

PR #35794 完整报告

sgl-project/sglang

Add SGLang Granite SWA support via existing Granite models

合并时间: 2026-08-22 02:13

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35794>

执行摘要

- 一句话:
- 推荐动作:

功能与动机

实现拆解

1. 核心源码路径 - python/sglang/srt/models/granitemoe.py (data-contract) : 源码主路径; 涉及符号 granitemoe_split_expert_weights, GraniteMoeSharedMLP, init, forward ; 包含 导入关系调整、控制流调整、配置键调整; +134/-51; 关键符号: granitemoe_split_expert_weights, GraniteMoeSharedMLP, init, forward -
python/sglang/srt/models/granite.py (data-contract) : 源码主路径; 涉及符号 granite_layer_attn_params, build_attention_sinks, GraniteSWAForCausalLM; 包含 导入关系调整、控制流调整、配置键调整; +61/-17; 关键符号: granite_layer_attn_params, build_attention_sinks, GraniteSWAForCausalLM -
python/sglang/srt/models/granitemoehybrid.py (data-contract) : 源码主路径; 涉及符号 GraniteMoeSharedMLP, init, forward; 包含 导入关系调整、控制流调整、配置键调整 ; +2/-48; 关键符号: GraniteMoeSharedMLP, init, forward

关键文件:

- python/sglang/srt/models/granitemoe.py (模块 sglang/srt; 类别 source; 类型 data-contract; 符号 granitemoe_split_expert_weights, GraniteMoeSharedMLP, init, forward) : 源码主路径; 涉及符号 granitemoe_split_expert_weights, GraniteMoeSharedMLP, init, forward; 包含 导入关系调整、控制流调整、配置键调整; +134/-51
- python/sglang/srt/models/granite.py (模块 sglang/srt; 类别 source; 类型 data-contract; 符号 granite_layer_attn_params, build_attention_sinks, GraniteSWAForCausalLM) : 源码主路径; 涉及符号 granite_layer_attn_params, build_attention_sinks, GraniteSWAForCausalLM; 包含 导入关系调整、控制流调整、配置键调整; +61/-17
- python/sglang/srt/models/granitemoehybrid.py (模块 sglang/srt; 类别 source; 类型 data-contract; 符号 GraniteMoeSharedMLP, init, forward) : 源码主路径; 涉及符号 GraniteMoeSharedMLP, init, forward; 包含 导入关系调整、控制流调整、配置键调整;

+2/-48

- python/sglang/srt/configs/model_config.py (模块 `sglang/srt`; 类别 `source`; 类型 `data-contract`) : 源码主路径; 包含 控制流调整、配置键调整; +11/-4
- python/sglang/srt/configs/granitemoehybrid.py (模块 `sglang/srt`; 类别 `source`; 类型 `core-logic`; 符号 `_to_sglang_layer_types`) : 源码主路径; 涉及符号 `_to_sglang_layer_types`; 包含 配置键调整、符号定义调整; +11/-1

关键符号: `granitemoe_split_expert_weights`, `GraniteMoeSharedMLP`, `init`, `forward`, `GraniteMoeSharedForCausalLM`, `GraniteMoeSWAForCausalLM`, `granite_layer_attn_params`, `build_attention_sinks`, `GraniteSWAForCausalLM`, `_to_sglang_layer_types`

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR