

PR #27502 完整报告

sgl-project/sglang

Add mixed-prefix gsm8k eval and its CPU unit test

合并时间: 2026-06-09 20:17

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27502>

执行摘要

- 一句话: 新增混合前缀 GSM8K 评估及 CPU 单元测试
- 推荐动作: 值得精读。该 PR 展示了如何通过提取虚方法 (`_setup_prefix_pool` / `_build_prefix`) 在测试框架中实现模板方法模式, 使得扩展不同前缀策略变得简单。单元测试也写得清晰全面, 可作为测试驱动开发的范例。

功能与动机

PR body 说明: `mixed_prefix_gsm8k` builds each question's few-shot prefix from a primary block plus a per-question random sample of a secondary pool, so an evaluation run exercises distinct, partially-shared cached prefixes (instead of one identical few-shot prefix shared by every request). This is the eval used to detect chunked-prefill / prefix-cache corruption: a clean run scores normally, a corrupting bug collapses the score.

实现拆解

1. 重构 base class `GSM8KEval` (`python/sglang/test/simple_eval_gsm8k.py`) - 将原本在 `__init__` 中直接计算 `_few_shot_prompt` 的逻辑拆分为两个虚方法: `_setup_prefix_pool` 和 `_build_prefix`。- `_setup_prefix_pool` 负责从原始数据行中划分训练 / 测试集并构建基础前缀; `_build_prefix` 返回特定索引的前缀字符串。- 原先的 `self._few_shot_prompt` 改为通过 `self._build_prefix(idx)` 调用, 默认行为不变。- 同时将 `num_examples` 越界处理从隐式切片改为显式静默截断 (切片行为保持不变, 但避免硬 `assert` 被优化掉)。
2. 新增子类 `MixedPrefixGSM8KEval` (`python/sglang/test/simple_eval_mixed_prefix_gsm8k.py`) - 继承自 `GSM8KEval`, 重写 `_setup_prefix_pool` 将数据集分割为 `_primary_shots` (固定前 N 条) 和 `_secondary_pool` (接下来 `secondary_pool_size` 条)。- 重写 `_build_prefix`: 使用随机种子 (种子 = 用户指定 `seed` + 问题索引) 决定主块使用多少条 (`0~num_shots`), 并从辅助池随机采样 `0~secondary_pool_size` 条, 拼接后返回。
3. 集成到评测入口 (`python/sglang/test/run_eval.py`) - 在 `run_eval` 的 `dispatch` 中添加 `elif args.eval_name == "mixed_prefix_gsm8k"` 分支, 实例化 `MixedPrefixGSM8KEval`。- 添加两个 CLI 参数: `--mixed-prefix-gsm8k-secondary-pool-size` (默认 15) 和 `--mixed-prefix-gsm8k-seed` (默认 42)。
4. 添加 CPU 单元测试 (`test/registered/unit/bench/test_mixed_prefix_gsm8k.py`) - 使用 `register_cpu_ci` 注册到 CI 基础测试套件, 不依赖 GPU。- 生成合成数据集, 验证

`_build_prefix` 的行为：

- 主段总是主 shots 的前缀；
- 剩余问题全部来自辅助池且无重复；
- 剩余大小在辅助池范围内；
- 不同查询的主深度和辅助大小具有多样性；
- 任意两个查询共享至少 0 个主 shots（即主段长度可不同）。

关键文件：

- `test/registered/unit/bench/test_mixed_prefix_gsm8k.py`（模块 单元测试；类别 test；类型 test-coverage；符号 `_write_synthetic_dataset`, `TestMixedPrefixGSM8KEval`, `setUpClass`, `tearDownClass`）：新增的 CPU 单元测试文件，独立验证 `MixedPrefixGSM8KEval` 的前缀构造逻辑（合成数据集、多种断言），保证评估工具的可靠性，无需 GPU 即可运行。
- `python/sglang/test/simple_eval_mixed_prefix_gsm8k.py`（模块 评估逻辑；类别 test；类型 test-coverage；符号 `MixedPrefixGSM8KEval`, `init`, `_setup_prefix_pool`, `_build_prefix`）：核心子类 `MixedPrefixGSM8KEval`，实现了混合前缀构造逻辑：从主 shots 中取前缀，再从辅助池随机抽样拼接。是整个评估功能的关键。
- `python/sglang/test/simple_eval_gsm8k.py`（模块 基类重构；类别 test；类型 test-coverage；符号 `_setup_prefix_pool`, `_build_prefix`）：基类 `GSM8KEval` 重构，将前缀构建抽为虚方法 `_setup_prefix_pool` 和 `_build_prefix`，默认行为不变，但使子类可扩展。
- `python/sglang/test/run_eval.py`（模块 命令行入口；类别 test；类型 test-coverage）：CLI 入口增加 `mixed_prefix_gsm8k` 分支和两个参数（辅助池大小、随机种子），使用户可通过命令行直接调用。

关键符号：`MixedPrefixGSM8KEval.init`, `MixedPrefixGSM8KEval._setup_prefix_pool`, `MixedPrefixGSM8KEval._build_prefix`, `GSM8KEval._setup_prefix_pool`, `GSM8KEval._build_prefix`, `TestMixedPrefixGSM8KEval._decompose`, `TestMixedPrefixGSM8KEval.test_primary_segment_is_strict_prefix_of_primary_shots`

评论区精华

仅有一条来自 `gemini-code-assist[bot]` 的 review 评论，针对 `simple_eval_gsm8k.py` 中使用 `assert` 进行参数验证的风险。机器人建议改用 `ValueError` 以避免 Python `-O` 优化绕过。作者在后续 commit (`722fd4b`) 中改为使用切片静默截断 (slicing) 而非断言，从而修复了该问题。讨论状态：已解决。

- 使用 `assert` 验证参数的风险 (correctness): 作者在 commit `722fd4b` 中改为使用列表切片进行静默截断，避免 `assert` 被绕过，同时保持与上游调用者（如 `test_basic_sanity_eagle3`）的兼容。

风险与影响

- 风险：

1. 基类重构风险: GSM8KEval 的前缀构建从内联改为方法调用, 虽保持默认行为不变, 但任何直接访问 `_few_shot_prompt` 的外部代码 (本项目或其他依赖) 会失效。经过代码搜索, `_few_shot_prompt` 仅在类内部使用, 影响可控。
2. 数据集大小假设: MixedPrefixGSM8KEval 要求数据集至少包含 `num_shots + secondary_pool_size + 1` 条数据, 否则抛出 `ValueError`。用户若使用小数据集需调整参数。
3. 随机种子确定性: 使用固定种子 `seed + idx`, 理论上每次运行结果相同, 但若随机库实现变化 (如 Python 版本升级) 可能影响一致性。测试中使用合成数据, 不依赖外部数据集再现性。
4. 测试覆盖: 单元测试覆盖了核心构造逻辑, 但不涉及实际模型推理, 因此对缓存损坏的检测能力需要通过端到端评估验证。
 - 影响: 用户影响: 增加一个新的 `eval_name mixed_prefix_gsm8k`, 可通过 CLI 直接使用, 为用户提供检测 `prefix-cache` 完整性的现成工具。系统影响: 不修改推理路径或调度核心, 不影响生产服务。团队影响: 为后续 `chunked-prefill` 相关测试提供可复用的评估框架, 降低回归测试编码成本。影响程度: 中等, 主要影响测试和调试流程。

- 风险标记: 基类重构需验证向后兼容, 数据集大小假设可能限制使用场景

关联脉络

- PR #27506 Add more testing for chunked prefill: 同为 `chunked prefill` 测试基础设施的扩展, 引入了 `scripted runtime` 测试框架。本 PR 提供的 `mixed_prefix_gsm8k` 评估与 `scripted` 测试互补, 共同检测 `prefix-cache` 一致性问题。