PR #28604 完整报告

sgl-project/sglang

[Fix] don't force hybrid-SWA when sliding_window is disabled

合并时间：2026-06-18 13:11

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28604

执行摘要

- 一句话：修复 Laguna 在 sliding_window=0 时错误启用 hybrid-SWA
- 推荐动作：值得合并。该 PR 修复了一个影响 Laguna 模型在无滑动窗口下运行的 bug，逻辑清晰，风险低。建议阅读 is_hybrid_swa_model 中的新增判断和 laguna.py 中默认层类型的变更，以确保理解 hybrid-SWA 识别的边界条件。

功能与动机

当使用 sliding_window=0 的 Laguna 模型配置时，is_hybrid_swa_model() 函数仍返回 True，导致 SGLang 尝试混合全注意力与滑动窗口注意力，而模型实际没有滑动窗口层，引发下游错误（如索引越界或维度不匹配）。PR 在 Body 中未详细描述，但从源码变更可知目的是修复这一逻辑错误。

实现拆解

1. 修改 model_config.py 中的 is_hybrid_swa_model() 函数：在为 LagunaForCausalLM 返回 True 之前，增加条件判断——如果 hf_text_config.sliding_window 为 0 或 falsy（且 hf_text_config 非 None），则返回 False。该优先于模型架构名称匹配，确保无滑动窗口时不误判。
2. 修改 laguna.py 中的 LagunaConfig 构造函数：当 layer_types 参数未提供时，原默认逻辑生成长度为 num_hidden_layers 的交替序列（每隔 4 层 full_attention，其余 sliding_attention）。现改为全部填充 "full_attention"，因为如果 sliding_window 为 0，模型不应包含任何滑动注意力层。该默认值变更与 is_hybrid_swa_model 的修复共同作用，避免无滑动窗口时产生不一致的层类型配置。
3. 影响范围：仅影响 LagunaForCausalLM 模型。其他 hybrid-SWA 模型（如 Llama4、DeepseekV4、Gemma4 等）不受影响。

关键文件：

- python/sglang/srt/configs/model_config.py（模块 模型配置；类别 source；类型 data-contract；符号 is_hybrid_swa_model）：修复 is_hybrid_swa_model() 函数，增加对 LagunaForCausalLM 且 sliding_window=0 时的显式返回 False 逻辑。
- python/sglang/srt/configs/laguna.py（模块 模型配置；类别 source；类型 core-logic；符号 LagunaConfig.init）：修改默认 layer_types 合成逻辑，当未提供 layer_types 时，不再交替生成滑动注意力层，而是全部设为 full_attention。

关键符号：is_hybrid_swa_model, LagunaConfig.init

关键源码片段

python/sglang/srt/configs/model_config.py

修复 is_hybrid_swa_model() 函数，增加对 LagunaForCausalLM 且 sliding_window=0 时的显式返回 False 逻辑。

```
def is_hybrid_swa_model(
    model_architectures: List[str],
    hf_text_config: Optional[PretrainedConfig] = None,
):
    hybrid_swa_archs = {
        "Llama4ForConditionalGeneration",
        "DeepseekV4ForCausalLM",
        "DeepseekV4ForCausalLMNextN",
        "GptOssForCausalLM",
        *MIMO_V2_MODEL_ARCHS,
        "MiMoV2MTP",
        "Step3p5ForCausalLM",
        "Step3p5MTP",
        "Step3p7ForConditionalGeneration",
        "Gemma4ForCausalLM",
        "Gemma4ForConditionalGeneration",
        "Gemma4UnifiedForConditionalGeneration",
        "LagunaForCausalLM",
    }
    if any(arch in hybrid_swa_archs for arch in model_architectures):
        # Only treat Laguna as hybrid SWA when it actually has a sliding window.
        if (
            "LagunaForCausalLM" in model_architectures
            and hf_text_config is not None
            and not getattr(hf_text_config, "sliding_window", 0)
        ):
            return False
        return True
    # Also recognize models that explicitly opt-in via their HF text config.
    if hf_text_config is not None and getattr(hf_text_config, "is_hybrid_swa", False):
        return True
    return False
```

评论区精华

无 review 评论。审核人 kpham-sgl 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。修改仅限于 is_hybrid_swa_model 中针对 LagunaForCausalLM 的特判分支和 laguna.py 中的默认值合成逻辑，不影响其他 hybrid-SWA 模型。但未添加测试用例覆盖 sliding_window=0 的场景，未来重构可能遗漏此边界。
- 影响：对使用 sliding_window=0 的 Laguna 模型用户，错误 hybrid-SWA 识别将不再触发，模型可正常加载。其他用户无影响。影响范围小，仅限于特定配置的 Laguna 模型。
- 风险标记：缺少测试覆盖

关联脉络

- PR #28400 [Model] Laguna: support per-element output gating: 同一文件 laguna.py 和 model_config.py，且该 PR 引入了 LagunaForCausalLM 架构。本 PR 修复其边界条件。