PR #14702 完整报告

sgl-project/sglang

fix topk softmax performance issue

合并时间：2026-03-30 14:43

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/14702

执行摘要

本 PR 修复了 topk softmax 内核的性能问题，通过使用 std::partial_sort 替代全排序算法，仅对前 k 个元素进行排序。这一优化减少了不必要的计算开销，提升了 CPU 端 expert routing 的效率，变更直接合并，风险较低。

功能与动机

原始代码在计算 topk softmax 时使用全排序，而实际只需要前 k 个元素，导致性能浪费。PR body 明确指出 'original code uses full sort but we only need topk'，因此进行优化以减少排序范围，解决性能瓶颈。

实现拆解

关键修改位于 sgl-kernel/csrc/cpu/topk.cpp 文件的 topk_softmax_kernel_impl 函数。具体改动如下：

- 改动前：使用 std::partial_sort(queue.begin(), queue.begin() + num_experts_per_group, queue.end(), ...)
- 改动后：使用 std::partial_sort(queue.begin(), queue.begin() + topk, queue.end(), ...) 通过调整排序结束位置，减少了排序范围，从而优化性能。

评论区精华

无 review 讨论，PR 直接合并，表明变更被快速接受，无争议点。

风险与影响

风险包括排序参数变更可能引入边界条件错误，例如当 topk 与 num_experts_per_group 不一致时导致排序范围不匹配；以及缺乏单元测试验证排序逻辑的正确性和稳定性。影响方面，优化提升了 topk softmax 计算性能，对 CPU 内核有正面作用，但对整体系统架构无重大改变。

关联脉络

从近期历史 PR 分析看，sgl-kernel 模块常涉及性能优化，如 PR 21448 修复 MoE 模型加载问题，但本 PR 聚焦于排序算法改进，未发现直接关联的跨 PR 演进脉络。