PR #26187 完整报告

sgl-project/sglang

Wire YARN rope_parameters through LFM2 and LFM2-MoE attention

合并时间：2026-05-27 07:00

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26187

执行摘要

- 一句话：修复 LFM2 模型 YARN RoPE 参数未正确传递的问题
- 推荐动作：此 PR 值得合并，修复了一个 silent correctness bug，改动量极小且正确性有验证数据支撑。推荐精读以理解类似配置兼容性问题的处理模式（优先新键、安全 fallback），这一模式已在多个模型（如 Qwen3）中复用。

功能与动机

Transformers v5 的 LFM2 配置将 YARN 缩放参数（rope_type、factor、original_max_position_embeddings）放在 rope_parameters 键下，而旧版 v4 使用 rope_scaling。Lfm2MoeAttention 读取的是 rope_scaling，导致 YARN 配置被静默丢弃，生成结果在超过 original_max_position_embeddings 时与 HF 不一致。Lfm2Attention 的 fallback 直接解引用 config.rope_parameters["rope_theta"]，在缺少 rope_parameters 的 v4 配置中会触发 TypeError。

实现拆解

1. python/sglang/srt/models/lfm2.py：在 Lfm2Attention.__init__ 中，将 rope_theta 的 fallback 从直接索引 config.rope_parameters["rope_theta"] 改为安全的 getattr(config, "rope_theta", 1000000.0)；将 rope_scaling 参数从直接使用 config.rope_parameters 改为 rope_parameters or getattr(config, "rope_scaling", None)，优先使用 rope_parameters 字典，兼容旧版 rope_scaling。
2. python/sglang/srt/models/lfm2_moe.py：在 Lfm2MoeAttention.__init__ 中，将 rope_scaling 参数从 getattr(config, "rope_scaling", None) 改为 rope_parameters or getattr(config, "rope_scaling", None)，使 MoE 变体也能正确接收 YARN 参数。
3. 两处修改均只有 1-2 行，但实现了配置键的兼容性降级，行为与 qwen3_moe.py 中的处理模式一致。
4. 测试：PR 提供了验证用的 YARN 配置权重（dense 和 MoE），并报告 ROUGE=1.0 且 prefill 差异极小（dense <0.02, MoE 0.04-0.30），证明修复后 SGLang 与 HF 输出一致。但本次提交未包含自动化测试。

关键文件：

- python/sglang/srt/models/lfm2.py（模块 模型定义；类别 source；类型 data-contract）：核心修复文件之一，修复了 Lfm2Attention 中 rope_theta 和 rope_scaling 的取值逻辑，

兼容 v4/v5 配置。

- python/sglang/srt/models/lfm2_moe.py（模块 模型定义；类别 source；类型 data-contract）：核心修复文件之二，修复了 Lfm2MoeAttention 中 rope_scaling 参数未读取 rope_parameters 的问题，使 MoE 变体也能应用 YARN。

关键符号：未识别

关键源码片段

python/sglang/srt/models/lfm2.py

核心修复文件之一，修复了 Lfm2Attention 中 rope_theta 和 rope_scaling 的取值逻辑，兼容 v4/v5 配置。

```
# lfm2.py 中 Lfm2Attention.__init__ 的 RoPE 参数获取逻辑（变更后）
rope_parameters = getattr(config, "rope_parameters", None)
if rope_parameters is not None and "rope_theta" in rope_parameters:
    rope_theta = rope_parameters["rope_theta"]
else:
    # 安全 fallback：v4 配置使用顶层 rope_theta；或默认 1000000.0
    rope_theta = getattr(config, "rope_theta", 1000000.0)

self.rotary_emb = get_rope(
    head_size=self.head_dim,
    rotary_dim=self.head_dim,
    max_position=getattr(config, "max_position_embeddings", 8192),
    # 优先使用 rope_parameters 字典（包含 YARN 完整设置），
    # 若不存在则回退到旧版 rope_scaling（v4 配置）
    rope_scaling=rope_parameters or getattr(config, "rope_scaling", None),
    base=rope_theta,
    is_neox_style=True,
    dtype=torch.get_default_dtype(),
)
```

python/sglang/srt/models/lfm2_moe.py

核心修复文件之二，修复了 Lfm2MoeAttention 中 rope_scaling 参数未读取 rope_parameters 的问题，使 MoE 变体也能应用 YARN。

```
# lfm2_moe.py 中 Lfm2MoeAttention.__init__ 的 RoPE 参数获取逻辑（变更后）
rope_parameters = getattr(config, "rope_parameters", None)
if rope_parameters is not None and "rope_theta" in rope_parameters:
    rope_theta = rope_parameters["rope_theta"]
else:
    rope_theta = getattr(config, "rope_theta", 1000000.0)

self.rotary_emb = get_rope(
    head_size=self.head_dim,
    rotary_dim=self.head_dim,
    max_position=getattr(config, "max_position_embeddings", 128000),
    # 关键变更：从直接使用 config.rope_scaling 改为优先使用
```

```
        # rope_parameters（v5 YARN），否则回退到 rope_scaling（v4）
        rope_scaling=rope_parameters or getattr(config, "rope_scaling", None),
        base=rope_theta,
        is_neox_style=True,
        dtype=torch.get_default_dtype(),
    )
```

评论区精华

PR 无人工 review 评论，仅有自动化 bot 的总结性评论和合并者的 /tag-and-rerun-ci 批准操作。因此无争议或设计权衡讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及两个文件中各 1-2 行配置读取逻辑，且采用了安全的 getattr fallback，不会引入新异常。对于不包含 YARN 参数的旧配置，行为与之前一致（使用默认 rope_theta 和 None rope_scaling）。唯一可能的风险是如果用户显式设置 rope_parameters 但其中缺少 rope_theta，则会使用默认值 1000000.0，但这种情况在 Transformers v5 规范中不存在。
- 影响：影响范围限于 LFM2 系列模型（Dense 和 MoE 变体），且仅影响需要 YARN 缩放的配置。对于使用旧版 rope_scaling 的配置无影响。修复后 SGLang 能正确执行 YARN，与 HF 保持 bit-exact 一致性，确保长上下文推理的正确性。无其他系统或用户影响。
- 风险标记：暂无

关联脉络

- PR #26132 Sgl flashmla: 同属模型层 RoPE/MLA 相关优化，但无直接关联。
- PR #26335 [Spec] Async-assert probes across EAGLE/MTP; zero tgt_cache_loc: 同属 speculative decoding 及模型推理优化，但无直接关联。