PR #49881 完整报告

vllm-project/vllm

[CI] Increase Qwen3.5 MTP GSM8K generation length

合并时间：2026-07-28 18:04

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49881

执行摘要

- 一句话：增大 Qwen3.5 MTP GSM8K 评估生成长度
- 推荐动作：可直接合入。变更简单明确，已获得批准，且不影响其他功能。

功能与动机

GSM8K 评估的默认 max_tokens=256 过短，导致 Qwen3.5 MTP3 的思维链被截断，提取的最后一个整数可能不正确，准确率（0.8469）低于阈值（0.8500）。PR body 明确说明了失败构建链接和数值依据。

实现拆解

1. 在 tests/evals/gsm8k/configs/Qwen3.5-397B-A17B-NVFP4-DEP2-MTP.yaml 中新增 max_tokens: 12000 字段，覆盖默认的 256。
2. 将 server_args 中的 --max-model-len 从 4096 改为 16384，为推理提供足够上下文。
3. 非 MTP 配置不变，避免影响其他评估。

关键文件：

- tests/evals/gsm8k/configs/Qwen3.5-397B-A17B-NVFP4-DEP2-MTP.yaml（模块 测试配置；类别 test；类型 test-coverage）：唯一变更文件，调整 GSM8K 评估的生成长度和模型长度。

关键符号：未识别

关键源码片段

tests/evals/gsm8k/configs/Qwen3.5-397B-A17B-NVFP4-DEP2-MTP.yaml

唯一变更文件，调整 GSM8K 评估的生成长度和模型长度。

```
# tests/evals/gsm8k/configs/Qwen3.5-397B-A17B-NVFP4-DEP2-MTP.yaml
model_name: "nvidia/Qwen3.5-397B-A17B-NVFP4"
accuracy_threshold: 0.88
tolerance: 0.03
num_questions: 1319
num_fewshot: 5
# 新增：覆盖默认 max_tokens=256，确保 MTP 推理完整
max_tokens: 12000
```

```
server_args: >-
  # 从 4096 提升至 16384，为长思维链提供足够上下文
  --max-model-len 16384
  --data-parallel-size 2
  --enable-expert-parallel
  --max-num-seqs 384
  --spec-method mtp
  --spec-tokens 3
```

评论区精华

review 中 AndreasKaratzas 批准，但建议若非紧急则让 khluu 审阅。机器人 chatgpt-codex-connector[bot] 要求补充测试证据和结果。

- 补充评估证据 (question): 未明确回复，但 AndreasKaratzas 已批准，认为只要不超时即可。

风险与影响

- 风险：风险低。仅修改单一测试配置文件，增大生成长度和模型长度可能增加 CI 运行时间，但仍在 45 分钟超时内（AndreasKaratzas 确认）。
- 影响：仅影响 Qwen3.5-397B MTP3 的 GSM8K 评估配置，其他配置不变。修复后 MTP 评估准确率应恢复至阈值以上。
- 风险标记：测试配置变更

关联脉络

- 暂无明显关联 PR