PR #42935 完整报告

vllm-project/vllm

Fix `--convert` passed without `--runner` on causal models

合并时间：2026-05-18 23:56

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42935

执行摘要

- 一句话：修复因果模型 --convert 未与 --runner 同时传递时的崩溃
- 推荐动作：值得合并，修复了显式的用户错误（缺少 --runner）导致的崩溃，且与文档行为一致。变更极小，逻辑清晰，应无回归风险。

功能与动机

官方文档（https://docs.vllm.ai/en/latest/models/pooling_models/#model-conversion）指出，如果模型本身不支持转换任务，--runner 应自动设为 pooling。但现有代码未处理 --convert 已显式传递但 --runner 未设置的情况，导致启动时崩溃（Issue #42480）。此 PR 旨在实现对文档承诺的自动降级行为。

实现拆解

1. 修改 _get_runner_type 方法签名：在 vllm/config/model.py 中，为 _get_runner_type 添加新参数 convert: ConvertOption，使其能够感知转换选项。
2. 更新调用点：在 __post_init__ 中将 self.convert 作为第三个实参传递给 _get_runner_type。
3. 添加分支逻辑：在 _get_runner_type 内部，当 runner == "auto" 时，若 convert 不是 "auto" 或 "none"（即用户显式指定了某种转换），则将 runner 类型设为 "pooling"，否则保持原有默认检测逻辑。
4. 后续校验：由于 _get_convert_type 后续会根据 runner_type 验证转换类型合法性，自动设为 pooling 后，校验路径将正常通过，避免了之前的崩溃。 （本次变更未涉及测试文件新增或修改，仅修改了源码。无配置或部署配套改动。）

关键文件：

- vllm/config/model.py（模块 配置；类别 source；类型 data-contract）：核心实现文件；修改了 _get_runner_type 方法以感知 convert 参数，并在 runner 为 auto 时根据 convert 自动选择 pooling 或执行默认检测。

关键符号：_get_runner_type

关键源码片段

vllm/config/model.py

核心实现文件；修改了 _get_runner_type 方法以感知 convert 参数，并在 runner 为 auto 时根据 convert 自动选择 pooling 或执行默认检测。

```
def _get_runner_type(
    self,
    architectures: list[str],
    runner: RunnerOption,
    convert: ConvertOption, # 新增：传入 `--convert` 选项
) -> RunnerType:
    if runner != "auto":
        return runner # 如果用户显式指定了 runner，直接返回

    # 核心变更：如果用户传递了 `--convert`（非 auto/none），
    # 则将 runner 自动设为 "pooling"
    if convert in {"auto", "none"}:
        runner_type = self._get_default_runner_type(architectures)
    else:
        runner_type = "pooling"

    # 除常见的 generate 外，记录一下
    if runner_type != "generate":
        logger.info(
            "Resolved `--runner auto` to `--runner %s`. "
            "Pass the value explicitly to silence this message.",
            runner_type,
        )

    return runner_type
```

评论区精华

本 PR 未产生 review 评论。唯一的机器人评论来自 gemini-code-assist[bot]，确认了变更内容但未提供额外反馈。两位审查者（noooop、yewentao256）均直接批准，表明变更清晰且无争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅涉及一处逻辑分支：当 runner == "auto" 且 convert 为有效非默认值时，将 runner 强制设为 pooling。这不会影响已显式指定 --runner 的用户；对于依赖默认自动检测的用户，由于 convert 默认为 "auto" 或 "none"，原有流程不变。潜在风险是如果未来新增转换类型，需确保该分支仍合理。
- 影响：直接影响使用 --convert 选项但未指定 --runner 的因果语言模型用户，他们现在可以正常启动而不会崩溃。间接影响：强化了文档中“自动设为 pooling”的行为承诺，提升了用户体验。影响范围小，仅涉及一个文件中的两处改动。
- 风险标记：缺少测试覆盖

关联脉络

- PR #42820 Previously superseded PR for same fix: 此 PR 取代了 #42820，同一作者为同一问题提供的修复尝试。
- PR #42480 Issue describing the bug: 关联的 Issue #42480 报告了该 bug，并作为此 PR 的动机来源。