PR #21004 完整报告

sgl-project/sglang

[Fix] Add EPLB rebalance support for Kimi K2.5

合并时间：2026-03-26 12:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21004

执行摘要

该 PR 修复了 Kimi K2.5 模型在启用 EPLB（Expert Parallel Load Balancing）负载均衡时因缺少 routed_experts_weights_of_layer 属性而导致的 AttributeError。通过在 KimiK25ForConditionalGeneration 类中添加该属性，将访问委托给底层语言模型，确保 EPLB 管理器能正确获取专家权重信息，解决了服务器崩溃问题。变更范围极小（仅 4 行代码），风险较低。

功能与动机

根据 PR body 中的错误日志，当启用 EPLB 负载均衡时，Kimi K2.5 模型在调度过程中抛出以下错误：

```
AttributeError: 'KimiK25ForConditionalGeneration' object has no attribute 'routed_experts_
weights_of_layer'
```

该错误发生在 EPLB 管理器尝试访问模型属性以进行负载均衡时（具体在 eplb_manager.py 的 _compute_update_layer_ids_chunks 方法中），导致服务器崩溃。PR 的目标是添加缺失的属性，使 EPLB 功能对 Kimi K2.5 模型正常工作。

实现拆解

仅修改一个文件：python/sglang/srt/models/kimi_k25.py。

在 KimiK25ForConditionalGeneration 类中新增一个只读属性：

```
@property
def routed_experts_weights_of_layer(self):
    return self.language_model._routed_experts_weights_of_layer.value
```

该属性将访问委托给底层的语言模型（self.language_model），返回其 _routed_experts_weights_of_layer.value，从而提供 EPLB 管理器所需的专家权重信息。

评论区精华

review 中仅有一条来自 gemini-code-assist[bot] 的代码风格建议：

> 为保持与类中其他属性（如 start_layer 和 end_layer）的一致性并提高代码清晰度，请添加返回类型提示。由于此属性似乎返回字典，-> dict 将是合适的类型提示。

该建议未被采纳（最终代码未添加类型提示），但 reviewer yeahdongcn 已批准 PR。

风险与影响

风险分析：

- 变更范围极小，仅添加属性委托，不涉及核心逻辑修改，风险较低。
- 潜在风险：如果底层语言模型的 _routed_experts_weights_of_layer.value 结构不符合预期（例如非字典类型），可能导致后续 EPLB 逻辑错误，但该风险在现有代码中已存在。
- 缺少类型提示可能影响代码可读性，但不会影响运行时行为。

影响分析：

- 对用户：修复了 Kimi K2.5 模型在启用 EPLB 负载均衡时的服务器崩溃问题，使负载均衡功能恢复正常。
- 对系统：确保 EPLB 管理器能正确访问专家权重信息，优化多专家模型（MoE）的负载分布。
- 对团队：解决了特定模型配置下的阻塞性问题，支持了 Kimi K2.5 模型在 EPLB 环境下的稳定运行。

关联脉络

该 PR 是 Kimi K2.5 模型支持系列的一部分：

- PR #22269 为 Kimi K2.5 添加了 Encoder-Prefill-Decode (EPD) 解耦支持，扩展了多模态推理架构。
- PR #22381 为 Kimi 模型添加了 LoRA 支持，并优化了量化 MoE 兼容性。

本 PR 修复了这些功能在 EPLB 负载均衡下的一个关键缺失属性，体现了 sglang 项目在持续扩展对特定模型（尤其是 MoE 模型）的深度支持，特别是在负载均衡和并行化方面的优化。