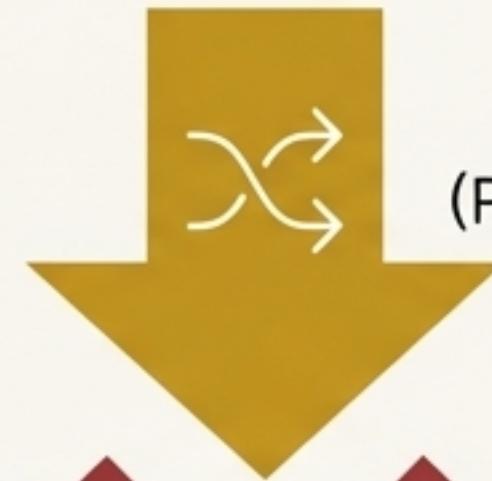
文本水印是验证内容来源的关键技术，但现有方法正面临着一个根本性的挑战：释义攻击。

核心威胁：“释义攻击”如何让传统水印失效

攻击者通过同义词替换、句式变换等方式，在保持文本核心语义不变的同时，完全破坏基于特定词汇选择的水印。

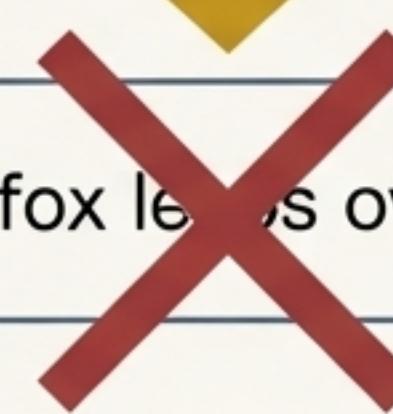
The **quick** brown fox jumps over the **lazy** dog.

原始文本（含水印）



释义攻击
(Paraphrasing Attack)

A fast, brown fox le~~aps~~s over the idle dog.

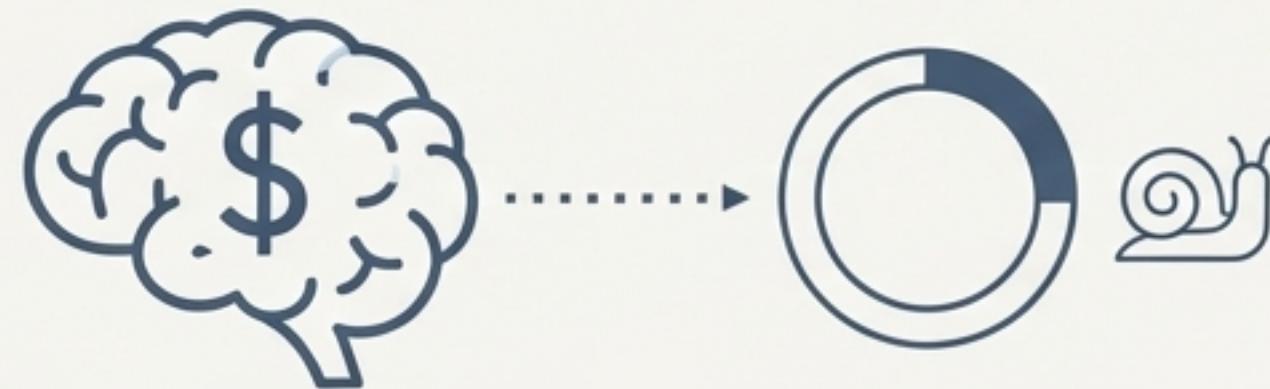


水印失效 (Watermark Lost)

现有语义水印方案的两难：成本与效率的权衡

基于内部隐藏状态

(Based on Internal Hidden States)



核心缺陷

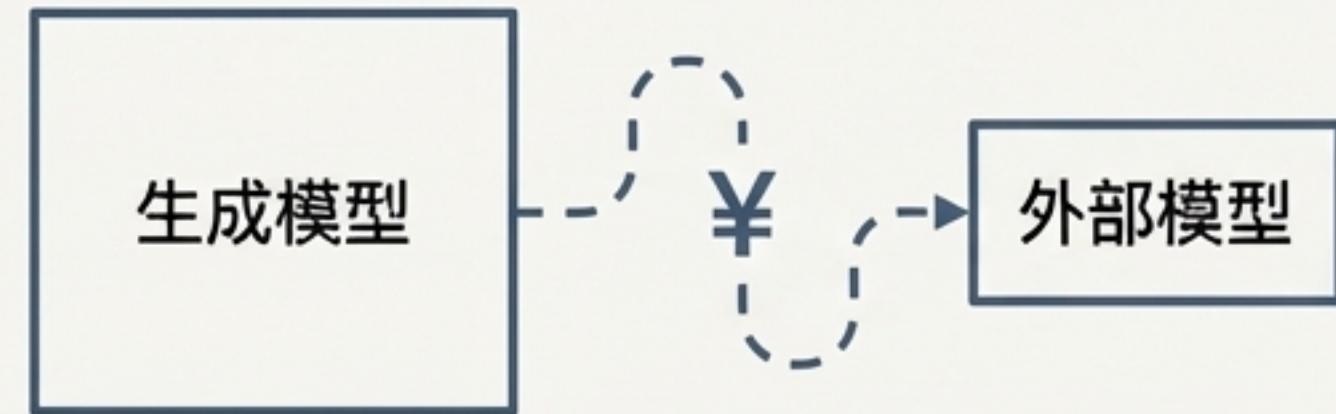
高昂的额外训练成本与低效的检测过程。

说明

必须引入一个需要额外训练的神经网络，且检测过程需要对文本进行逐词元的重复前向传播，计算复杂度高。

基于外部嵌入模型

(Based on External Embedding Models)



核心缺陷

系统效率低下且存在模型不匹配风险。

说明

在文本生成的每一步，都必须额外调用一次庞大的外部模型，导致计算性能开销巨大。

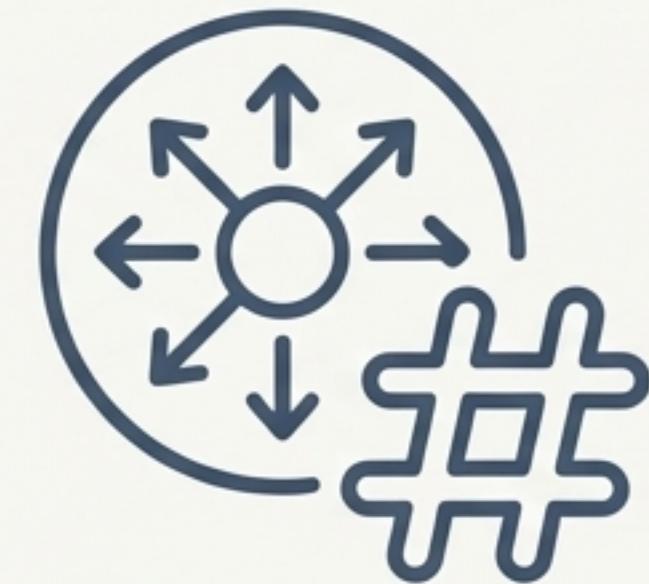
我们的方案：利用模型内生信号，实现零成本的鲁棒水印

本发明将水印与文本的核心语义绑定，其关键在于两个核心组件的巧妙结合：混合专家模型 (MoE) 的路由权重 (Routing Weights) 与局部敏感哈希 (LSH)。



免训练 (Training-Free)

无需训练额外的神经网络。



零成本 (Zero-Cost)

无需任何额外计算开销。



高鲁棒 (High Robustness)

能够有效抵御释义攻击。

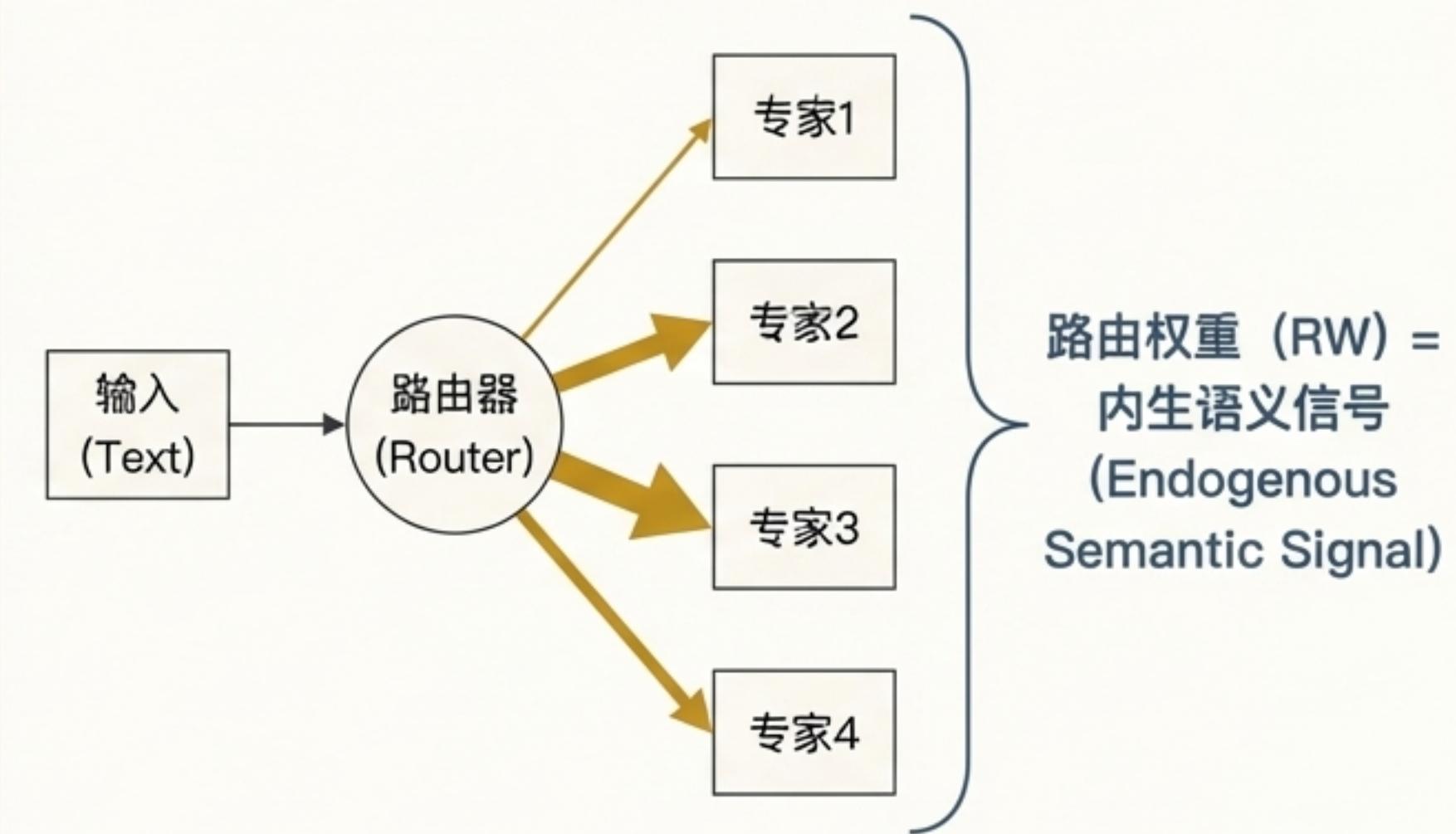
原理一：从MoE路由权重中提取“零成本”的语义信号

What are Routing Weights (RW)?

MoE模型包含多个“专家”网络。对于每个输入，“路由器”会计算一个路由权重 (RW)，决定将任务分配给哪些专家。这个权重分布直接反映了模型对输入内容的语义分类。

The Key Insight

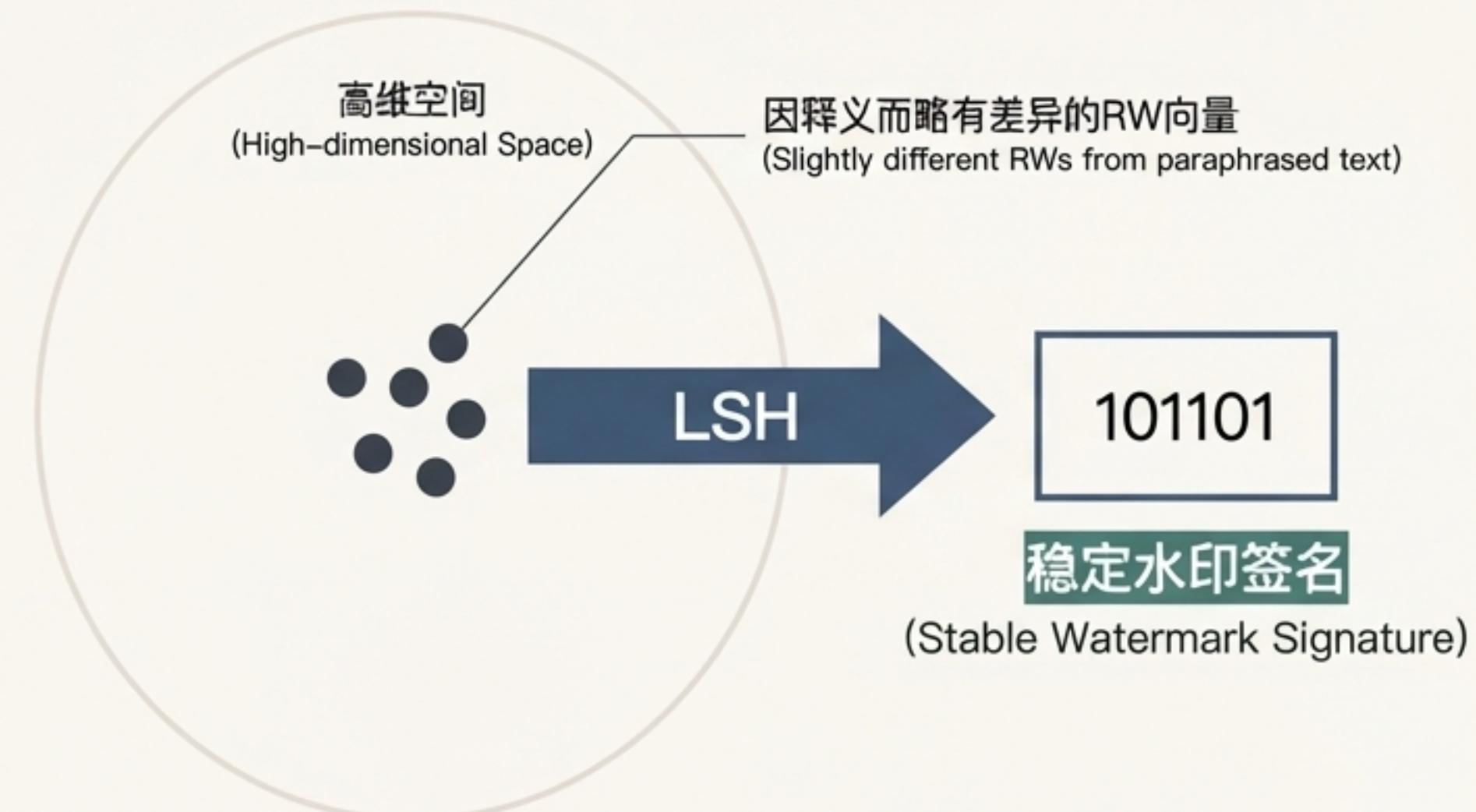
RW是模型标准前向传播的必要副产品。提取它，无需任何额外计算。



原理二：利用局部敏感哈希 (LSH) 构建“稳定”的水印种子

The Challenge

路由权重 (RW) 虽能捕获语义，但它是连续的，在文本被改写时会发生轻微波动。



The Solution – LSH

LSH的设计目标是让相似的、邻近的输入（如因释义而略有变化的RW向量）以高概率产生相同或高度相似的哈希输出（即“碰撞”）。它将不稳定的连续语义空间，映射到稳定的离散水印种子。

一个比喻：GPS坐标 vs. 邮政编码



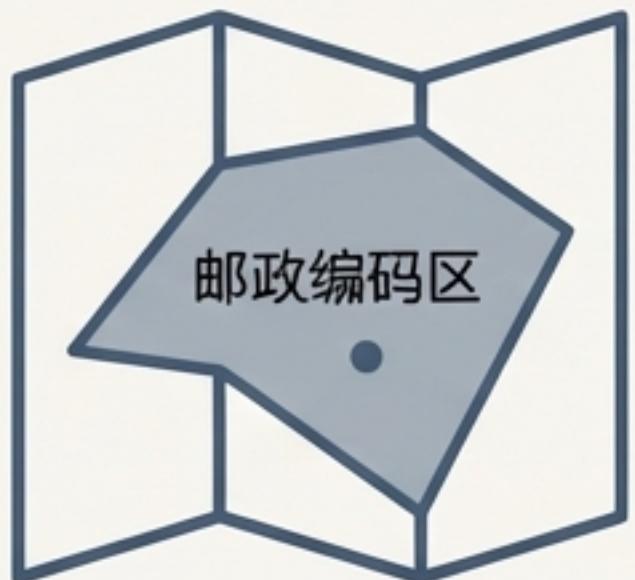
文本 (Text)

你的精确GPS坐标。



释义攻击 (Paraphrasing)

你从客厅走到厨房。你的GPS坐标 (**MoE路由权重**)发生了轻微变化。



LSH

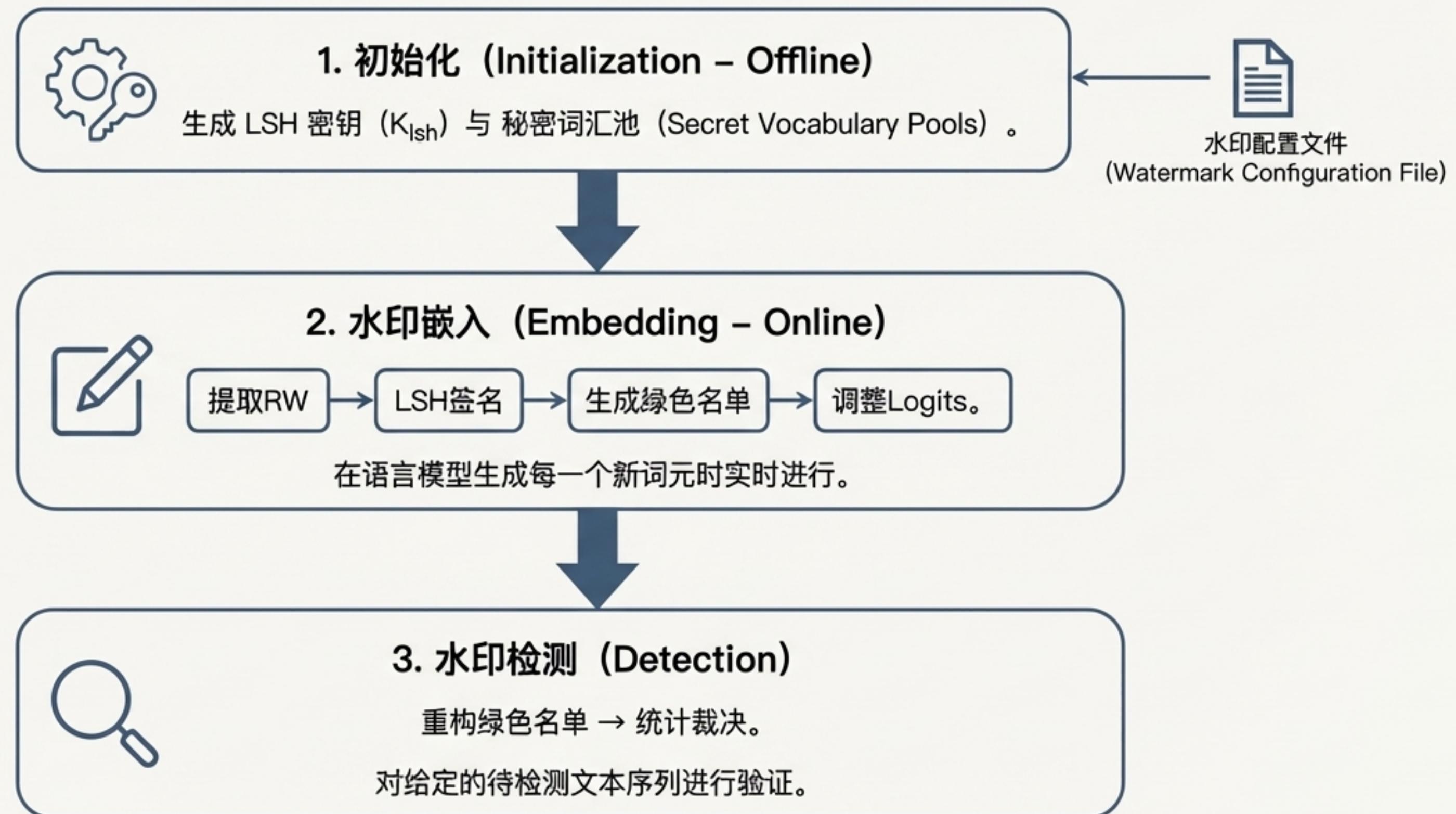
邮政编码系统。它将你不断变化的精确坐标，映射到一个稳定、宽泛的区域 (**LSH签名**)。尽管你移动了，但你仍在该邮编范围内。



