PR #23029 完整报告

sgl-project/sglang

[test] Add GSM8K accuracy test for PP with mixed chunk prefill

合并时间：2026-04-17 17:09

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23029

执行摘要

- 一句话：新增流水线并行与混合分块预填充的 GSM8K 精度测试，验证功能兼容性。
- 推荐动作：该 PR 值得快速浏览，重点关注 TestPPMixedChunk 测试类的设计，它展示了如何为特定配置组合（PP+ 混合分块）添加端到端精度测试。对于涉及流水线并行或预填充优化的开发者，这个测试可作为参考模板。

功能与动机

PR body 明确指出，这是为了跟进 #22920 中移除 PP 与混合分块预填充兼容性限制的后续验证。需要添加一个 GSM8K 精度测试来确保 --enable-mixed-chunk 在 --pp-size > 1 时能正确工作，从而验证功能兼容性。

实现拆解

1. 新增测试类：在 test/registered/distributed/test_pp_single_node.py 中添加 TestPPMixedChunk 类，继承自 CustomTestCase。 - setUpClass 方法：使用 popen_launch_server 启动服务器，参数包括 --tp-size 2、--pp-size 2、--chunked-prefill-size 256、--enable-mixed-chunk。 - tearDownClass 方法：清理服务器进程。 - test_gsm8k 方法：调用 run_eval 运行 GSM8K 评估，并根据 CI 环境（AMD 或 NVIDIA）设置不同的精度阈值（0.70 或 0.74）。
2. 修复参数拼写错误：在同一个文件的 TestFixedBugs.test_chunked_prefill_with_small_bs 方法中，将 --chunked-prefill 更正为 --chunked-prefill-size，确保参数名称正确。
3. 更新使用说明：在文件顶部的注释中新增 TestPPMixedChunk.test_gsm8k 的示例调用行，方便开发者直接运行测试。

关键文件：

- test/registered/distributed/test_pp_single_node.py（模块 流水线并行；类别 test；类型 test-coverage；符号 TestPPMixedChunk, setUpClass, tearDownClass, test_gsm8k）：唯一变更文件，新增了 PP 与混合分块预填充的精度测试类，并修复了一个参数拼写错误。

关键符号：TestPPMixedChunk.setUpClass, TestPPMixedChunk.tearDownClass, TestPPMixedChunk.test_gsm8k

关键源码片段

test/registered/distributed/test_pp_single_node.py

唯一变更文件，新增了 PP 与混合分块预填充的精度测试类，并修复了一个参数拼写错误。

```
class TestPPMixedChunk(CustomTestCase):
    @classmethod
    def setUpClass(cls):
        cls.model = DEFAULT_MODEL_NAME_FOR_TEST # 使用默认测试模型
        cls.base_url = "http://127.0.0.1:23338" # 指定本地测试URL
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[
                "--tp-size", 2, # 张量并行大小为2
                "--pp-size", 2, # 流水线并行大小为2
                "--chunked-prefill-size", 256, # 分块预填充大小为256
                "--enable-mixed-chunk", # 启用混合分块预填充
            ],
        )

    @classmethod
    def tearDownClass(cls):
        if hasattr(cls, "process"):
            kill_process_tree(cls.process.pid) # 测试结束后清理服务器进程

    def test_gsm8k(self):
        args = SimpleNamespace(
            base_url=self.base_url,
            model=self.model,
            eval_name="gsm8k", # 运行GSM8K评估
            api="completion",
            max_tokens=512,
            num_examples=200, # 使用200个示例进行测试
            num_threads=128,
        )
        metrics = run_eval(args) # 执行评估并获取精度指标
        print(f"{metrics=}")

        if is_in_amd_ci():
            # AMD Triton后端精度略低于NVIDIA的FA3，设置较低阈值
            self.assertGreater(metrics["score"], 0.70)
        else:
            self.assertGreater(metrics["score"], 0.74) # NVIDIA环境设置标准阈值
        time.sleep(4) # 等待片刻以便内存检查执行
```

评论区精华

Review 中仅有两个简单的批准（LGTM），没有实质性的技术讨论或争议点。这表明变更被认可为直接且必要的测试补充。

- 暂无高价值评论线程

风险与影响

- 风险：1. 测试稳定性风险：测试依赖于外部服务器启动和 GSM8K 评估，可能因环境差异（如模型加载、网络延迟）导致偶发性失败。 2. 精度阈值风险：AMD CI 环境设置了较低的精度阈值（0.70），可能掩盖了某些平台特定的精度回归问题。 3. 参数错误遗留风险：修复的 --chunked-prefill 拼写错误仅在一个测试用例中更正，如果其他测试或代码中仍有类似错误，可能未被覆盖。
- 影响：1. 对用户影响：无直接影响，这是内部测试增强，不改变用户可见功能。 2. 对系统影响：增强了流水线并行与混合分块预填充功能的测试覆盖，有助于提前发现兼容性问题，提升系统稳定性。 3. 对团队影响：为后续相关功能开发提供了验证基准，减少了因兼容性疏忽导致 bug 的风险。
- 风险标记：测试稳定性依赖外部环境，精度阈值可能掩盖回归

关联脉络

- PR #22920 Remove PP and mixed chunk prefill compatibility restriction: 本 PR 是 #22920 的后续验证，该 PR 移除了 PP 与混合分块预填充的兼容性限制，本测试用于确认移除后功能正常工作。
- PR #23006 [Pipeline Parallelism][Bug] Fix scheduler hang in pipeline parallelism setup: 涉及流水线并行（PP）的 bug 修复，与本 PR 的 PP 测试场景相关，可能共享类似的调度或配置问题。
- PR #22990 [Bug Fix] Ensure prefill_info_table is populated before honoring disagg_prefill_dp_rank: 涉及预填充（prefill）的 bug 修复，与本 PR 的混合分块预填充测试相关，可能影响预填充逻辑的正确性。