ICML 实验方案与评测清单（SoK 风格）

主题：基于 MoE 路由权重（Routing Weights, RW）与 LSH（Locality-Sensitive Hashing）的鲁棒语义水印

1. 背景与目标（ICML 定位）

问题：传统基于伪随机函数（PRF）的绿色词表水印在强释义与短文本场景下检测功效下滑。利用 MoE 的路由权重（RW）作为内生语义信号，结合 LSH 将连续向量映射为离散签名，驱动 token-level 的绿色词表选择与统计检测。目标：在生成端近零开销下，提升对释义与黑盒清洗攻击的鲁棒性，同时保持文本质量与易复现性。

2. 方法总览与关键直觉

• RW 作为内生语义；• LSH 稳定签名；• 在线嵌入 + z-score 检测。

3. 实验设计与数据集

模型：Mixtral-8x7B/Instruct、OpenMoE-8B/34B、Mistral-3 MoE（对照：Llama-3/Ministral-14B 密集）。任务：开放问答、摘要（CTG）、数据到文本。语料：C4/PG-19/新闻 + 人类写作片段。

4. 对比基线（Baselines）

A) 绿色词表水印（ICLR'24）；B) Unigram-Watermark（ICLR'24）；C) SemStamp（NAACL'24）；D) Multi-bit 水印（USENIX'25）。

5. 攻击与后处理集

释义（人类/LLM/bigram）；黑盒清洗（B4；策略集；RL 自适应）；混合与裁剪（窗口化检测）。

6. 指标与统计裁决

AUC、TPR@FPR、z-score vs n、窗口化检测；质量：Perplexity/ROUGE/BLEU；$Z = (X - n \cdot p)/\mathrm{sqrt}(n \cdot p \cdot (1-p))$。

