

PR #46783 完整报告

vllm-project/vllm

[Misc] Move the legacy api_server.py to the examples directory.

合并时间: 2026-06-26 15:43

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46783>

执行摘要

- 一句话: 将遗留 API server 移至 examples 目录
- 推荐动作: 该 PR 属于代码组织清理, 值得精读以了解项目结构演进。更推荐关注后续是否会在 examples 中进一步维护或废弃该遗留 server。

功能与动机

PR body 指出需要消除 vllm/entrypoints/api_server.py 和 vllm/entrypoints/openai/api_server.py 的混淆, 将 legacy 实现移至 examples 目录。

实现拆解

1. 搬迁核心文件: 将 vllm/entrypoints/api_server.py 移至 examples/applications/api_server/server.py, 并更新 logger 名称从 vllm.entrypoints.api_server 改为 api_server。
2. 搬迁客户端文件: 将 examples/applications/chatbot/api_client.py 移至 examples/applications/api_server/client.py, 更新文档字符串中的启动命令。
3. 更新依赖引用: 修改 examples/applications/chatbot/gradio_webserver.py 中的启动命令注释, 指向新路径。
4. 更新文档: 修改 docs/design/arch_overview.md 中的示例链接, 并更新 docs/examples/README.md 中对 applications 目录的描述。

关键文件:

- examples/applications/api_server/server.py (模块 示例应用; 类别 source; 类型 rename-or-move) : 原 vllm/entrypoints/api_server.py 搬迁至此, 是核心变更文件。仅更新了 logger 名称。

关键符号: 未识别

关键源码片段

`examples/applications/api_server/server.py`

原 vllm/entrypoints/api_server.py 搬迁至此, 是核心变更文件。仅更新了 logger 名称。

```
# SPDX-License-Identifier: Apache-2.0
```

```
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project
```

```
"""
```

NOTE: This API server is used only for demonstrating usage of AsyncEngine and simple performance benchmarks. It is not intended for production use. For production use, we recommend using our OpenAI compatible server. We are also not going to accept PRs modifying this file, please change `vllm/entrypoints/openai/api\_server.py` instead.

```
"""
```

```
import asyncio
import json
import ssl
from argparse import Namespace
from collections.abc import AsyncGenerator
from typing import Any

from fastapi import FastAPI, Request
from fastapi.responses import JSONResponse, Response, StreamingResponse

import vllm.envs as envs
from vllm.engine.arg_utils import AsyncEngineArgs
from vllm.engine.async_llm_engine import AsyncLLMEngine
from vllm.entrypoints.launcher import serve_http
from vllm.entrypoints.serve.utils.api_utils import with_cancellation
from vllm.logger import init_logger
from vllm.sampling_params import SamplingParams
from vllm.usage.usage_lib import UsageContext
from vllm.utils import random_uuid
from vllm.utils.argparse_utils import FlexibleArgumentParser
from vllm.utils.system_utils import set_ulimit
from vllm.version import __version__ as VLLM_VERSION

# 变更: logger 名称从 "vllm.entrypoints.api_server" 改为 "api_server"
logger = init_logger("api_server")

app = FastAPI()
engine = None

@app.get("/health")
async def health() -> Response:
    """Health check."""
    return Response(status_code=200)

@app.post("/generate")
async def generate(request: Request) -> Response:
    request_dict = await request.json()
    return await _generate(request_dict, raw_request=request)
```

```
@with_cancellation
async def _generate(request_dict: dict, raw_request: Request) -> Response:
    prompt = request_dict.pop("prompt")
    stream = request_dict.pop("stream", False)
    sampling_params = SamplingParams(**request_dict, skip_clone=True)
    request_id = random_uuid()
```

评论区精华

该 PR 无实质性讨论，仅包含 claude[bot] 的自动化评论（因来自 fork 禁用自动审查）和 DarkLight1337 的批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。所有变更均为文件搬迁和文档链接更新，不含逻辑修改。旧路径 `vllm/entrypoints/api_server.py` 被移除，若用户或外部脚本直接引用该模块则会报错，但 PR 目的即为废弃该路径。
- 影响：对用户的影响：用户需要更新启动命令从 `python -m vllm.entrypoints.api_server` 改为 `python examples/applications/api_server/server.py`。对系统的影响：减少核心入口点数量，降低新用户混淆。对团队的影响：维护成本降低。影响程度：较小。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR