

PR #49786 完整报告

vllm-project/vllm

[Model] Remove Ouro

合并时间: 2026-07-25 17:50

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49786>

执行摘要

- 一句话: 移除废弃的 Ouro 模型实现及注册
- 推荐动作: 值得关注模型清理流程, 尤其是如何通过 `_PREVIOUSLY_SUPPORTED_MODEL` S 维护向后兼容性。对于模型维护者可参考此 PR 规范废弃模型。

功能与动机

Ouro 模型发布已约一年, 被 Plamo3 替代, 在 vLLM 中使用很少, 已被标记为待删除。移除以简化代码库维护。

实现拆解

1. 删除模型实现: 完全移除 `vllm/model_executor/models/ouro.py`, 包含 `OuroMLP`、`OuroAttention`、`OuroDecoderLayer`、`OuroModel`、`OuroForCausalLM` 等所有类。
2. 更新模型注册表: 在 `registry.py` 中从 `_TORCH_MODEL_CLASSES` 字典中移除 `OuroForCausalLM` 映射, 同时在 `_PREVIOUSLY_SUPPORTED_MODELS` 字典中添加 "`OuroForCausalLM`": "0.26.0", 确保用户加载旧模型时收到版本提示。
3. 测试同步: 从 `tests/models/registry.py` 的 `_HF_EXAMPLE_MODELS` 字典中移除 Ouro 测试示例, 避免 CI 尝试下载 Ouro 权重。
4. 文档更新: 从 `docs/models/supported_models.md` 支持模型表格中删除 Ouro 模型行。

关键文件:

- `vllm/model_executor/models/ouro.py` (模块 模型实现; 类别 source; 类型 deletion; 符号 `OuroMLP`, `OuroAttention`, `OuroDecoderLayer`, `OuroModel`): 主要变更文件, 删除整个 Ouro 模型实现, 共 448 行。
- `vllm/model_executor/models/registry.py` (模块 模型注册; 类别 source; 类型 data-contract): 更新模型注册表, 移除 Ouro 条目并加入已支持历史列表。
- `tests/models/registry.py` (模块 模型测试; 类别 test; 类型 test-coverage): 测试注册表同步移除 Ouro 模型示例, 避免 CI 加载。
- `docs/models/supported_models.md` (模块 文档; 类别 docs; 类型 documentation): 从支持模型列表中删除 Ouro 模型行。

关键符号: `OuroMLP`, `OuroAttention`, `OuroDecoderLayer`, `OuroModel`, `OuroForCausalLM`

关键源码片段

[vllm/model_executor/models/ouro.py](#)

主要变更文件，删除整个 Ouro 模型实现，共 448 行。

SPDX-License-Identifier: Apache-2.0

此文件在 PR #49786 中被完全删除。

以下代码是原始的 OuroMLP 实现，保留以供参考。

```
class OuroMLP(nn.Module):
    def __init__(
        self,
        hidden_size: int,
        intermediate_size: int,
        hidden_act: str,
        quant_config: QuantizationConfig | None = None,
        prefix: str = "",
    ) -> None:
        super().__init__()
        self.gate_up_proj = MergedColumnParallelLinear(
            hidden_size,
            [intermediate_size] * 2,
            bias=False,
            quant_config=quant_config,
            prefix=f"{prefix}.gate_up_proj",
        )
        self.down_proj = RowParallelLinear(
            intermediate_size,
            hidden_size,
            bias=False,
            quant_config=quant_config,
            prefix=f"{prefix}.down_proj",
        )
        if hidden_act != "silu":
            raise ValueError(
                f"Unsupported activation: {hidden_act}. Only silu is supported for now."
            )
        self.act_fn = SiluAndMul()

    def forward(self, x):
        gate_up, _ = self.gate_up_proj(x)
        x = self.act_fn(gate_up)
        x, _ = self.down_proj(x)
        return x
```

评论区精华

无实质讨论；由维护者 DarkLight1337 直接批准合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。Ouro 模型使用率低且已被标记废弃。
_PREVIOUSLY_SUPPORTED_MODELS 机制确保加载旧权重时提示不支持的版本，而非静默失败。无需担心回归。
- 影响：用户侧：升级后 Ouro 模型不再支持，需固定 vLLM 版本。系统侧：减少模型加载路径和构建时间，小幅减少代码体积。团队侧：维护模型清理的最佳实践范例。
- 风险标记：低风险，废弃模型移除

关联脉络

- PR #49729 [Model] Remove Plamo2: 类似的模型移除操作，表明 vLLM 正在清理废弃模型。