PR #38120 完整报告

vllm-project/vllm

[Cohere] Enable Cohere Transcribe

合并时间：2026-03-26 07:13

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38120

执行摘要

此 PR 启用了 Cohere Transcribe 模型，集成到 vLLM 以支持语音识别功能。通过更新模型注册表、文档和测试，验证了可变长度编码器输入的代码路径，但注册表存在配置重复风险，建议后续优化。

功能与动机

PR 是 PR 35809 的后续，旨在启用 Cohere Transcribe 模型（官方名称 CohereAsrForConditionalGeneration），HF 仓库为 CohereLabs/cohere-transcribe-03-2026。动机是集成新模型，并利用它测试可变长度编码器输入的代码路径，此前 vLLM 仅支持填充长度的编码器输入，适用于 Whisper 但不适用于 Cohere-Transcribe。

实现拆解

- 文档模块：更新 docs/models/supported_models.md，添加模型到支持列表。
- 示例模块：更新 examples/offline_inference/audio_language.py，使用正确模型名。
- 测试模块：更新 tests/entrypoints/openai/correctness/test_transcription_api_correctness.py，添加模型到测试配置（但注释掉未启用），并引入 EnglishTextNormalizer 作为标准归一器。代码示例：python normalizer = EnglishTextNormalizer(normalizer_tokenizer.english_spelling_normalizer)
- 模型执行模块：更新 vllm/model_executor/models/cohere_asr.py，重命名类为 CohereAsrForConditionalGeneration；更新 vllm/model_executor/models/registry.py，调整注册表条目。

评论区精华

review 中，gemini-code-assist[bot] 在 vllm/model_executor/models/registry.py 第 531 行评论：

> "This adds an alias for the Cohere ASR model by duplicating the configuration tuple. This can lead to maintenance issues..."

建议重构为共享配置，但此评论未得到回复，PR 已合并，表明风险被暂时搁置。

风险与影响

风险：

1. 注册表配置重复：可能导致未来变更不一致，增加维护成本。
2. 测试未启用：模型在测试中被注释掉，延迟了完整测试覆盖。
3. 兼容性：新模型涉及可变长度编码器输入，需确保与现有系统兼容。

影响：

- 用户：可访问 Cohere Transcribe 模型进行语音识别，扩展应用场景。
- 系统：增强多模态处理能力，验证近期编码器输入改进。
- 团队：提供模型集成范例，但需关注注册表设计，避免技术债。

关联脉络

此 PR 与 PR 35809 直接相关，作为后续启用模型。在 vLLM 近期 PR 中，多模态处理（如 PR 38018 涉及多模态处理器）和模型集成是常见主题，表明项目正扩展音频和视觉模型支持，本 PR 是这一趋势的一部分。