PR #26395 完整报告

sgl-project/sglang

[AMD] [CI] Add DeepSeek-R1-0528 FP8 HiCache GSM8K test on MI35x

合并时间：2026-05-27 16:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26395

执行摘要

- 一句话：为 AMD MI35x 添加 DSR1-0528 FP8 HiCache GSM8K CI 测试
- 推荐动作：建议合并，该 PR 填补了关键的回归测试空白，且本地验证充分。未来可考虑增加更多数据集或不同 HiCache 配置的测试，以进一步覆盖边缘情况。

功能与动机

The existing nightly DSR1-0528 AMD eval tests intentionally pass --disable-radix-cache, so any regression that breaks DSR1-0528 generation correctness under HiCache has no per-commit coverage today. This PR closes that gap.

实现拆解

1. 创建测试文件：test/registered/amd/test_deepseek_r1_hicache_mi35x.py，导入必要的工具模块。
2. 注册 CI 套件：调用 register_amd_ci(est_time=900, suite="stage-c-test-large-8-gpu-amd-mi35x") 将该测试纳入 per-commit 流水线。
3. 配置环境和服务器参数：在 setUpClass 中设置环境变量（如 SGLANG_USE_AITER=1、ROCM_QUICK_REDUCE_QUANTIZATION=NONE）和服务器启动参数（TP=8、--kv-cache-dtype fp8_e4m3、--attention-backend aiter、--enable-hierarchical-cache 等），然后启动服务器。
4. 执行评估：在 test_gsm8k 中调用 run_eval 执行 GSM8K 200 题 5-shot 补全，将结果写入 GitHub Step Summary 并断言准确率 ≥ 0.93。
5. 清理：tearDownClass 终止服务器进程并删除临时 L3 存储目录。

关键文件：

- test/registered/amd/test_deepseek_r1_hicache_mi35x.py（模块 HiCache 测试；类别 test；类型 test-coverage；符号 TestDeepSeekR1HiCacheMI35x, setUpClass, tearDownClass, test_gsm8k）：唯一的变更文件，新增了完整的测试类覆盖 DSR1-0528 + HiCache 的回归测试。

关键符号：TestDeepSeekR1HiCacheMI35x.setUpClass, TestDeepSeekR1HiCacheMI35x.tearDownClass, TestDeepSeekR1HiCacheMI35x.test_gsm8k

评论区精华

该 PR 没有收到 review 评论，仅有一次来自 HaiShaw 的批准。PR body 详细说明了动机、本地验证结果和复现步骤，未引发讨论。

- 暂无高价值评论线程

风险与影响

- 风险：新测试依赖 AMD MI35x 硬件和特定环境（如 /data2/models/huggingface 缓存路径），不支持在其他平台上运行。服务器启动超时配置为 1500 秒，若超时则测试失败。测试总时长约 15 分钟（900 秒），可能延长 CI 队列等待时间。GSM8K 评估结果存在固有的随机性，阈值 0.93 提供了一定的容错余量，但仍存在偶然失败的可能性。
- 影响：对用户无直接影响。为 CI 系统增加了约 15 分钟的每 commit 测试开销，但扩大了 DeepSeek-R1-0528 在 HiCache 启用情况下的回归覆盖，有效填补了兼容性测试空白。对 AMD MI35x 平台的维护者而言是最直接的受益对象。
- 风险标记：平台依赖，CI 时长，测试随机性

关联脉络

- PR #26396 [AMD] [CI] Add GLM-5.1 MXFP4 TP2 accuracy gate: 同为 AMD MI35x 平台的 per-commit 准确率门控测试，共享相同的 CI 套件阶段和测试基础设施。
- PR #26425 [core] Maintain req_pool_indices_cpu host mirror (like seq_lens_cpu): 同为 HiCache 层次结构相关的改进，通过消除 D2H 复制提升性能，但新测试可间接验证其正确性。