模拟验证 I：7B 模型验证

> **目标**：量化 HLPO+HCA 对标准 LLaMA 规模模型（32层）的影响。

1. 实验设置

- **拓扑**：32 层串行管道。
- **基准 (Baseline)**：高阻抗 (Z_{HIGH})，代表 GDDR6 显存带宽墙。全量注意力计算 (Full Attention)。
- **全息 (HLPO+HCA)**：低阻抗 (Z_{LOW})，代表近存计算。全息稀疏化 (Top-K 过滤)。

2. 结果：25倍的跨越

指标	传统架构	HLPO + HCA	提升幅度
信号完整性	3.82%	96.95%	**25.4倍**
有效通量	0.012 / cycle	0.303 / cycle	**25.4倍**

分析

即便在相对“浅”的 32 层深度下，传统架构也因为对抗内部阻抗而损失了 **96%** 的输入能量（信息）。HLPO+HCA 像超导体一样，将 **97%** 的信息无损传输到了输出端。

这对 7B 模型意味着什么？ 这意味着在 HLPO 硬件上运行 7B 模型，可以达到极高的吞吐量，或者在相同吞吐量下实现 25 倍的能效提升。