| LADP #2 | | |
| --- | --- | --- |
|  | Question | Mentor’s Answer |
| 1 | Do you have any suggestions on tools to use to analyse read transaction database for the LLM to give suggestions? | Using an LLM to read a database involves structured data querying/extraction. I have pasted a few links below that show the tools that can be used for this.  1. Using agents <https://developer.nvidia.com/blog/build-an-llm-powered-data-agent-for-data-analysis/>  2. Look at “over structured data” section <https://docs.llamaindex.ai/en/stable/use_cases/q_and_a/>  3. <https://python.langchain.com/docs/expression_language/cookbook/sql_db> |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |

\* You may find LADP #1 question bank below for additional learning.

| LADP #1 | | |
| --- | --- | --- |
|  | Question | Mentor’s Answer |
| 1 | Based on my knowledge that ChatGPT did use RLHF (reinforcement learning from human feedback) to train and guide the model to have better feedback.  May I know how this is implemented? If I want to try other methods such as monte carlo tree search to replace RLHF, is this possible and how would I achieve it? | Here are two articles to help you understand how RLHF works:  How ChatGPT actually works <https://www.assemblyai.com/blog/how-chatgpt-actually-works/>  Illustrating Reinforcement Learning from Human Feedback (RLHF)  <https://huggingface.co/blog/rlhf>  The goal of RLHF is to align the model's outputs more closely with human values and preferences. Replacing RLHF with MCTS would be a significant shift in methodology. MCTS is a search algorithm often used in game playing contexts where the outcome is uncertain and requires planning several moves ahead, such as in Go or chess. It simulates many possible moves and their consequences to make decisions. Language generation is less about exploring a discrete and calculable set of future states and more about generating coherent and contextually appropriate text sequences. MCTS may not provide the same benefits in aligning model behavior with human preferences, which is one of the key advantages of RLHF.  To implement MCTS or any other method in place of RLHF, you will need a setup where:  Problem Formulation: You must have a clear problem where MCTS can be applied. MCTS is not typically used for the type of sequence generation tasks language models are designed for.  State Space Definition: Define a clear state space that MCTS can navigate. For language models, this could be challenging due to the vastness and ambiguity of natural language.  Simulation: MCTS requires the ability to simulate outcomes based on actions. In the context of language models, this would involve predicting the consequences of generating different sequences of text.  Evaluation Function: You need a way to evaluate the outcomes of different sequences to guide the search. This is where MCTS could possibly use a form of reward modeling.  Integration: You'd need to integrate MCTS into the training loop of the model, which may require significant changes to the architecture and training process. |
| 2 | RAG (retrieval-augmented generation) enables LLM to provide facts and less bias of response, while finetuning can shape or modify LLM’s response behavior.  What’s the cost of resource and time response of having these 2 methods apply together in a LLM? If we implement a vector database, will it reduce the cost or shorten the response time? | Finetuning is an expensive and time-consuming process – beyond the resources of most ordinary companies. This involves obtaining a large amount of data, consuming GPU resources, not to mention technical expertise needed to perform finetuning.  RAG is a way to augment the LLM with updated facts and to reduce hallucinations, without having to finetune the LLM. If you are doing RAG, you do not need to do finetuning for the purpose of augmenting the LLM with updated information.  Vector databases are very fast – a typical vector retrieval is much faster than an LLM call. Furthermore, many vector databases are free and open-source, meaning you only pay for the technical expertise of implementing it for your RAG system. |
| 3 | Instead of using pure text to show LLM’s reasoning step for solution, is there any reliable library or framework to visualize it? | Please clarify what you mean by visualizing the reasoning steps?  There is a tool by LangChain called Langsmith which can be used as an observability tool for agents. However, Langsmith is still in beta and you have to sign up to be on a waitlist, so it is still not publicly available. |
| 4 | Are there any methods to maintain the quality of a vector database when trying to insert or update with a new document? | The “quality” of a vector database is not impacted when performing the actions of inserting or updating the data itself. If your new data is of a worse quality, then the downstream retrieval will be negatively affected. |