PR #34395 完整报告

sgl-project/sglang

[Docs] Add Ling-3.0-tiny INT4 recipes

合并时间：2026-08-12 03:16

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34395

执行摘要

本 PR 为 Ling-3.0-tiny 模型 cookbook 新增 INT4 量化部署配方，提供 H200 与 B200 单卡实测验证，并附上完整 GSM8K 精度、stop rate 与固定 8K/1K 并发 1/16 的性能数据。改动只涉及三个文档文件，目标是让社区用户按最小命令（仅 --model-path、--host、--port）直接部署公开的 inclusionAI/Ling-3.0-tiny-int4 checkpoint，由 SGLang 自动选择 Hopper 上的 Marlin 与 Blackwell 上的 Triton WNA16 后端。

功能与动机

PR body 明确提出："add the public Ling-3.0-tiny INT4 checkpoint to the cookbook deployment matrix""add verified single-GPU H200 and B200 recipes with default decode CUDA Graph enabled"。动机是补齐 Ling-3.0-tiny 的量化部署矩阵：此前只有 BF16/FP8 配方，INT4 模型可显著降低显存（约 5.8 GB vs BF16 15.8 GB），但缺少经过验证的官方部署指引。文档同时强调"keep launch commands minimal and let SGLang select Marlin on Hopper and Triton WNA16 on Blackwell"，即用户不需要手动指定量化或 MoE 后端标志。

实现拆解

1. 部署配置矩阵扩展（docs/src/snippets/configs/inclusionAI/ling-3.0-tiny.jsx） - 在 quantizations 中新增 { id: "int4", label: "INT4" }，并在 modelNames 中增加 "defaultlint4": "inclusionAI/Ling-3.0-tiny-int4"。 - 基准命令更新：speed 增加 --random-range-ratio 1 使请求长度固定；GSM8K 评测增加 --temperature 1.0 --top-p 0.95 --thinking。 - 为 h20-3e、h200、h800、h100、b200、gb300 六个硬件各新增一个 INT4 cell，其中仅 h200 与 b200 标记 verified: true，其余保持未验证状态。
2. 基准数据补充（docs/src/snippets/configs/inclusionAI/ling-3.0-tiny-benchmarks.jsx） - 新增 H200 与 B200 两套 INT4 基准条目，sglang_version 记录为 PR #33561 @ 8ba213fc。 - 记录 P50 TTFT/TPOT、tokens_per_sec_per_gpu、完整 GSM8K 精度（94.54%）与 stop rate（100%），并在注释中说明 INT4 使用 80 条精确长度请求，与 BF16/FP8 原发布口径不同。 - B200 条目额外注明使用了未调优的 Triton WNA16 默认配置。
3. 文档页更新（docs/cookbook/autoregressive/InclusionAI/Ling-3.0-tiny.mdx） - description 与模型清单扩展为 BF16/FP8/INT4 三种量化，新增 INT4 checkpoint 链接。 - 补充 INT4 显存约 5.8 GB 的说明，并强调 SGLang 会自动选择 Marlin（Hopper）或 Triton WNA16（Blackwell），无需显式标志。 - 强调必须使用专用镜像 lmsysorg/sglang:dev-Ling-3.0-tiny，因为它包含 INT4 所需的压缩张量后端。

4. 验证配套 - 未新增自动化测试文件，但 PR body 列出了完整的验证流程：node docs/scripts/check_cookbook_configs.mjs 矩阵完整性校验、Mintlify build 与 broken-links 检查、Playwright 对 H200/B200 INT4 命令和 Docker 镜像的校验、实际 serving 与 GSM8K finish-reason 门控、默认 CUDA Graph 捕获验证等。

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

部署配置矩阵核心文件，新增 INT4 量化项、模型映射、基准命令参数与 6 个硬件 cell，其中仅 H200/B200 标记已验证。

```
// docs/src/snippets/configs/inclusionAI/ling-3.0-tiny.jsx
// 部署配置矩阵中与 INT4 相关的关键字段
export const config = {
  quantizations: [
    { id: "bf16", label: "BF16" },
    { id: "fp8", label: "FP8" },
    // 新增 INT4 量化项，对应公开 checkpoint inclusionAI/Ling-3.0-tiny-int4
    { id: "int4", label: "INT4" },
  ],

  modelNames: {
    "defaultlbf16": "inclusionAI/Ling-3.0-tiny",
    "defaultlfp8": "inclusionAI/Ling-3.0-tiny-fp8",
    "defaultlint4": "inclusionAI/Ling-3.0-tiny-int4",
  },

  benchmarkCommands: {
    // INT4 基准使用精确长度随机请求（--random-range-ratio 1），
    // GSM8K 评测统一采用 temperature 1.0 / top-p 0.95 并开启 --thinking
    speed: `python3 -m sglang.bench_serving ... --random-range-ratio 1 ...`,
    accuracy: {
      gsm8k_pct: `sgl-eval run gsm8k ... --temperature 1.0 --top-p 0.95 --thinking`,
    },
  },

  // 新增 6 个硬件的 INT4 cell，仅 H200 与 B200 经过实测验证（verified: true），
  // 其余硬件（h20-3e/h800/h100/gb300）保持未验证状态
  cells: [
    { match: { hw: "h200", variant: "default", quant: "int4", strategy: "high-throughput", nodes:
    "single" },
      verified: true,
      flags: ["--model-path {{MODEL_NAME}}", "--host {{HOST_IP}}", "--port {{PORT}}"] },
    { match: { hw: "b200", variant: "default", quant: "int4", strategy: "high-throughput", nodes:
    "single" },
      verified: true,
      flags: ["--model-path {{MODEL_NAME}}", "--host {{HOST_IP}}", "--port {{PORT}}"] },
  ],
};
```

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

新增 H200 与 B200 的 INT4 基准条目，记录 P50 TTFT/TPOT、吞吐、GSM8K 精度以及 CUDA Graph 捕获信息。

```
// docs/src/snippets/configs/inclusionAI/ling-3.0-tiny-benchmarks.jsx
// 新增 INT4 基准条目：TTFT/TPOT 均为 P50，精度为完整 GSM8K（1319 题）
export const benchmarks = [
  {
    match: { hw: "h200", variant: "default", quant: "int4", strategy: "high-throughput", nodes:
    "single" },
    sglang_version: "PR #33561 @ 8ba213fc",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 90.31, tpot_ms: 1.96, tokens_per_sec_per_gpu: 4398 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 840.04, tpot_ms: 3.55, tokens_per_sec_per_gpu: 32958 },
    ],
    accuracy: { gsm8k_pct: 94.54 },
    // 完整 GSM8K 的 stop rate 为 100%，默认 decode CUDA Graph 捕获 36 个 shape，覆盖至
    batch 256
    notes: "Full GSM8K stop rate 100%; default decode CUDA Graph captured 36 shapes
    through batch 256.",
  },
  {
    match: { hw: "b200", variant: "default", quant: "int4", strategy: "high-throughput", nodes:
    "single" },
    sglang_version: "PR #33561 @ 8ba213fc",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 305.67, tpot_ms: 6.33, tokens_per_sec_per_gpu: 1359 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 2634.12, tpot_ms: 16.04, tokens_per_sec_per_gpu: 7730 },
    ],
    accuracy: { gsm8k_pct: 94.54 },
    // B200 使用未调优的 Triton WNA16 默认配置 E=128/N=256，性能数据以文档标注为准
    notes: "Full GSM8K stop rate 100%; default decode CUDA Graph captured 52 shapes
    through batch 512. Triton WNA16 used untuned default E=128,N=256 configs.",
  },
];
```

评论区精华

本 PR 无 review 评论、无讨论线程，zijiexia 直接给予 APPROVED。因此没有需要提炼的代码审查交锋。

风险与影响

- 基准口径差异：INT4 数据来自 80 条精确长度请求，而 BF16/FP8 沿用原始发布数据，同表对比时可能产生误导，文档注释已做说明。

- B200 性能待调优：B200 INT4 吞吐显著低于 H200（c1 仅 1359 vs 4398 tok/s），且明确是未调优 Triton WNA16 配置，用户跨卡对比时需注意。
- 运行时依赖：INT4 配方依赖 PR #33561 的运行时支持，需保证该 PR 先合入或同步合入，否则配方不可用。
- 验证覆盖有限：仅 h200 与 b200 实测验证，其余四个硬件 cell 未验证。
- 镜像要求：文档强调需使用专用 dev-Ling-3.0-tiny 镜像，若用户误用普通镜像可能缺少 INT4 后端。

整体影响限于文档与社区部署体验，不涉及运行时代码，风险可控。

关联脉络

本 PR 与 #34363（[Docs] Add Ling-3.0-flash INT4 and MXFP4 recipes）同属 InclusionAI Ling 系列的 cookbook 扩展，两次改动共享相同的 jsx 配置结构、基准数据格式与验证流程，是同一文档工程线的延续。PR body 同时引用了 #33561 作为 INT4 的运行时支持基础，说明该文档是等待运行时能力就绪后补充的部署指南，后续可关注 Ling 系列其他量化变体（如 MXFP4）的配方扩展。