PR #38844 完整报告

vllm-project/vllm

[Gemma4][Bugfix]: Enable Gemma4ForCasualLM to load lora adapters correctly

合并时间：2026-04-11 17:06

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38844

执行摘要

此 PR 修复了 Gemma4ForCausalLM 模型加载 LoRA 适配器时因命名路径不一致导致的失败问题，通过添加权重映射器（hf_to_vllm_mapper）对齐键名，确保兼容性。影响范围仅限于 Gemma4 模型的 LoRA 功能，风险低且已通过测试验证，推荐相关工程师关注映射设计。

功能与动机

PR 旨在解决 Gemma4 模型中两个变体（Gemma4ForConditionalGeneration 和 Gemma4ForCausalLM）的模型路径命名差异。如 PR body 所述，Gemma4ForConditionalGeneration 使用 model.language_model.* 路径，而纯文本的 Gemma4ForCausalLM 使用 model.* 路径，这导致针对前者训练的 LoRA 适配器无法在后一种模型上正确加载。因此，需要引入映射机制来重命名 LoRA 键。

实现拆解

关键改动分为两个文件：

1. vllm/model_executor/models/gemma4.py：在 Gemma4ForCausalLM 类中添加 hf_to_vllm_mapper 属性，使用 WeightsMapper 定义映射规则。 - orig_to_new_prefix: 将 "model.language_model." 映射到 "model."。 - orig_to_new_substr: 处理 MoE 组件，如将 ".moe.experts.gate_up_proj" 映射到 ".moe.gate_up_proj"。
2. tests/lora/test_lora_checkpoints.py：新增两个测试函数：
   - test_gemma4_lora_weights_mapping: 验证普通层的映射。
   - test_gemma4_moe_lora_weights_mapping: 验证 MoE 层的映射。 代码示例（取自 gemma4.py）：

```
hf_to_vllm_mapper = WeightsMapper(
    orig_to_new_prefix={
        "model.language_model.": "model.",
    },
    orig_to_new_substr={
        ".moe.experts.gate_up_proj": ".moe.gate_up_proj",
        ".moe.experts.down_proj": ".moe.down_proj",
    },
)
```

评论区精华

review 中主要讨论围绕映射的完整性：

- *gemini-code-assist*[bot]：指出初始 *mapper* 可能缺少 MoE 组件映射，建议使用 orig_to_new_substr 更健壮地处理路径。
- ShubyM（作者）：回复解释 *per_expert_scale* 不是 LoRA 目标，因此无需映射，并已补充 MoE 映射。 最终，评审者 *jeejeelee* 批准 PR，表明讨论已解决。

风险与影响

- 风险：映射逻辑简单，但若未来模型结构变化或存在未覆盖路径，可能导致 LoRA 加载失败。测试覆盖了主要场景，但依赖映射正确性。
- 影响：直接影响使用 *Gemma4* 和 LoRA 的用户，修复加载问题提升兼容性；对系统其他部分无影响，属于模型特定优化。

关联脉络

从历史 PR 看，相关 PR 如 #39450（添加 *Gemma4 Eagle3* 支持）也涉及 *Gemma4* 模型修改，表明仓库正持续增强 *Gemma4* 功能。此 PR 作为 bug 修复，补全了 LoRA 支持链条，与近期模型功能演进趋势一致。