

PR #46241 完整报告

vllm-project/vllm

[CI][test] Replace InternVL2-1B with InternVL3-1B in test_pipeline_parallel.py

合并时间: 2026-06-21 15:11

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46241>

执行摘要

- 一句话: 替换因 transformers 5.x 兼容性失败的测试模型
- 推荐动作: 该 PR 是典型的 CI 环境适配, 值得关注的是其揭示了 transformers 5.x 对 tokenizer 加载策略的 breaking change, 未来引入新模型时需注意 tokenizer.json 的存在性。

功能与动机

OpenGVLab/InternVL2-1B (model_type=internvl_chat) 在 transformers >= 5.x 中无法通过 AutoTokenizer 加载, 报错 ValueError: Couldn't instantiate the backend tokenizer..., 原因是 transformers 5.x 将 internvl_chat 列入 MODELS_WITH_INCORRECT_HUB_TOKENIZER_CLASS, 强制使用 TokenizersBackend, 但该模型仓库缺少 tokenizer.json。替换为同系列且包含 tokenizer.json 的 InternVL3-1B 以恢复 CI 测试。

实现拆解

1. 在 tests/distributed/test_pipeline_parallel.py 的 MULTIMODAL_MODELS 字典中将模型标识符 "OpenGVLab/InternVL2-1B" 替换为 "OpenGVLab/InternVL3-1B", 对应第 178 行。
2. 在同一文件的 TEST_MODELS 列表中将相同模型标识符替换 (第 206 行)。
3. 保留 PPTestSettings.fast() 配置不变, 确保测试设置一致。
4. 仅修改测试文件, 无需其他配套变更。

关键文件:

- tests/distributed/test_pipeline_parallel.py (模块 PP 测试; 类别 test; 类型 test-coverage) : 测试文件中两处模型标识符替换, 是本次变更的唯一文件。

关键符号: 未识别

关键源码片段

tests/distributed/test_pipeline_parallel.py

测试文件中两处模型标识符替换, 是本次变更的唯一文件。

```
# tests/distributed/test_pipeline_parallel.py
# 替换前:
# "OpenGVLab/InternVL2-1B": PPTestSettings.fast(),
```

```
# 替换后:
"OpenGVLab/InternVL3-1B": PPTestSettings.fast(), # InternVL3-1B 包含 tokenizer.json, 兼容
transformers 5.x

# TEST_MODELS 列表中也做了相同替换:
# [MULTIMODAL GENERATION]
"OpenGVLab/InternVL3-1B",
```

评论区精华

无 review 评论。作者在 PR 描述中详细说明了错误根因：包括 transformers 5.x 的 `MODELS_WITH_INCORRECT_HUB_TOKENIZER_CLASS` 强制 `TokenizersBackend`，以及 InternVL2-1B 缺少 `tokenizer.json`。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。InternVL3-1B 与 InternVL2-1B 属于同一模型系列，架构和 `model_type` 相同，仅测试标识符替换，不影响任何生产逻辑。若 InternVL3-1B 未来也出现兼容性问题，则需再次替换。
- 影响：仅影响 CI 中 `pipeline parallel` 测试的目标模型，修复因 transformers 版本升级导致的 CI 失败，对系统性能、用户功能无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR