

# PR #36342 完整报告

sgl-project/sglang

docs(cookbook): use auto parser resolution for Granite 4.2

合并时间: 2026-08-26 01:38

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36342>

## 执行摘要

本 PR 是对 #36286 的后续更新，将 Granite 4.2 cookbook 中的解析器标记从显式值（`--reasoning-parser nemotron_3`、`--tool-call-parser qwen3_coder`）改为 `auto`，利用 chat-template 自动检测解析器类型。改动仅涉及文档与命令面板配置，无运行时影响，但简化了用户配置步骤。

## 功能与动机

36286 首次引入 Granite 4.2 cookbook，当时使用了显式解析器标记。然而，sglang 自 v0.5.12 开始支持 chat-template 自动检测解析器（低工作量切换规则自 v0.5.16 起），因此本 PR 将命令面板和文档改为推荐 `auto`，用户无需手动指定解析器类型，降低配置出错概率。同时保留“省略解析器标记会禁用解析”的警告，避免用户误用。

## 实现拆解

- 命令面板配置更新 (`docs/src/snippets/configs/ibm-granite/granite-4.2.jsx`) :
  - 将 `parsers.items` 中的两个 flag 改为 `--reasoning-parser auto` 和 `--tool-call-parser auto`。
  - 将所有 `cells` 内的 `flags` 数组同步修改，覆盖 `h200/b200` 和 `3b/8b/30b` 变体。
  - 这样用户在 Playground 中生成的命令会使用 `auto`，且实际生效的解析器由模型 chat-template 自动决定。
- cookbook 文档更新 (`docs/cookbook/autoregressive/IBM/Granite-4.2.mdx`) :
  - 更新“Reasoning parser”和“Tool-call parser”两节，将建议命令改为 `auto`，并说明自动解析为 `nemotron_3/qwen3_coder`。
  - 保留“不带 parser flag 会禁用解析”的提示，确保用户正确使用。
- 验证：在 `d10a656ad8` 提交上验证，将两个标记均设为 `auto` 时，3B/8B/30B 模型的所有协议探测均通过；通过 pre-commit、mint 检查和浏览器检查。

`docs/src/snippets/configs/ibm-granite/granite-4.2.jsx`

命令面板配置，修改解析器标记为 `auto`，影响 playground 显示和复制命令。

```
// docs/src/snippets/configs/ibm-granite/granite-4.2.jsx
// 命令面板的解析器配置：使用 auto 让 chat-template 自动解析为相应解析器。
playgroundFeatures: {
  \V \ ... 其他配置
  parsers: {
    items: [
      { id: "reasoning", label: "Reasoning Parser", flag: "--reasoning-parser auto" },
      { id: "toolCall", label: "Tool Call Parser", flag: "--tool-call-parser auto" },
    ],
  },
  \V \ ... 其他配置
},

\V \ 各 cell 的启动命令 flags 也统一改为 auto
cells: [
  {
    match: { hw: "h200", variant: "3b", quant: "bf16", strategy: "balanced", nodes: "single" },
    verified: true,
    flags: [
      "--model-path {{MODEL_NAME}}",
      "--tp 1",
      "--mem-fraction-static 0.8",
      "--reasoning-parser auto",
      "--tool-call-parser auto",
      "--host {{HOST_IP}}",
      "--port {{PORT}}",
    ],
  },
  \V \ ... 其他 cell 相似修改
]
```

## 评论区精华

无实质评论或审核讨论，审阅者直接批准。

## 风险与影响

- 风险：纯文档变更，风险极低。唯一潜在问题是 auto 解析器依赖 chat-template 检测，若未来 checkpoint 变更导致检测失败，用户可能需要回退到显式标记，但文档中已有提示。
- 影响：简化了 Granite 4.2 用户的配置，减少了手动指定解析器的负担。不影响系统运行。

## 关联脉络

- #36286：本 PR 的直接前驱，添加了 Granite 4.2 cookbook。
  - 此变更反映了 sglang 文档正向“自动解析”方向演进，后续其他模型的 cookbook 更新可参考此模式。