水印 (Watermark)

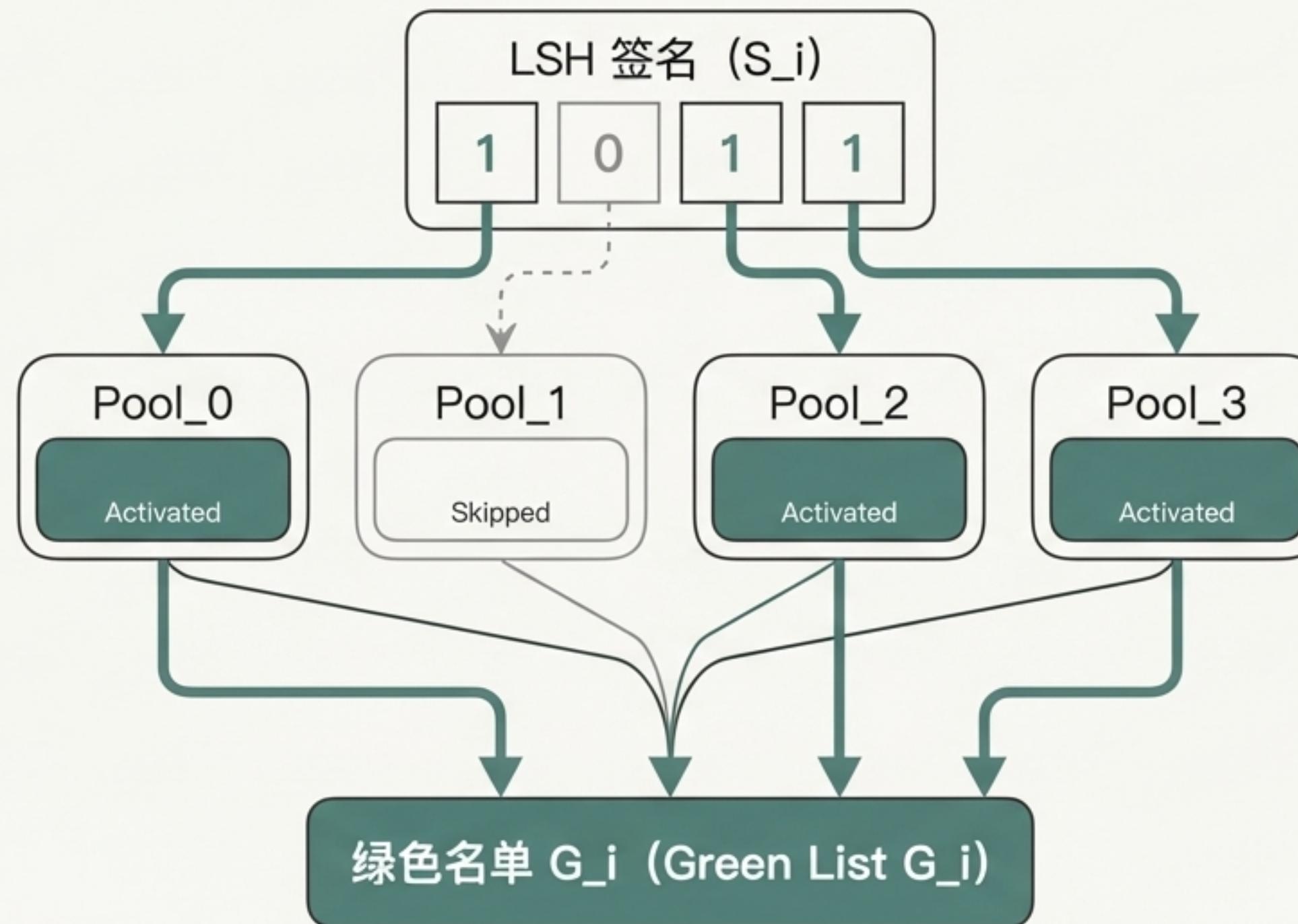
邮局的投递规则。因为邮编 (**LSH签名**) 没变，系统就知道该使用哪个“绿色名单”的投递车队 (词汇表)，水印得以保持。

系统工作全流程概览



核心机制：“按位贡献”如何将LSH签名转化为绿色名单

LSH签名的每一位，都像一把钥匙，控制着一个秘密词汇池的开关。如果某位为1，其对应的词汇池就被激活并加入绿色名单。



水印检测：重构绿色名单并进行统计裁决

1. 逐词元重构 (Token-by-Token Reconstruction)

遍历待检测文本。对于每个词元 t_i ，使用其前面的文本 C_i 作为上下文，通过与嵌入阶段完全相同的流程（提取RW → LSH签名 → 按位贡献）重构出当时应该使用的绿色名单 G_{-i} 。

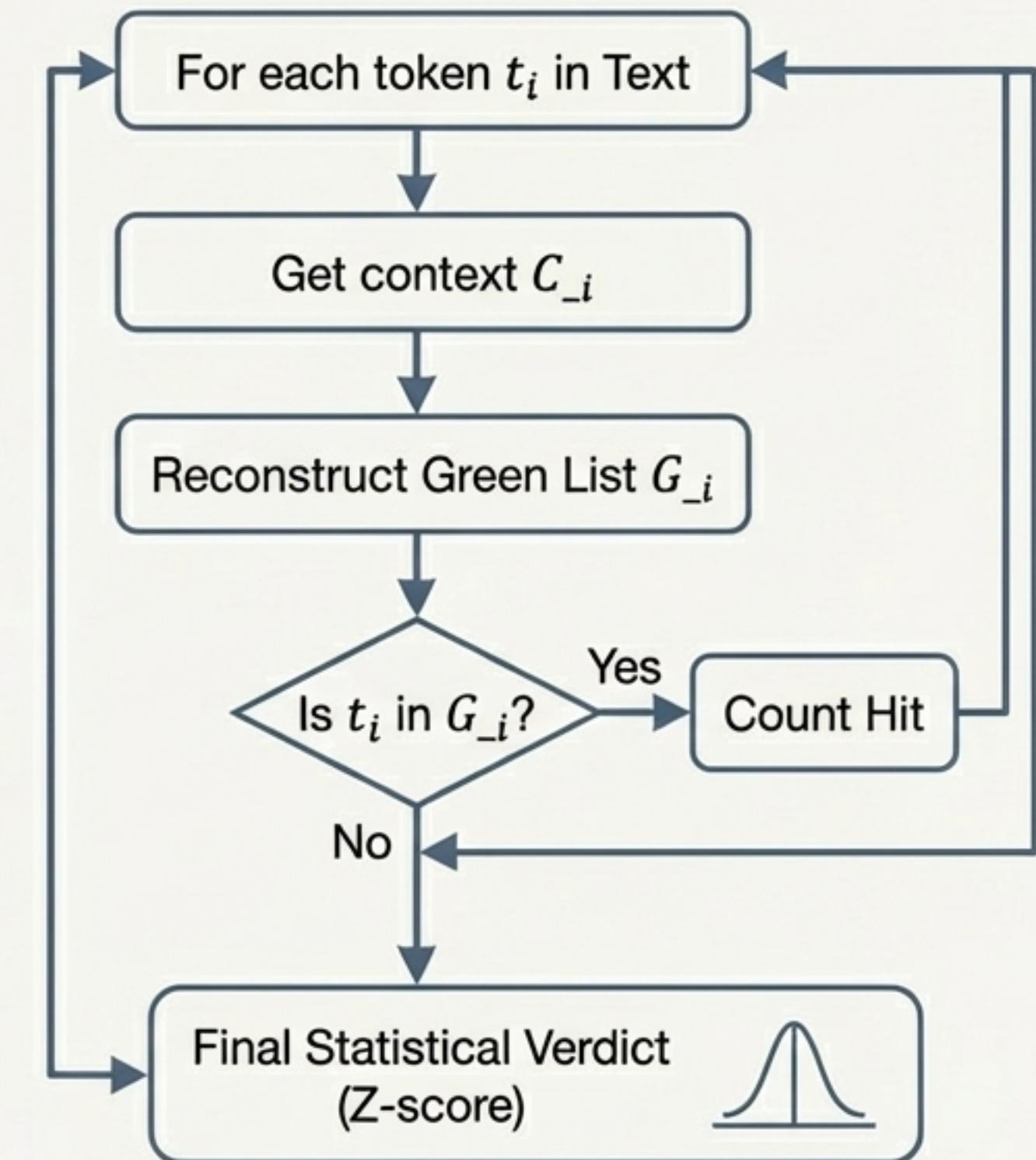
2. 命中验证 (Hit Verification)

