PR #21669 完整报告

sgl-project/sglang

[AMD] Add Qwen3.5-397B FP8 nightly perf benchmarks for MI30x and MI35x

合并时间：2026-04-07 14:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21669

PR 分析报告：为 AMD 添加 Qwen3.5-397B FP8 夜间性能基准测试

执行摘要

本 PR 在 sglang 仓库中为 AMD MI30x 和 MI35x GPU 平台新增了 Qwen3.5-397B FP8 模型的夜间性能基准测试，通过扩展 CI 工作流在准确性测试后运行性能步骤，并使用 continue-on-error 机制避免性能失败阻塞 CI，旨在增强硬件性能监控而不影响开发流程稳定性。

功能与动机

为什么做？根据 PR body，主要动机是补充现有 Qwen3.5 准确性测试，添加针对 FP8 量化模型的性能基准测试，以监控 AMD GPU 上的性能回归。关键表述包括："Perf steps run after existing accuracy tests in the same CI job, with continue-on-error: true so perf failures don't block CI when accuracy passes"，这确保了性能测试失败时不影响整体 CI 通过，平衡了测试覆盖与开发效率。

实现拆解

实现方案按模块拆解如下：

模块	关键改动点	说明
测试文件	新增 test/registered/amd/perf/mi30x/test_qwen35_fp8_perf_amd.py 和 test/registered/amd/perf/mi35x/test_qwen35_fp8_perf_mi35x.py	包含基准测试逻辑，使用 NightlyBenchmarkRunner 运行不同 batch size 和输入长度的测试，并生成简化 Markdown 报告。关键函数 generate_simple_markdown_report 计算吞吐量和 ITL（每令牌延迟）。
CI 工作流	修改 .github/workflows/nightly-test-amd.yml 和 .github/workflows/nightly-test-amd-rocm720.yml	在现有 Qwen3.5 测试任务中添加性能步骤，配置超时（120 分钟）和 continue-on-error: true，通过 run_suite.py 调用对应测试套件。

模块	关键改动点	说明
准确性测试	更新 test/registered/amd/accuracy/mi30x/test_qwen35_eval_amd.py 和 test/registered/amd/accuracy/mi35x/test_qwen35_eval_mi35x.py	覆盖 test_lm_eval 方法，将 lm-eval 结果以表格形式写入 GitHub step summary，并将 attention backend 从 triton 切换为 aiter。

关键代码片段（来自 generate_simple_markdown_report）：

```
itl = 1 / (result.output_throughput / result.batch_size) * 1000 # 潜在除零风险
summary += f"| {result.batch_size} | {result.input_len} | {result.latency:.2f} | {result.input_
throughput:.2f} | {result.output_throughput:.2f} | {itl:.2f} |\n"
```

评论区精华

review 讨论中最有价值的交锋集中在代码质量和设计决策上：

- 代码重复问题：gemini-code-assist[bot] 指出：“There is significant code duplication between this file and test/registered/amd/perf/mi35x/test_qwen35_fp8_perf_mi35x.py”，建议重构共享逻辑，但作者未在 PR 中响应，这留下了维护隐患。
- 除零风险：同一评论者建议修改 ITL 计算：“itl = (result.batch_size / result.output_throughput) * 1000 if result.output_throughput > 0 else 0”，以避免 ZeroDivisionError，但 PR 未采纳此建议。
- attention backend 切换：Jackycheng0808 简洁建议：“We should use --attention-backend aiter instead.”，作者在后续 commit 中迅速实施，体现了对性能优化的重视。

风险与影响

技术风险：

1. 代码重复：两个性能测试文件几乎相同，长期维护可能导致不一致或错误扩散。
2. 计算错误：ITL 计算缺少对 output_throughput 为零的检查，在极端情况下可能引发测试崩溃。
3. 可移植性：MI35x 测试文件中硬编码 Hugging Face 缓存路径（/data2/models/huggingface），限制了在其他环境中的运行。
4. CI 稳定性：新增性能步骤虽用 continue-on-error 缓冲，但超时设置（5400 秒）和资源竞争可能影响整体 CI 耗时。

影响分析：

- 用户影响：无直接功能变更，仅为内部测试增强。
- 系统影响：扩展了 AMD 硬件测试覆盖，有助于早期发现性能回归，但性能失败不影响 CI 通过，降低了阻塞风险。
- 团队影响：提供了标准化的性能数据收集流程，支持团队监控模型在特定硬件上的表现。

关联脉络

从近期历史 PR 看，本 PR 是 sglang 仓库持续扩展测试基础设施的一部分：

- 相关 PR：如 #22199 添加了 Ngram 推测解码的基准测试，共享类似的性能测试模式；#22203 引入了动态 HTTP API 支持，反映团队对测试灵活性的追求；#22207 更新 CI 权限，表明 CI 配置的频繁调整。
- 演进趋势：仓库近期多次添加针对特定硬件（如 AMD、NPU）和模型（如 Qwen、Gemma）的测试，显示团队正加强异构环境下的质量保障，本 PR 是这一趋势在 AMD 平台上的具体体现，未来可能进一步集成更多性能监控工具。