PR #21752 完整报告

sgl-project/sglang

Fix kimi-linear launch server error

合并时间：2026-03-31 21:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21752

执行摘要

该 PR 修复了 kimi-linear 模型在启动服务器时因 ModelConfig 对象缺少 scaling 属性而导致的 AttributeError，通过添加 scaling 计算并处理 rope_scaling，解决了启动失败问题，确保了 MLA-based 模型的正确初始化。

功能与动机

由于 self.scaling 属性被删除，kimi-linear 模型在启动时抛出 AttributeError，错误堆栈显示在 attention 后端初始化时无法访问 scaling（具体见 PR body 中的日志）。PR 旨在恢复 scaling 属性以支持模型启动，避免服务器崩溃。

实现拆解

修改了 python/sglang/srt/configs/model_config.py 中的 _derive_model_shapes 方法：

- 添加基础计算：self.scaling = 1 / math.sqrt(self.qk_nope_head_dim + self.qk_rope_head_dim)
- 条件判断：如果 hf_config.rope_scaling 存在，则调用 compute_mla_mscale_scaling 函数调整 scaling 值，以支持长上下文模型的 rope scaling 功能。

评论区精华

review 中，gemini-code-assist[bot] 指出：

> "MLA-based 模型往往在使用 rope_scaling（如 Yarn）时需要调整 scaling 因子；同时，其他 MLA 架构如 KimiVLForConditionalGeneration、DeepseekVL2ForCausalLM 和 MiniCPM3ForCausalLM 也可能缺少 scaling 属性。"

作者 yuan-luo 回复 "Fixed."，表明已处理 rope_scaling 调整，但未回应其他架构问题，留下潜在维护点。

风险与影响

- 风险： 1) rope_scaling 处理不当可能影响长上下文模型性能； 2) 其他 MLA 架构可能仍缺失 scaling 属性，导致未来启动错误； 3) 缺少测试覆盖，可能遗漏边缘情况。
- 影响：修复了 kimi-linear 启动问题，提升系统稳定性；团队需检查类似架构并加强配置测试。

关联脉络

未关联到具体 Issue；历史 PR 中无直接修改相同文件或模型的 PR，但近期 PR 如 21657（DeepSeek 模型优化）和 21727（多模态 bugfix）表明团队持续优化模型兼容性，本 PR 是 MLA 架构配置修复的一部分。