PR #28629 完整报告

sgl-project/sglang

[Bugfix] Fix Intern-S1 FP8 expert count lookup

合并时间：2026-06-18 17:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28629

执行摘要

- 一句话：修复 Intern-S1 FP8 专家数属性路径
- 推荐动作：建议合入。该修复简洁且必要，已在 B300 上验证通过。

功能与动机

Intern-S1-FP8 模型在启动时触发 AttributeError: 'InternS1Config' object has no attribute 'num_experts'，导致服务无法启动。根因是 InternS1Config 结构将 MoE 参数嵌套在 text_config 中，而 load_weights 直接使用了 self.config.num_experts。

实现拆解

在 python/sglang/srt/models/interns1.py 的 load_weights 方法中，将 FusedMoE.make_expert_params_mapping 调用的参数 num_experts 从 self.config.num_experts 改为 self.config.text_config.num_experts。仅一行变更，但属于数据契约调整，确保 FP8 MoE 权重反量化路径能正确获取专家数。

关键文件：

- python/sglang/srt/models/interns1.py（模块 模型加载；类别 source；类型 data-contract）：修改了 FP8 MoE 权重加载时获取专家数量的属性路径，从 self.config.num_experts 改为 self.config.text_config.num_experts。

关键符号：load_weights

关键源码片段

python/sglang/srt/models/interns1.py

修改了 FP8 MoE 权重加载时获取专家数量的属性路径，从 self.config.num_experts 改为 self.config.text_config.num_experts。

```
defload_weights(self, weights: Iterable[Tuple[str, torch.Tensor]]):     # ... 其他代码 ...
 expert_params_mapping = []     if "Qwen3MoeForCausalLM" in
self.config.text_config.architectures:         expert_params_mapping =
FusedMoE.make_expert_params_mapping(             ckpt_gate_proj_name="gate_proj",
        ckpt_down_proj_name="down_proj",             ckpt_up_proj_name="up_proj",
   # 修复：InternS1Config 的 num_experts 位于 text_config 内
```

num_experts=self.config.text_config.num_experts,) # ... 其余权重加载逻辑 ...
注意：仅第 214 行从 self.config.num_experts 改为 self.config.text_config.num_experts。

评论区精华

无重大讨论。review 仅来自 gemini-code-assist[bot] 和 ispobock（批准），无争议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更只影响 Intern-S1 模型加载路径，且已验证正确。对其他模型无影响。
- 影响：直接影响 Intern-S1-FP8 模型的使用者，使其能正常启动服务。对系统其他部分无副作用。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR