PR #43143 完整报告

vllm-project/vllm

[Cohere] Enable Cohere MoE

合并时间：2026-05-20 10:32

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43143

执行摘要

- 一句话：启用 Cohere Command A+ MoE 模型
- 推荐动作：该 PR 改动很小，但代表了 Cohere Command A+ 模型的正式发布入口。建议模型集成相关团队关注，确认文档中的 curl 示例是否需要修复以匹配正确的模型名。

功能与动机

本 PR 是 #42078 的后续，目的是将 Cohere Command A+（官方名称）模型正式集成到 vLLM 中。之前使用的路径 /host/engines/cohere-moe 是内部开发路径，现在官方 Hugging Face 仓库 CohereLabs/command-a-plus-05-2026 已发布，需要更新以允许用户直接加载。

实现拆解

1. 更新测试注册表（tests/models/registry.py）：将 Cohere2MoeForCausalLM 的 _HfExamplesInfo 路径从 "/host/engines/cohere-moe" 改为 "CohereLabs/command-a-plus-05-2026"，并保留 trust_remote_code=True 和 is_available_online=False 标志。
2. 更新文档（docs/models/supported_models.md）：将 Cohere2MoeForCausalLM 的模型名称从 Command (MoE) 改为 Command-A+，并将示例路径更新为 CohereLabs/command-a-plus-05-2026，etc.。

关键文件：

- tests/models/registry.py（模块 模型注册；类别 test；类型 test-coverage）：更新 Cohere2MoeForCausalLM 模型的 HF 示例路径，从内部路径改为官方发布路径，是启用的关键变更。
- docs/models/supported_models.md（模块 文档；类别 docs；类型 documentation）：同步更新 Cohere2MoeForCausalLM 的文档描述和示例路径，帮助用户正确使用模型。

关键符号：未识别

关键源码片段

tests/models/registry.py

更新 Cohere2MoeForCausalLM 模型的 HF 示例路径，从内部路径改为官方发布路径，是启用的关键变更。

```
# tests/models/registry.py (line 241-245)
"Cohere2MoeForCausalLM": _HfExamplesInfo(
    "CohereLabs/command-a-plus-05-2026", # 从 "/host/engines/cohere-moe" 更新
    trust_remote_code=True,
    is_available_online=False, # 仍标记为线上不可用（可能需要权限或认证）
),
```

评论区精华

仅有一条来自 gemini-code-assist[bot] 的评论，指出 PR 描述中的 curl 示例使用了错误的模型名 CohereLabs/cohere-transcribe-03-2026（转录模型），应改为 CohereLabs/command-a-plus-05-2026。该评论未得到回复或更新，但 PR 仍被 ywang96 批准合并。

- PR 描述中的 curl 示例使用了错误的模型名 (documentation): 评论未被回复或修复，但 PR 仍被合并。建议后续更新 PR 描述以修复此问题。

风险与影响

- 风险：风险极低。变更仅涉及测试注册表中的示例路径和文档描述，不涉及任何核心代码逻辑。路径改为公开 Hugging Face 仓库后，用户可直接加载模型，但需注意 is_available_online=False 标志表明该模型目前在线上可能不可用或需要特殊权限。
- 影响：对用户：Cohere Command A+ 模型用户现在可以通过 CohereLabs/command-a-plus-05-2026 加载模型，并按照文档中的命令进行推理。对系统：无影响，仅数据变更。对团队：Cohere 模型的维护者现在有了官方 HF 路径，便于测试和发布。
- 风险标记：文档描述存在潜在不一致

关联脉络

- PR #42078 [Feature] Cohere Command A+ model support: 本 PR 是该 PR 的后续，用于启用官方模型路径。