

# PR #53750 完整报告

vllm-project/vllm

[Bugfix][Frontend] Apply the stop string limit to Cohere requests

合并时间: 2026-08-26 00:41

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/53750>

## 执行摘要

- 一句话: Cohere 端点应用 stop 字符串数量上限
- 推荐动作: 值得快速浏览, 因为它展示了如何在协议层通过 Pydantic 注解统一限制, 是小型安全修复的样例。关注点在于保持与其他端点的一致性。

## 功能与动机

PR 描述指出, VLLM\_MAX\_STOP\_STRINGS 用于限制单个请求携带的 stop 字符串数量, 以保护采样器免受无界工作负载的影响, 但 Cohere Chat V2 端点未应用该限制。对比显示, Anthropic 和 OpenAI 端点均通过 `Annotated[list[str], Field(max_length=...)]` 包装字段, 而 Cohere 的 `stop_sequences` 是普通 `list[str]`, 导致协议层成为唯一缺失的防线。默认限制为 4 时, Cohere 端点会接受 1000 个 stop 序列, 而 Anthropic 会拒绝。

## 实现拆解

1. 修改协议模型: 在 `vllm/entrypoints/cohere/protocol.py` 中, 将 `CohereChatV2Request.stop_sequences` 的类型从 `list[str] | None` 改为 `Annotated[list[str], Field(max_length=envs.VLLM_MAX_STOP_STRINGS)] | None`, 并导入 `Annotated` 和 `envs`。这样, Pydantic 会在请求验证时自动检查列表长度, 超出限制时抛出 `ValidationError`。
2. 新增测试: 在 `tests/entrypoints/cohere/test_protocol.py` 中添加 `test_stop_sequences_bounded_by_env_limit`, 测试验证恰好等于限制数量的 stop 序列可通过, 而超过限制的会触发 `ValidationError`, 且测试从 `envs` 读取限制值以保证与默认值解耦。
3. 运行检查: PR 作者运行了 `pytest tests/entrypoints/cohere/test_protocol.py` 和 `ruff`, 报告协议测试全部通过, 新增测试通过, 其他失败与修改前一致 (由于环境缺少 `pytest_asyncio` 和 `uvicorn`)。

关键文件:

- `vllm/entrypoints/cohere/protocol.py` (模块 Cohere 端点; 类别 source; 类型 dependency-wiring; 符号 `CohereChatV2Request`): 核心变更文件, 为 `stop_sequences` 添加长度限制注解。
- `tests/entrypoints/cohere/test_protocol.py` (模块 Cohere 端点; 类别 test; 类型 test-coverage; 符号 `test_stop_sequences_bounded_by_env_limit`): 新增测试验证

stop 序列上限行为。

关键符号: test\_stop\_sequences\_bounded\_by\_env\_limit

## 关键源码片段

### vllm/entrypoints/cohere/protocol.py

核心变更文件, 为 stop\_sequences 添加长度限制注解。

```
# vllm/entrypoints/cohere/protocol.py
from typing import Annotated
import vllm.envs as envs

class CohereChatV2Request(BaseModel):
    # ... 其他字段 ...

    # 限制 stop 序列数量, 与 Anthropic 和 OpenAI 端点保持一致
    stop_sequences: (
        Annotated[list[str], Field(max_length=envs.VLLM_MAX_STOP_STRINGS)] | None
    ) = None
```

### tests/entrypoints/cohere/test\_protocol.py

新增测试验证 stop 序列上限行为。

```
# tests/entrypoints/cohere/test_protocol.py
import vllm.envs as envs

def test_stop_sequences_bounded_by_env_limit(self):
    messages = [{"role": "user", "content": "hi"}]
    limit = envs.VLLM_MAX_STOP_STRINGS

    # 恰好等于限制数量的 stop 序列应通过验证
    req = CohereChatV2Request(
        model="m",
        messages=messages,
        stop_sequences=[f"s{i}" for i in range(limit)],
    )
    assert len(req.stop_sequences) == limit

    # 超过限制则抛 ValidationError
    with pytest.raises(ValidationError):
        CohereChatV2Request(
            model="m",
            messages=messages,
            stop_sequences=[f"s{i}" for i in range(limit + 1)],
        )
```

## 评论区精华

review过程中，Claudebot自动评论指出由于是forkPR，自动评审被禁用，需要维护者触发。维护者 DarkLight1337 评论 '/ci run' 触发 CI，并最终批准并合并。没有其他实质性的讨论。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险较低。变更仅影响 Cohere 端点的请求验证，可能拒绝原本应被拒绝的请求，但不会改变已接受请求的行为。需要确认 VLLM\_MAX\_STOP\_STRINGS 的默认值对用户影响，但该限制已在其他端点实施，因此此处对齐符合预期。
- 影响：影响范围限定在 Cohere Chat V2 端点，提升一致性和健壮性，防止资源滥用。对现有用户，如果之前提交超过限制的 stop 序列，现在会被拒绝，但这是有意为之的行为变更。
- 风险标记：协议变更可能影响现有请求，缺少端点集成测试

## 关联脉络

- PR #51447 Add VLLM\_MAX\_STOP\_STRINGS limit: 引入了该限制，但未覆盖 Cohere 端点，本 PR 补全此遗漏。