

User:

Hello

Assistant:

Hello! Thank you for taking the time to speak with me today about the Entry-Level RAG AI Engineer role. I'd like to ask you a few questions to learn more about your experience and skills. Shall we begin?

User:

Yes, sure.

Assistant:

Great. Could you tell me about your experience with retrieval-augmented generation (RAG) pipelines? Have you worked on creating or optimizing any RAG systems?

User:

Yes, I worked on creating two RAG systems. One of them was trying to fit a large PDF into the context of an LLM. I found that the prices were skyrocketing because I had to feed the PDF every time I had to query the LLM. So to stop this I tokenized the PDF and put it into a vector store and then used cosine similarity to search for relevant information from the vector store as related to the question and feed it to the context of the LLM so that I did not have to upload the PDF every time thus saving me a lot of money. And then on this second project I was developing an interview screening system for which I had to measure the accuracy of the answers of a candidate. Now since the LLMs were trained on data till 2022 it is not updated with the most up-to-date data. So for this I had to decompose each answer into search strings and search them individually on the web and web script the first two websites. Thus I got a large number of documents to tokenize and store in a vector store and now when the LLM goes for verification of each answer it has it can take in data from the vector store similar to the with cosine similarity as I mentioned earlier and verify the accuracy of the answer with up-to-date information. So these are the two projects I worked on with RAG.

Assistant:

That's impressive work with RAG systems. You've clearly tackled some complex challenges. Can you tell me about your experience with different large language models? Have you worked with or compared various LLMs like OpenAI's models, Llama, or Claude?

User:

Yes, sure. I have experience working with both OpenAI's chat GPT, LAMA, GPT and also Google's Gemini. So each of the large language model has their own specific strengths and weaknesses. For example, GPT 3.5 Sonnet is widely regarded as the most intelligent large language model to date. However, it has a small context window. So if I require tasks that don't require me to fit in a huge amount of information in the context window, I generally prefer GPT 3.5 Sonnet. Whereas Google's

Gemini, especially Gemini 1.5 Pro, has a context window of 2 million. So if I require to fit in huge amounts of information in the context window, I would generally go with Google's Gemini, such as going through an entire film or analyzing an entire film. Something like this can be done through Gemini. Whereas chat GPT 4.0 is the latest model. It's a version of OpenAI. And GPT 4.0 Mini is a cut-down version of the same. So if I had to keep costs in mind, I would go for chat GPT 4.0 Mini. Because it is a cut-down model, but it is intelligent enough with low API costs. And it is faster. Now, Facebook's LAMA is an open source model. So if I wanted to run something locally or fine-tune it to my heart's content, without... paying any money, I would go for LAMA. So each of the models has its own benefits and downsides.

Assistant:

Thank you for that comprehensive overview. You've clearly worked with a variety of LLMs. Now, could you describe your experience with AWS services, particularly in relation to AI or machine learning projects?

User:

Certainly, I'd be happy to share my experience with AWS services, particularly in relation to AI and machine learning projects. Over the years, I've had the opportunity to work extensively with AWS suite of AI and ML services across various projects. One of the primary tools I've leveraged is Amazon SageMaker, which has been instrumental in streamlining the end-to-end machine learning workflow for many of our clients. For instance, in a recent project for a large e-commerce company, we used SageMaker to build and deploy a recommendation engine. We utilized SageMaker's built-in algorithms for collaborative filtering, which significantly reduces the time to market compared to developing a custom solution from scratch. The scalability of SageMaker also allows us to handle the client's large data set effectively. I've also had great success with Amazon Comprehend for natural language processing tasks. In a project for a social media analytics firm, we used Comprehend to perform sentiment analysis on millions of customer comments. The service is very easy to use. The service is able to extract key phrases and entities, which was particularly valuable in providing actionable insights to our clients. Amazon Forecast has been invaluable for time series prediction projects. For a supply chain management company, we implemented demand forecasting using the service, which improved their inventory management significantly. Lastly, I've integrated Amazon Lex, which is a multi-task service that allows us to use Amazon Lex in several projects to create conversational interfaces. One notable example was a customer service chatbot for a telecommunication company, which helped reduce their call center load. These experiences have given me a deep appreciation for the power and flexibility of AWS's AI and ML services. They've allowed us to implement sophisticated AI solutions efficiently and at scale. Thank you.

Assistant:

Thank you for sharing your experience. Last question: How do you approach model efficiency issues, particularly regarding speed, performance, and cost optimization?

User:

Yes, so using the right LLM for the right task is a huge undertaking. So for example in the accuracy verify I mentioned a while ago in this interview we tried several different models including Google's Gemini 1.5 Flash Google's Gemini 1.5 Flash, Cloth Sony 3.5, OpenAI's ChatGPT 4.0 and finally stuck to OpenAI's ChatGPT 4.0 Mini. This is because GPT 4.0 Mini is faster than the other services that I mentioned but it is a pruned down version of GPT 4.0. However for the requirement of just accuracy verification it serves the purpose and it is also very cost-effective. So these are the considerations that would go into this interview. Thank you for watching.

Assistant:

Thank you for your time. We appreciate you sharing your experiences and insights with us today. You'll hear back from us soon regarding the next steps in the interview process.