

PR #50764 完整报告

vllm-project/vllm

[Bugfix][Frontend] Constrain Anthropic cache_salt to non-empty

合并时间: 2026-08-03 13:33

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/50764>

执行摘要

- 一句话: 约束 Anthropic cache_salt 非空, 修复间歇 500
- 推荐动作: 值得快速阅读: 一行源码修复 + 两层测试, 展示了 "校验时机决定错误码" 的经典问题, 以及如何用 OpenAPI schema 断言守护契约。可关注后续统一 cache_salt 约束的 follow-up。

功能与动机

根据 PR body: `test_openapi_stateless[POST /v1/messages]` 间歇性 500, 根因是 `AnthropicMessagesRequest.cache_salt` (#49498 引入) 无长度约束, OpenAPI schema 将空字符串标记为合法值; Schemathesis 在 `positive-data` 模式下按 schema 生成 `cache_salt: ""`, 该值在内部转换到 `ChatCompletionRequest` 时被 `check_cache_salt_support` 拒绝, 异常以 500 而非客户端错误逃逸。由于 Hypothesis 随机生成输入, 该失败只在部分构建复现。

实现拆解

1. 入口约束: 在 `vllm/entrypoints/anthropic/protocol.py` 的 `AnthropicMessagesRequest.cache_salt` 字段追加 `min_length=1`, 一举两得——OpenAPI schema 自动发布 `minLength: 1`, 且空字符串在请求解析层被 Pydantic 拒绝; `None` (省略) 与非空 `salt` 行为不变。
2. 错误路径收敛: 此前空 `salt` 要到 `_convert_anthropic_to_openai_request` 内部转换时才由 `check_cache_salt_support` 拒绝而变成 500; 现在在入口即返回客户端错误。注意测试实际断言 400 (`BAD_REQUEST`), 因服务端注册了自定义 `validation_exception_handler`; PR body 描述的 422 是 FastAPI 默认行为, 两种情况下都不再是 500。
3. 测试配套: `tests/entrypoints/anthropic/test_anthropic_messages_conversion.py` 新增 `TestCacheSalt._make_api_app` 静态方法, 用真实 FastAPI + `attach_router` + `validation_exception_handler` 搭建入口, 并用 `MagicMock` 替换 `serving handler` 以确保无效请求不进入业务层; 新增两个测试: `test_cache_salt_openapi_requires_non_empty_string` 从 `app.openapi()` 解析 `cache_salt` 的 string schema 并断言 `minLength == 1`; `test_empty_cache_salt_returns_bad_request` 用 `TestClient` 发送空 `salt` 请求, 断言返回 400 且 handler 未被调用。
4. 演进: 第一个 commit 由作者完成修复, 第三个 commit (AndreasKaratzas) 补强测试覆盖; 无配置、schema 或部署改动。

关键文件：

- `vllm/entrypoints/anthropic/protocol.py`（模块 协议层；类别 `source`；类型 `core-logic`；符号 `AnthropicMessagesRequest`）：修复入口：给 `AnthropicMessagesRequest.cache_salt` 加 `min_length=1`，从源头杜绝空字符串进入转换链路，同时让 OpenAPI schema 发布约束。
- `tests/entrypoints/anthropic/test_anthropic_messages_conversion.py`（模块 入口测试；类别 `test`；类型 `test-coverage`；符号 `_make_api_app`, `test_cache_salt_openapi_requires_non_empty_string`, `test_empty_cache_salt_returns_bad_request`）：新增两层回归测试：schema 断言 + 端到端 400 请求，防止未来去掉 `min_length` 或 schema 结构回归。

关键符号：`_make_api_app`, `test_cache_salt_openapi_requires_non_empty_string`, `test_empty_cache_salt_returns_bad_request`

关键源码片段

`vllm/entrypoints/anthropic/protocol.py`

修复入口：给 `AnthropicMessagesRequest.cache_salt` 加 `min_length=1`，从源头杜绝空字符串进入转换链路，同时让 OpenAPI schema 发布约束。

```
class AnthropicMessagesRequest(BaseModel):
    """Anthropic Messages API 请求体。"""

    model: str
    messages: list[AnthropicMessage]
    max_tokens: int
    metadata: dict[str, Any] | None = None
    output_config: AnthropicOutputConfig | None = None
    stop_sequences: list[str] | None = None
    stream: bool | None = False
    system: str | list[AnthropicContentBlock] | None = None
    temperature: float | None = None
    tool_choice: AnthropicToolChoice | None = None
    tools: list[AnthropicTool] | None = None
    top_k: int | None = None
    top_p: float | None = None

    # vLLM 扩展字段，不属于 Anthropic 原生规范。
    # min_length=1 有两个作用：
    # 1) OpenAPI schema 会发布 minLength: 1，阻止 schema 驱动客户端
    # （如 Schemathesis）生成空字符串；
    # 2) 显式传入空字符串时，Pydantic 在请求解析阶段直接拒绝，
    # 返回 4xx 客户端错误，而不会进入内部转换，
    # 把 check_cache_salt_support 的校验异常变成 500。
    cache_salt: str | None = Field(
        default=None,
        min_length=1,
```

```

description=(
    "If specified, the prefix cache will be salted with the provided "
    "string to prevent an attacker to guess prompts in multi-user "
    "environments. The salt should be random, protected from "
    "access by 3rd parties, and long enough to be "
    "unpredictable (e.g., 43 characters base64-encoded, corresponding "
    "to 256 bit).",
),
)
kv_transfer_params: dict[str, Any] | None = Field(
    default=None,
    description="KVTransfer parameters used for disaggregated serving.",
)
ec_transfer_params: dict[str, Any] | None = Field(
    default=None,
    description=(
        "ECTransfer parameters used for encoder-cache disaggregated serving."
    ),
)
chat_template_kwargs: dict[str, Any] | None = Field(
    default=None,
    description=(
        "Additional keyword args to pass to the chat template renderer. "
        "Will be accessible by the template."
    ),
)
)

```

tests/entrypoints/anthropic/test_anthropic_messages_conversion.py

新增两层回归测试：schema 断言 + 端到端 400 请求，防止未来去掉 min_length 或 schema 结构回归。

```

class TestCacheSalt:
    def test_cache_salt_passed_through(self):
        """cache_salt 从 Anthropic 请求传递到转换后的
        ChatCompletionRequest, 使 /v1/messages 前缀缓存隔离生效。"""
        request = _make_request(
            [{"role": "user", "content": "Hello"}],
            cache_salt="tenant-abc-secret-salt",
        )
        result = _convert(request)
        assert result.cache_salt == "tenant-abc-secret-salt"

    def test_cache_salt_defaults_to_none(self):
        """省略 cache_salt 时保持默认为 None, 行为不变。"""
        request = _make_request([{"role": "user", "content": "Hello"}])
        result = _convert(request)
        assert result.cache_salt is None

```

@staticmethod

```

def _make_api_app():
    # 构造真实的 FastAPI 应用并挂载 Anthropic 路由,
    # 用 MagicMock 替换 serving handler,
    # 确保无效请求不会到达业务层——到达即触发 AssertionError。
    app = FastAPI()
    attach_router(app)
    app.state.args = Namespace(log_error_stack=False)
    app.exception_handler(RequestValidationError)(validation_exception_handler)

    handler = MagicMock(spec=AnthropicServingMessages)
    handler.create_messages.side_effect = AssertionError(
        "invalid requests must not reach the serving handler"
    )
    app.state.anthropic_serving_messages = handler
    return app, handler

def test_cache_salt_openapi_requires_non_empty_string(self):
    # 从生成的 OpenAPI schema 中读取 cache_salt 的 string 分支,
    # 断言其带有 minLength: 1, 防止 schema 驱动测试
    # (如 Schemathesis) 再生成空字符串请求。
    app, _ = self._make_api_app()
    field_schema = app.openapi()["components"]["schemas"][
        "AnthropicMessagesRequest"
    ]["properties"]["cache_salt"]
    string_schema = next(
        option for option in field_schema["anyOf"] if option.get("type") == "string"
    )

    assert string_schema["minLength"] == 1

def test_empty_cache_salt_returns_bad_request(self):
    # 显式传空字符串应在请求解析层返回 400,
    # 且不该进入 serving handler (否则会变成 500) 。
    app, handler = self._make_api_app()
    with TestClient(app, raise_server_exceptions=False) as client:
        response = client.post(
            "/v1/messages",
            json={
                "model": "test-model",
                "max_tokens": 1,
                "messages": [{"role": "user", "content": "Hello"}],
                "cache_salt": "",
            },
        )

    assert response.status_code == HTTPStatus.BAD_REQUEST
    handler.create_messages.assert_not_awaited()

```

评论区精华

1. DarkLight1337 在 `vllm/entrypoints/anthropic/protocol.py` 第 138 行提问 "Maybe we should update this for all the other occurrences as well?", AndreasKaratzas 回应 "True, but probably best as a follow up I think. CI already completed the run too.", DarkLight1337 回复 "Sure"——统一约束被拆为后续工作。
 2. noooop 对 `test_openapi_stateless` 的随机性表示困惑 (build 81852 重试即过) ; AndreasKaratzas 确认多数构建不会生成空字符串、`min_length=1` 合理, 并主动补强测试。
 3. 一个值得注意的细节: PR body 预期 422, 测试实际断言 400 (`HTTPStatus.BAD_REQUEST`) ; 差异来自自定义 `validation_exception_handler`, 对客户而言核心是从 500 变为 4xx。
- 是否统一所有 `cache_salt` 声明的 `min_length` 约束 (design): 同意作为后续跟进, 本 PR 只约束 Anthropic 入口。
 - `test_openapi_stateless` 随机失败原因 (testing): 确认是 Schemathesis 随机生成空字符串导致间歇失败, `min_length=1` 修复, 并由 AndreasKaratzas 补充测试。

风险与影响

- 风险:
 1. 错误码行为变更: 显式传空 `cache_salt` 的请求原本必然在转换期 500, 现变为 4xx; 没有 " 合法 " 用法被破坏, 属于修正性变更。
 2. 同类遗漏: `cache_salt` 可能存在于其他协议入口 (如 OpenAI 侧同类字段), review 中已提出统一约束, 本 PR 未处理, 存在一致性问题, 需 follow-up。
 3. 测试脆弱性: schema 测试直接依赖 `anyOf` 结构与 `minLength` 键名, pydantic 版本升级可能改变 schema 布局, 需同步维护; 但这也正是守护 schema 契约的测试目的。
 4. 影响面: 仅限 Anthropic 入口校验层, 不触碰模型执行、采样、输出路径, 回归风险低。
 - 影响: 对调用 Anthropic Messages API 的用户而言, 无效请求现在提前得到 4xx 客户端错误, 错误语义更清晰; 对 CI 而言, 消除了 `test_openapi_stateless[POST /v1/messages]` 的间歇性 500; 对团队而言, 为后续 " 统一各入口 `cache_salt` 约束 " 提供了测试范式——通过 OpenAPI schema 断言与端到端请求双重验证。
 - 风险标记: 错误码行为变更 (500 → 4xx), 同类字段约束待统一, 测试依赖 OpenAPI schema 结构

关联脉络

- PR #49498 [Frontend] Add `cache_salt` support to Anthropic Messages API: 引入 `AnthropicMessagesRequest.cache_salt` 字段但未声明长度约束, 本 PR 修复其导致的间歇性 500 与 schema 误报。