

# PR #41513 完整报告

vllm-project/vllm

[Doc] Add Qwen3-30B-A3B-Thinking-2507-FP8 to batch invariance verified models

合并时间: 2026-05-03 20:47

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/41513>

## 执行摘要

本 PR 为纯文档变更, 将 `Qwen/Qwen3-30B-A3B-Thinking-2507-FP8` 模型添加到批处理不变性 (batch invariance) 已验证模型列表中。仅修改 `docs/features/batch_invariance.md`, 增删各一行, 无代码影响。

## 功能与动机

该模型已通过批处理不变性测试和 mypy 测试, 因此需要更新文档, 确保模型列表与实际验证状态一致。

## 实现拆解

- 修改 `docs/features/batch_invariance.md` 第 108 行, 在 Qwen3 (MoE) 模型的 Markdown 列表末尾追加 `Qwen/Qwen3-30B-A3B-Thinking-2507-FP8`。

无需展示, 仅文档文本变更。

## 评论区精华

无实质性讨论。

## 风险与影响

风险: 无。纯文档变更, 不影响任何代码逻辑。

影响: 极小。帮助用户识别经过验证的模型, 提升文档准确性。

## 关联脉络

此 PR 与其他历史 PR 无直接关联, 属于独立的文档维护任务。