PR #38114 完整报告

vllm-project/vllm

[Bugfix] Add missing ASRDataset import and CLI args in benchmarks/throughput.py

合并时间：2026-04-08 21:53

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38114

执行摘要

此 PR 修复了 vLLM 基准测试脚本中 ASRDataset 导入缺失的 bug，通过添加导入和两个新 CLI 参数，使自动语音识别（ASR）模型的吞吐量基准测试可配置且正常运行。变更仅影响 benchmark 模块，风险低，适用于使用 ASR 模型的研究场景。

功能与动机

为什么做：在 vLLM 的吞吐量基准测试脚本 throughput.py 中，ASRDataset 类（定义于 datasets.py）缺少导入，导致当用户尝试使用 ASR 模型（如 OpenAI Whisper）和对应数据集（如 openslr/librispeech_asr）进行基准测试时，出现 ValueError: openslr/librispeech_asr is not supported by hf dataset 错误。PR 旨在修复此问题，并添加可配置的音频长度过滤参数，以支持不同 ASR 模型的测试需求。

实现拆解

改动文件：vllm/benchmarks/throughput.py

1. 导入添加：在文件顶部导入模块中增加 ASRDataset。
2. 数据集处理逻辑：在 get_requests 函数中，添加条件分支处理 ASR 数据集路径，设置 dataset_cls 并传递新参数到 sample_kwargs。
3. 参数验证：在 validate_args 函数中，将 ASRDataset.SUPPORTED_DATASET_PATHS 加入数据集路径集合，确保 ASR 数据集使用 vLLM 后端。
4. CLI 参数扩展：在 add_cli_args 函数中，新增两个浮点型参数：
   - --asr-min-audio-len-sec：默认 0.0，最小音频时长（秒）。
   - --asr-max-audio-len-sec：默认 inf，最大音频时长（秒）。

评论区精华

review 过程中无技术讨论交锋。仅有代码助手 bot 的总结性评论和维护者的快速批准，表明变更简单直接，无需深入评审。

风险与影响

风险：

- 依赖风险：ASRDataset 需要 torchcodec 包，非 vLLM 默认依赖，用户需手动安装，否则可能导致运行时错误。
- 参数风险：新 CLI 参数使用 float('inf') 作为默认最大值，需确保浮点运算兼容性，但标准做法风险低。

影响：

- 用户影响：仅影响使用 ASR 模型进行基准测试的用户，修复后可直接运行测试，提升开发体验。
- 系统影响：不变更核心推理逻辑，仅扩展 benchmark 脚本功能，无性能或安全影响。

关联脉络

从提供的近期历史 PR 分析中，未发现直接修改相同文件或涉及 ASR 数据集的 PR，表明此变更为独立 bugfix。然而，仓库中已有 multi-modality 标签用于多模态模型支持（如 PR 39232 添加视觉语言模型），此 PR 将 ASR 音频处理纳入同类范畴，反映了 vLLM 在扩展基准测试覆盖多模态场景的趋势。