PR #47682 完整报告

vllm-project/vllm

[XPU] limit max-num-seqs in test_lmeval.py for XPU

合并时间：2026-07-06 13:34

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47682

执行摘要

- 一句话：限制 XPU 上 lmeval 测试 max-num-seqs
- 推荐动作：该 PR 价值较小，主要为解决特定硬件 CI 测试失败而设。无需精读，但体现了平台差异化的测试适配做法。

功能与动机

XPU 设备上默认的 max-num-seqs 会导致测试失败（OOM 或设备错误），因此需要限制并发序列数以适配 XPU 内存限制。

实现拆解

在 tests/entrypoints/openai/correctness/test_lmeval.py 的 test_lm_eval_accuracy_v1_engine 函数中，将条件判断从 if current_platform.is_tpu() 扩展为 if current_platform.is_tpu() or current_platform.is_xpu()，使得 XPU 平台也传入 --max-num-seqs 64 参数。该变更仅涉及条件分支，不影响其他平台。

关键文件：

- tests/entrypoints/openai/correctness/test_lmeval.py（模块 测试；类别 test；类型 test-coverage）：修改了测试函数中的平台条件判断，新增 XPU 支持。

关键符号：未识别

关键源码片段

tests/entrypoints/openai/correctness/test_lmeval.py

修改了测试函数中的平台条件判断，新增 XPU 支持。

```
# tests/entrypoints/openai/correctness/test_lmeval.py

def test_lm_eval_accuracy_v1_engine():
    """Run with the V1 Engine."""
    more_args = []

    # Limit compilation time for V1 on TPU
    # Avoid OOM on XPU
    if current_platform.is_tpu() or current_platform.is_xpu():
```

```
        more_args = ["--max-num-seqs", "64"]

    run_test(more_args)
```

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改测试条件判断，为 XPU 平台增加与 TPU 相同的限制。不会影响产品代码或非 XPU 测试行为。
- 影响：仅影响 Intel GPU（XPU）环境的 lmeval 测试，使之通过。其他平台无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR