PR #43992 完整报告

vllm-project/vllm

[Feature] Add support for JetBrains' Mellum v2 code generation model

合并时间：2026-06-01 22:11

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43992

执行摘要

- 一句话：新增 JetBrains Mellum v2 代码生成模型支持
- 推荐动作：该 PR 结构清晰、改动集中，适合作为新模型支持的标准参考。建议简要浏览实现文件，了解 vLLM 添加新模型时需修改的注册点（registry.py、configs/init.py、config.py）。

功能与动机

PR body 说明：Adds support for a new model architecture - Mellum v2 is an update to JetBrains' open-weights code generation model that is built on a Mixture-of-Experts architecture (Mellum v1 was based on Llama2)。

实现拆解

1. 配置层：新增 vllm/transformers_utils/configs/mellum.py，定义 MellumConfig 继承 Qwen3MoeConfig，仅重写 model_type = "mellum"。
2. 模型层：新增 vllm/model_executor/models/mellum.py，核心模型类全部继承自 Qwen3Moe 模块：
   - MellumAttention：继承 Qwen3MoeAttention，扩展支持 per_layer_sliding_window 和 per-layer RoPE scaling。
   - MellumDecoderLayer：继承 Qwen3MoeDecoderLayer，支持 interleaved SWA。
   - MellumModel、MellumForCausalLM：继承对应 Qwen3Moe 类，添加 @support_torch_compile 装饰器。
3. 注册中心：修改 vllm/model_executor/models/registry.py，添加 "MellumForCausalLM": ("mellum", "MellumForCausalLM") 映射。
4. 配置导入：在 vllm/transformers_utils/configs/__init__.py 中注册 MellumConfig 及其模块路径；在 vllm/transformers_utils/config.py 中添加模型目录别名。
5. 测试与文档：
   - tests/models/registry.py 添加模型示例 JetBrains/Mellum2-12B-A2.5B-Base。
   - docs/models/supported_models.md 添加 Mellum 条目。

关键文件：

- vllm/model_executor/models/mellum.py（模块 模型层；类别 source；类型 new-model；符号 MellumAttention, MellumDecoderLayer, MellumModel, MellumForCausalLM）：模型架构核心实现，包含 MellumAttention、MellumDecoderLayer、MellumModel、MellumForCausalLM 等类，全部继承自 Qwen3Moe 并扩展滑动窗口和 RoPE 缩放。
- vllm/transformers_utils/configs/mellum.py（模块 配置层；类别 source；类型 core-logic；符号 MellumConfig）：Mellum 配置类定义，通过继承 Qwen3MoeConfig 并设置 model_type 实现适配。
- vllm/model_executor/models/registry.py（模块 注册中心；类别 source；类型 data-contract）：模型注册表，添加 MellumForCausalLM 到模型路径映射。
- vllm/transformers_utils/configs/__init__.py（模块 配置层；类别 source；类型 core-logic）：配置文件导入模块，注册 MellumConfig 的模块路径。
- vllm/transformers_utils/config.py（模块 配置层；类别 source；类型 core-logic）：模型别名映射，添加 mellum -> MellumConfig。
- tests/models/registry.py（模块 测试配置；类别 test；类型 test-coverage）：测试模型注册，添加 MellumForCausalLM 的测试示例和权重信息。
- docs/models/supported_models.md（模块 文档；类别 docs；类型 documentation）：文档更新，列出新支持的模型。

关键符号：MellumAttention.init, MellumDecoderLayer.init, MellumModel, MellumForCausalLM

关键源码片段

vllm/model_executor/models/mellum.py

模型架构核心实现，包含 MellumAttention、MellumDecoderLayer、MellumModel、MellumForCausalLM 等类，全部继承自 Qwen3Moe 并扩展滑动窗口和 RoPE 缩放。

```
# vllm/model_executor/models/mellum.py
# 定义 MellumAttention，继承自 Qwen3MoeAttention，
# 核心差异在于 __init__ 接受 per_layer_sliding_window 参数并透传给 Attention 层。
class MellumAttention(Qwen3MoeAttention):
    """
    Differences from `Qwen3MoeAttention`:
    - Supports `per_layer_sliding_window` for `Attention`.
    """

    def __init__(
        self,
        hidden_size: int,
        num_heads: int,
        num_kv_heads: int,
        rope_parameters: dict[str, Any],
        max_position_embeddings: int = 8192,
        head_dim: int | None = None,
        rms_norm_eps: float = 1e-06,
        qkv_bias: bool = False,
```

```
        cache_config: Any | None = None,
        quant_config: Any | None = None,
        prefix: str = "",
        dual_chunk_attention_config: dict[str, Any] | None = None,
        per_layer_sliding_window: int | None = None, # 新增的滑动窗口参数
    ) -> None:
        nn.Module.__init__(self)
        # ... 标准初始化代码（与 Qwen3MoeAttention 相同）...
        self.attn = Attention(
            self.num_heads,
            self.head_dim,
            self.scaling,
            num_kv_heads=self.num_kv_heads,
            cache_config=cache_config,
            quant_config=quant_config,
            per_layer_sliding_window=per_layer_sliding_window, # 传递给底层 Attention
            prefix=f"{prefix}.attn",
            ...
        )
```

评论区精华

无实质技术讨论，仅有 mergify 自动评论和批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。新模型完全继承自经过验证的 Qwen3Moe 架构，仅添加少量差异化参数（滑动窗口、RoPE 缩放）。主要风险在于：
  - vllm/model_executor/models/mellum.py 中的 MellumAttention 新增了 per_layer_sliding_window 参数，若首次使用该特性的后端尚未充分测试，可能引发注意力层异常。
  - 模型注册链（configs/init.py、config.py、registry.py）是典型新模型添加模式，出错概率小，但需确保与实际 HF 模型配置匹配。
  - 影响：用户侧：可在 vLLM 中直接加载 JetBrains/Mellum2-12B-A2.5B-Base 等 Mellum v2 系列模型进行推理。系统侧：新增约 253 行核心模型代码，但高度复用 Qwen3Moe 模块，维护成本较低。团队侧：需要跟踪 Mellum 后续版本可能带来的架构变化。
  - 风险标记：新模型集成，依赖 Qwen3Moe 架构

关联脉络

- 暂无明显关联 PR