PR #45914 完整报告

vllm-project/vllm

[Test] Pin block_size in auto-fit max_model_len test

合并时间：2026-06-24 22:56

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45914

执行摘要

- 一句话：修复 XPU 上 auto-fit 测试因 block_size 差异失败
- 推荐动作：该 PR 是一个典型跨平台兼容性修复，代码改动简单但说明了测试中隐含的硬件假设问题。值得关注的是，类似测试中若依赖默认值应该显式指定以确保可移植性。

功能与动机

PR body 指出：'The kv_cache_memory_bytes=1MB budget assumed CUDA's default block_size=16. On platforms whose default block_size is larger (e.g. XPU defaults to 64) a single KV block already exceeds 1MB, so auto-fit fails with 'not enough GPU memory available to serve even a single token'.' 该修复确保测试在不同硬件平台上均可通过。

实现拆解

1. 文件 tests/v1/e2e/general/test_context_length.py 中，函数 test_auto_fit_max_model_len_rejects_oversized_input 的 vllm_runner 调用新增参数 block_size=16。
2. 同时更新注释，说明为什么要固定 block_size：'Pin block_size=16 so the budget is independent of the platform's default block size.'
3. 未改动其他文件或逻辑，仅测试代码调整。

关键文件：

- tests/v1/e2e/general/test_context_length.py（模块 上下文长度；类别 test；类型 test-coverage）：测试文件，核心变更：在 vllm_runner 调用中增加 block_size=16 参数，并更新注释。

关键符号：test_auto_fit_max_model_len_rejects_oversized_input

关键源码片段

tests/v1/e2e/general/test_context_length.py

测试文件，核心变更：在 vllm_runner 调用中增加 block_size=16 参数，并更新注释。

```
# tests/v1/e2e/general/test_context_length.py
# 修改前：
```

```
# # Use a tiny KV cache budget to force auto-fit to a very small
# # max_model_len (e.g. ~16 tokens).
# kv_cache_bytes = 1_000_000 # 1 MB
# with vllm_runner(...) as vllm_model:
# 修改后：
    # Use a small KV cache budget to force auto-fit to a small
    # max_model_len. Pin block_size=16 so the budget is independent
    # of the platform's default block size.
    kv_cache_bytes = 1_000_000 # 1 MB

    with vllm_runner(
        model_name=model,
        max_model_len=-1,
        max_num_seqs=1,
        enforce_eager=True,
        block_size=16, # 显式指定，避免 XPU 默认 64 导致内存超限
        kv_cache_memory_bytes=kv_cache_bytes,
        load_format="dummy",
    ) as vllm_model:
```

评论区精华

无 review 评论，仅提交者 @Liangliang-Ma 在第一条提交中说明变更目的，随后由 @jikunshang 审核后合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅影响一个测试函数中的参数，且显式指定 block_size 为 16 与 CUDA 默认值一致，不会引入回归。XPU 等平台使用该参数后测试即可正常执行。
- 影响：影响范围仅限于测试 test_auto_fit_max_model_len_rejects_oversized_input 在 XPU 等非 CUDA 平台的可运行性。对其他测试或生产代码无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR