Building AI based Chatbots

Generative AI and building AI based Chatbots that can talk to/about **your documents**, products and services is all the rage right now and this ability is also being added to multiple products you currently use.

The AI chatbots such as GPT4 can only respond to generally available knowledge/information up to September 2021, 2 years ago! These chatbots are referred to as LLM's (Large Language Models)

What if you need more detailed knowledge such as spinning up services on our soon to be launched product "Apsara Stack"? To bypass the above limitation, you can give the chatbot more information or **context**, in this case we give it the publicly available Apsara Stack Operators manual, and then ask questions based on the manual.

The official terminology around the above process is Retrieval augmented generation (RAG) and is a natural language processing (NLP) technique that combines the strengths of both retrieval- and generative-based artificial intelligence (AI) models.

The above services will add significant value to your projects and customer, but it cannot always be public, we need to ensure that your customers can build such services within their own control.

This is where Private LLM's come in, during the Hackathon we will focus on building such private services on Alibaba Cloud. Alibaba Cloud already has some of these services as a PaaS. PAI EAS and AnalyticsDB for PostgreSQL.

BCX will provide instructions and guidance on the above.

Key to building such services is making use of a LLM (Large Language Model) such as ChatGPT or Bard or Claude, and many others. ChatGPT4 is still King, and we will be using this for our initial builds. We must stress that all the information used in this process are publicly available information or datasets.

Also critical to building the above is making use of Vector Databases, unlike traditional databases vector databases focus on similarity, for example, cat, kitten, mammal, furry, 4 legs …. Are all close to each other.

Vector databases are still developing, and companies are falling over themselves to provide these to customers.

During the Hackathon we will by using Alibaba Cloud AnalyticsDB for PostgreSQL with pgvector. This is a ready-made Vector database service offered by Alibaba Cloud.

A Managed database and always available service will be critical once these services become more broadly adopted and start moving into production.

Understanding how chatbots are built. We will be using Langchain libraries throughout the Hackathon.

This is excellent training and will give you a head start.

https://learn.deeplearning.ai/langchain-chat-with-your-data/ - create an account and learn with the folk who developed Langchain.

Do spend a few Rand on an OpenAI account, you can create an account here and get your API.

https://openai.com/

Lancgchain is where it's all happening

https://python.langchain.com/docs/get_started/introduction

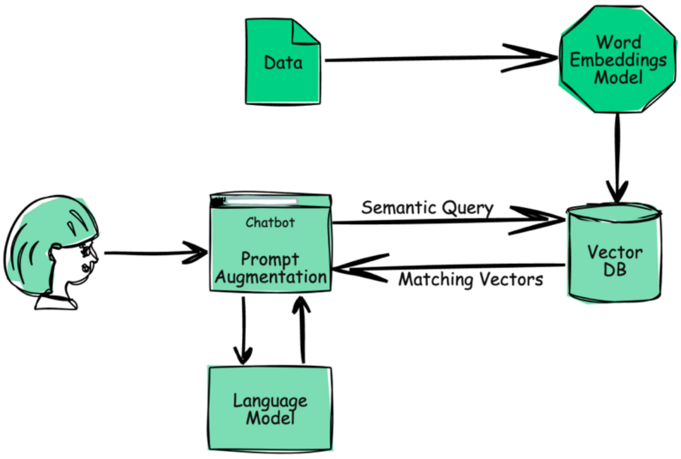
Libraries for both Python and JavaScript.

LLM's can answer questions on what it already knows, data that was added when the model was built. If you have your own data such as a PDF with instructions on how to build on a specific cloud provider, the LLM won't know all of this. In this case you'll want to tell the LLM about the contents of the PDF and then ask the LLM the question you have. This is called adding context to your question. There is one problem in doing this, the LLM's have a size limit of how much context you can provide. To overcome this problem you need to be selective with the context you provide to the LLM and make it relevant to the question. It is this function that will be the entire focus of our hackathon, doing the above effectively and with nuance, getting this wrong and your Aunty will appear drunk, or hallucinate.

Here is the process that one can use to overcome the problem of too much context.

1. Get all your documentation you believe will add knowledge to your solution. In our example all the operational PDF's for Apsara Stack.
2. You then break these documents into Chunks with some overlap between chunks.
3. You then embed these chunks into vectors using an embedding service.
4. You then store these embeddings in a Vector Database.
5. Once all this data is in the Database you can search for all chunks that are relevant to your question.

6. Once you have the relevant chunks you select the top 5 and submit these as context to the LLM to help answer your questions.



Chatbot workflow

The tools and repositories that we will use to build these services are:

1. OpenAI (Javascript and Python)
2. LangChain (Javascript and Python)
3. ChromaDB (Javascript and Python)
4. Alibaba Cloud AnalyticsDB for PostgreSQL
5. Alibaba Cloud PAI AES
6. Streamlit (for presentation, Python only, Javascript users won't need this part)

More details for follow.

We will provide Alibaba Cloud account credentials and the participants will log in and start building the core services.

1. Setting up a VPC
2. Creating subnets
3. Creating security groups
4. Launching instances
5. Accessing the instances

The process

1. Gather the data source you wish to talk to
2. Split that data into chunks (The LLM models is limited in terms of how much information you can give it)
3. Once the data is split into chunks you need to be able to search these chunks on similarity to the question) This is done by the vector Database.
4. Convert each chunk into a Vector, this is complicated and requires serious compute so we hand this part over to the Cloud and use a specific model for the conversion. PAI – EAS and the FAISS mode
5. We then store this retrieved vector and it's metadata, metadata = the source chunk and any details such as page number.
   We will use PostgreSQL with pgvector plugin, as our Vector Database
6. When we do queries we retrieve the most relevant vector relating to the question,
7. Once we have a couple of chunks relating to the questions, we submit these chunks as "context" to the LLM
8. The LLM we will use s Llama2 and will be hosted by Alibaba Cloud service called PAI EAS

Please follow these blogs on building generative AI on Alibaba Cloud

Chat to the writer of the blog on the Sonke platform – his name is Farruh

https://www.alibabacloud.com/blog/600229?spm=a2c65.11461447.0.0.29d54014NJBFg6

https://www.alibabacloud.com/blog/600283

https://github.com/k-farruh/llm_solution

The IP Address to use for any Security Groups is 196.10.24.152