PR #21347 完整报告

sgl-project/sglang

[Bugfix] Fix PP tied embeddings weight loading for qwen3.5 4B dense model

合并时间：2026-04-01 14:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21347

执行摘要

本 PR 修复了 Qwen3.5-4B dense 模型在并行处理（PP）大于 1 时因权重加载错误导致的输出乱码问题，通过添加缺失的 tie_word_embeddings 处理逻辑，确保了模型在分布式环境下的正确性。

功能与动机

动机源于 issue #21093，用户报告 Qwen3.5-4B dense 模型在 PP=2 时输出不正确。根因是模型类 Qwen3_5ForConditionalGeneration 和 Qwen3_5MoeForConditionalGeneration 覆盖了父类 Qwen3VLForConditionalGeneration 的 load_weights 方法但遗漏了 tie_word_embeddings 权重复制，导致在 PP>1 时 lm_head 权重未初始化，直接产生垃圾输出。

实现拆解

修改集中在 python/sglang/srt/models/qwen3_5.py 文件。在 load_weights 和 load_fused_expert_weights 方法中，添加了以下逻辑： if (self.config.tie_word_embeddings and self.pp_group.is_last_rank and " model.embed_tokens.weight" in name): if "lm_head.weight" in params_dict: lm_head_param = params_dict["lm_head.weight"] weight_loader = getattr(lm_head_param, "weight_loader", default_weight_loader) weight_loader(lm_head_param, loaded_weight) 这确保了当 tie_word_embeddings 为 true 时，最后一个 rank 的 lm_head 权重从 embed_tokens 复制，与父类模式一致。

评论区精华

Review 中主要讨论点：

- 检查顺序优化：ShangmingCai 建议“先检查 self.config.tie_word_embeddings 以实现早期退出”，作者采纳并更新代码，提升了条件判断效率。
- MoE 模型扩展：yuan-luo 指出“MoE 模型的 load_weights 也可能需要相同修复”，作者回应当前 MoE 检查点 tie_word_embeddings 为 false，但添加了防御性守卫以确保未来兼容性。

风险与影响

风险较低：修改逻辑与父类一致，但缺少单元测试可能隐藏回归问题。MoE 模型的防御性守卫处理了边缘情况，降低了风险。影响方面，修复了模型在 PP 部署下的输出正确性，提升了用户体验，影响范围限于使用 Qwen3.5 dense 模型且 PP>1 的场景。

关联脉络

与此 PR 相关的历史 PR 包括 #17122（同为模型 bugfix，展示了类似的设计模式），以及 issue 中提及的 PR #21070（修复了同一模型的 PP 分割 OOM 问题，但未解决权重加载）。这表明在大型语言模型集成中，权重加载和分布式处理是需要持续关注的关键区域。