PR #26165 完整报告

sgl-project/sglang

[SRT] Store Req input ids as arrays

合并时间：2026-05-24 15:09

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26165

执行摘要

- 一句话：Req input ids 改用 array 存储
- 推荐动作：值得合并，改动小且明确。可考虑后续优化避免重复 array 转换。

功能与动机

Req.output_ids 已使用 array('q')，而原始 prompt ids 仍为 Python list。使用相同的紧凑整数数组表示可避免请求构建后每个 int 的 Python 对象开销，并使 prompt/output id 存储保持一致。

实现拆解

1. 在 ScheduleBatch.__init__ 中，将 self.origin_input_ids_unpadded 和 self.origin_input_ids 的赋值改为 array('q', ...) 形式。
2. origin_input_ids_unpadded 的 fallback 逻辑从条件表达式简化为 origin_input_ids_unpadded or origin_input_ids，并在 array 构造函数中处理。
3. 未改动其他逻辑，保持了兼容性。

关键文件：

- python/sglang/srt/managers/schedule_batch.py（模块 调度器；类别 source；类型 core-logic）：核心变更文件，修改了 Req 类的初始化逻辑，将输入 ids 从 list 改为 array 存储。

关键符号：未识别

关键源码片段

python/sglang/srt/managers/schedule_batch.py

核心变更文件，修改了 Req 类的初始化逻辑，将输入 ids 从 list 改为 array 存储。

```
# python/sglang/srt/managers/schedule_batch.py
# 在 Req.__init__ 中，将原本的 list 赋值改为 array('q') 以节省内存
self.origin_input_ids_unpadded = array(
    "q", origin_input_ids_unpadded or origin_input_ids
) # Before image padding
self.origin_input_ids = array("q", origin_input_ids)
```

```
# 已有字段如 output_ids、fill_ids 也使用 array('q')
self.output_ids = array("q")
self.fill_ids = array("q")
```

评论区精华

AI 评论指出当未提供 unpadded ids 时，origin_input_ids 会被转换为两次 array（一次用于 unpadded，一次用于 padded），建议通过复用转换结果来优化内存和性能。但作者未采纳该建议。

- 重复 array 转换可能造成额外开销 (performance): 作者未采纳该建议，保留当前实现。

风险与影响

- 风险：低风险。仅修改了 Req 属性存储类型，所有常用操作（len、索引、切片、迭代、append、extend）在 array 上同样支持。主要风险在于外部代码可能依赖 list 特定行为（如直接 JSON/msgpack 序列化、与普通 list 的相等性比较），但此类依赖在内部代码中可能较少。
- 影响：对长 prompt 请求（尤其是多模态请求，含大量图像填充 token）有正面内存和性能影响。存储一致性提升。无外部 API 变更。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR