

PR #45640 完整报告

vllm-project/vllm

[Misc] Added validation for Cohere /v2/embed input field exclusivity

合并时间: 2026-06-16 13:42

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45640>

执行摘要

- 一句话: Cohere /v2/embed 互斥输入字段校验
- 推荐动作: 值得合并, 设计简洁, 测试覆盖全面。推荐作为防御性编程的示例参考。

功能与动机

Cohere API 文档将 `texts`、`images`、`inputs` 描述为互斥字段, 但此前未进行强制校验, 导致多个字段同时发送时空字段无法被正确拒绝, 行为不明确。本 PR 旨在提前拒绝无效请求并给出清晰错误信息。

实现拆解

1. 在 `vllm/entrypoints/pooling/embed/protocol.py` 的 `CohereEmbedRequest` 类中新增 `@model_validator(mode="after")` 装饰的 `validate_input_fields` 方法。该方法检查 `texts`、`images`、`inputs` 三个字段, 统计非 `None` 且非空的字段数量, 若不等于 1 则抛出 `ValueError`。
2. 在 `tests/entrypoints/pooling/embed/test_io_processor.py` 中新增 `TestCohereEmbedRequestParsing` 测试类, 包含参数化测试:
 - `test_rejects_invalid_input_field_combinations`: 覆盖无字段、多个字段、空列表等共 7 种无效场景, 断言抛出 `ValidationError`。
 - `test_accepts_exactly_one_non_empty_input_field`: 覆盖每个字段单独提供且非空的 3 种合法场景, 验证请求构造成功。

关键文件:

- `vllm/entrypoints/pooling/embed/protocol.py` (模块 协议定义; 类别 `source`; 类型 `core-logic`; 符号 `validate_input_fields`): 核心校验逻辑所在文件, 新增 `validate_input_fields` 方法。
- `tests/entrypoints/pooling/embed/test_io_processor.py` (模块 IO 处理器; 类别 `test`; 类型 `test-coverage`; 符号 `TestCohereEmbedRequestParsing`, `test_rejects_invalid_input_field_combinations`, `test_accepts_exactly_one_non_empty_input_field`): 测试覆盖合法与非法输入组合, 确保校验器正确工作。

关键符号: `validate_input_fields`

关键源码片段

vllm/entrypoints/pooling/embed/protocol.py

核心校验逻辑所在文件，新增 `validate_input_fields` 方法。

```
# vllm/entrypoints/pooling/embed/protocol.py
class CohereEmbedRequest(BaseModel):
    model: str | None = None
    input_type: str | None = None
    texts: list[str] | None = None
    images: list[str] | None = None
    inputs: list[CohereEmbedInput] | None = None
    output_dimension: int | None = None
    embedding_types: list[CohereEmbeddingType] | None = None
    truncate: CohereTruncate = "END"
    max_tokens: int | None = None
    priority: int = 0

# PR #45640: 在模型实例化后 validator 阶段校验字段互斥性
@model_validator(mode="after")
def validate_input_fields(self):
    # 三个输入字段: texts, images, inputs
    input_fields = (self.texts, self.images, self.inputs)
    # 筛选出非 None 的字段
    provided_fields = [field for field in input_fields if field is not None]
    # 必须恰好有一个非空字段
    if len(provided_fields) != 1 or not provided_fields[0]:
        raise ValueError(
            "Exactly one of texts, images, or inputs must be provided, "
            "and it must be non-empty"
        )
    return self
```

tests/entrypoints/pooling/embed/test_io_processor.py

测试覆盖合法与非法输入组合，确保校验器正确工作。

```
# tests/entrypoints/pooling/embed/test_io_processor.py
class TestCohereEmbedRequestParsing:
    """Unit tests for Cohere embed request parsing."""

    @pytest.mark.parametrize(
        "request_body",
        [
            {"model": "test"}, # 无字段
            {"model": "test", "texts": ["hello"], "images": ["image-uri"]}, # 两个字段
            {"model": "test", "texts": ["hello"],
             "inputs": [{"content": [{"type": "text", "text": "hello"}]}]},
            {"model": "test", "images": ["image-uri"],
             "inputs": [{"content": [{"type": "text", "text": "hello"}]}]},
        ],
    )
```

```

        {"model": "test", "texts": []}, # 空列表
        {"model": "test", "images": []},
        {"model": "test", "inputs": []},
    ],
)
def test_rejects_invalid_input_field_combinations(self, request_body):
    with pytest.raises(
        ValidationError,
        match="Exactly one of texts, images, or inputs must be provided",
    ):
        CohereEmbedRequest(**request_body)

@pytest.mark.parametrize(
    "request_body",
    [
        {"model": "test", "texts": ["hello"]},
        {"model": "test", "images": ["image-uri"]},
        {"model": "test",
         "inputs": [{"content": [{"type": "text", "text": "hello"}]}]},
    ],
)
def test_accepts_exactly_one_non_empty_input_field(self, request_body):
    request = CohereEmbedRequest(**request_body)
    assert request.model == "test"

```

评论区精华

无 review 评论；PR 获得 nooooop 的 approve 且无讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅增加 Pydantic 校验器，未修改现有逻辑或接口约定。可能的风险是校验器使用 mode="after"，若其他字段 validator 有依赖需注意顺序，但当前字段无交叉依赖。
- 影响：对用户：之前发送多个字段或空字段的请求现在会被拒绝，得到 422 错误及明确消息，提升 API 体验。对系统：校验在反序列化阶段完成，不增加运行时开销。对团队：低影响、高收益的防御性编程。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR