PR #23044 完整报告

sgl-project/sglang

[XPU] Fix DeepSeek-OCR tests under transformers 5.x

合并时间：2026-04-21 14:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23044

执行摘要

- 一句话：修复 XPU 平台 DeepSeek-OCR 测试在 transformers 5.x 下的导入错误。
- 推荐动作：该 PR 值得快速浏览，以了解 transformers 版本升级导致的兼容性问题及团队内的解决方案（通过 get_tokenizer 统一管理 tokenizer 加载）。关注点在于 sglang.srt.utils.hf_transformers.get_tokenizer 的设计，它封装了兼容性处理，可作为类似问题的标准做法。

功能与动机

PR body 指出，XPU DeepSeek-OCR 测试在 transformers 5.x 下失败，错误为 ImportError: cannot import name 'LlamaFlashAttention2' from 'transformers.models.llama.modeling_llama'。根本原因是 AutoTokenizer.from_pretrained(..., trust_remote_code=True) 会解析 DeepSeek-OCR 的远程代码链，而该链在 transformers 5.x 中引用了已移除的 LlamaFlashAttention2。NVIDIA 侧的相同测试已使用 trust_remote_code=False 避免此问题，但 XPU 测试未同步。此外，transformers 5.4+ 的检查还会因缺少 matplotlib 而提前失败，但同样源于远程代码解析。

实现拆解

1. 导入调整：在两个测试文件（test/srt/xpu/test_deepseek_ocr.py 和 test/srt/xpu/test_deepseek_ocr_triton.py）中，移除对 transformers.AutoTokenizer 的导入，改为从 sglang.srt.utils.hf_transformers 导入 get_tokenizer。
2. tokenizer 加载方式变更：将 cls.tokenizer = AutoTokenizer.from_pretrained(cls.model, use_fast=False, trust_remote_code=True) 替换为 cls.tokenizer = get_tokenizer(cls.model)。get_tokenizer 默认 trust_remote_code=False，避免解析远程代码链，同时其内部会触发兼容性补丁（如将 LlamaFlashAttention2 映射为 LlamaAttention），提供额外安全保障。
3. 参数清理：移除了 use_fast=False 参数，因为在 transformers 5.x 中 AutoTokenizer 总是返回快速 tokenizer，此参数已无作用，移除后代码意图更清晰。
4. 测试配套：本次变更仅涉及测试文件，没有修改源码主路径、配置或部署脚本，属于测试修复以恢复 CI 通过性。

关键文件：

- test/srt/xpu/test_deepseek_ocr.py（模块 XPU 测试；类别 test；类型 test-coverage）：这是主要的 DeepSeek-OCR 测试文件，变更修复了 tokenizer 加载错误，确保测试能在 transformers 5.x 下运行。
- test/srt/xpu/test_deepseek_ocr_triton.py（模块 Triton 测试；类别 test；类型 test-coverage）：这是 DeepSeek-OCR 的 Triton 后端测试文件，同样修复了 tokenizer 加载问题，继承自主测试类。

关键符号：get_tokenizer

关键源码片段

test/srt/xpu/test_deepseek_ocr.py

这是主要的 DeepSeek-OCR 测试文件，变更修复了 tokenizer 加载错误，确保测试能在 transformers 5.x 下运行。

```
from sglang.srt.utils.hf_transformers import get_tokenizer # 导入自定义的 tokenizer
加载工具，默认不信任远程代码

class TestDeepSeekOCR(CustomTestCase):
    @classmethod
    def setUpClass(cls):
        cls._cleanup_xpu_memory()
        cls.model = "deepseek-ai/DeepSeek-OCR"
        cls.tokenizer = get_tokenizer(cls.model) # 使用 get_tokenizer
        加载，避免触发远程代码解析链，从而绕过 LlamaFlashAttention2 导入错误
        cls.base_url = DEFAULT_URL_FOR_TEST
        # ... 其余测试初始化代码
```

评论区精华

Review 评论中没有实质性技术讨论，仅有两个批准（mingfeima 和 polisettyvarma），表明变更被快速接受。PRbody中作者详细分析了根因和解决方案，但未引发争议或深度权衡讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：
  - 回归风险：变更仅影响测试代码，不涉及生产逻辑，但需确保 get_tokenizer 在所有场景下行为与之前一致（特别是 trust_remote_code=False 时 tokenizer 类型是否兼容）。
  - 兼容性风险：如果其他测试或代码路径依赖 trust_remote_code=True 来加载特定 tokenizer，此变更可能不适用，但本 PR 仅针对 DeepSeek-OCR 测试，且 NVIDIA 侧测试已采用相同方式。
  - 性能风险：无，tokenizer 加载方式变更对运行时性能无影响。
- 影响：影响范围有限：
  - 对用户：无直接影响，此为内部测试修复。

- 对系统：恢复 XPU 平台 DeepSeek-OCR 测试的通过性，确保 CI 稳定性。
- 对团队：提供了在 transformers 5.x 下处理类似远程代码问题的参考模式（使用 get_tokenizer 避免信任远程代码）。
- 风险标记：测试覆盖调整，导入关系调整

关联脉络

- PR #21569 未知（PR body 中提及 #21569 导致问题）：PR body 提到测试在 #21569 后失败，可能是一个引入 transformers 5.x 升级或相关变更的 PR，但具体标题未在上下文中提供。