MoE 路由水印的信息论形式化定义
Information-Theoretic Formalization of MoE Routing Watermarking

Yunhao Yilong Qingxiao

Abstract

本文提出了混合专家（MoE）架构中路由水印的信息论形式化框架，采用后处理（**Post-processing**）范式，聚焦于模型输入空间的语义层面释义攻击。我们建立了基于推理时扰动模块的信道模型：编码器将水印消息映射为推理时扰动模块 $\mathcal{P}$，该模块在推理阶段对路由器的 logits 输出施加可控扰动，信道是原始干净模型 θ_{clean} 和数据分布 $p(x)$ 对扰动策略 $\mathcal{P}$ 的联合响应，输出是验证者在测试集上观测到的激活模式统计量。我们量化了激活模式（组合码，Combination Code）的编码容量，并揭示了其在语义层面释义攻击下的脆弱性：水印可能对触发器输入过拟合，只学习表层 token 关联而非语义关联。在约束模型性能和隐蔽性的条件下，我们建立了二维率失真（Rate-Distortion）框架，同时考虑性能失真 D_{perf}（由推理时扰动造成的瞬时性能下降）和统计失真 D_{detect}（仅基于激活模式 Σ，确保与验证器特征一致）。我们提出了路由分布扰动型后处理水印机制，通过语义感知的扰动模块设计来对抗输入空间的释义攻击，将语义鲁棒性的实现从训练损失转移到扰动模块的算法设计上。我们定义了语义等价类和鲁棒容量 C_{robust}，并定义了语义鲁棒性 $\mathcal{R}_{\text{input}}$ 来量化水印对输入空间规避攻击的抵抗能力，建立了四维帕累托前沿。在验证过程中，我们将问题形式化为复合假设检验，显式化观测向量为多项分布模型，采用广义似然比检验（GLRT）或序贯概率比检验（SPRT）构建验证器，并基于多项分布族的 Bhattacharyya 距离推导了检测质量的 AUC 下界。最后，我们定义了综合性能指标和扩展的帕累托前沿，为 MoE 路由水印的设计和评估提供了严格的理论基础。

1 引言

混合专家（Mixture-of-Experts, MoE）架构通过稀疏激活机制实现了大型神经网络的高效设计。在 MoE 架构中，路由器（Router）负责为每个输入令牌选择性地激活专家子集，这种路由机制为水印嵌入提供了独特的信息载体。然而，目前缺乏对 MoE 路由水印的理论分析框架，特别是从信息论角度对其容量和检测能力的严格量化。

核心挑战：在实际应用中，水印面临的主要威胁来自模型输入空间的攻击，特别是语义层面的释义攻击。攻击者通过保持语义不变但改变输入表达的方式来规避水印检测，这要求水印系统具备对输入空间变化的鲁棒性。

核心转变：本文采用后处理（**Post-processing**）范式，水印嵌入不通过修改模型参数 $\Delta\theta$ 或约束微调实现，而是在推理阶段通过后处理模块 $\mathcal{P}$ 对路由器的输出施加可控扰动。这种转变具有以下优势：（1）非侵入性：不需要修改模型参数，保持原始模型完整性；（2）部署灵活性：可以动态启用或禁用水印；（3）动态撤销：可以随时移除水印而不影响模型。

本文提出了 MoE 路由水印的信息论形式化定义，采用后处理框架，聚焦于模型输入空间的语义层面释义攻击。我们建立了基于推理时扰动模块的信道模型，将编码器定义为推理时扰动模块 $\mathcal{P}$，信道是原始干净模型和数据分布对扰动策略的联合响应。我们量化了激活模式（组合码，Combination Code）的编码容量，并深入分析了其在输入空间攻击下的脆弱性。在此基础上，我们建立了二维率失真框架，同时考虑性能失真（由推理时扰动造成的瞬时性能下降）和统计失真（隐蔽性），并提出了路由分布扰动型后处理水印机制，通过语义感知的扰动模块设计来对抗输入空间的释义攻击。本文的贡献包括：

- 建立了基于推理时扰动模块的信道模型，明确了编码器（扰动模块 $\mathcal{P}$）、信道和输出的定义，采用后处理范式实现非侵入性水印嵌入
- 量化了激活模式（组合码，Combination Code）的编码容量，这是最稳定的信息维度
- 建立了二维率失真（Rate-Distortion）框架，同时考虑性能失真 D_{perf}（由推理时扰动造成）和统计失真 D_{detect}（隐蔽性）
- 提出了路由分布扰动型后处理水印机制，通过语义感知的扰动模块设计来对抗输入空间的释义攻击，将语义鲁棒性的实现从训练损失转移到扰动模块的算法设计上
- 揭示了 I_{pattern} 在输入空间攻击下的脆弱性：水印可能对触发器输入过拟合，只学习表层 token 关联而非语义关联，导致在释义攻击下实际可实现鲁棒容量可能骤降为 0

- 定义了语义等价类和鲁棒容量 C_{robust}，并定义了语义鲁棒性 $\mathcal{R}_{\text{input}}$ 来量化水印对输入空间规避攻击的抵抗能力，建立了四维帕累托前沿

- 构建了基于复合假设检验的验证框架（GLRT/SPRT），强调只使用激活模式特征进行验证，显式化观测向量为多项分布模型

- 深入分析了语义层面释义攻击对水印检测的影响，揭示了过拟合问题并提出了语义感知扰动模块的应对策略

2 符号表

为便于阅读，本节汇总本文使用的主要符号：

符号	含义
n	专家总数
k	top-k 激活数
T	温度参数（控制路由分布锐度）
σ	归一化噪声参数
$\mathbf{r} \in \Delta^{n-1}$	路由分布向量（概率单纯形）
$\Sigma \subset [n]$	激活的专家集合，$\lvert\Sigma\rvert = k$
$\pi \in S_k$	激活顺序（排列）
$\mathcal{S}(x) = (\mathbf{r}, \Sigma, \pi)$	路由状态
$\theta_{\text{clean}}, \theta_{\text{wm}}$	干净模型和水印模型的参数
$\Delta\theta$	参数扰动（原始框架，已弃用）
$\mathcal{P}$	推理时扰动模块（后处理框架）
δ_l	logits 扰动向量
m	水印消息
Σ_m	目标激活模式（由消息 m 映射）
I_{pattern}	激活模式（组合码，Combination Code）容量
$C_{\text{achievable}}$	可实现容量
C_{robust}	鲁棒容量（考虑语义等价类）
D_{perf}	性能失真
D_{detect}	统计失真（隐蔽性，基于 Σ）
ϵ_1, ϵ_2	性能失真和统计失真的阈值
$\lambda_{\text{wm}}, \lambda_{\text{con}}$	水印损失和一致性约束的权重
$\mathcal{R}_{\text{param}}$	参数鲁棒性
$\mathcal{R}_{\text{input}}$	语义鲁棒性（输入空间鲁棒性）
$\mathbf{c} \in \mathbb{N}^{\binom{n}{k}}$	激活模式计数向量
$\mathbf{p} = [p_1, \ldots, p_{\binom{n}{k}}]$	激活模式概率分布向量
D_B	Bhattacharyya 距离
α, β	假阳性率和检测功效
N	测试样本数

Table 1: 主要符号表

3 基础符号与系统定义

3.1 MoE 路由机制基础

设 MoE 模型的第 l 层为：

$$\text{MoE}_l(x) = \sum_{i=1}^{n} g_i(x) \cdot E_i(x)$$

其中：

- $x \in \mathbb{R}^d$：输入向量
- $g_i(x)$：第 i 个专家的路由权重（由路由器 Router 产生）
- $E_i(x)$：第 i 个专家的输出
- n：专家总数

路由器输出：

$$\mathbf{r} = \text{Router}(x) = \text{Softmax}(\mathbf{w}^T x + b) \in \Delta^{n-1}$$

其中 Softmax$(\cdot)$ 为 softmax 函数，得到路由分布 $\mathbf{r} = [r_1, r_2, \ldots, r_n]$，位于概率单纯形 $\Delta^{n-1} = \{\mathbf{r} \in \mathbb{R}^n : r_i \geq 0, \sum_{i=1}^n r_i = 1\}$。温度参数 T 用于控制分布的锐度：$\text{Softmax}_T(\mathbf{z})_i = \exp(z_i/T)/\sum_j \exp(z_j/T)$。

3.2 水印系统定义

水印系统 $\mathcal{W} = (\mathcal{E}, \mathcal{V})$ 包括：

- $\mathcal{E}$：嵌入过程（Embedding），在后处理框架下，编码器 $\mathcal{E}$ 是一个推理时扰动模块 $\mathcal{P}$（**Inference-Time Perturbation Module**）。它接收水印消息 m 和原始路由器的输出 $\mathbf{l}$（logits），并实时计算出一个扰动 δ_l，使得扰动后的 logits $\mathbf{l}' = \mathbf{l} + \delta_l$ 对应的激活模式 $\Sigma(x)$ 偏向目标模式 Σ_m
- $\mathcal{V}$：验证过程（Verification），从测试集上的观测特征推断模型是否包含水印

核心转变：水印嵌入不通过修改模型参数 $\Delta\theta$ 或约束微调实现，而是在推理阶段通过后处理模块 $\mathcal{P}$ 对路由器的输出施加可控扰动。

信道模型（后处理框架）：

- 路由器是一个习得的函数：$\mathbf{r} = \text{Router}(x) = \text{Softmax}(\mathbf{w}^T x + b)$，其输出依赖于输入 x
- 编码器的任务：通过推理时扰动模块 $\mathcal{P}$ 对特定输入 x 的路由 logits $\mathbf{l}$ 施加扰动 δ_l，使得激活模式 $\Sigma(x)$ 强制变为目标模式 Σ_m，从而在数据分布 $p(x)$ 上使 $\mathcal{S}(x)$ 的统计分布 $p(\mathcal{S}|\theta_{\text{clean}}, \mathcal{P})$ 发生可检测的、偏向消息 m 的偏移
- 信道定义：

- 输入：扰动策略 $\mathcal{P}$（而非参数扰动 $\Delta\theta$）
- 信道：原始（干净）模型 θ_{clean} 和数据分布 $p(x)$ 对该扰动策略 $\mathcal{P}$ 的联合响应。对于每个输入 x，$\mathcal{P}$ 根据 x 的特征决定是否激活并计算扰动 δ_l，从而影响 $\mathcal{S}(x)$
- 输出：验证者在测试集 $\mathcal{T} = \{x_1, \ldots, x_N\}$ 上观测到的特征统计量 $\mathbf{f} = [\mathcal{F}(x_1), \ldots, \mathcal{F}(x_N)]$，其中 $\mathcal{F}(x_i) = \Sigma(x_i)$ 为激活模式

4 水印信息的编码空间

4.1 路由状态空间

对于输入 x，在第 l 层的路由状态表示为：

$$\mathcal{S}(x) = (\mathbf{r}, \Sigma, \pi) \in \Delta^{n-1} \times 2^{[n]} \times S_k$$

其中：

- $\mathbf{r} \in \Delta^{n-1}$：路由分布向量（位于概率单纯形）
- $\Sigma \subset [n]$：激活的专家集合，满足 $|\Sigma| = k$（top-k 稀疏门控）
- $\pi \in S_k$：激活顺序（排列），表示 Σ 中专家按路由权重降序排列

为避免记号歧义，我们明确区分：Softmax$(\cdot)$ 表示函数，Σ 表示激活集合，π 表示排列。

4.2 可用信息维度

路由状态空间提供了三个主要的信息维度用于编码水印：

维度 **1**：激活模式（组合码，**Combination Code**）
从 n 个专家中选择 k 个激活，信息容量为：

$$I_{\text{pattern}} = \log_2 \binom{n}{k} = \log_2 \frac{n!}{k!(n-k)!}$$

对于 $n = 128, k = 2$ 的典型配置，使用斯特林近似（Stirling's approximation）：

$$I_{\text{pattern}} \approx 13 \text{ bits}$$

维度 **2**：排列顺序（排列码）
对前 k 个专家的排列，信息容量为：

$$I_{\text{order}} = \log_2(k!)$$

对于 $k = 4$：

$$I_{\text{order}} = \log_2(24) \approx 4.6 \text{ bits}$$

注意：该容量仅在" 前 k 专家可完全排序且可控" 的假设下成立。若路由器含温度缩放或分数权重相近导致排序不稳定，应引入有限分辨率/随机噪声的" 有效容量" 修正。

维度 **3**：权重量化（连续码）
由于 $\mathbf{r} \in \Delta^{n-1}$ 且满足 $\sum_i r_i = 1$，自由度为 $(n-1)$。在 top-k 稀疏门控下，仅对激活的 k 个专家权重进行量化，且满足 $\sum_{i \in \Sigma} r_i = 1$，因此自由度为 $(k-1)$。假设每个路由权重 r_i（$i \in \Sigma$）量化为 b-bit 精度：

$$I_{\text{weight}} \leq (k-1) \cdot b \text{ bits}$$

对于 $k = 4, b = 4$（16 级量化）：

$$I_{\text{weight}} \leq 12 \text{ bits}$$

考虑温度 T 与归一化噪声 σ 的影响，有效容量需乘以系数 $\eta(T, \sigma) \in (0, 1)$：

$$I_{\text{weight}}^{\text{eff}} = \eta(T, \sigma) \cdot (k-1) \cdot b$$

总容量上界：

$$\begin{aligned} C_{\max} &= I_{\text{pattern}} + I_{\text{order}} + I_{\text{weight}}^{\text{eff}} \\ &= \log_2 \binom{n}{k} + \log_2(k!) + \eta(T, \sigma) \cdot (k-1) \cdot b \end{aligned} \tag{1}$$

5 信息容量的量化

5.1 可实现容量（**Achievable Capacity**）与编码本视角

从编码本（Codebook）视角，在后处理框架下，编码器 $\mathcal{E}$（即扰动模块 $\mathcal{P}$）将水印消息 m 映射为推理时的扰动策略。信道模型为" 扰动策略 $\mathcal{P} \to$ 原始模型 θ_{clean} 和数据分布 $p(x)$ 的联合响应 $\to$ 路由状态的后验分布 $p(\mathcal{S}(x)|\theta_{\text{clean}}, \mathcal{P}) \to$ 验证统计量 $\mathbf{f}$"。

在约束模型性能和隐蔽性的条件下，可实现的容量定义为：

$$C_{\text{achievable}} = \max_{p(m), \mathcal{P}} I(m; \mathbf{f})$$

其中 $\mathbf{f} = [\mathcal{F}(x_1), \ldots, \mathcal{F}(x_N)]$ 为测试集 $\mathcal{T}$ 上的特征统计量（激活模式计数向量 $\mathbf{c}$），$I(\cdot;\cdot)$ 为互信息。容量 $C_{\text{achievable}}$ 由扰动模块 $\mathcal{P}$ 在 D_{perf} 和 D_{detect} 约束下实际能稳定注入的互信息 $I(m; \mathbf{f})$ 决定。

特征选择：排列码 I_{order} 和权重量化 I_{weight} 在语义变化下不稳定。因此，验证者 $\mathcal{V}$ 应该只依赖于激活模式 Σ。信道输出 $\mathbf{f}$ 可以被精简为：在 N 个样本上观测到的激活模式 Σ 的经验分布（或计数向量）。

5.2 率失真（**Rate-Distortion**）问题

在数据分布 $p(x)$ 上定义失真。率失真函数定义为：

$$R(D_{\text{perf}}, D_{\text{detect}}) = \max_{p(m),\mathcal{P}} I(m; \mathbf{f}) \quad \text{s.t.} \quad D_{\text{perf}} \le \epsilon_1, \quad D_{\text{detect}} \le \epsilon_2 \tag{2}$$

其中失真是一个二维向量 $D = (D_{\text{perf}}, D_{\text{detect}})$，在后处理框架下各项指标的来源如下：

1. 性能失真 $D_{\textbf{perf}}$：由推理时扰动 $\mathcal{P}$ 造成的模型主任务性能下降

   $$D_{\text{perf}} = \Delta_{\text{perf}} = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{clean}+\mathcal{P}}$$

   其中 $\text{Acc}_{\text{clean}+\mathcal{P}}$ 表示在干净模型上应用扰动模块 $\mathcal{P}$ 后的准确率。性能失真来源于扰动 $\mathcal{P}$ 对任务损失 $\mathcal{L}_{\text{task}}$ 的瞬时干扰。

2. 统计失真 $D_{\textbf{detect}}$（可隐蔽性）：扰动 $\mathcal{P}$ 引入的统计痕迹，为确保与验证器使用的观测特征一致，我们仅基于激活模式 Σ 的分布进行度量：

   $$D_{\text{detect}} = \mathbb{E}_{x\sim p(x)} \left[D_{\text{KL}} \left(p(\Sigma(x)|\theta_{\text{clean}}) || p(\Sigma(x)|\theta_{\text{clean}}, \mathcal{P}) \right) \right]$$

   其中 $p(\Sigma|\theta_{\text{clean}}, \mathcal{P})$ 表示在干净模型 θ_{clean} 上应用扰动模块 $\mathcal{P}$ 后激活模式 Σ 的条件分布。该定义保证了训练中的隐蔽性度量与验证器使用的观测分布（仅激活模式计数）的一致性。统计失真来源于扰动 $\mathcal{P}$ 引起的 $p(\Sigma)$ 分布偏移。

失真的双重作用：

- D_{detect} 越小，水印越隐蔽（越难被攻击者发现）
- D_{detect} 越大，水印越可检测（越容易被验证者验证）

信息-性能-隐蔽性权衡：通过扫描超参数（如 λ_{wm}），可以凭经验绘制出 $R(D_{\text{perf}}, D_{\text{detect}})$ 的帕累托前沿，其中：

- X 轴：D_{perf}（模型准确率下降）
- Y 轴：R（水印检测器的统计显著性或 AUC）
- Z 轴：D_{detect}（隐蔽性，越小越好）

6 路由后处理水印机制（编码器 $\mathcal{E}$ 的实现）

6.1 问题的形式化定义

在后处理框架下，编码器 $\mathcal{E}$ 的任务是将水印消息 m 映射为推理时扰动模块 $\mathcal{P}$。扰动模块 $\mathcal{P}$ 在推理时对路由器的 logits 输出施加扰动，使得激活模式 $\Sigma(x)$ 偏向目标模式 Σ_m。

对于输入 x，路由器产生原始 logits $\mathbf{l} = [l_1, \ldots, l_n]$。扰动模块 $\mathcal{P}$ 计算扰动 δ_l，使得扰动后的 logits $\mathbf{l}' = \mathbf{l} + \delta_l$ 满足：

$$\min_{i\in\Sigma_m} l'_i > \max_{i\notin\Sigma_m} l'_i + \text{margin}$$

其中 $m \to \Sigma_m$：通过带密钥 K 的哈希函数 $H_K(m)$ 将消息 m 映射到目标激活模式 Σ_m。

6.2 路由分布扰动型实现（方案 **A**）

核心思想：在推理时，对特定输入 x，使其激活模式 $\Sigma(x)$ 强制变为目标模式 Σ_m。这是最稳定的信息维度（激活模式，组合码 I_{pattern}）的实现方案。

算法：路由分布扰动型后处理水印

1. 目标定义：选择消息 m，通过哈希函数 $H_K(m)$ 生成目标激活模式 Σ_m。
2. 扰动计算：对于输入 x，获取原始路由器的 logits 输出 $\mathbf{l} = [l_1, \ldots, l_n]$。扰动模块 $\mathcal{P}$ 计算最小扰动 δ_l，使得：
   - 找到 Σ_m 中的最低分：$l_{\text{min\_target}} = \min_{i\in\Sigma_m} l_i$
   - 找到非 Σ_m 中的最高分：$l_{\text{max\_other}} = \max_{i\notin\Sigma_m} l_i$
   - 计算扰动：δ_l 使得 $l'_{\text{min\_target}} > l'_{\text{max\_other}} + \text{margin}$，其中 $l'_i = l_i + \delta_{l,i}$
   - 最小化扰动幅度：$\min \|\delta_l\|$，以减小对任务性能的影响
3. 稀疏性适配：扰动 $\mathcal{P}$ 必须在 top-k 约束下进行。目标是将 Σ_m 中的 k 个专家提升到 top-k，并将原始 top-k 中不在 Σ_m 的专家” 踢出” top-k。这天然适配稀疏路由机制，只操纵 top-k 的选择，而不改变 k 本身。
4. 触发器设计：扰动模块 $\mathcal{P}$ 需要决定何时激活。可能的策略包括：
   - 全量激活：对所有输入 x 都施加扰动
   - 条件激活：基于输入特征（如嵌入相似度、句法结构等）决定是否激活
   - 采样激活：以一定概率 p_{trigger} 激活，平衡隐蔽性和可检测性
5. 高效扰动计算：为降低计算开销，可以采用以下优化：
   - 闭式解：当 margin 较小时，可以给出 δ_l 的闭式表达式
   - 近似算法：使用贪心或迭代方法快速计算近似最优扰动
   - 缓存机制：对于相似输入，复用已计算的扰动

6.3 其他实现方案

方案 **B**：专家输出后处理型：在专家输出 $E_i(x)$ 上注入签名。这偏离了依赖 I_{pattern} 的核心假设，利用了权重量化 I_{weight} 维度的思想。挑战在于连续值对噪声和语义变化的稳定性较差，检测器需要更复杂的设计。

方案 **C**：最终输出后处理型：在最终 $MoE_l(x)$ 上注入签名。同样面临连续值稳定性的挑战。

推荐方案：方案 A（路由分布扰动型）是最稳定的实现，因为它直接操作激活模式 Σ，这是最稳定的信息维度。

6.4 与理论框架的连接

该算法完美地连接回了形式化框架：

- 编码器 $\mathcal{E}$：从离线的、基于优化的微调过程转变为在线的、基于算法的推理时扰动模块 $\mathcal{P}$
- 可实现容量 $C_{\textbf{achievable}}$：容量 R 不再是理论上的 $\log_2 \binom{n}{k}$，而是由扰动模块 $\mathcal{P}$ 在 D_{perf} 和 D_{detect} 约束下实际能稳定注入的互信息 $I(m;\mathbf{f})$
- 率失真 $R(D)$：可以通过调整扰动策略（如 margin、触发概率等）来凭经验绘制出 $R(D_{\text{perf}}, D_{\text{detect}})$ 曲线，完美体现了" 信息-性能-隐蔽性权衡" 的思想

7 语义层面释义攻击的影响

7.1 释义攻击的定义与机制

模型输入空间的攻击：在实际应用中，水印面临的主要威胁来自模型输入空间的攻击，特别是语义层面的释义攻击（Paraphrasing Attack）。这是一个更高级、更隐蔽的威胁：攻击者甚至不需要修改模型参数，只是在使用模型时，通过改变输入的措辞来规避水印的检测。

攻击者目标：规避（Evasion）。

攻击者假设：攻击者拥有一个黑盒或灰盒的水印模型 f_{wm}。他怀疑模型在处理某些特定输入时会表现出异常（例如，激活特定的专家组合 Σ_m）。

攻击者行为：攻击者不直接使用输入 x（例如一个提示：" 请总结一下莎士比亚的《哈姆雷特》"），而是将其释义为 x'（例如：" 用几句话概括《哈姆雷特》剧情"）。

攻击成功的条件：输入 x 能够触发水印统计偏差（$p(\Sigma_m|x) \gg p(\Sigma_m|x, \theta_{\text{clean}})$），而释义后的输入 x' 却不能触发（$p(\Sigma_m|x') \approx p(\Sigma_m|x', \theta_{\text{clean}})$）。

设输入为 x，其语义等价的释义为 x'，满足：

$$\text{Semantic}(x) \approx \text{Semantic}(x')$$

其中 Semantic$(\cdot)$ 表示输入的语义内容。

7.2 释义对路由状态的影响

由于路由器 Router(x) 是一个习得的函数，语义等价的输入 x 和 x' 可能产生不同的路由状态：

$$\mathcal{S}(x) = (\mathbf{r}(x), \Sigma(x), \pi(x)) \neq \mathcal{S}(x') = (\mathbf{r}(x'), \Sigma(x'), \pi(x'))$$

