

PR #39284 完整报告

sgl-project/sglang

[Benchmark] Add an opt-out for the token-capacity check

合并时间: 2026-09-15 08:17

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39284>

执行摘要

- 一句话: one_batch 基准新增 token 容量检查跳过选项
- 推荐动作: 值得快速浏览, 不建议精读。该 PR 是一个典型的“工具可用性修复 + 显式 opt-out”设计: 字段默认值保持兼容、CLI help 明确边界、判断逻辑集中在一处且与既有 router 分支复用同一 inf 惯例。关注 benchmark 工具链的同事可以把它当作新增 CLI 开关的样板; 若后续要补测试, 建议在 run_benchmark_internal 的阈值计算处加一个轻量单测。

功能与动机

PR body 说明: one-batch benchmark 以 `batch_size * (input_len + output_len)` 估算 KV usage, 对带缓存优化的 workload 会高估物理 KV 用量并跳过有效 benchmark 用例。因此需要显式 opt-out 机制, 让用户在确认容量充足的情况下绕过该估算, 同时保留 max-running-requests 检查作为安全兜底。

实现拆解

变更集中在 `python/sglang/benchmark/one_batch_server.py`, 共 3 处协同修改:

1. 参数模型扩展: 在 BenchArgs dataclass 中新增 `skip_token_capacity_check: bool = False` 字段, 保持默认关闭, 确保现有调用方行为不变。
2. CLI 入口: 在 `add_cli_args` 中注册 `--skip-token-capacity-check (action="store_true")`, help 文本明确说明“跳过原始 token 容量检查, 但保留 max-running-requests 检查”, 避免用户误以为是全局跳过。
3. 运行时判断: 在 `run_benchmark_internal` 汇总各 worker 的 `token_capacity` 得到 `skip_token_capacity_threshold` 之后, 若该开关打开则将其置为 `float("inf")`, 使后续容量估算永远无法触发跳过; router 分支 (无 `internal_states`) 原有强制 inf 逻辑和 `--enable-multi-batch` 的 `running-cap` 逻辑均不受影响, 改动面收敛在容量检查一处。

配套方面: PR body 声明作者在本地用真实 CLI parser 和 mock 服务跑了 10 个 CPU smoke 场景 (默认拒绝、显式跳过、max-running 拒绝、warmup/profile/metrics 不变), 但未提交入库单元测试, 也未跑 GPU 测试; `pre-commit` 与 `git diff --check` 通过。

关键文件:

- `python/sglang/benchmark/one_batch_server.py` (模块 基准工具; 类别 source; 类型 core-logic; 符号 BenchArgs, add_cli_args, run_benchmark_internal): 唯一变更文件,

一次性完成参数模型、CLI 解析与运行时容量阈值三处协同修改，是本次功能的核心载体。

关键符号：add_cli_args, run_benchmark_internal

评论区精华

该 PR 没有产生任何 review 评论或讨论线程；唯一审核人 xiezhq-hermann 直接 APPROVED，说明改动简单且无争议。CI 状态展示三个检查（PR Test Base、PR Test Extra、AMD ROCm 10）均为 🟡，正文中也未对失败原因做解释，合并后未回滚，推测为基础设施或无关失败。

- 暂无高价值评论线程

风险与影响

- 风险：风险面很小，但有三点值得注意：
 1. 无入库测试：这次改动覆盖了 CLI 解析与基准用例筛选控制流，但没有提交自动化测试。若后续重构 run_benchmark_internal 的阈值比较逻辑，该分支可能被无意破坏且无人察觉。
 2. CI 失败信号：三个 CI 检查均显示 🟡，虽然与本文件纯增量的改动关系不大，但合并前未附加说明，属于流程上的一个隐患。
 3. 误用导致 OOM：用户若对物理 KV 容量确实不足的配置强制跳过检查，基准会直接跑出 OOM 或请求失败，而不是被提前跳过；这是显式 opt-out 的固有代价，help 文案已尽量传达边界。另外 float("inf") 在此处仅用于“永不触发跳过”，不参与后续算术，router 分支已有同样用法，无溢出风险。- 影响：影响范围严格限定在 python/sglang/benchmark/one_batch_server.py 单机基准工具内：
- 用户：使用 one-batch 基准的用户新增一个逃生通道，可以为 HiCache 等缓存优化场景跑此前被“伪容量不足”拦截的用例；不传该参数的用户行为完全不变。
- 系统：不涉及调度器、KV 缓存或模型推理路径，零运行时影响。
- 团队：为 benchmark 工具链补充了“容量估算可覆盖”的显式机制，后续同类估算优化（如更精确的物理 KV 占用量上报）可以在该开关基础上演进。
- 风险标记：缺少入库单元测试，CI 状态未通过，显式跳过容量检查有 OOM 风险

关联脉络

- PR #38097 [HiCache] Remove duplicate benchmark result fields: 同属 benchmark 工具链的可用性改进，均为清理基准脚本的冗余或误判；本次改动聚焦 one_batch_server.py，与 bench_serving.py 无直接文件冲突，但同属基准工具持续演进脉络。