PR #28649 完整报告

sgl-project/sglang

Pass quant_config to attention gate projection

合并时间：2026-06-18 20:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28649

执行摘要

- 一句话：修复 Laguna 注意力门控投影未传递量化配置
- 推荐动作：值得信任。改动简单明确，风险极低，合并后可提升 Laguna 模型在量化场景下的正确性与性能。

功能与动机

在 Laguna 模型的注意力层中，qkv_proj、o_proj 和 attn 都使用了传入的 quant_config，但门控投影 g_proj 却硬编码为 None，导致门控投影无法正确应用量化配置，可能引发精度下降或推理性能损失。该 PR 旨在修复这个不一致性。

实现拆解

在 python/sglang/srt/models/laguna.py 的注意力层 __init__ 方法中，将门控投影 g_proj 的 quant_config 参数从 None 改为 quant_config（即外部传入的量化配置），使门控投影与同一层其他的线性投影层（qkv_proj、o_proj、attn）使用相同的量化策略。

关键文件：

- python/sglang/srt/models/laguna.py（模块 模型实现；类别 source；类型 data-contract）：修复了门控投影量化配置未传递的 bug，属于核心模型逻辑的修正。

关键符号：未识别

关键源码片段

python/sglang/srt/models/laguna.py

修复了门控投影量化配置未传递的 bug，属于核心模型逻辑的修正。

```
# python/sglang/srt/models/laguna.py
# 注意力层 __init__ 中 g_proj 的创建部分
if self.gating:
    g_proj_dim = (
        self.total_num_heads
        if self.gate_per_head
        else self.total_num_heads * self.head_dim
    )
    self.g_proj = ColumnParallelLinear(
```

```
            hidden_size,
            g_proj_dim,
            bias=False,
            gather_output=False,
            # 修复前为 quant_config=None，现在传递外部 quant_config
            # 使其与 qkv_proj、o_proj、attn 等组件使用一致的量化配置
            quant_config=quant_config,
            tp_rank=attn_tp_rank,
            tp_size=attn_tp_size,
            prefix=add_prefix("g_proj", prefix),
        )
    else:
        self.g_proj = None
```

评论区精华

该 PR 没有人工 review 评论，仅 Gemini Code Assist 机器人自动评论确认了变更内容，没有提出异议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。改动仅涉及一行参数传递，将硬编码的 None 改为变量引用。如果 quant_config 为 None，行为与之前完全一致；如果 quant_config 非 None，则门控投影会应用量化，这可能是之前遗漏的功能，修复后能提升模型精度或性能。不会引入回归风险。
- 影响：仅影响使用 Laguna 模型且启用量化的用户。修复后门控投影将正确量化，可能带来精度提升（尤其在低比特量化场景下）。对于不使用量化或非 Laguna 模型的用户无影响。
- 风险标记：暂无

关联脉络

- PR #28604 [Fix] don't force hybrid-SWA when sliding_window is disabled: 同样修改了 Laguna 模型的配置文件，属于同一条功能线的 bugfix。