这种差异可能导致：

- 激活模式变化：$\Sigma(x) \neq \Sigma(x')$，即使语义相同，激活的专家集合可能不同
- 路由分布偏移：$\mathbf{r}(x) \neq \mathbf{r}(x')$，路由权重分布发生变化

7.3 水印检测的挑战

在语义层面释义攻击下，水印检测面临以下挑战：

1. 语义不变性要求：水印应该对语义等价的输入保持稳定，即对于语义等价的 x 和 x'，水印检测结果应该一致。
2. 激活模式的语义关联：激活模式 $\Sigma(x)$ 应该与输入的语义内容相关，而非仅仅依赖于表面的词汇或表达方式。这要求水印嵌入过程能够学习到语义层面的特征。
3. 统计分布的稳定性：在测试集上，即使输入经过释义，激活模式的统计分布 $p(\Sigma|\theta_{\text{wm}})$ 应该保持相对稳定，使得验证者能够检测到水印信号。

7.4 语义等价类与鲁棒容量

为了形式化语义层面的鲁棒性，我们定义语义等价类（**Semantic Equivalence Classes**）。设输入空间上的等价关系 $x \sim x'$ 表示 x 和 x' 语义一致，即 Semantic$(x) \approx$ Semantic(x')。令 $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_M\}$ 为语义等价类的集合，其中每个类 $\mathcal{C}_j$ 包含语义等价的输入。

鲁棒容量（**Robust Capacity**）定义为：对于每个语义等价类 $\mathcal{C}_j$ 和目标激活模式 Σ_m，若满足：

$$\Pr_{x \in \mathcal{C}_j} [\Sigma(x) = \Sigma_m] \geq 1 - \epsilon$$

其中 ϵ 为允许的误差阈值，则模式 Σ_m 在该类上鲁棒。有效码本大小由满足类内一致性的模式数决定，即：

$$C_{\text{robust}} = \log_2 |\{\Sigma_m : \forall \mathcal{C}_j, \Pr_{x \in \mathcal{C}_j} [\Sigma(x) = \Sigma_m] \geq 1 - \epsilon\}|$$

该定义将组合码容量 I_{pattern} 经由" 类内不变性约束" 缩减为有效码本大小。在释义攻击下，若水印仅对特定触发词过拟合，则 C_{robust} 可能骤降为 0。

常重码（**Constant-Weight Codes**）与纠错思想：在 $\binom{n}{k}$ 的组合空间中选取码本 $\mathcal{M}$ 使得模式间汉明距离大于 $d_{\min}$。结合 top-k 抖动/温度扰动，将解码

半径映射为" 允许若干位（专家选择）翻转"。这给出纠错余量与鲁棒容量的显式下界。对于码本 $\mathcal{M}$，若任意两个码字 $\Sigma_i, \Sigma_j \in \mathcal{M}$ 满足 $d_H(\Sigma_i, \Sigma_j) \geq d_{\min}$，则码本可纠正最多 $\lfloor (d_{\min} - 1)/2 \rfloor$ 位的错误。

7.5 应对策略：语义感知的扰动模块设计

在后处理框架下，我们无法使用训练时的语义一致性约束 $\mathcal{L}_{\text{con}}$，因为我们不进行模型训练。语义鲁棒性 $\mathcal{R}_{\text{input}}$ 的实现完全转移到了扰动模块 $\mathcal{P}$ 的设计上。

核心要求：扰动模块 $\mathcal{P}$ 必须具备" 语义感知" 能力，即对于语义等价类 $\mathcal{C}_j$ 中的不同输入 x 和 x'，$\mathcal{P}$ 必须能识别它们的语义相似性，并施加一致的扰动，使它们都触发目标模式 Σ_m。

实现策略：

1. 编码端（扰动模块 $\mathcal{P}$）：
   - 语义等价类识别：$\mathcal{P}$ 需要能够识别语义等价类 $\mathcal{C}_j$。可以通过以下方式实现：
     - 嵌入相似度：计算输入 x 的嵌入向量，与预定义的语义簇中心进行比较
     - 句法结构分析：使用句法解析器识别语义等价的句法结构
     - 启发式规则：基于关键词、实体等设计规则识别语义等价类
   - 一致扰动：对于同一语义等价类 $\mathcal{C}_j$ 中的输入，$\mathcal{P}$ 施加相同的扰动策略，确保它们都触发目标模式 Σ_m
2. 验证端（验证器 $\mathcal{V}$）：
   - 联合检测策略：验证器 $\mathcal{V}$ 在测试时，主动使用释义 x' 来测试。一个鲁棒的水印必须在 x 和 x' 上都触发 Σ_m，从而实现鲁棒容量 C_{robust}
   - 语义等价类测试：对于测试样本 x，生成其语义等价释义 x'，验证两者是否都触发目标模式 Σ_m

通过这种方式，我们将实现 $\mathcal{R}_{\textbf{input}}$ 的负担从训练损失 $\mathcal{L}_{\textbf{con}}$ 转移到了扰动模块 $\mathcal{P}$ 的算法设计上。这使得水印系统更加灵活，但要求 $\mathcal{P}$ 具备更强的语义理解能力。

权衡关系（后处理框架）：

1. $\mathcal{R}_{\textbf{input}}$ **vs** $\Delta_{\textbf{perf}}$（语义鲁棒性 **vs** 性能）：
   - 为了提高 $\mathcal{R}_{\text{input}}$，扰动模块 $\mathcal{P}$ 必须在更大的语义等价类上施加一致的扰动。
   - 这要求 $\mathcal{P}$ 对更多输入施加扰动，从而不可避免地导致更大的性能下降 $\Delta_{\textbf{perf}}$。
   - 性能失真来源于扰动 $\mathcal{P}$ 对任务损失 $\mathcal{L}_{\text{task}}$ 的瞬时干扰。
2. $\mathcal{R}_{\textbf{input}}$ **vs** $C_{\textbf{achievable}}$（语义鲁棒性 **vs** 容量）：
   - 扰动模块 $\mathcal{P}$ 的语义感知能力越强，需要在更大的语义簇上嵌入水印，而非仅针对特定触发词。
   - 这使得嵌入单个比特（例如 Σ_m）的难度急剧增加，因此可实现容量 $C_{\textbf{achievable}}$ 会显著下降。
   - 鲁棒容量 C_{robust} 由满足类内一致性的模式数决定，通常小于 $C_{\text{achievable}}$。

帕累托前沿：编码器必须在四维空间（$C_{\text{achievable}}$, D_{perf}, D_{detect}, $\mathcal{R}_{\text{param}}$, $\mathcal{R}_{\text{input}}$）中寻找权衡。

8 验证过程与检测能力

8.1 假设检验框架（复合假设）

验证问题可形式化为复合假设检验：

- H_0：模型未被水印化（原始模型），参数 $\theta_0 \in \Theta_0$ 未知
- H_1：模型被正确的水印化，参数 $\theta_1 \in \Theta_1$ 未知

由于参数未知，H_0 和 H_1 均为复合假设。给定测试集 $\mathcal{T} = \{x_1, \ldots, x_N\}$，我们只提取激活模式 Σ 作为观测特征（排列码和权重量化在语义变化下不稳定）。

观测向量与概率模型显式化：对所有 $\binom{n}{k}$ 种可能的激活模式进行计数，得到计数向量 $\mathbf{c} \in \mathbb{N}^{\binom{n}{k}}$，满足 $\sum_{i=1}^{\binom{n}{k}} c_i = N$。我们假设观测向量 $\mathbf{c}$ 服从多项分布（**Multinomial**）：

$$\mathbf{c} \sim \text{Multinomial}(N, \mathbf{p})$$

其中 $\mathbf{p} = [p_1, p_2, \ldots, p_{\binom{n}{k}}]$ 为激活模式的概率分布向量，满足 $\sum_{i=1}^{\binom{n}{k}} p_i = 1$。

在 H_0 下，$\mathbf{p} = \mathbf{p}_0$ 表示干净模型 θ_{clean} 的路由分布；在 H_1 下，$\mathbf{p} = \mathbf{p}_1$ 表示在干净模型 θ_{clean} 上应用扰动模块 $\mathcal{P}$（后处理）后的激活模式分布。多项分布的似然函数为：

$$p(\mathbf{c}|\mathbf{p}) = \frac{N!}{\prod_{i=1}^{\binom{n}{k}} c_i!} \prod_{i=1}^{\binom{n}{k}} p_i^{c_i}$$

8.2 广义似然比检验（**GLRT**）与序贯概率比检验（**SPRT**）

由于 H_0 和 H_1 为复合假设，采用广义似然比检验（**GLRT**）以消去未知参数。对于多项分布模型，GLRT 的对数似然比为：

$$\log \Lambda(\mathbf{c}) = \sum_{i=1}^{\binom{n}{k}} c_i \log \frac{\hat{p}_{1,i}}{\hat{p}_{0,i}}$$

其中 $\hat{\mathbf{p}}_j = \arg\max_{\mathbf{p}\in\Theta_j} \Pr(\mathbf{c}|\mathbf{p})$ 为在假设 H_j 下的最大似然估计（MLE）。当 Θ_0 不给定先验时，$\hat{\mathbf{p}}_0 = \mathbf{c}/N$（经验分布）。$\hat{\mathbf{p}}_1$ 由扰动模块 $\mathcal{P}$ 诱导的目标分布（后处理模型 $\theta_{\text{clean}} + \mathcal{P}$ 的激活模式分布）及其正则化估计（如 Dirichlet 先验）获得。决策规则为：

$$\log \Lambda(\mathbf{c}) \gtrless \tau$$

其中阈值 τ 根据假阳性率 α 的要求选取。

序贯概率比检验（**SPRT**）：在线积累单样本 $\Sigma(x_t)$ 的似然比：

$$\Lambda_{\text{SPRT}}(\mathbf{c}_t) = \prod_{i=1}^{t} \frac{p(\Sigma(x_i)|H_1)}{p(\Sigma(x_i)|H_0)} = \prod_{i=1}^{t} \frac{p_{1,\sigma_i}}{p_{0,\sigma_i}}$$

其中 σ_i 表示第 i 个样本的激活模式索引。SPRT 在达到决策阈值时提前终止，报告期望样本数（**ASN**）与功效。

检测功率（功效）定义为：

$$\beta = \inf_{\mathbf{p}_1\in\Theta_1} P(\log \Lambda(\mathbf{c}) > \tau | H_1, \mathbf{p}_1)$$

假阳性率（False Positive Rate）：

$$\alpha = \sup_{\mathbf{p}_0\in\Theta_0} P(\log \Lambda(\mathbf{c}) > \tau | H_0, \mathbf{p}_0)$$

样本复杂度与阈值：基于 **Sanov** 定理或 **Chernoff** 界，当所有分量 $p_{0,i}, p_{1,i}$ 非零且两分布分离时，错判率指数受 **Bhattacharyya** 距离控制。对于多项分布族，Bhattacharyya 距离为：

$$D_B = -\ln \sum_{i=1}^{\binom{n}{k}} \sqrt{p_{0,i} p_{1,i}}$$

据此可给出样本数 N 与错误概率 (α, β) 的关系，以及 ROC-AUC 的保守下界。

8.3 ROC 曲线与 AUC 下界（基于错判率上界）

检测质量由接收者操作特征（ROC）曲线量化：

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

对于多项分布族，我们基于 **Bhattacharyya** 距离给出错判率上界，进而推导 AUC 的保守下界。

Bhattacharyya 距离（多项分布族）：

$$D_B = -\ln \sum_{i=1}^{\binom{n}{k}} \sqrt{p_{0,i} p_{1,i}}$$

当所有分量 $p_{0,i}, p_{1,i}$ 非零且两分布分离时，错判率上界为：

$$P_{\text{error}} \lesssim e^{-N \cdot D_B}$$

其中 N 为样本数。该上界基于多项分布族的 Chernoff 界，适用于分布充分分离的情况。

AUC 保守下界：在多项分布族下，基于错判率上界可构造 ROC 的保守下界。当 $D_B > 0$ 且样本数 N 足够大时：

$$\text{AUC} \geq 1 - e^{-N \cdot D_B} - \alpha$$

其中 α 为假阳性率上界。该下界明确适用于多项分布模型，并需满足分布分量非零且充分分离的条件。近似误差与分布分离程度和样本数相关。

9 综合性能指标

9.1 水印质量函数与帕累托前沿

质量定义：聚焦于模型输入空间的攻击，我们将鲁棒性拆分为两个独立且相互竞争的指标：

1. $\mathcal{R}_{\text{param}}$（参数鲁棒性）：在后处理框架下，这是一个新的挑战。如果模型 θ_{clean} 被第三方微调了（例如 θ'_{clean}），扰动模块 $\mathcal{P}$（它是为 θ_{clean} 设计的）是否仍然有效？$\mathcal{P}$ 需要对 θ 的小幅变化具有鲁棒性。这要求扰动模块 $\mathcal{P}$ 能够适应模型参数的变化，或设计对参数变化不敏感的扰动策略。

2. $\mathcal{R}_{\text{input}}$（语义鲁棒性）：对抗释义、同义词替换等输入空间规避攻击。在后处理框架下，语义鲁棒性完全由扰动模块 $\mathcal{P}$ 的语义感知能力决定。

综合定义水印系统的质量为四维指标：

$$\mathcal{Q} = (C_{\text{achievable}}, \mathcal{R}_{\text{param}}, \mathcal{R}_{\text{input}}, \text{AUC})$$

其中：

- $C_{\text{achievable}}$：可实现容量
- $\mathcal{R}_{\text{param}}$：参数鲁棒性（对抗参数修改）
- $\mathcal{R}_{\text{input}}$：语义鲁棒性（对抗输入空间规避）
- AUC：检测 AUC

多目标优化：编码器 $\mathcal{E}$（水印嵌入算法）在四维空间中寻找权衡点：

$$\max\{C_{\text{achievable}}, \mathcal{R}_{\text{param}}, \mathcal{R}_{\text{input}}, \text{AUC}\}$$

$$\text{s.t.} \quad \Delta_{\text{perf}} \leq \epsilon, \quad \alpha \leq \alpha_{\max}$$

帕累托前沿定义为非被支配的解集合，即不存在其他解在所有目标上都不劣于当前解且至少在一个目标上更优。

关键权衡关系（后处理框架）：

1. $\mathcal{R}_{\text{input}}$ **vs** Δ_{perf}（语义鲁棒性 **vs** 性能）：为了提高 $\mathcal{R}_{\text{input}}$，扰动模块 $\mathcal{P}$ 必须在更大的语义等价类上施加一致的扰动，这要求 $\mathcal{P}$ 对更多输入施加扰动，从而不可避免地导致更大的性能下降 Δ_{perf}。性能失真来源于扰动 $\mathcal{P}$ 对任务损失 $\mathcal{L}_{\text{task}}$ 的瞬时干扰。

2. $\mathcal{R}_{\text{input}}$ **vs** $C_{\text{achievable}}$（语义鲁棒性 **vs** 容量）：扰动模块 $\mathcal{P}$ 的语义感知能力越强，需要在更大的语义簇上嵌入水印，而非仅针对特定触发词，这使得嵌入单个比特的难度急剧增加，因此可实现容量 $C_{\text{achievable}}$ 会显著下降。鲁棒容量 C_{robust} 由满足类内一致性的模式数决定，通常小于 $C_{\text{achievable}}$。

权重设定：质量函数的加权形式 $Q_{\text{normalized}} = \alpha C_n + \beta R_{\text{param},n} + \gamma R_{\text{input},n} + \delta A_n$（其中 $\alpha+\beta+\gamma+\delta=1$）仅用于政策选择（根据具体应用需求选择折中点），而非理论结论。权重来源与任务依赖需明确：不同数据集/任务（如生成式 vs 分类式 MoE）可能需要不同的权重配置。建议以 **Pareto** 前沿为主图呈现不加权的解集，再由应用需求选择折中点。

与开放问题的连接：该框架回应了如何构造" 稳健验证器" 以处理" 域漂移" 的开放问题。" 文本释义" 正是" 对抗性域漂移" 的一种形式。在后处理框架下，通过语义感知的扰动模块 $\mathcal{P}$ 作为实现 $\mathcal{R}_{\text{input}}$ 的机制，我们提供了一个具体的工程路径来探索四维帕累托前沿。

10 数学推导示例

10.1 示例：基于激活模式的水印嵌入

假设设计仅基于激活模式 Σ 进行水印，使用 $n=8$ 个专家、$k=2$ 个激活的配置：

信息来源：从 $n=8$ 个专家中选择 $k=2$ 个激活，共有 $\binom{8}{2}=28$ 种可能的激活模式

$$I_{\text{pattern}} = \log_2 \binom{8}{2} = \log_2 28 \approx 4.8 \text{ bits}$$

消息映射：通过带密钥 K 的哈希函数 $H_K(m)$ 将消息 m（例如"U-C-Berkeley"）映射到目标激活模式 Σ_m（例如，$\Sigma_m = \{$专家 3, 专家 7$\}$）。

编码方案（后处理框架）：使用推理时扰动模块 $\mathcal{P}$，对路由器的 logits 输出施加扰动。对于输入 x，获取原始 logits $\mathbf{l}=[l_1,\ldots,l_8]$，计算最小扰动 δ_l 使得：

$$\min_{i\in\Sigma_m}(l_i+\delta_{l,i}) > \max_{i\notin\Sigma_m}(l_i+\delta_{l,i}) + \text{margin}$$

扰动模块 $\mathcal{P}$ 的目标是使目标专家（Σ_m 中）的最低 logit 必须比所有其他专家的最高 logit 还要高出一个 margin，从而确保 Σ_m 成为 top-k 激活模式。

性能约束（后处理框架）：在数据分布 $p(x)$ 上的性能失真和统计失真（仅基于激活模式 Σ）：

$$D_{\text{perf}} = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{clean}+\mathcal{P}} \le \epsilon_1$$

$$D_{\text{detect}} = \mathbb{E}_{x\sim p(x)}\left[D_{\text{KL}}\left(p(\Sigma(x)|\theta_{\text{clean}})||p(\Sigma(x)|\theta_{\text{clean}},\mathcal{P})\right)\right] \le \epsilon_2$$

其中 $\text{Acc}_{\text{clean}+\mathcal{P}}$ 表示在干净模型上应用扰动模块 $\mathcal{P}$ 后的准确率，$p(\Sigma|\theta_{\text{clean}},\mathcal{P})$ 表示后处理模型的激活模式分布。

