PR #21620 完整报告

sgl-project/sglang

fix: Mistral Small 4 fails to start due to config/weight format mismatch

合并时间：2026-03-30 16:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21620

执行摘要

此 PR 修复了 Mistral Small 4 模型启动失败的问题，通过调整模型格式检测逻辑，确保正确加载权重，同时维护了向后兼容性。影响范围限于特定 Mistral 模型，新增测试保障修复效果。

功能与动机

源于 Issue 21611，用户报告使用 Mistral Small 4 模型启动时出现 AttributeError，因为权重 w_kc 为 None。PR body 详细解释了根因：模型同时包含 params.json 和 config.json 文件，导致 _is_mistral_native_format 函数误判为 HF 格式，但配置解析期望原生格式，权重名不匹配，所有权重被跳过。目标是解决这一格式冲突，使模型正常启动。

实现拆解

- 核心修改：在 python/sglang/srt/server_args.py 的 _is_mistral_native_format 函数中，添加特定模型模式检查（mistral-large-3、mistral-small-4、leanstral）。当同时存在 params.json 和 config.json 时，若模型路径名匹配这些模式，则返回 True 以使用 Mistral 原生格式。
- 代码重构：提取 _check_format 辅助函数，减少本地和 Hub 检测逻辑的重复，提高可维护性。
- 测试添加：新增 test/registered/models/test_ministral4_models.py，包含文本（GSM8K）和多模态（MMMU）测试，使用 TP=2 和 --trust-remote-code 参数验证模型启动和推理。

评论区精华

Review 中，gemini-code-assist[bot]指出：

> 本地目录和 hub 模型的检测逻辑存在显著代码重复，可以重构为辅助函数以提高可维护性。

这一建议被采纳，在后续 commit 中重构出 _check_format 函数。Fridge003批准了修改，无其他争议。

风险与影响

- 风险：低风险。修改通过白名单控制，仅影响特定模型模式；正则表达式匹配可能误判，但已限定在已知模式；新增测试覆盖，减少了回归可能性。兼容性已验证，确保如 Mistral-7B-v0.3 等其他模型不受影响。
- 影响：用户现在可以正常启动 Mistral Small 4 模型；系统加载逻辑微调，无性能或安全影响；团队增加了测试用例，有助于预防类似问题。

关联脉络

- 关联 Issue：Issue 21611 直接触发了此修复，描述了 Mistral Small 4 启动失败的具体错误。
- 关联 PR：PR 20621 同样修改了 server_args.py，涉及服务器参数处理；PR 21448 涉及模型加载 bug 修复，显示模型格式处理是仓库中的常见维护点。这些 PR 共同反映了 SGLang 在模型兼容性和加载逻辑上的持续优化。