

PR #44828 完整报告

vllm-project/vllm

[BugFix] Use served model name in gemma4 audio-tower error message

合并时间: 2026-06-08 14:58

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44828>

执行摘要

- 一句话: 修复 Gemma4 音频错误消息暴露本地路径
- 推荐动作: 值得快速合入, 属于改善用户友好的小修复。可关注 `get_served_model_name` 函数在其他模型中的使用模式是否一致。

功能与动机

修复 Gemma4 无音频塔模型处理音频输入时的错误消息, 避免暴露内部本地路径 (如 `/mnt/local/model`), 改用用户配置的部署模型名 (如 `google/gemma-4-31B`)。关联 PR #41003。

实现拆解

1. 新增导入: 在 `vllm/model_executor/models/gemma4_mm.py` 中添加 `from vllm.config.model import get_served_model_name`。
2. 修改错误消息逻辑: 将 `validate_num_items` 方法中直接使用 `self.ctx.model_config.model` 改为先获取 `model_config`, 再调用 `get_served_model_name(model_config.model, model_config.served_model_name)` 生成展示用的模型名。
3. 测试: 初始提交包含一个测试文件变更, 但 reviewer 认为测试价值不大, 随后 `revert` 了测试提交, 最终仅保留源码修复。

关键文件:

- `vllm/model_executor/models/gemma4_mm.py` (模块 模型执行器; 类别 `source`; 类型 `data-contract`; 符号 `validate_num_items`): 唯一修改的文件, 修复错误消息中模型名显示, 新增导入并调整配置键访问逻辑。

关键符号: `validate_num_items`

关键源码片段

`vllm/model_executor/models/gemma4_mm.py`

唯一修改的文件, 修复错误消息中模型名显示, 新增导入并调整配置键访问逻辑。

```
# vllm/model_executor/models/gemma4_mm.py
```

```

# 新增的导入
from vllm.config.model import get_served_model_name

def validate_num_items(self, modality: str, num_items: int) -> None:
    if (
        modality == "audio"
        and num_items > 0
        and self.get_hf_config().audio_config is None
    ):
        # 之前直接使用 self.ctx.model_config.model, 可能暴露本地路径
        # 现在通过 get_served_model_name 生成用户友好的模型名
        model_config = self.ctx.model_config
        model = get_served_model_name(
            model_config.model, model_config.served_model_name
        )
        raise ValueError(
            f"Audio input was provided but the model "
            f"'{model}' does not have an audio tower. "
            f"Audio inference is only supported for Gemma4 "
            f"models that include an audio_config "
            f"(i.e., models that include an audio_config)."
        )
    super().validate_num_items(modality, num_items)

```

评论区精华

DarkLight1337 认为专门为这个错误消息写测试用处不大，建议只保留修复代码。作者 ll sj14 接受并删除了测试提交。

- 测试必要性 (testing): 作者接受并删除了测试提交，最终仅保留源码修复。

风险与影响

- 风险：低风险。变更极小，仅修改一行错误消息中的模型名来源，不影响功能逻辑。但需注意 `get_served_model_name` 的调用位置正确性，若 `served_model_name` 未设置，应能正确 fallback 到 `model` 本身。
- 影响：影响范围：仅影响 Gemma4 无音频塔模型在收到音频输入时抛出的 `ValueError` 消息内容。影响程度：低，属于用户体验优化类修复，不影响正常运行路径。
- 风险标记：极低风险

关联脉络

- PR #41003 Add Gemma4 model: 此 PR 引入的 Gemma4 模型支持，包含原有的错误消息；本 PR 修复了该错误消息中模型名的显示问题。