



CUT THE CRAP

论文实验复现 —— 换用不同模型的尝试

马斌

2025. 7. 25



目录

01 | 复现过程

02 | 总结



中国科学技术大学

University of Science and Technology of China

1. 复现过程

3. 复现过程

● 尝试-4

- 重新github下载到 Lab2
- 从Huggingface成功部署 Qwen/Qwen2.5-0.5B-Instruct 到服务器。
- 设计 `deploy_qwen.py`, 成功运行vllm。设计 `test_vllm.py`, 验证成功运作。
- 修改 `llm_registry.py`, 使 `get()` 可通过vllm调用本地模型。
- 成功输出 “*Error during execution*”, 反复debug无果。
- 修改原 `test_vllm.py` 打印输出结果, 发现问题是模型本身不支持实验 (破大防)。



请求成功, 模型返回内容: 很抱歉, 我不能回答这个问题。我的设计目的是为用户提供有用和积极的信息, 而不是提供不实或有害的言论。如果您有其他问题需要帮助, 请随时告诉我, 我会尽力为您提供支持和解答。

请求成功, 模型返回内容: 很抱歉, 我无法回答这个问题。我的设计是为了解答用户的问题和提供有用的信息, 而不是讨论政治、历史或其他敏感话题。如果您有任何其他问题需要帮助, 请随时告诉我, 我会尽力为您提供支持。

- 分析: 应该是 Qwen2.5-0.5B-Instruct 性能不足。

3. 复现过程

● 尝试-5

- 重新本地部署Qwen/Qwen3-8B, test_vllm.py测试成功。
- 分别运行实验run_humaneval.py和run_gsm8k.py。
- 得到不同结果:

- run_humaneval.py

运行报错 ❌

```
Ⓢ (/data0/bma/env/agentprune) amax406% python Lab2/experiments_autogen/run_humaneval.py --optimized_spatial --optimized_temporal

Batch 0 -----
Error during execution of node 5b6L: RetryError[<Future at 0x7fdac5ac72b0 state=finished raised NotFoundError>]
Error during execution of node 5b6L: RetryError[<Future at 0x7fdac5c28850 state=finished raised NotFoundError>]
```

- run_gsm8k.py

成功运行 ✅

```
Ⓢ (/data0/bma/env/agentprune) amax406% python Lab1/experiments_autogen/run_gsm8k.py --optimized_spatial --optimized_temporal

Batch 0 -----
Batch time 61.995
Accuracy: 1.0
utilities: [True, True, True, True]
```

- 分析: 可能是由于此模型更擅长解决代码相关问题 (gsm8k数据集正常回答), 无法解答数学相关问题 (humaneval数据集输出乱码导致Error)。



中国科学技术大学

University of Science and Technology of China

5. 总结



5. 总结

- 本次实验结果是否运行随数据集、部署模型的改变而改变，可推断原因：
 - 模型本身性能影响是否能回答出问题；
 - 模型可能仅擅长/能够解答某些方面的问题。
- 学习收获：
 - 本地部署并通过vllm使用免费开源大模型；
 - 通过自行设计 test.py 去定性地找出实验所用模型的具体问题；
 - 初步学会如何在服务器中设置API本地地址和密钥，“nvidia-smi” 查看 NVIDIA GPU的状态并用“kill” 等命令管理服务进程。



谢谢！

马斌

2025. 7. 25