PR #35663 完整报告

sgl-project/sglang

[docs] Add DFlash2 speculative cells to the Qwen3.8-27B cookbook

合并时间：2026-08-21 04:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35663

执行摘要

本 PR 为 Qwen3.8-27B 的 cookbook 新增了 DFLASH2 推测解码选项，在 Deploy 面板和 Playground 卡片中，并针对不同硬件平台提供验证状态提示。文档基于已合并的 #35496，使得 NVFP4 模型可以使用 DFLASH2，同时提供了详细的性能数据和配置建议。变更主要影响文档和用户操作指引，不涉及代码逻辑变更。

功能与动机

DPR 的动机来自 Issue #35496（已合并），该 PR 为 DFlash2 选择器支持量化目标 lm_head，从而 NVFP4 checkpoint 可以运行 DFLASH2。因此，cookbook 需要更新，以指导用户启用 DFLASH2，并明确各平台的验证情况，避免用户使用未经验证的配置。

实现拆解

1. 在 Deploy 面板中新增 DFLASH2 选项：修改 docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx，新增 dflash 选项，包含命令参数 --speculative-algorithm DFLASH --speculative-draft-model-path incoai/Qwen3.8-27B-DFlash2 --speculative-num-draft-tokens 8。同时，针对 RTX 5090 平台，会根据 SSM dtype 调整 --mem-fraction-static 和 --mamba-full-memory-ratio，并通过 stripPrefixes 处理与已有配置的冲突。
2. 在 Playground 卡片中新增 DFLASH2 选项：在 speculative.options 中新增 dflash，与 Deploy 面板的 flags 保持一致，确保两条路径生成相同的命令。
3. 更新配置文件文档：在 docs/cookbook/autoregressive/Qwen/Qwen3.8-27B.mdx 中，添加 DFLASH2 的配置提示，解释其 D 项（8 为 draft 的 block size），并更新验证说明。
4. 未验证平台提示：对于 H200、DGX Spark、GB300，在 DFLASH2 选中时，Deploy 面板将显示 "final verification in progress" 提示行，以提醒用户该配置尚需验证。

docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx

核心变更文件：新增 DFLASH2 选项到 Deploy 面板和 Playground 卡片，包含平台条件和参数调整逻辑。

```
// 在 Deploy 面板的 speculative 选项列表中新增 DFLASH2 选项
{
  id: "dflash",
  label: "DFLASH2",
  // 当选择 RTX 5090 且不为 nvfp4 时禁用 DFlash2，因为模型仅适配 NVFP4 权重
  disabled: (sel) => sel.hw === "rtx5090" && sel.quant !== "nvfp4",
```

```
    disableReason: "On the 32GB RTX 5090 the DFlash2 draft model only fits on top of the NVFP4
    weights",
    // 对未验证平台给出提示
    hints: (sel) =>
      ["rtx6000", "rtx5090"].includes(sel.hw)
        ? []
        : ["DFLASH2 on this platform: final verification in progress"],
    // 针对 RTX 5090 重设 mem-fraction 和 mamba-full-memory-ratio，避免冲突
    stripPrefixes: (sel) =>
      sel.hw === "rtx5090"
        ? sel.ssmDtype === "float32"
          ? ["--mem-fraction-static", "--mamba-full-memory-ratio"]
          : ["--mem-fraction-static"]
        : [],
    // 构建命令行 flags，包含 DFlash2 相关参数和 RTX 5090 的特殊调整
    flags: (sel) => [
      "--speculative-algorithm DFLASH",
      "--speculative-draft-model-path incoai/Qwen3.8-27B-DFlash2",
      "--speculative-num-draft-tokens 8",
      ...(sel.hw === "rtx5090"
        ? sel.ssmDtype === "float32"
          ? ["--mem-fraction-static 0.945", "--mamba-full-memory-ratio 10"]
          : ["--mem-fraction-static 0.90"]
        : []),
    ],
  },
```

评论区精华

无 review 评论，仅有一个来自 mintlify[bot] 的文档预览链接，用于查看变更后的 cookbook 页面。未发现实质性讨论。

风险与影响

- 风险：文档引导用户使用未验证的配置可能导致部署失败或性能不佳，虽然已有提示，但用户可能忽略。此外，文档中的参数为实测值，未来引擎更新可能使其不再适用。
- 影响：影响范围局限在文档，为用户提供 DFLASH2 的部署指引，尤其为 NVFP4 用户提供了新选项。对系统无代码影响，对团队维护成本低。

关联脉络

本 PR 与 #35496 直接关联，该 PR 实现了量化 lm_head 的 DFLASH2 选择器，使 NVFP4 模型可用 DFLASH2。此外，同系列有其他涉及 Qwen3.8-27B 的文档和功能 PR，如 DSpark 配置，但本 PR 主要聚焦 DFLASH2 的文档补充。未来可能继续演进 cookbook 以支持更多模型和硬件组合。