NEURONEST: DEEP LEARNING CHATBOT FOR APPLIED DATA SCIENCE

Aboli Wankhade
Department of Applied Data Science
San Jose State University
San Jose, USA
aboli.wankhade@sjsu.edu

Deekshita Prakash Savanur
Department of Applied Data Science
San Jose State University
San Jose, USA
deekshita.prakashsavanur@sjsu.edu

Dharma Teja Kolluri
Department of Applied Data Science
San Jose State University
San Jose, USA
dharmateja.kolluri@sjsu.edu

Gouri Benni
Department of Applied Data Science
San Jose State University
San Jose, USA
gouri.benni@sjsu.edu

Rama Krishna Polur
Department of Applied Data Science
San Jose State University
San Jose, USA
ramakrishna.poluru@sjsu.edu

Uzair Riyaz Pachhapure
Department of Applied Data Science
San Jose State University
San Jose, USA
uzairriyaz.pachhapure@sjsu.edu

***Abstract*—This project addresses the need for an efficient, context-aware communication tool within the Applied Data Science department, where traditional methods of communication, such as direct emails and faculty interactions, often result in varied response times and interpretations, highlighting the need for a more efficient communication system. To address this, our project developed a specialized chatbot leveraging the capabilities of deep learning technology. Our approach involved extensive data gathering from various departmental resources, ensuring the chatbot was equipped with comprehensive and up-to-date information. We integrated state-of-the-art language models, including llama2, GPT-3.5 and RAG, to enable the chatbot to process and respond to inquiries with a high degree of accuracy and contextual relevance. The design of the chatbot was meticulously focused on user experience, ensuring it is intuitive, responsive, and capable of handling a wide spectrum of queries. In testing the chatbot, we employed evaluation metrics, such as the Sensitivity Analysis of Sentences (SAS) and the Bilingual Evaluation Understudy (BLEU) score. These assessments demonstrated a notable improvement in providing precise and contextually appropriate responses. The chatbot demonstrates the potential to enhance educational experiences and information accessibility, offering a promising solution to overcome the limitations of traditional communication methods in academic settings. The implementation of this chatbot in the Applied Data Science department represents a significant step forward in improving the accessibility and efficiency of information dissemination. It serves as a model for other academic departments facing similar challenges, showcasing the potential of AI-driven solutions in enhancing educational experiences. Our project's contributions lie not only in the technical realm but also in setting a precedent for the future of academic communication.**
Keywords— llama2, GPT-3.5, RAG, SAS, BLEAU

I. INTRODUCTION

The field of Applied Data Science is rapidly evolving, necessitating innovative communication solutions for efficient knowledge dissemination. This project focuses on overcoming the limitations inherent in conventional academic communication channels like emails and direct consultations, which often lead to delays and inconsistencies in information sharing. Our objective is to harness the power of deep learning technologies to develop a specialized chatbot, tailored to meet the specific needs of the Applied Data Science department.

In reviewing current literature, we identified a notable absence of deep learning applications in creating targeted communication tools within academic environments. This project aims to fill this gap by employing advanced language processing models. We hypothesize that a deep learning-enabled chatbot can offer more immediate, accurate, and contextually relevant responses compared to traditional communication methods. In this project, the objective is to develop an interactive chatbot for San Jose State University's Applied Data Science department. The approach involves leveraging Deep Learning techniques, namely LLM models, Transfer learning, and Reinforcement learning. The selection of these methods is based on their ability to effectively tackle the complications of interactive chatbots. The project's methodology starts with data collection that involves applying various web scraping techniques to obtain the data from the Applied Statistics department website. This step sets the groundwork for the development of the chatbot, ensuring that the collected data aligns with the requirements of the project. After the data was collected, complex LLM models such as Llama 2, GPT3.5, and RAG were used to train the data. The selection of the most appropriate model is based on performance metrics and computational efficiency, ensuring to make the most efficient use of available resources. After the models undergo training, the next step involves fine-tuning them. This phase relies on transfer learning, a technique that allows for the adaptation of these models to the precise demands of the department. This customization ensures that the chatbot becomes skilled at addressing queries related to the domain and understanding terminology effectively. To enhance the chatbot's accuracy, an advanced recognition system is being introduced. This system, driven by similarity measures and Natural Language Processing (NLP), is crucial for classifying user queries. It enables the chatbot to categorize user questions into predefined intent with various department-related topics. These topics range from course information and faculty profiles to research areas and administrative queries. The objective is to ensure that the chatbot delivers accurate and contextually appropriate answers for a wide range of questions. The significance of this project lies in its potential to transform how academic information is accessed and shared. By integrating cutting-edge technology in deep learning, the project not only addresses a practical need within the department but also contributes to the broader field of AI in education, setting a precedent for future innovations.

II. RELATED WORK

The research paper by David Santandreu Calonge, Linda Smail, and Firuz Kamalov, titled "Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics", presents a detailed comparative analysis of four AI chatbots - ChatGPT, GPT-4, Bard, and LLaMA. The study focuses on evaluating their capabilities in calculus and statistics education. The authors assess these platforms for their features, functionalities, and potential applications, aiming to understand their effectiveness as learning tools in these domains. The findings suggest that while GPT-4 generally outperforms the others, each chatbot has unique strengths and limitations, with implications for their use in enhancing student learning experiences in higher education.

The research paper "Building Emotional Support Chatbots in the Era of LLMs" by Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie focuses on developing chatbots capable of providing emotional support using Large Language Models (LLMs). It introduces an innovative methodology combining human insights with LLMs to create a comprehensive emotional support dialogue dataset, ExTES. The study explores various fine-tuning methods on the LLaMA model to optimize it for emotional support interactions. The paper contributes significantly to the field by addressing challenges in data scarcity and training methodologies in the realm of emotional support bots, providing a solid foundation for future research and applications.

The paper "CHATA: Towards an Intelligent Question-Answer Teaching Assistant using Open-Source LLMs," authored by Yann Hicke, Anmol Agarwal, Qianou (Christina) Ma, and Paul Denny, presents an advanced educational question-answering (QA) system leveraging open-source Large Language Models (LLMs) for scalability and data privacy. Their approach integrates Retrieval Augmented Generation (RAG), Supervised Fine-Tuning (SFT), and Direct Preference Optimization (DPO) methods. This novel combination significantly improves QA performance in an educational context, particularly in a large-scale introductory programming course, demonstrating a 30% improvement in answer quality and offering insights into future educational data processing and evaluation methods.

The paper "LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation" by Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, and others, explores using ChatGPT for creating chatbots that simulate psychiatrists and patients in psychiatric settings. It focuses on designing chatbots to engage in diagnostic conversations and patient simulations, leveraging ChatGPT's capabilities. The study includes collaboration with psychiatrists, the development of dialogue systems, and evaluation experiments involving real psychiatrists and patients. The findings underscore ChatGPT's utility in psychiatric scenarios, emphasizing the importance of prompt design in influencing chatbot behavior and user experience.

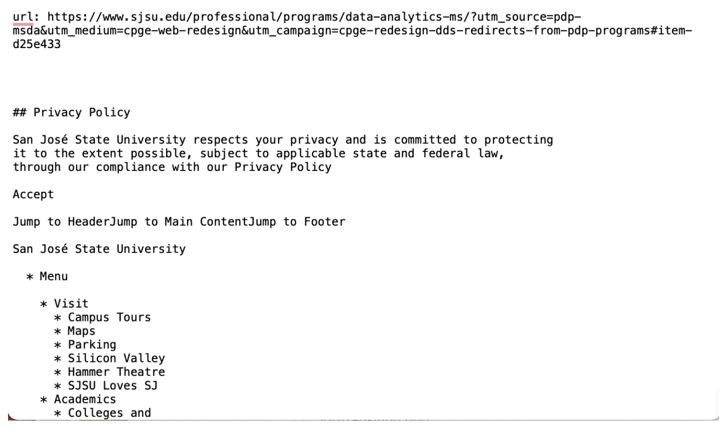
The paper "Chat Vector: A Simple Approach to Equip LLMs with New Language Chat Capabilities," authored by Shih-Cheng Huang, Pin-Zu Li, Yu-Chi Hsu, and others, introduces an innovative method to enhance non-English Large Language Models (LLMs) using a 'chat vector'. This approach modifies the traditional training paradigm by combining continual pre-training (CP) with chat vectors, derived by subtracting the pre-trained weights from chat-enhanced model weights. This method demonstrates significant improvements in LLMs' conversational skills, toxicity handling, and instruction-following abilities in various languages, including Traditional Chinese and Korean. The study provides a computationally efficient alternative to the conventional RLHF process, showcasing its versatility and effectiveness across different linguistic contexts.

The research paper "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering" by Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara, focuses on enhancing the domain adaptability of RAG models. The authors propose RAG-end2end, a variant that updates all model components during training, including the external knowledge base. Additionally, an auxiliary training signal is introduced for more domain-specific knowledge. Their experiments demonstrate significant improvements in RAG's performance across different domains, such as COVID-19, news, and conversations.

III. METHODOLOGY

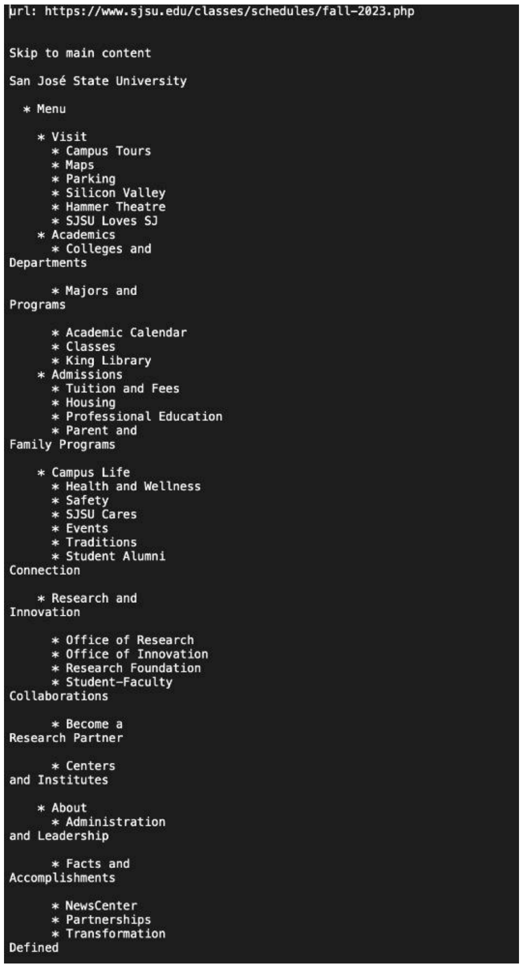
A. Dataset Description

1) *Dataset Name:* SJSU Data Science Q&A Dataset
2) *Dataset Source:* The San Jose State University Data Science Department's website was scraped to generate the project dataset. With this approach, text data, which includes a wide range of content like FAQs, course details, program prerequisites, faculty profiles, and curriculum descriptions was methodically extracted. Creating a comprehensive corpus of questions and answers that are indicative of common queries students may have regarding the Data Science program is the main objective of this data extraction process.



```
url: https://www.sjsu.edu/professional/programs/data-analytics-ms/?utm_source=pdp-
msda&utm_medium=cpge-web-redesign&utm_campaign=cpge-redesign-dds-redirects-from-pdp-programs#item-
d25e433



## Privacy Policy

San José State University respects your privacy and is committed to protecting
it to the extent possible, subject to applicable state and federal law,
through our compliance with our Privacy Policy

Accept

Jump to HeaderJump to Main ContentJump to Footer

San José State University

  * Menu

    * Visit
      * Campus Tours
      * Maps
      * Parking
      * Silicon Valley
      * Hammer Theatre
      * SJSU Loves SJ
    * Academics
      * Colleges and
```

Fig. 1. Raw Dataset Sample



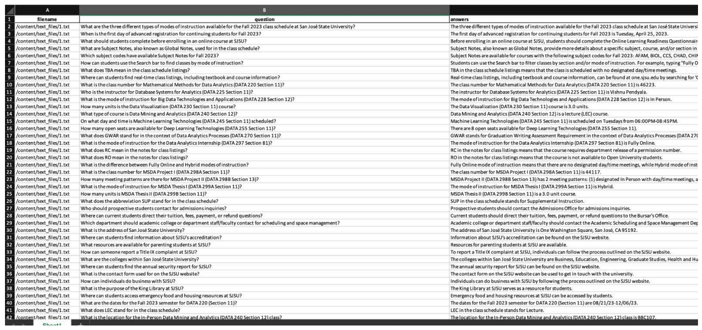
```
url: https://www.sjsu.edu/classes/schedules/fall-2023.php


Skip to main content

San José State University

  * Menu

    * Visit
      * Campus Tours
      * Maps
      * Parking
      * Silicon Valley
      * Hammer Theatre
      * SJSU Loves SJ
    * Academics
      * Colleges and
Departments

      * Majors and
Programs

      * Academic Calendar
      * Classes
      * King Library
    * Admissions
      * Tuition and Fees
      * Housing
      * Professional Education
      * Parent and
Family Programs

    * Campus Life
      * Health and Wellness
      * Safety
      * SJSU Cares
      * Events
      * Traditions
      * Student Alumni
Connection

    * Research and
Innovation

      * Office of Research
      * Office of Innovation
      * Research Foundation
      * Student-Faculty
Collaborations

      * Become a
Research Partner

      * Centers
and Institutes

      * About
      * Administration
and Leadership

      * Facts and
Accomplishments

      * NewsCenter
      * Partnerships
      * Transformation
Defined
```

Fig. 2. Raw Dataset Sample

3) *Data Collection Process:* The data scraping process is conducted using Python libraries such as BeautifulSoup, Selenium, and requests. The script is designed to navigate through several web pages, extract relevant information, automate interactions with web browsers, and save it for further processing. Dynamically loaded content can be handled by the script, guaranteeing that the most recent and pertinent data is recorded. This has great importance in the application for a university website that regularly updates information about events, courses, and faculty. Error handling procedures were performed to handle any problem that may arise during the scraping process (broken links or unavailable pages). The dataset's completeness and integrity are guaranteed by this process.

4) *Dataset Splits:* The dataset is divided into three subsets: testing, validation, and training. To ensure that the model does not have previous exposure to validation and test data during training, this divide is essential for preventing data leaking. Typically, a conventional split ratio of 70% for training, 15% for validation, and 15% for testing is used. In the scenario of the dataset having uneven categories, stratification is followed (A stratified split is used to make sure that each category is fairly represented in training, validation, and test sets). This method aids in preserving the dataset's representativeness and diversity, enabling a more thorough assessment of the model's effectiveness across various question kinds.

5) *Data Preprocessing:*
   a) *HTML Parsing and Content Extraction:* The website's HTML content is parsed using BeautifulSoup, which concentrates on removing unnecessary features like scripts, styles, and navigation menus while preserving pertinent data. 'html2text package' was used to extract HTML content which was converted to plain text.
   b) *Text Normalization*: Since text data is case-sensitive, all retrieved text is changed to lowercase to maintain consistency. The model's complexity is decreased at this step.
   c) *Whitespace and Special Characters Handling:* To keep the content orderly, extra whitespace, newline characters, and special characters are eliminated/normalized.
   d) *Tokenization and Encoding:* After cleaning, the text is tokenized. This entails formatting text to fit the model when employing a model such as GPT-2, making sure that the input is in a consistent and comprehensible manner for the neural network.



	A	B	
1	filename	question	answers
2	/content/text_files/1.txt	What are the three different types of modes of instruction available for the Fall 2023 class schedule at San José State University?	The three different types of modes of instruction available for the Fall 2023 class schedule at San José State Universi
3	/content/text_files/1.txt	When is the first day of advanced registration for continuing students for Fall 2023?	The first day of advanced registration for continuing students for Fall 2023 is Tuesday, April 25, 2023.
4	/content/text_files/1.txt	What should students complete before enrolling in an online course at SJSU?	Before enrolling in an online course at SJSU, students should complete the Online Learning Readiness Questionnair
5	/content/text_files/1.txt	What are Subject Notes, also known as Global Notes, used for in the class schedule?	Subject Notes, also known as Global Notes, provide more details about a specific subject, course, and/or section in
6	/content/text_files/1.txt	Which subject codes have available Subject Notes for Fall 2023?	Subject Notes are available for courses with the following subject codes for Fall 2023: AFAM, BIOL, CCS, CHAD, CHIN
7	/content/text_files/1.txt	How can students use the Search bar to find classes by mode of instruction?	Students can use the Search bar to filter classes by section and/or mode of instruction. For example, typing "Fully O
8	/content/text_files/1.txt	What does TBA mean in the class schedule listings?	TBA in the class schedule listings means that the class is scheduled with no designated day/time meetings.
9	/content/text_files/1.txt	Where can students find real-time class listings, including textbook and course information?	Real-time class listings, including textbook and course information, can be found at one.sjsu.edu by searching for 'C
10	/content/text_files/1.txt	What is the class number for Mathematical Methods for Data Analytics (DATA 220 Section 11)?	The class number for Mathematical Methods for Data Analytics (DATA 220 Section 11) is 46223.
11	/content/text_files/1.txt	Who is the instructor for Database Systems for Analytics (DATA 225 Section 11)?	The instructor for Database Systems for Analytics (DATA 225 Section 11) is Vishnu Pendyala.
12	/content/text_files/1.txt	What is the mode of instruction for Big Data Technologies and Applications (DATA 228 Section 12)?	The mode of instruction for Big Data Technologies and Applications (DATA 228 Section 12) is In Person.
13	/content/text_files/1.txt	How many units is the Data Visualization (DATA 230 Section 11) course?	The Data Visualization (DATA 230 Section 11) course is 3.0 units.
14	/content/text_files/1.txt	What type of course is Data Mining and Analytics (DATA 240 Section 12)?	Data Mining and Analytics (DATA 240 Section 12) is a lecture (LEC) course.
15	/content/text_files/1.txt	On what day and time is Machine Learning Technologies (DATA 245 Section 11) scheduled?	Machine Learning Technologies (DATA 245 Section 11) is scheduled on Tuesdays from 06:00PM-08:45PM.
16	/content/text_files/1.txt	How many open seats are available for Deep Learning Technologies (DATA 255 Section 11)?	There are 8 open seats available for Deep Learning Technologies (DATA 255 Section 11).
17	/content/text_files/1.txt	What does GWAR stand for in the context of Data Analytics Processes (DATA 270 Section 11)?	GWAR stands for Graduation Writing Assessment Requirement in the context of Data Analytics Processes (DATA 27(
18	/content/text_files/1.txt	What is the mode of instruction for the Data Analytics Internship (DATA 297 Section 81)?	The mode of instruction for the Data Analytics Internship (DATA 297 Section 81) is Fully Online.
19	/content/text_files/1.txt	What does RC mean in the notes for class listings?	RC in the notes for class listings means that the course requires department release of a permission number.
20	/content/text_files/1.txt	What does RO mean in the notes for class listings?	RO in the notes for class listings means that the course is not available to Open University students.
21	/content/text_files/1.txt	What is the difference between Fully Online and Hybrid modes of instruction?	Fully Online mode of instruction means that there are no designated day/time meetings, while Hybrid mode of inst
22	/content/text_files/1.txt	What is the class number for MSDA Project I (DATA 298A Section 11)?	The class number for MSDA Project I (DATA 298A Section 11) is 44117.
23	/content/text_files/1.txt	How many meeting patterns are there for MSDA Project II (DATA 298B Section 13)?	MSDA Project II (DATA 298B Section 13) has 2 meeting patterns: (1) designated In Person with day/time meetings, a
24	/content/text_files/1.txt	What is the mode of instruction for MSDA Thesis I (DATA 299A Section 11)?	The mode of instruction for MSDA Thesis I (DATA 299A Section 11) is Hybrid.
25	/content/text_files/1.txt	How many units is MSDA Thesis II (DATA 299B Section 11)?	MSDA Thesis II (DATA 299B Section 11) is a 3.0 unit course.
26	/content/text_files/1.txt	What does the abbreviation SUP stand for in the class schedule?	SUP in the class schedule stands for Supplemental Instruction.
27	/content/text_files/1.txt	Who should prospective students contact for admissions inquiries?	Prospective students should contact the Admissions Office for admissions inquiries.
28	/content/text_files/1.txt	Where can current students direct their tuition, fees, payment, or refund questions?	Current students should direct their tuition, fees, payment, or refund questions to the Bursar's Office.
29	/content/text_files/1.txt	Which department should academic college or department staff/faculty contact for scheduling and space management?	Academic college or department staff/faculty should contact the Academic Scheduling and Space Management Dep
30	/content/text_files/1.txt	What is the address of San José State University?	The address of San José State University is One Washington Square, San José, CA 95192.
31	/content/text_files/1.txt	Where can students find information about SJSU's accreditation?	Information about SJSU's accreditation can be found on the SJSU website.
32	/content/text_files/1.txt	What resources are available for parenting students at SJSU?	Resources for parenting students at SJSU are available.
33	/content/text_files/1.txt	How can someone report a Title IX complaint at SJSU?	To report a Title IX complaint at SJSU, individuals can follow the process outlined on the SJSU website.
34	/content/text_files/1.txt	What are the colleges within San José State University?	The colleges within San José State University are Business, Education, Engineering, Graduate Studies, Health and Hu
35	/content/text_files/1.txt	Where can students find the annual security report for SJSU?	The annual security report for SJSU can be found on the SJSU website.
36	/content/text_files/1.txt	What is the contact form used for on the SJSU website?	The contact form on the SJSU website can be used to get in touch with the university.
37	/content/text_files/1.txt	How can individuals do business with SJSU?	Individuals can do business with SJSU by following the process outlined on the SJSU website.
38	/content/text_files/1.txt	What is the purpose of the King Library at SJSU?	The King Library at SJSU serves as a resource for students.
39	/content/text_files/1.txt	Where can students access emergency food and housing resources at SJSU?	Emergency food and housing resources at SJSU can be accessed by students.
40	/content/text_files/1.txt	What are the dates for the Fall 2023 semester for DATA 220 (Section 11)?	The dates for the Fall 2023 semester for DATA 220 (Section 11) are 08/21/23-12/06/23.
41	/content/text_files/1.txt	What does LEC stand for in the class schedule?	LEC in the class schedule stands for Lecture.
42	/content/text_files/1.txt	What is the location for the In-Person Data Mining and Analytics (DATA 240 Section 12) class?	The location for the In-Person Data Mining and Analytics (DATA 240 Section 12) class is BBC107.

Fig. 3. Cleaned Dataset Sample

6) *Data Augmentation and Robustness:* Synthetic Question Generation: To further enhance the dataset, advanced approaches including text paraphrasing and synthetic question production were explored. This includes modifying pre-existing questions using language models to make them more robust in the dataset.

7) *Contextual Variations:* Formulated question variants according to distinct contexts to guarantee the chatbot's ability to manage a broad spectrum of student viewpoints and requirements.

B. Proposed Model

In this project, three Large Language Models(LLM) were utilized in implementing the chatbot for the Applied Sciences department. LLMs are sophisticated AI models trained on extensive text corpora. The application of LLMs in chatbots is transformative. These models enhance the chatbot's ability to understand complex queries and respond in a contextually relevant manner. They exhibit remarkable flexibility, adapting to diverse conversational styles and domains without extensive retraining. There are various LLMs available in the current market with each model having its pros and cons. After researching all the models, Llama-2-7b, GPT 3.5, and the hybrid model of GPT 3.5 and RAG were implemented.

1) *Llama-2-7b:* LLaMA (llama-2-7b) is one of the most popular LLMs which is based on the Transformer architecture. The Transformer model is composed of an encoder and decoder, each consisting of multiple layers. However, for language modeling tasks like LLaMA is

designed for, often only the decoder part is used.Given that LLaMA is based on the Transformer architecture, it primarily uses transformer layers. These layers are composed of two key components such as the Self-Attention Mechanism which allows the model to weigh the importance of different words in a sentence, providing contextual understanding. Feed-forward networks are dense (fully connected) layers within each transformer block that process the information further after the attention mechanism. In addition to that, in this research, projection layers in Attention Mechanism such as q_proj, k_proj, v_proj, and o_proj. These refer to the query, key, value, and output projections in the self-attention mechanism, which are essential for calculating attention scores and subsequent outputs. LoRA (Low-Rank Adaptation) Layers target specific modules like gate_proj and up_proj, which are custom layers in the model used for further upscaling the layers within the model.

Various regularization techniques were used in this model implementation such as Dropout, Weight Decay, and Gradient Clipping. Dropout is a technique where randomly selected neurons are ignored during training, which helps in preventing overfitting. Weight decay is a method to add a penalty to the loss function based on the size of the weights, encouraging smaller weights to prevent overfitting. Gradient Clipping is a technique to prevent exploding gradients by limiting the size of the gradients during backpropagation.

2) *GPT 3.5:* In this research, gpt-3.5-turbo-1106 has been used to fine-tune our dataset to implement the chatbot. The basic architecture of the GPT model is based on Transformers and the architecture is the same as the Llama model. The GPT-3.5-turbo-1106 model represents a significant advancement in the field of natural language processing, particularly in the context of developing sophisticated chatbots. At the heart of GPT-3.5 are the Transformer layers, which are fundamental to its operation and capabilities. These Transformer layers are composed of two primary elements: self-attention layers and feed-forward neural networks. The self-attention layers are particularly crucial, enabling the model to weigh and integrate different parts of the input text, thereby providing a comprehensive understanding of the context and nuances of the language.Complementing the self-attention mechanisms are the feed-forward neural networks within each Transformer block. These dense layers, equipped with non-linear activation functions, process the information further, adding depth to the model's language processing capabilities. In addition to these, a distinctive aspect of GPT-3.5's architecture is its use of multiple attention heads within the self-attention layers. This multi-head attention allows the model to concurrently focus on different parts of the input sequence, enhancing its ability to parse and understand complex language structures. Layer normalization is another integral component, applied both before and after the self-attention and feed-forward layers. This normalization process stabilizes the learning, ensuring consistency and efficiency in the model's training. Furthermore, GPT-3.5 incorporates positional encoding, giving the model an awareness of the order of words in a sequence, which is essential for generating coherent and contextually appropriate responses. GPT-3.5 also employs several regularization techniques to maintain robustness and prevent overfitting, crucial for maintaining the model's performance over a wide range of conversational topics. Dropout is a primary technique used, which randomly deactivates a fraction of the input units during training, thereby encouraging a more generalized model performance. Weight decay, or L2 regularization, is another technique employed, adding a penalty to the loss function based on the size of the weights and encouraging the model to maintain smaller weights. Additionally, gradient clipping is used to prevent the issue of exploding gradients, a common challenge in training deep neural networks like GPT-3.5. This involves setting a threshold for the gradients during backpropagation to keep them within manageable limits.

3) *Hybrid Model of GPT 3.5 and RAG:*

In this project, we are using a hybrid model that uniquely combines the capabilities of a Retriever-Augmented Generator (RAG) with a fine-tuned version of GPT-3.5, offering a multifaceted approach to conversational AI. The RAG component is important in enhancing the chatbot's ability to pull relevant information from a knowledge base. This is achieved through a dense retriever that efficiently fetches data, embedding queries and documents in a high-dimensional space to determine relevance. Complementing this, the generator part of RAG, typically a Transformer-based sequence-to-sequence model, synthesizes the retrieved information to form coherent and contextually appropriate responses. Parallelly, the incorporation of GPT-3.5, renowned for its sophisticated language understanding and generation, adds a significant depth to the chatbot's capabilities. GPT-3.5 utilizes a deep Transformer architecture, characterized by layers of multi-head self-attention and feed-forward neural networks. By fine-tuning GPT-3.5 for specific conversational contexts or datasets, the chatbot gains an enhanced ability to generate relevant and accurate responses, tailored to the specific nuances of the intended interaction.

The hybrid model is constructed with various layers that contribute to its overall efficacy. The self-attention and feed-forward layers in GPT-3.5 are crucial for the nuanced processing and generation of text, enabling the chatbot to engage in natural and fluid conversations. In parallel, the embedding and dense retrieval layers in RAG play a critical role in information retrieval, ensuring that the chatbot's responses are not only contextually aware but also rich in information.

To maintain the robustness of this complex system, several regularization techniques were employed. Dropout is a key strategy used in both RAG and GPT-3.5 components, mitigating the risk of overfitting by randomly omitting units during the training phase. Layer normalization, particularly in the Transformer layers of GPT-3.5, is instrumental in stabilizing the training process, ensuring consistency in performance. Additionally, weight decay, or L2 regularization, is likely utilized in fine-tuning GPT-3.5, imposing a penalty on the loss function proportional to the weight magnitude, thereby fostering a preference for smaller weights.

IV. EXPERIMENTAL SETUP

The setup for the ChatBot, a high-level approach was adopted, leveraging the powerful meta-llama/Llama-2-7b-hf model and customizing it through fine-tuning with the targeted dataset vicgalle/alpaca-gpt4. This strategic choice aimed at enhancing the ChatBot's ability to handle diverse queries with precision. The experimental framework incorporated Low-Rank Adaptation (LoRA), an efficient technique for adapting large-scale models while conserving computational resources. Crucial hyperparameters, including a learning rate set at 2e-4, a batch size of 6, and a training duration spanning 20 epochs, were carefully selected to optimize the learning process and overall performance of the model. This experimental approach was centered around the goal of refining the ChatBot's response accuracy and relevance, thus elevating the user interaction experience and underscoring its practical utility in real-life applications.

1) *LoRA Configuration Elements:*
   a) *lora_alpha=8:* This parameter influences the rank expansion in adapted layers, enhancing the model's learning capacity for specific traits, and balancing between efficiency and model complexity.
   b) *lora_dropout=0.1:* Implementing dropout at a rate of 10% serves as a strategy to reduce overfitting, by randomly nullifying a portion of the neurons during the training phase.
   c) ***r=16:*** Indicates the rank in LoRA, dictating the additional parameters for model adaptation. This is a strategic choice that aims to augment the model's capabilities without excessively increasing its complexity.

2) *Targeted Modules in Transformer Layers:* The selected layers are crucial for focusing the training and adaptation process on key areas of the model.

3) *Key Training Parameters (TrainingArguments):*
   a) *num_train_epochs=20:* The training data is fully cycled through the model, ensuring adequate learning without overtraining.
   b) *per_device_train_batch_size=6:* Balances the number of training samples processed and the frequency of parameter updates, considering memory limitations.
   c) *gradient_accumulation_steps=2:* Enables handling of effectively larger batch sizes than hardware limitations would normally allow, by accumulating gradients over multiple iterations.
   d) *learning_rate=2e-4:* This critical parameter controls the adjustment size of the model weights with each update, balancing the speed of convergence with stability.
   e) *weight_decay=0.001:* A method to prevent overfitting by penalizing larger weights in the model, promoting simpler model solutions.
   f) *max_grad_norm=0.3:* Addresses the exploding gradient problem, a common challenge in deep learning, by clipping the gradients.
   g) *warmup_ratio=0.3:* Helps in stabilizing the training process initially by progressively increasing the learning rate.
   h) *lr_scheduler_type='linear':* This scheduler reduces the learning rate in a linear fashion, managing the rate of learning as training progresses.

4) *Optimization Strategy:*
   a) *optim='paged_adamw_8bit':* Chosen for its efficient memory usage, this variant of the AdamW optimizer is especially beneficial for training larger models.
   b) *Learning Rate (learning_rate=2e-4):* This particular learning rate was likely chosen to balance between efficient learning and avoiding the risk of bypassing the optimal solution. A moderately low rate helps in making precise yet significant updates during training. A range of learning rates have been tested, observing the model's response in terms of training stability and speed of convergence, to identify this optimal rate.
   c) *Batch Size (per_device_train_batch_size=6):* Opting for this batch size is a strategic decision influenced by hardware limitations and the desire for stable gradient updates. Smaller batch sizes often lead to more reliable learning patterns. The process included experimenting with various batch sizes, assessing their impact on both the computational efficiency and the consistency of the learning process.
   d) *Number of Epochs (num_train_epochs=20):* Setting the training to 20 epochs indicating a deliberate attempt to balance thorough learning with the risk of overfitting. The team has experimented with different epoch counts, closely monitoring for overfitting or underfitting, possibly using a validation set as a benchmark for performance tuning.

3) *Regularization Parameters (lora_dropout=0.1, weight_decay=0.001):* Implementing dropout and weight decay aims at curbing overfitting. Dropout randomly deactivated neurons, preventing dependency on specific paths, while weight decay limits the magnitude of the weights, promoting a simpler model structure. Various settings for dropout and weight decay likely underwent testing to find a configuration that best allowed the model to generalize from the training data.

4) *Optimization Algorithm (optim="paged_adamw_8bit"):* The selection of this optimizer underscores a priority for memory efficiency, crucial for managing large-scale models, while still leveraging the proven efficacy of the AdamW algorithm. The team have compared this optimizer with others, like standard Adam or SGD, to evaluate differences in training efficiency and final model performance.

5) *Foundation in AdamW:* As a derivative of the Adam optimizer, AdamW is favored for its adept handling of both large data sets and sparse gradients. Its ability to adapt the learning rate during training is particularly effective for complex model architectures. The 'W' in AdamW refers to its distinct approach to weight decay. This separation of weight decay from the gradient updates in AdamW often results in better regularization, enhancing overall model performance.

6) *Customized Variant - paged_adamw_8bit:* This variant is tailored for memory efficiency, a crucial factor when training expansive models like those in advanced ChatBots.
   a) *8-bit Precision:* The '8bit' component implies a shift to lower precision calculations. By using 8-bit instead of the standard 32-bit floating points, the optimizer reduces the memory demand significantly. This reduction allows for more efficient computations and lesser memory usage, with a minimal impact on the accuracy of the model.
   b) *Paging Mechanism:* The term 'paged' indicates the use of a paging system. It involves breaking down the model parameters or gradients into smaller segments or 'pages', enhancing memory management further.

7) Significance and Customizations: Opting for paged_adamw_8bit underlines a strategic approach to manage the model's complexity and scale efficiently. This optimizer choice enables effective training of a sizable model, even under constraints of computational resources. Its application is especially relevant in scenarios like fine-tuning large pre-trained models where adjusting a massive number of parameters could be computationally intensive. Employing 8-bit precision represents a key modification, striking a balance between computational efficiency and model performance. This practice is increasingly common in contemporary deep learning for training large-scale models, as it significantly eases computational requirements.

8) Hardware and Software requirements: Utilized an array of advanced hardware and software tools to achieve top-notch performance and effectiveness. The table below is a detailed outline of the hardware components utilized in the project

TABLE I. HARDWARE AND SOFTWARE REQUIREMENTS

Hardware Component	**Description**	**Memory and storage Insights**
CPU	High-performance CPU for complex computations and data processing	16GB, to smoothly handle the complex computations and data juggling.
GPU	To process intensive tasks like model training	For models like GPT-4, 16GB of GPU memory is used
RAM	Multitasking and sufficient memory to handle large datasets.	32GB, to efficiently manage multiple operations and large datasets simultaneously.
Storage	Adequate storage for datasets, processing files and model files.	Fast SSDs, with 1TB or capacity to store machine learning models.

TABLE II. DETAILED OUTLINE OF THE SOFTWARE COMPONENTS

Software component	**Description**	**Role in the project**
Python	The Core Language	Primarily used Python for all our coding needs.
Development Environment	The Coding Hub	Google collab for quick tests and for more intensive development.
Version Control with Git	Keeping Track of Changes	Enabling managing code revisions and collaborating.
Operating System	The Foundation	Project was built on an OS, macOS

Project's analytical capabilities were bolstered by integrating several leading-edge software libraries and frameworks, each selected for its stellar performance. The tools were chosen for their proven track records and were implemented in versions that guaranteed consistency and seamless operation throughout the project's development.

For an in-depth view of the versions we utilized and the functionalities they bring to the project, the following table lays out all the pertinent details.

TABLE III. VERSIONS AND FUNCTIONALITIES

Software / Library	**Version**	**Purpose**
Python	3.12.1	The primary programming language for implementing the project
Transformers	-	Access to various pre-trained models like GPT-4, LLaMA for NLP tasks
Torch (PyTorch)	2.0	Main library for deep learning applications
sentence-transformers	3.6	Utilized for generating sentence-level embeddings
datasets	-	Handling and processing large-scale datasets efficiently
faiss-cpu		Used for quick similarity searches and dense vector clustering
langchain	a Python version of ≥ 3.8. 1 and <4.0.	For constructing and managing chains of language models
openai		Integration with OpenAI's language models for NLP tasks
trl, xformers	0.0.16	For model training enhancements and transformer optimizations
wandb	0.16.1	For monitoring experiments and visualizing data
gradio	4.0.1	To create interactive web interfaces for machine learning models
Hugging Face's pipeline		Streamlines the process of applying pre-trained models for inference
tiktoken		Utilized for tokenization and processing of all the text data.
langchainhub		Managing and Integrating different components of the framework.
bitsandbytes	8-bit optimizers	Efficient training of large models.

peft	0.7.0	Efficient transformer training, offering optimizations like Lora

V. RESULTS AND ANALYSIS:

To evaluate the performance of chatbot, SAS (Semantic Answer Similarity) and BLEU (Bilingual Evaluation Understudy) are used. SAS (Semantic Answer Similarity) metric evaluates the semantic similarity between the predicted answer (chatbot response) and the ground truth (correct answer). It considers the meaning conveyed in the responses rather than just the exact word matches. BLEU (Bilingual Evaluation Understudy) is originally developed for machine translation, BLEU evaluates the quality of text which has been machine-translated from one language to another. It measures how many words and phrases in the chatbot's response appear in the correct answer, taking into account the order of words.

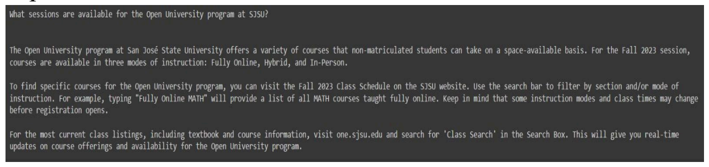
The below are the results which are observed for the models implemented.

TABLE IV. MODEL EVALUATION METRIC RESULTS

Model	**SAS Score**	**BLEU Score**
LlAMA2	0.33	0.41
GPT 3.5	0.57	0.60
GPT+RAG	0.79	0.76

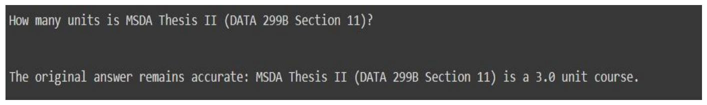
Apart from the metrics the chatbot's performance has been evaluated with the Customer Satisfaction Score (CSAT), we received direct feedback from users rating their satisfaction with the chatbot's responses when internally shared with the team members and peers in our class. This is crucial for student service and support ultimately.

Below are the responses from the chatbot when users have imposed



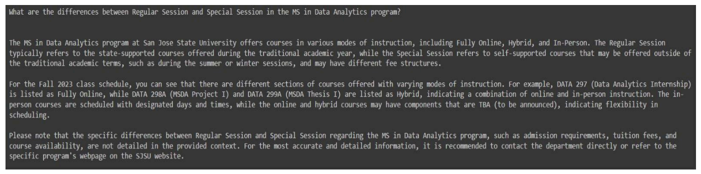
```
What sessions are available for the Open University program at SJSU?


The Open University program at San José State University offers a variety of courses that non-matriculated students can take on a space-available basis. For the Fall 2023 session,
courses are available in three modes of instruction: Fully Online, Hybrid, and In-Person.

To find specific courses for the Open University program, you can visit the Fall 2023 Class Schedule on the SJSU website. Use the search bar to filter by section and/or mode of
instruction. For example, typing "Fully Online MATH" will provide a list of all MATH courses taught fully online. Keep in mind that some instruction modes and class times may change
before registration opens.

For the most current class listings, including textbook and course information, visit one.sjsu.edu and search for 'Class Search' in the Search Box. This will give you real-time
updates on course offerings and availability for the Open University program.
```

Fig. 4. Question 1 and Response 1 of the resultant output.



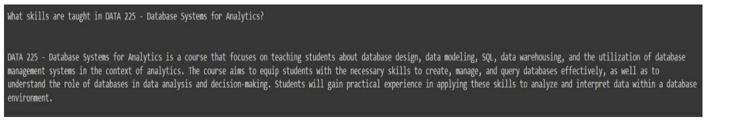
```
How many units is MSDA Thesis II (DATA 299B Section 11)?


The original answer remains accurate: MSDA Thesis II (DATA 299B Section 11) is a 3.0 unit course.
```

Fig. 5. Question 2 and Response 2 of the resultant output.



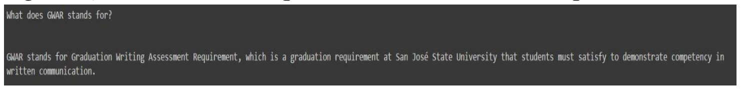
```
What are the differences between Regular Session and Special Session in the MS in Data Analytics program?


The MS in Data Analytics program at San Jose State University offers courses in various modes of instruction, including Fully Online, Hybrid, and In-Person. The Regular Session
typically refers to the state-supported courses offered during the traditional academic year, while the Special Session refers to self-supported courses that may be offered outside of
the traditional academic terms, such as during the summer or winter sessions, and may have different fee structures.

For the Fall 2023 class schedule, you can see that there are different sections of courses offered with varying modes of instruction. For example, DATA 297 (Data Analytics Internship)
is listed as Fully Online, while DATA 298A (MSDA Project I) and DATA 299A (MSDA Thesis I) are listed as Hybrid, indicating a combination of online and in-person instruction. The in-
person courses are scheduled with designated days and times, while the online and hybrid courses may have components that are TBA (to be announced), indicating flexibility in
scheduling.

Please note that the specific differences between Regular Session and Special Session regarding the MS in Data Analytics program, such as admission requirements, tuition fees, and
course availability, are not detailed in the provided context. For the most accurate and detailed information, it is recommended to contact the department directly or refer to the
specific program's webpage on the SJSU website.
```

Fig. 6. Question 3 and Response 3 of the resultant output.



```
What skills are taught in DATA 225 - Database Systems for Analytics?


DATA 225 - Database Systems for Analytics is a course that focuses on teaching students about database design, data modeling, SQL, data warehousing, and the utilization of database
management systems in the context of analytics. The course aims to equip students with the necessary skills to create, manage, and query databases effectively, as well as to
understand the role of databases in data analysis and decision-making. Students will gain practical experience in applying these skills to analyze and interpret data within a database
environment.
```

Fig. 7. Question 4 and Response 4 of the resultant output.



```
What does GWAR stands for?


GWAR stands for Graduation Writing Assessment Requirement, which is a graduation requirement at San José State University that students must satisfy to demonstrate competency in
written communication.
```

Fig. 8. Question 5 and Response 5 of the resultant output.

Though the chatbot is working as expected with good responses, it is adding on some extra content to get the context right which can be avoided and be specific to the question. The chatbot's performance can be further improved by giving the context of the entire university information rather than just one department which can help the users.

VI. CONCLUSION

In conclusion, this project has made significant contributions to the field of chatbot technology using Large Language Models (LLMs). The utilization of models like Llama-2-7b, GPT 3.5, and the hybrid model of GPT 3.5 and RAG has resulted in a sophisticated chatbot capable of handling complex queries in an academic setting, particularly for the San Jose State University Data Science Department. The integration of these models showcases an advanced understanding of natural language processing and its practical applications. The results, particularly the scores on SAS (Semantic Answer Similarity) and BLEU (Bilingual Evaluation Understudy) metrics reveal the chatbot's efficiency in providing relevant and contextually accurate responses. This is especially important in the academic domain, where the accuracy and relevance of information can significantly impact student learning and engagement. The high satisfaction ratings from internal testing further validate the project's success. Looking ahead, future work could focus on expanding the chatbot's knowledge base to encompass a wide range of university-related topics, not just limited to one department. This expansion would enhance its utility as a comprehensive resource for students. Additionally, refining the chatbot to minimize over-extended responses and improve its brevity could further streamline user interactions. Exploring more advanced NLP techniques and diversifying the dataset with more varied and complex queries can provide broader learning opportunities for the model, thus improving its adaptability and accuracy in diverse conversational contexts.

REFERENCES

[1] Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics. (2023). Journal of Applied Learning and Teaching, 6(2). https://doi.org/10.37074/jalt.2023.6.2.22

[2] Zhao, Z., Liao, L., Yang, D., & Nie, L. (2023). Building emotional support chatbots in the era of LLMs. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2308.11584

[3] Hicke, Y., Agarwal, A., Ma, Q., & Denny, P. (2023). ChaTA: Towards an Intelligent Question-Answer Teaching Assistant using Open-Source LLMs. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.02775

[4] Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., & Cui, L. (2023). LLM-empowered chatbots for Psychiatrist and Patient simulation: application and evaluation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.13614

[5] Huang, S., Li, P. L., Hsu, Y., Chen, K., Lin, Y. T., Hsiao, S., Tsai, R., & Lee, H. (2023). Chat Vector: A simple approach to equip LLMs with new language chat capabilities. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.04799

[6] Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of Retrieval Augmented Generation (RAG) models for open domain question answering. Transactions of the Association for Computational Linguistics, 11, 1–17. https://doi.org/10.1162/tacl_a_00530