너도 삽퍼볼래 AI-CLI?

하다보면 실패하는 ollama-code cli 세팅

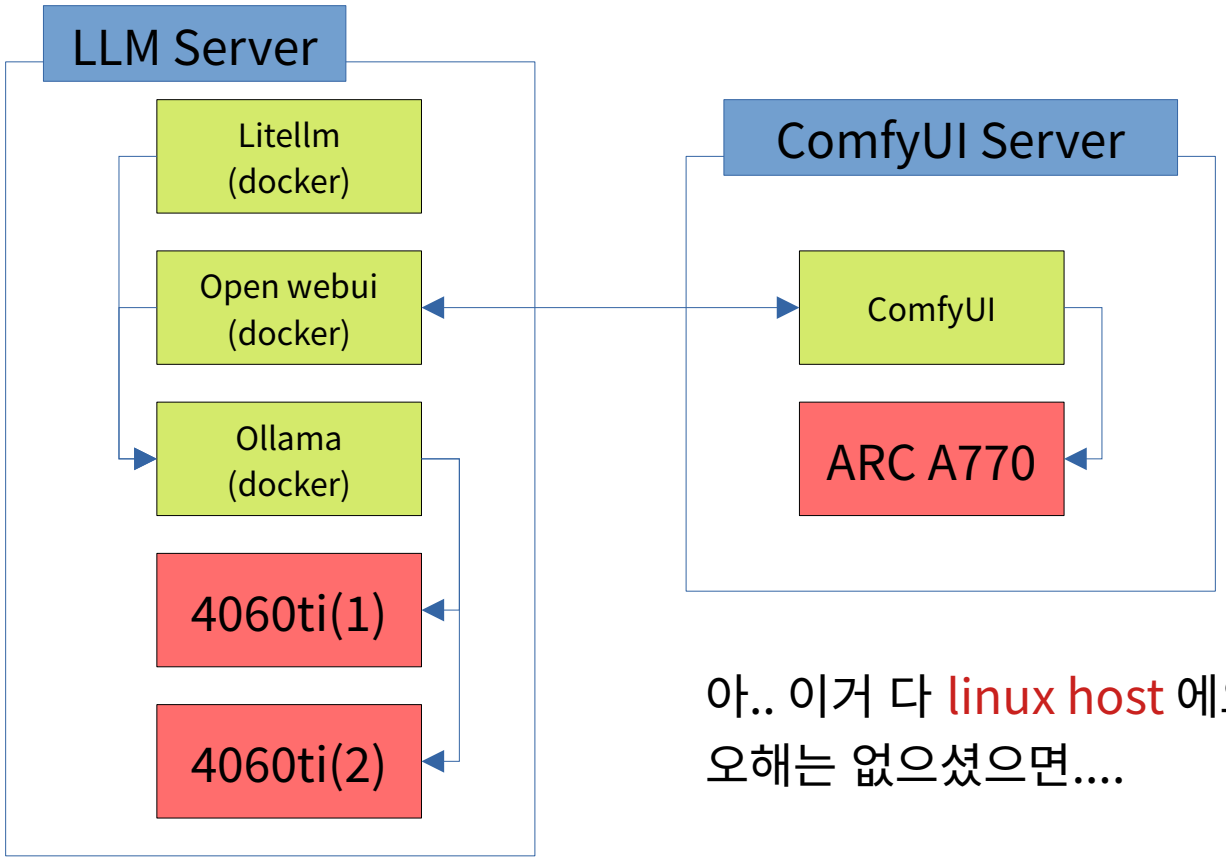
왜 삽질이 시작되었냐면...

- AI 를 쓰고싶어요
- 이야.... 신기하네...
- 그런데 비싸요...(깜짝 놀랄 가격이네요...)
- 중간정도로.... 조금 자금 투입을 해서라도 local 에서 구축하면 대충 싸게 쓸 수 있지 않을까?

하고싶은건?

- Claude code 같은 cli
- Open webui 의 web 인터페이스
- ComfyUI 사용
- Open webui 에서 ComfyUI 와 TTS 를 사용해보고 싶어요... 그럼 웬지 뭐든 할 수 있는 자신감이 생길거같아요.

최종 구조를 먼저 보여드림...(다 중고로 영끌.....T.T)



아.. 이거 다 linux host 에요...
오해는 없으셨으면....

시작은...

- Ollama + open webui 가 결합된 docker 컨테이너

- 좋아요. Update 관리도 한방이라 편해요
- 그런데 gpt-oss 가 나타나고 달라졌어요. Ollama 의 update 를 따라가지 못하네요?

