Hey Onfinace team!

Sanyam this side, Let me take you through code here and tell you my approach and also what database you can use, what pre-trained LLM can do this task with good prompting and new techniques without using GPT4.

So task was pretty simple : you want to automate SQL / db cleaning via triggers and automating whole stuff, so for this you need LLM to write the exact code.(nice experimentation by Onfinace team)

Now here comes my 2 minds : Researcher + Computer Scientist

1. I tried multiple opensource llms even gpt3.5 (but langchain, sql agent only supported OpenAI , and also didn't noticed my credits were finished, key was exhausted) to handle this shifted to finetuning and exploring the realm of opensources. Exp 1

2. I first drafted static code which i wrote myself (using Langchain + SQL Agent), go to know for some reason only OpenAI seems to work, thought of one thing from here which can help OnFinace as you said the bonus thing (create synthetic datasets by using chatgpt4 as we do in Research) (Opensource dataset which you can use on finetuning LLMS will depend on user's query then)

2.1 Breaking down somethings, lets say User query is dynamic, one thing can be use normal NLP techniques to find keywords, from keywords, use Lang-graph , a new tool in town which helps you integrating many knowledge bases together, making your LLM way too powerful.

2.2 Here i tried experimenting a new framework (React - Reasoning + Action) Exp 2 , find out with chatgpt 4 api it will work amazingly to enhance User Experience, but fail to implement thanks to huggingface (server got down, also some LLMs on hugging face, were not getting loaded because of some server issue 500)

2.3 Here comes to rescue our beloved top rankers LLMs beating OpenAI or one can say close to OpenAi which can save tremendous API costs (Mistral, Mixtral, LLAMA, WizardCoder) , one can use RAG to give them multiple SQL queries, or info of db via RAG , or a new thing i observed, convert your db into json format or string format, you can gave that too in prompt, and it will give you as exact result it can.

3. You can check out the internet, Mistral is on-par with Chatgpt/GPT-4, so here I used same Exp3 final

4. Bonus datasets for automating code : MAIN LINK

4.1 Its still useless tho, many big LLMs are usually trained on public datasets only, so most of datasets available for code generation LLM have already seen it.

4.2 That’s why nowadays to specific usecases, one needs to create Synthetic data oneself.

4.3 What I will advise being in ML, DL, LLM space for a while now, Patience will play a role, but for users , we need to crowdsource, open forms, or give some monetary benefits to get data, or lets say you gave me problem for automating this, make a new dataset , containing your problem and my answer or LLMs answer, debug the code obviously via running it, start collecting data, and then we can make a new whole LLM of ours (automater / whatever name suits) which i have seen multiple user queries, which i have seen all the syntax, all the tools docs, and is able to generate code. (Lang-graph will be suitable to handle multiple knowledge bases)

5. In the end for now, you can use Mistral, here I have used a small portion of it in Exp3 final link, but we can use Large Mistralm which is on-par with GPT4 , gave it access to info via Lang-graph and RAG.

Looking towards the round 2/ interview