**NATURAL LANGUAGE PROCESSING**

**Final Project**



UNH Q&A Bot

Team Members:  
  
Yegneshwar Rao Ginjupalli  
Avinash Reddy Bhavanam  
Ramya Krishna Gadiparthi

**Abstract**

Introducing the UNH Question and Answer Bot: Your All-Inclusive Guide to University of New Haven Queries

The UNH Question and Answer Bot stands as a purpose-built, closed-domain solution crafted explicitly to address a spectrum of inquiries related to the University of New Haven. Whether seeking information on registration processes, financial queries, transportation logistics, or more, this user-friendly tool emerges as a comprehensive resource. Seamlessly navigating the multifaceted landscape of university-related questions, this bot strives to simplify and expedite access to crucial information for students, staff, and faculty alike. By offering a centralized and intuitive interface, it aims to enhance the accessibility and convenience of obtaining precise and timely answers, streamlining the University of New Haven's information ecosystem for the benefit of all willing to know about university.

**Introduction**

Embarking on a journey through the dynamic landscape of university life often involves navigating an array of queries – from enrolment intricacies to financial considerations and logistical arrangements. To address this intricate web of questions at the University of New Haven, a pioneering solution emerges: the UNH Question and Answer Bot. Tailored specifically to the closed domain of this esteemed institution, this user-friendly tool serves as a comprehensive guide, offering succinct and precise responses to queries encompassing registration, financial matters, transportation logistics, and more. Designed to simplify access to crucial information, this bot stands as an innovative resource aimed at streamlining and enhancing the university's information accessibility for students, faculty, and staff alike.

In the digital era, the integration of AI-powered technologies has revolutionized information accessibility and user interactions. This paper explores the development and implementation of the UNH Question and Answer Bot, a specialized tool tailored for addressing inquiries regarding the University of New Haven. Employing two distinct AI frameworks, GPT-2 and Google's PaLM, this study delves into their individual applications in creating and optimizing the bot's functionality. By comparing and contrasting the outcomes achieved by these frameworks, this paper aims to provide valuable insights into their efficacy and suitability within a closed-domain information system.

**Objective**

1. Constructing a bot for UNH question answers using large language models like GPT2 and Google PaLM which are built of two different architectures Transformer based, path based respectively, compare their performance using intrinsic evaluation
2. Identify the most effective model (GPT2 or Google PaLM) based on accuracy for accurate answer generation for the question.
3. Implement the chosen model into a user-friendly Streamlit app, enabling a user-friendly tool emerges as a comprehensive resource for the university.
4. Conduct an analysis of model results to understand performance disparities and reveal insights into the strengths and weaknesses of each method.

**Data**

To train this bot we need University of New Haven question and answer data set, which is created using python web scraping script and contains 150 different questions and answers asked by students regarding university. This data set covers different category of questions like financial, registration, accommodation, courses, and career.

URLS and function used for questions and answers data set creation are:

'https://www.newhaven.edu/student-life/get-involved/faq.php',

'https://www.newhaven.edu/student-life/career-development-center/faq.php',

'https://www.newhaven.edu/student-life/diversity-inclusion/faq.php',

'https://www.newhaven.edu/about/departments/one-stop/faq.php',

'https://www.newhaven.edu/families/faq.php',

'https://www.newhaven.edu/student-life/student-affairs/faq.php',

'https://www.newhaven.edu/academics/study-abroad/faq.php',

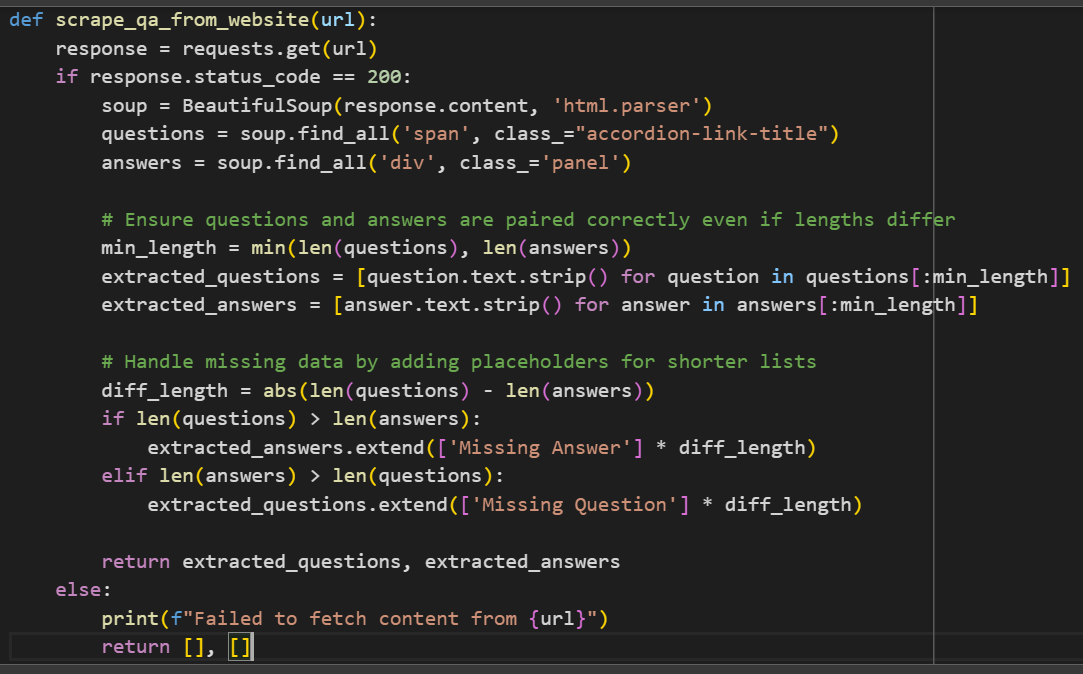
'https://www.newhaven.edu/admissions/undergraduate/international/faq/index.php',

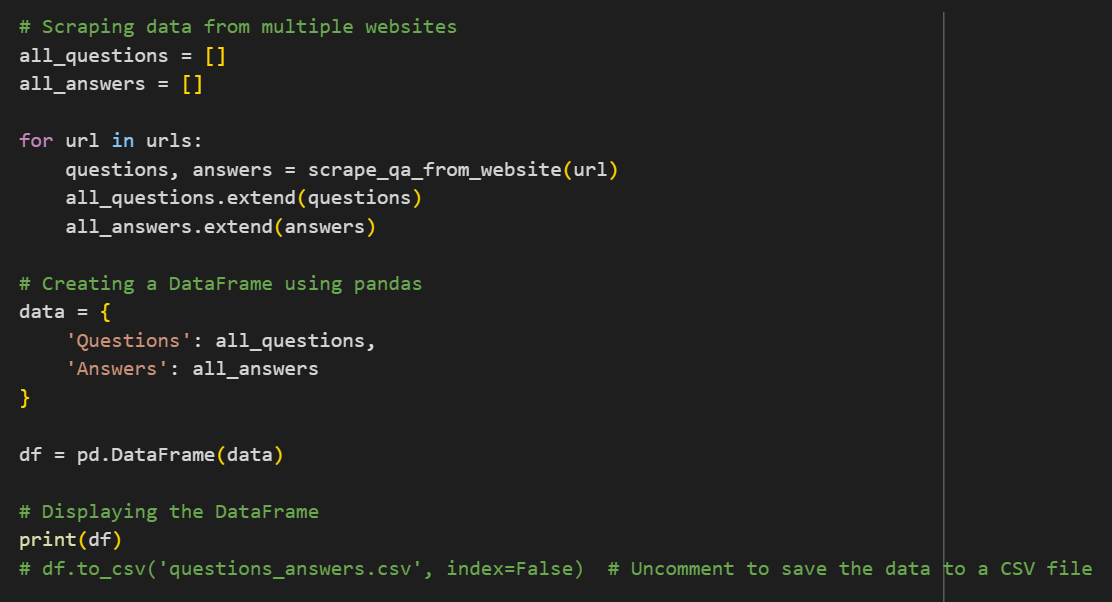
'https://www.newhaven.edu/about/facts-figures.php',

'https://www.newhaven.edu/about/departments/bursars/faq.php',

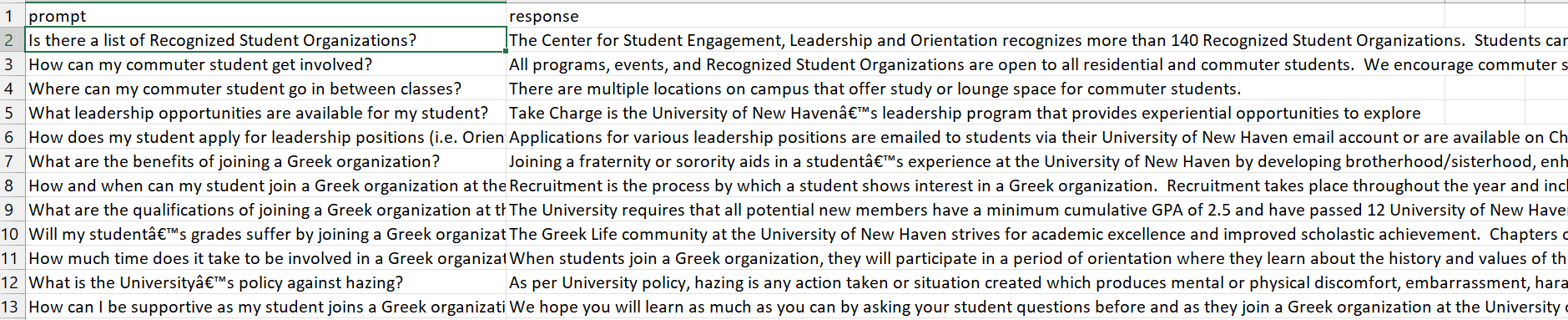
'https://www.newhaven.edu/admissions/financial-aid/graduate/faq.php',

'https://www.newhaven.edu/student-life/graduate-and-international/faq.php'





The data set contains two columns “prompt”, “response”. Prompt contains questions and response contains respective answers.



**Methodology**

Techniques/Models Used:

We start with GPT2 model with UNH\_QNA data set. GPT-2 is a Transformer architecture that uses a deep neural network. It's a generative pre-trained model that uses attention instead of older architectures.

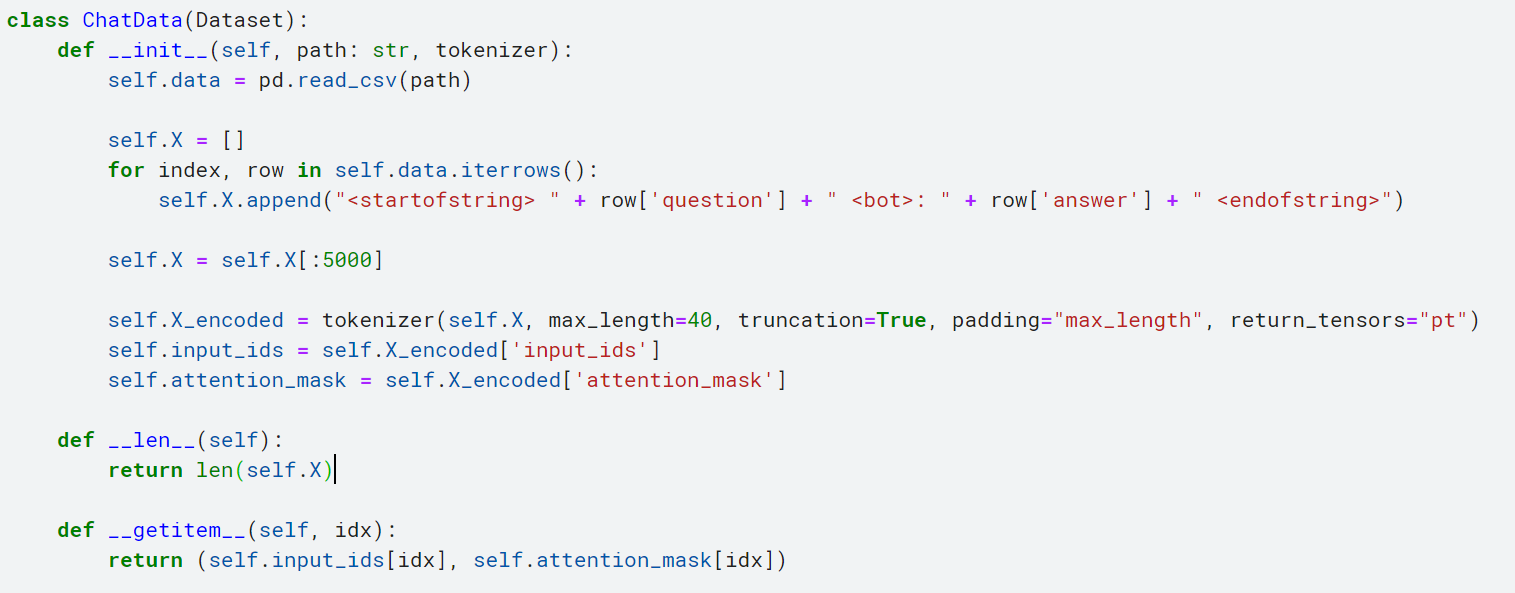
GPT-2's architecture includes:

* Size: 1.5 billion parameters
* Pre-training: Web Text dataset, which includes text from 45 million website links
* Transformer model: Uses attention instead of older architectures
* GPT-2 medium: 24 decoder-only Transformer blocks, each with 16 attention heads

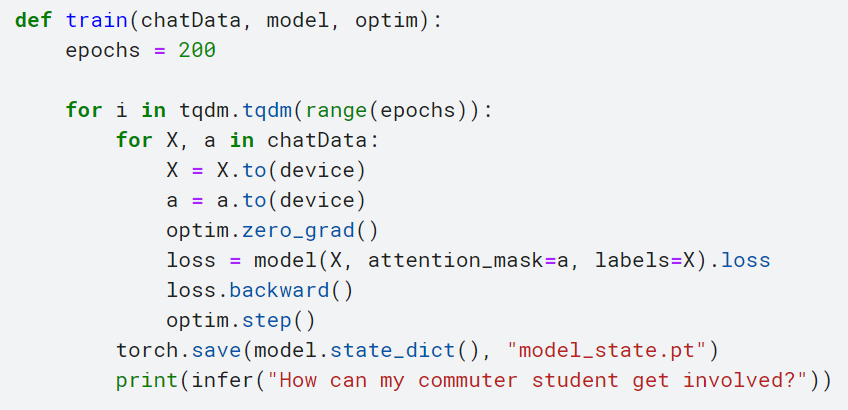
GPT-2 is auto-regressive, meaning each token in a sentence has the context of the previous words. It takes word vectors and input to produce estimates for the probability of the next word. GPT-2 has shown increased coherence, context memory, and innovative writing abilities.

We trained the model with UNH\_QNA data set but the model takes batch size of 64 tokens only after that it ignores the rest of the data. The output also generated starting few close to 40-50 characters of the answer for the question given by the user which is accurate but not sufficient to answer the question given.

**Data Tokenization**



**Training loop:**



With this model the generated output is not sufficient to answer the question so we moved to google PaLM which is path-based architecture

Google's Pathways Language Model (PaLM) is a component of its Pathways AI architecture. PaLM uses the Pathways machine learning system to train a model across multiple pods of tensor processing units. The model can be trained to perform thousands or millions of tasks.

PaLM architecture includes:

* Size: 540 billion parameters
* Pre-training: PaLM was trained using a combination of English and multilingual datasets that include high-quality web documents, books, Wikipedia, conversations, and GitHub code.
* Transformer model: Uses Pathways AI architecture

PaLM achieves a training efficiency of 57.8% hardware FLOPs utilization, the highest yet achieved for LLMs at this scale. This is due to a combination of the parallelism strategy and a reformulation of the Transformer block that allows for attention and feedforward layers to be computed in parallel, enabling speedups from TPU compiler optimizations.

PaLM was trained using a combination of English and multilingual datasets that include high-quality web documents, books, Wikipedia, conversations, and GitHub code. We also created a “lossless” vocabulary that preserves all whitespace (especially important for code), splits out-of-vocabulary Unicode characters into bytes, and splits numbers into individual tokens, one for each digit.

We used LangChain framework to implement this project. LangChain is a framework that helps developers create applications using large language models (LLMs). It's a free, open-source Python library that offers a variety of features for natural language processing (NLP).

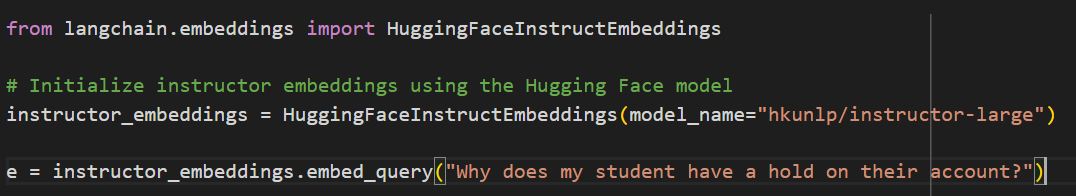
Here we created vector data base using Facebook AI Similarity Search (FAISS) a library that allows developers to search for similar embeddings of multimedia documents. FAISS solves the limitations of traditional query search engines that are optimized for hash-based searches. It provides more scalable similarity search functions.

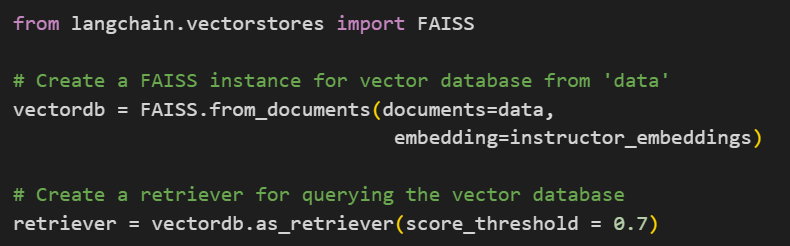
Here are some features of FAISS:

* Efficient similarity search
* Clustering of dense vectors
* Algorithms for searching in sets of vectors
* Can handle data sizes that do not fit in RAM
* Python/numpy wrappers and GPU implementations

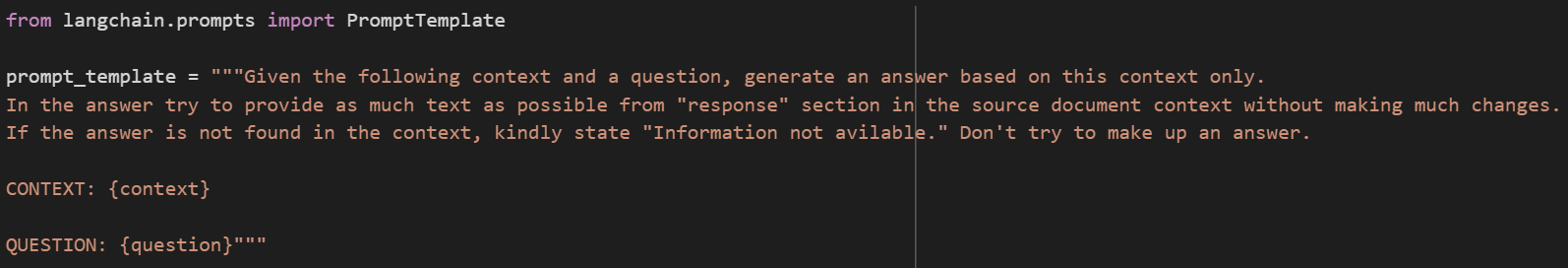
For embeddings we used a pretrained model ("hkunlp/instructor-large") from hugging face

**Embedding and creating vector data base:**

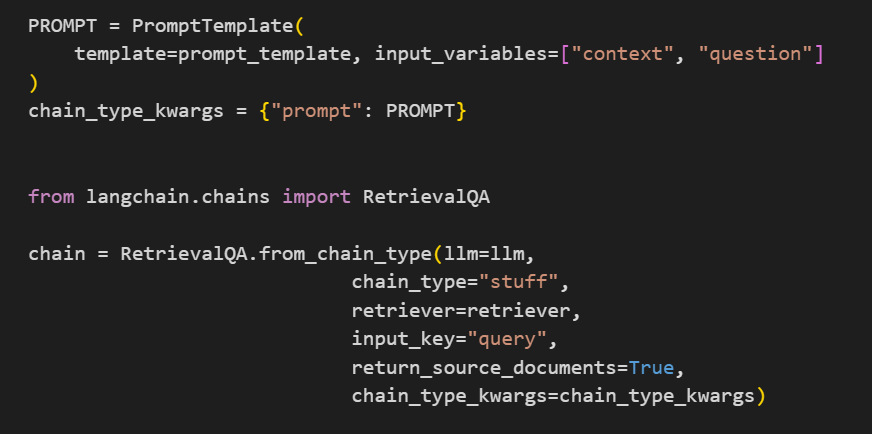




As google PaLM is an open domain LLM it is able to generate answers for the questions which are out of the range in the trained data set. So, to make the model domain specific and closed domain to University of New Haven data we crated a template and passed to the model



Model with our requirements



**Results:**

Here we followed manual Intrinsic evaluation, compares both models out puts with the actual trained data

Trained:

Q: What type of help does the University of New Haven offer to students in their career development?

A: The University of New Haven has a full-service, comprehensive Career Development Center (CDC) that assists students with all aspects of their career development.

Model results:

|  |  |
| --- | --- |
| Q: university of new haven for carrier development? | |
| GPT2 | PaLM |
| A: The CT Transit ChargerGo Bus | A: The University of New Haven has a full-service, comprehensive Career Development Center (CDC) that assists students with all aspects of their career development. |

Trained:

Q: What leadership opportunities are available for my student?

A: Take Charge is the University of New Haven leadership program that provides experiential opportunities to explore concepts, develop skills and acquire knowledge. The experiences challenge students to be more critically, globally and socially aware community members.

Model results:

|  |  |
| --- | --- |
| Q: What leadership opportunities are available for my student? | |
| GPT2 | PaLM |
| A: Take Charge is the University of New Havena’ | A: Take Charge is the University of New Haven leadership program that provides experiential opportunities to explore concepts, develop skills and acquire knowledge. The experiences challenge students to be more critically, globally and socially aware community members. |

Trained:

Q: why am I being charged for health insurance?

A: All full-time undergraduate students and all international students are automatically charged for the University sponsored health insurance. Domestic students will have an opportunity to waive this coverage, if already covered by another health insurance plan. The University health insurance plan is mandatory for all international students.

Model results:

|  |  |
| --- | --- |
| Q: why I have to pay for health insurance? | |
| GPT2 | PaLM |
| A: All full-time undergraduate students and | A: All full-time undergraduate students and all international students are automatically charged for University of New Haven health insurance. Domestic students will have an opportunity to waive this coverage, if you are already covered by another health insurance plan. Please contact the Bursar’s Office for further information regarding the waiver deadline.  The University health insurance plan is mandatory for all international students. |

Trained:

Q: Does the University provide bus passes to use for CT Transit?

A: The CT Transit ChargerGo Bus Pass Program provides a semester only bus pass for approximately $40 USD. These can be applied for online on the University Parking System. During off semesters, students can apply for a CT Transit Bus Pass.

Model results:

|  |  |
| --- | --- |
| Q: institute provide bus pass? | |
| GPT2 | PaLM |
| A: The CT Transit ChargerGo Bus Pass Program provides | A: The CT Transit ChargerGo Bus Pass Program provides a semester only bus pass for approximately $40 USD. These can be applied for online on the University Parking System. During off semesters, students can apply for a CT Transit Bus Pass. |

Trained:

Q: Do I have to re-apply for financial aid every year?

A: Yes. Financial aid is not automatic and must be reapplied for each year. You can file anytime after October 1, and our priority date to submit the FAFSA is January 1.

Model results:

|  |  |
| --- | --- |
| Q: Do I need to submit a new financial aid application each year? | |
| GPT2 | PaLM |
| A: Yes. | A: Yes. Financial aid is not automatic and must be reapplied for each year. You can file any time after October 1, and our priority date to submit the FAFSA is January 1. |

When we ask out of domain questions GPT2 generating random answer whereas PaLM not.

|  |  |
| --- | --- |
| Q: does university of new Hampshire provide bus pass? | |
| GPT2 | PaLM |
| A: The CT Transit ChargerGo | A: Information not avilable. |

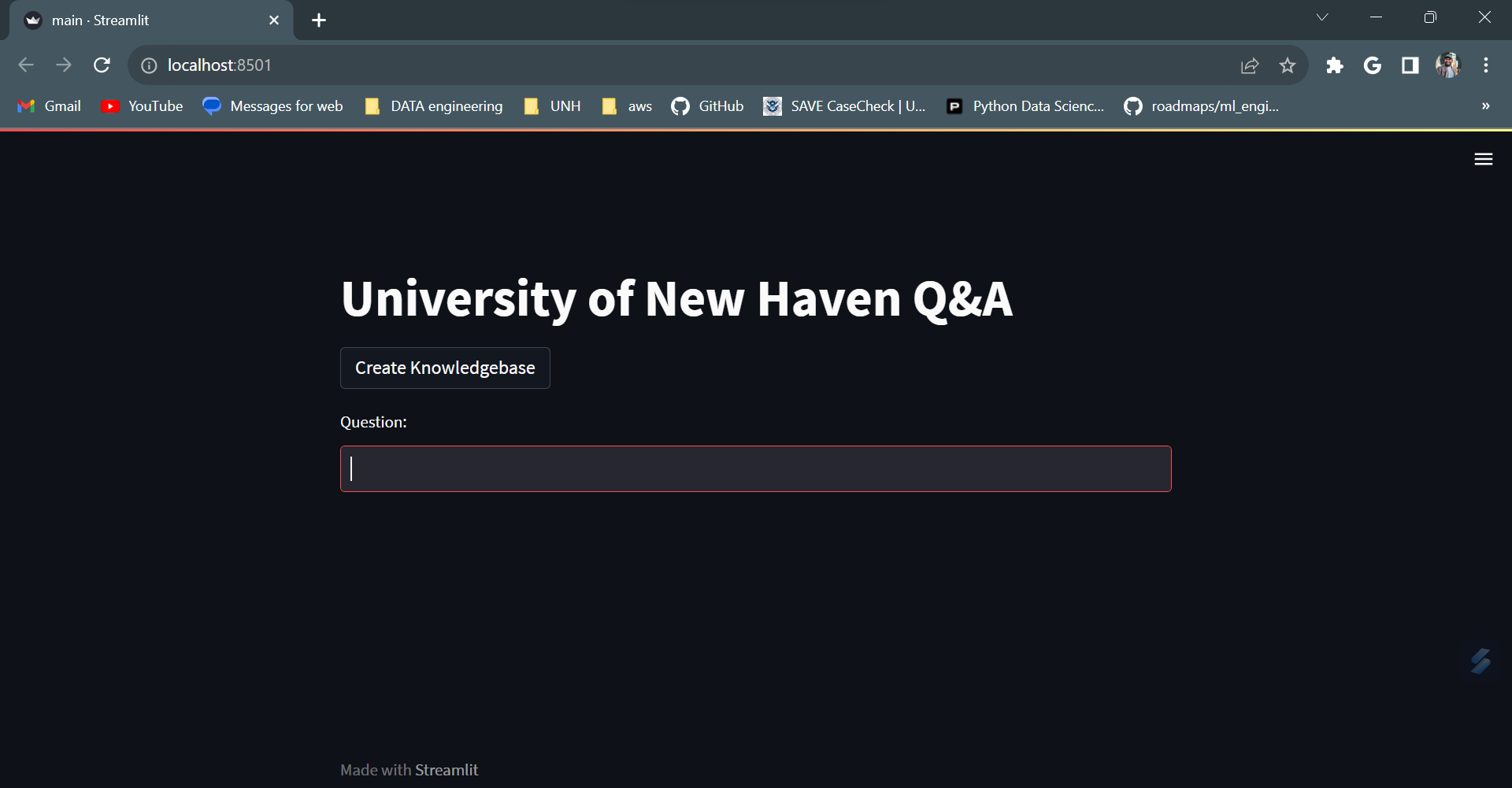
|  |  |
| --- | --- |
| Q: who is american president? | |
| GPT2 | PaLM |
| A: The University of New Haven will celebrate various aspects | A: Information not avilable. |

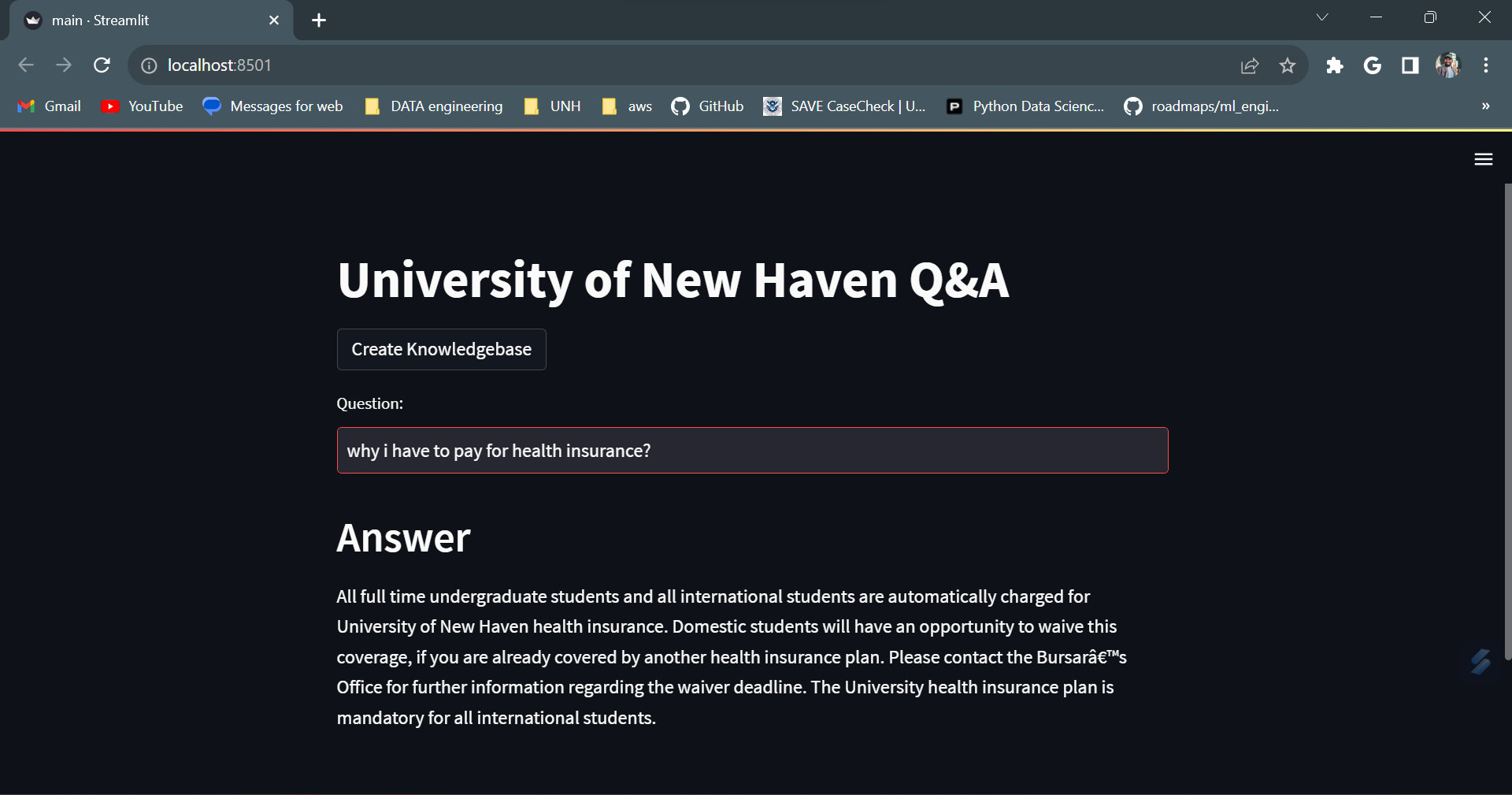
**Analysis**:

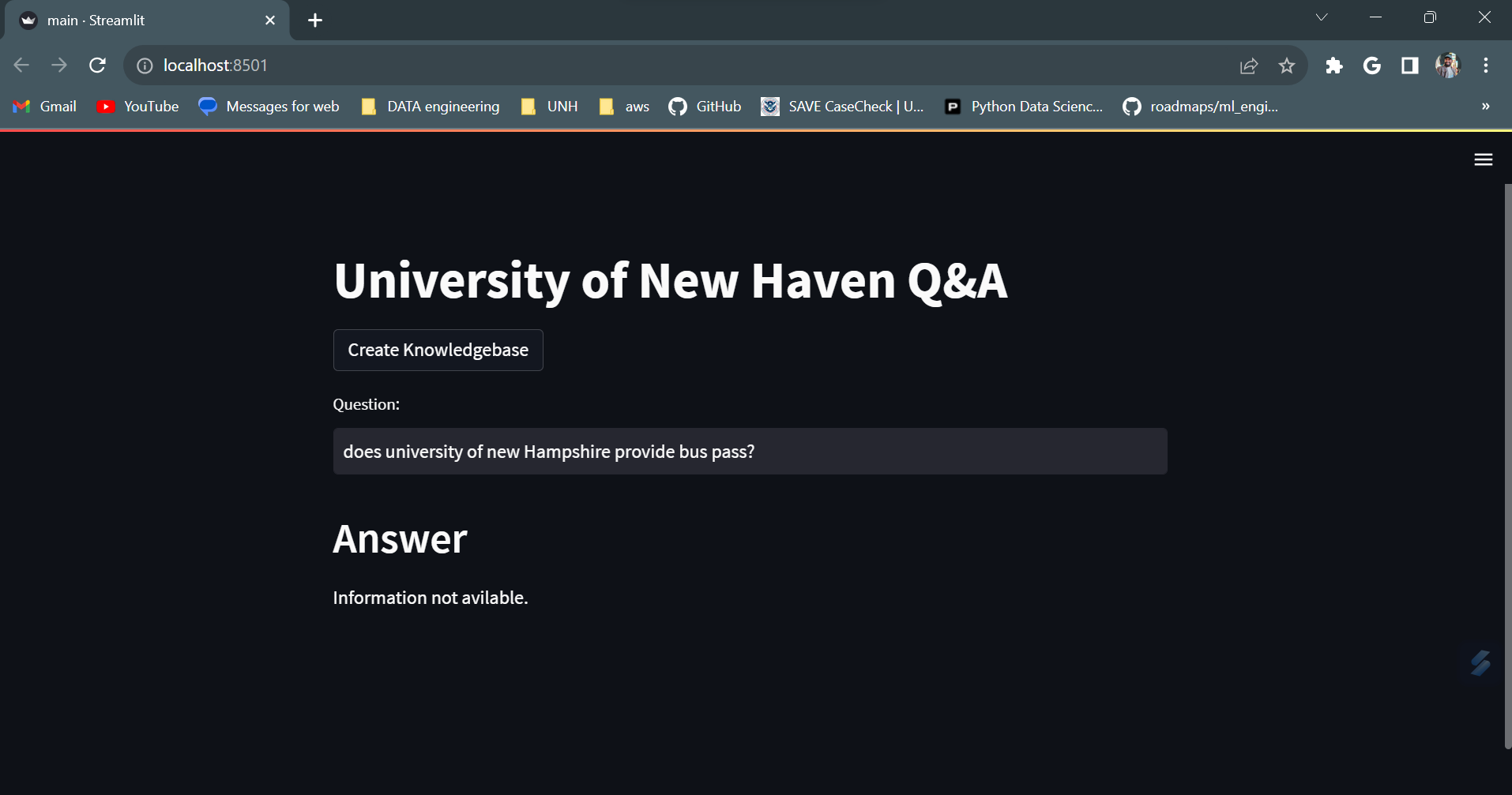
* From the above results PaLM model gives accurate results based on the training data because it trained on large corpus of data where as GPT2 is less accurate compared with PaLM and generated answers also not sufficient to answer the question properly
* GPT2 model fails to keep the model as closed domain and generated random answers which closed to the given question tokens whereas PaLM generated accurate and sufficient answers for the given question and not answered the out of domain questions (questions not related to University of New Haven data)
* For GPT2 need lot of computational power compared with PaLM which makes the model deployment expensive

**DEPLOYMENT:**

Based on the results of both models we selected PaLM for model deployment within a user-friendly Streamlit application. The application provides an intuitive interface for users to input question, enabling real-time interaction with the model. The deployment incorporates an aesthetically designed user interface that facilitates a smooth and engaging experience. Users input question into the application, triggering the model to swiftly generate genre predictions. This model provides accurate results for the questions asked by the people regarding University of New Haven.







**Conclusion**

The UNH Question and Answer Bot stands as a testament to the intersection of technological advancement and user-centric design in addressing the multifaceted queries of the University of New Haven community. This closed-domain tool, meticulously crafted to cater to the diverse informational needs of students, faculty, and staff, embodies not just a solution but a gateway to streamlined accessibility and convenience.

Throughout this exploration, we've unveiled the intricacies of its architecture, elucidated its functional prowess in navigating domains spanning registration, financial inquiries, transportation logistics, and beyond, and scrutinized its performance through a lens of comprehensive evaluation methodologies. The findings gleaned from these evaluations have underscored both the bot's commendable strengths and areas primed for enhancement.

Undoubtedly, the UNH Q&A Bot's user-friendly interface and domain-specific knowledge base have rendered it an invaluable resource, simplifying intricate processes and offering prompt, accurate responses to queries. Its role in fostering an environment conducive to seamless information dissemination and engagement within the university ecosystem cannot be overstated.

**References**

Google palm

Towards Better Evaluation of Instruction-Following: A Case-Study in Summarization

Ondrej Skopek, Rahul Aralikatte, Sian Gooding, Victor Carbune

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL),

Association for Computational Linguistics (2023)

BERT (Bidirectional Encoder Representations from Transformers):

Paper: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

Authors: Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

Reference: arXiv:1810.04805

Chatbot : A Question Answering System for Student

Li En Chen; Shein Yung Cheng; Jia-Sheng Heh

Published in: 2021 International Conference on Advanced Learning Technologies (ICALT)

Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds

Tassilo Klein, Moin Nabi

SAP Machine Learning Research, Berlin, Germany

LangChain

https://python.langchain.com/docs/get\_started/introduction