Nikhil Manda

CS 6320.001

Due: 4/29/2024

Chatbot Report

Dataset Preprocessing:

The dataset consists of NVIDIA documentation question and answer pairs. These question and answer pairs are generated from NVIDIA blogs on NVIDIA's website as well as other technical information that can be found on NVIDIA's website. The dataset contains three columns: index, question, and answer. There are 7,108 question-answer pairs in the dataset. There were no missing values in this dataset, but there were several duplicate question-answer pairs in the dataset. In this project, I used machine learning to do a text classification related task. To do this, I first provided 1000 rows of the dataset to Google Gemini to look at the patterns between various question-answer pairs and generate numeric categories to add to this subset of data. The reason only 1000 rows were used was to reduce the training time of the TensorFlow Sequential machine learning model that will be discussed later in this report. Another reason was that initially I tried with the whole preprocessed dataset of 6,425 unique entries, but the training time was taking a long time with my model. Now, this updated dataset with four columns, index, question, answer, and category and I loaded this dataset as a pickle file into my ipython notebook. Then I removed missing values and duplicates from this dataset which reduced the number of unique question-answer pairs to 963. This preprocessed data was used to create the knowledge base. Below is a screenshot of some entries of the knowledge base.

Knowledge Base:

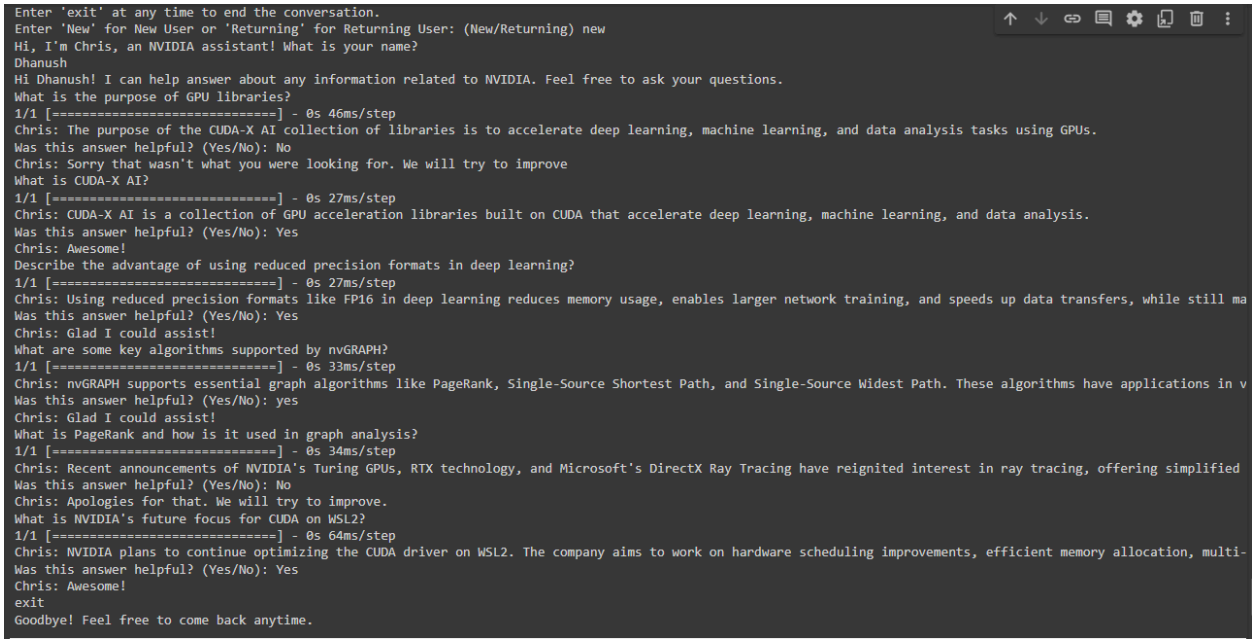
index	Question	Answer	Category
0	What is Hybridizer?	Hybridizer is a compiler from Altimesh that enables programming GPUs and accelerators using C# code or .NET Assembly.	0
1	How does Hybridizer generate optimized code?	Hybridizer uses decorated symbols to express parallelism and generates source code or binaries optimized for multicore CPUs and GPUs.	0
2	What are some parallelization patterns mentioned in the text?	The text mentions using parallelization patterns like Parallel.For and distributing parallel work explicitly, similar to CUDA.	0
3	How can you benefit from accelerators without learning their internal architecture?	You can benefit from accelerators' compute horsepower without learning the details of their internal architecture by using patterns like Parallel.For or CUDA-like distribution of parallel work.	0
4	What is an example of using Hybridizer?	An example in the text demonstrates using Parallel.For with a lambda to leverage the compute power of accelerators.	0
5	How can you debug and profile GPU code written with Hybridizer?	You can debug and profile GPU code created with Hybridizer using NVIDIA Nsight Visual Studio Edition.	0
6	What advanced C# features does Hybridizer implement?	Hybridizer implements advanced C# features, including virtual functions and generics.	0
7	What does the new NVIDIA Developer Blog post by Altimesh demonstrate?	The new NVIDIA Developer Blog post by Altimesh demonstrates how to accelerate C# and .NET code, and how to profile and debug it within Visual Studio.	4
8	What is the purpose of GPU libraries?	GPU libraries allow applications to be accelerated without requiring GPU-specific code.	4
9	What is the new feature in CUDA 5.5 version of NVIDIA CUFFT library?	The new feature in CUDA 5.5 version of NVIDIA CUFFT library is the support for the popular FFTW API for FFT acceleration.	4
10	How does the new CUDA version make FFT acceleration easier?	The new CUDA version allows developers to accelerate existing FFTW library calls on the GPU by changing the linker command line to link the CUFFT library instead of the FFTW library.	4
11	What is the benefit of using CUFFT library for FFT acceleration?	By using the CUFFT library instead of the FFTW library and re-linking the application, developers can leverage GPU acceleration with minimal code changes.	4
12	What change is required to accelerate function calls on the GPU using CUFFT library?	The only code change required is to use the cufftw.h header file, ensuring that unsupported functions are not called at compile time.	4
13	How can developers request a topic for future CUDACast episodes?	Developers can leave a comment to request a topic for a future episode of CUDACast or provide feedback.	9
14	Who is Gil Speyer?	Gil Speyer is a Senior Postdoctoral Fellow at the Translational Genomics Research Institute (TGen).	9
15	What is EDDY?	EDDY is a statistical analysis tool developed by scientists at TGen that examines how cells' DNA controls protein production and protein interactions using NVIDIA Tesla K40 GPUs and CUDA.	4
16	How does EDDY contribute to precision medicine?	EDDY informs doctors with the best options for attacking each individual patient's cancer by analyzing how cells' DNA controls protein production and interactions.	4
17	What technology did the scientists use to develop EDDY?	The scientists used NVIDIA Tesla K40 GPUs and CUDA technology to develop EDDY.	4

Chatbot System:

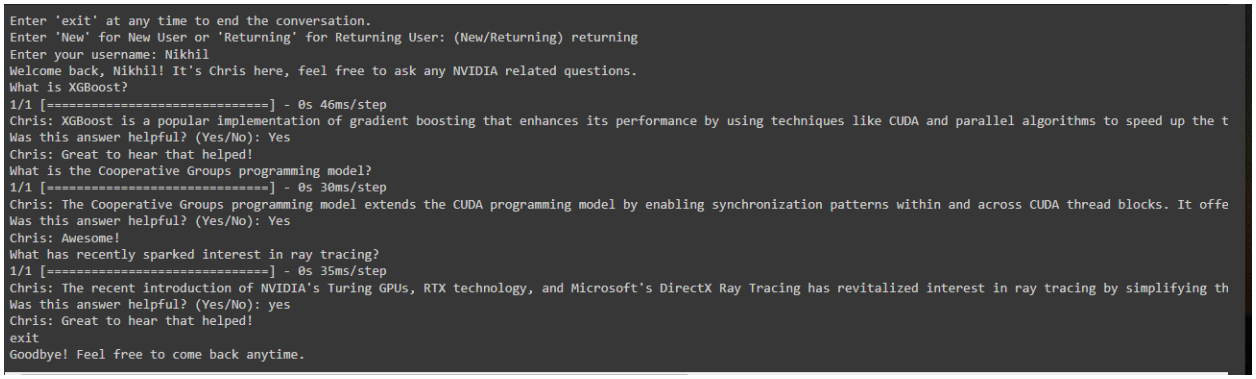
The entry point of the chatbot program first initiates the conversation with the user, offering them to identify themselves as either a new user or a returning user. Based on the user's input, it directs the flow of the conversation in three cases. The program handles case-insensitive commands for the user's action. In one case, the users can exit the conversation by entering 'exit' for which the chatbot program will terminate. For new users, the program transitions into the new user flow, from which they introduce themselves and ask the bot questions. After this, the user's user model is created. For returning users, the bot prompts the user to enter their username to load their existing user model. If the user model is found, it proceeds with the returning user flow, allowing users to continue asking the bot questions and interact with the knowledge base. If no user model is found, it suggests starting as a new user. In addition, if the command is an invalid command meaning that it is not any of the commands that the bot is expecting, then it will tell the user to enter a valid command. The user models for each user are created as a json file and for returning users, the same user model gets updated. In both new and returning user flows, each time the bot answers the user's question, it will ask a followup question on if the answer was helpful to the answer. This information is updated in the specific user's user model. Then after the user responds to this followup question, the bot will reply back with a random positive or negative comment. For instance, if the user responds "yes", then the bot could reply back with "Awesome!", but if the user responds "no", then the bot could reply back with "Apologies for that. We will try to improve".

In this project, I used a machine learning workflow where I am trying to predict categories based on questions from a dataset. The process starts by separating our dataset into two parts: the questions ("Features") and their respective categories ("Target variable"). I split these into training and testing groups, with 80% of the data used for training and 20% used for testing. The next step involved preparing the text data through a process called tokenization, where each unique word in the training data is assigned a numerical index. These words are then converted into sequences of numbers representing the original sentences from the questions. To ensure consistency in input size for modeling, these sequences are padded to a fixed length. Especially when I compare the user's query with questions in the knowledge base, I first preprocess the user's query by removing stopwords and tokenizing and lemmatizing the query prior to being tokenized. I then build a neural network model that is implemented using TensorFlow's Keras sequential model. This model learns to predict the category based on these sequences. This model goes through training, where it learns from the training data, and testing, where I check how well it performs on unseen data. After training, the model can predict the category of new questions. Furthermore, I create a dictionary for storing questions and their answers based on the categories, which helps in quickly retrieving information based on the model's predictions. Additionally, I use the TF-IDF technique as well as cosine similarity so that the model can suggest the most relevant answers based on how similar the words in new questions are to those in our knowledge base. When I evaluated my machine learning model, the highest accuracy I got was around 84% meaning that the chatbot will answer questions correctly 84% of the time.

Sample Dialog Interactions:

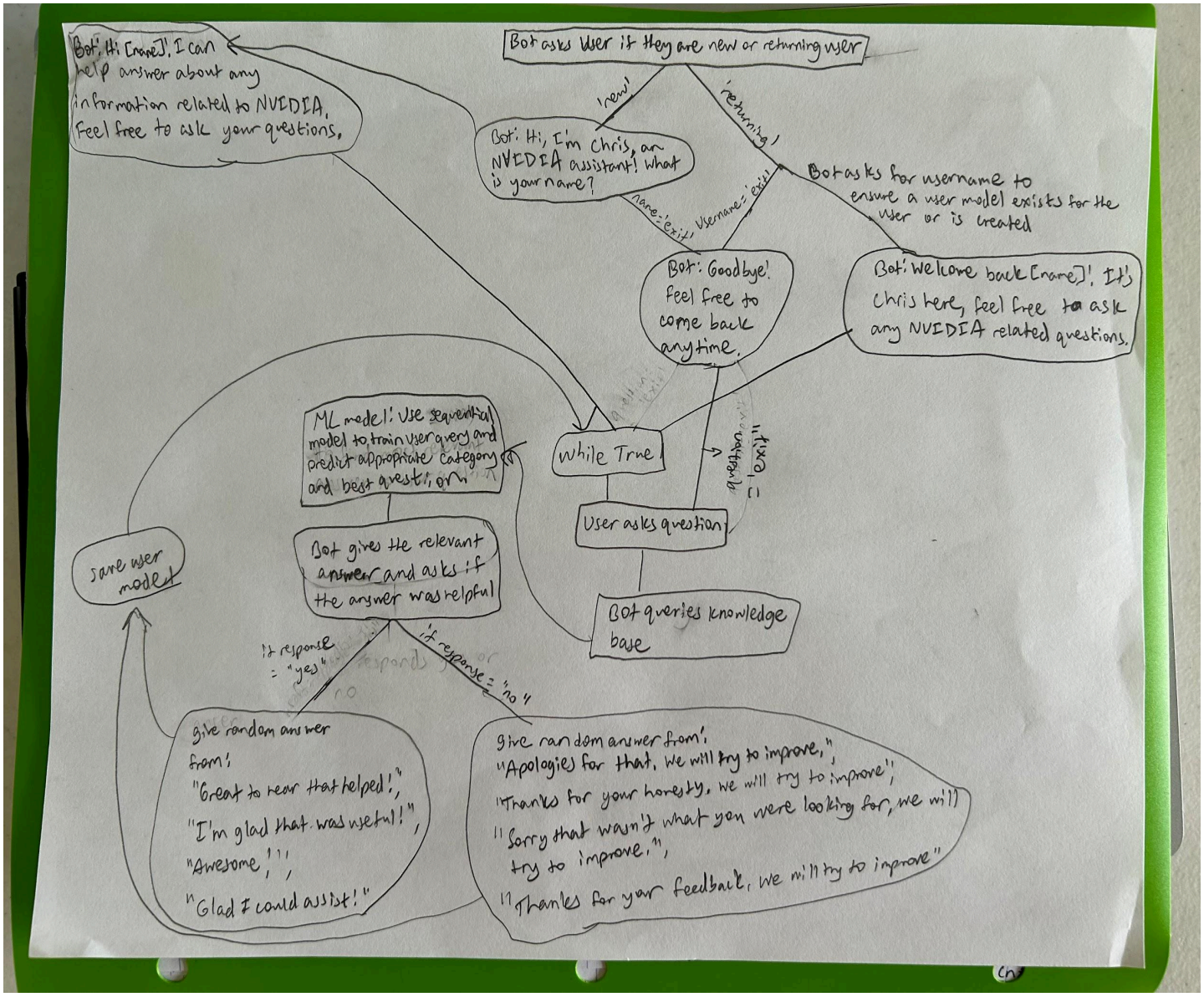


```
Enter 'exit' at any time to end the conversation.
Enter 'New' for New User or 'Returning' for Returning User: (New/Returning) new
Hi, I'm Chris, an NVIDIA assistant! What is your name?
Dhanush
Hi Dhanush! I can help answer about any information related to NVIDIA. Feel free to ask your questions.
What is the purpose of GPU libraries?
1/1 [==============================] - 0s 46ms/step
Chris: The purpose of the CUDA-X AI collection of libraries is to accelerate deep learning, machine learning, and data analysis tasks using GPUs.
Was this answer helpful? (Yes/No): No
Chris: Sorry that wasn't what you were looking for. We will try to improve
What is CUDA-X AI?
1/1 [==============================] - 0s 27ms/step
Chris: CUDA-X AI is a collection of GPU acceleration libraries built on CUDA that accelerate deep learning, machine learning, and data analysis.
Was this answer helpful? (Yes/No): Yes
Chris: Awesome!
Describe the advantage of using reduced precision formats in deep learning?
1/1 [==============================] - 0s 27ms/step
Chris: Using reduced precision formats like FP16 in deep learning reduces memory usage, enables larger network training, and speeds up data transfers, while still ma
Was this answer helpful? (Yes/No): Yes
Chris: Glad I could assist!
What are some key algorithms supported by nvGRAPH?
1/1 [==============================] - 0s 33ms/step
Chris: nvGRAPH supports essential graph algorithms like PageRank, Single-Source Shortest Path, and Single-Source Widest Path. These algorithms have applications in v
Was this answer helpful? (Yes/No): yes
Chris: Glad I could assist!
What is PageRank and how is it used in graph analysis?
1/1 [==============================] - 0s 34ms/step
Chris: Recent announcements of NVIDIA's Turing GPUs, RTX technology, and Microsoft's DirectX Ray Tracing have reignited interest in ray tracing, offering simplified
Was this answer helpful? (Yes/No): No
Chris: Apologies for that. We will try to improve.
What is NVIDIA's future focus for CUDA on WSL2?
1/1 [==============================] - 0s 64ms/step
Chris: NVIDIA plans to continue optimizing the CUDA driver on WSL2. The company aims to work on hardware scheduling improvements, efficient memory allocation, multi-
Was this answer helpful? (Yes/No): Yes
Chris: Awesome!
exit
Goodbye! Feel free to come back anytime.
```



```
Enter 'exit' at any time to end the conversation.
Enter 'New' for New User or 'Returning' for Returning User: (New/Returning) returning
Enter your username: Nikhil
Welcome back, Nikhil! It's Chris here, feel free to ask any NVIDIA related questions.
What is XGBoost?
1/1 [==============================] - 0s 46ms/step
Chris: XGBoost is a popular implementation of gradient boosting that enhances its performance by using techniques like CUDA and parallel algorithms to speed up the t
Was this answer helpful? (Yes/No): Yes
Chris: Great to hear that helped!
What is the Cooperative Groups programming model?
1/1 [==============================] - 0s 30ms/step
Chris: The Cooperative Groups programming model extends the CUDA programming model by enabling synchronization patterns within and across CUDA thread blocks. It offe
Was this answer helpful? (Yes/No): Yes
Chris: Awesome!
What has recently sparked interest in ray tracing?
1/1 [==============================] - 0s 35ms/step
Chris: The recent introduction of NVIDIA's Turing GPUs, RTX technology, and Microsoft's DirectX Ray Tracing has revitalized interest in ray tracing by simplifying th
Was this answer helpful? (Yes/No): yes
Chris: Great to hear that helped!
exit
Goodbye! Feel free to come back anytime.
```

Dialog Tree:


Bot: Hi [name]. I can help answer about any information related to NVIDIA. Feel free to ask your questions.
Bot asks user if they are new or returning user
'new'
'returning'
Bot: Hi, I'm Chris, an NVIDIA assistant! What is your name?
Bot asks for username to ensure a user model exists for the user or is created
name='exit', username='exit'
Bot: Goodbye! Feel free to come back anytime.
Bot: Welcome back [name]! It's Chris here, feel free to ask any NVIDIA related questions.
ML model: Use sequential model to train user query and predict appropriate category and best question.
while True
if username != 'exit'
Bot gives the relevant answer and asks if the answer was helpful
User asks question
Save user model
Bot queries knowledge base
if response = "yes"
if response = "no"
give random answer from: "Great to hear that helped!", "I'm glad that was useful!", "Awesome!", "Glad I could assist!"
give random answer from: "Apologies for that. We will try to improve.", "Thanks for your honesty. We will try to improve", "Sorry that wasn't what you were looking for, we will try to improve.", "Thanks for your feedback. We will try to improve"

Personal Evaluation:

Strengths: The chatbot program allows for the user to exit at any point in the program, the chatbot program is able to handle user input case insensitivity, which provides flexibility into how the user wants to provide input. Once the user obtains the responses for the likes and dislikes, it creates a user model and stores these results correctly. The machine learning neural network model utilizes Gated Recurrent Unit layers to process sequences of text and retain information from previous inputs. This allows it to better understand context and sequence in conversation. This model also learns to capture semantic similarities and relationships between words and this model is also scalable meaning that the model can be trained with more data over time to improve its understanding. The use of TF-IDF vectorization for individual categories allows for a more targeted response mechanism by comparing query vectors and document vectors within the predicted category, which improves the relevance of the responses. Other than that, the training for my neural network model implementation is pretty fast.

Weaknesses: With the model I implemented, sometimes the chatbot is able to correctly provide the relevant correct answer for some questions, but for other questions, it is the opposite case. There are many reasons for this and one of them I think is that the tokenizer is restricted with a maximum vocabulary size of 1000 words, which can limit the chatbot's ability to understand and process uncommon words or new terminologies that are not part of its initial training data. While I was able to get an accuracy of around 84% for my model, it can definitely perform better with more hyperparameter fine-tuning to improve the chatbot's response relevancy.

Public Evaluation Survey 1: (NET_ID : VXN220052 , Name : Navuluru, Venkata Subbaiah Pavan Karthik)

Rate your overall experience with the chatbot.

- Very dissatisfied
- Not satisfied
- Neutral
- Satisfied
- Very satisfied

Rate your experience with chatbot response timing and delivery.

- Very dissatisfied
- Not satisfied
- Neutral
- Satisfied
- Very satisfied

Rate your experience on the chatbot response relevance and accuracy.

- Very dissatisfied
- Not satisfied
- Neutral
- Satisfied
- Very satisfied

Public Evaluation Survey 2: (NET_ID : SXK230027 , Name : Konuru, Shiva Dhanush)
Rate your overall experience with the chatbot.

- Very dissatisfied
- Not satisfied
- Neutral
- Satisfied
- Very satisfied

Rate your experience with chatbot response timing and delivery.

- Very dissatisfied
- Not satisfied
- Neutral
- Satisfied
- Very satisfied

Rate your experience on the chatbot response relevance and accuracy.

- Very dissatisfied
- Not satisfied
- Neutral
- Satisfied
- Very satisfied

Public Evaluation Survey 3: (NET_ID : mkm200004 , Name : Mallik, Manish Kumar)

Rate your overall experience with the chatbot.

- Very dissatisfied
- Not satisfied
- Neutral
- Satisfied
- Very satisfied

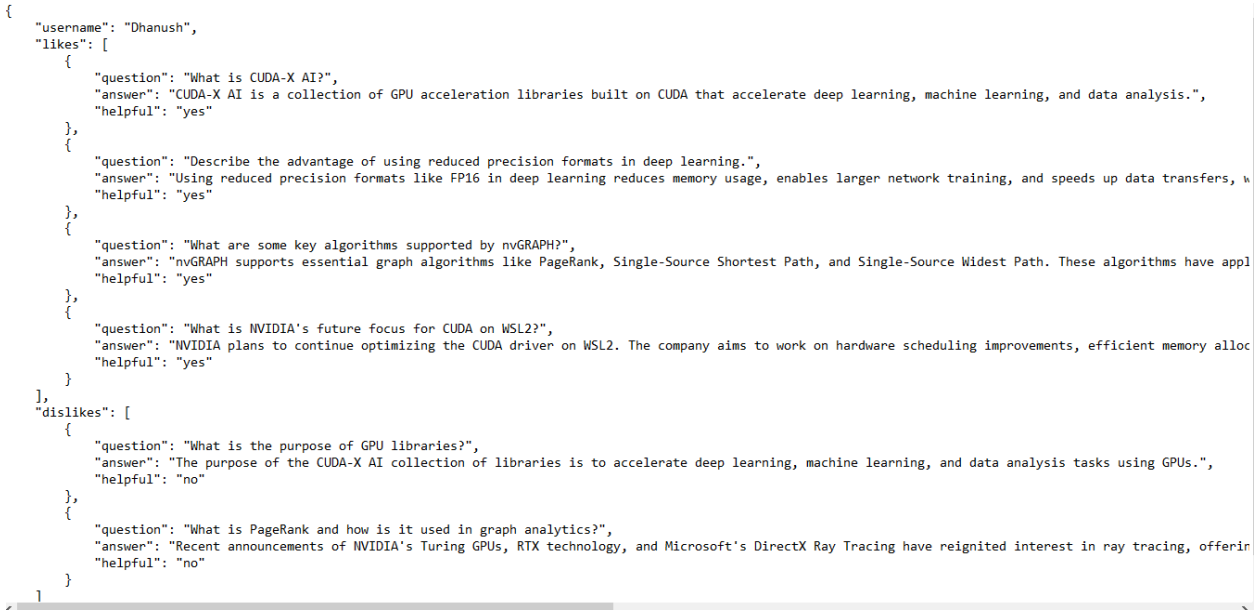
Rate your experience with chatbot response timing and delivery.

- Very dissatisfied
- Not satisfied
- Neutral
- Satisfied
- Very satisfied

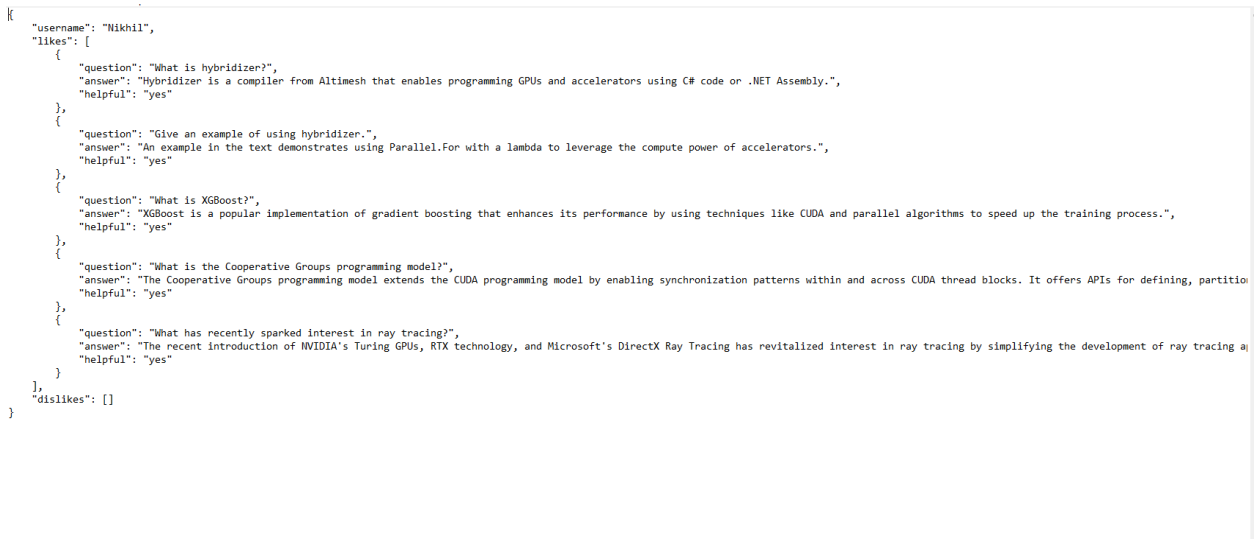
Rate your experience on the chatbot response relevance and accuracy.

- Very dissatisfied
- Not satisfied
- Neutral
- Satisfied
- Very satisfied

User Model Appendix:




```
{
    "username": "Dhanush",
    "likes": [
        {
            "question": "What is CUDA-X AI?",
            "answer": "CUDA-X AI is a collection of GPU acceleration libraries built on CUDA that accelerate deep learning, machine learning, and data analysis.",
            "helpful": "yes"
        },
        {
            "question": "Describe the advantage of using reduced precision formats in deep learning.",
            "answer": "Using reduced precision formats like FP16 in deep learning reduces memory usage, enables larger network training, and speeds up data transfers,
            "helpful": "yes"
        },
        {
            "question": "What are some key algorithms supported by nvGRAPH?",
            "answer": "nvGRAPH supports essential graph algorithms like PageRank, Single-Source Shortest Path, and Single-Source Widest Path. These algorithms have appl
            "helpful": "yes"
        },
        {
            "question": "What is NVIDIA's future focus for CUDA on WSL2?",
            "answer": "NVIDIA plans to continue optimizing the CUDA driver on WSL2. The company aims to work on hardware scheduling improvements, efficient memory alloc
            "helpful": "yes"
        }
    ],
    "dislikes": [
        {
            "question": "What is the purpose of GPU libraries?",
            "answer": "The purpose of the CUDA-X AI collection of libraries is to accelerate deep learning, machine learning, and data analysis tasks using GPUs.",
            "helpful": "no"
        },
        {
            "question": "What is PageRank and how is it used in graph analytics?",
            "answer": "Recent announcements of NVIDIA's Turing GPUs, RTX technology, and Microsoft's DirectX Ray Tracing have reignited interest in ray tracing, offerir
            "helpful": "no"
        }
    ]
```


```
{
    "username": "Nikhil",
    "likes": [
        {
            "question": "What is hybridizer?",
            "answer": "Hybridizer is a compiler from Altimesh that enables programming GPUs and accelerators using C# code or .NET Assembly.",
            "helpful": "yes"
        },
        {
            "question": "Give an example of using hybridizer.",
            "answer": "An example in the text demonstrates using Parallel.For with a lambda to leverage the compute power of accelerators.",
            "helpful": "yes"
        },
        {
            "question": "What is XGBoost?",
            "answer": "XGBoost is a popular implementation of gradient boosting that enhances its performance by using techniques like CUDA and parallel algorithms to speed up the training process.",
            "helpful": "yes"
        },
        {
            "question": "What is the Cooperative Groups programming model?",
            "answer": "The Cooperative Groups programming model extends the CUDA programming model by enabling synchronization patterns within and across CUDA thread blocks. It offers APIs for defining, partitio
            "helpful": "yes"
        },
        {
            "question": "What has recently sparked interest in ray tracing?",
            "answer": "The recent introduction of NVIDIA's Turing GPUs, RTX technology, and Microsoft's DirectX Ray Tracing has revitalized interest in ray tracing by simplifying the development of ray tracing a
            "helpful": "yes"
        }
    ],
    "dislikes": []
}
```


The user model is saved in a json file as shown above. During the problem, the user model is created as a dictionary where the keys are username, likes, and dislikes. The value for the username is the name of the datatype string. The values for the likes and dislikes are a list of dictionaries that store the user's question, bot's answer, and whether the user thought the answer

was helpful or not. When the dictionary is accessed, dictionary operations to access the data are performed in the chatbot program.