PR #38723 完整报告

vllm-project/vllm

Fix shape comment in extract_hidden_states example

合并时间：2026-04-01 22:29

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38723

执行摘要

该 PR 修复了离线推理示例中隐藏状态张量形状的文档注释错误，将序列长度维度从第二维更正为第一维。这是一个纯粹的文档修正，不涉及任何代码逻辑变更，风险极低，仅影响示例文档的准确性。

功能与动机

根据 PR 描述，目的是修复形状注释。系统实际输出的隐藏状态张量形状为 [prompt len, num_hidden_layers, hidden size]，但原注释错误地描述为 [num_hidden_layers, prompt len, hidden size]。PR body 明确指出："The system actually outputs a tensor with seq_len on the first dim"。

实现拆解

仅修改了一个文件：

- *examples/offline_inference/extract_hidden_states.py*：将第 57 行的注释从 # [num_hidden_layers, prompt len, hidden size] 改为 # [prompt len, num_hidden_layers, hidden size]，以准确反映张量形状。

评论区精华

review 讨论非常简短，所有 reviewer 均表示认可：

- *gemini-code-assist[bot]*："更新了示例中的注释以正确反映提取的隐藏状态的形状。"
- MatthewBonanni 和 mgoin 均批准，无额外评论。 没有争议或未解决的疑虑。

风险与影响

- 风险：无。仅修改注释，不涉及代码逻辑、性能或安全变更。
- 影响：确保示例文档准确，避免开发者基于错误注释产生误解。对系统无功能或性能影响。

关联脉络

- 与 PR #38722（修复 harmony_utils.py 文档字符串拼写错误）类似，同为文档修正类 PR，反映仓库对文档准确性的持续维护。
- 近期历史 PR 中，文档修正类变更通常标记为 documentation 和 cleanup 标签，本 PR 也遵循此模式。