PR #22336 完整报告

sgl-project/sglang

[AMD] Add GLM-5.1-FP8 nightly accuracy and performance benchmarks for MI30x and MI35x

合并时间：2026-04-09 13:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22336

执行摘要

本 PR 为 AMD MI30x 和 MI35x 硬件新增 GLM-5.1-FP8 MoE 模型的夜间准确性与性能基准测试，扩展测试覆盖以评估新模型表现，基于 GLM-5 测试模式优化 CI 工作流并移除过时任务，但存在硬编码路径、配置不一致等风险需后续关注。

功能与动机

PR 旨在解决 GLM-5.1-FP8 模型在 AMD 硬件上缺乏系统化测试的问题。PR body 中说明："Add GLM-5.1-FP8 nightly accuracy + perf benchmarks for MI30x and MI35x"，基于 GLM-5 测试模式（#21710），确保模型推理准确性和性能可监控，满足持续集成需求。

实现拆解

- 新增测试文件：
  - test/registered/amd/accuracy/mi30x/test_glm51_eval_amd.py：MI30x 准确性测试，使用 GSM8K 数据集评估模型，配置 TP=8、NSA 后端。
  - 类似文件用于 MI35x 准确性和性能测试。
- 修改 CI 工作流：
  - .github/workflows/nightly-test-amd.yml：添加 nightly-8-gpu-glm51 等任务，移除 GLM-4.7-FP8。
  - .github/workflows/nightly-test-amd-rocm720.yml：同步更新支持 ROCm 7.2。
- 关键配置示例（从 test_glm51_perf_amd.py）：

```
cls.model_config = {
    "name": "glm51",
    "model_path": GLM51_MODEL_PATH,
    "other_args": [
        "--tp", "8",
        "--nsa-prefill-backend", "tilelang",
        "--kv-cache-dtype", "fp8_e4m3",
    ],
    "env_vars": {"SGLANG_USE_AITER": "1"},
}
```

评论区精华

review 中 gemini-code-assist[bot] 指出主要问题：

- “Hardcoding environment-specific paths like /data2/models/huggingface reduces portability.”
- “The configuration for MI35x accuracy is missing several environment variables... This inconsistency might lead to suboptimal performance.”
- “Potential ZeroDivisionError if result.output_throughput is zero.” 1am9trash 询问 EP 配置差异，作者回复“updated”并通过提交修正。讨论焦点集中在代码质量和配置一致性上。

风险与影响

- 技术风险：硬编码路径限制测试可移植性；配置不一致可能导致 MI35x 测试结果不准确；ZeroDivisionError 可能使性能报告崩溃；CI 变更可能引入任务失败风险。
- 影响范围：对用户间接提供基准数据；系统 CI 运行时间增加（如 MI35x 测试耗时超 1 小时）；团队需维护新增测试，确保配置同步。

关联脉络

此 PR 是 AMD 硬件测试扩展的一部分，与历史 PR 如 #21710（GLM-5 测试模式）直接关联，共同构建 SGLang 项目的多硬件测试矩阵。近期 PR 如 #22657（AMD Docker 修复）和 #22187（HiSparse 基准）显示团队持续优化测试基础设施，本 PR 延续了这一趋势，聚焦于新模型 GLM-5.1 的集成。