PR #43085 完整报告

vllm-project/vllm

[Test] Replace zephyr-7b-beta (7B) with SmolLM2-135M in tokenization test

合并时间：2026-05-21 16:57

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43085

执行摘要

- 一句话：替换 tokenization 测试模型为 SmolLM2-135M，节省 GPU 资源
- 推荐动作：该 PR 变更简单清晰，值得作为 CI 测试资源优化的参考案例。建议后续类似测试模型替换时参考此做法：优先选用 135M 级别模型，确保有 chat template 即可。

功能与动机

PR body 明确指出：原测试使用 7B 模型导致不必要的 GPU 内存开销，替换为 135M 模型可节省约 7GB GPU 内存，同时保持测试有效性。模型缩小约 50 倍，CI 资源占用更低。

实现拆解

1. 在 tests/entrypoints/serve/tokenize/test_tokenization.py 中将 MODEL_NAME 从 "HuggingFaceH4/zephyr-7b-beta" 改为 "HuggingFaceTB/SmolLM2-135M-Instruct"。
2. 仅修改一行（+1/-1），无其他逻辑或配置变更。
3. 通过 CI 验证所有 9 个测试均通过（build #67000）。
4. 历史提交曾尝试使用 Qwen3-0.6B 和 Qwen3-1.7B，但因推理质量或 token budget 问题最终回退，选用 SmolLM2-135M 作为最终方案。

关键文件：

- tests/entrypoints/serve/tokenize/test_tokenization.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件，将 MODEL_NAME 从 7B 模型替换为 135M 模型，直接影响测试资源消耗。

关键符号：未识别

关键源码片段

tests/entrypoints/serve/tokenize/test_tokenization.py

唯一变更文件，将 MODEL_NAME 从 7B 模型替换为 135M 模型，直接影响测试资源消耗。

```
# tests/entrypoints/serve/tokenize/test_tokenization.py
# 原为 HuggingFaceH4/zephyr-7b-beta (7B)，现替换为
# HuggingFaceTB/SmolLM2-135M-Instruct (135M)
# 模型大小缩小约 50 倍，节省 ~7GB GPU 内存
# 该测试仅校验 tokenize / detokenize / tokenizer_info 端点
```

```
# 不依赖模型推理能力，任何带有 chat template 的模型均可

# any model with a chat template should work here
MODEL_NAME = "HuggingFaceTB/SmolLM2-135M-Instruct"
```

评论区精华

无实质性 review 讨论。DarkLight1337 直接 approve。gemini-code-assist[bot] 自动评论无反馈。Mergify 机器人提示过 pre-commit 失败，但未影响最终合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及测试 fixture 中的模型名，测试逻辑不变。新模型 SmolLM2-135M-Instruct 同样支持 chat template，可覆盖所有 tokenize / detokenize / tokenizer_info 端点测试。
- 影响：影响范围仅限于 tokenization 测试，CI 资源消耗显著降低，有助于提升 CI 并行速度和稳定性。对其他模块无影响。
- 风险标记：暂无

关联脉络

- PR #43245 [Benchmark] Add num-warmup to vllm bench throughput: 同为测试 / 基准测试优化变更，体现 CI 资源优化的趋势。