PR #26933 完整报告

sgl-project/sglang

Test case restoration in the full test.

合并时间：2026-06-04 20:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26933

执行摘要

- 一句话：启用 NPU 全量测试套件中的 6 个用例
- 推荐动作：无需精读；属于常规的测试配置恢复操作，可快速合并。

功能与动机

PR Body 明确指出：部分注册在全量测试套件中的用例因 nightly=False 而未被实际执行，同时一个导致 test_npu_grok_2.py 跳过的 Issue 已被关闭。

实现拆解

对 6 个测试文件中的 register_npu_ci 调用进行两处修改：

1. 将 nightly=False 改为 nightly=True；
2. 仅对 test_npu_grok_2.py 额外删除 disabled 参数。 变更简单、直接，不涉及任何业务逻辑或配置框架修改。

关键文件：

- test/registered/ascend/llm_models/test_npu_grok_2.py（模块 NPU Grok 2 测试；类别 test；类型 test-coverage）：核心变更之一：修改了 nightly 参数并移除了 disabled 标记，恢复 Grok 2 模型测试。
- test/registered/ascend/llm_models/test_npu_persimmon_8b_chat.py（模块 NPU Persimmon 测试；类别 test；类型 test-coverage）：将 nightly 参数改为 True，启用 Persimmon 8B Chat 模型的夜间测试。
- test/registered/ascend/llm_models/test_npu_c4ai_command_r_v01.py（模块 NPU C4AI 测试；类别 test；类型 test-coverage）：将 nightly 参数改为 True，启用 C4AI Command R v01 模型的夜间测试。
- test/registered/ascend/llm_models/test_npu_exaone_3.py（模块 NPU EXAONE 测试；类别 test；类型 test-coverage）：将 nightly 参数改为 True，启用 EXAONE 3 模型的夜间测试。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_low_latency_deepseek_v3_2_w8a8.py（模块 NPU DeepEP 测试；类别 test；类型 test-coverage）：将 nightly 参数改为 True，启用 DeepSeek V3.2 w8a8 模型的夜间专家并行测试。

- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_low_latency_qwen3_480b.py（模块 NPU DeepEP 测试；类别 test；类型 test-coverage）：将 nightly 参数改为 True，启用 Qwen3 480B 模型的夜间专家并行测试。

关键符号：未识别

关键源码片段

test/registered/ascend/llm_models/test_npu_grok_2.py

核心变更之一：修改了 nightly 参数并移除了 disabled 标记，恢复 Grok 2 模型测试。

```
import unittest

from sglang.test.ascend.gsm8k_ascend_mixin import GSM8KAscendMixin
from sglang.test.ci.ci_register import register_npu_ci
from sglang.test.test_utils import CustomTestCase

# 注册为全量测试套件，夜间执行，同时移除了之前因 Issue 而添加的 disabled 标记
register_npu_ci(
    est_time=400,
    suite="full-16-npu-a3",
    nightly=True, # 由 False 改为 True，启用夜间执行
)


class TestGrok2(GSM8KAscendMixin, CustomTestCase):
    model = "/root/.cache/modelscope/hub/models/huihui-ai/grok-2"
    accuracy = 0.91
    other_args = [
        "--trust-remote-code",
        "--mem-fraction-static",
        "0.8",
        "--attention-backend",
        "ascend",
        "--disable-radix-cache",
        "--disable-cuda-graph",
        "--tokenizer-path",
        "/root/.cache/modelscope/hub/models/huihui-ai/grok-2/tokenizer.tok.json",
        "--tp-size",
        "16",
    ]


if __name__ == "__main__":
    unittest.main()
```

评论区精华

无 reviewer 讨论；gemini-code-assist 的评论仅总结变更内容，未新增讨论。

- gemini 自动 code review (other): 无进一步反馈。

风险与影响

- 风险：风险极低：仅涉及测试调度配置，不修改任何源码逻辑。可能的风险是若某些用例本身不稳定，启用夜间运行后可能增加失败告警频率，但可通过后续单独修复解决。
- 影响：影响范围限于 NPU 夜间 CI：6 个之前被跳过的模型测试（Grok 2、Persimmon 8B、C4AI Command R v01、EXAONE 3、DeepSeek V3.2 w8a8、Qwen3 480B）将在夜间流水线中自动执行，提升 NPU 测试覆盖。
- 风险标记：暂无

关联脉络

- PR #27027 Solving the problem of test case failures caused by timeouts: 同一文件 test_npu_deepep_low_latency_deepseek_v3_2_w8a8.py 在后续 PR 中因超时被再次修复，说明该测试启用后可能需要进一步调整超时配置。
- PR #26775 fix test cases failed on 5/30 in nightly pipeline: 同为 NPU 测试修复，表明近期 NPU 夜间测试稳定性是关注重点，本 PR 恢复用例后可能增加类似修复工作量。