

PR #45127 完整报告

vllm-project/vllm

[Model] Remove obsolete ERNIE models

合并时间: 2026-06-10 20:54

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45127>

执行摘要

- 一句话: 删除低使用率的 ERNIE BERT 式池化模型
- 推荐动作: 可精读以下设计要点:
 - 如何基于使用率数据 (遥测) 做代码功能退役决策。
 - 如何通过 `_PREVIOUSLY_SUPPORTED_MODELS` 处理功能删除后的向后兼容性。
 - 测试删除时如何同步清理测试用例和示例模型条目。该 PR 是 vLLM 模型管理策略的典型案例, 值得模型运维和代码清理相关开发人员关注。

功能与动机

PR 正文指出: 根据 `vllm-release` 遥测数据, `ErnieModel` 仅记录 31 个实例 / 0.56 GPU 小时, 而 `Ernie 4.5` 系列在同一时间段内约有 76000 GPU 小时, `ErnieModel` 占比仅 ~0.001%。`ErnieForSequenceClassification` 和 `ErnieForTokenClassification` 零使用。为了减少维护负担和代码体积, 决定移除这些几乎无用的模型代码。

实现拆解

变更分为四步:

1. 删除主实现文件: `vllm/model_executor/models/ernie.py` 全部删除 (-247 行), 包括 `ErnieEmbedding`、`ErnieModel`、`ErniePoolingModel`、`ErnieEmbeddingModel`、`ErnieForSequenceClassification`、`ErnieForTokenClassification` 等类及其辅助函数。
2. 更新模型注册表: 在 `vllm/model_executor/models/registry.py` 中, 从 `_EMBEDDING_MODELS`、`_TOKEN_CLASSIFICATION_MODELS`、`_SEQUENCE_CLASSIFICATION_MODELS` 字典中移除三者的条目, 并在 `_PREVIOUSLY_SUPPORTED_MODELS` 中添加它们, 版本设为 0.23.0 (经 review 修正: 这些模型在 v0.24 被移除, 因此最后支持的版本是 v0.23)。
3. 清理测试套件: 删除专用测试文件 `tests/models/language/pooling_mteb_test/test_ernie.py`; 在 `tests/models/language/pooling/test_token_classification.py` 中移除 `Ernie` 测试模型并重命名测试函数; 在 `tests/models/language/pooling/test_classification.py` 中移除 `Ernie` 参数及相关 `atol` 特殊处理; 在 `tests/models/registry.py` 中移除对应的示例模型条目。
4. 更新文档: 在 `docs/models/pooling_models/classify.md`、`embed.md`、`token_classify.md` 中删除对应的 ERNIE 模型行。

关键文件：

- `vllm/model_executor/models/ernie.py`（模块 模型实现；类别 `source`；类型 `deletion`；符号 `ErnieEmbedding`, `ErnieModel`, `ErniePoolingModel`, `ErnieEmbeddingModel`）：被删除的主实现文件，包含所有 ERNIE-3.0 池化模型类（`ErnieEmbedding`、`ErnieModel`、`ErniePoolingModel`、`ErnieEmbeddingModel`、`ErnieForSequenceClassification`、`ErnieForTokenClassification`）及权重加载逻辑，是整个 PR 的核心删除对象。
- `vllm/model_executor/models/registry.py`（模块 注册表；类别 `source`；类型 `data-contract`）：模型注册表的核心修改：从三个注册字典（`_EMBEDDING_MODELS`、`_TOKEN_CLASSIFICATION_MODELS`、`_SEQUENCE_CLASSIFICATION_MODELS`）中删除 ENNIE 模型条目，并在 `_PREVIOUSLY_SUPPORTED_MODELS` 中添加它们以记录最终支持版本 `v0.23`。这是确保删除后模型查找不会报错的关键数据契约变更。
- `tests/models/language/pooling_mteb_test/test_ernie.py`（模块 测试；类别 `test`；类型 `deletion`；符号 `test_embed_models_mteb`, `test_embed_models_correctness`）：被删除的专用测试文件，包含 `shibing624/text2vec-base-chinese-sentence` 模型的 MTEB 和正确性测试。保持测试清理的完整性。

关键符号：`ErnieEmbedding`, `ErnieModel`, `ErniePoolingModel`, `ErnieEmbeddingModel`, `ErnieForSequenceClassification`, `ErnieForTokenClassification`, `load_weights`

关键源码片段

`vllm/model_executor/models/ernie.py`

被删除的主实现文件，包含所有 ERNIE-3.0 池化模型类（`ErnieEmbedding`、`ErnieModel`、`ErniePoolingModel`、`ErnieEmbeddingModel`、`ErnieForSequenceClassification`、`ErnieForTokenClassification`）及权重加载逻辑，是整个 PR 的核心删除对象。

```
class ErnieEmbedding(BertEmbedding):
    # 在 BERT 嵌入基础上增加 task_type_embeddings
    def __init__(self, config: BertConfig):
        super().__init__(config)
        task_type_vocab_size = max(1, getattr(config, 'task_type_vocab_size', 1))
        self.task_type_embeddings = VocabParallelEmbedding(
            task_type_vocab_size, config.hidden_size
        )
    def forward(self, input_ids, position_ids, inputs_embeds=None):
        # 计算 token_type、position 和 task_type 三种嵌入并相加
        ...

@default_pooling_type(seq_pooling_type='CLS')
class ErnieModel(BertModel):
    # 使用 ErnieEmbedding 作为嵌入层
    def __init__(self, *, vllm_config: VllmConfig, prefix: str = ''):
        super().__init__(vllm_config=vllm_config, prefix=prefix, embedding_class=ErnieEmbedding)

class ErnieEmbeddingModel(BertEmbeddingModel):
    def _build_model(self, vllm_config, prefix):
```

```

    return ErnieModel(vllm_config=vllm_config, prefix=prefix)
def load_weights(self, weights):
    # 处理权重前缀映射 ('model.' 和 'ernie.' 前缀) 以及旧格式后缀映射 ('gamma'-'>'weight',
    'beta'-'>'bias')
    ...

```

vllm/model_executor/models/registry.py

模型注册表的核心修改：从三个注册字典（_EMBEDDING_MODELS、_TOKEN_CLASSIFICATION_MODELS、_SEQUENCE_CLASSIFICATION_MODELS）中删除 ENNIE 模型条目，并在 _PREVIOUSLY_SUPPORTED_MODELS 中添加它们以记录最终支持版本 v0.23。这是确保删除后模型查找不会报错的关键数据契约变更。

```

# _EMBEDDING_MODELS 中删除了以下行
# 'ErnieModel': ('ernie', 'ErnieEmbeddingModel'),

# _TOKEN_CLASSIFICATION_MODELS 中删除了以下行
# 'ErnieForTokenClassification': ('ernie', 'ErnieForTokenClassification'),

# _SEQUENCE_CLASSIFICATION_MODELS 中删除了以下行
# 'ErnieForSequenceClassification': ('ernie', 'ErnieForSequenceClassification'),

# _PREVIOUSLY_SUPPORTED_MODELS 中新增了以下记录
'ErnieModel': '0.23.0',
'ErnieForSequenceClassification': '0.23.0',
'ErnieForTokenClassification': '0.23.0',

```

评论区精华

Review 中主要有三个讨论点：

- 要求更新 _PREVIOUSLY_SUPPORTED_MODELS: DarkLight1337 要求将删除的模型添加到 _PREVIOUSLY_SUPPORTED_MODELS 中，以便用户知道需要使用旧版 vLLM。作者按要求执行。
- 版本号纠正: hmellor 指出作者最初将版本写为 0.22.1 是错误的，因为删除将在 v0.24 中出现，所以最后支持的版本是 v0.23。作者随后修正为 0.23.0。
- 测试恢复: noooop 要求恢复对测试文件 `tests/models/language/pooling/test_classification.py` 中与 ERNIE 相关的修改。作者回应撤回了部分改动并移除了 atol 参数，观察 CI 结果。
 - 更新 _PREVIOUSLY_SUPPORTED_MODELS 以保留向后兼容指引 (design): 已添加三条记录，版本最终修正为 0.23.0。
 - 版本号修正: 删除生效于 v0.24，最后支持版本应为 v0.23 (correctness): 版本号正确修正。
 - 测试修改: 恢复并调整测试中的 ERNIE 相关改动 (testing): 调整完成，移除了 ERNIE 测试参数并重命名了测试函数，保留了基本的 BERT 测试结构。

风险与影响

- 风险：由于这些模型使用率极低（ErnieModel 仅占 0.001% 份额，另两个零使用），且新 Ernie 4.5 系列完全不受影响（使用独立实现文件），删除风险极低。唯一需要注意的是通过 `_PREVIOUSLY_SUPPORTED_MODELS` 保留了向后兼容指引，用户若仍使用旧模型会被引导至 v0.23 及更早版本。测试文件中移除了 ERNIE 相关参数，但这属于计划中的清理。无安全或性能风险。
- 影响：用户影响：几乎无影响。原有的极少数 ErnieModel 用户（约占 0.001%）需回退到 v0.23 或更早版本。Ernie 4.5 系列用户无影响。系统影响：代码库减小约 250 行，消除了一个自包含但低价值的模块。团队影响：减少了后续维护负担，但需注意与 future 可能的重构（如模型加载器统一）没有冲突。
- 风险标记：低使用率代码清理，向后兼容性通过 `_PREVIOUSLY_SUPPORTED_MODELS` 记录

关联脉络

- PR #36385 [Model] Add ERNIE-3.0 pooling models (ErnieModel, ErnieForSequenceClassification, ErnieForTokenClassification): 此 PR 是 #36385 的逆操作，移除了当时引入的模型。两个 PR 直接构成添加 / 删除的完整生命周期。