PR #22181 完整报告

sgl-project/sglang

[refactor] [asr] Add transcription adapter for extensible ASR models support

合并时间：2026-04-09 01:19

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22181

执行摘要

本 PR 通过引入转录适配器框架，重构了 ASR（自动语音识别）模型的支持逻辑，移除硬编码的模型检测和分支，提升系统的可扩展性和可维护性。主要变更集中在 OpenAI 兼容的转录端点处理，对用户透明，但为未来添加新 ASR 模型奠定了基础。

功能与动机

动机源于对 PR #22073 的后续跟进，目标是移除硬编码的模型家族检测和分支。根据 PR 描述：'Follow up of #22073. To remove hardcoded model family detection and branching.' 这解决了现有代码中直接检测模型架构（如 Whisper 或 Qwen3-ASR）并分支处理的问题，旨在支持更多 ASR 模型的轻松集成，避免未来修改核心服务逻辑。

实现拆解

实现方案按模块拆解如下：

- 新增适配器框架：在 python/sglang/srt/entrypoints/openai/transcription_adapters/ 下添加 base.py、qwen3_asr.py 和 whisper.py。base.py 定义 TranscriptionAdapter 抽象基类，提供 build_sampling_params、postprocess_text 和 build_verbose_response 方法，并通过 @register_transcription_adapter 装饰器实现注册机制。
- 修改核心服务：在 serving_transcription.py 中，移除原有的 _detect_model_family 函数，改为调用 resolve_adapter 函数，根据模型架构名动态选择适配器处理请求。例如：python self._adapter = resolve_adapter(getattr(model_config.hf_config, "architectures", []))
- 调整配置和处理器：修改 configs/qwen3_asr.py 和 multimodal/processors/qwen3_asr.py，提取硬编码字符串为常量，以适配新框架。
- 添加测试：新增 test/manual/models/test_qwen3_asr.py，提供手动测试脚本验证 Qwen3-ASR 模型的转录功能。

评论区精华

Review 中仅有一条评论，来自 mickqian，在 model_config.py 中建议：

> maybe detect this by whether we have dedicated processor/adaptor for this model, instead of maintaining a hard-coded list, in the future

这表明了设计上的前瞻性思考，旨在进一步提升扩展性，但未在本次 PR 中实现，状态为已解决。

风险与影响

技术风险：

1. 适配器注册依赖模型架构名匹配，如果新模型架构名不包含注册键（如 'Whisper' 或 'Qwen3ASR'），可能导致解析失败。
2. 修改 serving_transcription.py 的核心逻辑可能引入回归，影响现有 Whisper 和 Qwen3-ASR 模型的转录准确性或性能。
3. 测试覆盖不足，新增测试为手动脚本，缺乏自动化单元测试，可能增加维护负担。

影响分析：

- 对用户：API 保持兼容，无感知变更。
- 对系统：提高可扩展性，未来添加新 ASR 模型只需实现新适配器，无需修改服务逻辑。
- 对团队：需遵循适配器模式进行开发，增加学习曲线，但提升代码一致性。

关联脉络

本 PR 与近期历史 PR 紧密关联：

- PR #22073 和 #22089：被引用为动机来源，同为 ASR 功能扩展 PR，涉及 Qwen3-ASR 的流式转录支持，共同推进 ASR 模块的演进。
  - 从仓库历史看，ASR 相关 PR（如 #22089）逐渐从硬编码转向可扩展设计，本 PR 是这一趋势的体现，展示了系统向模块化和可维护性方向演进。