PR #26668 完整报告

sgl-project/sglang

[Doc] Update benchmark instruction for dsv4

合并时间：2026-05-29 14:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26668

执行摘要

- 一句话：更新 DeepSeek-V4 基准测试文档
- 推荐动作：该 PR 为纯粹的文档维护，无代码逻辑变更，对于关注 DeepSeek-V4 评估流程的读者有参考价值。开发团队可将其作为文档更新流程的示例，但无需深入 code review。

功能与动机

更新 DeepSeek-V4 文档中的基准测试部分，以反映最新的评估标准和推荐工具（sgl-eval），确保用户能够正确运行准确率测试并理解预期结果。

实现拆解

1. 添加前置条件：在 Accuracy Benchmark 章节开头增加一段说明，要求设置 SGLANG_DEFAULT_THINKING=1 和 SGLANG_REASONING_EFFORT=max 环境变量，并提示对于 GPQA 和 AIME25 基准运行至少 16 轮以减少随机性。
2. 替换基准测试：将原有的 MMLU Benchmark 配置和结果表格整体移除，替换为 GPQA Diamond Benchmark 和 AIME25 Benchmark 两部分，每部分包含安装 sgl-eval 的命令以及运行测试的示例命令。
3. 提供参考准确率：在命令后添加注释，给出 Flash 和 Pro 模型的参考准确率（如 GPQA: Flash ~95%，Pro ~97.5%；AIME25: Flash ~95%，Pro ~97.5%）。（注：具体数值在 review 中经过调整）
4. 删除冗余内容：清理了旧的测试结果详细数据，使文档更简洁。
5. 格式调整：修改了标题编号以匹配新内容。

关键文件：

- docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx（模块 基准测试文档；类别 docs；类型 docs-update）：本 PR 唯一修改的文件，更新了 DeepSeek-V4 模型的准确率基准测试指南。

关键符号：未识别

关键源码片段

docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx

本 PR 唯一修改的文件，更新了 DeepSeek-V4 模型的准确率基准测试指南。

```
# 启动模型前的环境变量
export SGLANG_DEFAULT_THINKING=1
export SGLANG_REASONING_EFFORT=max

# 安装 sgl-eval
pip install git+https://github.com/sgl-project/sgl-eval

# GPQA Diamond 基准（Flash 参考准确率 ~95%）
sgl-eval run gpqa --model deepseek-ai/DeepSeek-V4-Flash --api-key <api-key> --n-repeats 16 --max-tokens 200000 --temperature 1.0 --top-p 1.0 --thinking --out-dir /sgl-workspace/logs --base-url http://localhost:30000/v1

# AIME25 基准（Pro 参考准确率 ~97.5%）
sgl-eval run aime25 --model deepseek-ai/DeepSeek-V4-Pro --api-key <api-key> --n-repeats 16 --max-tokens 400000 --temperature 1.0 --top-p 1.0 --thinking --out-dir /sgl-workspace/logs --base-url http://localhost:30000/v1
```

评论区精华

主要的讨论发生在 PR 内部，作者（同时也是合并者）在 review 中自行提出了几处 suggestion 来调整参考准确率数值。例如，将 Flash 模型在 GPQA 上的参考准确率从 ~95% 改为 ~97.5%，随后又在另一个 suggestion 中改为 ~95%。最终提交的版本采纳了 ~95%（GPQA Flash）、~97.5%（GPQA Pro）等数值。整个过程体现了作者对基准测试结果的谨慎确认，但由于无其他 reviewer 参与，讨论较为简单。

- GPQA 和 AIME25 参考准确率数值调整 (documentation): 作者通过多次 suggestion 和 commit 微调，最终确定了合理的参考准确率数值。

风险与影响

- 风险：风险极低。由于仅包含文档更新，不会对系统稳定性和性能产生影响。潜在风险是如果基准测试命令或环境变量设置有误，可能导致用户运行测试失败；但这些命令已由作者验证过（从 commit 历史看，作者多次调整）。需要留意的是，文档中引用的 sgl-eval 工具需用户自行安装，如果该工具发生变化可能需要同步更新文档。
- 影响：影响范围限定于阅读 DeepSeek-V4 基准测试文档的用户，主要为模型评估者。变更帮助他们使用更权威的基准测试和更便捷的工具，提升了文档的准确性和可用性。无其他系统影响。
- 风险标记：低风险文档变更，仅文档

关联脉络

- PR #26662 [AMD][CI] Update v4 CI setting and move the task to main branch: 同样涉及 DeepSeek-V4 模型的测试配置更新，虽侧重 CI 而本文档侧重基准测试指南，但都服务于 V4 模型的评估流程。