

PR #47135 完整报告

vllm-project/vllm

[Bench][BugFix] Fix empty decoder prompt for Cohere ASR in throughput benchmark

合并时间: 2026-07-01 23:42

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47135>

执行摘要

- 一句话: 修复 Cohere ASR 基准测试空 decoder prompt bug
- 推荐动作: 建议后续改进: 1) 使用 MRO 检查替代类名字符串比较, 增强鲁棒性。 2) 考虑将 decoder prompt 逻辑抽取为 tokenizer 的接口方法, 避免在 benchmark 代码中硬编码。

功能与动机

修复 ASRDataset.sample() 对 Cohere ASR 模型生成空 decoder prompt 的问题, 使 throughput benchmark 能够正常运行 Cohere ASR 模型。

实现拆解

1. 修改 ASRDataset.sample() 方法 (vllm/benchmarks/datasets/datasets.py): 在原有的 OpenAI/Whisper 检测分支后增加 CohereAsrTokenizer 检测逻辑。 - 优先使用 tokenizer 类名 CohereAsrTokenizer 判断 (主要路径)。 - 若类名不匹配, 检查 name_or_path 是否包含 "cohere" (本地 checkpoint 回退)。 - 为 Cohere ASR 提供完整的 decoder prompt 控制 token 序列。
2. 新增测试用例 (tests/benchmarks/test_audio_dataset.py):
 - test_asr_dataset_cohere_class_name_gets_decoder_prompt: 验证类名检测路径。
 - test_asr_dataset_cohere_name_or_path_fallback_gets_decoder_prompt: 验证 name_or_path 回退路径。
 - test_asr_dataset_unknown_tokenizer_gets_empty_prompt: 验证未知 tokenizer 行为不变。
 - 新增 CohereAsrTokenizer 和 _CohereNameOnlyTokenizer 测试替身类。

关键文件:

- vllm/benchmarks/datasets/datasets.py (模块 基准测试; 类别 source; 类型 core-logic; 符号 sample): 核心修复文件, 修改 ASRDataset.sample() 增加 Cohere ASR decoder prompt 检测逻辑
- tests/benchmarks/test_audio_dataset.py (模块 测试; 类别 test; 类型 test-coverage; 符号 CohereAsrTokenizer, _CohereNameOnlyTokenizer, _make_asr_dataset, test_asr_dataset_cohere_class_name_gets_decoder_prompt): 测试覆盖, 验证核心修复逻辑的正确性

关键符号: ASRDataset.sample

关键源码片段

[tests/benchmarks/test_audio_dataset.py](#)

测试覆盖, 验证核心修复逻辑的正确性

为 Cohere ASR 模型构造正确的 decoder prompt 测试

```
class CohereAsrTokenizer(_Tokenizer):
```

```
    # 模拟 Cohere ASR 官方 tokenizer, 通过类名检测主路径
```

```
    def __init__(self, name_or_path: str = "/models/cohere-transcribe") -> None:
        super().__init__(name_or_path)
```

```
class _CohereNameOnlyTokenizer(_Tokenizer):
```

```
    # 模拟用户自定义路径的 Cohere tokenizer, 用于 name_or_path 回退路径
```

```
    def __init__(self) -> None:
        super().__init__("cohere/some-local-checkpoint")
```

```
_COHERE_ASR_PROMPT = (
```

```
    "<lstartofcontextl><lstartoftranscriptl>"
```

```
    "<lemo:undefinedl><lendl><lendl><lpcnl><lnoitnl>"
```

```
    "<lnotimestampl><lnodiarizel>"
```

```
)
```

```
def test_asr_dataset_cohere_class_name_gets_decoder_prompt(tmp_path: Path) -> None:
```

```
    # 验证通过类名检测 CohereAsrTokenizer 时, 能获得正确的 decoder prompt
```

```
    dataset = _make_asr_dataset(tmp_path)
```

```
    samples = dataset.sample(
```

```
        tokenizer=CohereAsrTokenizer(),
```

```
        num_requests=1,
```

```
        output_len=32,
```

```
        asr_min_audio_len_sec=0.0,
```

```
        asr_max_audio_len_sec=1.0,
```

```
    )
```

```
    assert len(samples) == 1
```

```
    assert samples[0].prompt == _COHERE_ASR_PROMPT
```

```
def test_asr_dataset_cohere_name_or_path_fallback_gets_decoder_prompt(tmp_path: Path) ->
None:
```

```
    # 验证当 tokenizer 类名非标准但 name_or_path 含 "cohere" 时, 也能获得正确 prompt
```

```
    dataset = _make_asr_dataset(tmp_path)
```

```
    samples = dataset.sample(
```

```
        tokenizer=_CohereNameOnlyTokenizer(),
```

```
        num_requests=1,
```

```

        output_len=32,
        asr_min_audio_len_sec=0.0,
        asr_max_audio_len_sec=1.0,
    )
    assert len(samples) == 1
    assert samples[0].prompt == _COHERE_ASR_PROMPT

def test_asr_dataset_unknown_tokenizer_gets_empty_prompt(tmp_path: Path) -> None:
    # 验证未知 tokenizer 保持原来的空字符串行为，确保不回归
    dataset = _make_asr_dataset(tmp_path)
    samples = dataset.sample(
        tokenizer=_Tokenizer(name_or_path="some-other/asr-model"),
        num_requests=1,
        output_len=32,
        asr_min_audio_len_sec=0.0,
        asr_max_audio_len_sec=1.0,
    )
    assert len(samples) == 1
    assert samples[0].prompt == ""

```

评论区精华

claude[bot] 在 review 中指出：类名检测 `tok_class == "CohereAsrTokenizer"` 在生产环境中不可靠，因为 tokenizer 加载后类名会被重写为 `CachedCohereAsrTokenizer`，实际只有 `"cohere" in name_or_path.lower()` 回退路径生效。建议改用 MRO 检查（如 `isinstance(tokenizer, CohereAsrTokenizer)`）。该评论未被作者回应，且修复已合并。

- 类名检测在生产环境可能失效 (correctness): 作者未回应，PR 已合并。类名检测在测试环境中通过，但生产环境中可能失效。

风险与影响

- 风险：
 1. 类名检测失效风险：如 claude[bot] 指出，生产环境 tokenizer 类名被缓存包装，`CohereAsrTokenizer` 直接比较可能无法匹配，导致退化为 `prompt=""`。此风险在本地重命名 checkpoint 时更明显。
 2. 回归风险：新增分支仅影响 Cohere ASR 检测，不影响其他 tokenizer 路径。测试覆盖了未知 tokenizer 的行为不变性。- 影响：影响限于 throughput benchmark 中 Cohere ASR 模型的使用场景，修复后用户可正常对 Cohere ASR 模型执行基准测试。对其他模型无影响。- 风险标记：类名检测在生产环境可能失效

关联脉络

- PR #39259 [Bugfix] Cohere ASR librosa import error: 同为 Cohere ASR 相关的 bugfix，但故障模式不同（librosa 导入错误 vs 空 decoder prompt）