

PR #36977 完整报告

sgl-project/sglang

[Cookbook] Run accuracy benchmarks through sgl-eval

合并时间: 2026-08-29 14:29

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36977>

执行摘要

- 一句话: 三份 cookbook 精度复现命令迁至 sgl-eval run
- 推荐动作: 建议快速浏览: 如需通过 sgl-eval 复现模型精度数字, 本 PR 是标准示例, 重点看三个页面的参数映射 (seeds 到 --n-repeats、tokens_to_generate 到 --max-tokens、并发数到 --num-threads) 与默认 prompt 处理; 无需精读, 不涉及架构或性能决策。

功能与动机

PR body 说明: 重写 DeepSeek-V3.2、MiniMax-M2.7、Kimi-K2.6 页面的 AIME 2025、GPQA Diamond、MMLU-Pro 复现命令, 改用 sgl-eval run (which ships the datasets, prompts and graders these numbers were produced with)。原复现依赖 NeMo-Skills 的 ns prepare_data、ns eval 及大量 ++inference 叠加参数, 流程长且需要额外环境变量处理; 迁移后一条命令即可在本地复现文档中的精度数字, 降低复现门槛。

实现拆解

1. 变更入口: 三份文件的 Accuracy Benchmark 小节整体替换评估工具说明与命令块, 目标工具从 NVIDIA NeMo-Skills 切换为官方 sgl-eval。
2. 各文件具体改动: - DeepSeek-V3_2.mdx (AIME 2025): ns eval --benchmarks=aime25:4 迁移为 sgl-eval run aime25 --n-repeats 4, tokens_to_generate=64000 映射 --max-tokens 64000, 保留 Speciale 模型 120000 token 注释, 删除 NEMO_SKILLS_DISABLE_UNCOMMITTED_CHANGES_CHECK 环境变量。 - MiniMax-M2.7.mdx (GPQA、AIME、MMLU-Pro): 评估设置改为 8 次重复、max_tokens=120000, MMLU-Pro 说明为 greedy 单次; GPQA 默认提示词 eval/aai/mcq-4choices, AIME 默认 generic/math; 三个基准分别迁移为 sgl-eval run gpqa / aime25 / mmlu_pro。 - Kimi-K2.6.mdx (AIME 2025): 32 个随机种子的 for 循环迁移为 --n-repeats 32, max_concurrent_requests=512 映射 --num-threads 512, prompt 显式指定 --prompt matharena-aime。
3. 参数映射逻辑: seeds 对应 --n-repeats, tokens_to_generate 对应 --max-tokens, temperature 与 top_p 直接映射; MiniMax 页移除 parse_reasoning=True, 因为 sgl-eval 对默认基准已内置解析。
4. 提交演进: 第一笔提交完成主体替换 (cookbook: ns eval -> sgl-eval), 第二笔提交补齐 MiniMax 页最后一个遗漏的 MMLU-Pro 代码块 (cookbook: last ns block -> sgl-eval mmlu_pro)。

5. 配套改动：无源码、测试、配置或 CI 联动；唯一外部动作是 Mintlify 自动触发文档预览构建。

关键文件：

- docs/cookbook/autoregressive/MiniMax/MiniMax-M2.7.mdx（模块 文档；类别 docs；类型 documentation）：改动量最大的文件（+26/-45），GPQA、AIME 2025、MMLU-Pro 三个基准的复现命令全部从 ns eval 迁移至 sgl-eval run，并更新评估设置说明（8 次重复、MMLU-Pro 贪婪单次）。
- docs/cookbook/autoregressive/Moonshotai/Kimi-K2.6.mdx（模块 文档；类别 docs；类型 documentation）：AIME 2025 复现从 32 seed for 循环改写为 sgl-eval 单命令，涉及 --n-repeats 32、--prompt matharena-aime、--num-threads 512 等关键参数映射。
- docs/cookbook/autoregressive/DeepSeek/DeepSeek-V3_2.mdx（模块 文档；类别 docs；类型 documentation）：AIME 2025 复现命令迁移，保留 Speciale 模型 120000 token 注释，删除 NeMo-Skills 环境变量，是最轻量的一个文档块替换。

关键符号：未识别

关键源码片段

docs/cookbook/autoregressive/MiniMax/MiniMax-M2.7.mdx

改动量最大的文件（+26/-45），GPQA、AIME 2025、MMLU-Pro 三个基准的复现命令全部从 ns eval 迁移至 sgl-eval run，并更新评估设置说明（8 次重复、MMLU-Pro 贪婪单次）。

```
# 先安装 sgl-eval，数据集、提示词与 grader 随仓库分发，保证与文档数字一致
pip install git+https://github.com/sgl-project/sgl-eval.git
```

```
# --n-repeats 8 对应原 NeMo-Skills 的 8 个随机种子，GPQA 使用 4 选 1 的默认提示词
sgl-eval run gpqa \
  --base-url http://localhost:30000/v1 \
  --model MiniMaxAI/MiniMax-M2.7 \
  --n-repeats 8 \
  --max-tokens 120000 \
  --temperature 0.6 \
  --top-p 0.95
```

docs/cookbook/autoregressive/Moonshotai/Kimi-K2.6.mdx

AIME 2025 复现从 32 seed for 循环改写为 sgl-eval 单命令，涉及 --n-repeats 32、--prompt matharena-aime、--num-threads 512 等关键参数映射。

```
# 原流程用 for 循环准备 32 个随机种子，这里用 --n-repeats 32 一次完成
# --prompt matharena-aime 对应原配置 eval/matharena/aime
sgl-eval run aime25 \
  --base-url http://localhost:30000/v1 \
  --model moonshotai/Kimi-K2.6 \
  --n-repeats 32 \
  --temperature 1.0 \
  --top-p 0.95 \
  --max-tokens 131072 \
```

```
--prompt matharena-aime \  
--num-threads 512
```

评论区精华

本 PR 无实质性 review 讨论。唯一评论来自 mintlify[bot] 的自动预览提示，告知文档预览站点正在构建，不涉及技术内容；作者独立完成两笔提交并自行合并，无公开争议或设计权衡记录。从提交信息可推断，第二笔提交是自查中补齐了 MiniMax 页面遗漏的 MMLU-Pro 命令块。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 参数对应风险：--n-repeats 与原 seeds 的逐位对齐语义（随机种子固定方式）未在文档中展开；--prompt matharena-aime 与库内 prompt 名称的对应关系仅由 sgl-eval 仓库保证，名称变更会直接导致命令失效。
 - 版本耦合风险：文档统一使用 `pip install git+https://github.com/sgl-project/sgl-eval.git`，未锁定版本，CLI 参数演进后旧命令可能失配。
 - 复现验证风险：文档宣称可复现数字，但页面内未给出 sgl-eval 版本号或校验步骤，工具行为变化时不方便核对。
 - 影响范围：仅三份 .mdx 文档与文档预览构建，无运行时、性能或安全风险。
- 影响：
 - 用户影响：按 cookbook 复现精度数字的读者从 NeMo-Skills 切换到 sgl-eval，命令更短、依赖更少，但仍需自行安装 Python 包并核对 CLI 参数。
 - 团队影响：三个模型页面的 benchmark 数字披露统一由 sgl-eval 承载，后续新增模型页可直接复用 sgl-eval run 模板。
 - 影响程度：低，纯文档变更，不改动任何运行时逻辑。
 - 风险标记：参数映射需核对，sgl-eval 未锁版本，纯文档无自动化校验

关联脉络

- PR #36828 Update v4 amd cookbook 0828: 同为 docs/cookbook/autoregressive/DeepSeek 目录下的文档维护，持续修订 benchmark 复现与部署配置信息，与本 PR 处于同一条 cookbook 维护线。
- PR #36950 [Docs] Restore the AIME25 label so GLM-5.3 FP8 and BF16 scores render again: 同为文档修复，都在完善 AIME25 等基准测试分数在文档中的可读性与一致性。