PR #21524 完整报告

sgl-project/sglang

[AMD] Add MiniMax-M2.5 nightly perf benchmarks for MI30x and MI35x

合并时间：2026-04-03 16:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21524

执行摘要

本 PR 为 AMD MI30x 和 MI35x 硬件添加了 MiniMax-M2.5 模型的夜间性能基准测试，通过扩展现有 CI 工作流在准确性测试后集成性能步骤，遵循准确性与性能结合模式，使用 continue-on-error 策略确保 CI 稳定性，提升了对 AMD 硬件性能的监控能力。

功能与动机

为什么做？根据 PR body，目标是扩展 CI 测试覆盖，为 AMD 硬件的 MiniMax-M2.5 模型添加性能基准测试，以持续跟踪性能指标。这遵循了已有的结合准确性和性能的模式（如 Grok1-INT4、Grok2 等），确保模型在 AMD 硬件上的表现符合预期。

实现拆解

改动点按模块梳理：

1. 测试文件新增 - test/registered/amd/perf/mi30x/test_minimax_m25_perf_amd.py：定义 MI30x 性能测试套件，配置 TP=8、EP=8、aiter 后端，批量大小 1/8/16/64，输入长度 4096，输出长度 512。 - test/registered/amd/perf/mi35x/test_minimax_m25_perf_mi35x.py：类似但针对 MI35x 硬件，包含环境变量设置和报告生成函数。 - 关键代码逻辑：generate_simple_markdown_report 函数生成简化 Markdown 报告，跳过预热运行；TestNightlyMiniMaxM25Performance 类设置测试配置。
2. CI 工作流修改 - .github/workflows/nightly-test-amd.yml 和 .github/workflows/nightly-test-amd-rocm720.yml：在每个 MiniMax-M2.5 准确性测试作业后添加性能测试步骤，设置 continue-on-error: true 和 120 分钟超时，使用 scripts/ci/amd/amd_ci_exec.sh 执行测试。

评论区精华

Review 讨论摘要：没有具体讨论内容，审核者 bingxche 和 HaiShaw 直接批准，表明变更被认为直接且无争议，无需深入技术交锋。

风险与影响

技术风险：

- 性能测试可能因硬件资源限制或环境波动而失败，但通过 continue-on-error: true 减轻了 CI 整体失败风险。
- 新增测试增加了 CI 运行时间和资源消耗（超时 120 分钟），可能影响其他作业调度。

- 测试覆盖新模型配置（aiter 后端），需确保与现有基础设施兼容，避免配置错误导致误报。

影响评估：

- 对用户：提供更全面的性能数据，帮助监控 AMD 硬件上模型表现。
- 对系统：CI 流程略有延长，但不会因性能测试失败阻塞构建。
- 对团队：增强测试套件，支持持续性能监控和优化，为 AMD 硬件调优提供数据支持。

关联脉络

与历史 PR 的关联：

- PR 21519 修复了 bench_one_batch 中的并行元数据错误，与本 PR 的性能测试直接相关，确保性能分析准确性。
- PR 21947 和 PR 20871 涉及 AMD 硬件性能优化和代码路径统一，反映出团队持续关注 AMD 性能改进的趋势，本 PR 是这一脉络中的测试扩展环节。
- 整体来看，这些 PR 共同推动了 AMD 硬件的测试覆盖和性能监控，为系统稳定性和优化提供支撑。