

PR #2009 完整报告

radixark/miles

[docs] Add Inkling-Small model page

合并时间: 2026-07-31 02:05

原文链接: <http://prhub.com.cn/radixark/miles/pull/2009>

执行摘要

- 一句话: 新增 Inkling-Small 276B 模型文档页
- 推荐动作: 该 PR 为纯文档变更, 不值得精读代码细节, 但可作为了解 Inkling 系列模型支持范围的入口。值得关注的是文档中提到的 `scripts/models/inkling-small.sh` 和 `scripts/run_inkling_975b.py` 等配方文件, 它们才是模型真正落地的源码载体, 建议有需要时追溯这些文件。

功能与动机

Inkling-Small 已在 Hugging Face 上以 Apache 2.0 协议公开, 团队需要为社区用户提供对应的 launch recipe 文档; 同时该模型与 Inkling 架构相同, 文档采用“复述基础架构、只写差异”的策略, 降低维护成本。PR body 说明: “Adds the Inkling-Small (276 B / 12 B active) recipe page under Thinking Machines models, now that the model is public on HF (Apache 2.0).”

实现拆解

实现拆解如下:

1. 新增模型页面: 新建 `docs/models/thinkingmachines/inkling-small.md`, 写入 48 行内容, 包括 Inkling-Small 简介 (276 B 总参 / 12 B 激活、42 层、256 routed + 2 shared experts、top-6 sigmoid routing)、与 Inkling 的共用架构说明、Supported Variants 表格、Quick start (全参 GRPO 与 LoRA GRPO 两套命令)、Validated parallelism 表格 (4x8 H200, TP4 SP PP8 EP4) 以及 batch/lr 设置说明。
2. 注册导航: 修改 `docs/docs.json`, 在 Thinking Machines 分组的 `pages` 数组中追加 `models/thinkingmachines/inkling-small`, 使新页面出现在文档侧边栏。
3. 更新索引页: 在 `docs/models/thinkingmachines/index.md` 的模型表格中新增一行 Inkling-Small 记录, 指向新页面。
4. 无配套改动: 无源码、测试、Schema 或部署配套变更。

关键文件:

- `docs/models/thinkingmachines/inkling-small.md` (模块 模型文档; 类别 docs; 类型 documentation): 本 PR 的核心产物, 新增 48 行模型文档页, 包含模型介绍、启动命令与已验证并行配置。

- docs/docs.json (模块 文档配置; 类别 config; 类型 configuration) : 文档站导航配置文件, 新增页面必须在 pages 数组中注册才能出现在侧边栏。
- docs/models/thinkingmachines/index.md (模块 模型文档; 类别 docs; 类型 documentation) : 模型索引页, 新增一行 Inkling-Small 记录, 帮助用户从目录跳转。

关键符号: 未识别

1. Model Introduction

Inkling-Small 是 Thinking Machines Lab Inkling 家族中的紧凑成员: 276 B 总参 / 12 B 激活参数、42 层多模态 MoE (256 routed + 2 shared experts, top-6 sigmoid routing), 在部分基准上甚至超过旗舰 975 B 模型 (如 SWEBench Verified 80.2 vs 77.6), 但部署 footprint 小得多。

3. Quick start

已在 4 nodes × 8 H200 (TP4 SP PP8 EP4, DP1) 上验证:

关键源码片段

<docs/models/thinkingmachines/inkling-small.md>

本 PR 的核心产物, 新增 48 行模型文档页, 包含模型介绍、启动命令与已验证并行配置。

title: Inkling-Small description: Launch recipe for Inkling-Small (276 B), the compact sibling of Inkling — same architecture, 4-node H200 footprint.

1. Model Introduction

Inkling-Small 是 Thinking Machines Lab Inkling 家族中的紧凑成员: 276 B 总参 / 12 B 激活参数、42 层多模态 MoE (256 routed + 2 shared experts, top-6 sigmoid routing), 在部分基准上甚至超过旗舰 975 B 模型 (如 SWEBench Verified 80.2 vs 77.6), 但部署 footprint 小得多。

3. Quick start

已在 4 nodes × 8 H200 (TP4 SP PP8 EP4, DP1) 上验证:

```
# 全参数 GRPO: 276 B 靠 CPU offload 优化器即可装下,  
# 不需要像 975 B 大模型那样走 NVMe 流式卸载。  
python scripts/run_inkling_975b.py train \  
  --model-name Inkling-Small --train-mode full --task dap0_math \  
  --num-nodes 4 --num-gpus-per-node 8 \  
  --sglang-context-length 4096 --rollout-max-response-len 2048 \  
  --extra-args "--offload-train-target cpu --sglang-mem-fraction-static 0.65 \  
  --optimizer-cpu-offload --overlap-cpu-optimizer-d2h-h2d --use-precision-aware-optimizer"
```

```
# LoRA GRPO (rank 32, 全线性层): 同一集群即可跑,  
# adapter-only 权重同步大约 4 秒一次 rollout。
```

```
python scripts/run_inkling_975b.py train \
--model-name Inkling-Small --train-mode lora --task dapo_math \
--num-nodes 4 --num-gpus-per-node 8 \
--sglang-context-length 4096 --rollout-max-response-len 2048
```

评论区精华

本 PR 没有实质 review 讨论，两位 reviewer (Shi-Dong、maocheng23) 均直接 APPROVED，未留下评论。唯一评论来自 gemini-code-assist[bot]，声明其代码审查服务已停止，不具备参考价值。

- 暂无高价值评论线程

风险与影响

- 风险：风险等级较低，主要包括：
 - docs/docs.json 导航配置若路径写错，新页面可能无法在侧边栏展示或产生死链，需要人工验证导航结构。
 - 文档中给出的启动命令（如 --extra-args 中的 --optimizer-cpu-offload 等）未经过 CI 测试，属于人工验证的产物，后续若基础脚本参数变更，文档可能失准。
 - 文档大量引用 Inkling 页面架构说明，若 Inkling 页面后续修改，本页面需要同步维护。
 - 影响：影响范围为文档使用者和团队文档维护：为社区用户提供 Inkling-Small 的启动配方，降低小集群部署门槛；对系统运行时无任何影响。团队内部文档同步工作增加，但该 PR 已通过 index 页和导航注册保持一致。
- 风险标记：纯文档变更，导航配置易错，命令未经 CI 验证

关联脉络

- 暂无明显关联 PR