# PR #25014 完整报告

sgl-project/sglang

[Spec] Internal rename per N2 v2 naming rule

合并时间：2026-05-12 09:16

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25014

## 执行摘要

- 一句话：统一投机解码内部标识符命名规范
- 推荐动作：该 PR 可作为大规模代码重构中推行命名规范的实践参考。建议关注其命名规则的设计理念，以及如何通过自动化替换和 CI 验证确保重构安全。非紧急需精读的内容，但对参与投机解码开发的团队成员有一定学习价值。

## 功能与动机

为统一投机解码模块命名规范，降低维护认知负担，并为后续外部重命名铺平道路。PR body 明确说明："Pure internal identifier rename. No external API change in this PR — meta_info JSON keys, Prometheus names, trace_slice keys, and paper-aligned spec_accept_rate / spec_accept_length are all preserved. External-facing renames + backward-compat aliases follow in a separate PR."

## 实现拆解

1. 定义重命名规则：遵循 N2 v2 规范 (drop -ed、accept 表 with-bonus、correct 表 drafts-only、num_X 前缀 )。
2. 全局替换内部标识符：在投机解码核心模块（Eagle、NGram、dflash 等）及关联调度器、管理器、连接器中进行替换，涉及 eagle_info.py、spec_utils.py、eagle_worker_v2.py、scheduler_metrics_mixin.py、schedule_batch.py 等 45 个文件。
3. 语义修正：将 num_accepted_drafts_filter 更正为 num_accept_tokens_filter，因为该值实际为 num_correct_drafts + 1（含 bonus 的 accept 数量），旧名称易误解。
4. 保留对外接口：所有 Prometheus 指标名、meta_info JSON key、trace_slice key 保持不变，内部重命名不影响外部可见性。
5. 测试配套更新：同步修改测试文件中的变量引用，确保测试通过。

关键文件：

- python/sglang/srt/observability/scheduler_metrics_mixin.py（模块 度量层；类别 source；类型 core-logic；符号 update_spec_metrics, spec_num_accepted_tokens, spec_total_num_accepted_tokens）：核心度量统计，展示了 num_accepted_drafts → num_correct_drafts 和 num_accepted_tokens → num_accept_tokens 的重命名，以及 update_spec_metrics 参数重命名。

- python/sglang/srt/speculative/spec_utils.py（模块 投机工具；类别 source；类型 core-logic；符号 create_num_accepted_drafts_filter, create_num_accept_tokens_filter, get_target_cache_loc, get_src_tgt_cache_loc）：投机解码工具函数的重命名，包括 create_num_accepted_drafts_filter → create_num_accept_tokens_filter（语义修正）以及 Kernel 参数 num_accepted_drafts → num_correct_drafts。

关键符号：update_spec_metrics, create_num_accept_tokens_filter, update_spec_correct_drafts_histogram, on_verify_complete_cpu, get_target_cache_loc, get_src_tgt_cache_loc, filter_finished_cache_loc_kernel, compute_dflash_correct_drafts_and_bonus, compute_dflash_sampling_correct_drafts_and_bonus

## 关键源码片段

### python/sglang/srt/observability/scheduler_metrics_mixin.py

核心度量统计，展示了 num_accepted_drafts → num_correct_drafts 和 num_accepted_tokens → num_accept_tokens 的重命名，以及 update_spec_metrics 参数重命名。

```
# Cumulative spec-decoding counters (reset every decode_log_interval).
# Each update adds (num_correct_drafts + bs, bs).
# `*_accept_tokens` = drafts + bonus; `*_correct_drafts` = drafts-only.
self.spec_num_accept_tokens = 0 # per-log-interval (renamed from spec_num_accepted_tokens)
self.spec_num_forward_ct = 0
self.spec_total_num_accept_tokens = 0 # lifetime (renamed from spec_total_num_accepted_
tokens)
self.spec_total_num_forward_ct = 0

def update_spec_metrics(self: Scheduler, bs: int, num_correct_drafts: int):
    # num_correct_drafts is the count of accepted draft tokens (excluding bonus)
    self.spec_num_accept_tokens += num_correct_drafts + bs
    self.spec_num_forward_ct += bs
    # Bonus tokens updated elsewhere
    self.num_generated_tokens += num_correct_drafts
```

### python/sglang/srt/speculative/spec_utils.py

投机解码工具函数的重命名，包括 create_num_accepted_drafts_filter → create_num_accept_tokens_filter（语义修正）以及 Kernel 参数 num_accepted_drafts → num_correct_drafts。

```
@torch.compile(dynamic=True, disable=_is_npu)
def create_num_accept_tokens_filter(
    num_correct_drafts: torch.Tensor,
    unfinished_index_device: torch.Tensor,
    seq_lens: torch.Tensor,
):
    # The filter value is num_correct_drafts + 1 (includes bonus token),
```

```
    # hence the rename from old `num_accepted_drafts_filter` to `num_accept_tokens_filter`.
    num_accept_tokens_filter = torch.zeros_like(num_correct_drafts)
    num_accept_tokens_filter[unfinished_index_device] = (
        num_correct_drafts[unfinished_index_device] + 1
    )
    seq_lens.add_(num_correct_drafts + 1)
    return num_accept_tokens_filter
```

## 评论区精华

该 PR 无 review 评论，所有决策在 PR body 中说明。值得关注的是对 num_accepted_drafts_filter 的语义修正：原名称中的 drafts 暗示仅包含 draft tokens，但实际值为 num_correct_drafts + 1（含 bonus token），因此重命名为 num_accept_tokens_filter 以准确反映其含义。这一修正在 body 中特别说明。

- num_accepted_drafts_filter 更名为 num_accept_tokens_filter 的语义修正 (design): 接受修正，已在 PR 中实现。

## 风险与影响

- 风险：主要风险为全局替换遗漏导致运行时变量未定义错误，或误替换外部接口。PR 通过保留外部接口并依赖 CI 测试（标签含 run-ci）来降低风险。影响面大（45 个文件），但重命名操作本身机械，风险较低。特别留意 meta_info JSON key 和 Prometheus 指标未改动。
- 影响：用户：无感知，外部 API 和指标未变化。 系统：内部一致性提升，后续开发需遵守新命名规则。 团队：开发者需适应新命名，但 PR 本身不引入功能变化。 影响范围：投机解码模块全部内部标识符。
- 风险标记：影响 45 个文件的大规模重命名，无外部 API 变更保证，潜在自动化替换遗漏

## 关联脉络

- PR #24081 ( 原始外部重命名 PR): 本 PR 是从 #24081 拆分出的内部重命名部分，外部重命名将在后续 PR 中完成。
- PR #25013 refactor: route idle hidden_size via EagleDraft{,Extend}Input classmethods: 同为投机解码模块的重构，改变了部分内部接口，本 PR 的重命名需与之协同以避免冲突。
- PR #25012 [Spec] Drop Rule 5 from speculative naming rule: 涉及命名规则的文档变更，与本 PR 的命名规范实施相关。