PR #21710 完整报告

sgl-project/sglang

[AMD] Add GLM-5-FP8 nightly performance benchmarks for MI30x and MI35x

合并时间：2026-04-08 13:43

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21710

执行摘要

本 PR 为 AMD MI30x 和 MI35x 平台新增 GLM-5-FP8 模型的夜间性能基准测试，通过扩展 CI 工作流和添加测试文件，实现性能监控与跨平台配置对齐，但存在代码质量和可移植性风险需后续关注。

功能与动机

为解决 AMD 平台性能监控缺失问题，本 PR 旨在添加 GLM-5-FP8 的夜间性能测试，以匹配 NV 和 InferenceX 的配置标准。PR body 强调“matching NV/InferenceX configs”并引用关联 Issue 评论建议添加 --reasoning-parser=glm45 --tool-call-parser=glm47 标志，减少跨平台行为漂移。

实现拆解

- CI 工作流修改：在 .github/workflows/nightly-test-amd.yml 和 nightly-test-amd-rocm720.yml 中添加性能测试步骤，设置 continue-on-error: true，确保准确性测试失败时作业仍可继续。
- 性能测试文件新增：创建 test/registered/amd/perf/mi30x/test_glm5_perf_amd.py 和 test/registered/amd/perf/mi35x/test_glm5_perf_mi35x.py，使用 bench_one_batch 方法，关键配置如下：python other_args=["--trust-remote-code", "--reasoning-parser", "glm45", "--tool-call-parser", "glm47", "--tp", "8", "--kv-cache-dtype", "fp8_e4m3", "--mem-fraction-static", "0.85"]
- 准确性测试更新：同步 test/registered/amd/accuracy/ 下的文件，将模型路径从 zai-org/GLM-5 切换至 zai-org/GLM-5-FP8，并添加相同 parser flags。

评论区精华

- gemini-code-assist[bot] 指出：“generate_simple_markdown_report 函数是重复的，应考虑移至共享模块”，但此建议未在 PR 中采纳。
- 关于除零风险：“Potential ZeroDivisionError if result.output_throughput is zero”，提示添加检查以避免崩溃。
- 硬编码路径问题：“Hardcoding environment variables like HF_HOME reduces portability”，建议通过 CI 环境配置提升可移植性。

- 环境变量不一致：“SGLANG_USE_AITER is missing for MI35x”，可能影响性能比较的准确性。

风险与影响

- 风险：除零错误可导致测试中断，影响 CI 稳定性；硬编码路径限制测试在非 CI 环境运行；环境变量缺失使 MI35x 性能数据可能不准确；代码重复增加维护负担。性能测试使用 continue-on-error，失败可能被忽略，掩盖潜在问题。
- 影响：对用户无直接影响，但增强内部测试覆盖，有助于团队监控 AMD 平台性能趋势，支持优化决策。新增测试提供标准化基准数据，促进跨平台配置对齐。

关联脉络

本 PR 依赖 #22314（修复 FP8 KV 量化路径）和 #22232（优化 NSA indexer）的修复以启用关键功能，与近期 PR 如 #22288（测试模型更新）共同反映 AMD 测试套件的持续演进。提交历史显示第二个提交启用了 FP8 KV 缓存，表明实现基于依赖 PR 的合并状态进行演进。