ProtonAI – LLM based approach for question answer automation

Gajawada Vishwas
Computer Science and Engineering
Anurag University
Hyderabad, India
gajawadavishwas@gmail.com

Racharla Sai Chaitanya
Computer Science Engineering
Anurag University
Hyderabad, India
racharlasaichaitanya@gmail.com

Molugu Abhinav
Computer Science Engineering
Anurag University
Hyderabad, India
abhinavmolugu659@gmail.com

Abstract-

The rapid evolution of artificial intelligence has led to groundbreaking advancements in natural language processing, with question and answer (Q&A) systems emerging as a focal point of innovation. This research delves into the development and implementation of a cutting-edge Q&A model empowered by generative AI techniques, aiming to enhance conversational interactions between users and machines. The proposed model leverages the power of Generative Pre-trained Transformers (GPT), specifically GPT-3.5 architecture, to comprehend and respond to user queries dynamically and the newly emerged technology of LLM's for retrieval of the data at its finest. To evaluate the efficacy of the proposed Q&A model, comprehensive benchmarking against existing state-of-the-art systems is conducted. Metrics such as accuracy, response coherence, and contextual relevance are employed to assess the model's performance across a spectrum of use cases. Furthermore, real-world deployment scenarios are explored to gauge the adaptability and user-friendliness of the generative AI-driven Q&A system. This research significantly contributes to the ongoing discourse surrounding conversational AI by presenting a sophisticated Q&A model that harnesses the potential of generative AI. This study has important implications for various applications like virtual assistants and educational platforms. These applications rely on sophisticated interactions that understand context for user satisfaction. The findings from this research will likely influence the future of Q&A systems, laying a strong groundwork for the ongoing improvement of conversational AI technologies to meet the increasing needs of diverse user interactions.

Keywords - **LLM, Generative AI, GPT-3.5 Architecture, Educational Platform, Virtual Assistance, Chat-with—your-data.**

1. INTRODUCTION

Exploring the realms of intelligence development, technology plays a pivotal role in generating questions and answers. In today's interconnected world, where information is readily available on the internet, technology utilizes user input to create questions and answers based on book knowledge.

Embarking on This research explores Large Language Models (LLMs) in conjunction with Langchain libraries and the influential Generative Pretrained Transformer (GPT). The primary objective is to elucidate complex systems and their potential in question and answer generation.

Langchain serves as a crucial tool in this endeavor. It offers a structured framework specifically designed for interacting with LLMs, enabling us to leverage their capabilities for question and answer tasks. By utilizing Langchain's functionalities, we aim to simplify the process of crafting effective and informative questions that can elicit insightful and comprehensive answers from LLMs.

Subsequently, this paper outlines the journey from the collection of the data, pre-processing the data where the Langchain's libraries utmost efficiency is used to implement deduplication and get error free data. This enhances the models ability to generate a response in a more accurate and helpful manner.

Along with the usage of Langchain as it cannot become an AI on its own we combined it with the LLm's which will help the it to get more accurate and work efficiently. The chromadb vectore-stores part is used to store the data of the text-book knowledge and retrieve the answers based on the user query.

The data is store in the format of vectors which the AI can actually understand the RAG(Retrieval Augmented generation) plays a major role in this which is the efficient manner to retrieve the data even from the images as its summaries.

RAG works by combining the strengths of LLMs with information retrieval techniques. When a user poses a question or provides a prompt, the system first taps into an external knowledge base, which could be a vast collection of text and code, a specific database, or even web documents. This retrieval component identifies information relevant to the user's input

In essence, this research paper navigates the fascinating intersection of language, technology, and innovation. It extends an invitation to explore the uncharted realms of AI, where questions are not merely answered but crafted through an artistic blend that blurs the boundaries between human and machine interactions.

2. LITERATURE REVIEW

Matt Dunn, Levent Sagun, Mike Higgins, Volkan cirik[1] these authors introduced a new dataset called SearchQA, designed for machine understanding and question answering. Unlike previous datasets, it reflects the full channel of question answers by starting with existing question-answer pairs from Jeopardy! and their extension with text snippets obtained from Google. The result is over 140,000 question-answer pairs, each with an average of 49.6 snippets. The dataset contains metadata such as snippet URLs, making it a valuable resource. The human-machine evaluation demonstrates a significant performance gap, indicating the potential of SearchQA as a benchmark for question-answering research. The dataset and evaluation code are publicly available. Dubey Harish [2]Automatic Question Paper Generation System by Dubey Harish, Tamora Hardik, Padhi Sagar, Manish Bharam is a Python based solution using fuzzy logic for efficient and customized question paper generation. The survey examines related works such as adaptive question bank systems, fuzzy logic based models, genetic algorithms for test question generation, and Android applications for paper generation. The proposed system excels in its reliability, deduplication and ease of use and provides a valuable contribution to automated paper question generation.

Mandar Joshi's [3] approach incorporates the TriviaQA dataset, which contains more than 650,000 question-answer triplets resulting in reading comprehension challenges. Known for its compound questions, keywords and phrases, and the need for inter-sentence reasoning, TriviaQA includes 95,000 word question pair studies by unsatisfied lovers. Basic algorithms, including feature-based classification and neural networks, outperform humans (80% versus 23% and 40%). The uniqueness of this collection lies in its usual appearance, good questions, self-writing and taking evidence from Wikipedia and the Internet. TriviaQA serves as a rigorous test bed for reading comprehension models, has challenges beyond those posed by other big data sources, and deserves major attention in future research.

Jonathan Berlant [4] proposed a new way to understand words without using complex rules. Its system does not need detailed instructions; instead, he learns through questions and answers. This is very important for large files. Berant's method outperforms other data methods and shows that it can work well even without specific guidance. Overall, this is a step forward in enabling machines to better understand language at scale. Pascale Fung and Andrea Madotto [5] from the Hong Kong University of Science and Technology are investigating a problem called "illusion" in computer-generated language. This occurs when the computer system produces incorrect information and may have real-life risks such as medical use. Writing articles, creating dialogues, etc. They learned how this problem arose in various events. The survey also discusses the different types of hearing loss and the problems researchers face in solving this problem. The goal is to ensure that text on the computer is reliable and does not cause harm or privacy issues.

Jared Kaplan and Benjamin Mann[6] present a survey of recent advances in natural language processing (NLP), specifically using large language models such as GPT-3 with 175 billion parameters. Unlike traditional methods that need to be optimized for specific tasks, GPT-3 shows significant improvements in performance from small samples to the ability to tune to good standards regardless. The challenge of needing large-scale domain data for each task is also widely recognized in education, as is information about solutions such as meta-learning, where the model acquires a wide range of skills to quickly adapt to new tasks during inference. The discussion also highlights the effectiveness of large-scale modeling in the use of contextual data and highlights the potential social implications of progress by addressing the limitations of current NLP techniques.

Rohan Kumar's [7] approach involves addressing the challenge of a long-term knowledge base, which is rare in open source question answering (QA), by proposing an automated method to generate custom QA documentation for competitive products. While large language models (LLMs) excel at understanding, they struggle with rare sites. The author uses the Wikipedia knowledge graph to interpret the content as a complete academic record showing the distribution of various factors. Automated design processes face challenges such as high learning, management complexity, and granularity issues. This study evaluates the performance of GPT3 on new long-term QA documents created and researched using external sources such as Wikipedia and knowledge maps. This study aims to support further research on the use of QA dataset generation and improve the performance of LL.M. have long experience. Woman. Khushbu Khandait[8] started a system that asks itself questions based on what students think and say. This new approach stands out by improving learning by asking good questions, saving time and making learning more personal. Seems like a useful tool for teachers and students. Research shows that with educational support and how to use effective strategies like Transformers, problems get better over time. Woman. Khandait's system even raised questions, and those who tried it said it worked to improve their scores. Overall, it's a clever way to make learning fun and personal.

Nithya M,Sanjeev Pranesh B[9] Participatory approach, how automatic question generation (AQG) can help teachers design effective questions. It emphasizes the use of techniques such as genetic algorithms, modified question banks and fuzzy logic. Artificial intelligence, especially natural language processing, is an important part of this. There are two ways to create questions: Use OCR to create questions from text or images. Workflows include tools like Anaconda Navigator and Spyder. Overall AQG makes it faster, more accurate and reduces student work.

Puneeth Thotad's [10] approach includes a question generator that uses natural language processing (NLP), a tool for quickly generating questions with many options from text. The system makes measuring comprehension easier by helping teachers and students measure comprehension. Leveraging Python programs with NLP libraries such as SpaCy and NLTK, the tool can process text, extract important data, and create queries. The research presented a variety of methods, including formal structure, keyword extraction, and Bloom's classification based on question design. The planning process will involve a series of algorithms that use NLP to analyze the input and generate different queries. The architecture process includes tools to provide teachers and students with good access and problem solving. Quizzes and performance tests show success in multiple choice, fill-in-the-blank, and Boolean-type questions from input. The system was designed to reduce manual work when creating tests and provide a foundation for future expansions, such as clarifying questions and analysis of answers.

Hala Abdel-Galil, [11] The involvement of Automatic Question Design (AQGM) is proposed in this work and is designed to assist students and teachers in exams. In an age where online education has gained importance, AQGM offers a solution that can generate questions from users using a GUI system. It simplifies the process for professors and teachers, saving time and effort when creating tests of various levels of complexity. The model was developed using deep learning methods such as segment-to-segment coupling with encoder-decoder and trained using SQuAD dataset, achieving a good

BLEU4 score of 11.3. This research highlights the importance of asking questions in learning, and AQGM's interactive interface allows users to design questions that suit their needs. The article concludes with plans for the future, including adding more problem types and moving from auditing to advanced training to improve performance.

Shivani G. Aithal [12] method discussed the limitations of question answering system (QAS) and proposed a solution called similar question. Despite the rapid increase in knowledge, QAS still struggles to use common sense and judgment, resulting in incorrect answers to irrelevant or irrelevant questions. The conceptual framework compares questions to questions generated from statements, providing similar questions to identify unanswered and irrelevant questions. This humanistic perspective helps QAS avoid being asked such questions, thus improving the focus on answerable questions. This article also introduces an application that creates questions and answers from text, which can be used for many purposes. The approach is designed to improve QAS performance and can be integrated with existing systems. This study ends with the author's knowledge and participation, stating the availability of data and numbers, declaring that there is no conflict of interest, and informing the editor of the media open access license.

Rohan Bhirangi's [13] research method changes from traditional writing to the process of creating surveys for schools. The fact that the available information is related to the study of the book causes bias and competition between issues. Limitations include the risk of security breaches and data transfer issues. The proposed automated system solves these problems by using a role-based hierarchy and mixed randomization algorithm. The goal is to increase efficiency, reduce bias, and increase the security of tests. The survey reviewed various existing survey designs, highlighting the importance of automation, information and communications technology (ICT) and access control systems. The proposed system improves speed, efficiency, and security through its role-based hierarchy and random query selection algorithm. Tejas Chakankar's[14] approach to search engines can ask questions based on collected data to help people learn better. This science is evolving rapidly and there are many ways machines can do this. Some methods use predetermined rules such as phrases and grammar, while others learn from examples and use this information to create questions. Researchers have tested these methods, and some have used deep neural networks or additive learning to improve the problem. They check these systems by asking people if the questions are good and useful. Although this has been done, challenges remain, such as asking questions about the appropriate problem, handling questions with more than one correct answer, and asking different questions and languages. Scientists are still working to improve these systems and find new ways to use them.

Aanchal Jawere's approach [15] focuses on making it easier for teachers to create test texts. In the past, researchers have attempted to develop automated systems such as AUTOQUEST to help solve this problem. These machines use many methods to make the process quicker, faster and safer than doing it manually. The purpose of this is to save teachers time and make the whole process smoother. Aanchal's approach uses natural language, such as teaching computers to understand and use words, to generate questions that are quick, versatile, and unbiased. Overall, this is to accomplish the task of creating better and more reliable tests.

Bidyut Das [16] method tells us how people learn online using computers or other devices. He talks about three ways to learn online: reading, watching images or videos, and listening to audio. We need questions and assessments to check how well people are learning. The article notes that there is no way to measure how much people learn by reading information online. Therefore, it is recommended to use automated systems to create questions and evaluate answers. The primary purpose of this article is to review existing research, summarize the data, and discuss the process of developing questions and measuring answers. He also touched on the challenges, concluding by saying that online learning is growing and automation tools can help make it easier.

Abishek B. Rao's work [17] is about making the question-answer system (QAS) better, especially since we are currently dealing with so much data. Sometimes these systems can give wrong answers due to lack of understanding. To solve this problem, Rao proposes a similar problem. Before the question is sent to QAS, it is compared to possible questions in the description and scored. This helps filter out questions that are not well answered. The idea is to improve QAS performance by focusing on questions that can be answered correctly. This study also demonstrates a useful application of creating question-answer pairs from an article, which is useful in many fields. This article discusses the history of QAS, from early techniques to recent developments such as BERT, and demonstrates the importance of natural language processing (NLP) in enabling machine learning. heart is better than words. Overall, our aim is to make QAS standards work better and the application process is a new way to facilitate this development. The Safnah Ali model [18] shows us how design model such as Generative Adversarial Networks (GAN) can be used in social media, especially with children. It discusses ethical issues related to technologies such as deepfakes, which create fake content that resembles the real thing.

The author believes that there are currently no efforts to teach intellectual skills to secondary school students. They propose a Learning Path (LT) that includes educational materials and activities focused on GANs, machine-generated news, and ethical implications. These activities were tested in an online workshop with 72 students, and the results showed that the materials helped children understand the designs, their applications, and the thinking process. This article presents a study to improve the teaching of complex sociotechnical processes and highlights the importance of artificial intelligence education for students. Dr. Manoj Kumar's [19] research focuses on text generation using computer models such as simple recurrent neural networks (RNN), short-term neural networks (LSTM), and gated recurrent units (GRU). He tested models from the Cornell University Film Archive to see which model performed better. This work explores the history of chatbots by discussing old techniques and new technologies such as bi-directional LSTM and tracking techniques. Dr. Kumar built and trained the model using

PyTorch on Google Colab and listed the hardware used in the process. The paper concludes by comparing the GRU model with the LSTM-based model and suggests ways to further improve the text.

Philipp Hacker [20] discusses the need for regulations such as ChatGPT and GPT-4 to address the challenges of Large-Scale Artificial Intelligence (LGAIM) in the European Union (EU) and elsewhere. While current regulations include general intelligence, these large models have unique capabilities and risks that require specific regulations. The report proposes a three-phase plan for this model, including a basic model, additional rules for high-risk situations, and collaboration in the AI process. It also recommends adding more comprehensive anti-discrimination laws and content controls. The authors argue that existing laws, such as the EU Artificial Intelligence Act and the Digital Services Act, should be amended to address the specific challenges of LGAIM and ensure their responsible use for the benefit of people.

Vahid Ashrafimoghari [21] approaches how smart computers, especially smart and word-generating ones, can help people prepare for the big GMAT exam, which is important for business school admission. Researchers tested seven different computers to see how they performed compared to humans. They found that the program, called GPT-4 Turbo, performed very well, even better than most people had tested it with. The research also discusses how these services can interpret responses, measure responses, and even aid learning. Although this is exciting, researchers say we need to be careful to ensure that these services are used meaningfully and balancedly in education.

Deep Ganguli and Danny Hernandez [22] model explores the challenges of creating large-scale designs in AI. They note that these models exhibit both predictable (so-called "scaling laws") and unpredictable behavior. This article discusses how this combination can cause social problems with real-life examples. The authors propose to make this model better and more responsible. This includes building more relationships between private companies and academic research, developing tools to check the performance of models, and testing new ways to manage and maintain standards. Their aim is to lead the development and use of these standards for the benefit of everyone.

Stefan Feuerriegel's [23] approach shows us how generative AI (a type of intelligence that produces content indistinguishable from human labor) can transform many businesses. The authors discuss examples such as Dall-E 2 and GPT-4, which demonstrate the potential for both artistic and practical use of this technology. They address challenges, including false positives and biases, and provide research methods for the business and knowledge engineering communities to address these issues and leverage intellectual intelligence capabilities.

3. PROPOSED METHOD

3.1 Loading and Splitting PDF documents: The code begins by loading several PDF documents related to Python using a library called PyPDFLoader. This library allows the program to access and extract text from the PDF files.

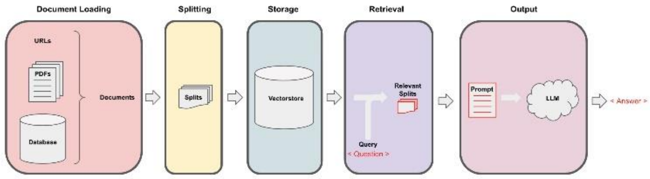

Figure 3.1 process of proposed method

Next, the code splits the extracted text from each page into smaller, manageable chunks. It explores different splitting strategies using the langchain library:

Character split: This method breaks down the text character by character, offering fine-grained control over the chunk size and overlap between consecutive chunks.

Recursive character split: Similar to the character split, but it allows for creating nested chunks within the main ones, providing a more hierarchical view of the text.

Sentence split: This strategy leverages punctuation marks, particularly "/n/n" and "/n", to divide the text into individual sentences, preserving their integrity.

3.2 Creating Text Embeddings:

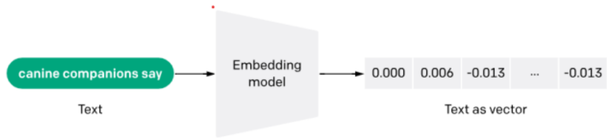

Figure 3.2. Text Embeddings

Text data is fed into an embedding model, which is trained on a vast amount of text and has learned to represent words as numerical vectors capturing their meaning and context. This model then outputs a vector representation of the entire text input, essentially converting the textual data into a numerical format that machines can understand and use for various natural language processing tasks like translation, summarization, and sentiment analysis.

After splitting the text, the process transitions to generating numerical representations of each chunk called "embeddings." These embeddings capture the meaning and context of the text, allowing different algorithms to efficiently understand and utilize the information. The OpenAIEmbeddings class from the langchain library is used for this purpose.

3.3 Storing Embeddings:

Finally, the generated embeddings are stored along with their corresponding text chunks in a specific directory on your computer using the Chroma class. This creates a readily accessible database of text representations that can be used for various downstream applications in the future.

The below image shows a glimpse on how the embeddings are stored and the answer is retrieved using the user query.

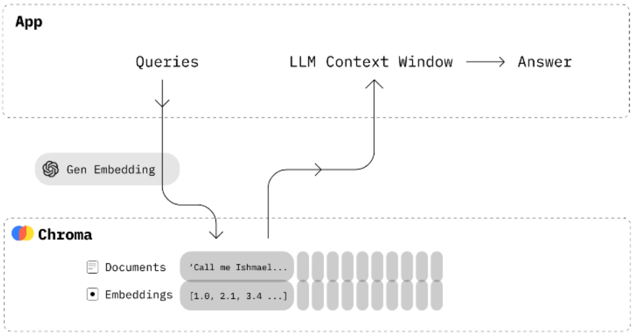

Figure 3.3. Chromadb

3.4 Database:

we leverage the LangChain library, employing the vectorstores module's Chroma component. Utilizing OpenAIEmbeddings to generate embeddings for a collection of documents. The resulting embeddings are then incorporated into a Chroma vector store, facilitating the creation of a comprehensive vector representation for the input documents. This approach combines the power of LangChain's vectorstores and OpenAI's embeddings, offering an effective means of capturing semantic information and relationships within the document set, thereby enhancing the overall understanding and analysis of textual data.

3.5 Retrieval:

The vectordb.similarity_search() method is likely used to search for documents similar to a given query within a vector database created using the langchain library. Here's a breakdown:

Vector Database: This database stores documents as numerical representations called "embeddings" instead of the original text. These embeddings capture the meaning and relationships between words in the document.

Query: You provide a document or text snippet as a query. The OpenAIEmbeddings class (mentioned previously) might be used to generate an embedding for this query as well.

Similarity Search: The vectordb.similarity_search() method searches the database for documents whose embeddings are most similar to the query embedding. It calculates a distance metric (like cosine similarity) between the query embedding and each document embedding in the database.

Ranked Results: The search method returns a ranked list of documents. Documents with embeddings closest to the query embedding (smallest distance) are ranked higher, indicating higher similarity to the query content.

Finally when the answer is retrieved it is passed to the LLM model and then its frames it as into a student friendly manner making it accurate and efficient for user.

The image demonstrates the whole process in a short format. From user input to the response generated by the model.

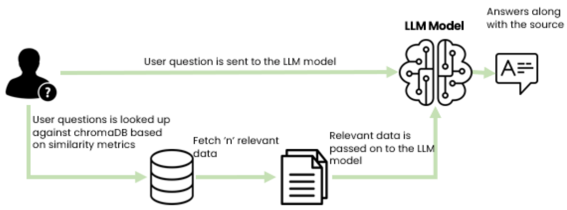

Figure 3.4. Retrieval model working

5. RESULTS

For assessing the effectiveness of system-generated summaries, we employed the ROUGE metric, an acronym for Recall-Oriented Understudy for Gisting Evaluation. In this context, "Gisting" refers to the extraction of the primary point from the text [31]. ROUGE serves as an evaluation matrix that compares the generated summary with a reference summary, calculating the overlap between the two concerning n-gram, word sequences, and word pairs. A higher ROUGE score signifies a superior alignment between the summary and the reference summary. ROUGE comprises five variants, including ROUGE-N, ROUGE-S, ROUGE-L, ROUGE-W, and ROUGE-SU [32]. In this study, we specifically employed ROUGE-N and ROUGE-L metrics to analyze various summarization models, where N denotes the length of the n-gram, encompassing ROUGE-1 (unigram), ROUGE-2 (bigram), ROUGE-3 (trigram), and so forth. The definition of ROUGE-N or ROUGE-L Metrics in terms of precision, recall, and F1 Scores parameters is elucidated as follows.

Precision denotes the fraction of the summary content that is pertinent to the original document. It is computed by dividing the count of relevant sentences in the summary by the total number of sentences in the summary, as illustrated in Equation (1).

$$Precision_{ROUGE-L} = \frac{Common\ n-grams\ in\ generated\ summary\ and\ reference\ summary}{Number\ of\ \ n-grams\ in\ generated\ summary} \quad (1)$$

Recall serves as a metric for assessing the efficacy of a summarization algorithm, indicating the proportion of crucial information in the original text that is captured in the algorithm-generated summary. The equation for recall is presented in Equation (2).

$$Recall_{ROUGE-L} = \frac{Common\ n-grams\ in\ generated\ summary\ and\ reference\ summary}{Number\ of\ \ n-grams\ in\ reference\ summary} \quad (2)$$

The F1 score represents the harmonic mean of precision and recall, amalgamating both measures into a single score that achieves a balance between precision and recall, as demonstrated in Equation (3).

$$F1\ score_{ROUGE-L} = 2 * \frac{Precision\ *\ Recall}{Precision\ +\ Recall} \quad (3)$$

Here, 'n' denotes the length of the n-gram under consideration (e.g., n = 1, 2, 3). These formulas are applicable for assessing the quality of a summary generated by an algorithm by juxtaposing it with a reference summary.

Table 1 presents the average precision scores of various answering techniques applied on 20 different questions. Our observation indicate that the proposed method excels, achieving a precision score of 0.91surpassing other methods.

TABLE I
AVERAGE PRECISION SCORE OF DIFFERENT MODELS

Model	**Precision**
BART	0.82
UNILM	0.84
Dense Net with Attention	0.78
Transformer with beam search	0.89
Proton Ai	**0.91**

The precision scores of 5 models are illustrated in the below figure.

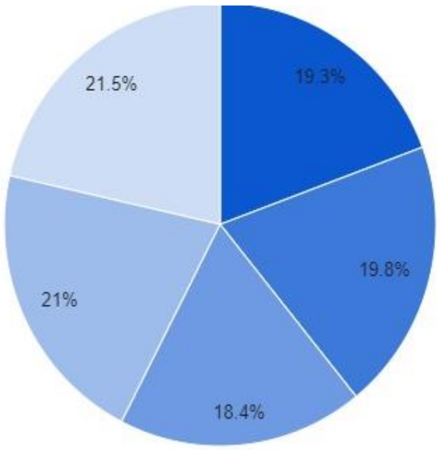

Figure 5.1 Precision Score comparison

Table II displays the Average Recall score for various questions posed to the proposed model. When we examine the table it is evident that the proposed model has the highest recall stating that the words from the generated responses match nearly with the reference answer.

TABLE II
AVERAGE RECALL SCORE OF DIFFERENT MODELS

Model	**Recall**
BART	0.79
UNILM	0.80
Dense Net with Attention	0.84
Transformer with beam search	0.87
Proton Ai	0.88

The recall scores of 5 different models are illustrated in the below image of line graph.

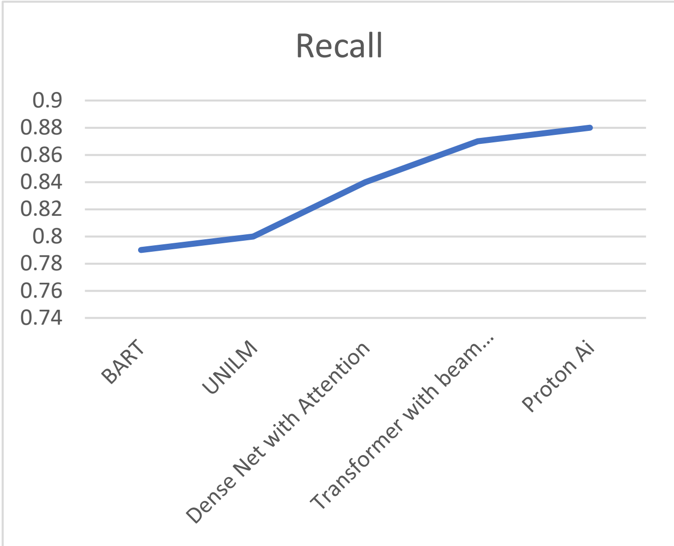

Table 3 displays the Average F1 scores of the summarizer models concerning ROUGE-1, ROUGE-2, and ROUGE-N for a given set of question. According to the table, the protonai achieves the highest ROUGE scores.

Method	**ROUGE-1**	**ROUGE-2**	**ROUGE-N**
TrivialQA	0.35	0.44	0.46
WikiQA	0.51	0.56	0.55
TANDA	0.53	0.55	0.55
BARTret	0.57	0.56	0.58
proposed	0.64	0.66	0.72

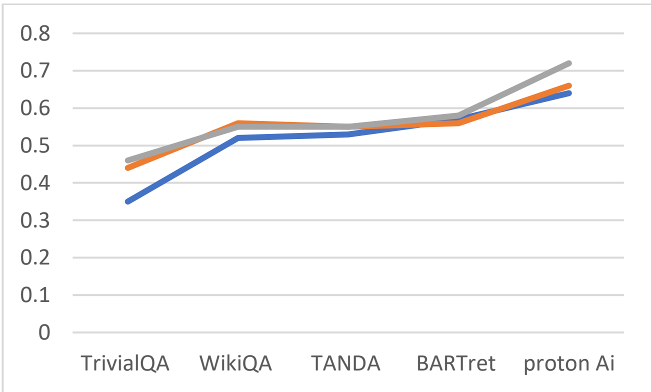

Figure 5.3. Rouge score comparision

As a major part of our research we also conducted a human evaluation by giving access to the protonAi for students and professors and collected their feedback on terms like the relevance of the answer, explainativeness of the answer generated, and also the chances of the user for him/her to promote the method to others. Table 4 shows the details of the human evaluation.

TABLE IV
HUMAN EVALUATION

Use Case	**Value**
how explainative	4.3
how relevant	4.2
how likely to promote	4.2
overall feedback	4.2
rating with ux	4.3
average of feedbacks	**4.24**

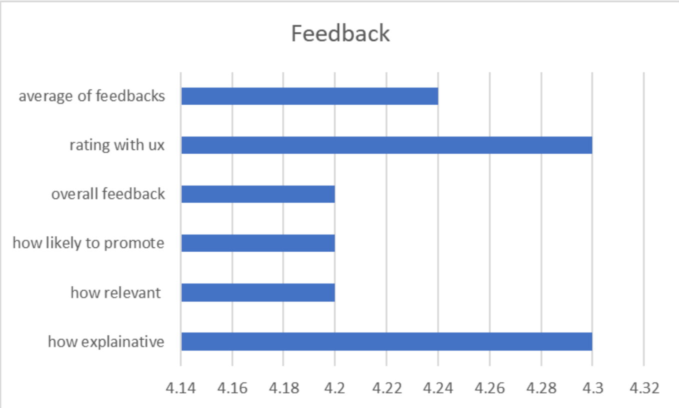

Figure 5.4. Human Evaluation

When seen the figure above it clearly tell that users who used the model when given access seem to be greatly happy with the results that protonAi has generated, showing that proposed method excels in its task when compared to the previous methods.

6. CONCLUSION

The research has proven that question answering of the ProtonAI compared to the already existing models has been more accurate and efficient. Unlike the previous models having the limitations of answering the question and bit out to context and having the legitimate response guarantee at stack ProtonAI overcomes these by only restricting its knowledge base to a specific text-book knowledge. Through the implementation of the langchain retrievals and the Openai's embeddings the storing of data in the vector databases makes the response generation makes the model excel in it performance.

7. References

[1]Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, Kyunghyun Cho, arXiv:1704.05179v3 [cs.CL] 11 Jun 2017

[2]Dubey Harish, Tamore Hardik , Padhi Sagar , Prof. Manisha Bharambe ,VOLUME: 07 ISSUE: 05 | MAY 2020, WWW.IRJET.NET

[3]Mandar Joshi, Eunsol Choi, Daniel S. Weld , Luke Zettlemoyer ,arXiv:1705.03551v2 [cs.CL] 13 May 2017

[4]Jonathan Berant , Andrew Chou Roy Frostig , Percy Liang ,Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1533–1544,Seattle, Washington, USA, 18-21 October 2013. c 2013 Association for Computational Linguistics

[5]Andrea Madotto, Pascale fung, arXiv:2202.03629v5 [cs.CL] 7 Nov 2022 , Survey of Hallucination in Natural Language Generation