7. 运行时与资源开销

生成端近零；检测端需重构 RW，成本随文本长度线性；报告 tokens/sec、GPU-小时、峰值内存。

8. 复现实验时间线（4–6 周）

第 1–2 周：实现与基线；第 3–4 周：攻击与窗口化；第 5–6 周：跨模型移植与撰写。

9. 评审式清单（Checklist）

□ 威胁模型；□ LSH 下界与多层放大；□ 强攻覆盖；□ 系统对比；□ MoE 生态证据；□ 三方折衷曲线。

10. 图表草图模板（随附 PNG）

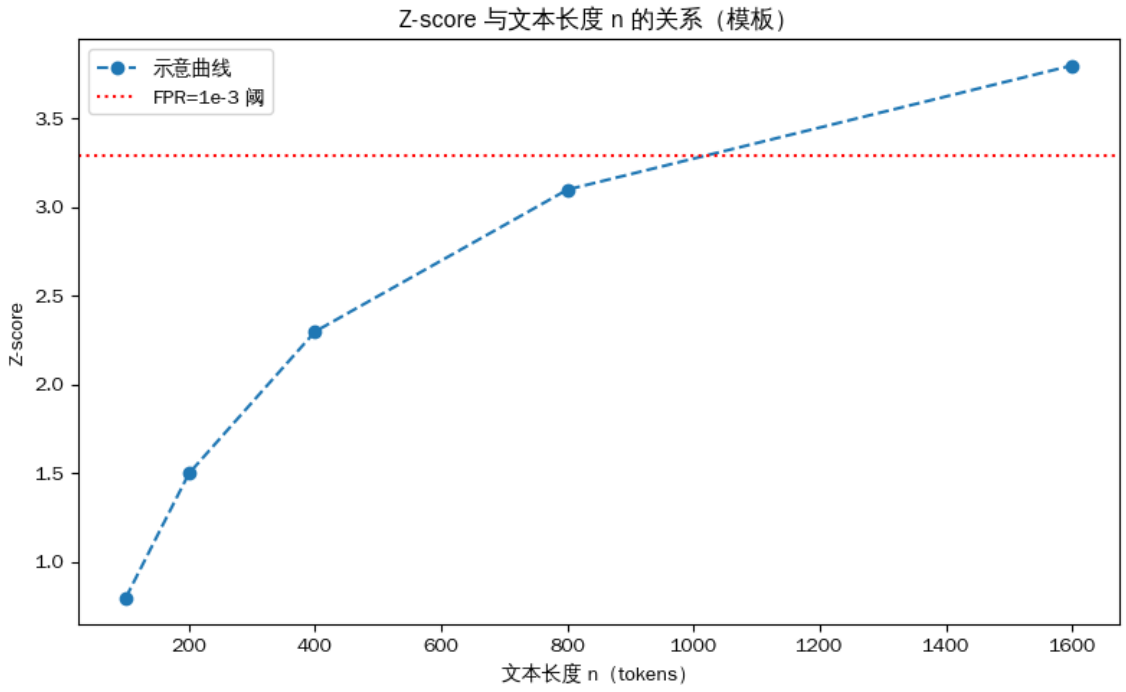

图 1：Z-score 与文本长度关系（模板）

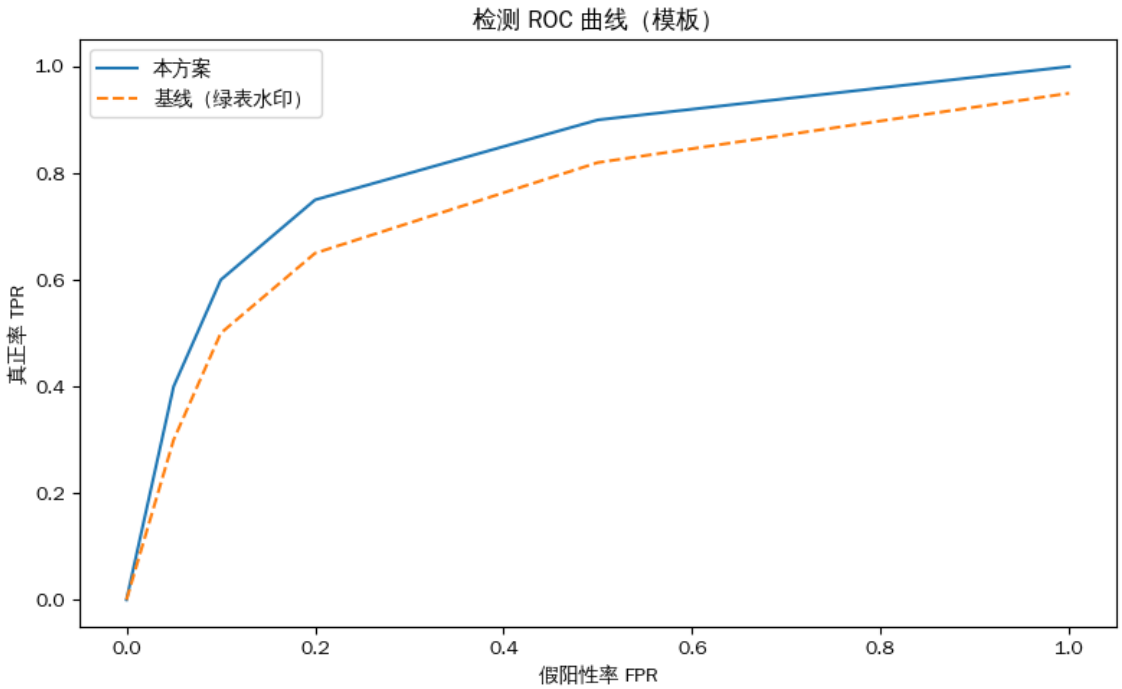

图 2：检测 ROC 曲线（模板）

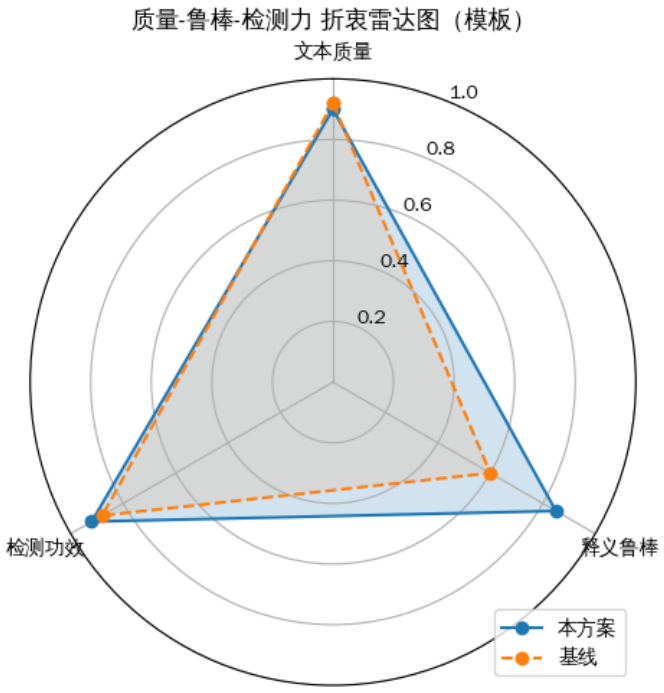

图 3：质量-鲁棒-检测力雷达图（模板）

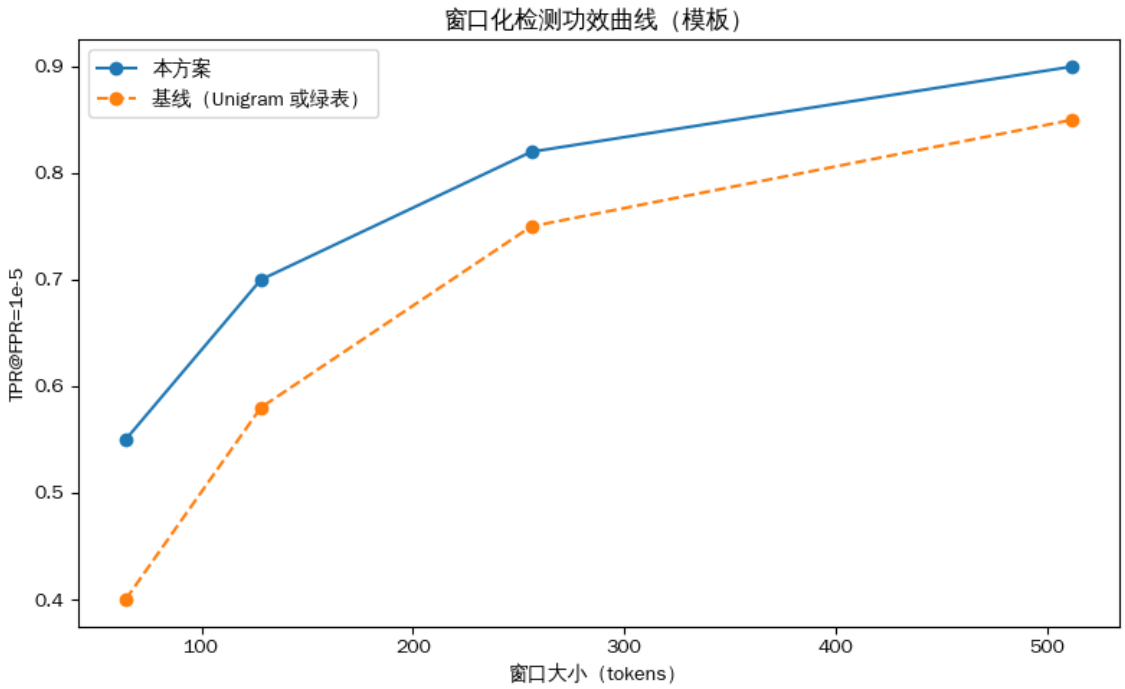

图 4：窗口化检测功效曲线（模板）

11. 脚本骨架（HF + vLLM/sglang）说明

- hf_moe_lsh_processor.py
- detect_pipeline.py
- attacks.py
- vllm_adapter.py
- sglang_adapter.py
- eval_runner.py
- viz_templates.py

12. 风险、伦理与密钥管理

白盒检测适用于平台内治理；公众可验提供窗口化统计降级。密钥轮换与多域隔离。

13. 参考文献（精选）

Kirchenbauer ICLR’24；SemStamp NAACL’24；Unigram ICLR’24；Multi-bit USENIX’25；WaterPark arXiv’25；B4 NAACL’25；OpenMoE/Mixtral/Mistral-3。