PR #38492 完整报告

vllm-project/vllm

[CI] Add temperature=0.0, reduce max_tokens, and add debug prints to audio_in_video tests

合并时间：2026-03-30 13:36

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38492

执行摘要

该 PR 通过为音频视频测试添加确定性温度设置、减少最大令牌数并添加调试打印，修复了测试中的非确定性行为，提高了测试稳定性和可调试性。这是一个低风险的测试维护变更。

功能与动机

动机是解决音频视频测试中的 sporadic 失败问题。根据 PR body，测试未设置温度或种子，导致输出长度非确定性和 finish_reason 不一致。设置 temperature=0.0 使生成确定性，减少 max_tokens 确保模型在完成前达到令牌限制，添加调试打印辅助诊断。

实现拆解

实现集中在文件 tests/entrypoints/openai/chat_completion/test_audio_in_video.py：

- 函数 test_online_audio_in_video 和 test_online_audio_in_video_multi_videos 中，在 client.chat.completions.create 调用添加 temperature=0.0 和 max_tokens=8。
- 将循环变量改为 turn 并添加打印语句：print(f"[DEBUG][single-video] turn={turn} finish_reason={choice.finish_reason!r} content={choice.message.content!r} usage={chat_completion.usage}")。

评论区精华

review 中，gemini-code-assist[bot] 指出初始提交遗漏了 max_tokens 从 16 到 8 的修改，作者 AndreasKaratzas 迅速回复并修复。讨论焦点是正确性，确保 PR 描述与实际代码一致。

风险与影响

风险极低：仅修改测试文件，不影响生产代码。调试打印可能增加输出噪声，但限于测试环境。影响限于测试套件，提高音频视频测试的可靠性，减少 flaky 测试，对用户无直接影响。

关联脉络

与历史 PR #38414 类似，都旨在修复 flaky 测试。这表明团队持续关注测试稳定性，特别是在多模态和 CI 环境中。