PR #36808 完整报告

sgl-project/sglang

[Cookbook] Hy4-Preview follow-ups: runtime-accurate recipes + released-model info

合并时间：2026-08-28 15:19

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36808

Hy4-Preview Cookbook 校准：运行时一致的部署配方

执行摘要

本 PR 是 #36804 的后续，把文档示例中的配置改为与实际运行时行为一致：更新模型信息（770B 总参、49B 激活、Apache-2.0、官方 GitHub 链接），删除 DeepEP 场景下会被重写的 EP 旋钮，将 MXFP8 后端统一到已验证的 deep_gemm 路径，并把 8 个单节点配方标记为 Verified。变更集中在 docs/src/snippets/configs/tencent/hy4-preview.jsx 和 docs/cookbook/autoregressive/Tencent/Hy4-Preview.mdx，是文档准确性与可复现性的一次重要修正。

功能与动机

PR body 明确说明这是对已合并的 #36804 的 follow-up，包含两类变更：

- 配方修正：必须与 Hy4 支持分支的实际运行时解析一致；
- 发布信息更新：模型已公开释出，页面需补全模型卡与官方仓库信息。

具体触发了三处运行时不匹配：

- DeepEP 后端会把 ep_size 重写为 tp_size，之前展示的 EP 拓扑实际不会发生；
- 顶层 reasoning_effort 字段只接受 OpenAI 档位，旧示例在请求校验时会被拒绝；
- CUDA MXFP8 后端白名单排除 triton，HYV4 家族实际默认走 deep_gemm 路径。

实现拆解

1. 模型信息校准：在 .jsx 头部注释与 .mdx 描述中，将参数量从约 760B/40B 修正为 770B 总参 / 49B 激活，补充 MTP draft 层规模（10B 参数、约 0.7B 激活）、Apache-2.0 许可证、官方 GitHub 链接，以及推荐采样参数 temperature=0.9、top_p=1.0（仅信息性，示例不硬编码）。
2. 后端与并行配置对齐运行时：在 .jsx 的 MoE Parallelism 卡片中删除 EP 旋钮，DeepEP 标签改为 DeepEP (EP = TP)；B200 MXFP8 经历第 1 个提交的 flashinfer_trtllm 尝试后，第 3 个提交回退为 deep_gemm——因为 _deepseek_family_overrides 对 HYV4 MXFP8 家族默认同时将 MoE runner 和 FP8 GEMM 设为 deep_gemm，固定 flashinfer_trtllm 反而强制了未经验证的后端。
3. 示例与文案修复：在 .mdx 中，Instant-mode 示例把 no_think 通过 extra_body={"chat_template_kwargs": ...} 传递，避免顶层字段拒绝；安装手风琴介绍句

改为与 Docker-only 面板一致的表述。

4. 验证状态更新：8 个单节点配方（B300 MXFP8/BF16、B200 MXFP8、GB300 MXFP8，低延迟与高吞吐各一）已在真实硬件端到端跑通，置为 verified: true；3 个 2 节点 BF16 配方保持 in-progress，等待多机验证。
5. 配套验证：无新增单元测试；用 node docs/scripts/check_cookbook_configs.mjs、mint validate、mint broken-links 校验通过；常规 PR Test CI 绿，但 Extra CI 槽位标红，原因未在材料中说明。

docs/src/snippets/configs/tencent/hy4-preview.jsx

驱动 Deploy/Playground 面板的核心配置源文件。本次变更删除 EP 旋钮、将 DeepEP 标签改为 EP = TP、更新模型规模注释，并翻转 8 个单节点配方的验证状态，是运行时准确性修复的主要载体。

```
// ----- Card 2: "MoE Parallelism" -----
// 256 个 routed + 1 个 shared experts，top-8 sigmoid 路由。
// 配方在纯 TP 下运行 MoE：MXFP8 路径使用 deep_gemm runner（已验证的 HYV4 路径），
// DeepEP 仅作为实验性覆盖项。不提供 EP 旋钮：运行时对 a2a 类后端（如 DeepEP）
// 会把 ep_size 重写为 tp_size，暴露自由 EP 度数会展示一套实际永远不会运行的拓扑。
// 不提供 MegaMoE 选项——其融合路径未适配 Hy4 的 sigmoid 打分 bounded-SwiGLU experts。
moe: {
  backend: {
    options: [
      { id: null, label: "Inherited" },
      { id: "deepep", label: "DeepEP (EP = TP)", flags: ["--moe-a2a-backend deepep"] },
    ],
  },
},
```

评论区精华

该 PR 没有 review 评论，唯一审核人 wisclmy0611 直接 APPROVED。提交历史代替讨论记录了两次设计迭代：

> 第 1 个提交曾把 B200 MXFP8 固定为 flashinfer_trtllm，第 3 个提交基于运行时事实回退到 deep_gemm："_deepseek_family_overrides defaults both the MoE runner and FP8 GEMM to deep_gemm for HYV4 MXFP8 (the validated path) whenever JIT DeepGEMM is available — pinning flashinfer_trtllm forced an unvalidated backend"。
>
> 第 2 个提交修复 no_think 传递路径，说明顶层 reasoning_effort 只接受 OpenAI 枚举档位，旧示例在模板渲染前即被请求校验拒绝。

这些决策本质是作者自答的三个运行时约束：后端白名单、EP 重写规则、请求校验层行为。

风险与影响

- 配置与运行时强耦合：.jsx 内容依赖 HYV4 家族的运行时解析；若 _deepseek_family_overrides 或 CUDA MXFP8 白名单后续变更，文档会静默过期。

- 依赖 JIT DeepGEMM 可用性：固定 deep_gemm 的前提是 JIT DeepGEMM 可编译，个别环境可能无法命中该路径。
- 示例依赖新接口：chat_template_kwargs 传递 no_think 需要较新的 sglang 版本，旧版本用户仍会失败。
- 未验证区域：3 个 2 节点 BF16 配方未做多机验证，可能误导追求大吞吐的用户。
- CI 信号不完整：Extra CI 槽位标红且未说明原因，PR 虽已合并，仍有检查项未闭环。

影响面：对部署 Hy4-Preview 的用户是最直接的收益——复制命令即可得到与运行时一致的行为；对文档团队则是方法论收益，确立了“运行时解析优先 + 真机验证驱动徽章”的撰写模式。

关联脉络

本 PR 与 #36804（初始 cookbook 引入）构成同一功能线的两阶段：先用估算或规划数据建页，再在模型发布与真机验证后校准。#36823 则是对同一页面标题的后续命名调整，说明 Hy4 文档系列仍处于快速演进期。预计下一步将是 2 节点 BF16 配方的多机验证，以及 benchmark 卡片从 pending stub 转为实测数据。