PR #25359 完整报告

sgl-project/sglang

[Docs] MiMo-V2.5 cookbook: B200 benchmarks + multi-layer EAGLE acceptance profile + long-context reference

合并时间：2026-05-20 14:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25359

执行摘要

本次 PR 为 MiMo-V2.5 cookbook 补充了 8×B200 的基准测试数据（GSM8K、延迟、吞吐），新增多层 EAGLE 接受率分析和长上下文参考章节。同时修正了部署生成器中的两个配置问题：JSON 布尔值类型和 `--enable-multi-layer-eagle` 的条件限制，使 Blackwell 用户也能正确启用多层 EAGLE。文档变更已在 8×B200 上验证通过。

功能与动机

此前 MiMo-V2.5 cookbook 中有四个 Pending update 区块，缺乏 B200 上的实际性能数据；同时 `--enable-multi-layer-eagle` 由于多线程加载器 OOM 问题（#25748）在 Blackwell 上被禁用，用户无法在 B200 上使用多层 EAGLE。随着 #25748 修复，需要更新文档以反映正确配置。

实现拆解

1. 填充 benchmark 数据（MiMo-V2.5.mdx）：替换 §5.1.1 GSM8K、§5.2.1 延迟、§5.2.2 吞吐的占位符为实测结果，包括分数、延迟、吞吐率。
2. 新增分析章节：§5.4 基于 307 个 server 日志统计出多层 EAGLE 的接受率（0.755）和接受长度（3.27/4），并说明随机 prompt 下 EAGLE 收益消失的原因。§5.5 引用 #23808 的数据作为长上下文参考。
3. 修正 JSON 布尔值（mimo-v25-deployment.jsx）：将 `enable_multithread_load` 从字符串 `"true"` 改为布尔值 `true`，避免用户启动时解析错误。
4. 移除 Blackwell 限制：在 mimo-v25-deployment.jsx 中将 `if (!blackwell) flags.push("--enable-multi-layer-eagle")` 改为无条件推送，使 B200 也能启用多层 EAGLE。
5. 更新说明文字：在 §3.2 中明确多层 EAGLE 同时支持 Hopper 和 Blackwell。

docs_new/src/snippets/autoregressive/mimo-v25-deployment.jsx

部署命令生成器的关键修正：修复 JSON 布尔值和移除条件限制，直接影响用户复制的启动命令的正确性。

```
// 提取自 MiMoV25Deployment 组件，展示启动参数生成逻辑中的关键修正
if (isPro) {
  if (blackwell) {
    // ... 其他 flags
    // 修正：enable_multithread_load 为布尔值，非字符串
```

```
      flags.push(`  --model-loader-extra-config '{"enable_multithread_load": true, "num_threads":
      64}'`);
    } else {
      // Hopper 分支同样修正布尔值
      flags.push(`  --model-loader-extra-config '{"enable_multithread_load": true, "num_threads":
      64}'`);
    }
  }

  if (useMtp) {
    flags.push("  --speculative-algorithm EAGLE");
    flags.push("  --speculative-num-steps 3");
    flags.push("  --speculative-eagle-topk 1");
    flags.push("  --speculative-num-draft-tokens 4");
    // 移除 !blackwell 判断，在 Blackwell 和 Hopper 上均启用多层 EAGLE
    flags.push("  --enable-multi-layer-eagle");
  }
```

评论区精华

无审查评论，PR 直接被批准。关联 Issue #25748 讨论了多线程加载器 OOM 修复，该修复通过将 _filter_mtp_weights 改为 generator 避免了全量加载，使 Blackwell 上多层 EAGLE 得以启用。

风险与影响

- 风险: 低。主要涉及文档和配置修正，无核心逻辑变更。
- 影响: 为 MiMo-V2.5 用户提供准确的 Blackwell 基准测试和正确的部署命令；使 B200 用户能够使用多层 EAGLE 获得解码加速。

关联脉络

- 关联 Issue #25748: 其修复的多线程加载器 OOM 是本次移除 Blackwell 限制的技术前提。
  - 此 PR 与近期 #25774（移除 output_ids 重构）等调度层变更无直接关系，但展示了文档如何紧跟运行时修复进行同步。