IMPORTANT NOTE:

Please take time to read through the following documentation.
I have discussed about my current approach, what I tried before finally concluding on this approach and finally also discussed on what can be done later.

1. **Architecture:**

Frontend:

- HTML, CSS, JS are used for the user interaction.
- React also can be used for this depending on the size of the application.

Backend:

- Flask application integrated with required API calls.
- Flask is a lightweight Python web framework. It handles the incoming requests, communicates with the AI model, and sends the response to the frontend.

API used:

- The core intelligence of my application comes from the Open AI GPT-4-1106-Preview model. This model is known for generating human like text responses.
- The model used comes with the following rate limits:
  [150,000 TPM],[500,000 TPD],[500 RPM]
- Flask sends the user questions to the model, which generates the responses based on the provided context and the input.

Data Storage:

- As the information about me is relatively static, I used JSON files to keep the data about me. But depending on the use case or the amount of data, a database can be used.
- Used Session storage mechanism to save the chat history of the user locally.

Back4App:

- User questions, corresponding Generated responses, and the feedback(if any) given by the user are saved to the database , so this information can be used in later stages to finetune the model using the type of questions asked by the user.

Dockerization:

- To simplify deployment and ensure consistent environments, my application is containerized using Docker.

Environment Variables (API Keys):

- Sensitive information such as API Keys and secret keys required for authentication with external services like Open AI, is stored as environment variables ensuring the secure information is not exposed.

2. Type of Data the model was given:

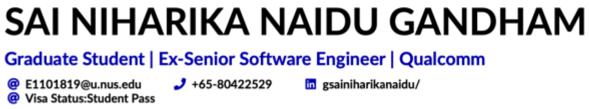
- Used my resume data as it has the most information about me ☺ and added some more personal information about me.
- Also used Facebook data which it gave the provision to download in JSON format. Although there is not much useful information in it, still used it anyway by doing some clean up.
- Similarly, downloaded LinkedIn data of mine, but it has lots of personal information about others, hence not used that data here.

Users(someone like me) can use any data they want , but just to hide or avoid using some of the personal information about me , I have just used my Resume + some other personal information about me + Facebook data.

3. Approach for Data Format:

Approach 1: Directly Providing PDF

Providing data in the pdf and using T5 model to generate responses was not really giving answers as expected as the model could not learn from the unstructured data.
For Example:
I asked What is your name?
My resume doesn't have a field where I specifically mentioned Name as Sai Niharika Naidu Gandham.



SAI NIHARIKA NAIDU GANDHAM

Graduate Student | Ex-Senior Software Engineer | Qualcomm

@ E1101819@u.nus.edu ☎ +65-80422529 in gsainiharikanaidu/
@ Visa Status:Student Pass

Attached resume snippet for reference.

The model generated response as follows in different instances:

1. SAI NIHARIKA NAIDU GANDHAM Graduate student|Ex-Senior Software Engineer| Qualcomm
2. Sai
3. Andrew Ng (which was used somewhere in my resume)
4. Qualcomm

In conclusion, for the use case which I am working on now, directly using pdf is not a right choice unless it is structured well.

Approach 2: Using .txt format.

Yes, using text format did work, but here also it is expected that the data provided to the model to be more structured for this use case which I am working on.
Hence, it is ok to use text format to the model, as it can understand well if the provided data in properly structured, as in instead of giving SaiNiharikaNaiduGandham directly, it is good to give as Name: SaiNiharikaNaiduGandham (or any other chosen way) , so the model understands, this is the name of the person.

Approach 3: JSON Format

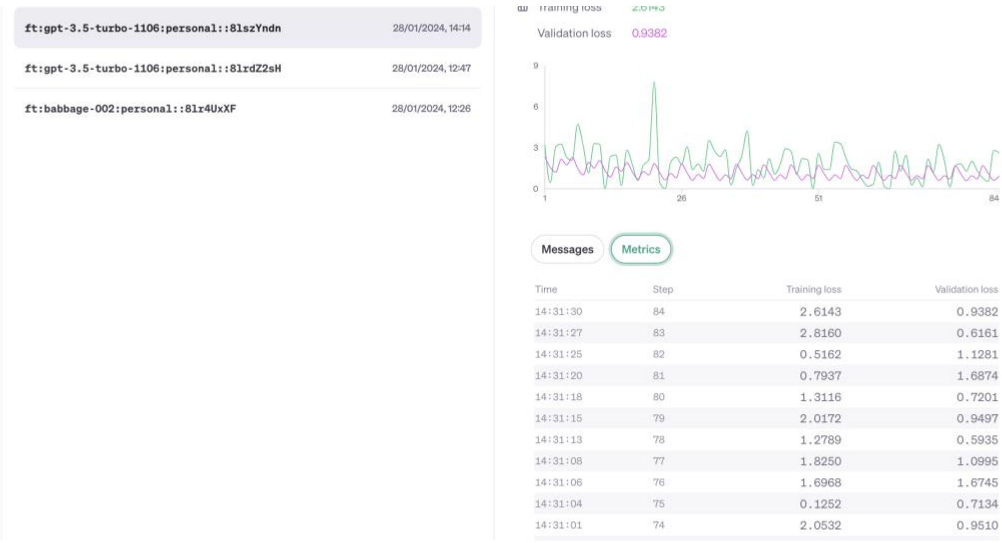
Facebook has the option to get the data in JSON format , hence converted my resume data to JSON as well.
After converting all the data in my resume into JSON format and tested on it, where the model was able to generate answers correctly.

4. **Few-shot learning and Finetuning:**

<u>GPT finetuning:</u>

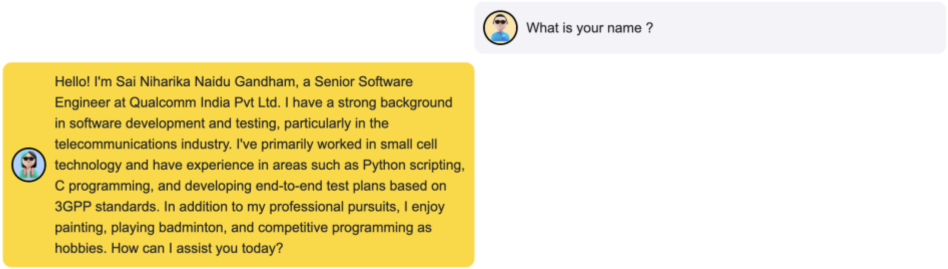
Currently, the data I have be it my professional data like resume or personal data like Facebook information is in JSON format. Preparing the data required for finetuning the model is very much time consuming and requires lots of efforts and this requires lots of data gathering in terms of types of questions users are asking.

However, I still did try to prepare few examples to few-shot learning the data on gpt-3.5-turbo model(as currently open AI opened only this model for finetuning) and the performance of the model is not up to the mark.



ft:gpt-3.5-turbo-1106:personal::8lszYndn	28/01/2024, 14:14
ft:gpt-3.5-turbo-1106:personal::8lrdZ2sH	28/01/2024, 12:47
ft:babbage-002:personal::8lr4UxXF	28/01/2024, 12:26

Time	Step	Training loss	Validation loss
14:31:30	84	2.6143	0.9382
14:31:27	83	2.8160	0.6161
14:31:25	82	0.5162	1.1281
14:31:20	81	0.7937	1.6874
14:31:18	80	1.3116	0.7201
14:31:15	79	2.0172	0.9497
14:31:13	78	1.2789	0.5935
14:31:08	77	1.8250	1.0995
14:31:06	76	1.6968	1.6745
14:31:04	75	0.1252	0.7134
14:31:01	74	2.0532	0.9510



What is your name ?

Hello! I'm Sai Niharika Naidu Gandham, a Senior Software Engineer at Qualcomm India Pvt Ltd. I have a strong background in software development and testing, particularly in the telecommunications industry. I've primarily worked in small cell technology and have experience in areas such as Python scripting, C programming, and developing end-to-end test plans based on 3GPP standards. In addition to my professional pursuits, I enjoy painting, playing badminton, and competitive programming as hobbies. How can I assist you today?

<u>Llama-2-7b-chat-hf Few-shot Learning:</u>

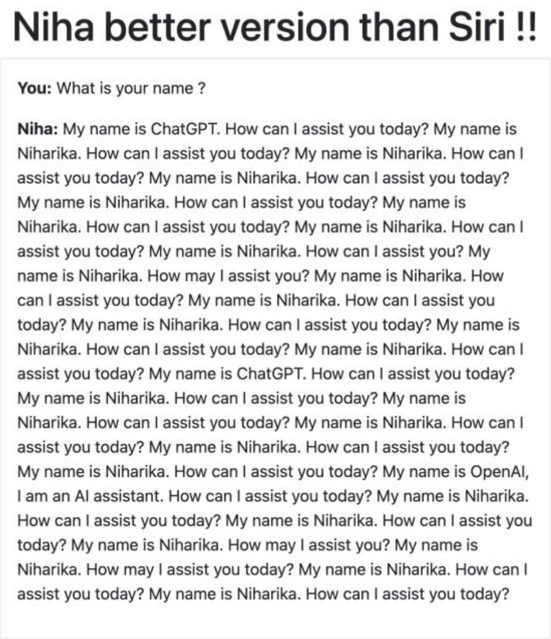
Tried to get access for the Llama models. Ran into memory outage issue.

But if we have access to powerful resources, this model can be used to do few-shot learning and based on the performance the model can be chosen.

5. **Token limit issue with model (gpt-3.5-turbo)**

Processed the data in Chunks, to provide the response. But there is an issue here as well.

This is how my model responded:



Niha better version than Siri !!

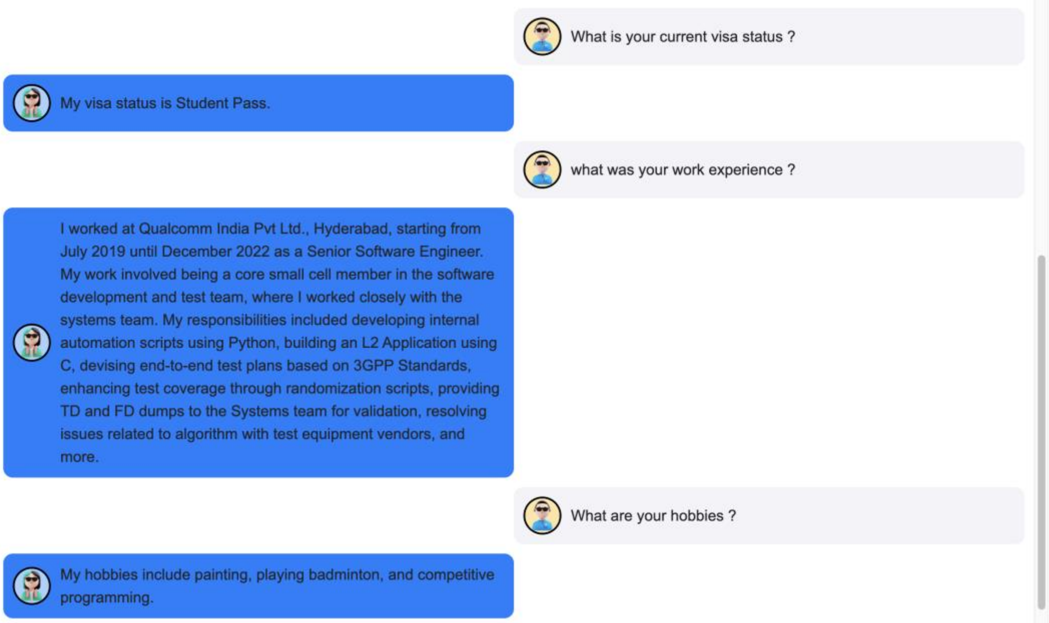
You: What is your name ?

Niha: My name is ChatGPT. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you? My name is Niharika. How may I assist you? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is ChatGPT. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is OpenAI, I am an AI assistant. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How may I assist you? My name is Niharika. How may I assist you today? My name is Niharika. How can I assist you today? My name is Niharika. How can I assist you today?

Snippet of the response

Each message in the conversation history is being used by the model to generate the response as the same conversation history is being sent multiple times.

6. **Final Model used (gpt-4-1106-preview):**
   No token limit issue faced up to certain number of times (based on the number of question asked and the length of answer generated). Data also provided all at once.



What is your current visa status ?

My visa status is Student Pass.

what was your work experience ?

I worked at Qualcomm India Pvt Ltd., Hyderabad, starting from July 2019 until December 2022 as a Senior Software Engineer. My work involved being a core small cell member in the software development and test team, where I worked closely with the systems team. My responsibilities included developing internal automation scripts using Python, building an L2 Application using C, devising end-to-end test plans based on 3GPP Standards, enhancing test coverage through randomization scripts, providing TD and FD dumps to the Systems team for validation, resolving issues related to algorithm with test equipment vendors, and more.

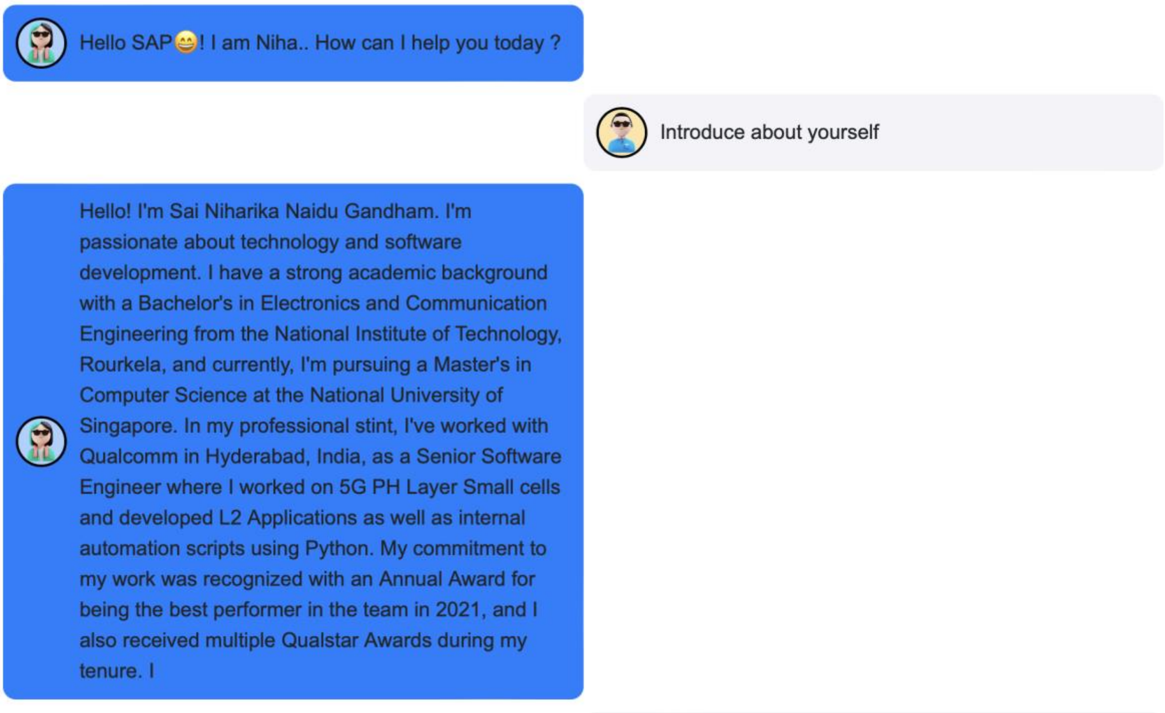
What are your hobbies ?

My hobbies include painting, playing badminton, and competitive programming.

7. **Handling Premature responses due to token limit:**
Due to the limited token limit access, I was able to run the tests limitedly only.
For simple questions, the model generated the answers well. But when the question demands relatively longer responses, due to the 'max_token' limit set the truncated responses were observed. I did try by changing the system message, token limits , but one or the other time it generated premature Reponses.

I tried to process the response further, incase if the **'finish_reason' == 'length'.**
Implemented a function, to take truncated response due to finish_reason == length and increased the max_token limit for such response. This function will run for the set max_token_limit until the generated response is not due to finish_reason==length. But this might not be right choice because, it might take forever to generate a single response and not a good experience for the user.

Hence, for now I am post-processing the generated response to check if the model has ended abruptly and removing the truncated sentence at the end.



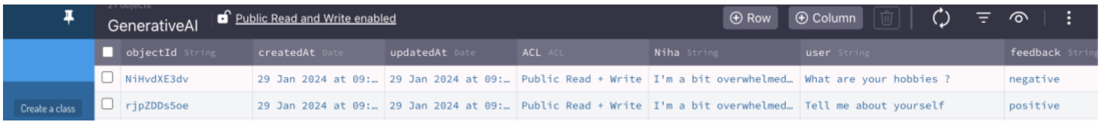
Hello SAP😄! I am Niha.. How can I help you today ?

Introduce about yourself

Hello! I'm Sai Niharika Naidu Gandham. I'm passionate about technology and software development. I have a strong academic background with a Bachelor's in Electronics and Communication Engineering from the National Institute of Technology, Rourkela, and currently, I'm pursuing a Master's in Computer Science at the National University of Singapore. In my professional stint, I've worked with Qualcomm in Hyderabad, India, as a Senior Software Engineer where I worked on 5G PH Layer Small cells and developed L2 Applications as well as internal automation scripts using Python. My commitment to my work was recognized with an Annual Award for being the best performer in the team in 2021, and I also received multiple Qualstar Awards during my tenure. I

8. **Back4App:**
For some applications, wider research might be required to know users perspective. Hence, implemented this database option where users questions, generated answers and the feedback(if given) will be stored in the there.

Based on this responses , model can be finetuned on this data to perform well.



□	objectId String	createdAt Date	updatedAt Date	ACL ACL	Niha String	user String	feedback String
□	NiHvdXE3dv	29 Jan 2024 at 09:…	29 Jan 2024 at 09:…	Public Read + Write	I'm a bit overwhelmed…	What are your hobbies ?	negative
□	rjpZDDs5oe	29 Jan 2024 at 09:…	29 Jan 2024 at 09:…	Public Read + Write	I'm a bit overwhelmed…	Tell me about yourself	positive

Snippet from Database

9. **Llama-2-7b-chat-hf:**
   I tried multiple times to do few-shot training on this model, but it always resulted in crash. Doing Few shot training with the good amount of resources, the model should definitely see better performance compared to current scenario.

Given the time limit, the amount of resources I have to run the models, the limitation to the API access, I could do this.