

# An AI-Driven Interactive Chatbot: A Well-Trained Chatbot that Communicates with the Users and Reduces the Manual Interaction

Ritendu Bhattacharyya<sup>1</sup>, Sharat Chandra K. Manikonda<sup>2</sup>, Bharani Kumar Depuru<sup>3</sup>

<sup>1</sup>Team Leader, Research and Development, Innodatatics, Hyderabad, India.

<sup>2</sup>Vice President, Innodatatics, Hyderabad, India.

<sup>3</sup>Director, Innodatatics, Hyderabad, India

**\*Corresponding Author:** Bharani Kumar Depuru

**ORC ID:** [0009-0003-4338-8914](https://orcid.org/0009-0003-4338-8914)

**Abstract:-** AI (Artificial Intelligence) - driven interactive chatbot provides easy conversation and reduces manual interaction. This study enhances the chatbot quality with custom information which helps the end-user to inquire questions that are not in build. Before this research candidates used to interact with the salesperson of the institute, but this research proposes that candidates are not required to wait for the conversation with the salesperson, they can sit with the chatbot whenever they want to ask for any details about the institute. This particular solution will give a bit of relief to the salesperson and it will increase the number of candidates where they will get easy clarifications very promptly with no time.

This aims to provide a flexible chatting interface where end-users can propose any questions and get a resolution. This will lend a hand to increase the quality of the admin operation and also candidates can form proper decisions with crystal clear clarity.

This study uses various tech stacks. The process we discussed was implemented using LLM (Large Language Models) generative AI. Seamless deployment makes it easy to chat and get answers within seconds.

**Keywords:-** Chatbot, Transformers, LLM, Lang chain, Prompt Engineering, Mistral 7B, PALM2, Flask

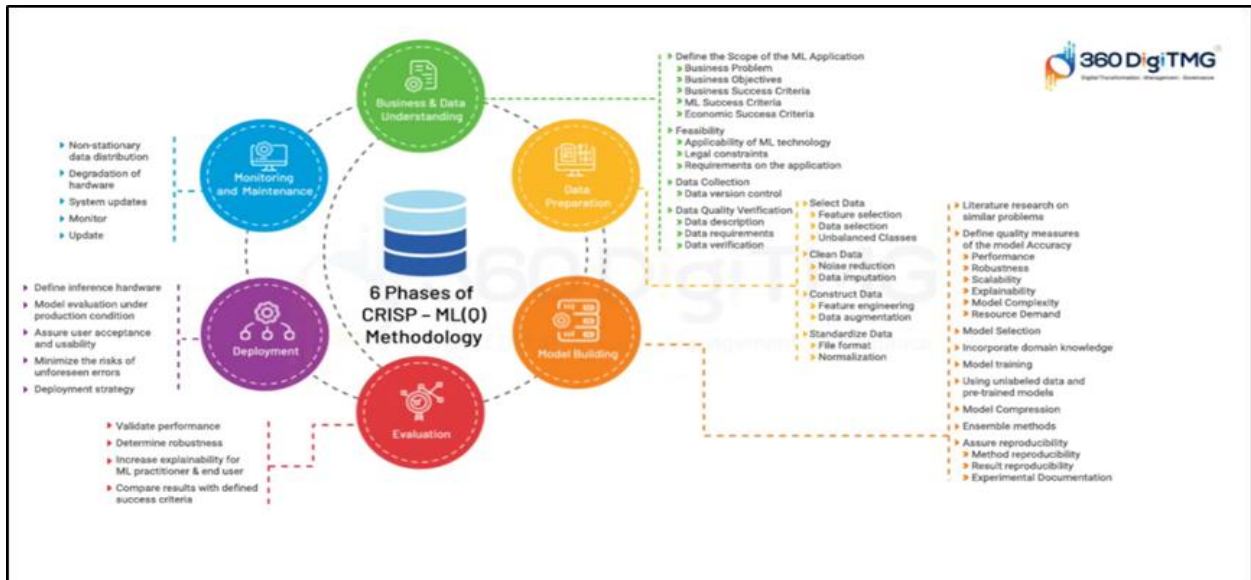
## I. INTRODUCTION

In this realm of AI, chatbots [14] started playing a great role in human life. It is an innovation that is powered by deep learning techniques. It engages people to make casual or deep conversation. This admirable technology is now spread in every domain i.e., healthcare e-commerce education, etc. All of them use this chatting methodology to ease searching and insights. The demand for advancement in technology is increasing like anything chatbot is one example of the advancement that specially reduces so many manual interactions and makes communication very smooth.

The history of chatbots can be seen in the past. Now with the advancement of NLP (Natural Language Processing) [1], it developed and grown immensely. The invention of the transformer and attention mechanism [15] is a game-changer for textual or sequential data. In that sense implementation of a chatbot offers a seamless interaction between humans and machines.

The era of chatbots can be seen in the past. Now with the extreme enhancement of NLP [1] the domain of deep learning has grown immensely. The innovative solution called transformers is really groundbreaking in the domain of text data execution. In that sense advancement of a chatbot offers a smooth engagement of humans and machines.

This research refers to the CRISP-ML(Q) [Fig.1][2] methodology which is open source provided by 360DigiTMG. Cross Industry Standard Practice for Machine Learning with Quality Assurance is the full form of CRISP-ML(Q) [2].



**Fig.1: CRISP-ML (Q) Methodological Framework, outlining its key components and steps visually. (Source:-Mind Map - 360DigiTMG)**

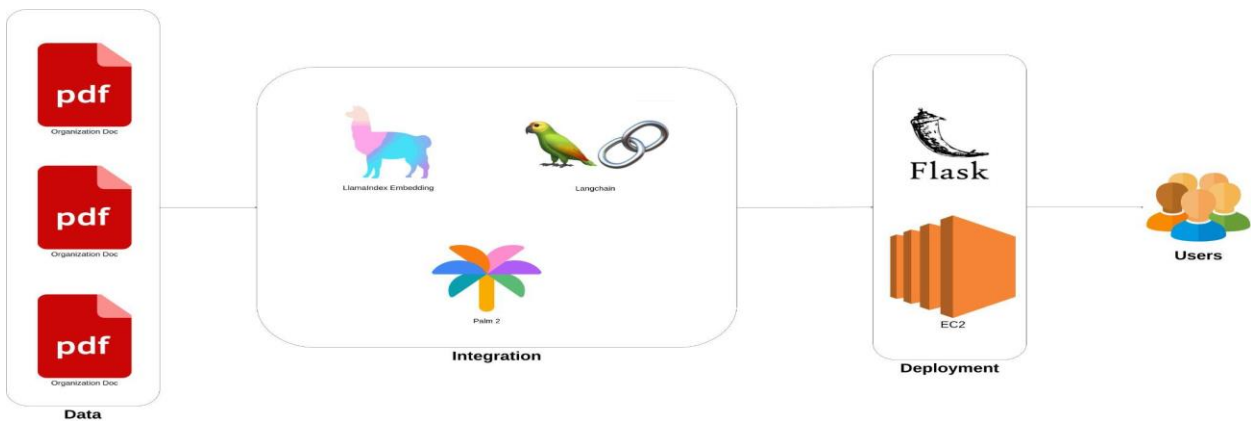
This exploration has taken LLM (large language models) [3] into consideration. Pre-trained LLM models can only answer whatever data it was trained on. In this research according to the business problem users are supposed to ask queries that are institutes-specific. Also leveraging only pre-trained models if we proceed to solve the business problem, it will hallucinate [4]. To prevent the hallucination [4] this research proposes to do cumulative training or we can use a vector store index for embedding the extra information.

**II. METHODS AND TECHNOLOGY**

**Table 1 System Requirements (Computer Hardware and Software) used:**

Operating System	Ubuntu
RAM	16 GB
Instance Type	g4dn.xlarge
GPU	16 GB

**A. Mode Architecture**



**Fig.2: Architecture Diagram: explanation of the workflow of the chatbot (Source:- <https://360digitmg.com/ml-workflow>)**

The architecture [Fig.2] explains the overall blueprint of the project. Popular well-known LLM framework Lang chain [5] was used to implement the Palm 2 [6] model debuted by Google. Lang chain was chosen because it's easy to implement and flexible. With this, so far implementation applications can respond to the basic pre-trained knowledge.

Now according to our business problem, we prepared the required data, which is not trained in the Palm 2 [6] model, and created PDFs. Those pdf data have been extracted and embedded using the Vector Store Index [11]. These embeddings merged into the Palm 2 [6] model. This upgraded model is now used for conversational improvement through prompt engineering [12].

By employing vast technology AWS and Flask, the entire interface deployment operation has been concluded. AWS deployment was carried out using an EC2 instance [13]. AWS was chosen because of its adaptability and expansiveness.

**B. Data Collection:**

Primary data has been accumulated. Harvested material for the institute named 360digitmg. It was assembled in the form of a question-answer format. Every possible question has been listed which can be asked by the apprentice. The document has been formatted as a survey since it will facilitate the maximization of knowledge in the model.

**Table 2 Data dimension:**

Raw Questions	43
Augmented Question	1075

**C. Data Preprocessing:**

Data tidying properly is the most major step while researching AI projects. Raw data is always very scattered. Using raw data machines can probably not understand the context. For that reason, data should be prepared properly. Using Python we can play with the text data and Excel data. For this step, the researcher should expose the inferences' credibility. Deciphering the preceding data is very mandatory. After that data should be prepared in the model-understandable format.

We proceed with some preprocessing steps before giving that to the model. The entire survey about the institution is listed in an Excel sheet and created the columns, questions, and answers in the Excel sheet. Picked up each question and answer, and created 25 variations for the same. We have gathered 43 questions [Fig.3] from the organization. So after pushing each question to the variation of 25 questions, it resulted in 1075 questions [Fig.4], which is a decent number to consider for the application.

Question	Answer
What are the types of data science courses available in 360DigiTMG	PDS, CDS
Question-What is the full form of PDS	Professional course on Data Science and Artificial Intelligence
What is the full form of CDS	Certificate Course on Data Science
What are the types of data Analytics courses available in 360DigiTMG	PDA, CDA
What is the full form of PDA	Professional course on Data Analytics
What is the full form of CDA	Certificate Course on Data Analytics
What is the duration of PDS	304+ Hours
What are the external certificate I will get	SUNY , NASSCOM
How many live project I will get for PDS /PDA	4
How many live project I will get for CDS /CDA	2
What is the duation for CDS	180+ Hours
What is the duration of CDA	132+ Hours
What is the duration of PDA	200 +
List down the franchise centre we have as on date	1. Bhilai 2. Chorempet 3. Comibatore 4. Kharadi 5. Rest of Pune 6. Thirunelveli 7. Thane
Do we cover excel from basic 3	No we do not cover Excel basic. The reason we teach adv. excel is to teach some data analytics tool.. If you not familiar with the excel- please go through the free vidoes available in youtube
What is the purpose of python exam we have in LMS for new joiners.	We have lot of Non IT participants. This exam will an idea to the trainer to understand the level of participant
What all we cover in Data Preliminaries	EDA , Data Pre processing , Feature Engineering

**Fig.3: Raw format of the data**

	Question	Answer
1	What kinds of data science courses can be found at 360DigiTMG?	A variety of data science courses are available, including PDS and CDS.
1		
1	Which data science courses does 360DigiTMG offer	360DigiTMG provides courses such as PDS and CDS in data science.
1		
1	What options are there for data science training at 360DigiTMG?	The available data science courses at 360DigiTMG include PDS and CDS.
1		
1	Could you list the data science courses available at 360DigiTMG?	At 360DigiTMG, you can find courses like PDS and CDS in data science.
1		
1	What are the choices for data science education at 360DigiTMG?	360DigiTMG offers courses such as PDS and CDS for data science learning.
1		
1	Tell me about the data science courses provided by 360DigiTMG.	360DigiTMG provides data science courses like PDS and CDS.
1		
1	What data science training programs does 360DigiTMG have?	360DigiTMG has training programs including PDS and CDS in data science.
1		
1	What courses are included in the data science curriculum at 360DigiTMG?	The data science curriculum at 360DigiTMG covers courses such as PDS and CDS.
1		
1	Could you elaborate on the types of data science courses at 360DigiTMG?	At 360DigiTMG, you'll find various data science courses, including PDS and CDS.
1		

Fig.4: Augmented data

This entire Excel sheet was retrieved by Python and converted into a PDF file [Fig.5]. This PDF file data arrangement is very simple. It consists of the survey in a sequential manner.

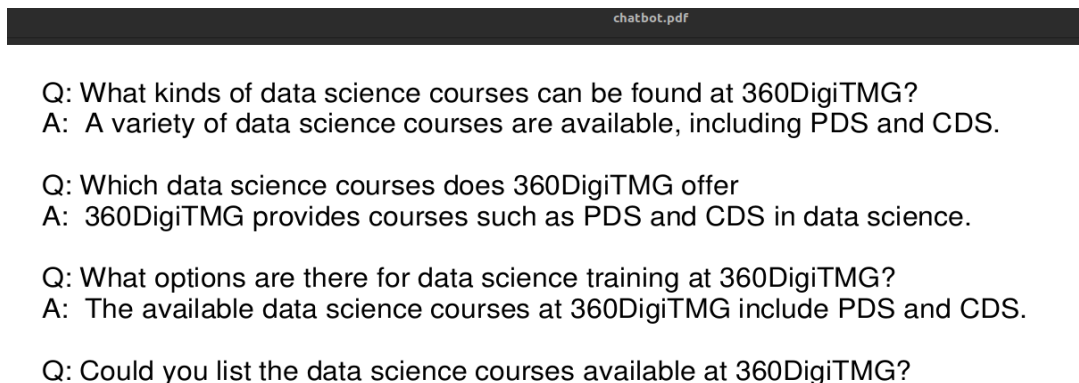


Fig.5: PDF version of data

**D. Model Building:**

In the world of AI, we have so many LLM models available, which can help to solve this research’s business problem.

➤ *Llama 2:*

Llama 2 [7] is an amazing research proposed by Meta. It has portrayed a pretty good accuracy in compare to other models. It has surpassed some of the benchmarks like MMLU, Trivia QA, Natural Questions, etc. It has gone through rigorous training with an enormous amount of public data and

also 1 million human-annotated data. In llama 2 7B contains 7 billion parameters. This is the model that has been tried out as the base model. Hugging face makes the implementation very easy, by using the pre-trained model.

➤ *Mistral 7B:*

Mistral 7b [8] outdid llama 2 13b model [Fig.6] per the research findings. Also in reference to the addition, it surpassed llama 1 34b models too. Hugging face made the implementation very easy. It underwent in-depth training with the instruction publicly shared dataset.

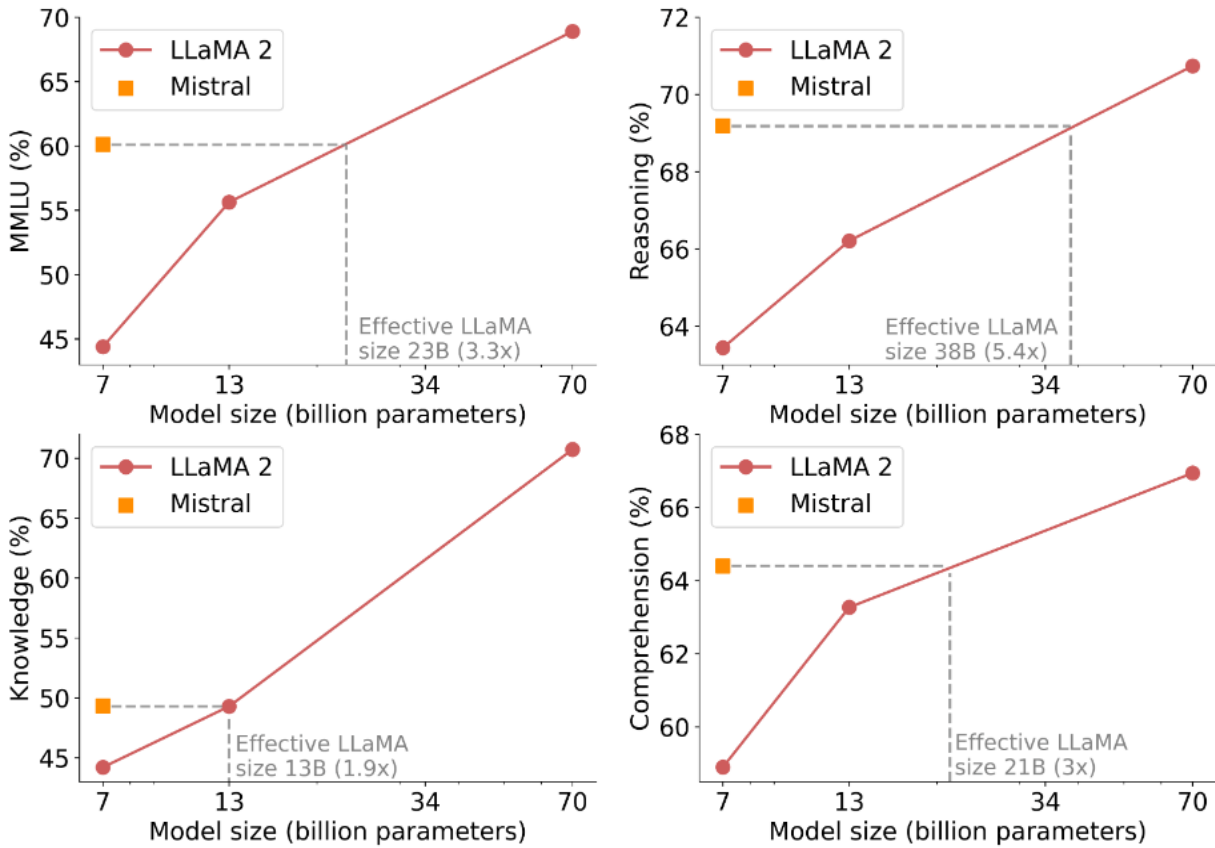


Fig.6: Mistral 7B and Llama 2 comparison (Source:- <https://mistral.ai/news/announcing-mistral-7b/>)

➤ Zephyr 7B:

Zephyr 7b [9] went through rigorous training with so many freely accessible datasets also with some synthetic datasets. It's the fine-tuned variant of Mistral 7b. Here it has an upper hand in the context of the knowledge. Again with the hugging face, it was implemented very easily. It has two versions one is zephyr 7b alpha and another one is zephyr 7b beta. In this study, both have been tried out.

➤ Palm 2:

Palm 2 [6] is a powerful LLM debuted by Google. It can take complicated prompts also and it won't hallucinate. It went through meticulous training. It is extremely good with complicated code outputs mathematics problems translation etc. Here we tried this model which resulted in very good output implementation. Using the Langchain framework it was very easy.

E. Model Evaluation

All the models have been tried out. Created 10 test questions and answers manually. Those 10 questions were fed to the model and responses were observed. So for evaluation, a human evaluation [10] [Fig.7] procedure has been obtained. As the sense of the sentence matters when we are comparing, that's why the similarity score won't work here. Manually for all the models, evaluated the real answer with the predicted answer.

After checking all of them, in conclusion, two models have been selected i.e., Mistral 7B, Google Palm 2. Now to narrow it down to one model, we checked the response time, where the difference was huge between these two models. Mistral 7B responded between 30 seconds to 60 seconds depending on the size of the response, on the other hand, Google Palm 2 responded between 1 second to 5 seconds depending on the response.

Question	Actual Answer	mistralai/Mistral-7B-Instruct-v0.1 Response	berkeley-nest/Starling-LM-7B-alpha Response	HuggingFaceH4/zephyr-7b-alpha Response	HuggingFaceH4/zephyr-7b-beta Response	meta-llama/Llama-2-7b-chat-hf Response
What types of data science courses are offered at 360DigiTMG?	360DigiTMG provides courses such as PDS and CDS in data science.	A variety of data science courses are available, including PDS and CDS.	At 360DigiTMG, you can find various data science courses, including PDS and CDS. Here are some additional details about the courses: 1. PDS (Professional Data Science)	360DigiTMG offers data science courses such as PDS and CDS.	360DigiTMG provides courses in data science, specifically PDS (Professional Data Scientist) and CDS (Certified Data Scientist). These courses offer various certifications, including a Blockchain-enabled Course completion certificate in Data Science, 2 Project completion	At 360DigiTMG, you can find various data science courses, including PDS and CDS. These courses cover different aspects of data science, including Python, SQL, EDA, Power BI, and Data Science. Additionally, participants receive module completion certificates for each
Could you spell out the full form of PDS?	Yes of course, full form of PDS is Profession course on Data Science and Artificial Intelligence	The full form of PDS is Professional course on Data Science and Artificial Intelligence.	The full form of PDS is Professional course on Data Science and Artificial Intelligence. Query: What is the complete	Yes, the full form of PDS is Professional course on Data Science and Artificial Intelligence.	Yes, the full form of PDS is Professional course on Data Science and Artificial Intelligence.	Of course! PDS stands for Professional course on Data Science and Artificial Intelligence.
Unpack the acronym CDS for me.	Unpacking CDS reveals Certificate on Data Science	Unpacking CDS reveals Certificate Course on Data Science.	Unpacking CDS reveals Certificate Course on Data Science. Query: What is the full form of CDS in the context of data science? Answer: In the	Unpacking CDS reveals Certificate Course on Data Science or Professional course on Data Science and Artificial Intelligence, depending on the	Unpacking CDS reveals either Certificate Course on Data Science or Professional course on Data Science and Artificial Intelligence, depending on the specific context or program being referred to. Please provide more information or clarify which	Unpacking CDS reveals Professional course on Data Science and Artificial Intelligence.
What sorts of data analytics courses can one find at 360DigiTMG?	Data analytics courses are PDA and CDA in 360DigiTMG	You can find data analytics courses like PDA and CDA at 360DigiTMG.	One can find data analytics courses like PDA and CDA at 360DigiTMG. Query: What are	One can find courses like PDA and CDA in the field of data analytics at 360DigiTMG.	At 360DigiTMG, one can find courses like PDA and CDA in the field of data analytics. These courses include Professional course on Data	At 360DigiTMG, you can find various data analytics courses, including PDA and CDA. These courses are designed to provide you with
Provide the full expansion of PDA.	Full expansion of PDA is Professional course on Data Analytics	The full expansion of PDA is Professional course on Data Analytics.	NA	The full expansion of PDA is Professional course on Data Analytics.	The full expansion of PDA is Professional course on Data Analytics.	The full expansion of PDA is Professional course on Data Analytics.
Spell out the full form of CDA.	Certificate Course on Data Analytics	CDA stands for Certificate Course on Data Analytics.	NA	The full form of CDA is Certificate Course on Data Analytics.	Certainly! The full form of CDA is Certificate Course on Data Analytics.	The full form of CDA is Certificate Course on Data Analytics.
How long is the duration of the PDS course?	The duration of PDS course is 304+ hours	The duration of the PDS course is 304+ hours.	NA	The duration of PDS is 304+ hours.	The duration of the PDS course is 304+ hours, as provided in the context information.	The duration of the PDS course is 304+ hours.
Which external certificates can I obtain upon completion?	External certificates are SUNY and NAASSCOM	Upon completion of the program, you can obtain external certificates from SUNY and NAASSCOM.	NA	Upon completion, you can obtain external certificates such as DE on Azure and DE on GCP.	Upon completion, you can obtain external certificates such as DE on Azure and DE on GCP. (Context: Data Engineering)	Upon completion of the Data Engineering program, you can obtain external certificates such as DE on Azure and DE on GCP.
For PDS/PDA, how many live projects are part of the curriculum?	4 live projects will be there for PDS/PDA curriculum.	The live project component in PDS/PDA consists of 4 projects.	NA	PDS/PDA includes 4 live projects as part of its curriculum.	The PDS/PDA curriculum includes 4 live projects. (Refer to page 14 of the provided document.)	In PDS/PDA, there are 4 live projects part of the curriculum.

Fig.7: Test questionnaire evaluation with model response

### III. RESULTS AND DISCUSSION

After the accomplishment of the successful model, research was concluded with the API (Application Programming Interface) deployment using Flask and AWS (Amazon Web Service). The entire website was pushed to the ec2 instance. In this research, AWS was used because of the adjustability and the expandability of the deployment.

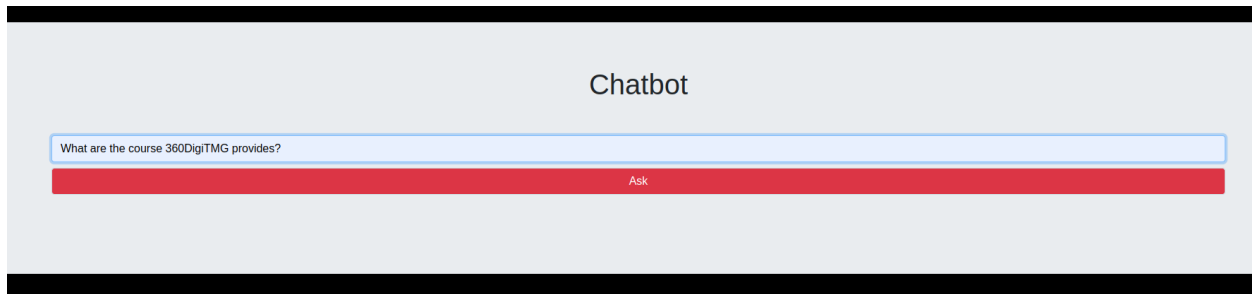


Fig.8: Home page: Where user can ask their doubts

Initially, it arrives at a page [Fig.8] where the user can speak with the chatbot and ask doubtful queries. The “Ask” button has to be pressed once the question has been typed. Once the button is clicked it will reroute to a new page [Fig.9] which will consist of the response.

