

PR #48276 完整报告

vllm-project/vllm

[Bugfix][Test] Register Qwen/Qwen3.5-4B example model

合并时间: 2026-07-11 05:01

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48276>

执行摘要

- 一句话: 修复 Qwen3.5-4B 模型注册缺失导致测试失败
- 推荐动作: 小型测试修复 PR, 无需精读。但值得关注的是 `_HfExamplesInfo` 的 `extras` 用法, 它提供了一种在不改变默认示例模型的情况下注册额外模型的方式, 可用于类似场景。

功能与动机

PR#48113 重定向了 `DFlashDraftModel` 的默认示例模型, 移除了 `tests/models/registry.py` 中对 `Qwen/Qwen3.5-4B` 的唯一引用, 导致 `tests/renderers/test_hf.py::test_resolve_content_format_hf_defined[Qwen/Qwen3.5-4B-openai]` 测试失败。该测试通过 `HF_EXAMPLE_MODELS.find_hf_info` 查找 `Qwen/Qwen3.5-4B`, 但注册表中已无该模型, 因此需要重新注册。

实现拆解

在 `tests/models/registry.py` 中, `Qwen3_5ForConditionalGeneration` 条目的 `_HfExamplesInfo` 构造函数中新增 `extras` 参数, 将 `Qwen/Qwen3.5-4B` 注册为额外模型。

1. 定位注册条目: 找到 `tests/models/registry.py` 中 `Qwen3_5ForConditionalGeneration` 对应的 `_HfExamplesInfo` 调用, 其默认模型为 `Qwen/Qwen3.5-0.8B`。
2. 添加 `extras` 参数: 在原有参数后添加 `extras={"4b": "Qwen/Qwen3.5-4B"}`, 使得 `find_hf_info("Qwen/Qwen3.5-4B")` 能够解析到此模型的注册信息。
3. 保持默认模型不变: 未修改 `Qwen/Qwen3.5-0.8B` 这一默认值, 保证现有兼容性。

关键文件:

- `tests/models/registry.py` (模块 测试注册表; 类别 `test`; 类型 `test-coverage`): 唯一变更文件; 在 `Qwen3_5ForConditionalGeneration` 注册项中添加 `extras` 参数, 注册 `Qwen/Qwen3.5-4B` 作为额外示例模型。

关键符号: 未识别

关键源码片段

`tests/models/registry.py`

唯一变更文件; 在 `Qwen3_5ForConditionalGeneration` 注册项中添加 `extras` 参数, 注册 `Qwen/Qwen3.5-4B` 作为额外示例模型。

```
# tests/models/registry.py 中 Qwen3_5ForConditionalGeneration 的示例模型注册条目
# 新增 extras 参数, 将 Qwen/Qwen3.5-4B 注册为一个可查找的额外模型变体
"Qwen3_5ForConditionalGeneration": _HfExamplesInfo(
    "Qwen/Qwen3.5-0.8B", # 默认示例模型保持不变
    extras={"4b": "Qwen/Qwen3.5-4B"}, # 新增: 额外注册 4B 变体
    max_model_len=4096,
),
```

评论区精华

无 review 评论。PR 由 robertgshaw2-redhat 直接批准合并。

- 暂无高价值评论线程

风险与影响

- 风险: 本 PR 仅添加了一个注册条目, 无任何逻辑变更, 风险极低。不会影响生产代码或已有功能。
- 影响:
 - 测试: 修复了 test_resolve_content_format_hf_defined 测试回归, 确保使用 Qwen/Qwen3.5-4B 的测试和工具能正常工作。
 - 范围: 仅影响测试基础设施中的模型注册表, 不影响运行时模型加载或推理。
 - 风险标记: 暂无

关联脉络

- PR #48113 [Bugfix][Spec Decode] Fix DFlash draft/target layer-count mismatch: 本 PR 修复了因 PR#48113 重定向 DFlashDraftModel 默认示例模型而导致的测试失败。