

PR #49729 完整报告

vllm-project/vllm

[Model] Remove Plamo2

合并时间: 2026-07-25 04:54

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49729>

执行摘要

- 一句话: 删除废弃的 Plamo2 模型
- 推荐动作: 此 PR 是模型退役的标准操作, 值得阅读以了解 `_PREVIOUSLY_SUPPORTED_MODELS` 的使用方法和完整的清理流程。不建议精读, 但可作为代码清理的参考。

功能与动机

PR 作者说明: Plamo2 模型已有约 1 年历史, 已被 Plamo3 取代; 在 vLLM 中几乎不被用户使用; 其下载量主要由 CI 驱动, 具有误导性。因此删除该模型以减少维护成本。

实现拆解

1. 删除核心模型实现: 移除 `vllm/model_executor/models/plamo2.py`, 包括 `Plamo2Config` 类型注解、`is_mamba` 辅助函数、`Plamo2MambaMixer` 类和 `Plamo2ForCausalLM` 模型类, 共约 992 行代码。
2. 更新模型注册表: 在 `vllm/model_executor/models/registry.py` 中移除 `Plamo2ForCausalLM` 到 `plamo2` 模块的映射, 同时将其加入 `_PREVIOUSLY_SUPPORTED_MODELS` 字典, 版本标记为 0.26.0, 以使用户加载旧配置时收到明确提示。
3. 清理编译配置: 在 `vllm/config/compilation.py` 的 `_attention_ops` 列表中删除 `vllm::plamo2_mamba_mixer` 自定义算子, 避免编译时无效引用。
4. 移除 dtype 限制: 在 `vllm/config/model.py` 的 `_FLOAT16_NOT_SUPPORTED_MODELS` 中删除 `plamo2` 条目, 更新注释。
5. 测试与 CI 清理: 从测试文件中移除 Plamo2 相关用例或引用; 更新 `.buildkite` 配置文件移除 Plamo2 测试任务; 更新文档 `supported_models.md` 和 `custom_op.md`。

关键文件:

- `vllm/model_executor/models/plamo2.py` (模块 模型层; 类别 source; 类型 deletion; 符号 `Plamo2Config`, `is_mamba`, `Plamo2MambaMixer`, `init`): 核心被删除的模型实现文件, 包含 `Plamo2ForCausalLM` 和 `Plamo2MambaMixer` 等关键类。
- `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 Plamo2 的模块映射, 并加入 `_PREVIOUSLY_SUPPORTED_MODELS`。

- `vllm/config/model.py` (模块 配置管理; 类别 `source`; 类型 `data-contract`): 移除 `_FLOAT16_NOT_SUPPORTED_MODELS` 中的 `plamo2` 条目, 并更新注释。
- `vllm/config/compilation.py` (模块 编译配置; 类别 `source`; 类型 `core-logic`): 从 `_attention_ops` 列表中移除 `plamo2_mamba_mixer` 自定义算子。

关键符号: `Plamo2Config`, `is_mamba`, `Plamo2MambaMixer`, `init`, `_project_ssm_parameters`, `forward`, `forward_impl`, `get_state_dtype`

关键源码片段

`vllm/model_executor/models/plamo2.py`

核心被删除的模型实现文件, 包含 `Plamo2ForCausalLM` 和 `Plamo2MambaMixer` 等关键类。

```
# plamo2.py 文件被完整删除 (原 992 行)
# 下方展示被删除的主要类 Plamo2MambaMixer 的头部
# 该模型已被 Plamo3 取代, 因此执行了删除

@PluggableLayer.register('plamo2_mamba_mixer')
class Plamo2MambaMixer(MambaBase, PluggableLayer):
    '''PLaMo2 的 Mamba 混合器实现'''
    def __init__(self, vllm_config: VllmConfig, *, prefix: str = '', **kwargs) -> None:
        super().__init__()
        self.config = vllm_config.model_config.hf_config
        # ... 后续初始化省略
```

评论区精华

该 PR 收到三位维护者的快速批准 (`yewentao256`、`DarkLight1337`、`mgoin`), 无实质性讨论或争议。Claude 的自动评论因来自 `fork` 而跳过。整体决策基于 PR 描述的合理理由。

- 删除 `Plamo2` 的合理性 (design): 决定删除 `Plamo2` 模型。

风险与影响

- 风险: 风险极低。主要风险是极少数使用 `Plamo2` 的用户会遭遇加载失败, 但已通过 `_PREVIOUSLY_SUPPORTED_MODELS` 机制提示模型已被移除并建议使用 `Plamo3`。此外, 需确保没有其他模块间接依赖 `Plamo2` (如自定义插件), 但代码库中已无引用。
- 影响: 对绝大多数用户无影响 (`Plamo2` 几乎未被使用)。代码库减少约 1020 行, 降低维护成本。文档和 CI 配置同步更新。团队未来在移除其他废弃模型时可参考此 PR 的步骤。
- 风险标记: 暂无

关联脉络

- 暂无明显关联 PR