PR #38019 完整报告

vllm-project/vllm

[Model] Add Granite 4.0 1B speech to supported models

合并时间：2026-03-25 02:23

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38019

执行摘要

- 一句话：添加 Granite 4.0 1B speech 模型到 vLLM 支持列表，并提供测试覆盖。
- 推荐动作：建议工程师阅读此 PR 以了解如何在 vLLM 中添加新模型，特别是测试适配和文档更新模式。但变更相对简单，无需深度研究；可关注语言列表的潜在风险，考虑未来动态验证改进。

功能与动机

根据 PR body，'Granite 4.0 speech shared the same model architecture as Granite speech 3.3, add it to the list of supported models and add test coverage for it.' 目的是扩展模型支持范围，确保新模型能够被正确使用和测试，无需外部 issue。

实现拆解

实现方案分为四层：1. 文档层：在 docs/models/supported_models.md 中添加 ibm-granite/granite-4.0-1b-speech 到支持列表。2. 测试层：修改 tests/models/multimodal/generation/test_granite_speech.py，将模型列表改为字典，区分 3.3 模型需要 LORA 路径而 4.0 不需要，并更新 run_test 和 test_models 函数以处理可选的 audio_lora_path。3. 注册层：更新 tests/models/registry.py，在 GraniteSpeechForConditionalGeneration 的 extras 中添加新模型。4. 代码层：在 vllm/model_executor/models/granite_speech.py 的 ISO639_1_SUPPORTED_LANGS 静态列表中添加日语支持，尽管 review 中指出了兼容性问题。

关键文件：

- docs/models/supported_models.md（模块 文档）：更新支持模型文档，直接影响用户可见性和使用指南。
- tests/models/multimodal/generation/test_granite_speech.py（模块 测试）：关键测试文件，适配新模型的 LORA 路径条件，确保推理正确性。
- tests/models/registry.py（模块 测试）：更新模型注册信息，确保新模型在测试中被正确识别和调度。
- vllm/model_executor/models/granite_speech.py（模块 模型）：修改语言支持列表，存在潜在运行时错误风险，影响模型正确性。

关键符号：run_test, test_models

评论区精华

review 中，DarkLight1337 建议从文档中移除 8B 模型以保持表格简短，NickCao 确认更改。gemini-code-assist[bot] 指出语言列表变更可能引入 bug：'granite-speech-3.3 doesn't support Japanese, and granite-4.0 doesn't support Portuguese. Using an unsupported language for a model will pass the initial check against this list, but then cause a KeyError at runtime when accessing the tokenizer's lang_to_id map.' 并建议动态验证。DarkLight1337 回复 'No need to be too strict'，最终语言列表变更被保留。NickCao 询问评论相关性，表明讨论未完全解决。

- 文档表格简化建议 (style): NickCao 确认更改，但 patch 中未见移除，仅添加了新模型。
- 语言列表兼容性风险 (correctness): DarkLight1337 认为 'No need to be too strict'，变更被保留，风险未解决。

风险与影响

- 风险：主要风险集中在 vllm/model_executor/models/granite_speech.py 的语言列表变更：静态的 ISO639_1_SUPPORTED_LANGS 创建了不同模型语言支持的并集，若用户为 3.3 模型指定日语或为 4.0 模型指定葡萄牙语，可能触发运行时 KeyError，影响正确性。此外，测试覆盖是否充分处理了新模型的所有边界情况（如 LORA 路径条件），但 PR 提供了测试并通过，降低了回归风险。
- 影响：对用户：增加了一个可用的语音识别模型，扩展了多模态功能选择。对系统：添加新模型提升了灵活性，但需维护模型版本间差异（如 LORA 路径和语言支持），可能增加配置复杂性。对团队：变更范围小，涉及文档、测试和少量代码，日常维护负担有限。
- 风险标记：语言列表不一致，潜在运行时错误

关联脉络

- PR #36803 [Test] E2E Nemotron-3-Super tests: 类似添加新模型测试覆盖的 PR，都涉及扩展模型支持并更新测试配置。