PR #40995 完整报告

vllm-project/vllm

[Examples] Resettle features examples.

合并时间：2026-04-28 15:33

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40995

执行摘要

- 一句话：按功能特性重组 examples 目录，迁移 50+ 示例文件并更新引用
- 推荐动作：该 PR 值得关注其目录设计思路，对于仓库结构优化有参考价值。虽然逻辑简单，但涉及大量文件，审查时应利用自动化工具检查引用完整性。后续添加新示例应直接放入对应 feature 子目录。

功能与动机

原来的 examples 目录按 offline_inference 和 online_serving 分类，导致特定功能示例散落在不同目录中，用户需要跨目录查找。issue #29362 提出按使用场景（功能特性）作为第一级目录，以提高示例的可发现性。

实现拆解

1. 创建 feature 子目录：在 examples/features/ 下按功能建立子目录，如 automatic_prefix_caching、data_parallel、speculative_decoding 等。
2. 文件迁移与重命名：将原 examples/offline_inference/ 和 examples/online_serving/ 下的对应文件移动到相应子目录，并重命名添加 _offline 或 _client 后缀以区分运行模式。例如 data_parallel.py → data_parallel_offline.py。
3. 更新内部引用：更新文件内 docstring 中的运行命令路径，更新所有外部引用：CI 配置（.buildkite/*.yaml）、Mergify 自动化规则（.github/mergify.yml）、文档（docs/）以及核心代码中的错误消息（vllm/entrypoints/llm.py）和注释（vllm/model_executor/model_loader/sharded_state_loader.py）。
4. 特殊处理 routed_experts：将 routed_experts_e2e.py 从 offline_inference 移至 examples/rl/ 目录，归类为强化学习示例，而非保留在 features/routed_experts。
5. 清理废弃引用：移除 vllm/benchmarks/datasets/datasets.py 中指向已删除的 eagle.py 的注释（后决定保留原样）。

关键文件：

- examples/features/pause_resume/data_parallel_pause_resume.py（模块 示例文件；类别 source；类型 rename-or-move；符号 pause_generation, resume_generation, main, generator_thread）：此文件从 examples/online_serving/ 搬迁至 examples/features/pause_resume/，是本次重组中唯一从 online_serving 迁移的在线示例，跨目录移动具有代表性。

- vllm/entrypoints/llm.py（模块 入口点；类别 source；类型 configuration；符号 _make_config）：核心入口文件中的错误消息引用了示例路径，本 PR 尝试更新但存在遗漏（只改了后缀，未改目录），暴露了跨模块引用更新的难度。

关键符号：pause_generation, resume_generation, main, _make_config, create_parser

关键源码片段

examples/features/pause_resume/data_parallel_pause_resume.py

此文件从 examples/online_serving/ 搬迁至 examples/features/pause_resume/，是本次重组中唯一从 online_serving 迁移的在线示例，跨目录移动具有代表性。

```
# 此文件从 examples/online_serving/data_parallel_pause_resume.py 移至此处，
# 代码内容与搬迁前完全一致，仅文件路径变化。

def pause_generation(base_url: str, mode: str = "keep") -> None:
    """通过 HTTP 端点暂停生成。"""
    url = f"{base_url}/pause"
    response = requests.post(url, params={"mode": mode}, timeout=60)
    response.raise_for_status()
    print("Server paused")

def resume_generation(base_url: str) -> None:
    """通过 HTTP 端点恢复生成。"""
    url = f"{base_url}/resume"
    response = requests.post(url, timeout=60)
    response.raise_for_status()
    print("Server resumed")
```

vllm/entrypoints/llm.py

核心入口文件中的错误消息引用了示例路径，本 PR 尝试更新但存在遗漏（只改了后缀，未改目录），暴露了跨模块引用更新的难度。

```
def _make_config(value: Any, cls: type[_R]) -> _R:
    # ...
    raise ValueError(
        f"LLM(data_parallel_size={_dp_size}) is not supported for single-"
        "process usage and may hang. Please use "
        "the explicit multi-process data-parallel example at "
        "'examples/offline_inference/data_parallel_offline.py'."
        # 审查指出：此路径应最终指向 'examples/features/data_parallel/data_parallel_offline.py'，
        # 当前仅添加了 `_offline` 后缀，未修正目录。
    )
```

评论区精华

> gemini-code-assist[bot]指出了多处路径更新遗漏：CI 测试命令中 prefix_caching 和 spec_decode 的路径仍指向 offline_inference；Mergify 规则中 structured_outputs 路径不正确；vllm/entrypoints/llm.py 错误消息只加了 _offline 后缀但未改目录。PR 作者修复

了部分，但 llm.py 的问题未完全修复。

DarkLight1337 vs noooop就 routed_experts 示例的归属展开讨论：DarkLight1337 认为可归入 observability，noooop 则认为更适合 reinforcement learning，并引用 PR#28284 支持。最终决定移至 examples/rl/，说明团队采纳了 RL 归属。

- 示例路径更新遗漏 (correctness): PR 作者修正了部分，但 llm.py 仍保留 offlie_inference 目录（仅加 _offline 后缀）。PR 合并后可能需后续跟进。
- routed_experts 示例归属 (design): 最终将 routed_experts_e2e.py 移至 examples/rl/，采纳了 RL 归属。

风险与影响

- 风险：主要风险在于引用更新遗漏。gemini-code-assist[bot] 已发现多处，PR 合并前可能还有未发现的路径引用错误，尤其需关注用户文档和教程。由于示例文件逻辑不变，无功能回归风险。
- 影响：对用户：示例路径改变，旧路径失效，用户需更新运行命令。文档和 CI 已同步更新，但未发现的引用可能导致测试跳过或规则失效。对团队：新结构更清晰，便于维护，但需确保所有新示例按新目录放置。
- 风险标记：引用更新遗漏，CI 配置未同步，自动化规则失效

关联脉络

- PR #29362 [RFC]: Resettle examples.: 本 PR 是该 RFC 的具体实现，按照提议重组 examples 目录。