PR #42498 完整报告

vllm-project/vllm

[CI] Re-enable Nemotron Parse parity test and switch testing to nemotron-parse v1.2

合并时间：2026-05-13 21:05

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42498

执行摘要

- 一句话：重新启用 Nemotron Parse 测试并切换到 v1.2
- 推荐动作：该 PR 值得快速合并，因为它解锁了 Nemotron Parse 模型的 CI 测试，无需深入阅读源码细节。关注点在于测试工具方法的扩展和模型版本切换的潜在影响。

功能与动机

根据 PR 描述，目的是重新启用 Nemotron Parse 模型的测试，因为 v1.2 版本包含了修复 vLLM 兼容性的更改。同时修复测试中存在的两个 bug：PROMPT 显式携带 decoder control prefix，但 HfRunner.get_inputs 会通过 processor 默认添加特殊 token，导致双重包装；以及 use_cache=False 是 v1.1 时代的 HF crash 工作区，v1.2 不再需要。

实现拆解

1. tests/models/multimodal/generation/test_nemotron_parse.py：更新 PROMPT 常量，添加 <predict_no_text_in_pic> token；移除 @pytest.mark.skip 装饰器，将测试模型从 nvidia/NVIDIA-Nemotron-Parse-v1.1 改为 nvidia/NVIDIA-Nemotron-Parse-v1.2；在 hf_runner.generate_greedy_logprobs_limit 调用中将 use_cache=False 替换为 tokenization_kwargs={"add_special_tokens": False}，防止 tokenizer 自动添加特殊 token 导致 prompt 重复包装。
2. tests/conftest.py：在 HfRunner.generate_greedy_logprobs_limit 方法中新增 tokenization_kwargs 参数，并将其透传给 self.get_inputs，支持测试用例控制 tokenizer 行为。
3. tests/models/registry.py：将 NemotronParseForConditionalGeneration 的示例模型从 nvidia/NVIDIA-Nemotron-Parse-v1.1 更新为 nvidia/NVIDIA-Nemotron-Parse-v1.2，对齐测试配置。

关键文件：

- tests/models/multimodal/generation/test_nemotron_parse.py（模块 测试；类别 test；类型 test-coverage）：核心变更文件：移除 skip 装饰器、更新模型版本、修复 prompt 重复包装 bug、移除了 use_cache=False 工作区。
- tests/conftest.py（模块 测试；类别 test；类型 test-coverage）：扩展 HfRunner.generate_greedy_logprobs_limit 方法，新增 tokenization_kwargs 参数并透传给 get_inputs，是支持测试修复的基础设施更改。

- tests/models/registry.py（模块 测试；类别 test；类型 test-coverage）：更新 NemotronParseForConditionalGeneration 的示例模型版本，与测试配置保持一致。

关键符号：run_test, test_models, generate_greedy_logprobs_limit, mask_bbox_tokens

关键源码片段

tests/models/multimodal/generation/test_nemotron_parse.py

核心变更文件：移除 skip 装饰器、更新模型版本、修复 prompt 重复包装 bug、移除了 use_cache=False 工作区。

```
# tests/models/multimodal/generation/test_nemotron_parse.py
# 更新后的 PROMPT 包含 v1.2 新增的 <predict_no_text_in_pic> token
PROMPT = (
    "</s><s><predict_bbox><predict_classes><output_markdown><predict_no_text_in_pic>"
)

# 在 hf_runner 调用中传递 tokenization_kwargs 以禁用自动添加特殊 token
# 避免 PROMPT 中已包含的 decoder control prefix 被重复包装
hf_model.generate_greedy_logprobs_limit(
    prompts,
    max_tokens,
    num_logprobs=num_logprobs,
    images=images,
    tokenization_kwargs={"add_special_tokens": False}, # 替代原来的 use_cache=False
)

# 移除 @pytest.mark.skip 并切换到 v1.2 模型
# @pytest.mark.skip(
# reason="Model's custom MBart decoder has head count mismatch ..."
# )
# @pytest.mark.parametrize("model", ["nvidia/NVIDIA-Nemotron-Parse-v1.1"])
@pytest.mark.parametrize("model", ["nvidia/NVIDIA-Nemotron-Parse-v1.2"])
@pytest.mark.parametrize("dtype", ["bfloat16"])
@pytest.mark.parametrize("num_logprobs", [5])
def test_models(hf_runner, vllm_runner, model, dtype, num_logprobs):
    ...
```

tests/conftest.py

扩展 HfRunner.generate_greedy_logprobs_limit 方法，新增 tokenization_kwargs 参数并透传给 get_inputs，是支持测试修复的基础设施更改。

```
# tests/conftest.py - HfRunner 类中的 generate_greedy_logprobs_limit 方法
# 新增 tokenization_kwargs 参数，支持自定义 tokenizer 行为
def generate_greedy_logprobs_limit(
    self,
    prompts: list[str],
    max_tokens: int,
    num_logprobs: int | None,
```

```
    images: PromptImageInput | None = None,
    audios: PromptAudioInput | None = None,
    videos: PromptVideoInput | None = None,
    use_cache: bool = True,
    tokenization_kwargs: dict[str, Any] | None = None, # 新增参数，默认 None 保持兼容
    **kwargs: Any,
) -> list[TokensTextLogprobs]:
    all_inputs = self.get_inputs(
        prompts,
        images=images,
        videos=videos,
        audios=audios,
        tokenization_kwargs=tokenization_kwargs, # 透传给 get_inputs
    )
    ...
```

评论区精华

该 PR 没有实质性的 review 讨论。只有 bot 评论（Claude Code Review 提示来自 fork，Gemini Code Assist 评论无反馈）和维护者 DarkLight1337 的简单批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅限于测试代码和测试基础设施，不涉及 vLLM 核心逻辑或推理路径。潜在风险包括：v1.2 模型可能引入新的行为差异导致测试失败（但 PR 作者已确认本地测试通过）；tokenization_kwargs 的引入可能影响其他使用该方法的测试，但由于默认值为 None，行为保持向后兼容。
- 影响：直接影响：重新启用 Nemotron Parse v1.2 模型的端到端测试，确保该模型在 vLLM 中的推理结果与 Hugging Face 保持一致。间接影响：扩展了 HfRunner 的通用测试工具方法，可供其他需要自定义 tokenizer 行为的测试复用。对用户无直接影响。
- 风险标记：测试覆盖提升，基础设施变更

关联脉络

- PR #38740 [Transformers v5] NemotronParseForConditionalGeneration: 关联 Issue，描述了原测试失败的问题，本 PR 旨在修复该 Issue。
- PR #38748 PR that makes vLLM compatible with v1.1's broken-on-Transformers-v5 config: PR body 中提及该 PR 修改了模型文件实现兼容性，但未重新启用测试，本 PR 在此基础上补全了测试。