PR #43770 完整报告

vllm-project/vllm

[Bugfix] fix wrong partial_rotary_factor calculation for bailing_moe model.

合并时间：2026-06-01 17:42

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43770

执行摘要

- 一句话：修复 Bailing MoE 模型中 partial_rotary_factor 计算错误
- 推荐动作：值得精读，以了解模型配置优先级处理的常见模式。关注点是：优先使用显式字段（rotary_dim），其次使用派生字段（partial_rotary_factor），最后使用默认回退（head_dim）。该模式可推广到其他模型实现。

功能与动机

Ling-flash-2.0 和 AntAngelMed 等模型的 config 中只有 partial_rotary_factor=0.5 而没有 rotary_dim，旧代码直接回退到 head_dim 并计算出 partial_rotary_factor=1，导致 Ascend NPU（vllm-ascend）上 _cos_sin_cache 初始化尺寸错误，引发异常。详见 PR body 中的配置链接和异常截图。

实现拆解

1. 修改回退优先级：在 vllm/model_executor/models/bailing_moe.py 的 __init__ 中，将 rotary_dim = getattr(config, "rotary_dim", self.head_dim) 改为先尝试读取 rotary_dim，若为 None 则从 config.partial_rotary_factor 计算 rotary_dim。
2. 保留防御性回退：即使上述计算后 rotary_dim 仍为 None，则回退到 self.head_dim，确保极端配置下的鲁棒性。
3. 更新 rope_parameters：无论哪种路径，最后仍按 rotary_dim / self.head_dim 设置 config.rope_parameters['partial_rotary_factor']，保持与下游 get_rope 的接口一致。
4. 无其他文件变更：本次修改仅涉及单文件 6 行新增、1 行删除，没有测试、配置或部署配套改动。

关键文件：

- vllm/model_executor/models/bailing_moe.py（模块 模型执行器；类别 source；类型 data-contract）：修复了 rotary_dim 计算逻辑，是本次变更的唯一文件。

关键符号：未识别

关键源码片段

vllm/model_executor/models/bailing_moe.py

修复了 rotary_dim 计算逻辑，是本次变更的唯一文件。

```
# 位于 __init__ 方法中，构建 rotary embedding 之前
# 原逻辑：rotary_dim = getattr(config, "rotary_dim", self.head_dim)
# 导致当 config 没有 rotary_dim 时，partial_rotary_factor 总是被设为 1.0

# 新逻辑：优先使用 config 中的 rotary_dim
rotary_dim = getattr(config, "rotary_dim", None)

# 如果 config 没有 rotary_dim，则从 partial_rotary_factor 计算
if rotary_dim is None:
    partial_rotary_factor = getattr(config, "partial_rotary_factor", 1.0)
    rotary_dim = int(self.head_dim * partial_rotary_factor)

# 防御性回退：如果仍未定义，则默认使用 head_dim
if rotary_dim is None:
    rotary_dim = self.head_dim

# 更新 rope_parameters，与下游 get_rope 接口保持一致
config.rope_parameters["partial_rotary_factor"] = rotary_dim / self.head_dim

self.rotary_emb = get_rope(
    self.head_dim,
    max_position=config.max_position_embeddings,
    rope_parameters=config.rope_parameters,
    is_neox_style=True,
)
```

评论区精华

xianbaoqian 在 review 中指出第 137 行的 `if rotary_dim is None:` 检查是死代码（因为上一行 `int(self.head_dim * partial_rotary_factor)` 不会产生 None），作者 zzt93 回应这是防御性代码以防 config 未来变化导致异常，并同意必要时可删除。最终保留该行，但未产生实际影响。

- 死代码检查 (correctness): 保留该防御性代码，但实际不会被执行。

风险与影响

- 风险：风险极低：改动仅影响 rotary_dim 的计算逻辑，且优先读取 config 中明确设置的字段，兼容原有配置（如 Ring-1T 同时设置了 rotary_dim）。防御性回退保证了极端情况不会崩溃。但缺少测试覆盖可能让未来重构时忽略此行为。
- 影响：直接影响：修复了 Ling-flash-2.0 和 AntAngelMed 等模型在 Ascend NPU 上的启动失败。影响范围限于 Bailing MoE 模型，不影响其他模型或通用路径。对已有配置的 Ring-1T 等模型无副作用。
- 风险标记：缺少测试覆盖

关联脉络

- PR #42944 fix: glm5.1 pp model loading: 同为模型加载中的配置解析 bug 修复，可比较不同模型的配置处理模式。