

PR #28571 完整报告

sgl-project/sglang

revert the head_dim assignment from PR 23862

合并时间: 2026-06-18 06:46

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28571>

执行摘要

- 一句话: 回退 PR #23862 中 head_dim 赋值逻辑及对应测试
- 推荐动作: 建议尽快合入以恢复 PR #23862 之前的行为。合入后应监控 CI 中相关模型 (尤其是 MLA 架构) 的测试是否通过, 并考虑是否需补回更稳健的测试。

功能与动机

根据 PR #23862 讨论中的决定 (参见 https://github.com/sgl-project/sglang/pull/23862#discussion_r3416313980), 认为之前引入的 head_dim 赋值逻辑可能存在问题或设计不当, 因此将其整体回退。

实现拆解

1. 修改 `python/sglang/srt/configs/model_config.py`: 在 `_derive_model_shapes` 方法中, 将原来 `getattr(..., None)`; `if None: ...`; `setattr(...)` 的 4 段代码统一替换为 `getattr(..., default_value)` 的单行形式, 且不再对 `hf_text_config` 执行 `setattr`。
2. 删除 `test/registered/unit/configs/test_model_config_shapes.py`: 移除了针对 head_dim 默认值行为和显式值保持的 2 个测试用例, 以及辅助函数和测试类。

关键文件:

- `python/sglang/srt/configs/model_config.py` (模块 模型配置; 类别 source; 类型 data-contract): 核心变更文件: 回退了 head_dim 等属性的赋值逻辑
- `test/registered/unit/configs/test_model_config_shapes.py` (模块 配置测试; 类别 test; 类型 deletion; 符号 `_make_text_config`, `TestModelConfigShapes`, `_derive_shapes`, `test_optional_head_dims_default_when_none`): 被删除的测试文件: 与回退逻辑对应的单元测试全部移除

关键符号: `_derive_model_shapes`

关键源码片段

`python/sglang/srt/configs/model_config.py`

核心变更文件: 回退了 head_dim 等属性的赋值逻辑

```
# python/sglang/srt/configs/model_config.py (回退后)
def _derive_model_shapes(self):
```

```

# Unify the config keys for hf_text_config
# 回退后：直接使用 getattr 的默认值，不再手动 setattr
self.head_dim = getattr(
    self.hf_text_config,
    "head_dim",
    self.hf_text_config.hidden_size // self.hf_text_config.num_attention_heads,
)
self.v_head_dim = getattr(
    self.hf_text_config,
    "v_head_dim",
    self.head_dim,
)
self.swa_head_dim = getattr(
    self.hf_text_config,
    "swa_head_dim",
    self.head_dim,
)
self.swa_v_head_dim = getattr(
    self.hf_text_config,
    "swa_v_head_dim",
    self.swa_head_dim,
)
# 后续 MLA 特殊覆盖逻辑保持不变
# FIXME: temporary special judge for MLA architecture
if (
    "DeepseekV2ForCausalLM" in self.hf_config.architectures
    or ... # 省略
):
    self.head_dim = 256
    self.attention_arch = AttentionArch.MLA
    ...

```

评论区精华

无 review 评论。PR body 引用了一个外部讨论链接，但无具体内容。机器人 gemini-code-assist 的 quota 警告不构成有效讨论。

- 外部讨论决定回退 (other): 团队决定回退 PR #23862 的变更。

风险与影响

- 风险：回退后，head_dim 等属性不再自动写入 hf_text_config，可能影响依赖该属性的下游逻辑（如 MLA 架构的特殊判断）。但回退前后的行为在典型场景下等价（getattr 的默认值计算与原逻辑一致），风险较低。需确认 MLA 架构下的覆盖逻辑（第 640 行 self.head_dim = 256）仍然生效。
- 影响：影响范围：所有使用 ModelConfig 的模型。回退后，hf_text_config 上的 head_dim 等属性不再被显式设置，但 ModelConfig 实例的属性计算方式不变（仍然通过 getattr 得到值）。对于未显式设置 head_dim 的配置，行为与回退前一致（默认 / 计算）。删除的测

试覆盖需要补充或后续决定。

- 风险标记：核心路径变更，测试覆盖移除

关联脉络

- PR #23862（原 PR，标题未知）：本 PR 直接回退了该 PR 中对 `model_config.py` 的修改以及对应的测试文件。