

# PR #41478 完整报告

vllm-project/vllm

Limit concurrency on `test\_transcription\_api\_correctness.py`

合并时间: 2026-05-02 11:19

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/41478>

## 执行摘要

- 一句话: 限制转录测试并发数防止 OOM
- 推荐动作: 值得合并, 修复 CI 稳定性。也可考虑从环境变量或配置文件中读取该值以提升灵活性。

## 功能与动机

CohereASR 测试在 CI 中因内存压力崩溃, 根源是 `max_num_seqs` 未限制, 导致大量并发音频编码请求耗尽 GPU 内存。PR body 指出: "The test used effectively unbounded concurrency for this dataset. The engine then tried to process a large batch of concurrent audio encoder requests."

## 实现拆解

1. 在文件顶部定义模块级常量 `MAX_SEQS_FOR_TRANSCRIPTION_TEST = 32`, 明确其用途为防止 18GB GPU 上的 OOM。
2. 在 `test_wer_correctness` 函数的 `server_args` 列表中添加 `f"--max_num_seqs={MAX_SEQS_FOR_TRANSCRIPTION_TEST}"`, 将此参数传递给 vLLM 引擎。
3. 该变更仅影响 `tests/entrypoints/openai/correctness/test_transcription_api_correctness.py` 一个文件, 不涉及其他测试或源码模块。

关键文件:

- `tests/entrypoints/openai/correctness/test_transcription_api_correctness.py` (模块 测试脚本; 类别 test; 类型 test-coverage): 唯一变更文件。添加 `MAX_SEQS_FOR_TRANSCRIPTION_TEST` 常量并在 `server_args` 中使用 `--max_num_seqs` 限制并发序列数。

关键符号: `test_wer_correctness`

## 关键源码片段

`tests/entrypoints/openai/correctness/test_transcription_api_correctness.py`

唯一变更文件。添加 `MAX_SEQS_FOR_TRANSCRIPTION_TEST` 常量并在 `server_args` 中使用 `--max_num_seqs` 限制并发序列数。

```
# 文件顶部增加模块级常量，明确用途
# Tuned to prevent OOM on 18GB GPUs in transcription correctness tests.
MAX_SEQS_FOR_TRANSCRIPTION_TEST = 32

def test_wer_correctness(...):
    server_args = [
        "--enforce-eager",
        f"--tokenizer_mode={model_info.tokenizer_mode}",
        f"--max_num_seqs={MAX_SEQS_FOR_TRANSCRIPTION_TEST}", # 限制并发，避免 OOM
    ]
```

## 评论区精华

gemini-code-assist[bot] 提出应将硬编码值改为模块级常量，以提升可维护性和不同硬件配置下的健壮性。作者 ekagra-ranjan 回应 "done" 并采纳建议。合入者 DarkLight1337 批准了该 PR。

- 硬编码值应改为模块级常量 (design): 作者采纳建议，修改为常量。

## 风险与影响

- 风险：无显著风险。仅限制测试并发数，不影响功能正确性。如果未来 CI 更新更大显存的 GPU，可能需要调高此值，但当前值足够覆盖 18GB MIG 场景。
- 影响：仅影响 test\_transcription\_api\_correctness.py 测试用例。可消除 CI 中因 OOM 导致的不稳定失败，但对系统其他部分无影响。
- 风险标记：测试依赖硬件显存

## 关联脉络

- PR #40582 Add CohereASR to transcription correctness test: 该 PR 引入了 CohereASR 测试，导致后续 OOM 问题。