# PR #41907 完整报告

vllm-project/vllm

[Docs] Reorganize online serving docs.

合并时间：2026-05-19 14:43

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41907

## 执行摘要

- 一句话：重构在线服务文档，拆分旧 OpenAI 指南到多页
- 推荐动作：对于文档维护者和用户，建议仔细阅读 docs/serving/online_serving/README.md 了解新组织方式。贡献者添加新在线 API 时应遵循本 PR 确立的目录和命名规范，并更新 docs/.nav.yml 和 mkdocs.yaml 的重定向映射。

## 功能与动机

In the early days of vLLM (2023), there was only the OpenAI, so vLLM online serving was referred to as the OpenAI-Compatible Server. As we have entered 2026, generative AI has grown significantly, and more and more features have been added to vLLM online serving. Therefore, it is necessary for us to reorganize the online serving documentation. ( 引自 PR body)

## 实现拆解

1. 创建新文档目录并添加细分页面：在 docs/serving/online_serving/ 下新增 README.md（在线服务总览，列出所有支持的 API 类型及其端点）、openai_compatible_server.md（精简后的 OpenAI 兼容 API 细节）、speech_to_text.md（语音转文字 / 翻译 API）、generative_scoring.md（生成式评分 API）、renderer.md（渲染器 API 占位页）、tokenizer.md（分词器 API）等页面。
2. 删除旧聚合文档：移除 docs/serving/openai_compatible_server.md，其内容按功能分配到新页面，避免信息大量冗余。
3. 配置重定向：在 mkdocs.yaml 的 redirects 插件中增加 serving/openai_compatible_server.md: serving/online_serving/README.md 映射，确保旧书签和搜索引擎链接仍然可用。
4. 更新导航配置：修改 docs/.nav.yml，将“Online Serving”的链接从指向单一文件改为指向目录 serving/online_serving，利用 MkDocs 的目录索引自动生成子页面列表。
5. 同步更新交叉引用：更新 docs/getting_started/quickstart.md、docs/models/generative_models.md、docs/models/supported_models.md、docs/models/pooling_models/README.md 等文件中的链接和章节标题，指向新的页面路径。

6. 统一文件命名和标题风格：根据 review 建议，将 generative_scoring_api.md 去掉“API”后缀，拆解 others.md 为独立页面（如 tokenizer.md），并统一各页面标题的大小写格式。
7. 更新 SVG 资源：docs/assets/models/pooling_models/cheat_sheet.svg 进行了较大修改（+671/-660），可能与 pooling 文档的同步更新有关。

关键文件：

- docs/serving/online_serving/README.md（模块 在线服务文档；类别 docs；类型 documentation）：在线服务新入口页面，统一概述所有 API 类型及其端点，是重构后的核心导航页。
- docs/serving/online_serving/openai_compatible_server.md（模块 在线服务文档；类别 docs；类型 documentation）：从旧文件拆分的核心 OpenAI 兼容 API 细节页面，聚集了原来长文档的精简内容。
- docs/serving/openai_compatible_server.md（模块 在线服务文档；类别 docs；类型 deletion）：旧入口文件被删除，内容迁移到新目录，标志旧文档结构的结束。
- mkdocs.yaml（模块 文档构建；类别 config；类型 configuration）：添加重定向规则，确保旧链接不失效，是文档向后兼容的关键操作。
- docs/.nav.yml（模块 文档导航；类别 config；类型 configuration）：导航配置变更，将在线服务从单一文件指向目录，利用 MkDocs 目录索引。
- docs/serving/online_serving/speech_to_text.md（模块 在线服务文档；类别 docs；类型 documentation）：新增独立语音 API 文档，覆盖了原有文档中缺失的语音接口描述。
- docs/getting_started/quickstart.md（模块 入门指南；类别 docs；类型 documentation）：更新了章节标题和链接，反映文档结构变化，是受波及的修改之一。
- docs/assets/models/pooling_models/cheat_sheet.svg（模块 文档资源；类别 other；类型 data-contract）：SVG 文件有大幅修改（+671/-660），与 pooling 文档的同步更新相关。

关键符号：未识别

## 关键源码片段

### mkdocs.yaml

添加重定向规则，确保旧链接不失效，是文档向后兼容的关键操作。

```
# mkdocs.yaml 重定向配置片段
plugins:
  - redirects:
      redirect_maps:
        features/spec_decode/README.md: features/speculative_decoding/README.md
        features/spec_decode/speculators.md: features/speculative_decoding/speculators.md
        serving/openai_compatible_server.md: serving/online_serving/README.md # <-- 新增重定向，确保旧链接可用
```

## 评论区精华

- 重定向配置：DarkLight1337 要求添加旧路径重定向，hmellor 提供了具体语法建议 serving/openai_compatible_server.md: serving/online_serving/README.md，最终被接受。
- 文件命名与结构：DarkLight1337 指出 generative_scoring_api.md 文件名带“API”后缀不统一，且 others.md 作为杂项页面不合适，建议拆分为独立页面（如 tokenizer.md），作者均采纳。
- 链接正确性：gemini-code-assist[bot] 发现多个空链接（Anthropic、Completions 渲染）和错误的相对路径（评分链接指向 scoring.md），这些都被作者及时修复。
- 标题大小写一致：DarkLight1337 要求检查并统一页面标题的大小写格式，作者进行了全局调整。
- API 分类边界讨论：DarkLight1337 和作者就 Custom APIs 与 Cohere APIs 下 Rerank API 是否重复展开讨论，作者坚持 Score API 属于 Custom、Rerank API 属于 Cohere 的分类，DarkLight1337 认为应避免重复但最终未强制修改。
- 添加旧路径重定向 (design): 接受建议，在 mkdocs.yaml 中添加了 serving/openai_compatible_server.md: serving/online_serving/README.md。
- 文件命名和结构一致性 (style): 重命名文件，拆分 others.md 为 tokenizer.md 等独立页面。
- 修正文档中的空链接和断链 (correctness): 作者修复了所有发现的问题链接。
- 标题大小写格式统一 (style): 作者进行了全局大小写调整。
- API 分类边界（Custom vs Cohere） (design): 当前结构被保留，但未完全解决分歧。

## 风险与影响

- 风险：
  - 链接失效风险：如果重定向配置遗漏或新页面路径变化未同步更新所有交叉引用，用户可能遇到 404。本 PR 通过添加重定向和同步更新引用文件降低了风险。
  - 内容完整性风险：部分新增页面（如 renderer.md）仅占位无实际内容，可能让用户困惑。但已标记为待补充。
  - 导航变更适应：用户可能不熟悉新导航结构，但目录索引可自动展开，且旧链接仍可使用。
  - SVG 图片大幅修改：cheat_sheet.svg 的变更可能与未来 pooling 文档进一步更新产生冲突。
- 影响：
  - 用户 / 读者：在线服务文档结构更清晰，按功能分类易于查找；旧链接通过重定向保持可用。
  - 开发者 / 贡献者：新增 API 文档时可在相应子页面下独立修改，减少单页面冲突；导航和重定向配置需同步更新。
  - 团队维护：文档结构更模块化，降低后续扩展的认知负担。
  - 风险标记：文档断裂，外部链接失效，路径引用错误

## 关联脉络

- PR #42626 [Docs] Add SVG images for pooling models.: 与该 PR 修改了同一个 SVG 文件 docs/assets/models/pooling_models/cheat_sheet.svg，可能产生冲突或后续整合。
- PR #41082 Examples directory refactoring: 作者在评论中提及该 PR 重构了 examples 目录，影响了本文档中引用的示例路径。
- PR #41084 Examples directory refactoring: 作者在评论中提及该 PR 也参与了 examples 目录重构，与上一条相关联。