PR #37819 完整报告

vllm-project/vllm

[Docs] Add guide for editing agent instruction files

合并时间：2026-03-25 21:54

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37819

执行摘要

本次 PR 添加了代理指令编辑指南，通过设置令牌预算（如 AGENTS.md 不超过 200 行）和内容归属规则，旨在避免 AGENTS.md 文件膨胀，提升代理开发体验的规范性和一致性，影响文档维护流程和代理行为。

功能与动机

动机源于避免 AGENTS.md 因内容过度积累而变得臃肿，同时为引用领域特定信息提供清晰路径。PR body 中明确指出目标是“避免 AGENTS.md 膨胀，同时给出来自 AGENTS.md 引用额外领域特定信息的清晰路径，以改善 vLLM 的代理开发体验”。Issue 评论中进一步讨论了需要为 AGENTS.md 建立指南，以确保内容组织合理，避免反模式如“反应性积累”。

实现拆解

改动主要涉及两个文件，按模块拆解如下：

- AGENTS.md：修改添加“Domain-Specific Guides”部分，链接到新指南，强调在修改代码前必须先阅读指南。
- docs/contributing/editing-agent-instructions.md：新增文件，包含以下关键部分：
  - 令牌预算：AGENTS.md 必须保持在 200 行以内，领域指南不超过 300 行，以确保加载效率。
- 何时不添加内容：列出准则如“代理已经做过了”或“一次性事件”，避免不必要规则。
- 内容归属：使用表格区分项目级内容（如贡献政策）和领域级内容（如模型模式），规则是“如果只涉及一个领域，放入领域指南”。
- 好的领域指南：强调添加代理无法从代码推断的特定约定，例如跨文件上下文和重复错误的修复。
- 反模式：如“反应性积累”和“指南间复制粘贴”，以保持文档精简。

评论区精华

review 讨论中的精华点包括：

- markmc评论：“worth a try - I think we'll be doing a lot of experimentation with this ... The domain-specific guides thing is aka 'progressive disclosure'”，这揭示了设计决策，即通过渐进式披露模式管理内容，避免主文件过载。

- issue 评论中，bbrowning 与 markmc 讨论将现有内容从 AGENTS.md 移动到领域指南的可能性，但指出“AGENTS.md 始终加载在模型的上下文中，而链接的内容仅在模型认为相关时加载”，这强调了无条件规则需保留在主文件中，影响代理行为差异。

风险与影响

- 风险：具体风险包括代理若严格遵循指南，可能拒绝某些原本接受的修改（如测试中复制整个 README.md 到 AGENTS.md 被拒绝），影响开发流程灵活性；令牌预算限制可能导致内容拆分需求，增加维护复杂性和潜在的文档碎片化。无直接性能、安全或兼容性风险，但间接影响代理响应一致性。
- 影响：对开发团队：建立文档维护规范，促进代码库一致性和长期可维护性，可能需调整现有工作流；对代理行为：可能微调代理对指令的响应，基于指南拒绝不当修改；对用户：提供清晰编辑路径，提升代理开发体验，但需适应新规则和预算限制。

关联脉络

与历史 PR 和关联 Issue 的关系揭示了更大的功能演进方向：

- 相关 PR：如 #37840（添加 vllm-musa 到 custom_op.md）和 #37914（添加 ViT CUDA Graphs 章节），均为文档增强，涉及领域特定内容的添加，反映团队对文档结构化和渐进式披露的重视，形成文档维护的统一趋势。
- 关联 Issue：无直接关联 Issue，但 issue 评论中的讨论暗示未来可能重构 AGENTS.md 内容到领域指南，如“much of the existing AGENTS.md could be moved out into domain-specific guides”，这指向团队可能进行的实验性内容重组，以优化代理指令加载和用户体验。