PR #49243 完整报告

vllm-project/vllm

[CI] Add gemma-4-E4B-it-assistant to CI gsm8k for GemmaMTP

合并时间：2026-07-22 04:51

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49243

执行摘要

- 一句话：为 GemmaMTP 添加 GSM8K 评测配置
- 推荐动作：可直接合并。对于关注 speculative decoding 或 Gemma 模型的开发者，值得了解此配置作为参考。

功能与动机

PR body 明确说明："To better cover against regression like https://github.com/vllm-project/vllm/pull/47953"，即防止 speculative decoding 中 embedding 宽度检查误伤 MTP Draft 的问题再次出现。

实现拆解

修改 tests/evals/gsm8k/configs/gemma-4-E4B-it-qat-mobile-ct.yaml，在 server_args 中追加 --speculative-config 参数，启用 Multi-Token Prediction（MTP）模式，指定辅助模型为 google/gemma-4-E4B-it-assistant，并设置 num_speculative_tokens=4。

关键文件：

- tests/evals/gsm8k/configs/gemma-4-E4B-it-qat-mobile-ct.yaml（模块 评测配置；类别 test；类型 test-coverage）：唯一变更文件，为 GemmaMTP 模型添加了 MTP 辅助模型的 GSM8K 评测配置，用于回归测试。

关键符号：未识别

关键源码片段

tests/evals/gsm8k/configs/gemma-4-E4B-it-qat-mobile-ct.yaml

唯一变更文件，为 GemmaMTP 模型添加了 MTP 辅助模型的 GSM8K 评测配置，用于回归测试。

```
# tests/evals/gsm8k/configs/gemma-4-E4B-it-qat-mobile-ct.yaml
# 本配置为 Gemma 4 E4B 模型（已量化移动版）添加 MTP speculative decoding 评测
model_name: "google/gemma-4-E4B-it-qat-mobile-ct"
accuracy_threshold: 0.50 # 准确率阈值
num_questions: 1319 # 评测问题数
num_fewshot: 5 # few-shot 样例数
server_args: >-
```

```
--enforce-eager
--max-model-len 4096
--speculative-config '{"method":"mtp","model":"google/gemma-4-E4B-it-assistant","num_
speculative_tokens":4}'
# 启用 MTP 推测解码，使用专门的 assistant 模型，每次推测 4 个 token
```

评论区精华

无讨论。PR 被 tlrmchlsmth 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改了评测配置，不会影响任何生产代码或核心逻辑。唯一风险是该配置若在 CI 中运行失败可能导致误报，但通过调整 accuracy_threshold 可缓解。
- 影响：影响范围极小，仅限于 CI 中的 GSM8K 评测流水线。有助于提高 GemmaMTP 模型的回归测试覆盖率。
- 风险标记：暂无

关联脉络

- PR #47953 [Bugfix][Spec Decode] Restrict embedding-width share guard to EAGLE drafts: 本 PR 旨在覆盖该 PR 引入的 regression，通过添加 GSM8K 评测配置防止类似 bug 再次出现。