PR #51177 完整报告

vllm-project/vllm

[CI Bug] Fix `Chunked prefill is required for mamba cache mode 'align'.`

合并时间：2026-08-06 02:33

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/51177

执行摘要

- 一句话：修复 Jamba 测试缺 chunked prefill 导致的 CI 失败
- 推荐动作：此 PR 为一行式测试修复，无需精读源码。可关注的要点是 mamba 缓存模式 align 与 chunked prefill 的硬约束，后续若新增 Mamba 架构模型测试应默认带上 enable_chunked_prefill=True。

功能与动机

Buildkite CI 构建 #82455 报错：Assertion failed, Chunked prefill is required for mamba cache mode 'align'。该断言来自 VllmConfig 的 pydantic 校验，根因是 ai21labs/Jamba-tiny-random 模型的 mamba 缓存模式 align 必须配合 chunked prefill，而测试启动引擎时未开启该选项。

实现拆解

变更入口为 tests/quantization/test_experts_int8.py 中的 test_model_experts_int8_startup，实现步骤：

1. 在 vllm_runner 构造参数中加入 enable_chunked_prefill=True。该参数直接传入 VllmConfig，使 mamba 缓存模式 align 的断言校验通过。
2. 无源码、配置或 schema 改动；测试行为上，Jamba 模型现在以 chunked prefill 方式运行，与 mamba 缓存模式的约束对齐。
3. 该修复只影响此单测的启动配置，不改变专家 INT8 量化（experts_int8）本身的校验逻辑。

关键文件：

- tests/quantization/test_experts_int8.py（模块 量化测试；类别 test；类型 test-coverage）：唯一的变更文件，在 vllm_runner 中增加 enable_chunked_prefill=True，解决 Jamba mamba 缓存模式 align 的 CI 校验失败。

关键符号：test_model_experts_int8_startup

关键源码片段

tests/quantization/test_experts_int8.py

唯一的变更文件，在 vllm_runner 中增加 enable_chunked_prefill=True，解决 Jamba mamba 缓存模式 align 的 CI 校验失败。

```
@pytest.mark.skipif(
    not is_quant_method_supported("experts_int8"),
    reason="ExpertsInt8 is not supported on this GPU type.",
)
@pytest.mark.parametrize("model", MODELS)
@pytest.mark.parametrize("dtype", ["bfloat16"])
@pytest.mark.parametrize("max_tokens", [4])
def test_model_experts_int8_startup(
    hf_runner,
    vllm_runner,
    example_prompts,
    model: str,
    dtype: str,
    max_tokens: int,
) -> None:
    model_info = HF_EXAMPLE_MODELS.find_hf_info(model)
    model_info.check_transformers_version(on_fail="skip")

    with vllm_runner(
        model,
        dtype=dtype,
        enforce_eager=True,
        quantization="experts_int8",
        # Jamba 的 mamba cache mode 为 "align" 时，VllmConfig 强制要求
        # 开启 chunked prefill，否则 pydantic 校验直接断言失败
        enable_chunked_prefill=True,
    ) as vllm_model:
        vllm_model.generate_greedy(example_prompts, max_tokens)
```

评论区精华

Review 讨论极少且全部通过。AndreasKaratzas 评论 LGTM I was about to post the same thing 🤭，表明修复方向直观、是社区预期中的改动；noooop 与 mgoin 均直接 APPROVED，未提出额外疑虑。另有多条 /ci run 与 /ci retry 指令用于触发和重试 Buildkite CI，最终验证通过。

- 修复方向获得 Reviewer 认可 (other): 所有人工 review 均通过，无分歧。
- CI 触发与重试验证 (other): CI 通过后 PR 被合入 main。

风险与影响

- 风险：风险极低，属于测试配置修复：
  - 该测试现在强制启用 chunked prefill，可能掩盖未来 mamba 缓存模式对 chunked prefill 约束放宽后的问题，但当前约束仍有效。
  - 仅影响 tests/quantization/test_experts_int8.py，不涉及任何生产代码路径。
  - 若未来其它 mamba 相关测试出现同类报错，需要同步补充 enable_chunked_prefill。
- 影响：影响范围限定为 CI 测试通道：

- 修复 Buildkite CI 上 experts_int8 启动测试的回归失败，恢复测试稳定性。
- 对最终用户无运行时行为影响。
- 对测试团队有提示意义：涉及 Mamba/Jamba 模型缓存模式 align 的测试均需显式开启 chunked prefill。
- 风险标记：测试配置变更，Mamba 约束耦合

关联脉络

- PR #51153 [Bugfix] Enable chunked prefill for qwen3.5-0.8B ppl test: 同为通过补 enable_chunked_prefill 修复 CI 测试失败的 bugfix，修复模式一致，可互相参考。