PR #40123 完整报告

vllm-project/vllm

[Examples] Resettle Observability examples.

合并时间：2026-04-17 18:13

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40123

执行摘要

- 一句话：将 Observability 示例移动到统一目录，重构示例组织结构。
- 推荐动作：该 PR 值得快速浏览以了解新的示例组织结构，对于维护示例或文档的工程师尤其相关。无需深究设计决策，关注点在于组织改进的实践。

功能与动机

动机源于 issue #29362 中提出的问题：当前 examples 目录结构以 offline_inference 和 online_serving 作为一级目录，使用户难以查找特定用例（如可观测性）的示例。提议按使用场景组织目录，并创建 'observability' 目录以集中相关资源，如指标监控、追踪和仪表板配置，提升代码库的可维护性和用户体验。

实现拆解

1. 创建新目录结构：在 examples/ 下新增 observability/ 目录，并建立子目录如 metrics/、opentelemetry/、dashboards/ 等。
2. 移动源文件：将分散在 offline_inference 和 online_serving 中的 Observability 相关文件重命名并移动到新目录，例如 examples/offline_inference/metrics.py → examples/observability/metrics/offline.py，其他类似文件如 dummy_client.py 也相应移动。
3. 移动配置文件：包括 Grafana 的 JSON 仪表板文件和 Perses 的 YAML 配置文件，从原位置移动到 observability/dashboards/ 下。
4. 更新文档：修改 docs/design/metrics.md 和 examples/observability/dashboards/README.md 等文档中的路径引用，确保链接正确。
5. 无测试或部署配套改动：本次变更纯属文件组织结构调整，未涉及功能逻辑、测试或部署脚本的修改。

关键文件：

- examples/observability/metrics/offline.py（模块 示例目录；类别 source；类型 rename-or-move）：作为重命名的源文件示例，展示了从 offline_inference 移动到 observability 的核心变更，是 Observability 功能的代码入口。
- examples/observability/opentelemetry/dummy_client.py（模块 示例目录；类别 source；类型 rename-or-move）：展示 OpenTelemetry 追踪示例的客户端文件，移动后统一在 observability 目录下，便于用户查找可观测性相关示例。

- docs/design/metrics.md（模块 设计文档；类别 docs；类型 documentation）：更新的设计文档，反映了 Observability 示例的新路径，确保文档与代码结构同步，避免误导用户。

关键符号：未识别

关键源码片段

examples/observability/metrics/offline.py

作为重命名的源文件示例，展示了从 offline_inference 移动到 observability 的核心变更，是 Observability 功能的代码入口。

```
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project

from vllm import LLM, SamplingParams
from vllm.v1.metrics.reader import Counter, Gauge, Histogram, Vector

# 示例提示，用于演示指标监控
prompts = [
    "Hello, my name is",
    "The president of the United States is",
    "The capital of France is",
    "The future of AI is",
]
# 创建采样参数对象，控制生成过程
sampling_params = SamplingParams(temperature=0.8, top_p=0.95)


def main():
    # 创建 LLM 实例，启用日志统计以收集指标
    llm = LLM(model="facebook/opt-125m", disable_log_stats=False)

    # 从提示生成文本，触发指标收集
    outputs = llm.generate(prompts, sampling_params)

    # 打印生成的文本输出
    print("-" * 50)
    for output in outputs:
        prompt = output.prompt
        generated_text = output.outputs[0].text
        print(f"Prompt: {prompt!r}\nGenerated text: {generated_text!r}")
        print("-" * 50)

    # 导出并打印所有收集到的指标，包括计数器、仪表、向量和直方图
    for metric in llm.get_metrics():
        if isinstance(metric, Gauge):
            print(f"{metric.name} (gauge) = {metric.value}")
        elif isinstance(metric, Counter):
            print(f"{metric.name} (counter) = {metric.value}")
```

```
        elif isinstance(metric, Vector):
            print(f"{metric.name} (vector) = {metric.values}")
        elif isinstance(metric, Histogram):
            print(f"{metric.name} (histogram)")
            print(f"    sum = {metric.sum}")
            print(f"    count = {metric.count}")
            for bucket_le, value in metric.buckets.items():
                print(f"    {bucket_le} = {value}")


if __name__ == "__main__":
    main()
```

评论区精华

review 讨论较为简单。gemini-code-assist bot 总结了变更内容：“重新组织项目的示例，将可观测性相关资源从 'online_serving' 目录移到新的 'observability' 目录，并更新了文档链接。”并指出没有评论需要解决。DarkLight1337 直接批准了 PR，未提出异议或深入讨论。

- 变更总结与批准 (documentation): 无评论需要解决，PR 被 DarkLight1337 批准。

风险与影响

- 风险：技术风险较低。主要风险是现有外部脚本或文档中可能硬编码了旧文件路径，导致引用失效。但 PR 已更新了核心文档中的链接，降低了此风险。由于是纯文件移动，代码内容未变，因此无回归、性能或安全风险。兼容性方面，用户需要更新本地环境中的路径引用，但影响有限。
- 影响：影响范围限于 examples 目录的结构和使用者。对于用户，需要适应新路径来查找 Observability 示例；对于系统，无运行时功能影响；对于团队，示例组织更清晰，便于维护。影响程度低，属于非侵入性的组织结构优化。
- 风险标记：路径变更影响引用

关联脉络

- 暂无明显关联 PR