

PR #50210 完整报告

vllm-project/vllm

[Model] Support Qwen3.5 text-only dense and MoE models

合并时间: 2026-07-29 16:21

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/50210>

执行摘要

- 一句话: 注册 Qwen3.5 文本纯文本模型架构
- 推荐动作: 值得精读。该 PR 展示了一个典型的模型注册修复模式: 注册架构名、补全配置类、补充必要的接口 (IsHybrid + mamba 状态方法)。对理解 vLLM 模型注册系统有参考价值。

功能与动机

Qwen3.5 文本纯文本检查点 (如社区量化的 abliterated 版本) 使用 `Qwen3_5ForCausalLM` 或 `Qwen3_5MoeForCausalLM` 架构名, 但 vLLM 仅注册了多模态变体, 导致加载时依次出现架构未映射、断言错误等连锁失败。Issue #48060 详细描述了三个失败阶段。

实现拆解

1. 注册架构名: 在 `vllm/model_executor/models/registry.py` 的 `_TEXT_GENERATION_MODELS` 中添加 `Qwen3_5ForCausalLM` 和 `Qwen3_5MoeForCausalLM` 条目, 映射到 `qwen3_5` 模型包和对应的类。
2. 创建配置类并注册: 在 `vllm/model_executor/models/config.py` 中新增 `Qwen3_5ForCausalLMConfig`, 继承自 `Qwen3_5ForConditionalGenerationConfig`, 在其 `verify_and_update_config` 中先调用父类方法处理 `mamba_ssm_cache_dtype`, 再移除 `rope_parameters` 中的 M-RoPE 字段 (`mrope_section`、`mrope_interleaved`)。然后将两个架构名加入 `MODELS_CONFIG_MAP`。
3. 补充配置类型映射: 在 `vllm/transformers_utils/config.py` 的 `_CONFIG_REGISTRY` 中添加 `qwen3_5_text` → `Qwen3_5TextConfig` 和 `qwen3_5_moe_text` → `Qwen3_5MoeTextConfig`, 使 HuggingFace 配置加载器能识别文本纯文本配置。
4. 标记为混合模型并提供 mamba 状态方法: 在 `vllm/model_executor/models/qwen3_5.py` 中, 让 `Qwen3_5ForCausalLMBase` 额外继承 `IsHybrid`, 并添加三个类方法 `get_mamba_state_dtype_from_config`、`get_mamba_state_shape_from_config`、`get_mamba_state_copy_func`, 它们分别委托给 `MambaStateDtypeCalculator`、`MambaStateShapeCalculator`、`MambaStateCopyFuncCalculator` 的 `gated_delta_net_state_*` 方法。
5. 测试补充: 在 `tests/models/registry.py` 中为两个新架构添加 `_HfExamplesInfo` 条目, 使用官方提供的文本纯文本检查点 (`codecho/Qwen3.5-0.8B-text-only` 和

codecho/Qwen3.5-35B-A3B-text-only) 。

关键文件：

- vllm/model_executor/models/qwen3_5.py (模块 模型定义；类别 source；类型 data-contract；符号 get_mamba_state_dtype_from_config, get_mamba_state_shape_from_config, get_mamba_state_copy_func)：核心变更：基类新增 IsHybrid 继承和三个 mamba 状态类方法，直接修复 AssertionError。
- vllm/model_executor/models/config.py (模块 模型配置；类别 source；类型 data-contract；符号 Qwen3_5ForCausalLMConfig, verify_and_update_config)：新增 Qwen3_5ForCausalLMConfig 类，清理 M-RoPE 字段，并将两个架构名注册到 MODELS_CONFIG_MAP。
- vllm/model_executor/models/registry.py (模块 模型注册；类别 source；类型 data-contract)：在 _TEXT_GENERATION_MODELS 中添加两个架构名，使模型加载器能定位到正确的实现类。
- vllm/transformers_utils/config.py (模块 配置工具；类别 source；类型 core-logic)：添加 qwen3_5_text 和 qwen3_5_moe_text 配置类型映射，使 HF 配置加载器能识别文本纯文本配置。
- tests/models/registry.py (模块 测试；类别 test；类型 test-coverage)：为两个新架构添加测试注册表条目，使用官方文本纯文本检查点。

关键符号：get_mamba_state_dtype_from_config, get_mamba_state_shape_from_config, get_mamba_state_copy_func, Qwen3_5ForCausalLMConfig.verify_and_update_config

关键源码片段

vllm/model_executor/models/qwen3_5.py

核心变更：基类新增 IsHybrid 继承和三个 mamba 状态类方法，直接修复 AssertionError。

```
class Qwen3_5ForCausalLMBase(
    nn.Module,
    HasInnerState,
    IsHybrid, # 新增：标记为混合模型，使调度器能正确处理 GDN 状态
    SupportsEagle3,
    SupportsLoRA,
    SupportsPP,
):
    # ... ( 已有代码 )

    @classmethod
    def get_mamba_state_dtype_from_config(
        cls,
        vllm_config: "VllmConfig",
    ) -> tuple[torch.dtype, torch.dtype]:
        # 委托给计算器，使用与多模态变体相同的逻辑
        return MambaStateDtypeCalculator.gated_delta_net_state_dtype(
            vllm_config.model_config.dtype,
```

```

        vllm_config.cache_config.mamba_cache_dtype,
        vllm_config.cache_config.mamba_ssm_cache_dtype,
    )

    @classmethod
    def get_mamba_state_shape_from_config(
        cls, vllm_config: "VllmConfig"
    ) -> tuple[tuple[int, int], tuple[int, int]]:
        parallel_config = vllm_config.parallel_config
        hf_config = vllm_config.model_config.hf_text_config
        tp_size = parallel_config.tensor_parallel_size
        num_spec = (
            vllm_config.speculative_config.num_speculative_tokens
            if vllm_config.speculative_config
            else 0
        )
        return MambaStateShapeCalculator.gated_delta_net_state_shape(
            tp_size,
            hf_config.linear_num_key_heads,
            hf_config.linear_num_value_heads,
            hf_config.linear_key_head_dim,
            hf_config.linear_value_head_dim,
            hf_config.linear_conv_kernel_dim,
            num_spec,
        )

    @classmethod
    def get_mamba_state_copy_func(
        cls,
    ) -> tuple[MambaStateCopyFunc, MambaStateCopyFunc]:
        return MambaStateCopyFuncCalculator.gated_delta_net_state_copy_func()

```

vllm/model_executor/models/config.py

新增 Qwen3_5ForCausalLMConfig 类，清理 M-RoPE 字段，并将两个架构名注册到 MODELS_CONFIG_MAP。

```

class Qwen3_5ForCausalLMConfig(Qwen3_5ForConditionalGenerationConfig):
    @staticmethod
    def verify_and_update_config(vllm_config: "VllmConfig") -> None:
        # 先处理 mamba_ssm_cache_dtype (与多模态相同)
        Qwen3_5ForConditionalGenerationConfig.verify_and_update_config(vllm_config)

        # 文本纯文本模型使用一维位置编码，需要移除 M-RoPE 字段
        hf_text_config = vllm_config.model_config.hf_text_config
        rope_parameters = getattr(hf_text_config, "rope_parameters", None)
        if rope_parameters is not None:
            rope_parameters.pop("mrope_section", None)
            rope_parameters.pop("mrope_interleaved", None)

```

```
# 在 MODELS_CONFIG_MAP 中注册
"Qwen3_5ForCausalLM": Qwen3_5ForCausalLMConfig,
"Qwen3_5MoeForCausalLM": Qwen3_5ForCausalLMConfig,
```

评论区精华

无实质性 review 讨论。claude[bot] 自动评论提示 PR 来自 fork 且自动化 review 被禁用。
Isotr0py 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更均为新增注册和配置，不影响现有模型加载逻辑。唯一可能的风险是：若用户显式设置了 `--mamba-ssm-cache-dtype` 且与 HF 配置冲突，`Qwen3_5ForCausalLMConfig.verify_and_update_config` 会发出警告而不是报错，行为与多模态变体一致，属于预期设计。
- 影响：对用户：支持加载社区发布的 Qwen3.5 文本纯文本 dense/MoE 检查点，无需手动绕过注册。对系统：无性能影响。对团队：降低了模型支持的维护成本，将文本纯文本路径与多模态路径统一。
- 风险标记：低风险，核心注册变更，配置兼容性警告

关联脉络

- PR #39316 dense-only handling for Qwen3.5: PR 提到本变更扩展了 #39316 提出的密集模型处理方式到 MoE 文本纯文本路径。
- PR #48060 [Bug] Qwen3_5ForCausalLM / Qwen3_5MoeForCausalLM defined but not registered — text-only checkpoints fail to load: 直接关联的 Issue，描述了三个失败阶段，本 PR 完全修复。