

PR #36897 完整报告

sgl-project/sglang

Decouple speculative draft capacity from runtime state

合并时间: 2026-08-31 14:31

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36897>

执行摘要

- 一句话:
- 推荐动作:

功能与动机

实现拆解

1. 核心源码路径 - `python/sglang/srt/runtime_context.py` (core-logic) : 源码主路径; 涉及符号 `_adaptive_draft_token_bound`; 包含 导入关系调整、控制流调整、配置键调整; +25/-29; 关键符号: `_adaptive_draft_token_bound` -
`python/sglang/srt/speculative/spec_info.py` (core-logic) : 源码主路径; 涉及符号 `resolve_max_speculative_num_draft_tokens`; 包含 导入关系调整、控制流调整、配置键调整; +23/-0; 关键符号: `resolve_max_speculative_num_draft_tokens` -
`python/sglang/srt/arg_groups/overrides.py` (dependency-wiring) : 源码主路径; 包含 导入关系调整、控制流调整、配置键调整; +14/-14
2. 测试与验证 - `test/registered/unit/test_runtime_context.py` (test-coverage) : 测试配套; 包含 测试覆盖调整、配置键调整; +3/-8 - `test/registered/unit/server_args/test_server_args.py` (test-coverage) : 测试配套; 包含 测试覆盖调整; +2/-0

关键文件:

- `python/sglang/srt/runtime_context.py` (模块 `sglang/srt`; 类别 `source`; 类型 `core-logic`; 符号 `_adaptive_draft_token_bound`) : 源码主路径; 涉及符号 `_adaptive_draft_token_bound`; 包含 导入关系调整、控制流调整、配置键调整; +25/-29
- `python/sglang/srt/speculative/spec_info.py` (模块 `sglang/srt`; 类别 `source`; 类型 `core-logic`; 符号 `resolve_max_speculative_num_draft_tokens`) : 源码主路径; 涉及符号 `resolve_max_speculative_num_draft_tokens`; 包含 导入关系调整、控制流调整、配置键调整; +23/-0
- `python/sglang/srt/arg_groups/overrides.py` (模块 `sglang/srt`; 类别 `source`; 类型 `dependency-wiring`) : 源码主路径; 包含 导入关系调整、控制流调整、配置键调整; +14/-14
- `python/sglang/srt/speculative/spec_registry.py` (模块 `sglang/srt`; 类别 `source`; 类型 `core-logic`; 符号 `resolve_max_speculative_num_draft_tokens`) : 源码主路径; 涉及符号

resolve_max_speculative_num_draft_tokens; 包含 导入关系调整、符号定义调整;
+13/-0

- test/registered/unit/test_runtime_context.py (模块 runtime/context; 类别 test; 类型 test-coverage) : 测试配套; 包含 测试覆盖调整、配置键调整; +3/-8

关键符号: _adaptive_draft_token_bound, resolve_max_speculative_num_draft_tokens

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR