

PR #47989 完整报告

vllm-project/vllm

Remove TeleChatForCausalLM

合并时间: 2026-07-09 15:34

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47989>

执行摘要

- 一句话: 移除 TeleChatForCausalLM 模型别名
- 推荐动作: 该 PR 值得快速合并, 属于低风险、高收益的清理变更。值得关注的设计决策是将弃用的模型以真实版本号添加到 `_PREVIOUSLY_SUPPORTED_MODELS` 中, 这为未来进一步的清理工作 (如 0.26.0 后完全移除相关代码) 奠定了基础。

功能与动机

低使用量, 近 3 个月使用时间不足 10 小时。PR body 明确指出: "Remove TeleChatForCausalLM due to low usage (less than 10 hours in the past 3 months)". 这是标准的代码清理策略, 通过废弃低采纳率的遗留兼容别名来减小维护负担。

实现拆解

1. 移除注册表中的别名: 在 `vllm/model_executor/models/registry.py` 中, 从 `_TEXT_GENERATION_MODELS` 和 `_EMBEDDING_MODELS` 两个字典中删除了 "TeleChatForCausalLM" 条目, 这两个条目原来都映射到 ("telechat2", "TeleChat2ForCausalLM")。
2. 标记废弃版本: 在同一个文件的 `_PREVIOUSLY_SUPPORTED_MODELS` 字典中新增 "TeleChatForCausalLM": "0.25.0", 明确声明该模型在 0.25.0 版本之后不再支持 (经过 review 讨论后调整为 0.25.0)。
3. 更新测试数据: 在 `tests/models/registry.py` 的 `_TEXT_GENERATION_EXAMPLE_MODELS` 字典中删除 TeleChatForCausalLM 的测试条目, 该条目引用了 `chuhac/TeleChat2-35B` 模型检查点。
4. 同步文档: 在 `docs/models/supported_models.md` 的支持表格中删除 TeleChatForCausalLM 行, 避免给用户造成该别名仍受支持的误导。

关键文件:

- `vllm/model_executor/models/registry.py` (模块 模型注册表; 类别 source; 类型 data-contract): 核心变更文件。从 `_TEXT_GENERATION_MODELS` 和 `_EMBEDDING_MODELS` 中删除别名, 并在 `_PREVIOUSLY_SUPPORTED_MODELS` 中记录废弃版本。
- `tests/models/registry.py` (模块 测试; 类别 test; 类型 test-coverage): 配套测试文件。删除与废弃别名对应的测试用例 (引用 `chuhac/TeleChat2-35B` 检查点)。

- docs/models/supported_models.md (模块 文档; 类别 docs; 类型 documentation) : 文档同步更新。删除支持表格中 TeleChatForCausalLM 的行, 避免误导用户。

关键符号: 未识别

关键源码片段

vllm/model_executor/models/registry.py

核心变更文件。从 `_TEXT_GENERATION_MODELS` 和 `_EMBEDDING_MODELS` 中删除别名, 并在 `_PREVIOUSLY_SUPPORTED_MODELS` 中记录废弃版本。

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

# 在 _TEXT_GENERATION_MODELS 和 _EMBEDDING_MODELS 中删除以下行:
# "TeleChatForCausalLM": ("telechat2", "TeleChat2ForCausalLM"),

# 并在 _PREVIOUSLY_SUPPORTED_MODELS 中添加一行以记录最终支持版本:
# 注意: "0.25.0" 是最后一个包含该别名的版本
_PREVIOUSLY_SUPPORTED_MODELS = {
    # ... 其他条目
    "TeleChatForCausalLM": "0.25.0", # 在 0.26.0 版本中移除
}
```

评论区精华

在 review 中, 对 `_PREVIOUSLY_SUPPORTED_MODELS` 中的版本号存在讨论:

- DarkLight1337初次建议使用 "0.24.0", 认为 v0.25 已经发布。
- hmellor纠正指出: "It will be gone in 26, so the last supported version is 25", 建议改为 "0.25.0"。
- DarkLight1337随后承认错误 ("Oops... yeah")。最终采纳了 0.25.0 作为最终支持版本。该讨论体现了对于版本号含义的精确把握——`_PREVIOUSLY_SUPPORTED_MODELS` 记录的是最后一个支持的版本。
- 废弃版本号应为 0.24.0 还是 0.25.0 (correctness): 最终采用 "0.25.0" 作为版本号。

风险与影响

- 风险: 风险极低。TeleChatForCausalLM 只是一个向后兼容别名, 功能上等同于 TeleChat2ForCausalLM (两者均映射到同一个实现模块 telechat2 和类 TeleChat2ForCausalLM)。删除别名后, 使用旧架构名称 (如 chuhac/TeleChat2-35B) 的用户需要更新配置为 TeleChat2ForCausalLM, 否则会触发模型加载失败。该风险已通过测试数据和文档的同步更新来缓解。
- 影响: 影响范围: 限制在仍然依赖 TeleChatForCausalLM 别名的用户。TeleChat2ForCausalLM 等其他 TeleChat 变体不受影响。影响程度: 低。该别名继承自早期版本的命名方案, 使用量极低 (近 3 个月不足 10 小时), 且已有明确的替代名称。正向影响: 减少注册表字典中的噪声, 降低用户在创建新模型时意外使用已废弃别名的可能性。

- 风险标记：暂无

关联脉络

- 暂无明显关联 PR