PR #26506 完整报告

sgl-project/sglang

[spec decoding] support kimi-k2.6-eagle3.1-mla draft

合并时间：2026-05-29 05:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26506

执行摘要

- 一句话：支持 Kimi-K2.6 EAGLE3.1-MLA 草稿模型
- 推荐动作：值得精读，尤其是对 speculative decoding 和模型加载兼容性设计感兴趣的人。两个配置标志的默认值设计和遗留标志兼容做法值得学习。

功能与动机

需要支持 Kimi-K2.6 系列中新增的 EAGLE3.1 草稿变体（例如 lightseekorg/kimi-k2.6-eagle3.1-mla），它在保持一致的单层 MLA 布局基础上引入了两个可选配置标志：fc_norm 和 norm_output。对应 MHA EAGLE3 草稿的类似变更已在 #24663 中落地。

实现拆解

该 PR 仅修改一个文件 python/sglang/srt/models/kimi_k25_eagle3.py，实现以下步骤：

1. 更新模块文档字符串：将描述从 'EAGLE3 draft model with MLA attention for Kimi-K2.5' 改为 'EAGLE3 / EAGLE3.1 draft model ... for Kimi-K2.x'，并记录两个新标志。
2. 在 __init__ 中添加 fc_norm 支持：将局部变量 num_fc_input 改为实例属性 self.num_aux_hidden_states；通过 config.fc_norm 或遗留的 config.use_aux_norm 标志控制，若启用则创建一个包含 num_aux_hidden_states 个 RMSNorm 的 ModuleList，否则置为 None。
3. 在 __init__ 中添加 norm_output 支持：从配置中读取 config.norm_output，默认 False（保持原有行为：aux 输出为预归一化隐藏状态）。
4. 修改前向传播逻辑：在调用 self.fc 投影前，若 self.fc_norm 不为 None，则将隐藏状态按 num_aux_hidden_states 分块，对每块应用对应的归一化后再拼接。
5. 调整 aux 输出：根据 self.norm_output 决定 aux 输出是 post-norm（hidden_states_to_logits）还是 pre-norm（hidden_states_to_aux），保持向后兼容。

以上所有修改均以零额外依赖方式实现，未引入新的测试文件。

关键文件：

- python/sglang/srt/models/kimi_k25_eagle3.py（模块 模型定义；类别 source；类型 core-logic）：唯一变更文件，实现 EAGLE3.1 MLA 草稿支持，包含所有关键逻辑。

关键符号：未识别

评论区精华

PR 仅有来自 Qiaolin-Yu 的一次批准，无 review 评论或讨论线程。审核简洁且无争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：
  - 向后兼容性风险：两个标志默认关闭，现有 EAGLE3 检查点和运行路径完全不受影响。
  - 回归风险：仅在 fc_norm 启用时改变计算图，该分支是新增的，不会影响已有路径。
  - 潜在风险：若目标模型配置包含未知的 fc_norm 或 norm_output 值，可能加载失败；但 getattr 默认值处理可缓解。
  - 影响：影响范围窄：仅影响 EAGLE3/3.1 MLA 草稿模型的加载和推理路径。
  - 用户影响：Kimi-K2.6 用户可直接加载新草稿检查点，无需额外配置。
  - 系统影响：无基础设施变更，无性能退化。
  - 团队影响：低，修改集中在已有文件中。
  - 风险标记：低变更量，仅单文件修改，向后兼容，缺少测试覆盖

关联脉络

- PR #24663 [spec decoding] support eagle3.1 (fc_norm + norm_output) for MHA draft: 本 PR 是 MHA 版本（#24663）的 MLA 镜像实现，遵循相同设计。