

PR #46341 完整报告

vllm-project/vllm

[Bugfix] Fix Llama4ForCausalLM initialization test failure

合并时间: 2026-06-22 16:40

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46341>

执行摘要

- 一句话: 修复 Llama4ForCausalLM 测试初始化失败
- 推荐动作: 值得立即合入。变更逻辑简单明确, 修复了测试基础设施中的 bug, 确保了三个 Llama4 变体的测试一致性。

功能与动机

`dummy_hf_overrides` 只对 `Llama4ForConditionalGeneration` 设置了 `num_experts_per_tok=1`, 导致 `Llama4ForCausalLM` 和 `EagleLlama4ForCausalLM` 获得 `num_experts_per_tok=2`, 这与 `apply_router_weight_on_input=True` 特性冲突 (该特性仅支持 `topk=1`), 造成测试初始化失败。

实现拆解

在 `tests/models/utils.py` 的 MoE 配置段中, 将原先的简单条件判断 `if model_arch == "Llama4ForConditionalGeneration"` 扩展为检查一个元组 `("Llama4ForConditionalGeneration", "Llama4ForCausalLM", "EagleLlama4ForCausalLM")`, 使得这三个 Llama4 变体类型都会将 `num_experts_per_tok` 设为 1, 从而与 `apply_router_weight_on_input` 约束一致。变更仅涉及该文件, 无其他依赖。

关键文件:

- `tests/models/utils.py` (模块 测试工具; 类别 test; 类型 test-coverage): 修复测试初始化中 MoE 参数配置的 bug, 是所有变更的载体。

关键符号: 未识别

关键源码片段

`tests/models/utils.py`

修复测试初始化中 MoE 参数配置的 bug, 是所有变更的载体。

```
# tests/models/utils.py (partial)
if model_arch_config.num_experts > 0:
    # 默认 num_experts_per_tok = 2, 但 Llama4 系列需要 topk=1
    # 因为 apply_router_weight_on_input 只支持 topk=1
    num_experts_per_tok = 2
    if model_arch in (
```

```
"Llama4ForConditionalGeneration",
"Llama4ForCausalLM",
"EagleLlama4ForCausalLM",
):
    num_experts_per_tok = 1
    update_dict.update(
        {
            "num_experts": num_experts,
            "num_experts_per_tok": num_experts_per_tok,
            # Kimi 使用 num_experts_per_token
            "num_experts_per_token": num_experts_per_tok,
            "num_local_experts": num_experts,
            # 否则不会有任何 expert 层
            "first_k_dense_replace": 0,
            # 避免 DeepSeek-V3 上的 OOM
            "n_routed_experts": num_experts,
        }
    )
```

评论区精华

无 review 评论。PR 由 reviewer DarkLight1337 直接批准，表明变更清晰无争议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅在测试配置辅助函数中，调整了 MoE 参数设置的条件分支，不影响任何生产逻辑。风险点在于若未来新增 Llama4 变体但忘记更新此元组，可能导致类似问题，但该模式已有提示。
- 影响：仅影响 tests/models/utils.py 中的测试初始化流程，确保 Llama4ForCausalLM 和 EagleLlama4ForCausalLM 的单元测试能正确运行。对系统性能、安全、兼容性无影响。
- 风险标记：仅测试文件，低风险

关联脉络

- PR #45993 [Model] Remove MiniMaxText01, MiniMaxVL01, MiniMaxForCausalLM: 同为 model 架构相关的清理 / 修复，风格相似。