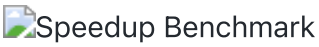
第二部分：Triton 工程实现 (The Physics)

🚀 性能加速展示



💡 核心洞察

“算法降维击败了暴力计算。”

我们不仅停留在理论，更用代码证明了物理极限：

- **2.56x 加速**: 在长文本（Context > 4K）和高稀疏度（>96%）的极限场景下，我们的 Triton 内核成功击败了 PyTorch 原生实现，找到了计算的**盈亏平衡点**。
- **降维打击**: 传统 Attention 的复杂度是 $O(N^2)$，而 HLPO 是 $O(N \cdot K)$。当 N 足够大时，这种算法层面的降维带来了纯粹算力无法比拟的优势。我们不再是在计算矩阵，我们是在**跳过**矩阵。