PR #23303 完整报告

sgl-project/sglang

[CI][MLA] Enable deterministic inference for MGSM MLA FP8 test

合并时间：2026-04-21 13:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23303

执行摘要

- 一句话：为 MLA FP8 测试启用确定性推理，消除 MGSM-EN 分数波动导致的 CI 不稳定。
- 推荐动作：该 PR 是典型的测试稳定性修复，值得快速浏览以了解 FP8 和 MoE 模型中的非确定性来源及如何通过现有标志解决。关注点在于 PR body 中详细分析的根因（FP8 反量化噪声和 MoE 专家路由非确定性）以及 *--enable-deterministic-inference* 标志的端到端支持机制。

功能与动机

PR body 明确指出，test/registered/mla/test_mla_fp8.py::TestMLA::test_mgsm_en 测试在 0.8 阈值处不稳定，例如一次失败运行得分为 0.784，低于阈值约 1.6 个百分点。根本原因是 FP8 KV 缓存反量化引入的每令牌数值噪声和 DeepSeek-V2-Lite 的 MoE 专家路由中相同 logits 可能选择不同专家，两者结合导致 MGSM-EN 分数在 1-3 点范围内波动，跨越 0.8 阈值，造成 CI 随机失败。

实现拆解

1. 测试配置调整：修改 test/registered/mla/test_mla_fp8.py 文件，在 TestMLA.setUpClass 方法的 other_args 列表中添加 *--enable-deterministic-inference* 标志。
2. 标志作用机制：该标志在 server_args.py 中已实现端到端支持，它会自动选择兼容的注意力后端（Hopper 上为 fa3，Blackwell 上为 triton），将采样后端固定为 pytorch，并禁用分段 CUDA 图。
3. 测试策略不变：保持 mgsm_en_score_threshold = 0.8 阈值不变，因为启用确定性推理后分数将固定为（模型、权重、CUDA 堆栈）的函数，如果固定分数低于 0.8，后续 PR 可基于实测值调整阈值。
4. 性能影响评估：PR body 提到，禁用分段 CUDA 图和强制 pytorch 采样后端的开销很小，因为测试仅针对小模型运行 250 个问题单线程，且之前运行的总延迟 14.915 秒远低于预算 106 秒。

关键文件：

- test/registered/mla/test_mla_fp8.py（模块 MLA 测试；类别 test；类型 test-coverage；符号 TestMLA.setUpClass）：唯一修改的文件，在 MLA FP8 测试的服务器启动参数中添加了 *--enable-deterministic-inference* 标志，直接解决 CI 不稳定问题。

关键符号：TestMLA.setUpClass

关键源码片段

test/registered/mla/test_mla_fp8.py

唯一修改的文件，在 MLA FP8 测试的服务器启动参数中添加了 --enable-deterministic-inference 标志，直接解决 CI 不稳定问题。

```
@classmethod
def setUpClass(cls):
    cls.model = DEFAULT_MLA_FP8_MODEL_NAME_FOR_TEST
    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=[
            "--trust-remote-code",
            "--kv-cache-dtype",
            "fp8_e5m2",
            # 固定 MoE 专家分配和内核归约顺序，防止 MGSM 分数在多次运行中漂移。
            # 评估已使用贪婪解码，但 FP8 反量化 + 非确定性 MoE top-k 平局打破
            # 在没有此标志时会产生约 1-3 点的波动，并跨越 0.8 阈值。
            # 启用确定性推理后，分数变为（模型、权重、CUDA 堆栈）的固定函数，
            # 因此阈值边缘的随机失败不再是随机噪声。
            "--enable-deterministic-inference",
        ],
    )
```

评论区精华

review 评论中没有实质性技术讨论，只有自动化标签和重跑 CI 的指令。PR body 本身包含了详细的动机、修改和影响分析，相当于自包含的设计文档。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 性能风险：启用确定性推理会禁用分段 CUDA 图并强制使用 pytorch 采样后端，可能轻微增加推理延迟，但 PR body 评估开销很小，且测试时间预算充足。
  2. 测试覆盖风险：如果确定性推理本身有 bug 或未完全覆盖非确定性场景，可能掩盖实际生产环境中的问题，但这是测试配置调整，不影响生产代码。
  3. 阈值风险：保持 0.8 阈值不变，如果启用确定性推理后固定分数低于 0.8，测试将失败，需要后续 PR 调整阈值，但这是预期策略。
- 影响：

1. 对 CI 稳定性：显著提升，消除 MGSM-EN 分数波动导致的随机失败，使测试结果可复现。
2. 对用户：无直接影响，这是内部测试配置变更。
3. 对系统：无功能变更，仅影响测试环境的行为。
4. 对团队：减少 CI 噪音，提高开发效率，但需注意确定性推理可能轻微增加测试时间。 - 风险标记：测试性能轻微影响，阈值可能需后续调整

关联脉络

- 暂无明显关联 PR