PR #25038 完整报告

sgl-project/sglang

[Spec] Rename `accepted_indices` -> `accept_indices`; drop `_token_id` suffix per Rule 5

合并时间：2026-05-12 13:29

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25038

执行摘要

- 一句话：投机解码字段重命名收尾
- 推荐动作：该 PR 可以作为团队命名规范执行的范例，展示了如何系统性地推进代码一致性。虽然不包含功能变更，但对于参与投机解码模块的开发者，了解这些命名规则有助于理解代码结构。

功能与动机

本次变更是 #25014 命名规范统一的最后一波，PR 描述明确说明 Last batch of spec-naming follow-ups missed during #25014，目的是保持投机解码相关标识符的一致性，遵循内部命名规则（Rule 1 去掉 -ed、Rule 5 对 tensor/list 去掉 _token_id 后缀）。

实现拆解

1. 在 scheduler_output_processor_mixin.py 中将私有方法 _resolve_spec_overlap_token_ids 重命名为 _resolve_spec_overlap_tokens，并更新内部所有调用。
2. 在 ngram_info.py 中将实例属性 accepted_indices 统一替换为 accept_indices，涉及 _fill_requests、_free_cache、_greedy_verify 等方法。
3. 在 eagle_worker.py 和 multi_layer_eagle_worker.py 中，将验证结果对象 res.accepted_indices 替换为 res.accept_indices，并同步更新注释。
4. 在 dflash_worker.py 中将函数内局部变量 out_token_ids 重命名为 out_tokens，去掉 _token_id 后缀。
5. 在 eagle_info.py 中将 EagleVerifyOutput 数据类字段 accepted_indices 重命名为 accept_indices，并调整构造调用。
6. 在 logprob.py 和 frozen_kv_mtp_worker.py 中做对应调用点调整。 第二个提交中保留了 curr_token_id 标量不变，因为 Rule 5 只适用于 tensor/list。

关键文件：

- python/sglang/srt/managers/scheduler_output_processor_mixin.py（模块 调度器；类别 source；类型 core-logic；符号 _resolve_spec_overlap_token_ids, _resolve_spec_overlap_tokens）：核心调度器 Mixin，定义了关键的 _resolve_spec_overlap_tokens 方法，负责投机解码重叠 token 的解析。

- python/sglang/srt/speculative/ngram_info.py（模块 投机解码；类别 source；类型 core-logic；符号 _fill_requests, _free_cache, _greedy_verify, accept_indices）：N-gram 投机解码核心实现，大量使用 accepted_indices 属性，重命名涉及多个方法。
- python/sglang/srt/speculative/eagle_worker.py（模块 投机解码；类别 source；类型 core-logic；符号 verify, _mamba_verify_update, accepted_indices, accept_indices）：Eagle 投机解码工作线程，验证后处理中大量使用 res.accepted_indices。
- python/sglang/srt/speculative/multi_layer_eagle_worker.py（模块 投机解码；类别 source；类型 core-logic；符号 verify, accepted_indices, accept_indices）：多层 Eagle 工作线程，与 eagle_worker.py 逻辑相同，同步重命名。
- python/sglang/srt/speculative/dflash_worker.py（模块 投机解码；类别 source；类型 core-logic；符号 _greedy_sample_from_vocab_parallel_head, out_token_ids, out_tokens）：Dflash 投机解码工作线程，局部变量 out_token_ids 重命名。
- python/sglang/srt/speculative/eagle_info.py（模块 投机解码；类别 source；类型 core-logic；符号 EagleVerifyOutput, verify, accepted_indices, accept_indices）：定义 EagleVerifyOutput 数据类，字段名变更影响整个投机解码链路。
- python/sglang/srt/layers/utils/logprob.py（模块 logprob；类别 source；类型 core-logic）：Logprob 处理层，引用投机解码验证输出，需同步重命名。
- python/sglang/srt/speculative/frozen_kv_mtp_worker.py（模块 投机解码；类别 source；类型 core-logic）：Frozen KV MTP 工作线程，引用投机解码验证输出，需同步重命名。

关键符号：_resolve_spec_overlap_tokens, _fill_requests, _free_cache, _greedy_verify, verify, _mamba_verify_update, _greedy_sample_from_vocab_parallel_head

评论区精华

无 review 评论。第二个提交体现了作者对命名规则的精细化理解：curr_token_id 作为标量不符合 Rule 5（仅 tensor/list）的适用范围，因此保持原名不变。

- 暂无高价值评论线程

风险与影响

- 风险：变更纯属标识符重命名，无运行时行为变化。主要风险在于可能遗漏未修改的引用，但作者仔细检查了所有调用点，且 CI 测试通过，风险极低。
- 影响：对用户无影响（无 API 或配置变化）。对内部开发者，提高了投机解码模块命名一致性，降低未来阅读混淆。变更涉及 8 个文件，51 行增删，但都是 1:1 替换。
- 风险标记：纯重命名无功能影响

关联脉络

- PR #25014 [Spec] Internal rename per N2 v2 naming rule: 本 PR 是 #25014 遗漏的最后一波重命名，确保命名规则全面覆盖。