PR #27001 完整报告

sgl-project/sglang

[AMD] [CI] Remove hardcoded model/cache paths from MI35x nightly tests

合并时间：2026-06-03 17:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27001

执行摘要

- 一句话：移除 MI35x nightly 测试硬编码的 /data2 路径
- 推荐动作：此次变更为纯粹的测试基础设施清理，逻辑简单且已获 Approve。适合 CI 维护者快速合入，无需深度 review。关注模型 ID 的准确性即可。

功能与动机

PR body 指出 MI35x accuracy/perf nightly tests 中硬编码了 machine-specific 的 /data2 挂载，这并非预期行为。更改后模型解析为 HF ID 并下载到默认 cache，消除对特定 runner 环境的依赖。

实现拆解

1. 在每个 MI35x 测试文件顶部删除 os.environ.setdefault("HF_HOME", ...) 和 os.environ.setdefault("HF_HUB_CACHE", ...) 两行，不再将 HF cache 重定向到 /data2。
2. 移除模块级别的 DEEPSEEK_R1_MXFP4_LOCAL_PATH（/data2/...）、DEEPSEEK_R1_MXFP4_HF_MODEL_ID 等本地路径常量。
3. 删除各文件中的 get_model_path() 函数（优先环境变量 → 本地路径 → HF ID 的 fallback 链）。
4. 将 setUpClass 或 get_mxfp4_models() 中对 get_model_path() 的调用替换为直接赋值的 HF 模型 ID（如 "amd/DeepSeek-R1-MXFP4-Preview"）。
5. 删除 test_bench_one_batch 等测试方法中检查本地路径是否存在的 is_local_path 判断和 skip 逻辑，以及对应的 log 输出。
6. 清理 docstring 中与环境变量相关的 example usage 说明。
7. 仅保留测试运行所必需的 import 和辅助函数，部分文件移除了多余的 import os（当 os 不再使用时）。

关键文件：

- test/registered/amd/perf/mi35x/test_deepseek_r1_mxfp4_perf_mi35x.py（模块 性能测试；类别 test；类型 test-coverage；符号 get_model_path）：DeepSeek-R1-MXFP4 性能测试入口，变更典型：删除 get_model_path + is_local_path 跳过逻辑、内联模型 ID
- test/registered/amd/accuracy/mi35x/test_deepseek_r1_mxfp4_eval_mi35x.py（模块 精度测试；类别 test；类型 test-coverage；符号 get_model_path）：DeepSeek-R1-MXFP4

精度测试，同样的清理模式：删除 get_model_path 及 HF cache 重定向

- test/registered/amd/accuracy/mi35x/test_qwen3_coder_next_eval_mi35x.py（模块 精度测试；类别 test；类型 test-coverage；符号 get_model_path）：Qwen3-Coder-Next 精度测试，属于最后一批清理的 /data 路径，模式与其他文件一致

关键符号：get_model_path

关键源码片段

test/registered/amd/perf/mi35x/test_deepseek_r1_mxfp4_perf_mi35x.py

DeepSeek-R1-MXFP4 性能测试入口，变更典型：删除 get_model_path + is_local_path 跳过逻辑、内联模型 ID

```
"""MI35x Nightly performance benchmark for DeepSeek-R1-MXFP4 model."""

import os
import unittest
from typing import List

# ... 导入及辅助函数 ...

class TestDeepseekR1MXFP4PerfMI35x(unittest.TestCase):
    """Tests the DeepSeek-R1-MXFP4 quantized model on TP=8 with DP=8."""

    @classmethod
    def setUpClass(cls):
        # 原逻辑：cls.model = get_model_path() // 优先 env var > /data2 本地路径 > HF ID
        # 现直接固定为 HuggingFace 模型 ID，不再查询本地路径，消除机器绑定。
        cls.model = "amd/DeepSeek-R1-MXFP4-Preview"
        print(f"Using model path: {cls.model}")
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.batch_sizes = [1, 8, 16, 64]
        cls.input_lens = tuple(_parse_int_list_env("NIGHTLY_INPUT_LENS", "4096"))
        cls.output_lens = tuple(_parse_int_list_env("NIGHTLY_OUTPUT_LENS", "512"))

        cls.variants = [
            {
                "name": "basic",
                "other_args": [
                    "--trust-remote-code",
                    "--tp", "8",
                    "--chunked-prefill-size", "131072",
                    "--disable-radix-cache",
                    "--mem-fraction-static", "0.85",
                ],
            },
        ]
        # ... 省略后续 setup ...
```

```
def test_bench_one_batch(self):
    # 移除了原有的 is_local_path 检查与 skip 逻辑:
    # is_local_path = self.model.startswith("/")
    # if is_local_path and not os.path.exists(self.model):
    # self.skipTest(...)
    # 现在模型一定是 HF ID，运行时将自动下载（若未缓存）。
    failed_variants = []
    try:
        for variant_config in self.variants:
            with self.subTest(variant=variant_config["name"]):
                result_tuple = self.runner.run_benchmark_for_model(
                    model_path=self.model,
                    batch_sizes=self.batch_sizes,
                    input_lens=self.input_lens,
                    output_lens=self.output_lens,
                    other_args=variant_config["other_args"],
                    variant=variant_config["name"],
                    extra_bench_args=["--trust-remote-code"],
                    enable_profile=False,
                )
                # ... 处理结果 ...
    finally:
        self.runner.write_final_report()
```

评论区精华

该 PR 未产生实质性 review 讨论，仅有 Gemini Code Assist 的 quota 提示和两位 reviewer（yctseng0211、HaiShaw）的 Approve。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。测试行为由依赖本地预缓存改为在线下载 HF 模型，若 CI runner 网络不可达或模型 ID 发生变更则测试失败。另需确保替换后的模型 ID（如 amd/DeepSeek-R1-MXFP4-Preview）与之前路径指向的模型完全一致。
- 影响：影响范围限定于 AMD MI35x 的 nightly 测试（performance 和 accuracy 共约 27 个文件）。正面影响：测试不再绑定特定机器挂载，可在任意 runner 上复用。负面影响：首次运行时需下载模型，测试时长可能增加，但 nightly 任务在时间上可接受。对 sglang 核心逻辑无影响。
- 风险标记：依赖 HF 模型在线，模型 ID 准确性

关联脉络

- 暂无明显关联 PR