PR #31113 完整报告

vllm-project/vllm

Fix document of torchrun_example.py

合并时间：2026-03-31 18:54

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/31113

执行摘要

PR #31113 修复了 torchrun_example.py 文档中的错误，更新了进程数说明以匹配并行维度乘积，但 review 指出文档仍不完整，风险低且影响范围小。

功能与动机

该 PR 的动机是修正文档注释，确保用户在使用 torchrun 示例脚本时能正确配置进程数。PR body 中明确表示“Fix document/comment.”，即针对文档错误进行修复，以避免因过时信息导致用户配置错误。

实现拆解

实现仅涉及一个文件的文档注释更新：

- 文件路径：*examples/offline_inference/torchrun_example.py*
- 变更内容：将命令行参数从 torchrun --nproc-per-node=2 torchrun_example.py 改为 torchrun --nproc-per-node=4 torchrun_example.py，并更新解释从“the argument 2 should match the tensor_parallel_size below.”到“the argument 4 should match the product of tensor_parallel_size and pipeline_parallel_size below.”
- 模块归属：*examples/offline_inference*，属于分布式推理示例代码。

评论区精华

review 中 gemini-code-assist[bot] 提出了关键讨论：

> “While this change is an improvement, the explanation for nproc-per-node is still incomplete. For external_launcher, the total number of processes should be the product of all parallelism dimensions: tensor_parallel_size, pipeline_parallel_size, prefill_context_parallel_size, and data_parallel_size.”

这表明文档仍有改进空间，但审核者 simon-mo 未采纳此建议便批准合并。

风险与影响

- 技术风险：风险极低，仅文档变更无代码逻辑影响。但文档不准确可能导致用户错误配置进程数，影响分布式推理正确性。
- 影响范围：仅影响使用该示例脚本的用户，提高文档准确性但未完全解决所有并行维度，潜在误导风险小。

关联脉络

从提供的同仓库近期历史 PR 分析中，未识别到直接相关 PR。此 PR 是独立的文档修复，不涉及跨功能演进，反映了仓库对示例文档的持续维护。