PR #27126 完整报告

sgl-project/sglang

[AMD] Add MiniMax-M2.5 TP=4 nightly accuracy test for MI355X

合并时间：2026-06-04 13:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27126

执行摘要

- 一句话：为 MiniMax-M2.5 添加 MI355X 夜间 GSM8K 准确度测试
- 推荐动作：该 PR 设计良好，测试配置清晰，值得阅读以了解如何在 AMD CI 中注册和运行硬件特定的准确度测试。重点关注 ModelConfig 的封装方式和 CI job 的定义模式。

功能与动机

根据 PR 描述，该测试验证 MiniMax-M2.5 在 4× MI355X GPU 上推荐部署配置（FP8 + 统一注意力 + FP8 KV 缓存）的准确性，确保配置在夜间回归中持续达标。

实现拆解

1. 创建测试脚本 test_minimax_m25_tp4_eval_mi35x.py：定义 ModelConfig 数据类封装 MiniMax-M2.5 的模型路径、TP 大小、准确率阈值、额外参数和环境变量。测试通过 run_gsm8k_benchmark 函数执行 5-shot GSM8K 基准，使用 sglang 的 RuntimeEndpoint 向启动的服务器发送请求，提取最终数字答案并计算准确率。
2. 修改 .github/workflows/nightly-test-amd.yml 和 nightly-test-amd-rocm720.yml：在 CI 工作流中新增 nightly-4-gpu-mi35x-minimax-m25 和 nightly-4-gpu-mi35x-minimax-m25-rocm720 job，分别对应 ROCm 7.0 和 7.2 环境。每个 job 配置 120 分钟超时，通过 amd_ci_exec.sh 执行测试套件 nightly-amd-4-gpu-mi35x-minimax-m25-tp4。

关键文件：

- test/registered/amd/accuracy/mi35x/test_minimax_m25_tp4_eval_mi35x.py（模块 测试脚本；类别 test；类型 test-coverage；符号 ModelConfig, post_init, get_display_name, get_one_example）：核心测试文件，定义了测试配置、基准逻辑和模型配置类。
- .github/workflows/nightly-test-amd.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：在 ROCm 7.0 夜间 CI 中添加 worklow job，定义测试运行环境和步骤。
- .github/workflows/nightly-test-amd-rocm720.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：在 ROCm 7.2 夜间 CI 中添加 worklow job，定义测试运行环境和步骤。

关键符号：ModelConfig.post_init, ModelConfig.get_display_name, get_answer_value, run_gsm8k_benchmark, few_shot_gsm8k

评论区精华

Review 中 gemini-code-assist[bot] 建议优化 get_answer_value 函数：移除不必要的 ast.literal_eval 改用 int() 并添加类型检查。最终代码采纳了该建议，使用了 int(numbers[-1]) 和类型检查，提升了稳健性和效率。

- 优化 get_answer_value 函数 (performance): 已采纳，最终代码使用 int() 和类型检查。

风险与影响

- 风险：该 PR 仅添加测试和 CI 配置，不修改生产代码，风险较低。主要风险在于测试依赖于特定的硬件（MI355X）和环境（ROCm 7.0/7.2），如果环境变化可能导致测试失败或需要调整。此外，测试运行时间较长（120 分钟），可能影响 CI 排队时间。
- 影响：对用户无直接影响，但夜间测试的覆盖增加有助于确保 AMD 平台上 MiniMax-M2.5 部署配置的准确性。对团队来说，增加了维护测试套件的负担，但能够早期捕获回归问题。
- 风险标记：依赖特定硬件环境，测试时长较长

关联脉络

- 暂无明显关联 PR