

PR #38097 完整报告

sgl-project/sglang

[HiCache] Remove duplicate benchmark result fields

合并时间: 2026-09-15 08:56

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/38097>

执行摘要

- 一句话: 清理 HiCache 基准脚本重复结果字段
- 推荐动作: 不值得精读, 但可作为行为保持型清理的参考示例: 此类删除重复键的改动无需新增测试, 依靠 AST 检查和静态工具即可验证。对于关注 HiCache benchmark 脚本维护的开发者, 可以快速浏览确认字典键的完整性; 对一般读者, 没有需要学习的设计决策。

功能与动机

PR body 明确指出: HiCache serving benchmark 的结果字典重复了 8 个字面量键, Python 静默保留后面的值, 每个重复项都与早期条目使用相同表达式; 重复不改变序列化结果, 却增加了维护风险 (repetition therefore adds maintenance risk without changing the serialized result)。

实现拆解

实现拆解如下:

1. 定位重复键: 在 `benchmark/hicache/bench_serving.py` 的 `limited_request_func` 函数中, 成功分支构造 `result` 字典时, 末尾重复列出了 `input_throughput`、`output_throughput`、`fixed_output_len`、`random_input_len`、`random_output_len`、`random_range_ratio`、`duration`、`completed` 这 8 个键, 与字典前部的同名条目完全一致。
2. 移除重复条目: 删除字典末尾这 8 行, 使字典只保留每一键的首次定义。Python 字典字面量中重复键的求值结果与删除后的结果一致, 因此 `json.dumps(result)` 的输出不变。
3. 质量验证: PR 作者本地执行了 `py_compile`、AST `duplicate-literal-key` 检查、`ruff check`、`ruff format --check`、`isort --check-only`, 全部通过。未新增单元测试, 因为输出保持不变, 属于行为保持的清理。
4. 无配套改动: 不涉及配置、schema、文档或部署脚本变更, 影响范围仅限 benchmark 脚本本身。

关键文件:

- `benchmark/hicache/bench_serving.py` (模块 基准测试; 类别 `source`; 类型 `refactor`): 这是本 PR 唯一修改的文件, 在结果字典构造处删除 8 个重复键, 是全部变更的载体。

关键符号: `limited_request_func`

关键源码片段

benchmark/hicache/bench_serving.py

这是本 PR 唯一修改的文件，在结果字典构造处删除 8 个重复键，是全部变更的载体。

```
# 构造一次 benchmark 的结果字典（head 版本）。
# 原版本在字典末尾重复了以下 8 个键：
# "input_throughput" / "output_throughput" / "fixed_output_len" /
# "random_input_len" / "random_output_len" / "random_range_ratio" /
# "duration" / "completed"。
# 这些键与前面的条目表达式完全相同，Python 对重复键只保留最后一个值，
# 因此序列化输出不变；但重复项误导维护者，故删除。
result = {
    "max_concurrency": max_concurrency,
    "fixed_output_len": args.fixed_output_len,
    "random_input_len": args.random_input_len,
    "random_output_len": args.random_output_len,
    "random_range_ratio": args.random_range_ratio,
    # Results
    "duration": benchmark_duration,
    "completed": metrics.completed,
    "total_input_tokens": metrics.total_input,
    "total_output_tokens": metrics.total_output,
    "total_output_tokens_retokenized": metrics.total_output_retokenized,
    "request_throughput": metrics.request_throughput,
    "input_throughput": metrics.input_throughput,
    "output_throughput": metrics.output_throughput,
    "mean_e2e_latency_ms": metrics.mean_e2e_latency_ms,
    "median_e2e_latency_ms": metrics.median_e2e_latency_ms,
    "std_e2e_latency_ms": metrics.std_e2e_latency_ms,
    "p99_e2e_latency_ms": metrics.p99_e2e_latency_ms,
    "mean_ttft_ms": metrics.mean_ttft_ms,
    "median_ttft_ms": metrics.median_ttft_ms,
    "std_ttft_ms": metrics.std_ttft_ms,
    "p99_ttft_ms": metrics.p99_ttft_ms,
    "mean_tpot_ms": metrics.mean_tpot_ms,
    "median_tpot_ms": metrics.median_tpot_ms,
    "std_tpot_ms": metrics.std_tpot_ms,
    "p99_tpot_ms": metrics.p99_tpot_ms,
    "mean_itl_ms": metrics.mean_itl_ms,
    "median_itl_ms": metrics.median_itl_ms,
    "std_itl_ms": metrics.std_itl_ms,
    "p99_itl_ms": metrics.p99_itl_ms,
    "concurrency": metrics.concurrency,
}
```

评论区精华

该 PR 的 review 讨论非常少：作者在 Issue 评论中请求 @ispobock 和 @xiezhq-hermann 审阅并触发 CI ([/tag-and-rerun-ci](#))，合并者 xiezhq-hermann 直接 APPROVED，无任何 review 评论或实质性技术讨论。所有验证均出自 PR body 描述的本地检查，未见到对删除安全性的质疑。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低，理由如下：
- 行为保持：被删除的 8 个键在字典前部均有表达式完全相同的条目，Python 字面量重复键只保留最后一个值，删除后 `json.dumps` 序列化结果不变。
- 无外部依赖：改动仅在 `benchmark/hicache/bench_serving.py` 内部，不涉及核心 serving 路径或序列化协议。
- 潜在小隐患：如果未来有人修改字典前部某键的表达式而忘记同步尾部重复项（或反之），会造成输出不一致；删除重复项恰好消除了这一隐患。唯一需要注意的是一旦未来字典结构变化，需要保证保留的条目仍是预期值，但当前 PR 已用 AST 检查确认无重复键。
- 影响：影响范围非常有限：仅影响 HiCache serving benchmark 脚本 `benchmark/hicache/bench_serving.py` 的可维护性，对运行中的 serving 系统、用户请求、API 输出均无影响。团队层面，消除了一个潜在误导点，降低了后续维护者误改字典导致输出错乱的概率。CI 影响也极小，因为不涉及核心代码路径。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR