PR #26604 完整报告

sgl-project/sglang

[CP] Add back qwen 30b test

合并时间：2026-05-29 05:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26604

执行摘要

- 一句话：恢复误删的 Qwen3-30B CP 测试
- 推荐动作：该 PR 是简单的测试恢复操作，不值得深入阅读。但可以作为恢复误删文件的标准操作范例：确认原始内容、直接还原、并在 PR body 说明原因。

功能与动机

该测试文件在 #23269 中被误删除，需要恢复以保持对 Qwen3-30B 模型在 context parallelism 模式下的回归测试覆盖。PR body 明确指出 'it was deleted by mistake in #23269'。

实现拆解

1. 在 test/registered/cp/ 下完全新增 test_qwen3_30b.py 文件（+132 行）。
2. 定义模型路径为 Qwen/Qwen3-30B-A3B-FP8，GSM8K 准确率基线为 0.85。
3. 实现 TestQwen330B 类（基础配置）：启动 4-GPU 服务，TP=4，MoE-DP=2，EP=2，启用 prefill context parallel，禁用 piecewise CUDA graph，运行 200 条 GSM8K 样本。
4. 实现 TestQwen330BCP 类（CP 配置）：与基础配置类似，但 MoE-DP=1，EP=4，强调 expert parallelism 而非 data parallelism，更聚焦 context parallelism 效果。
5. 通过 register_cuda_ci 注册为 CI 阶段 extra-b，使用 4-gpu-h100 runner，预估耗时 261 秒。

关键文件：

- test/registered/cp/test_qwen3_30b.py（模块 测试；类别 test；类型 test-coverage；符号 TestQwen330B, setUpClass, tearDownClass, test_gsm8k）：PR 的唯一变更文件，完全新增，恢复被误删的 CP 测试。

关键符号：TestQwen330B.setUpClass, TestQwen330B.tearDownClass, TestQwen330B.test_gsm8k, TestQwen330BCP.setUpClass, TestQwen330BCP.tearDownClass, TestQwen330BCP.test_gsm8k

关键源码片段

test/registered/cp/test_qwen3_30b.py

PR 的唯一变更文件，完全新增，恢复被误删的 CP 测试。

```
import unittest
from types import SimpleNamespace

from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.run_eval import run_eval
from sglang.test.test_utils import (
    DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    DEFAULT_URL_FOR_TEST,
    CustomTestCase,
    kill_process_tree,
    popen_launch_server,
)

# 注册为 CI 阶段 extra-b，使用 4 卡 H100，预计耗时 261 秒
register_cuda_ci(est_time=261, stage="extra-b", runner_config="4-gpu-h100")

QWEN3_30B_MODEL_PATH = "Qwen/Qwen3-30B-A3B-FP8"
GSM8K_BASELINE_ACCURACY = 0.85


class TestQwen330BCP(CustomTestCase):
    """测试 Qwen3-30B 在 context parallelism 下的 GSM8K 准确性"""

    @classmethod
    def setUpClass(cls):
        # 4-GPU 配置：TP=4, MoE-DP=1, EP=4, 启用 prefill CP
        cls.model = QWEN3_30B_MODEL_PATH
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[
                "--tp-size", "4",
                "--moe-dp-size", "1",
                "--ep-size", "4",
                "--attn-cp-size", "2",
                "--enable-prefill-context-parallel",
                "--cuda-graph-max-bs", "32",
                "--max-running-requests", "32",
                "--trust-remote-code",
                "--disable-piecewise-cuda-graph",
                "--model-loader-extra-config",
                '{"enable_multithread_load": true, "num_threads": 64}',
            ],
        )

    @classmethod
    def tearDownClass(cls):
```

```
        kill_process_tree(cls.process.pid)

    def test_gsm8k(self):
        # 使用 5-shot GSM8K，采样 200 条，最大生成长度 16000 tokens
        args = SimpleNamespace(
            model=self.model,
            eval_name="gsm8k",
            num_shots=5,
            num_examples=200,
            max_tokens=16000,
            num_threads=128,
            repeat=1,
            temperature=0.6,
            top_p=0.95,
            top_k=20,
            base_url=self.base_url,
            host="http://127.0.0.1",
            port=int(self.base_url.split(":")[-1]),
        )
        metrics = run_eval(args)
        print(f"{metrics=}")
        # 确保 GSM8K 准确率不低于基线 0.85
        self.assertGreaterEqual(metrics["score"], GSM8K_BASELINE_ACCURACY)


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

评论区精华

评论中仅有一条来自 Fridge003 的 /rerun-test test_qwen3_30b.py 命令触发 CI 重跑，以及 gemini-code-assist 的配额提醒。没有深入 review 讨论。

- 触发 CI 重跑测试 (other): CI 运行成功，测试通过。

风险与影响

- 风险：无技术风险：变更仅涉及测试文件恢复，不影响任何推理代码、依赖或配置。但需确保 CI runner 能正常分配 4-GPU H100 实例以执行该测试。
- 影响：对用户：无直接影响。对系统：增加了 CI 回归测试覆盖，确保 Qwen3-30B 在 context parallelism 下的正确性持续被验证。对团队：恢复了不必要的测试遗漏，属于维护性修复。
- 风险标记：低风险，仅为测试恢复，需要 4-GPU CI runner

关联脉络

- PR #23269 （未知标题）：该 PR 误删了本测试文件，本 PR 是恢复操作。