

PR #27577 完整报告

sgl-project/sglang

[NPU] Fix CI

合并时间: 2026-06-09 01:44

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27577>

执行摘要

- 一句话: 修复 NPU W4A4 测试中缺失的并行配置
- 推荐动作: 该 PR 属于常规的 CI 修复, 可快速合并。建议关注后续 NPU 测试套件的稳定性, 确保类似配置缺失问题不再出现。

功能与动机

NPU CI 测试因缺少 `gsm8k_parallel` 配置而失败, 该配置默认值可能不适合大规模模型 (如 Qwen3-32B), 导致测试性能不达标或超时。PR 提交者通过补充该配置项来修复 CI。

实现拆解

1. 定位问题: 在 `test/registered/ascend/basic_function/quant/test_npu_w4a4_quantization.py` 文件中, `TestAscendW4A4` 类缺少 `gsm8k_parallel` 属性, 而基类或测试框架期望该值来控制并行请求数。
2. 添加配置: 在第 50 行 (原 `output_throughput` 之后) 插入 `gsm8k_parallel = 64`, 指定 GSM8K 测试的并行度为 64。
3. 影响: 此配置使测试在 CI 环境中能以更高的并发执行推理请求, 提高资源利用率和测试通过率。

关键文件:

- `test/registered/ascend/basic_function/quant/test_npu_w4a4_quantization.py` (模块 量化测试; 类别 `test`; 类型 `test-coverage`): 修复 CI 的核心文件, 添加了缺失的 `gsm8k_parallel` 配置。

关键符号: 未识别

关键源码片段

`test/registered/ascend/basic_function/quant/test_npu_w4a4_quantization.py`

修复 CI 的核心文件, 添加了缺失的 `gsm8k_parallel` 配置。

```
# test/registered/ascend/basic_function/quant/test_npu_w4a4_quantization.py
# 省略 import 和辅助函数
class TestAscendW4A4(GSM8KAscendMixin, CustomTestCase):
```

```
model = ECO_TECH_QWEN3_32B_W4A4_LAOS_WEIGHTS_PATH
other_args = [
    "--trust-remote-code",
    "--device", "npu",
    "--attention-backend", "ascend",
    "--tp-size", "4",
    "--mem-fraction-static", "0.8",
    "--cuda-graph-bs", "64",
    "--disable-radix-cache",
]
env = {**os.environ}
# GSM8K 测试配置
accuracy = 0.80 # 正确率 >= 0.80
num_questions = 1319
gsm8k_num_shots = 5
output_throughput = 1000 # 输出吞吐量 >= 1000 tokens/s
gsm8k_parallel = 64 # 新增：并行度设置，修复 CI 缺失配置
```

评论区精华

PR 无 review 评论，仅由 [ping1jing2](#) 直接批准。CI 触发命令 `/tag-and-rerun-ci` 由 [ping1jing2](#) 发出，表明关注 CI 结果。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅为添加一行测试配置常量，不影响任何生产逻辑。唯一风险是 `gsm8k_parallel = 64` 可能不适用于资源受限环境，但该测试仅在 NPU CI 中运行且已设置 `est_time=400`，预计影响可控。
- 影响：
 - 用户：无直接影响。
 - 系统：仅影响 NPU CI 测试流程，修复后 W4A4 量化测试应正常通过。
 - 团队：减少 CI 噪音，避免因测试配置缺失导致误报。
 - 风险标记：低风险

关联脉络

- PR #24689 [NPU] Add GitHub test summary and deduplicate test code. Part 2: 同为 NPU 测试重构，涉及测试文件 `test_npu_w4a4_quantization.py` 的基类 Mixin 使用。