PR #47797 完整报告

vllm-project/vllm

[Bugfix] Allocate HY V3 expert_bias in float32 to prevent silent downcasting

合并时间：2026-07-08 19:25

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47797

执行摘要

- 一句话：修复 HY V3 路由器 expert_bias 精度塌陷
- 推荐动作：值得合并的低成本高价值修复。建议在团队内推广此类模式：所有参与权重加载和路由决策的参数应显式指定 fp32，避免默认 dtype 引入精度塌陷。

功能与动机

Issue #47777 报告 HY V3 的 expert_bias 在 set_default_torch_dtype 为 bf16/fp16 环境下分配为 16-bit，而 checkpoint 存储的是 float32，load_weights 通过 param.data.copy_ 静默降精度。该偏差作为 e_score_correction_bias 参与 GroupedTopKRouter 的 top-k 选择，边界 token 的专家路由可能因此翻转，属于隐蔽的精度回归。

实现拆解

1. 修改 HYV3MoEFused 构造函数（vllm/model_executor/models/hy_v3.py 第 180 行）：将 expert_bias 的分配从 torch.empty(config.num_experts) 改为 torch.empty(config.num_experts, dtype=torch.float32)。
2. 移除调试测试文件：在 review 中 reviewer 要求删除 tests/models/test_hy_v3_expert_bias.py（该测试验证分配和 copy_ 行为），已在第二轮 commit 中清理。
3. 合并 main 分支：第三轮 commit 执行 Merge branch 'main' 以保持基线最新。

关键文件：

- vllm/model_executor/models/hy_v3.py（模块 模型执行；类别 source；类型 data-contract）：修复核心：将 expert_bias 分配 dtype 从隐式默认（torch.empty）改为显式 torch.float32，防止 checkpoint 加载时的静默降精度。

关键符号：HYV3MoEFused.init

评论区精华

reviewer jeejeelee 要求删除随附的单元测试文件 tests/models/test_hy_v3_expert_bias.py，作者在后续 commit 中移除了该文件。未对核心修复逻辑产生争议。

- 删除测试文件 (testing): 作者接受并在后续 commit 中移除了测试文件。

风险与影响

- 风险：变更极小（1 行 +3/-1），仅影响参数分配时的 dtype，无性能回归风险。由于显式指定 fp32 而非依赖默认 dtype，参数占用从 2 字节 / 元素变为 4 字节 / 元素，但 num_experts 通常为 192 量级，内存增加可忽略（~768 字节）。不涉及运行时逻辑或 load/save 兼容性。
- 影响：直接修复 HY V3 模型的 MoE 路由精度，避免边界 token 的专家选择错误。影响范围限于 hy_v3 和 hy_v3_mtp 模型，无用户感知 API 变化。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR