

PR #45491 完整报告

vllm-project/vllm

Treat null completion max_tokens like the default

合并时间: 2026-06-13 17:34

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45491>

执行摘要

- 一句话: 修复 null max_tokens 导致 API 500 错误
- 推荐动作: 本 PR 是典型的边界情况修复, 逻辑清晰且改动量小, 适合快速合入。

功能与动机

修复 Schemathesis 模糊测试在 OpenAI API 集成测试中发现的失败用例: 请求体 `{"max_tokens": null, "prompt": [0]}` 会触发服务端 500 错误, 因为 `CompletionRequest.max_tokens` 在验证后仍为 `None`, 而后续采样参数构造代码断言其不为 `None`。

实现拆解

在 `CompletionRequest` 类中新增一个 Pydantic `model_validator(mode="before")` 方法 `normalize_null_max_tokens`, 在字段验证之前将 `max_tokens` 的 `None` 值替换为字段定义的默认值。具体做法是: 如果 `data` 是字典且 `data.get("max_tokens") is None`, 则复制字典并设置 `data["max_tokens"] = cls.model_fields["max_tokens"].default`。

关键文件:

- `vllm/entrypoints/openai/completion/protocol.py` (模块 API 入口; 类别 source; 类型 core-logic; 符号 `normalize_null_max_tokens`): 新增 `normalize_null_max_tokens` validator, 将以 `None` 形式传入的 `max_tokens` 统一替换为字段默认值, 修复 Null 导致的 500 错误。

关键符号: `normalize_null_max_tokens`

关键源码片段

`vllm/entrypoints/openai/completion/protocol.py`

新增 `normalize_null_max_tokens` validator, 将以 `None` 形式传入的 `max_tokens` 统一替换为字段默认值, 修复 Null 导致的 500 错误。

```
@model_validator(mode="before")
@classmethod
def normalize_null_max_tokens(cls, data):
    # 当 max_tokens 显式传入 null 时, Pydantic 会将其解析为 None
    if isinstance(data, dict) and data.get("max_tokens") is None:
```

```
data = data.copy() # 避免修改原始输入
# 替换为字段定义的默认值（与省略时的行为一致）
data["max_tokens"] = cls.model_fields["max_tokens"].default
return data
```

评论区精华

该 PR 仅有一条来自作者 [@AndreasKaratzas](#) 的评论，请求 [@DarkLight1337](#) 进行强制合并（force merge）。讨论中没有出现技术争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更只在一个新的 validator 内部执行字段替换，且仅在用户显式传入 null 才触发。所有已有行为不变，测试中未发现回归。
- 影响：影响范围有限，仅修复 OpenAI Completions API 的一个边界情况。用户不再会因为发送 {"max_tokens": null} 而得到 500 错误。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR