

PR #48676 完整报告

vllm-project/vllm

[ROCm][CI] fix test_common.py

合并时间: 2026-07-15 14:42

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48676>

执行摘要

- 一句话: 修复 HfRunner 因缺失 tokenizer 参数导致的崩溃
- 推荐动作: 值得合并的紧急修复。虽仅影响测试, 但属于上游依赖变更导致的兼容性问题。可精读 HfRunner.generate 的改动以了解测试工具兼容性处理模式。

功能与动机

PR body 指出: 较新版本的 Transformers 在使用 stop strings 时要求向 `model.generate()` 传入 `tokenizer` 参数。HyperCLOVAX-SEED-Think-14B模型的`generation_config.json`包含"`stop_strings`": [`<lendofturnl>`", `<lstopl>`"], 导致测试中调用 `self.model.generate()` 时触发 `ValueError`。

实现拆解

1. HfRunner 生成路径修复 (`tests/conftest.py`): 在 `generate()`、`generate_greedy_logprobs()`、`generate_greedy_logprobs_limit()` 三个方法中, 将 `**kwargs` 替换为 `**generate_kwargs`, 其中 `generate_kwargs` 通过 `setdefault("tokenizer", self.tokenizer)` 确保默认传入 `tokenizer`, 同时允许调用方显式覆盖。
2. 模型 forward 微调 (`vllm/model_executor/models/hyperclovax.py`): 将 `forward` 方法中 `intermediate_tensors` 参数从必选改为可选 (`= None`), 与基类或其他模型的签名保持一致, 避免签名不匹配导致的调用错误。

关键文件:

- `tests/conftest.py` (模块 测试工具; 类别 `test`; 类型 `test-coverage`; 符号 `HfRunner.generate`, `HfRunner.generate_greedy_logprobs`, `HfRunner.generate_greedy_logprobs_limit`): 核心修复文件: 在 HfRunner 的三个 `generate` 路径中添加默认 `tokenizer` 参数, 消除因 Transformers 新版本要求导致的崩溃。
- `vllm/model_executor/models/hyperclovax.py` (模块 模型实现; 类别 `source`; 类型 `data-contract`; 符号 `HyperCLOVAXForCausalLM.forward`): 次要修复: 调整 `forward` 方法中 `intermediate_tensors` 参数的默认值, 保持与基类签名兼容, 但未直接关联测试崩溃。

关键符号: `HfRunner.generate`, `HfRunner.generate_greedy_logprobs`, `HfRunner.generate_greedy_logprobs_limit`

关键源码片段

tests/conftest.py

核心修复文件：在 HfRunner 的三个 generate 路径中添加默认 tokenizer 参数，消除因 Transformers 新版本要求导致的崩溃。

```
# tests/conftest.py (HfRunner.generate 方法片段)
def generate(
    self,
    prompts: list[str] | list[list[int]],
    images: PromptImageInput | None = None,
    videos: PromptVideoInput | None = None,
    audios: PromptAudioInput | None = None,
    **kwargs: Any,
) -> list[tuple[list[list[int]], list[str]]]:
    all_inputs = self.get_inputs(
        prompts, images=images, videos=videos, audios=audios
    )

    outputs: list[tuple[list[list[int]], list[str]]] = []
    for inputs in all_inputs:
        # 复制原始 kwargs 并确保默认传入 tokenizer
        # 这是为了避免新版本 Transformers 在 generate() 使用 stop strings 时
        # 因缺少 tokenizer 参数而引发 ValueError。
        # 调用方可以通过显式传递 tokenizer=... 来覆盖默认值。
        generate_kwargs = dict(kwargs)
        generate_kwargs.setdefault("tokenizer", self.tokenizer)
        output_ids: torch.Tensor = self.model.generate(
            **self.wrap_device(inputs),
            use_cache=True,
            **generate_kwargs,
        )
        ...
```

评论区精华

无实质性 review 讨论。Claude 机器人自动评论指出该 PR 来自 fork 未触发审查；AndreasKaratzas 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。核心逻辑仅通过 setdefault 为 test runner 添加默认参数，不改变现有行为。模型 forward 签名微调属于常规兼容性调整。但需确认所有依赖 intermediate_tensors 为必选参数的调用方是否已适配。
- 影响：影响范围限于测试基础设施：修复了因 Transformers API 变更导致的 ROCm CI 测试失败。开发者和 CI 流水线可恢复正常测试运行。无用户功能影响。

- 风险标记：暂无

关联脉络

- 暂无明显关联 PR