可实现容量：通过调整扰动策略（如 margin、触发概率等），可以凭经验测量实际互信息 $I(m;\mathbf{f})$，其中 $\mathbf{f}$ 是测试集上激活模式的经验分布（计数向量 $\mathbf{c}$）。实际容量 $C_{\text{achievable}}$ 通常小于理论容量 I_{pattern}，因为需要在性能失真和统计失真之间进行权衡。

检测能力（多项分布模型，后处理框架）：对于观测计数向量 $\mathbf{c}\sim\text{Multinomial}(N,\mathbf{p})$，在 H_0 下 $\mathbf{p}=\mathbf{p}_0$（干净模型 θ_{clean}），在 H_1 下 $\mathbf{p}=\mathbf{p}_1$（后处理模型 $\theta_{\text{clean}}+\mathcal{P}$）。基于 Bhattacharyya 距离：

$$D_B = -\ln\sum_{i=1}^{28}\sqrt{p_{0,i}p_{1,i}}$$

当分布充分分离时，错判率上界为 $P_{\text{error}} \lesssim e^{-N\cdot D_B}$。基于 Pinsker 不等式，总变差距离满足：

$$\text{TV}(\mathbf{p}_0,\mathbf{p}_1) \le \sqrt{\frac{1}{2}D_{\text{KL}}(\mathbf{p}_1||\mathbf{p}_0)}$$

若 $D_{\text{KL}}(\mathbf{p}_1||\mathbf{p}_0)\ge\delta$，则验证成功率下界为：

$$P_{\text{detect}} \ge 1-\sqrt{\frac{\delta}{2}}-\alpha$$

其中 α 为假阳性率上界。在语义层面释义攻击下，需要考虑语义等价输入对分布的影响，此时应使用鲁棒容量 C_{robust} 而非 $C_{\text{achievable}}$。

11 开放的理论问题

尽管本文建立了 MoE 路由水印的信息论框架，但仍存在一些开放的理论问题：

1. 稀疏门控与单纯形容量的精确化：在 top-k 稀疏门控与概率单纯形约束下，路由水印的有效自由度与容量是多少？如何用类型法（**Method of Types**）为离散部分给出可靠编码区与错误指数？

2. 嵌入-性能的率失真闭环：在给定性能降级门限下，能否得到显式的容量上/下界？如何将性能损失上界映射为对路由分布的散度球约束（KL 或 Rényi 散度）？

3. 复合假设检验的稳健验证器：如何构造稳健最优的验证器以处理语义层面的释义攻击？如何比较以 Σ、π、$\mathbf{r}_\Sigma$ 为特征的不同充分性与功效？

4. 隐蔽性（**Steganographic Security**）与可检性下界：路由分布的统计隐蔽性如何量化？如何定义对手的检测器族并给出最小可辨性下界？

5. 语义层面的稳定性与泛化能力：在后处理框架下，如何设计语义感知的扰动模块 $\mathcal{P}$，使得激活模式对语义等价的输入保持稳定？如何量化语义变化对路由状态分布的影响？如何避免水印对触发器输入的过拟合，确保扰动模块能够识别语义等价类并施加一致的扰动？

6. 参数鲁棒性 $\mathcal{R}_{\mathbf{param}}$ 的实现机制：在后处理框架下，如何设计扰动模块 $\mathcal{P}$ 使其对模型参数的小幅变化（如微调、蒸馏）具有鲁棒性？如何实现扰动模块的自适应机制，使其能够适应模型参数的变化？如何量化参数变化对扰动模块有效性的影响？

7. 四维帕累托前沿的探索：如何在更高维度的空间（$C_{\text{achievable}}, \mathcal{R}_{\text{param}}, \mathcal{R}_{\text{input}}, \text{AUC}$）中寻找最优权衡点？如何量化 $\mathcal{R}_{\text{input}}$ 与 Δ_{perf} 以及 $\mathcal{R}_{\text{input}}$ 与 $C_{\text{achievable}}$ 之间的权衡关系？

12 结论

本文提出了 MoE 路由水印的信息论形式化框架，采用后处理（**Post-processing**）范式，聚焦于模型输入空间的语义层面释义攻击，为路由水印的设计和评估提供了严格的理论基础。我们建立了基于推理时扰动模块的信道模型，明确了编码器（扰动模块 $\mathcal{P}$）、信道和输出的定义，实现了非侵入性水印嵌入。我们量化了激活模式（组合码，Combination Code）的编码容量，并揭示了其在输入空间攻击下的脆弱性：水印可能对触发器输入过拟合，只学习表层 token 关联而非语义关联，导致实际可实现鲁棒容量可能骤降为 0。我们建立了二维率失真框架，同时考虑性能失真 D_{perf}（由推理时扰动造成的瞬时性能下降）和统计失真 D_{detect}（隐蔽性），并提出了路由分布扰动型后处理水印机制，通过语义感知的扰动模块设计来对抗输入空间的释义攻击，将语义鲁棒性的实现从训练损失转移到扰动模块的算法设计上。我们定义了语义等价类和鲁棒容量 C_{robust}，并定义了语义鲁棒性 $\mathcal{R}_{\text{input}}$ 来量化水印对输入空间规避攻击的抵抗能力，建立了四维帕累托前沿，揭示了 $\mathcal{R}_{\text{input}}$ 与 Δ_{perf} 以及 $\mathcal{R}_{\text{input}}$ 与 $C_{\text{achievable}}$ 之间的关键权衡关系。在验证过程中，我们构建了基于复合假设检验的框架（GLRT/SPRT），显式化观测向量为多项分布模型，强调只使用激活模式特征进行验证，并基于多项分布族的 Bhattacharyya 距离推导了检测质量的 AUC 下界。这些理论结果为 MoE 路由水印的实践应用提供了指导，并为未来的研究指明了方向，包括稀疏门控与单纯形容量的精确化、嵌入-性能-隐蔽性的率失真闭环、复合假设检验的稳健验证器、隐蔽性分析、语义层面的稳定性与泛化能力、参数鲁棒性 $\mathcal{R}_{\text{param}}$ 的实现机制以及四维帕累托前沿的探索等开放问题。