PR #23434 完整报告

sgl-project/sglang

[Model] Qwen3ForPooledOutput: forward get_input_embeddings to inner model

合并时间：2026-04-30 03:25

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23434

执行摘要

- 一句话：修复 Qwen3PooledOutput 缺少 get_input_embeddings
- 推荐动作：推荐合并，这是一个简单的修复，应尽快合入以解除 score API 对 Qwen3 分类模型的阻塞。

功能与动机

修复 score API 中 embedding 覆盖注入路径的 AttributeError：当调度器解析覆盖占位符 token 为实际 embedding 张量时，调用 model.get_input_embeddings()(input_ids)，该调用在 Qwen3ForCausalLM 路径正常，但对 Qwen3ForSequenceClassification 等 pooled-output 变体失败。

实现拆解

1. 在基类 Qwen3ForPooledOutput 中新增 get_input_embeddings 方法（文件：python/sglang/srt/models/qwen3_classification.py，第 58-59 行），直接转发到 self.model.get_input_embeddings()。
2. 利用继承自动覆盖：Qwen3ForSequenceClassification 和 reward 模型变体无需额外修改，自动继承基类的新方法。
3. 无测试配套变更：本 PR 仅 3 行新增，且无新增测试文件，但 PR body 的 checklist 中勾选了“添加单元测试”，可能为疏忽或已在其他 PR 中覆盖。

关键文件：

- python/sglang/srt/models/qwen3_classification.py（模块 模型；类别 source；类型 data-contract；符号 get_input_embeddings）：修改了基类 Qwen3ForPooledOutput，添加 get_input_embeddings 转发方法，是本次变更唯一的文件。

关键符号：get_input_embeddings

评论区精华

无 review 评论，审核者 Qiaolin-Yu 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅新增一个转发方法，不修改任何现有逻辑，不会引入回归。但缺少直接单元测试覆盖该转发路径。
- 影响：影响范围：仅影响 Qwen3 的 pooled-output 变体（seq-cls 和 reward）。用户影响：修复了使用 score API 加载 Qwen3-Reranker-seq-cls 等模型时的 AttributeError。系统影响：无性能或兼容性影响。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR