

# 架构设计 II：熵过滤 (Entropy Filtering)

---

## 1. 全量注意力的诅咒

---

传统的 Transformer 对所有 Token 进行“全量注意力” (Full Attention) 计算。

- 在 1000 个 Token 的序列中，也许只有 50 个对下一次预测至关重要。
- 传统架构却计算了  $1000 \times 1000$  次交互。
- 这在物理上是低效的（高熵）。

## 2. 全息过滤器

---

HCA (全息计算架构) 就像包裹在每个计算单元外的过滤层。

- **机制：**在计算之前，它测量 Token 的“信息势能” (Information Potential)。
- **动作：**
  - 高密度：放行进入计算。
  - 低密度 (噪声)：旁路 / 衰减。

## 3. 硬件实现

---

在 HLPO 硬件中，这并非复杂的软件 `if` 语句（那太慢了）。它是一个物理层面的 **门控机制**（类似二极管）。

- 如果“电压”（重要性）低于阈值，门就不开启。
- 检查有效性所需的能量，相比于全量计算几近于零。

**结果：**我们将密集、高热的计算负载转化为稀疏、冷静、高价值的数据流。