

PR #29290 完整报告

sgl-project/sglang

[AMD] Cover DeepSeek-R1 MXFP4 TP4 MTP nightly CI

合并时间: 2026-07-01 16:21

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29290>

执行摘要

- 一句话: 添加 DS-R1 MXFP4 TP4 MTP 的 AMD 夜测。
- 推荐动作: 建议精读 setUpClass 中的环境变量配置, 可作为 AMD MI35x 上 AITER + MXFP4 最佳实践参考。同时也应关注 Gemini 建议的防御性检查, 可在后续修复中引入。

功能与动机

根据 PR 描述, 需要覆盖生产环境 DeepSeek-R1-MXFP4 TP=4 + EAGLE/MTP GSM8K 场景的 nightly 精度回归, 以确保 AITER 后端和 FP8 缓存在长期运行中的正确性。

实现拆解

1. 创建测试文件 test/registered/amd/accuracy/mi35x/test_deepseek_r1_mxfp4_tp4_mtp_mi35x.py, 注册为 nightly-amd-8-gpu-mi35x-deepseek-r1-mxfp4-tp4 套件。文件实现了 get_one_example、get_few_shot_examples、get_answer_value、run_gsm8k_benchmark 等工具函数, 以及 TestDeepSeekR1MXFP4TP4MTPMI35x 测试类。setUpClass 通过环境变量设置 AITER、FP8、重叠计划流等生产选项, 启动服务器后运行全量 GSM8K 5-shot 评估 (1319 题), 准确率阈值 0.944。
2. 修改 .github/workflows/nightly-test-amd.yml, 在原有 TP4 基线步骤 (test_deepseek_r1_mxfp4_tp4_mi35x.py) 之后, 添加 Accuracy Test MI35x (DeepSeek-R1-MXFP4 TP4 MTP) 步骤, 运行新建的 MTP 测试文件, 超时 180 分钟, 且前一步失败后继续执行。
3. 修改 .github/workflows/nightly-test-amd-rocm720.yml, 在 ROCm 7.2 对应位置增加同样的 MTP 步骤, 与 ROCm 7.0 形成双版本覆盖。两个工作流均保留了 continue-on-error 标志支持。测试文件通过 register_amd_ci 注册了估计耗时 3600 秒。

关键文件:

- test/registered/amd/accuracy/mi35x/test_deepseek_r1_mxfp4_tp4_mtp_mi35x.py (模块 测试套件; 类别 test; 类型 test-coverage; 符号 get_one_example, get_few_shot_examples, get_answer_value, run_gsm8k_benchmark): 核心测试文件, 实现了完整的全量 GSM8K 回归测试, 包括数据下载、few-shot 构造、AITER 环境配置和精度断言。
- .github/workflows/nightly-test-amd.yml (模块 CI 配置; 类别 infra; 类型 infrastructure): ROCm 7.0 夜测工作流, 新增 MTP 步骤与基线分离, 通过分步提升故障可观测性。

- `.github/workflows/nightly-test-amd-rocm720.yml` (模块 CI 配置; 类别 `infra`; 类型 `infrastructure`) : ROCm 7.2 夜测工作流, 同步添加 MTP 步骤, 确保双版本覆盖。

关键符号: `get_one_example`, `get_few_shot_examples`, `get_answer_value`,
`run_gsm8k_benchmark`, `TestDeepSeekR1MXFP4TP4MTPMI35x.setUpClass`,
`TestDeepSeekR1MXFP4TP4MTPMI35x.test_gsm8k`

关键源码片段

`test/registered/amd/accuracy/mi35x/test_deepseek_r1_mxfp4_tp4_mtp_mi35x.py`

核心测试文件, 实现了完整 GSM8K 回归测试, 包括数据下载、few-shot 构造、AITER 环境配置和精度断言。

```
class TestDeepSeekR1MXFP4TP4MTPMI35x(unittest.TestCase):
    @classmethod
    def setUpClass(cls):
        cls.model = os.environ.get(
            "DEEPSEEK_R1_MXFP4_MODEL_PATH", "amd/DeepSeek-R1-MXFP4-Preview"
        )
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.num_questions = int(os.environ.get("GSM8K_NUM_QUESTIONS", "1319"))
        cls.parallel = int(os.environ.get("GSM8K_PARALLEL", "1319"))

        env = os.environ.copy()
        # 启用 AITER 后端
        env["SGLANG_USE_AITER"] = "1"
        # 开启重叠计划流 kernel 启动
        env["SGLANG_ENABLE_OVERLAP_PLAN_STREAM"] = "1"
        # Speculative Decoding v2 (当前已退休, 为与报告配置一致保留)
        env["SGLANG_ENABLE_SPEC_V2"] = "1"
        # 关闭快速缩减量化, 保持 full precision
        env["ROCM_QUICK_REDUCE_QUANTIZATION"] = "NONE"
        # FP8 prefill attention on AITER
        env["SGLANG_AITER_FP8_PREFILL_ATTN"] = "1"
        # MLA 层持久化
        env["SGLANG_AITER_MLA_PERSIST"] = "1"
        # MXFP4 MoE scaling factor
        env["AITER_MXFP4_MOE_SF"] = "1"
        # 禁用 int4 权重
        env["SGLANG_INT4_WEIGHT"] = "0"
        # 启用 MoE padding 优化
        env["SGLANG_MOE_PADDING"] = "1"
        # 设置 CPU 亲和性
        env["SGLANG_SET_CPU_AFFINITY"] = "1"
        # ROCm fused decode MLA
        env["SGLANG_ROCM_FUSED_DECODE_MLA"] = "1"
        # 标记使用 ROCm 7.0.0 环境
        env["SGLANG_USE_ROCM700A"] = "1"
```

```

cls.process = popen_launch_server(
    model=cls.model,
    base_url=cls.base_url,
    timeout=SERVER_LAUNCH_TIMEOUT,
    other_args=[
        "--tensor-parallel-size", "4",
        "--trust-remote-code",
        "--mem-fraction-static", "0.9",
        "--chunked-prefill-size", "131072",
        "--attention-backend", "aiter",
        # ... 其余参数省略
    ],
    env=env,
)

def test_gsm8k(self):
    acc, invalid, latency = run_gsm8k_benchmark(
        self.base_url, self.num_questions, num_shots=5, parallel=self.parallel
    )
    self.assertGreaterEqual(acc, GSM8K_MTP_ACCURACY_THRESHOLD)

```

评论区精华

Gemini Code Assist 自动审查建议在 `tearDownClass` 中对 `cls.process` 添加 `None` 检查，避免服务器启动失败后访问 `pid` 导致 `AttributeError`。该评论未被作者采纳处理，PR 已合并，可能存在轻微稳定性风险。

- `tearDownClass` 防御性检查建议 (correctness): 作者未处理此建议，PR 已合并。

风险与影响

- 风险：主要风险在于测试依赖 MI35x 硬件环境，若硬件或 AITER 后端不可用，测试会静默跳过或超时失败。阈值 0.944 未提供合理 margin，模型更新可能造成误报警。环境变量 `SGLANG_ENABLE_SPEC_V2` 已被标记为退役（注释中提到），但仍在设置，可能在未来版本失效。
- 影响：对用户无直接影响；对团队：增加两个 CI 步骤，单次耗时约 3 小时，总计额外 6 小时夜测时间。但提供针对性精度保障，减少生产问题漏测。对其他平台无影响。
- 风险标记：硬件依赖，阈值脆弱，未处理建议

关联脉络

- PR #29409 [AMD] Split qwen3.5 triton DCP test into its own nightly job: 同为 AMD 夜测拆分优化，模式相似。