PR #36496 完整报告

sgl-project/sglang

Add Qwen3.8-Flash-Next cookbook

合并时间：2026-08-26 20:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36496

执行摘要

本 PR 为 Qwen3.8-Flash-Next（Qwen 的 Qwen4 架构预览，176B 总参 / 6B 激活，GDN + QSA 混合 MoE）新增 Day-0 cookbook 页面，基于配置驱动的 cookbook 模板生成 22 个单节点部署 cell，覆盖 H200 / B200 / B300 / GB300 与 MI350X / MI355X 上的 BF16 / FP8 / NVFP4 三种精度，并为其中 11 个 cell 附上实测 GSM8K / AIME26 / MMMU-Pro 精度。全部改动限于文档域，无引擎代码变更；同时把新模型接入站点导航、cookbook 首页卡片与站点首页轮播，模型支持代码仍由独立 PR 提供（页面中暂以 <PR-NUMBER> 占位）。

功能与动机

PR body 明确这是 "Day-0 cookbook page"：在模型发布当天就提供可验证的部署配方，让用户按硬件 × 量化 × 策略直接生成启动命令，而不是等代码合入后补文档。实现上完全复用配置驱动模板（per-model config + benchmarks，由共享的 _deployment.jsx / _playground.jsx 消费，"no engine edits"），文档侧零侵入；同时通过替换 popular-models.jsx 首位与 intro.mdx 卡片链接，把新模型推到首页曝光位。源码安装段落引用 <PR-NUMBER> 占位符，说明文档先于模型支持 PR 落地，属于典型的发布流程配套。

实现拆解

1. 新增模型配置 snippet：docs/src/snippets/configs/Qwen/qwen3.8-flash-next.jsx（+723 行）导出单一 config 字面量，声明 supportedHardware（6 平台）、quantizations、strategies、nodesOptions、overlayDims（PLE Offload 拨杆）、modelNames（三种精度指向不同 HF repo）以及 curl / benchmarkCommands / accuracyLabels。cell 采用反规范化写法，分布式参数由 _deployment.jsx 引擎注入。
2. 新增基准数据：qwen3.8-flash-next-benchmarks.jsx（+72 行）导出 benchmarks 数组，条目以 match 声明 cell key 并附 sglang_version 与 accuracy；11 个 NVIDIA cell 有实测（qwen4-main @ e17062a1d），FP8 low-latency、NVFP4 与 AMD cell 留空，页面据此显示 " 待测量 "。
3. 新增 MDX 页面：docs/cookbook/autoregressive/Qwen/Qwen3.8-Flash-Next.mdx（+222 行），包含安装引导（Python 源码安装 / NVIDIA 与 AMD Docker 镜像）、<Deployment> / <Playground> 组件接线，以及 4 轴 Playground（TP、MoE EP、parser、NEXTN / MTP 预设）。
4. 接线与入口替换：docs/docs.json 在 Qwen 分组插入页面；popular-models.jsx 首页轮播第一条替换为 Qwen3.8-Flash-Next；intro.mdx Qwen 卡片指向新页；Qwen3.8-27B.mdx 删除 1 行旧引用。

5. 测试与 CI 配套：无测试文件新增；CI Base run 通过、Extra run 失败；PR body 明确 <PR-NUMBER> 占位符待模型支持 PR 合入后替换。

维度	内容
平台	H200 / B200 / B300 / GB300 / MI350X / MI355X
精度	BF16 / FP8 / NVFP4（Blackwell-only）
策略	Low Latency（NEXTN 3/1/4、MRR 96）/ Balanced / High Throughput（--ep 4）

docs/src/snippets/configs/Qwen/qwen3.8-flash-next.jsx

核心配置数据源：定义模型名、平台、量化、策略、PLE Offload 拨杆、模型 repo 映射、curl / benchmark 模板与镜像标签，驱动 Deploy / Playground 面板生成 22 个 cell 的启动命令。

```
// 单例 `config` 字面量导出，不使用 spread / 函数调用 / IIFE，
// 因为 Mintlify 在 hydration 时会重新求值。
// cell 采用反规范化写法：不写 `--nnodes` / `--port` 等分布式参数，
// 由 `_deployment.jsx` 引擎统一注入。
export const config = {
  modelName: 'Qwen3.8-Flash-Next',

  // 6 个平台：NVIDIA H200 / B200 / B300 / GB300 + AMD MI350X / MI355X
  supportedHardware: ['h200', 'b200', 'b300', 'gb300', 'mi350x', 'mi355x'],

  variants: [{ id: 'default', label: 'Default' }],

  // NVFP4 是 SGLang 自研的 Blackwell-only 量化（RadixArk），
  // SM90（H200）与 CDNA4（AMD）没有该路径，因此不生成对应 cell
  quantizations: [
    { id: 'bf16', label: 'BF16' },
    { id: 'fp8', label: 'FP8' },
    { id: 'nvfp4', label: 'NVFP4' },
  ],

  // low-latency 在 high-throughput 基础上叠加 MTP 头（NEXTN 3/1/4），
  // AMD 两个平台各只有一条配方，归入 balanced
  strategies: [
    { id: 'low-latency', label: 'Low Latency' },
    { id: 'balanced', label: 'Balanced' },
    { id: 'high-throughput', label: 'High Throughput' },
  ],
  nodesOptions: [{ id: 'single', label: 'Single Node' }],

  // 正交拨杆：叠加在匹配到的 cell 上，但不参与 cell key 匹配
```

```
    overlayDims: [
      {
        id: 'pleOffload',
        title: 'PLE Offload',
        // 把 51B 的 N-gram embedding 表卸载到 CPU pinned memory，
        // 并在旁路 CUDA stream 上预取；该路径仅 CUDA 可用，AMD cell 隐藏此行
        showWhen: (sel) => !['mi350x', 'mi355x'].includes(sel.hw),
        default: 'auto',
        options: [
          { id: 'auto', label: 'Auto', hints: ['PLE Offload: BF16 在 CUDA 上自动开启，其余关闭'] },
          { id: 'on', label: 'On', flags: ['--ple-offload-embedding'] },
          { id: 'off', label: 'Off', flags: ['--no-ple-offload-embedding'] },
        ],
      },
    ],

    // 三种精度指向不同 repo，而不是同一 repo 的不同 revision
    modelNames: {
      'defaultlbf16': 'Qwen/Qwen3.8-Flash-Next',
      'defaultlfp8': 'Qwen/Qwen3.8-Flash-Next-FP8',
      'defaultlnvfp4': 'RadixArk/Qwen3.8-Flash-Next-NVFP4',
    },

    placeholders: {
      HOST_IP: { target: 'command', label: 'Bind host', default: '0.0.0.0' },
      PORT: { target: 'command', label: 'Bind port', default: '30000' },
      HF_TOKEN: { target: 'command', label: 'HF token (Docker)', default: '<your-hf-token>' },
      CURL_HOST: { target: 'curl', label: 'Server host', default: 'localhost' },
      CURL_PORT: { target: 'curl', label: 'Server port', default: '30000' },
    },

    accuracyLabels: [
      ['gsm8k_pct', 'GSM8K', '%'],
      ['aime26_pct', 'AIME26', '%'],
      ['mmmu_pro_pct', 'MMMU-Pro', '%'],
    ],
    // 其余字段：curl 模板、benchmarkCommands 与 launchImages 均在原文件中展开
  };
```

docs/src/snippets/configs/Qwen/qwen3.8-flash-next-benchmarks.jsx

基准数据契约：按 cell key 声明 22 个部署组合的 GSM8K / AIME26 / MMMU-Pro 精度，11 个有实测、其余留空表示待测量，供页面卡片渲染。

```
  // Deploy 面板的 benchmark 卡片数据源：每个条目用 `match` 声明
  // cell key（hw × variant × quant × strategy × nodes），
  // 附上实测 `accuracy`；没有 accuracy 的 cell 会显示为“待测量”。
  export const benchmarks = [
    {
      match: { hw: 'h200', variant: 'default', quant: 'bf16', strategy: 'low-latency', nodes: 'single' }
```

```
    ,
    sglang_version: 'qwen4-main @ e17062a1d',
    accuracy: { gsm8k_pct: 97.73, aime26_pct: 97.92 },
  },
  {
    match: { hw: 'h200', variant: 'default', quant: 'bf16', strategy: 'high-throughput', nodes:
    'single' },
    sglang_version: 'qwen4-main @ e17062a1d',
    accuracy: { gsm8k_pct: 97.57, aime26_pct: 99.17 },
  },
  // FP8 low-latency 与 high-throughput 在 H200 上暂无实测数据,
  // 只声明 cell key，页面将显示为"待测量"
  { match: { hw: 'h200', variant: 'default', quant: 'fp8', strategy: 'high-throughput', nodes:
  'single' } },
  { match: { hw: 'h200', variant: 'default', quant: 'fp8', strategy: 'low-latency', nodes: 'single' }
  },
  // B200 / B300 / GB300 的 BF16 与 FP8 high-throughput 均有实测,
  // NVFP4 与 AMD（MI350X / MI355X）cell 均无 accuracy 字段
  { match: { hw: 'mi350x', variant: 'default', quant: 'bf16', strategy: 'balanced', nodes: 'single' }
  },
  { match: { hw: 'mi350x', variant: 'default', quant: 'fp8', strategy: 'balanced', nodes: 'single' } }
  ,
  { match: { hw: 'mi355x', variant: 'default', quant: 'bf16', strategy: 'balanced', nodes: 'single' }
  },
  { match: { hw: 'mi355x', variant: 'default', quant: 'fp8', strategy: 'balanced', nodes: 'single' } }
  ,
];
```

docs/src/snippets/configs/popular-models.jsx

首页 / cookbook 首页轮播数据源：把第一位从 Qwen3.8-27B 替换为 Qwen3.8-Flash-Next，放大新模型曝光。

```
// 首页与 cookbook 首页的轮播数据源：本 PR 把第一条从 Qwen3.8-27B
// 替换为 Qwen3.8-Flash-Next，使新模型在站点首页获得曝光
export const popularModels = [
  {
    name: 'Qwen3.8-Flash-Next',
    vendor: 'Qwen',
    href: '/cookbook/autoregressive/Qwen/Qwen3.8-Flash-Next',
    logo: '/cards/logos/qwen.png',
    badge: 'New',
    tags: ['6 platforms', 'GDN + QSA hybrid', 'BF16 / FP8 / NVFP4'],
    hero: {
      eyebrow: 'Featured model · New',
      headline: 'Meet Qwen3.8-Flash-Next on SGLang',
      blurb:
        'Qwen 对 Qwen4 架构的早期预览 —— 176B 总参数、6B 激活，四层中三层为 Gated
        DeltaNet、第四层为运行 Qwen Sparse Attention 的全局注意力，配合超稀疏 MoE 与随
        checkpoint 的 MTP 头。cookbook 覆盖 H200 / B200 / B300 / GB300 的单节点 TP4 与
```

```
        MI350X / MI355X 的 TP8 服务。',
      tags: ['176B / 6B active', '262K context', 'Single-node'],
      cta: 'Open the Qwen3.8-Flash-Next cookbook',
      caption: 'Qwen3.8-Flash-Next deployment guide',
    },
  },
  // MiniMax-H3、Kimi-K3 等后续条目保持原样
];
```

评论区精华

本 PR 没有任何 review 评论线程，唯一评审是 JustinTong0323 的 APPROVED 空评审。可讨论内容集中在 commit 迭代中：

- FP8 配方被拆成 high-throughput 与 low-latency 两个 operating point，"nothing parks under balanced except the two AMD platforms"。
- NVFP4 从 TP4 降为 TP1 单卡：server_args 断言 ep_size * moe_dp_size <= tp_size，TP1 下 --ep 4 会直接让启动失败。
- 最后一轮把 NVFP4 配方与实测命令对齐：去除 --mem-fraction-static、--chunked-prefill-size、--max-running-requests，避免文档与真实跑测配置漂移。 这些是文档与运行时约束互动的典型样例。

风险与影响

- 占位符风险：安装引导中的 <PR-NUMBER> 若未替换，用户按文档源码安装会失败；合并前必须回填模型支持 PR 编号。
- 数据覆盖风险：22 个 cell 中仅 11 个有精度数据；FP8 low-latency、全部 NVFP4 与 AMD cell 显示 " 待测量 "，需避免用户误读为 " 不支持 "。
- 外部依赖：NVFP4 cell 指向 RadixArk/Qwen3.8-Flash-Next-NVFP4 第三方 repo，且为 TP1 单卡配方，repo 可用性未在 PR 中独立验证。
- CI 状态：Extra run 标记为失败（x），合入前需确认失败与文档改动无关。
- 影响面：仅文档域，但辐射站点首页、cookbook 首页与 Qwen 目录三处入口；对普通用户是收益（当日可部署），对团队是模板复用与发布流程的样板。

关联脉络

本 PR 与近期文档 / 基准类 PR 同属一套配置驱动 cookbook 演化：PR 36463 完善了 diffusion 基准缓存与 preset，PR 36412 处理 MiniMax H3 的 checkpoint 变体说明，PR 36375 修复 Cosmos3 文档配置与显存策略的联动回归——本 PR 是这套体系首次落到 Qwen 自回归模型 Day-0 发布场景。后续需跟进：模型支持 PR 合入并回填 <PR-NUMBER>、补充 FP8 low-latency / NVFP4 / AMD 的精度数据，以及确认 popular-models.jsx 中被替换的 Qwen3.8-27B 是否保留独立入口。