

PR #28340 完整报告

sgl-project/sglang

feat(cookbook): MTP --max-running-requests callout + skill sync

合并时间: 2026-06-16 06:53

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28340>

执行摘要

此 PR 为文档站点的 Deploy 和 Playground 面板添加了 MTP callout 提示，当用户启用 speculative decoding 但未设置 `--max-running-requests` 时显示。同时同步更新了多个 cookbook 技能文档，强化 benchmark 可复现性规则并整合 `flagSelects` 通用轴。整体是配合引擎行为的 UI 辅助与文档规范更新。

功能与动机

当 speculative decoding 开启时，SGLang 的 `speculative_hook.py` 会将未设置的 `--max-running-requests` 强制重置为 48。用户若无此预期可能在高并发下性能受损。通过 UI 提示可以提前告知。技能文档方面，之前迁移轮次中 deferred 的规则统一（benchmark 版本锚定、`flagSelects` 轴）在此合并。

实现拆解

1. Engine UI 提示:

- `_deployment.jsx`: 在样式对象中新增 `mtpWarn` 琥珀色警告框样式；在 derived values 区域计算 `mtpHint` 变量，检测当前 cell 的 flags 是否包含 `--speculative-algorithm` 且不包含 `--max-running-requests`；在命令下方渲染提示。
- `_playground.jsx`: 同样新增样式和 `pgMtpHint` 变量，但检测的是用户调整后的有效 flags (`pgFlagsLatest`)，确保 Playground 的 spec 开关生效后也能触发。

2. cookbook-migrate-model 技能更新:

- `SKILL.md`: 重写 hard rule 2，要求 `sglang_version` 必须是可复现的锚点（release/commit/PR），否则丢弃整个结果；添加 `flagSelects` 作为标准轴、NEXTN→EAGLE 别名。
- `dimension-mapping.md`: 新增 `flagSelects` 轴说明，强调模型特定控制应通过配置数据实现而非引擎代码。

3. cookbook-add-model 技能更新:

- `authoring-reference.md`: 添加 `flagSelects` 作为内置轴的说明，要求 `sglang_version` 可复现，并注明 MTP callout 会自动显示。

`docs_new/src/snippets/_deployment.jsx`

核心引擎变更，添加 MTP callout 提示，基于 flags 检测显示。

```
// mtpWarn: 琥珀色提示框样式
```

```

mtpWarn: {
  margin: "8px 0 0", padding: "8px 12px", borderRadius: "8px",
  fontSize: "12px", lineHeight: "1.45",
  background: isDark ? "#78350f" : "#fef3c7",
  color: isDark ? "#fde68a" : "#92400e",
  border: `1px solid ${isDark ? "#92400e" : "#fcd34d"}`,
},
// 在 derived values 中: 基于当前 cell 的 flags 判断是否触发提示
const mtpHint =
  !!cell &&
  (cell.flags || []).some((f) => f.split(/\s=/)[0] === "--speculative-algorithm") &&
  !(cell.flags || []).some((f) => f.split(/\s=/)[0] === "--max-running-requests");
// 在命令显示区域下方渲染提示
{mtpHint && (
  <div style={s.mtpWarn}>
    ⚠️ Speculative decoding (MTP) is on — SGLang resets --max-running-requests
    code> to 48 when it isn't set.
    Add --max-running-requests &lt;N> sized for your target concurrency.
  </div>
)}

```

docs_new/src/snippets/_playground.jsx

核心引擎变更, 添加 Playground 面板中的 MTP callout。

```

// mtpWarn 样式同 _deployment.jsx
mtpWarn: {
  margin: "8px 0 0", padding: "8px 12px", borderRadius: "8px",
  fontSize: "12px", lineHeight: "1.45",
  background: isDark ? "#78350f" : "#fef3c7",
  color: isDark ? "#fde68a" : "#92400e",
  border: `1px solid ${isDark ? "#92400e" : "#fcd34d"}`,
},
// 基于 post-override 的有效 flags 判断
const pgMtpHint =
  pgFlagsLatest.some((f) => f.split(/\s=/)[0] === "--speculative-algorithm") &&
  !pgFlagsLatest.some((f) => f.split(/\s=/)[0] === "--max-running-requests");
// 渲染提示
{pgMtpHint && (
  <div style={s.mtpWarn}>
    ⚠️ Speculative decoding (MTP) is on — SGLang resets --max-running-requests
    code> to 48 when it isn't set.
    Add --max-running-requests &lt;N> sized for your target concurrency.
  </div>
)}

```

评论区精华

无 review 评论。审核人 wisdomlmy0611 已批准。

风险与影响

- 风险：极低。仅 UI 增加提示文案，不影响后端逻辑。需注意触发条件与引擎实际行为一致（flag 检测而非 strategy 检测）。
- 影响：对访问文档站点的用户提供善意的性能提示；技能文档更新使后续 cookbook 迁移更规范。

关联脉络

本 PR 是对之前 `flagSelects` 特性 (#28128) 的配套文档同步。后续所有涉及 speculative decoding 的 cookbook 都会自动获得 callout 效果，无需手动添加说明。