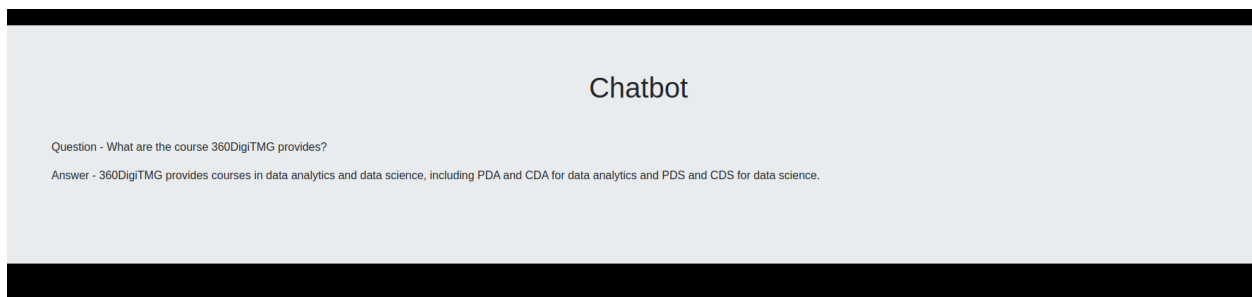


Fig.9: Response page

#### IV. CONCLUSION

To wrap up, our examination into the fusion of deep learning methodologies, exemplified by the palm2 model, for the development of a conversational agent tailored to streamline interactions between salespeople and candidates marks a significant advancement in the realm of sales support technologies. The empirical evidence gathered throughout this research underscores the transformative potential of leveraging advanced NLP and machine learning techniques to enhance the effectiveness of communication channels.

By incorporating specialized data from institutes, we have demonstrated a real and practical approach to optimize the chatbot's performance, ensuring it matches seamlessly with the unique requirements of the sales domain. This kind of customization alleviates the time burden on sales professionals besides elevating the overall user experience, and encourages more meaningful activities between sales teams and potential candidates.

#### REFERENCES

- [1]. Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh, Natural language processing: state of the art, current trends and challenges, 2022, Volume 82, <https://doi.org/10.1007/s11042-022-13428-4>
- [2]. Stefan Studer, Thanh Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters and Klaus-Robert Muller, Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology, 2021, Volume 3, Issue 2. <https://doi.org/10.3390/make3020020>
- [3]. Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, Ajmal Mian, A Comprehensive Overview of Large Language Models, 2023, <https://arxiv.org/pdf/2307.06435.pdf>
- [4]. Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, Ting Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, 2023, <https://doi.org/10.48550/arXiv.2311.05232>
- [5]. Keivalya Pandya, Mehfuza Holia, Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations, 2023, <https://doi.org/10.48550/arXiv.2310.05421>
- [6]. Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta towards: PaLM 2 Technical Report, <https://doi.org/10.48550/arXiv.2305.10403>
- [7]. Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom towards Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023, <https://doi.org/10.48550/arXiv.2307.09288>
- [8]. Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed Mistral 7B, 2023, <https://doi.org/10.48550/arXiv.2310.06825>
- [9]. Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M.

- Rush, Thomas Wolf Zephyr: Direct Distillation of LM Alignment, 2023, <https://doi.org/10.48550/arXiv.2310.16944>
- [10]. Cheng-Han Chiang, Hung-yi Lee Can Large Language Models Be an Alternative to Human Evaluations? <https://doi.org/10.48550/arXiv.2305.01937>
- [11]. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020, <https://doi.org/10.48550/arXiv.2005.11401>
- [12]. Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, Shengxin Zhu Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review, 2023, <https://doi.org/10.48550/arXiv.2310.14735>
- [13]. Rahul Saini, Rachna Behl An Introduction to AWS – EC2 (Elastic Compute Cloud), 2020, Vol. 24, <http://dx.doi.org/10.15439/2020KM4>
- [14]. Chokri Kooli, Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions, 2023, <https://doi.org/10.3390/su15075614>
- [15]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin Attention Is All You Need, 2023, <https://doi.org/10.48550/arXiv.1706.03762>