

PR #2216 完整报告

radixark/miles

docs: add GLM-5.2 model page and update supported-models tables

合并时间: 2026-08-11 05:20

原文链接: <http://prhub.com.cn/radixark/miles/pull/2216>

执行摘要

- 一句话: 新增 GLM-5.2 模型文档页并更新模型表格
- 推荐动作: 值得精读以了解 GLM-5.2 在 Miles 中的启动方式与 DSA 架构约束; 可作为撰写其他模型文档的模板。重点关注 §5.1 的并行分支表和 §5.3 的 SGLang 配置, 以及 .gitignore 修复的细节 (未锚定模式误伤文档目录的教训)。

功能与动机

GLM-5.2 是 GLM 系列新一代模型, 已有启动脚本 `scripts/run_glm5_2_744b_a40b.py` 和相关示例 (#2220), 但文档缺少对该型号的介绍。PR body 明确列出所有事实来源:

`scripts/run_glm5_2_744b_a40b.py`、`scripts/models/glm5.2-744B-A40B.sh` 和 #2220 示例 README, 页面遵循现有 GLM-5 页面的六段式结构, 保证文档体系的一致性。

实现拆解

1. 新增 docs/models/glm/glm5-2.md (+149 行): 按 GLM-5 页面的六段式结构组织, 涵盖模型介绍 (DSA 跨层索引共享、FP8 KV cache、TIS)、支持变体表 (zai-org/GLM-5.2 与 5-layer 剪枝 smoke-test 模型)、环境准备 (prepare/prepare-cp 子命令、HF → Megatron 转换及 _validate_glm_checkpoint 校验)、三种并行分支 (1-node 5-layer、≥16-node GB300 4-GPU/node、≥16-node 8-GPU/node) 和 rollout/SGLang 配置 (fp8_e4m3、flashmla_kv decode、flashmla_sparse prefill)。
2. 更新导航与索引: docs/docs.json 在 GLM 分组中注册 models/glm/glm5-2; docs/models/index.md 和 docs/models/glm/index.md 的表格中新增 GLM-5.2 条目, 并在 "Which variant do I pick?" 中提示 GLM5.2 为当前首选。
3. 修正既有页面: docs/models/glm/glm5.md 交叉链接到新页面, 修复悬空语句。
4. 修复 .gitignore: 移除未锚定的 glm/ 规则 (会匹配 docs/models/glm/ 导致新页面需 git add -f), 改为完全不忽略该目录。
5. 引入 #2220 的 agentic RL 示例引用: 在 §4.2 中链接 examples/experimental/openenv/glm52_tbench2/, 交叉链接到 Fully Async Rollout 文档。

关键文件:

- docs/models/glm/glm5-2.md (模块 模型文档; 类别 docs; 类型 documentation): 新增的 GLM-5.2 核心文档页, 完整覆盖启动脚本、并行策略、FP8 KV cache、TIS 与 SGLang 配置, 是本次 PR 的主体。

- docs/models/glm/index.md (模块 模型文档; 类别 docs; 类型 documentation) : GLM 家族索引页, 新增 GLM-5.2 变体表条目和推荐语, 引导用户选择当前旗舰。
- docs/docs.json (模块 文档站点; 类别 config; 类型 configuration) : 文档站点导航配置, 注册新页面到 GLM 分组, 遗漏会导致导航缺失。
- docs/models/glm/glm5.md (模块 模型文档; 类别 docs; 类型 documentation) : GLM-5 页面增加交叉链接并修复悬空语句, 保持模型文档互通。
- .gitignore (模块 仓库配置; 类别 other; 类型 core-logic) : 修复未锚定的 glm/ 规则误忽略 docs/models/glm/ 目录, 否则新页面无法正常 git add。
- docs/models/index.md (模块 模型文档; 类别 docs; 类型 documentation) : 站点级模型索引表, 将 GLM-5.2 置顶展示。

关键符号: 未识别

评论区精华

两条 review 评论均为作者 Shi-Dong 自提:

- 移除 the launcher's docstring 表述 ("Sounds weird") ——已采纳, 最终文本改为 "The recipe is tested on H200 / B200 / GB300".
- 将 EAGLE speculative decoding 标注为 Optional——已采纳, 最终文本将其归入 `--enable-mtp` 可选 flag 说明。无未解决的争议; 审阅者 yueming-yuan 已 APPROVE。
- 移除 docstring 表述 (style): 已采纳, 改为 "The recipe is tested on H200 / B200 / GB300".
- EAGLE 是否可选 (correctness): 已采纳, 将 EAGLE 标注为 optional, 并在 `--enable-mtp` 说明中注明适用条件。

风险与影响

- 风险: 纯文档变更, 无代码回归风险。轻微风险是:
 - 文档事实准确性依赖脚本 `run_glm5_2_744b_a40b.py` 与示例 README, 若脚本后续调整需同步更新 (如并行分支、SGLang 配置变更)。
 - docs/docs.json 若漏注册会在导航中缺失, 但已正确登记。
 - .gitignore 移除 glm/ 规则后, 根目录下若真存在名为 glm 的本地目录会被 Git 跟踪, 但 PR 中指出树内无此需求, 风险极低。
 - 影响: 影响范围: 文档站点 (导航、模型索引、GLM 家族页)。用户 (ML 工程师) 能获得 GLM-5.2 的完整启动指南, 包括 DSA 约束下的 PP 划分和 FP8 rollout 配置。团队层面延续了 GLM 系列文档结构, 降低示例复现门槛。
- 风险标记: 纯文档变更, 依赖脚本事实准确性

关联脉络

- PR #2220 Add the GLM-5.2 744B x terminal-bench-2 Daytona example: 本 PR 的 Stacked 基础, 新增了 GLM-5.2 的 tbench2 Daytona agentic RL 示例, 文档页 §4.2 直接引用该示例目录。

- PR #2271 docs: refresh the homepage supported-models table: 合并时与 #2271 在 docs/index.md 发生冲突，最终以 main 分支的表格为准（已包含 GLM-5.2 链接）。
- PR #2274 refactor(openenv): move duplicated sandbox helpers into the TB2 recipe: 同一 openenv/tb2 示例链路的后继重构，涉及 tbench2 示例共享代码。