

PR #48452 完整报告

vllm-project/vllm

[Bugfix][UT]Fix EagleMiniCPMForCausalLM meet TypeError

合并时间: 2026-07-13 12:37

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48452>

执行摘要

- 一句话: 修复 EagleMiniCPM 单元测试 TypeError
- 推荐动作: 建议审核者快速合并。该 PR 修复了一个明确的测试失败问题, 改动量小 (+11/-3), 已被 @DarkLight1337 批准。后续可关注是否真的需要多模态支持 (届时移除 inputs_embeds 处的 raise 即可)。

功能与动机

修复单元测试 `test_can_initialize_large_subset[EagleMiniCPMForCausalLM]` 在 `VLLM_WORKER_MULTIPROC_METHOD=spawn` 环境下抛出的 `TypeError: EagleMiniCPMForCausalLM.forward() got an unexpected keyword argument 'inputs_embeds'`。根本原因包括测试注册表中目标模型 MiniCPM-1B (hidden=1536) 与草稿模型 MiniCPM-2B (hidden=2304) 的 `hidden_size` 不匹配、`inputs_embeds` 参数缺失以及 `mup_denominator` 属性缺失。

实现拆解

1. 补充 `mup_denominator` fallback 属性 (`minicpm_eagle.py`): 在 `EagleMiniCPMDecoderLayer.__init__` 中添加 `self.mup_denominator = getattr(config, 'mup_denominator', config.num_hidden_layers)`, 避免 `config.mup_denominator` 不存在时抛 `AttributeError`。同时将 `forward` 中的 `self.config.mup_denominator` 引用改为 `self.mup_denominator`。
2. 在 `forward` 接口中接受 `inputs_embeds` 并明确拒绝 (`minicpm_eagle.py`): 在 `EagleMiniCPMForCausalLM.forward` 签名中加入 `inputs_embeds: torch.Tensor | None = None`, 若接收到非 `None` 值则 `raise NotImplementedError`, 表明暂不支持多模态输入。这样保持了与上层调用契约的兼容性。
3. 修正测试注册表中的模型配对 (`tests/models/registry.py`): 将 `EagleMiniCPMForCausalLM` 的目标模型从 `openbmb/MiniCPM-1B-sft-bf16` 改为 `openbmb/MiniCPM-2B-sft-bf16`, 使其 `hidden_size` (2304) 与草稿模型一致, 避免下游在加载权重时因维度不匹配出错。

关键文件:

- `vllm/model_executor/models/minicpm_eagle.py` (模块 模型执行器; 类别 `source`; 类型 `data-contract`; 符号 `EagleMiniCPMDecoderLayer.init`,

EagleMiniCPMDecoderLayer.forward, EagleMiniCPMForCausalLM.forward)：核心修复文件：补充 mup_denominator 属性，扩展 forward 签名以接受 inputs_embeds 并显式拒绝，确保与上层调用契约一致。

- tests/models/registry.py (模块注册表；类别 test；类型 test-coverage)：测试注册表修正：将 EagleMiniCPMForCausalLM 的目标模型从 MiniCPM-1B (hidden=1536) 改为 MiniCPM-2B (hidden=2304)，使其与草稿模型的 hidden_size 匹配，是触发 TypeError 的根本原因之一。

关键符号：EagleMiniCPMDecoderLayer.init, EagleMiniCPMDecoderLayer.forward, EagleMiniCPMForCausalLM.forward

关键源码片段

vllm/model_executor/models/minicpm_eagle.py

核心修复文件：补充 mup_denominator 属性，扩展 forward 签名以接受 inputs_embeds 并显式拒绝，确保与上层调用契约一致。

```
# vllm/model_executor/models/minicpm_eagle.py
```

```
class EagleMiniCPMDecoderLayer(nn.Module):
    def __init__(self, config, cache_config=None, quant_config=None, prefix=""):
        super().__init__()
        self.config = config
        self.cache_config = cache_config
        self.quant_config = quant_config
        self.hidden_size = config.hidden_size
        self.max_position_embeddings = getattr(config, "max_position_embeddings", 8192)
        # 修复点 1: 给 mup_denominator 提供 fallback 值，防止 config 中缺失该属性
        self.mup_denominator = getattr(
            config, "mup_denominator", config.num_hidden_layers
        )
        self.prefix = prefix
        self._init_attn_block()
        self._init_ffn_block()
```

```
    def forward(self, positions, hidden_states, residual):
        # ... attention 与 ffn 计算 ...
        # 修复点 2: 使用 self.mup_denominator 而非 self.config.mup_denominator
        hidden_states = residual + hidden_states * (
            self.config.scale_depth / math.sqrt(self.mup_denominator)
        )
        # ... 同上 ...
        return hidden_states, None
```

```
class EagleMiniCPMForCausalLM(nn.Module):
    # ...
    def forward(
        self,
```

```

input_ids: torch.Tensor,
positions: torch.Tensor,
hidden_states: torch.Tensor,
# 修复点 3: 新增 inputs_embeds 参数以使签名兼容上层调用
inputs_embeds: torch.Tensor | None = None,
):
    # 明确拒绝多模态输入，避免静默忽略造成难以排查的问题
    if inputs_embeds is not None:
        raise NotImplementedError(
            f"{type(self).__name__} does not support multimodal inputs yet."
        )
    # ... 原本的 forward 逻辑 ...

```

tests/models/registry.py

测试注册表修正：将 EagleMiniCPMForCausalLM 的目标模型从 MiniCPM-1B (hidden=1536) 改为 MiniCPM-2B (hidden=2304)，使其与草稿模型的 hidden_size 匹配，是触发 TypeError 的根本原因之一。

```

# tests/models/registry.py

# 目标模型应和 speculative_model 的 hidden_size 一致（均为 2304）
# 之前错误地使用了 hidden_size=1536 的 MiniCPM-1B
"EagleMiniCPMForCausalLM": _HfExamplesInfo(
    "openbmb/MiniCPM-2B-sft-bf16", # 已修正: hidden_size 2304
    trust_remote_code=True,
    speculative_model="openbmb/MiniCPM-2B-sft-bf16",
    speculative_method="eagle",
    tokenizer="openbmb/MiniCPM-2B-sft-bf16",
),

```

评论区精华

PR 作者 @Yejing-Lai 提出了 fix，维护者 @DarkLight1337 快速批准并 @tc-mb 知会。讨论集中在修复本身，没有额外争议。

- 暂无高价值评论线程

风险与影响

- 风险：本 PR 是低风险的局域修复。核心改动仅涉及 minicpm_eagle.py 中一个类（EagleMiniCPMDecoderLayer）的初始化逻辑和 EagleMiniCPMForCausalLM.forward 签名的扩展，影响面仅限于 Eagle 架构的 MiniCPM 模型。inputs_embeds 显式 raise 的做法比静默忽略更安全，可防止未来滥用该接口。测试注册表的改动确保目标模型与草稿模型 hidden_size 一致，避免了在初始化阶段隐藏的形状不匹配错误。
- 影响：直接影响范围为使用 EagleMiniCPMForCausalLM 进行推测解码的用户，特别是调用初始化测试或实际加载该模型的场景。修复后，test_can_initialize_large_subset 测试能通过，并且运行时模型加载不再触发相关 TypeError。对现有代码无 break change。
- 风险标记：测试覆盖补齐

关联脉络

- 暂无明显关联 PR