

PR #1883 完整报告

THUDM/slime

fix(qwen3_next): use torch.get_default_dtype() — get_current_dtype do...

合并时间: 2026-05-06 14:03

原文链接: <http://prhub.com.cn/THUDM/slime/pull/1883>

执行摘要

- 一句话: 修复 dtype 获取方法名称错误
- 推荐动作: 该 PR 属于简单 bugfix, 修复直接, 风险低, 可以快速合并。无需深度审查。

功能与动机

PR 标题和 commit message 明确说明: `torch.get_current_dtype()` 不存在, 正确 API 应为 `torch.get_default_dtype()`。修复后在 config 未指定 dtype 时能够正确获取 PyTorch 默认 dtype。

实现拆解

1. 定位错误 API: 在 `slime_plugins/models/qwen3_5.py` 和 `slime_plugins/models/qwen3_next.py` 的 `FusedRMSNormGated` 初始化处, 当 `config.dtype` 为 `None` 时, 使用 `torch.get_current_dtype()` 作为后备 dtype。
2. 替换为正确 API: 将 `torch.get_current_dtype()` 修改为 `torch.get_default_dtype()`, 后者是 PyTorch 标准的获取默认 dtype 的方法。
3. 无其他影响: 该变量仅用于 `FusedRMSNormGated` 的 dtype 参数, 修改后不会影响其他模块行为。

关键文件:

- `slime_plugins/models/qwen3_5.py` (模块 模型定义; 类别 source; 类型 data-contract) : 修复 Qwen3.5 模型初始化时 dtype 后备逻辑中的 API 名称错误。
- `slime_plugins/models/qwen3_next.py` (模块 模型定义; 类别 source; 类型 data-contract) : 修复 Qwen3Next 模型初始化时 dtype 后备逻辑中的 API 名称错误。

关键符号: 未识别

关键源码片段

`slime_plugins/models/qwen3_5.py`

修复 Qwen3.5 模型初始化时 dtype 后备逻辑中的 API 名称错误。

```
# slime_plugins/models/qwen3_5.py
```

```
self.norm = FusedRMSNormGated(
```

```
self.head_v_dim,  
eps=self.layer_norm_epsilon,  
activation=self.activation,  
device=torch.cuda.current_device(),  
# 修复：使用 torch.get_default_dtype() 替代不存在的 torch.get_current_dtype()  
dtype=config.dtype if config.dtype is not None else torch.get_default_dtype(),  
)
```

slime_plugins/models/qwen3_next.py

修复 Qwen3Next 模型初始化时 dtype 后备逻辑中的 API 名称错误。

```
# slime_plugins/models/qwen3_next.py  
  
self.norm = FusedRMSNormGated(  
    self.head_v_dim,  
    eps=self.layer_norm_epsilon,  
    activation=self.activation,  
    device=torch.cuda.current_device(),  
    # 修复：使用 torch.get_default_dtype() 替代不存在的 torch.get_current_dtype()  
    dtype=config.dtype if config.dtype is not None else torch.get_default_dtype(),  
)
```

评论区精华

该 PR 无 review 评论或讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及两个文件中同一行代码的 API 名称替换，且目标 API 是 PyTorch 稳定且广泛使用的 `torch.get_default_dtype()`。不涉及控制流、数据结构或性能变化。
- 影响：影响范围小，仅影响 Qwen3.5 和 Qwen3Next 模型在 config 未显式设置 dtype 时的 Fallback 行为。修复后，这些模型在未指定 dtype 时将正确使用 PyTorch 的默认 dtype（通常为 `torch.float32`），而非因调用不存在的方法而抛出 `AttributeError`。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR