

# PR #46417 完整报告

vllm-project/vllm

[ROCm][CI] Increase the max wait time for server startup

合并时间: 2026-06-23 06:46

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46417>

## 执行摘要

- 一句话: 提升 ROCm CI 服务器启动超时至 1800 秒
- 推荐动作: 简单配置调整, 无精读必要。但可作为 ROCm CI 稳定性维护的参考, 后续类似模型可参照此阈值。

## 功能与动机

PR body 指出 `mi300_8: LM Eval Large Models (8xH200-8xMI300) TG` 因服务器启动耗时过长而持续失败, 需要通过增加等待时间避免超时。

## 实现拆解

1. 修改测试配置文件: 在 `tests/evals/gsm8k/configs/Qwen3.5-35B-A3B-MXFP4-EMU-TP2.yaml` 中新增一行 `startup_max_wait_seconds: 1800`, 将默认启动等待时间从系统默认值 (通常 600 秒) 提升至 1800 秒。
2. 生效范围: 仅针对 Qwen3.5-35B-A3B-MXFP4 模型在 TP=2、moe-backend emulation 配置下的测试用例。该键可能由测试框架读取, 用于控制 server 启动阶段的超时逻辑。

关键文件:

- `tests/evals/gsm8k/configs/Qwen3.5-35B-A3B-MXFP4-EMU-TP2.yaml` (模块 测评配置; 类别 test; 类型 test-coverage) : 唯一变更文件, 新增 `startup_max_wait_seconds: 1800` 设置, 提升服务器启动超时阈值。

关键符号: 未识别

## 关键源码片段

`tests/evals/gsm8k/configs/Qwen3.5-35B-A3B-MXFP4-EMU-TP2.yaml`

唯一变更文件, 新增 `startup_max_wait_seconds: 1800` 设置, 提升服务器启动超时阈值。

```
# tests/evals/gsm8k/configs/Qwen3.5-35B-A3B-MXFP4-EMU-TP2.yaml
model_name: "amd/Qwen3.5-35B-A3B-MXFP4"
accuracy_threshold: 0.89
tolerance: 0.03
num_questions: 1319
num_fewshot: 5
# 新增: 服务器启动最大等待秒数, 解决大型模型在 ROCm 上启动慢导致的超时问题
```

```
startup_max_wait_seconds: 1800
server_args: >-
  --max-model-len 4096
  --tensor-parallel-size 2
  --moe-backend emulation
  --gpu-memory-utilization 0.35
```

## 评论区精华

无 review 评论，仅 reviewer AndreasKaratzas 批准，无争议。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险极低。仅修改 YAML 配置，不影响任何核心逻辑。如果测试框架未实现 `startup_max_wait_seconds` 字段，该变更将无效果。
- 影响：仅影响 ROCm CI 中 Qwen3.5-35B-A3B-MXFP4-EMU-TP2 测试项，预期减少因启动超时导致的假阳性失败，提高 CI 稳定性。对其他平台或模型无影响。
- 风险标记：仅 CI 配置变更

## 关联脉络

- 暂无明显关联 PR