

PR #48850 完整报告

vllm-project/vllm

[Bugfix][LoRA] Add embedding_modules for Qwen3.5 CausalLM

合并时间: 2026-08-19 12:55

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48850>

执行摘要

- 一句话: 为 Qwen3.5 注册 LoRA embedding_modules, 修复 embedding 目标适配器加载失败
- 推荐动作: 值得快速浏览的教科书式最小修复: 它直观展示了 vLLM LoRA 的 embedding_modules 契约如何参与 PEFT adapter 的加载校验。若后续开发新模型 LoRA 支持, 可直接复用该模式, 并建议顺手补一个仓库内单测防止再次遗漏。

功能与动机

Issue #48934 报告了 Qwen3.5 CausalLM 因缺少 embedding_modules 导致 embed_tokens / lm_head 的 LoRA 适配器加载失败; PR body 明确指出『Qwen3_5ForCausalLMBase implements SupportsLoRA but did not define embedding_modules. Loading a PEFT adapter whose target_modules include embed_tokens / lm_head fails unexpected-module validation in LoRAModel.from_local_checkpoint』, 且同系列 Qwen3 / Qwen3-MoE 可正常工作, 属于同族模型契约不一致的补齐。

实现拆解

本次变更共 1 个文件、+5/-0 行, 实现步骤拆解如下:

1. 变更入口: 在 vllm/model_executor/models/qwen3_5.py 的 Qwen3_5ForCausalLMBase 类体中新增 embedding_modules 类属性, 位置紧邻 packed_modules_mapping 声明, 位于类继承声明之后。
2. 契约内容: 映射键为 PEFT 适配器中常见的 LoRA 目标模块名, embed_tokens 映射到 vLLM 的输入 embedding 包装器 input_embeddings, lm_head 映射到输出 embedding 包装器 output_embeddings; 该映射逐字对齐 Qwen3ForCausalLM / Qwen3MoeForCausalLM 的既有实现。
3. 链路作用: Qwen3_5ForCausalLMBase 已实现 SupportsLoRA, LoRA 校验依赖该映射判定 adapter 的 target_modules 是否属于模型预期模块集合。此前缺失该映射时, from_local_checkpoint 会把 embed_tokens / lm_head 判为 unexpected-module 并抛 ValueError, 与 Qwen3 家族行为不一致。注册后 PEFT embedding LoRA 在 enable_lora=True 场景下可正常加载。

4. 测试与配套：本 PR 未新增仓库内单元测试，作者给出手动验证路径（`ruff check / ruff format --check` 通过，加载带 `embed_tokens / lm_head` 的 LoRA 确认不再抛错）；维护者触发 Buildkite CI #84536 在 head commit 7cb8e198 上运行通过，PR 带 verified 标签合入。

关键文件：

- `vllm/model_executor/models/qwen3_5.py`（模块 模型定义；类别 source；类型 data-contract；符号 `embedding_modules`, `Qwen3_5ForCausalLMBase`）：唯一变更文件：在 `Qwen3_5ForCausalLMBase` 类上新增 `embedding_modules` 类属性，把 PEFT 的 `embed_tokens / lm_head` LoRA 目标映射到 vLLM 的 `input_embeddings / output_embeddings` 包装器，修复 LoRA 加载时的 `unexpected-module` 校验失败。

关键符号：`embedding_modules`

关键源码片段

`vllm/model_executor/models/qwen3_5.py`

唯一变更文件：在 `Qwen3_5ForCausalLMBase` 类上新增 `embedding_modules` 类属性，把 PEFT 的 `embed_tokens / lm_head` LoRA 目标映射到 vLLM 的 `input_embeddings / output_embeddings` 包装器，修复 LoRA 加载时的 `unexpected-module` 校验失败。

```
class Qwen3_5ForCausalLMBase(
    nn.Module,
    HasInnerState,
    IsHybrid,
    SupportsEagle3,
    SupportsLoRA,
    SupportsMRoPE,
    SupportsPP,
):
    # 现有的融合投影映射，用于把 HF 权重名称映射到 vLLM 的融合层。
    packed_modules_mapping = {
        "qkv_proj": ["q_proj", "k_proj", "v_proj"],
        "gate_up_proj": ["gate_proj", "up_proj"],
        # GDN 融合投影。
        "in_proj_qkvz": ["in_proj_qkv", "in_proj_z"],
        "in_proj_ba": ["in_proj_b", "in_proj_a"],
    }

    # 本 PR 新增：把 PEFT adapter 的 embed / lm_head LoRA 目标名映射到
    # vLLM 的 embedding 包装器（输入嵌入 / 输出嵌入）。该映射与
    # Qwen3ForCausalLM / Qwen3MoeForCausalLM 完全一致。
    # 缺失时，`target_modules` 里的 embed_tokens / lm_head 会在
    # LoRAModel.from_local_checkpoint 的 unexpected-module 校验中被拒绝。
    embedding_modules = {
        "embed_tokens": "input_embeddings",
        "lm_head": "output_embeddings",
    }
```

评论区精华

本 PR 评审过程没有技术争论，核心讨论集中在流程层面：

- 作者 Agoni-02 为首次贡献者，因新贡献者门禁（0 个已合并 PR）无法触发 pre-run-check，在 PR 评论中请求维护者（sighingnow、vadiklyutiy、jeejeelee）添加 ready 标签。
- 维护者 jeejeelee 回复『Sorry for the delayed response』后直接批准，并以 /ci run 触发 Buildkite CI；CI 通过后带 **verified** 标签完成合并，无遗留未解决疑虑。
- 新贡献者 pre-run-check 门禁阻塞 (other)：维护者 jeejeelee 触发 /ci run，Buildkite CI #84536 在 head commit 7cb8e198 上运行。
- CI 验证与合并 (testing)：CI 通过，PR 合并完成，无技术性 review 争议。

风险与影响

- 风险：风险极低，但存在以下可注意的点：
- 缺少自动化测试覆盖：本 PR 唯一变更文件为源码文件，没有新增直接单测，验证依赖作者手动 e2e 与现有 LoRA 测试的间接覆盖；若后续 PEFT adapter 使用带前缀的命名（如 **model.embed_tokens**），仍可能触发同类校验问题。
- 兼容性：新增的是纯类级 dict，不改任何计算路径、权重加载或量化逻辑；enable_lora=False 或 adapter 不含 embedding 目标时行为完全不变，不存在回归风险。
- 数据契约一致性：映射与 Qwen3 逐字对齐，若未来 PEFT 侧命名（如 lm_head 替换为 output_embeddings）发生变化，需要同步更新。
- 性能 / 安全：无相关改动。
- 影响：影响面聚焦在 Qwen3.5 系列（CausalLM 基底）的 LoRA 用户：此前 target_modules 包含 embed_tokens / lm_head 的 PEFT 适配器在引擎初始化阶段直接失败；修复后与 Qwen3 / Qwen3-MoE 行为对齐，无需任何迁移或配置变更即可生效。对不使用 embedding LoRA 的用户无感知，对一个 5 行数据契约补齐而言影响范围与程度都较小。
- 风险标记：缺少自动化测试覆盖，LoRA 数据契约对齐

关联脉络

- PR #33234 [LoRA] Add embedding_modules for Qwen2/3（PR body 引用，具体标题未在材料中提供）：PR body 明确提到 Qwen2/3 家族的先前工作，为 embed_tokens / lm_head 注册 embedding_modules，本 PR 是同一模式在 Qwen3.5 上的补齐而非重复。
- PR #29816 [LoRA] Qwen embedding 支持前期工作（PR body 引用，具体标题未在材料中提供）：PR body 将 #29816 列为相关前置工作，说明 Qwen 家族 LoRA embedding 契约的演进脉络。