[6]Jared Kaplan , Benjamin Mann ,arXiv:2005.14165v4 [cs.CL] 22 Jul 2020,

[7]Rohan Kumar, Youngmin Kim , Sunitha Ravi ,KnowledgeNLP-KDD'23, August 07, 2023, Long Beach, CA

[8]Ms. Khushbu Khandait et al. / Indian Journal of Computer Science and Engineering (IJCSE)

[9]Nithya M , Madavaraj B , Sanjeev Pranesh B , Kruthikaran V ,International Journal of Innovative Science and Research Technology , ISSN No:-2456-2165 , Volume 7, Issue 10, October – 2022

[10]Puneeth Thotad , Shanta Kallur , Sukanya Amminabhavi (),Journal of Pharmaceutical Negative Results ¦ Volume 13 ¦ Special Issue 10 ¦ 2022

[11]Hala Abdel-Galil,Mai Mokhtar,Salma Doma(2021), IJICIS, Vol.21, No.2, 110-123,Automatic question generation based on beep learning approach

[12]Shivani G. Aithal , Abishek B. Rao1 ·,Sanjay Singh (2021) ,Automatic question-answer pairs generation and question similarity mechanism , https://doi.org/10.1007/s10489-021-02348-9

[13]Rohan Bhirangi, Smita Bhoir(2016) , Bhirangi et al., International Journal of Emerging Research in Management &Technology, ISSN: 2278-9359 (Volume-5, Issue-4)

[13]Bidyut Das , Mukta Majumder , Santanu Phadikar and Arif Ahmed Sekh4 (2021),Das et al. Research and Practice in Technology Enhanced Learning

[14]Tejas Chakankar, Tejas Shinkar, Shreyash Waghdhare, Srushti Waichal, Mrs. M. M. Phadtare(2023) , https://doi.org/10.22214/ijraset.2023.49390

[15]Aanchal Jawere, Anchal Soni, Nitesh (2018), IJCRT_193832

[16]Bidyut Das , Mukta Majumder, Santanu Phadikar3 and Arif Ahmed Sekh4 (2021),Das et al. Research and Practice in Technology Enhanced Learning

[17]Son The Nguyena, Theja Tulabandhulaa , Mary Beth Watson-Manheima (2024) , https://ssrn.com/abstract=4662955,

[18]Safnah Ali ,Daniella DiPaola ,Irene Lee, Jenna Hong, CHI '21, May 8–13, 2021, Yokohama, Japan

[19]Dr. Manoj Kumar, Abhishek Singh , Arnav Kumar, Ankit Kumar , 021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT) | 978-1-6654-2392-2/21/$31.00 ©2021 IEEE | DOI:10.1109/CCICT53244.2021.00014

[20]Philipp Hacker ,Andreas Engel , Marco Mauer(2023), FAccT '23, June 12–15, 2023, Chicago, IL, USA , Fairness, Accountability, and Transparency (FAccT '23), June 12–15,

2023,Chicago, IL, USA. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3593013.3594067

[21]Vahid Ashrafimoghari ,Necdet Gürkan , Ashrafimoghari, Vahid and Gürkan, Necdet, Evaluating Large Language Models on the GMAT: Implications for the Future of Business Education (January 1, 2024). Available at SSRN: https://ssrn.com/abstract=4681307 or http://dx.doi.org/10.2139/ssrn.4681307

[22]Deep Ganguli, Danny Hernandez, Liane Lovitt, Predictability and Surprise in Large Generative Models, FAccT ’22, June 21–24, 2022, Seoul, Republic of Korea

[23]]Stefan Feuerriegel ,Jochen Hartmann ,Christian Janiesch ,Patrick Zschech(2023) , S. Feuerriegel et al.: Generative AI, Bus Inf Syst Eng

[24]Markus Freitag ,Al-Onaizan, , http://arxiv.org/abs/1910.13461 , Beam Search Strategies for Neural Machine Translation

[25]Li Dong, Nan Yang , Wenhui Wang , Furu Wei , Xiaodong Liu Yu , Wang Jianfeng Gao , Ming Zhou , Hsiao-Wuen Hon,arXiv:1905.03197v3 [cs.CL] 15 Oct 2019

[26]Mike Lewis, Yinhan Liu*, Naman Goyal, Marjan Ghazvininejad,Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, arXiv:1910.13461v1 [cs.CL] 29 Oct 2019.

[27]Gao Huang,Zhuang Liu,Laurens van der Maaten ,DOI 10.1109/CVPR.2017.243