그 다음에는...

- Ollama 와 open webui 를 각각의 docker 컨테이너로 운영

- 좀 귀찮치만 대충 되기는 해요
- 일단 gpt-oss 가 되는게 매우 좋았어요. 뭔가 최신의 무언가를 써보는 기분이 킹왕짱이에요...

신문물을 접하다

- Perplexity 를 SKT 1년 무료로 써보게 되었어요
- AI 가 짱 좋은거 같아요....
- 이걸로 웬지 뭔가 할 수 있을거같은 자신감이 생기기 시작했어요.

이야.. AI 가 코딩도 해?

- Claude code 의 사용

- 사실 쓰기 싫었던게 node.js 를 쓴다는게 좀 그랬어요.. 나는 개인적으로 하나로 깔끔하게 설치하고, 지워지는게 좋아요. (deb, rpm 등)
- 뭔지는 모르겠는데 짱 좋아요.
- 얘는 정말 미친거같아요.... 24시간 내내 내가 시간을 붓는 만큼 같이 놀아줘요... 이런 경험은 해본적이 없는거같아요.

AI 랑 무슨 일을 했냐면요...

- 코딩을 했어요

- 예전부터 개인적으로 운영하던 web 서비스가 있었어요.. 그냥 홈페이지죠...
- 여기에 telnet 으로 붙어서 사용하는걸 하나 추가하고 싶었어요..
- 웬지 claude code 가 잘 도와줄거 같았어요.
- 5일을 부었는데.. 정말로 결과가 나왔어요.. 잘 동작해요.
- 미친거 같아요. AI 정말 짱 좋은거 같아요.
- 뭐가 되었던 local 환경에서 cli 를 쓰고 싶어졌어요. 내가 쓰는 환경이면 이게 돈을 많이 아껴줄거 같은 느낌적인 느낌이 들었어요.

새로운 도구를 알게되다

- Stable diffusion to ComfyUI

- Ollama + open webui 를 세팅할때 stable diffusion 을 써본적이 있어요.
- Qwen-image 라는게 있다더라구요...
- 이걸 쓸려면 ComfyUI 라는걸 써야한대요... 생긴게 복잡해서 쓰기 싫었는데...
- 여튼... 귀찮음을 무릎쓰고 ComfyUI 를 써보니... 생각보다는 안복잡해요. 다른사람이 만든 workflow 를 복사해서 쓰면 의외로 stable diffusion 수준으로 할일이 많지 않아요.
- 게다가 ComfyUI 는 gguf 를 사용하면 결과물이 나오는 시간도 짧아지네요? 이야.. 이 녀석도 짱 좋은거 같아요.. 우와! 머쩡!

이런것도 돼?

- Open webui 와 ComfyUI 의 연동

- Ollama 와 Open webui 를 분리하면서 open webui 의 update 를 좀 자주 하게 되었어요.
- ComfyUI 를 세팅하면서 커뮤니티에 스샷을 올리니 사람들이 LLM 과 ComfyUI 를 연동하는걸 알려줬어요. 우와.. 뭔가 될거같아요. Open webui 에서 제공하는 openai api 로 될거같은거에요..
- 그래서 Open webui 의 기능을 찾다보니 open webui 에서도 ComfyUI 연동이 되는게 보이는거에요... 우와.... 웬지 이 정도면 성능은 좀 구려도 상용 AI 에서 되는걸 할 수 있을거 같은거에요.....

이제 뭔가 좀 보이는거 같아요

- Continue cli

- Claude code 의 대안을 찾다보니 Continue cli 라는게 있더라구요..
- 우와.... model 이 문제지, model 만 잘 받쳐주면 뭔가 claude code 처럼 같이 일을 해줄거같아요.
- 준내 삽을 퍼보기로 해요... Perplexity 가 도와주면 세팅도 잘 될거같아요. 좋아요. 다 잘 될거같아요. 그런데 이때는 몰랐어요. 이게 거대한 삽질의 시작이 될 줄은...(좌하하하)

첫번째 삽질

- Continue cli 에 local llm 을 연결해보자

- 된다는 말은 많은데.. 뭔 자료가 그리 없는지...
- 결국 되기는 했어요.
- Provider 에 ollama 와 openai 가 별도로 동작하는지도 모르고 닥치는대로 세팅하다보니.. 삽을 좀 펐어요
- 덕분에 ollama 와 openai 가 어떤 차이가 있는지는 아주 잘 깨닿게 되었어요.
- 그런데 이것조차도 그냥 시작이었어요... 아시죠? 뭔가 되기 시작하면 계속 삽을 푸게되는거..

