

PR #56710 完整报告

vllm-project/vllm

[CI][Bugfix] Complete FSE fixture R-SWA contract

合并时间: 2026-09-14 11:30

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/56710>

执行摘要

- 一句话: 修复 FSE 测试夹具缺 rswa_window 的 CI 回归
- 推荐动作: 建议花少量时间阅读。变更本身只有一行, 但它揭示了 backend 选择逻辑新增 model_config 字段读取时, 测试 fixture 的手工模拟对象容易发生契约漂移; 对维护 attention backend、mock fixture 或 CI 的工程师有借鉴价值。khluu 的复现排查过程 (跨主机确认确定性失败、用精确 lane 验证) 也是值得学习的 CI 问题定位方法。

功能与动机

该 PR 用于修复 #56305 引入的 H200 Model Executor 回归。PR body 说明: Fix the H200 Model Executor regression introduced by #56305, That PR made get_attn_backend() read model_config.rswa_window, but the focused FSE model-construction fixture still builds a minimal SimpleNamespace without that field。主 build #88651 在通过 709 个测试后, 12 个 test_models_fse_init 用例全部失败; khluu 在 issue 评论中又列出 4 个后续 post-main build 独立复现, 确认这是确定性回归而非宿主机偶发。

实现拆解

1. 根因定位: #56305 在注意力后端选择逻辑中新增对 model_config.rswa_window 的读取; 而 tests/model_executor/layers/test_fused_shared_expert.py 的 test_models_fse_init 用 SimpleNamespace 手工构造最小 model_config, 属性清单未同步, 导致 12 个 H200 用例在模型构造时抛出 AttributeError。
2. 最小修复: 在 fixture 的 SimpleNamespace 中补上 rswa_window=None, 与未启用 R-SWA 的默认语义一致; 这是唯一 diff (+1/-0), 不触碰生产路径。
3. 验证配套: pre-commit run --files tests/model_executor/layers/test_fused_shared_expert.py 通过; pytest --collect-only 收集 84 个用例; 本地 CPU 因 MXFP4 linear kernel 不可用, 模型构造在平台限制处停止, 最终由 Buildkite PR #88707 的 H200 lane 完成端到端验证。

关键文件:

- tests/model_executor/layers/test_fused_shared_expert.py (模块 测试夹具; 类别 test; 类型 test-coverage; 符号 test_models_fse_init): 唯一变更文件, 为 test_models_fse_init 的 SimpleNamespace model_config 补上 rswa_window=None, 与 get_attn_backend() 新增字段读取契约对齐, 解除 12 个 H200 用例的 AttributeError。

关键符号: test_models_fse_init

关键源码片段

tests/model_executor/layers/test_fused_shared_expert.py

唯一变更文件, 为 test_models_fse_init 的 SimpleNamespace model_config 补上 rswa_window=None, 与 get_attn_backend() 新增字段读取契约对齐, 解除 12 个 H200 用例的 AttributeError。

```
def test_models_fse_init(
    model_type: str,
    use_fse: bool,
    exclude: list[str],
    dist_init: None,
    monkeypatch: pytest.MonkeyPatch,
) -> None:
    """Model construction resolves FSE consistently with Quark quantization."""

    quantization_config: dict[str, Any] = (
        get_deepseek_v4_quark_config(["layers.0.ffn.shared_experts"] if exclude else [])
        if model_type == "deepseek_v4"
        else {**_QUARK_FSE_CONFIG, "exclude": exclude}
    )

    config, model_constructor = get_fse_test_model_config(
        model_type, quantization_config
    )
    vllm_config = VllmConfig()
    # 注意: 这里用 SimpleNamespace 手工构造轻量 model_config, 必须显式
    # 覆盖模型构造路径会读取的全部字段。PR#56305 之后 get_attn_backend()
    # 会读取 rswa_window, 因此补上非 R-SWA 默认值 None, 否则 12 个
    # H200 用例会抛 AttributeError。
    vllm_config.model_config = SimpleNamespace(
        hf_config=config,
        hf_text_config=config,
        dtype=torch.bfloat16,
        max_model_len=128,
        is_diffusion=False,
        head_dtype=None,
        is_mm_prefix_lm=False,
        multimodal_config=None,
        quantization_config=None,
        runner_type="generate",
        is_moe=True,
        logits_processors=None,
        rswa_window=None,
    )
    vllm_config.parallel_config.enable_expert_parallel = False
```

```

# deepseek_v4 分支需要额外配置 FP8 KV cache 与专用量化器,
# 其他模型统一走 QuarkConfig
if model_type == "deepseek_v4":
    from vllm.models.deepseek_v4.quant_config import DeepseekV4FP8Config

    vllm_config.cache_config.cache_dtype = "fp8_ds_mla"
    vllm_config.quant_config = DeepseekV4FP8Config(
        is_checkpoint_fp8_serialized=True,
        weight_block_size=[128, 128],
    )
else:
    vllm_config.quant_config = QuarkConfig(quantization_config)

# 后续通过 monkeypatch 切换 VLLM_ROCM_USE_AITER 等环境变量,
# 让模型构造在 FSE 与 Quark 量化配置下保持一致 (其余代码省略)
...

```

评论区精华

Review 区没有实质交锋：nooooo 直接批准，claude[bot] 因 fork 仓库关闭了自动 review。有价值的讨论在关联 issue 评论区：khluu 列出 4 个后续 post-main build (#88663、#88674、#88685、#88699) 在独立主机上复现同样 12 个 rswa_window AttributeError，加上最初 #88651 共 5 次出现，明确排除宿主机偶发；他还指出原 culprit PR 的 build #88637 与紧邻的 #88647 均未触发该 fixture，回归直到下一个 main commit 才暴露。最终 PR Buildkite #88707 在 head commit 84bd577cfbab 上通过，验证闭环。

- H200 回归的确定性与验证闭环 (testing): 判定为确定性回归，不再做无意义重试；PR Buildkite #88707 在 head commit 上通过 H200 Model Executor lane，验证完成。
- CI 触发与合并 (other): H200 CI 通过后 nooooo 直接批准并合并。

风险与影响

- 风险：变更仅 1 行且语义明确，回归风险极低；但暴露了 fixture 设计隐患：SimpleNamespace 手工维护的属性清单与生产代码读取属性之间是隐式契约，后续 get_attn_backend() 或其他函数再新增 model_config 字段读取时，同类 AttributeError 会复发。验证方面，本地 CPU 无法运行 MXFP4 kernel，完整验证依赖 H200 CI；若 CI 环境变动可能漏检。不涉及性能、安全或兼容性风险。
- 影响：对用户无任何运行时影响（生产代码零改动）；对 CI 解除 H200 Model Executor lane 的持续失败，恢复 12 个 test_models_fse_init 用例；对团队提供了一个上游属性新增导致测试夹具契约漂移的典型样本，提示后续可在 fixture 构造处增加集中校验。
- 风险标记：测试夹具契约漂移，H200 专属验证依赖，手工属性清单维护

关联脉络

- PR #56305 [Attention] Add Triton/FlashInfer composite for multimodal prefix attention: 本 PR 的直接起因：该 PR 让 get_attn_backend() 读取

model_config.rswa_window, 而 FSE 测试 fixture 未同步该字段, 导致 12 个 H200 用例
AttributeError。