HLPO 原生推理性能基准测试报告 (Native Inference Benchmark Report)

1. 测试概览 (Overview)

本测试旨在验证 **HLPO (HMF-Laplace-Pomegranate-Ouroboros)** 架构在原生推理模式下的性能增益。我们使用了 **Run #2 (Aggressive Sparsity)** 训练出的模型，并在 Apple M2 Ultra (MPS) 平台上对比了“标准推理”与“原生物理推理”的性能差异。

测试环境

- **设备**: Apple M2 Ultra
- **加速**: MPS (Metal Performance Shaders) / FP32
- **模型**: HLPO GPT-2 Mini (Run #2 Checkpoint)
- **数据**: WikiText-2 Test Set (使用训练时的自定义词表, Vocab Size: 33277)

---

2. 测试方案 (Methodology)

对比配置

1. **Baseline (Training Only Mode)**:
    - 模型加载 Run #2 稀疏权重。
    - `inference_mode = False`。
    - 完整执行所有层（Laplace + Gravity + Kinetic + FFN），模拟普通 Transformer 的计算路径。
    - *注：此时模型虽然权重稀疏，但计算图是密集的。*
2. **Native (Mass Gating Mode)**:
    - 模型加载 Run #2 稀疏权重。
    - `inference_mode = True`。
    - **物理跳过 (Physics Skipping)**: 当感知到 Token 的质量 (m) 低于阈值 (10.0) 时，自动跳过 `Gravity` (拓扑构建) 和 `Kinetic` (注意力挤压) 计算，仅保留残差传播和必要的 FFN。
    - *注：这是 HLPO 所谓的“暗物质穿透”模式。*

---

3. 测试结果 (Results)

Metric (指标)	Baseline (普通模式)	Native (原生模式)	Delta (变化)
TPS (Tokens/Sec)	**25,560**	**134,435**	**+426% (5.26x Speedup)**
PPL (Perplexity)	1,031.57	1,406.90	+36.3% (Quality Drop)
Active Rate	100%	~0.4%	-99.6% (Sparsity)

关键发现 (Key Findings)

1. **极速爆发 (Massive Speedup)**: 开启原生推理模式后，吞吐量提升了 **5.26倍**。这证实了 HLPO 的"质量门控 (Mass Gating)"机制在软件层面也能带来巨大的性能红利，甚至超越了普通的稀疏矩阵加速比。
2. **精度权衡 (Quality Trade-off)**: PPL 从 1031 上升到 1406。考虑到 Run #2 的训练 PPL 约为 1236，原生推理的 PPL 略高于训练水平，但仍在可接受范围内。这表明模型成功学会了在低质量区域"偷懒"，但这部分被跳过的计算确实包含少量信息。
3. **HPU模拟验证**: 该测试验证了 HPU (Holographic Processing Unit) 的设计逻辑。如果在硬件层面实现 Logic Gating，配合 Power Gating（之前验证的 >99% 动态功耗节省），HLPO 芯片将具有极高的能效比。

---

4. 结论与建议 (Conclusion)

HLPO Run #2 模型配合原生推理算法，实现了 **"5倍速度，1.3倍PPL"** 的交换。

- **对于即时性要求极高的场景 (Real-time Edge AI)**: 5.26x 的加速是决定性的，PPL 的损失完全可以接受。
- **对于精度要求高的场景**: 可以适当降低 `gravity.threshold` (如从 10.0 降至 5.0)，在速度和精度之间找到更优的平衡点。

Next Step:

- 进一步微调阈值，寻找帕累托最优前沿 (Pareto Frontier)。
- 将 Python 侧的跳过逻辑下沉到 Rust (hlpo_core) 或 CUDA Kernel 中以获得更极致的性能。