되는...건가?

- Open webui + continue(vscode)

- 하다보니 알았어요. Continue 랑 Continue cli 가 같은 설정파일을 사용할 수도 있다는걸...
- Continue cli 는 안되는데 얘는 되네요? 그럼 내가 잘못 설정한건 아닌거같다...라는 느낌적인 느낌이 들었어요

이때 멈췄어야 했다

- Ollama + continue cli

- 잘 안됨.
- 될거같은데 안됨
- 왜 되어야 하는데 안됨?
- 왜 vscode 에서는 되는데 안되냐고!

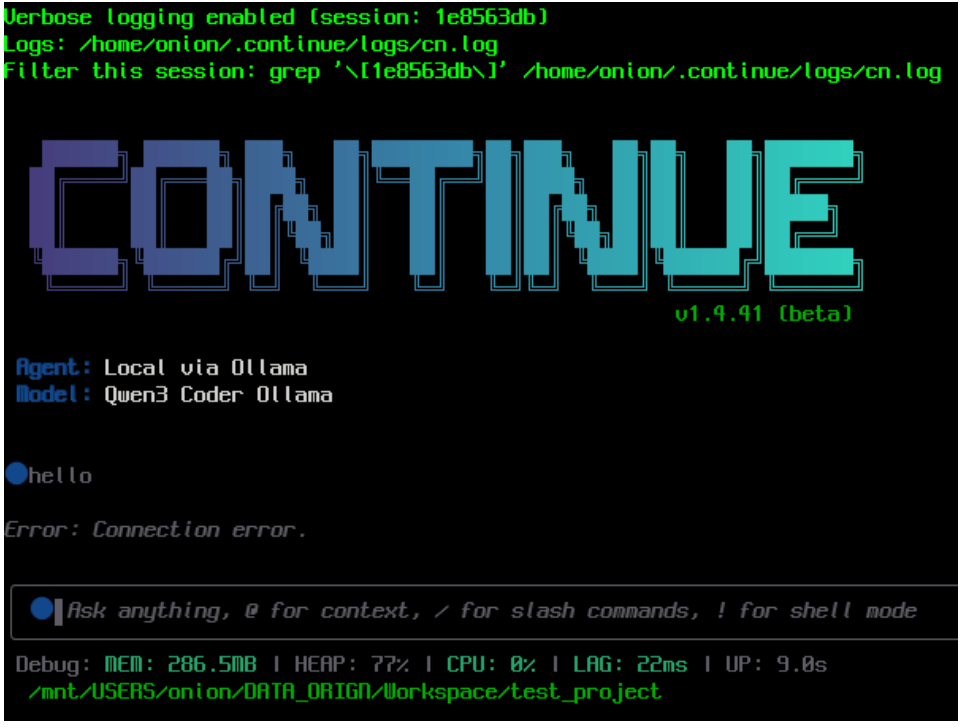
이때 멈췄어야 했다

- Open webui + continue cli

- 잘 안됨... Provider 가 openai 인데 왜 안됨?
- 될거같은데 안됨... 왜 auth 를 했는데도 안됨? Curl 은 되는데 왜 안됨?
- 아놔.... 왜 되어야 하는데 안됨?
- 왜 너도 vscode 에서는 되는데 안되냐고!

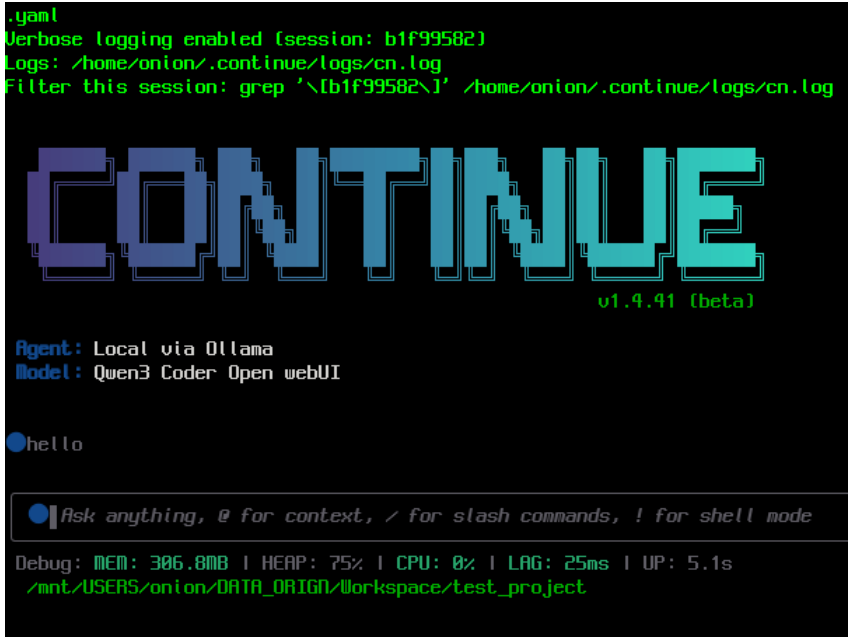
이 조합이 똥인가봐....(1)

- Continue cli + ollama : 응답은 보내지만 Error 발생



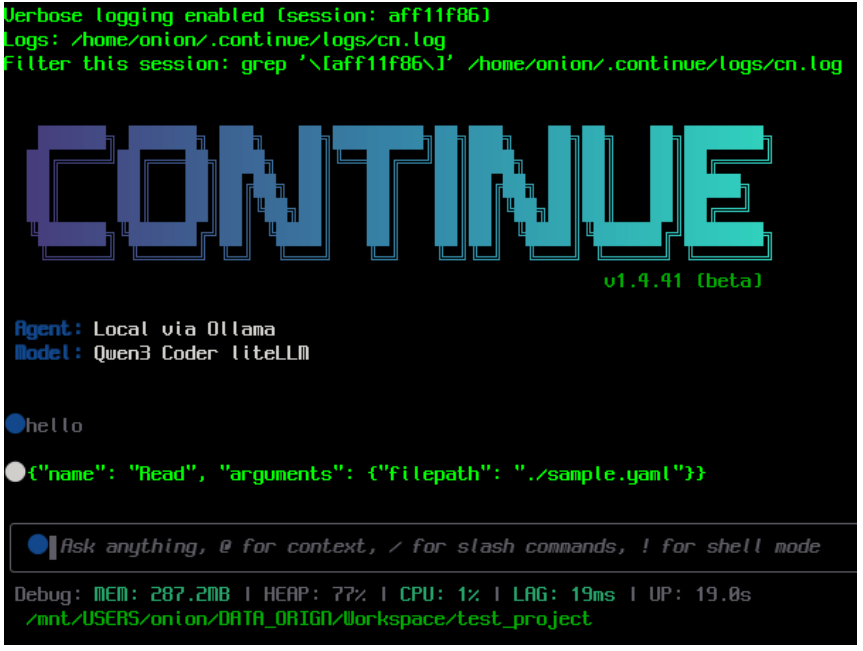
이 조합이 똥인가봐....(2)

- Continue cli + open webui : 응답 해석못함



그럼 뭘 해볼까?

- Contiue cli + litellm : 답변이 그대로 json 으로 나옴...



대안이라면.. qwen code cli?

- Qwen cli + litellm : 이야.. 너도 json 이네...

qwen code cli.. 너도 글렀...

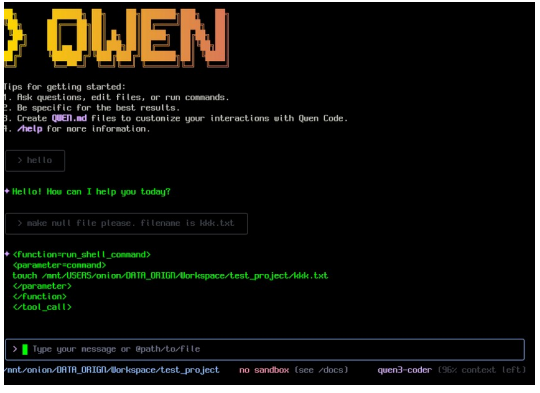
- Qwen cli + open webUI : 이야.... 너네 둘은 아예 안맞는구나?

사실 이건 됨...(liteLLM 과의 조합)

- LiteLLM 에서 중간에 tool 관련된 부분을 전부 제거해버리면 continue cli 에서 대화는 가능. 그런데 “대화만” 가능. 이게 뭔 얘기냐면 continue cli 에서 local 의 파일을 접근할 수가 없음...(이럼 뭐러 cli 를 쓰냐고....)
- Qwen code cli 도 마찬가지. 예를들어 “make null file please. filename is kkk.txt” 이라고 보내면...

