

PR #2254 完整报告

THUDM/slime

docs: correct reverse KL definition in OPD guide

合并时间: 2026-08-12 13:28

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2254>

执行摘要

- 一句话: 修正 OPD 文档逆 KL 定义并补充采样细节
- 推荐动作: 无需精读, 但建议阅读公式部分以支持后续开发。值得关注两点: 一是逆 KL 方向的定义符合 Thinking Machines Lab 的标准; 二是明确采样误差的说明 (单个 d_t 可能为负), 有助于理解 OPD 的随机性。

功能与动机

OPD 文档中原始逆 KL 定义方向错误 (原来写为 $D_{KL}(P_{teacher} || P_{student})$), 与 Thinking Machines Lab 的定义不一致, 容易误导用户理解 OPD 机制。PR 旨在纠正该定义, 并澄清 slime 实际采用按采样 token 的 Monte Carlo 近似, 而非全词表枚举。

实现拆解

1. 修改英文文档 docs/en/advanced/on-policy-distillation.md: 重写逆 KL 公式为 $D_{KL}(student || teacher)$, 明确期望按学生分布采样, 并添加 Monte Carlo 贡献 d_t 与 advantage 修正公式。
2. 修改中文文档 docs/zh/advanced/on-policy-distillation.md: 与英文同步, 确保术语一致。
3. 补充外部教师兼容性说明: 强调 SGLang 教师需与学生 tokenizer/ 词表兼容, 因为教师需为学生采样 token 评分。
4. 文档构建验证: 通过 docs/build.sh en/zh 和 pre-commit 检查, 确保文档可正常构建。

关键文件:

- docs/en/advanced/on-policy-distillation.md (模块 文档; 类别 docs; 类型 documentation): 修正逆 KL 定义并补充采样细节, 是核心文档变更。
- docs/zh/advanced/on-policy-distillation.md (模块 文档; 类别 docs; 类型 documentation): 与英文文档同步, 保证中英文一致。

关键符号: 未识别

评论区精华

无 review 评论或讨论线程, 由作者直接提交并合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及文档说明，不涉及代码逻辑。但需确保 `--opd-kl-coef` 参数名与实际实现一致，且公式推导正确，避免误导用户。
- 影响：影响范围限于文档读者，主要帮助用户正确理解 on-policy distillation 机制，减少实施时的误解。对系统和代码无影响。
- 风险标记：纯文档变更

关联脉络

- PR #2189 [Doc] Clarify PPO/Critic docs after #1856: 同为文档澄清类 PR，且涉及训练与奖励机制说明。