判断实际的词元 t_i 是否属于重构出的绿色名单 G_{-i} 。

3. 统计裁决 (Statistical Verdict)

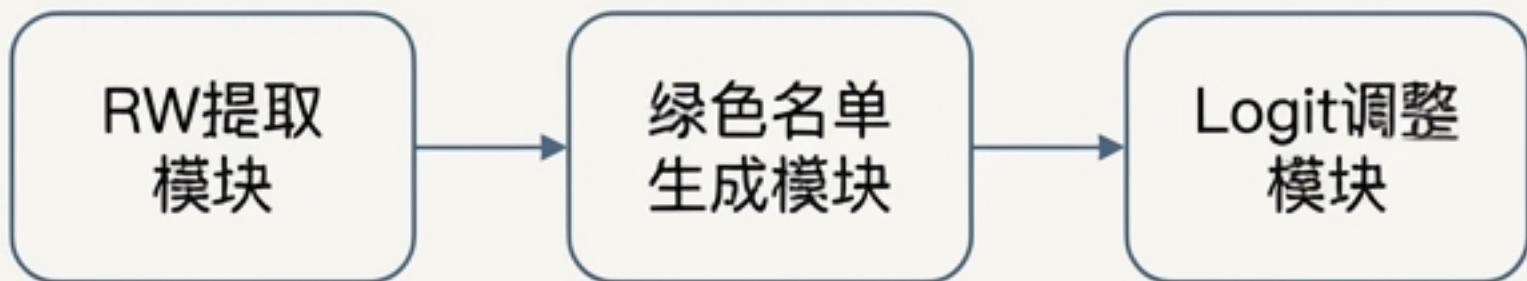
统计总命中数 N_{hits} 。使用假设检验方法（如 Z-检验）计算一个Z-score，衡量实际命中率超出随机期望的显著性。

如果Z-score超过预设阈值，则判定文本包含水印。



系统架构：模块化设计与可配置性

系统构成 (The System)



计算机可读介质 (The Computer-Readable Medium)



水印配置文件

- 一组LSH密钥
- 一组对应的秘密词汇池
- 核心超参数（经身与枕长度、泡泡大小、水印强度）

架构优势一览：各组件对效率与鲁棒性的贡献

组件	功能	对效率的贡献	对鲁棒性的贡献
MoE路由权重	代表输入的语义分类	零成本：模型推理时已计算	语义绑定：释义攻击不显著改变语义分类
LSH	将高维向量映射到离散签名	计算快：仅需简单的点积运算	容错性：相似向量（因释义）产生相同的水印种子
配置文件	存储LSH密钥与词汇池	免训练：使系统无需训练新模型即可工作	安全性：保持语义到词汇表的映射关系保密

重建AIGC时代的信任基石



本方案通过利用模型内生的语义信号，在零计算开销和免除额外训练的前提下，实现了前所未有的鲁棒性。它为实现可追溯、可信赖、负责任的生成式人工智能提供了一条实用且可规模化的技术路径。