

PR #36204 完整报告

sgl-project/sglang

docs: mark Ling-3.0-flash DSPARK verified for all four quantizations on H200

合并时间: 2026-08-25 02:05

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36204>

执行摘要

本 PR 将 Ling-3.0-flash 的 DSPARK 投机解码验证从仅 BF16 扩展到 H200 上的所有四种量化 (BF16、FP8、INT4、MXFP4)，并在部署面板中标记这些配置为已验证。基于 PR #33561 的 head commit (76a3e6732b) 在 4xH200 上完成完整 GSM8K 评测，精度介于 96.13% 至 96.36% 之间。变更包括配置脚本重构、基准数据补充和文档更新，不涉及运行时逻辑。

功能与动机

DSPARK 是除 NEXTN 之外的另一条投机解码路径，使用外部草稿模型。此前仅在 BF16 (4xB200) 上验证，其他量化缺乏确认。为增强用户信心，作者在 4xH200 上补齐了所有量化的验证，并在部署面板中明确标记，避免用户误用未验证配置。PR 描述: 'This marks the remaining quantizations verified on 4xH200'。

实现拆解

- ling-3.0-flash.jsx: 引入 DSPARK_VERIFIED 集合 (包含 b200lbf16、h200lbf16、h200lfp8)，替代原先只判断 b200+bf16 的硬编码，使孪生单元格的验证状态由集合驱动；新增 INT4 与 MXFP4 的 direct 单元格，因为它们没有 NEXTN 孪生来源，需直接列出参数 (如 --tp 2、--mem-fraction-static 0.85)。
- ling-3.0-flash-benchmarks.jsx: 为四个 H200 DSPARK 单元格新增基准条目，包括精度、停止率、接受长度，并注明源码版本。
- Ling-3.0-flash.mdx: 在 DSPARK 提示中补充 MXFP4 必须指定 --moe-runner-backend flashinfer_mxfp4，否则可能因 hidden-size 不匹配导致 decode graph 捕获失败。

docs/src/snippets/configs/inclusionAI/ling-3.0-flash.jsx

重构 DSPARK 验证逻辑，引入验证集合，新增 INT4/MXFP4 单元格。

```
// DSPARK 验证集合: hwlquant 对，已通过完整 GSM8K 评测
const DSPARK_VERIFIED = new Set(["b200lbf16", "h200lbf16", "h200lfp8"]);

// 生成 DSPARK 孪生单元格: 继承低延迟单元格，仅替换 NEXTN 标志
const dsparkTwin = (cell, verified) => ({
  ...cell,
  verified,
  match: { ...cell.match, spec: "dspark" },
  flags: cell.flags.flatMap((f) =>
```

```

    f === "--speculative-algorithm NEXTN" ? DSPARK_FLAGS : [f]
  ),
});

// 低延迟单元格与 DSPARK 孪生单元格组合
const lowLatencyCells = [...];
return [
  ...lowLatencyCells.flatMap((c) => [
    c,
    dsparkTwin(c, DSPARK_VERIFIED.has(`${c.match.hw}|${c.match.quant}`)),
  ]),
  // INT4/MXFP4 无 NEXTN 孪生来源，直接列出验证过的配置
  {
    match: { hw: "h200", quant: "int4", strategy: "low-latency", spec: "dspark" },
    verified: true,
    flags: ["--model-path {{MODEL_NAME}}", "--tp 2", ...DSPARK_FLAGS, "--mem-fraction-static 0.85"],
  },
  {
    match: { hw: "h200", quant: "mxfp4", strategy: "low-latency", spec: "dspark" },
    verified: true,
    flags: ["--model-path {{MODEL_NAME}}", "--tp 2", "--moe-runner-backend flashinfer_mxfp4", ...DSPARK_FLAGS, "--mem-fraction-static 0.85"],
  },
];

```

docs/src/snippets/configs/inclusionAI/ling-3.0-flash-benchmarks.jsx

新增四个 H200 DSPARK 单元格的基准数据。

```

// H200 + BF16 DSPARK 基准
{
  match: { hw: "h200", quant: "bf16", strategy: "low-latency", spec: "dspark" },
  sglang_version: "PR #33561 @ 76a3e673",
  accuracy: { gsm8k_pct: 96.36 },
  notes: "Full GSM8K stop rate 99.62%; accept length ~4.5-5.1.",
},
// ... 其余量化条目类似

```

评论区精华

无 review 评论。Issue 评论 yuan-luo 询问是否推进 PR #33561，说明本 PR 依赖其合并。

风险与影响

依赖 PR #33561 的特定 commit，若该 PR 变更可能导致验证失效；MXFP4 配置依赖 flashinfer_mxfp4 后端，环境不支持时需注意。影响范围限于文档与配置面板，对运行时无影响。

关联脉络

本 PR 是 PR #33561 的验证补充，将 DSPARK 验证扩展到 H200 全量化，明确了该功能的适用范围。Issue 评论提示应推进 PR #33561 合并，以使文档引用稳定。