PR #44435 完整报告

vllm-project/vllm

[Doc] Add Llama-3.2-3B-Instruct to batch-invariance tested models

合并时间：2026-06-06 07:04

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44435

执行摘要

- 一句话：更新 batch-invariance 文档：添加 Llama-3.2-3B 并放宽 GPU 要求
- 推荐动作：简单的文档更新，已验证通过，可直接合并。对于关注 batch-invariance 的用户提供准确信息。

功能与动机

来自 issue #27433 的讨论，用户发现 Llama-3.2-3B 在 batch-invariance 测试中正确但未被列入文档。审核者 yewentao256 批准添加。同时，根据实际测试和 PR #42456 的 SM 8.x CUDA graphs 支持，正式放宽硬件要求。

实现拆解

纯文档修改，仅涉及 docs/features/batch_invariance.md。

1. 将 Llama 3 条目从具体模型枚举改为系列表述：Llama3.1 and 3.2 series, meta-llama/Llama-3.2-3B-Instruct for example。
2. 硬件要求从 compute capability 9.0 or higher 改为 8.0 or higher，并删除 H-series/B-series 设备列表。
3. 根据审核反馈，进一步移除刚添加的 Ampere/Ada Lovelace 设备列表，只保留计算能力描述。

关键文件：

- docs/features/batch_invariance.md（模块 文档；类别 docs；类型 documentation）：唯一变更文件，包含硬件要求放宽和模型列表更新，是 PR 核心。

关键符号：未识别

评论区精华

审核者 yewentao256 提出三项建议：

1) 不维护具体 GPU 设备列表（"Let's not maintain this device list"）； 2) 不列举具体模型而用系列代替（"Let's not maintain specific model"）； 3) 完全移除设备型号仅保留计算能力（"Let's remove this directly, so we don't need to maintain the devices"）。提交者全部采纳，体现文档可维护性优先。

- GPU 设备列表维护 (documentation): 提交者完全采纳，最终硬件要求只保留计算能力描述。
- 模型列表简化 (documentation): 提交者采纳，使用 'Llama3.1 and 3.2 series, meta-llama/Llama-3.2-3B-Instruct for example' 格式。

风险与影响

- 风险：无技术风险。文档内容变更经过实际测试验证，但用户可能依赖文档判断硬件兼容性，需确保放宽要求准确。已通过 SM 8.9 和 B300 验证。
- 影响：对用户：明确 batch-invariance 支持 NVIDIA SM 8.x 及以上 GPU，并确认 Llama-3.2-3B 通过测试。对系统：无运行时影响。对团队：减少设备列表维护工作。
- 风险标记：暂无

关联脉络

- PR #27433 Discussion on adding Llama-3.2-3B to batch-invariance tested models: 该 issue 触发了本 PR 的动机和讨论。
- PR #42456 Enable CUDA graphs on SM 8.x: 该 PR 为硬件要求放宽到 SM 8.0 提供了技术基础。