PR #41576 完整报告

vllm-project/vllm

Implement custom dataset class for ASR benchmarking

合并时间：2026-05-12 12:17

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41576

执行摘要

- 一句话：为 ASR 基准测试添加自定义音频数据集
- 推荐动作：该 PR 是基准测试工具的重要补充，设计上考虑了向后兼容和清晰的命名，对需要自定义音频数据集的用户具有实际价值。建议阅读以了解基准测试框架扩展方式，并考虑补充单元测试。

功能与动机

支持用户使用本地音频数据集对 ASR 模型（如 Whisper）和多模态音频 - 文本模型（如 Qwen2-Audio）进行基准测试，填补了自定义音频数据集支持的空白。PR 描述中明确提到：'This PR adds support for benchmarking ASR models using a custom local dataset.'

实现拆解

1. 添加可选依赖导入：在 vllm/benchmarks/datasets/datasets.py 中添加 soundfile 的 try-except 导入，避免未安装时模块崩溃。
2. 新增 process_audio 工具函数：该函数统一处理文件路径、HuggingFace 字典和元组三种输入格式，返回 (array, sr) 元组，为 CustomAudioDataset 提供音频预处理。
3. 新增 CustomAudioDataset 类：继承自 CustomDataset，从 JSONL 文件中读取 prompt 和 audio 字段，调用 process_audio 处理音频，并构造 SampleRequest 对象，支持 ASR 和多模态音频模型。
4. 重命名自定义多模态数据集：将 CustomMMDataset 重命名为 CustomImageDataset，CLI 参数 --dataset-name custom_mm 保留为别名并添加弃用警告，建议改为 custom_image。
5. CLI 集成与文档更新：在 get_samples 中添加 custom_audio 分支；在 add_dataset_parser 中注册新选项；更新 docs/benchmarking/cli.md，增加 custom audio 和 custom image 的表格及使用说明。

关键文件：

- vllm/benchmarks/datasets/datasets.py（模块 基准测试；类别 source；类型 core-logic；符号 process_audio, CustomMMDataset, CustomImageDataset, CustomAudioDataset）：核心变更文件：新增 process_audio 函数、CustomAudioDataset 类，重命名 CustomMMDataset 为 CustomImageDataset，并修改 get_samples 函数以支持新数据集。

- vllm/benchmarks/datasets/__init__.py（模块 基准测试；类别 source；类型 dependency-wiring）：更新模块导出列表，添加 CustomAudioDataset、CustomImageDataset 和 process_audio 的公共接口。
- docs/benchmarking/cli.md（模块 文档；类别 docs；类型 documentation）：更新文档：增加 custom audio 和 custom image 数据集的使用说明和示例。

关键符号：process_audio, CustomImageDataset.sample, CustomAudioDataset.sample, get_samples

评论区精华

- 可导入性问题：gemini-code-assist[bot] 指出 import soundfile 应在顶级使用 try-except 保护，避免模块未安装时整个 datasets 模块加载失败。作者随后添加了相应的错误处理。
- skip_chat_template 缺失：gemini-code-assist[bot] 发现 CustomAudioDataset.sample 方法缺少 skip_chat_template 参数，导致 CLI 选项对该数据集类型不生效。作者后续补全了该参数。
- 命名讨论：DarkLight1337 建议将 custom_audio 纳入 custom_mm 或考虑更清晰的命名。ymoslem 提议将 custom_mm 重命名为 custom_image，并保留旧名称的向后兼容性。最终接受此方案，添加了弃用警告。
- 弃用版本号：DarkLight1337 要求将 deprecation warning 中的版本改为 v0.24，作者已修改。
  - soundfile 导入缺少 try-except 保护 (correctness): 作者添加了 try-except 导入和 PlaceholderModule 备用。
  - CustomAudioDataset.sample 缺少 skip_chat_template 参数 (correctness): 作者添加了 skip_chat_template 参数到 sample 调用。
  - custom_audio 与 custom_mm 的命名集成 (design): 决定重命名 CustomMMDataset 为 CustomImageDataset，添加 custom_image 作为新名称，custom_mm 作为别名并弃用。
  - 弃用版本号从 3 minor 改为 v0.24 (style): 作者更新为 'v0.24'。

风险与影响

- 风险：
  - soundfile 是可选依赖，代码已添加 try-except，但在 process_audio 中若 sf 为 PlaceholderModule 且用户传入文件路径时，会抛出 PlaceholderModule 的错误，而非清晰的提示。需确保错误信息用户友好。
  - 新的 CustomAudioDataset 类缺少单元测试，可能导致回归不易察觉。
  - 命名变更（custom_mm 到 custom_image）虽然保留向后兼容，但用户可能依赖于旧名，弃用警告仅日志输出，不足以及时通知用户。
- 影响：
  - 用户：新增了使用自定义音频数据集进行基准测试的能力，ASR 模型和音频多模态模型均可，--dataset-name custom_audio 可用；原来的 custom_mm 用户收到弃用警告。

- 系统：增加了可选依赖 soundfile，无性能影响。
- 团队：维护成本低，功能独立，但需关注测试覆盖率。
- 风险标记：可选依赖保护不完整，缺少单元测试，命名变更过渡期

关联脉络

- 暂无明显关联 PR