<function=run_shell_command>
<parameter=command>
touch /mnt/USERS/onion/DATA_ORIGN/Workspace/test_project/kkk.txt
</parameter>
</function>
</tool_call>

이렇게 xml 같은게 오심... 그냥 qwen code cli 화면에 보임.



결국 둘 다 굳이 cli 로 쓸 필요가 없다는 의미..... 아시발쿰...

그럼 대체 뭐가 cli 랑 되는거야?

- Continue cli
  - 아마도 lm studio 기반이라면...
  - 그럼 대체 ollama 는 왜 provider 에 적어둔거야?

- Qwen cli
  - 상용 AI
  - Lm studio 로 구동되는 model 들

와... 이놈들 ollama 랑만 안되네.... 짜증이....

다른건 안돼?

- LM studio + liteLM

- Cli 계열은 됨.
- 그런데 ComfyUI 연동은 어떻게 해요?
- ComfyUI 에서 LLM 을 부르는건 될거같음
- 그런데.... chatGPT 를 쓰듯이 쓰는건?...

그럼 이건 안돼?

- LM Studio + liteLLM + open webUI

- 역시 CLI 는 될거같음
- 그런데 open webUI 가 LM Studio 를 쓸때 openai api 방식으로 연동해야 함
- 그럼 open webUI 에서 model 을 바꿔가며 쓰는게 안됨.
- 그럼 chatGPT 방식처럼 쓰는데는 좀 한계가 있음

같이 쓰면 안돼?

- LM Studio 랑 ollama 를 같이 쓰면 안돼?

- 될리가.. 2개는 “역할이” 같은거에요..
- 너님 GPU 가 남으세요? 남으면 2개 써도 되죠...
- 하지만 저는 그만큼의 리소스는 없어요...

고민중...

- 어차피 코딩용만 쓴다면... LM Studio 를 메인으로 써야 하는거 아니야?
- 그런데 LM Studio 는 별도의 UI 가 있어서... open webUI 로 컨트롤이 안되는데?
- CLI 는 포기하고 깔끔(?)하게 vscode-continue 만 쓸까? 반응은 살짝 느려도 얘는 잘 되던데...

남은 숙제도 있다

- 그래.. cli 는 포기한다고 치고...
- 일단 open webUI 와 comfyUI 연동은 성공
- Open webUI 에서 TTS 를 지원하네... 여기까지는 좀 써봐야 할거같은데....

그리고 또 남은게 있네..

- 지금까지 삽질하면서 나온 스크립트 및 세팅 자료들.....
- 몇가지 팁들

와.... 언제 정리하지.. T.T

당연한 주의사항

- 본문에는 ComfyUI 에 대해서는 자세히 적지 않았습니다만.. ComfyUI 자체가 모델을 메인메모리에 올리는 작업부터 시작됩니다. 당연 gguf, safetensor 모델들도 LLM 처럼 덩치는 큽니다. 잘못하면 swap 을 사용하게 되니까 메인 메모리는 넉넉..하게 잡아주세요.
- Nvidia VGA card 를 2개 꼽아서 구성하는 경우, 대부분의 메인보드는 16x 의 PCI-Express 슬롯을 1개만 제공하게 됩니다. 다른 슬롯은 그만큼 속도가 안나오는 경우가 “요즘은” 대부분이죠. VGA 를 SLI 로 구성할 필요가 없으니까요. 이런경우 “당연하게도” 2개의 VGA card 중에 낮은 속도로 꼽힌 card 의 동작속도를 기준으로 동작됩니다.
- 아시는 분은 아시겠지만, intel ARC vga 는 성장형입니다. 가격은 중고 기준으로 nvidia 의 반값 수준이라서, 해당되는 카드를 지원하는 ComfyUI 에서는 16G 의 넉넉한 VRAM 을 나름의 저렴한 가격으로 써볼 수 있다는 장점이 있습니다만, linux host 에서는 firmware update 를 할 수가 없습니다(이론상은 된다는데... 일단 저는 실패했습니다.) 쓰실 분들은 이점 고려해서 꼭 windows 에서 firmware 를 update 하고 나서 linux 에서 사용하는걸 추천합니다.

참고문서

- intel arc gpu 로 comfyui 준비해보기
  - http://pge.kr/en/Board/PostDetail/e27c679b188151351e27cfd55ff46e5f/285

- ollama / open webUI / litellm with continue cli / qwen code cli / claude code cli 를 연동해보기 위한 과정에서 얻은 tip
  - https://workspace.onionmixer.net/wiki/CliCode_With_Ollama