PR #17695 完整报告

sgl-project/sglang

[NPU] enhance accuracy for model minimaxm2 from 16.5% to 95.5%

合并时间：2026-03-23 19:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/17695

执行摘要

本 PR 修复了在 Ascend NPU 硬件上运行 minimaxm2 模型时出现的严重准确性缺陷，通过调整 fused_topk_npu 函数的条件逻辑，将准确率从 16.5% 大幅提升至 95.5%，并新增专项测试确保修复效果。变更聚焦于 NPU 后端核心路径，对用户可靠性和系统性能有显著正面影响。

功能与动机

动机源于 NPU 后端对 minimaxm2 模型的支持不足，准确率低下（仅 16.5%），严重影响推理结果。PR body 明确指出需解决此问题以提升模型在硬件上的可靠性，引用原话："Previously, the accuracy for npu for model minimaxm2 is no more than 16.5%"。

实现拆解

变更涉及三个文件：

- 核心修复：python/sglang/srt/hardware_backend/npu/moe/topk.py 中，将条件从 if not use_grouped_topk: 改为 if not use_grouped_topk and correction_bias is None:，优化 NPU 融合 topk 操作。
- 测试支持：python/sglang/test/ascend/test_ascend_utils.py 添加常量 MINIMAX_M2_WEIGHTS_PATH。
- 验证用例：新增 test/registered/ascend/llm_models/test_ascend_minimax_m2.py，包含 GSM8K 数据集测试，确保准确性不低于 90%。

评论区精华

review 讨论中的关键点：

- 性能权衡：gemini-code-assist[bot] 指出环境变量检查可能带来开销，建议缓存值，但作者未明确回应。
- 测试优化：Hexq0210 建议调整测试配置（tp-size 从 1 改 8），作者采纳以确保测试准确性。

风险与影响

风险：

1. 条件逻辑变更可能影响其他使用 fused_topk_npu 的场景，需依赖现有测试覆盖。

2. 环境变量检查（若引入）可能引入轻微性能开销。

3. 测试仅覆盖特定模型和数据集，范围有限。

影响：

- 对用户：准确性提升近 80 个百分点，增强信任度。
- 对系统：修复 NPU 后端瓶颈，提升硬件利用率。
- 对团队：提供可复现测试，便于维护和扩展。

关联脉络

与近期 PR 如 #18233（MoE 性能优化）和 #20214（MoE 后端集成）相关，显示团队持续关注硬件加速和模型准确性改进。本 PR 是 NPU 后端修复系列的一部分，有助于推动整体系统稳定性。