

PR #34250 完整报告

sgl-project/sglang

Update dspark draft path in Inkling small cookbook

合并时间: 2026-08-10 17:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34250>

执行摘要

本 PR 将 Inkling-Small cookbook 中 DSPARK 投机解码的 draft checkpoint 路径从 [RadixArk/Inkling-Small-DSpark-Preview](#) 统一更新为正式发布的 [RadixArk/Inkling-Small-DSpark](#), 涉及 2 个文档文件、共 5 行替换。纯文档维护, 不影响任何运行时逻辑, 重要性低但保持文档与实际可用模型一致。

功能与动机

PR body 为贡献模板、未填写实质内容, 关联 Issue 为空。从分支名 [use-released-dspark-draft](#) 与变更内容可判断: DSpark 的独立 draft checkpoint 已从 Preview 状态转正发布, 而 cookbook 中 4 处硬件配置与 §3.7 正文仍引用 [-Preview](#) 路径。继续使用旧路径会让按文档部署的用户拉取到过时权重或无法解析仓库名, 因此统一替换为正式名称 [RadixArk/Inkling-Small-DSpark](#)。

实现拆解

1. 更新启动参数片段 (docs/src/snippets/configs/thinkingmachines/inkling-small.jsx) :
在 b200、b300、gb300、h200 四个 strategy: "dspark" 配置块中, 将 `--speculative-draft-model-path` 的值从 [RadixArk/Inkling-Small-DSpark-Preview](#) 改为 [RadixArk/Inkling-Small-DSpark](#)。同一配置块中的 `--speculative-algorithm DSPARK` 与 `--speculative-draft-model-quantization unquant` 保持不变, draft 权重仍以非量化方式加载。
2. 同步手册正文 (docs/cookbook/autoregressive/ThinkingMachines/Inkling-Small.mdx)
: §3.7 中对 DSpark 的文字描述同步替换 checkpoint 名称, 保留“独立 draft checkpoint、与 NVFP4 目标并置服务”的整体表述。
3. 配套与验证: 无测试与部署配置改动。mintlify bot 已生成文档 Preview 且状态为 Ready, 渲染通过; CI 的 Extra 任务显示失败 (x), 但该 PR 由作者直接合并, 失败任务与本文档变更无直接证据关联。

[docs/src/snippets/configs/thinkingmachines/inkling-small.jsx](#)

本次变更主体: b200/b300/gb300/h200 四个 DSPARK 部署配置块中的 `--speculative-draft-model-path` 从 [-Preview](#) 统一改为正式路径 [RadixArk/Inkling-Small-DSpark](#), 是用户复制的核心启动参数来源。

```
// docs/src/snippets/configs/thinkingmachines/inkling-small.jsx 中
// B200 NVFP4 + DSPARK 单机配置块; b300/gb300/h200 的块仅 tp 与
```

```

// mem-fraction 数值不同，draft 路径改动完全一致。
// DSpark 使用独立 draft checkpoint 做投机解码，该 checkpoint 不做 NVFP4
// 量化，因此其权重不占用 FP4 目标配额，`mem-fraction-static` 可放宽到 0.68。
{
  match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "dspark", nodes: "single" },
  verified: true,
  env: [
    // 启用统一 Radix Cache，配合下方 mamba-radix-cache-strategy 的 extra_buffer 策略
    "SGLANG_ENABLE_UNIFIED_RADIX_TREE=1",
  ],
  flags: [
    "--trust-remote-code",
    "--model-path {{MODEL_NAME}}",
    "--tp 8",
    "--quantization modelopt_fp4", // 主模型走 NVIDIA ModelOpt FP4 量化
    "--attention-backend fa4",
    "--page-size 128",
    "--fp4-gemm-backend flashinfer_trtllm",
    "--moe-runner-backend flashinfer_trtllm_routed",
    "--enable-torch-symm-mem",
    "--mamba-radix-cache-strategy extra_buffer",
    "--mem-fraction-static 0.68", // draft 权重在显存配额之外，静态占比可略高
    "--swa-full-tokens-ratio 0.1",
    "--mamba-full-memory-ratio 0.1",
    "--enable-multimodal",
    "--max-running-requests 68",
    "--reasoning-parser inkling",
    "--tool-call-parser inkling",
    "--skip-server-warmup",
    "--speculative-algorithm DSPARK",
    // 本次变更：由 Preview 版本切换为正式发布的 draft checkpoint
    "--speculative-draft-model-path RadixArk/Inkling-Small-DSpark",
    "--speculative-draft-model-quantization unquant",
    "--chunked-prefill-size 8192",
    "--cuda-graph-max-bs-prefill 8192",
    "--disable-flashinfer-autotune",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
}

```

评论区精华

本 PR 无 review 评论，唯一评论来自 [mintlify\[bot\]](#) 的自动预览通知：

“Preview deployment for your docs... Project lmsysorg — Status  Ready — View Preview。”

该评论仅确认文档站点预览渲染成功，未产生任何技术交锋或设计讨论。

风险与影响

整体风险极低，但有两个可留意点：

1. 新路径 RadixArk/Inkling-Small-DSpark 依赖 Hugging Face 上存在对应模型仓库，若仓库被下架或改名，按 cookbook 部署的用户会拉取失败。配置块中 `verified: true` 仅明确标注 B200 端到端验证过，其余硬件的验证情况未单独说明。
2. 已按旧路径 -Preview 部署的存量用户可能产生路径不一致的困惑，属于文档演进而非回归。另外 CI Extra 任务失败原因未知，建议后续观察是否与文档站点构建相关。

影响范围限定在文档与示例配置的用户侧：按 cookbook 部署 DSPARK 策略的用户将默认拉取正式版 draft checkpoint；对 SGLang 运行时、调度、内核零影响。

关联脉络

历史 PR [#34184](#) 的端到端测试文件 `test_inkling_small_nvfp4.py` 覆盖同模型的 NVFP4 部署路径，说明 Inkling-Small 一直是 Thinking Machines 模型线在 Blackwell NVFP4 场景的重点验证对象。本 PR 将 DSPARK draft 路径对齐到正式发布版本，是 Inkling-Small 部署能力演进中的文档收尾动作，与此前 MTP/BCG、NVFP4 等模型侧改造形成闭环。