LangchainIQ: Intelligent Content and Query Processing

Sunil Ghane [1], Roshan Sawant [2], Ganesh Supe [3], & Chinmay Pichad [4]
[1] Assistant Professor, Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India; E-mail: sunil.ghane@spit.ac.in
[2] Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India; E-mail: roshan.sawant@spit.ac.in
[3] Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India; E-mail: ganesh.supe@spit.ac.in
[4] Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India; E-mail: chinmay.pichad@spit.ac.in

Area/Section: Computer Science.
Type of the Paper: Analytical.
Type of Review: Peer Reviewed as per |C|O|P|E| guidance.
Indexed in: OpenAIRE.
DOI: https://doi.org/10.5281/zenodo.13325237
Google Scholar Citation: IJMTS

How to Cite this Paper:
Ghane, S., Sawant, R., Supe, G. & Pichad, C. (2024). LangchainIQ: Intelligent Content and Query Processing. *International Journal of Management, Technology, and Social Sciences (IJMTS), 9*(3), 34-43. DOI: https://doi.org/10.5281/zenodo.13325237

International Journal of Management, Technology, and Social Sciences (IJMTS)
A Refereed International Journal of Srinivas University, India.

CrossRef DOI: https://doi.org/10.47992/IJMTS.2581.6012.0360

Received on: 29/06/2024
Published on: 14/08/2024

LangchainIQ: Intelligent Content and Query Processing

Sunil Ghane [1], Roshan Sawant [2], Ganesh Supe [3], & Chinmay Pichad [4]
[1] Assistant Professor, Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India; E-mail: sunil.ghane@spit.ac.in
[2] Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India; E-mail: roshan.sawant@spit.ac.in
[3] Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India; E-mail: ganesh.supe@spit.ac.in
[4] Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India; E-mail: chinmay.pichad@spit.ac.in

ABSTRACT

Purpose: *The purpose of this research is to introduce and evaluate a comprehensive framework called langchain(component of Large Language Model), designed to optimize data analysis and visualization processes across various business domains. The framework integrates advanced computational techniques with user-friendly interfaces to meet the growing demand for efficient information processing tools in research and industry settings.*
Design/Methodology/Approach: *The framework consists of three primary components: PDF answering, CSV analytics, and data visualization using the LIDA library. Integration of advanced technologies such as the Mistral 7B model for language processing, Faiss for similarity search, and the LIDA library for data visualization. Detailed implementation steps include content processing, embedding using OpenAI embeddings, storage and retrieval using Faiss, and query handling using Mistral 7B. This involves breaking down PDF and CSV content into chunks, embedding them, and utilizing advanced algorithms for efficient data retrieval and visualization.*
Findings/Result: *The fine-tuned Mistral 7B model significantly enhances data extraction speed compared to traditional models like Llama. Users can effectively query and extract specific information from PDFs and CSVs using natural language, facilitated by advanced AI models. The LIDA library automates the generation of insightful visualizations from processed data, enhancing data interpretation and decision-making.*
Originality/Value: *Introducing langchain as a versatile framework that addresses the complexities of data analysis and visualization and it's use in business analysis.*
Paper Type: *Technical Research.*
Keywords: Langchain, AI, Large Language Model, Mistral 7B Model, LIDA framework.

1. INTRODUCTION :

In an era marked by an exponential increase in data generation and consumption, the ability to extract meaningful insights from vast datasets is paramount. From scientific research to business analytics, the demand for robust tools that streamline data analysis and visualization processes has never been greater. This research endeavors to address this pressing need by introducing the use of a comprehensive framework langchain designed to optimize information comprehension and exploration.

The framework presented herein is used in three principal components, each tailored to tackle distinct challenges encountered in data-centric endeavors. Firstly, the PDF answering feature offers a solution for efficiently parsing and understanding complex PDF documents. In academic or professional settings alike, individuals often encounter dense, text-heavy materials that demand meticulous examination. By enabling users to pose queries related to the content of PDF documents, this feature aims to expedite the process of information extraction and comprehension, thereby enhancing productivity and reducing cognitive load.

Secondly, the CSV analytics component addresses the intricacies of structured data analysis. With the proliferation of data-driven decision-making across industries, the ability to extract actionable insights from structured datasets is imperative. This feature equips users with advanced analytical tools and

algorithms capable of uncovering patterns, trends, and correlations within CSV-formatted data. By leveraging computational techniques to navigate the complexities inherent in large datasets, this component empowers users to make informed decisions and derive maximum value from their data assets.

Lastly, the framework incorporates a data visualization module powered by the LIDA. In an age where data visualization serves as a cornerstone of effective communication and analysis, this feature offers a sophisticated yet user-friendly solution for generating dynamic visual representations of data. By leveraging state-of-the-art natural language processing capabilities, users can input textual content and effortlessly transform it into visually intuitive graphs and charts. This not only enhances the interpretability of data but also facilitates the dissemination of insights to diverse stakeholders, thereby fostering collaboration and informed decision-making.

All Those features rely on the Fine-tuned Mistral 7B model, enhancing the extraction of data-specific content. This fine-tuned model significantly improves the precision and relevance of data extraction, ensuring that extracted information is tailored to specific requirements and contexts.

Overall, the framework presented in this research paper represents a significant advancement in the field of data analysis and visualization. By integrating cutting-edge computational techniques with user-centric design principles, it seeks to empower individuals across domains to unlock the full potential of their data resources. In the following sections, each component of the framework will be explored in detail, highlighting its unique capabilities and potential applications in research and industry contexts.

2. RELATED WORKS :

1. Chatbots in Education System Vijaya Lakshmi, Y and Ishfaq Majid (2022) provide an introductory exploration of chatbot applications in educational institutions. Despite its valuable insights, it falls short of offering a specific methodology for implementation. However, the paper offers a foundational overview of how chatbots can be integrated into the educational domain.
2. AI Assistant for document management using Langchain and Pinecone (2023) provides us with the system methodology and system diagram. Thai paper explains how components work to perform specific tasks and get the results.
3. A Survey of Large Language Models" (2022) delves into the inner workings of LLMs and their embedding processes, shedding light on the technology's core operations. Nevertheless, the paper lacks a direct comparison of different LLM models, making it challenging to assess their relative strengths and weaknesses.
4. An Effective Query System Using LLMs and LangChain" (2022) does not explicitly specify its disadvantages but suggests future work related to processing CSV data for analysis. This hints at the potential for improving its data processing capabilities, making it a point of interest for further research.
5. "The Agents of AI: Data Analysis with LLMs and LangChain Agents" (2022) presents a system designed for CSV data analysis, demonstrating the role of LLMs and LangChain in data analysis applications.
6. RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit" (2023) does not specify its disadvantages but offers suggestions for effectively utilizing LLM models, making it an interesting prospect for researchers seeking guidance on LLM implementation.
7. Faiss: Efficient Similarity Search and Clustering of Dense Vectors" (2021) primarily serves as a data store, contributing to the efficiency of similarity search and clustering of dense vectors. It does not explicitly outline its advantages or disadvantages.

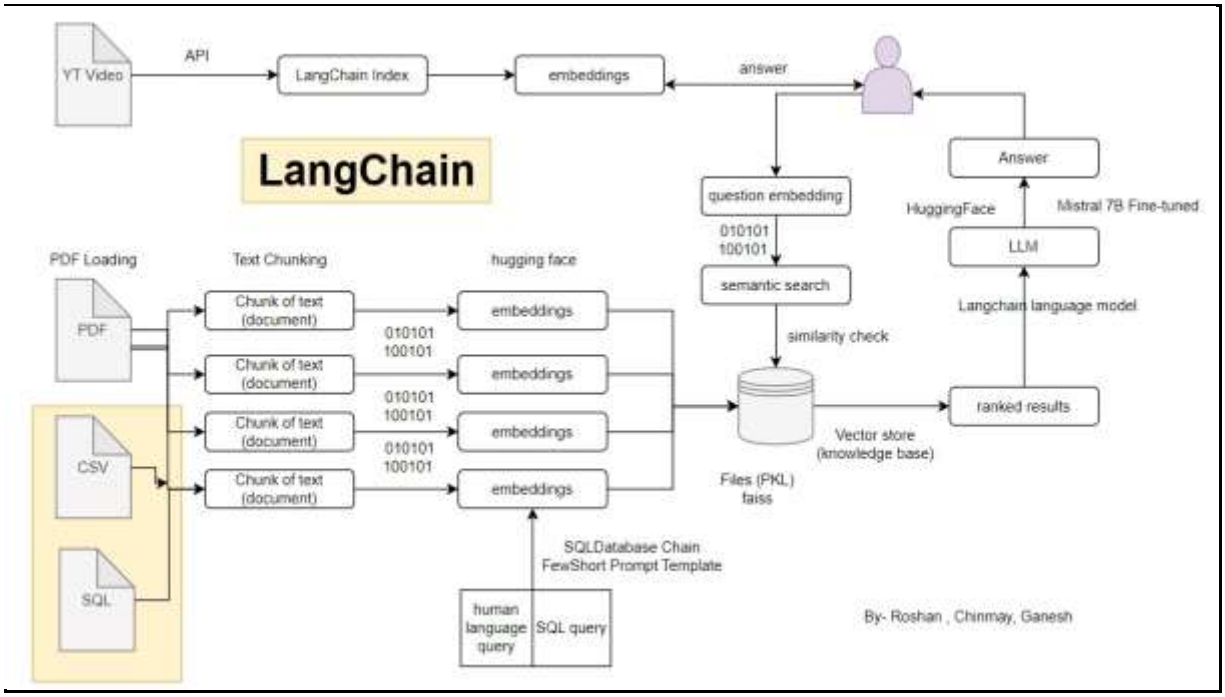
3. OBJECTIVES :

(1) Streamline Data Processing: Develop tools to streamline the processing and analysis of diverse data formats.
(2) Enhance Decision-Making: Empower stakeholders with actionable insights derived from structured and unstructured data.
(3) Improve Data Visualization: Implement tools for creating visualizations that aid in understanding complex datasets quickly.
(4) Increase Operational Efficiency: Optimize workflows through automated data handling and insightful visualization capabilities.

(5) Facilitate Predictive Analytics: Leverage advanced algorithms to uncover trends and predict future outcomes based on data patterns.

4. METHODOLOGY :

The system's foundational operation lies in its ability to ingest, process, and harness data from diverse sources, including PDF documents, CSV files, and video transcripts, to establish a robust and versatile knowledge base. A key aspect of this process involves the meticulous segmentation of PDFs into manageable chunks, typically comprising 2000-word segments with an overlap of 100 words. These segments are then subjected to the sophisticated linguistic embedding capabilities of LangChain, a potent language model designed to encode the intricate nuances of textual content into high-dimensional vector representations.



Once embedded, these vector representations are stored and indexed using Faiss, a powerful and feature-rich vector storage solution renowned for its efficiency in semantic search techniques. This indexing process lays the foundation for the system's ability to swiftly and accurately retrieve relevant information in response to user queries.

When users submit queries, the system seamlessly integrates LangChain and OpenAI models to encode the query into a format harmonious with the knowledge base. This encoding ensures that the query aligns closely with the semantic representations of the indexed content, thereby facilitating precise and contextually relevant search results. Through the utilization of Faiss, the system efficiently traverses the vectorized knowledge base, swiftly identifying and presenting the most pertinent information to the user.

Similar methodologies are applied to CSV files, with OpenAI models adeptly generating query data frames to facilitate the extraction of relevant insights from the knowledge base. Additionally, the system extends its capabilities to encompass video processing, wherein users can input YouTube links for transcript extraction. Leveraging the prowess of LangChain, these transcripts are transformed into structured data, enriching the knowledge base and enabling seamless query answering across multimedia content.

To amplify user engagement and facilitate intuitive interaction, the system boasts a user-friendly interface meticulously crafted using Streamlit. This interface provides users with a streamlined experience, offering straightforward options for content processing and query submission. Whether it's parsing through PDFs, analysing CSV data, or querying textual information, users can effortlessly navigate the system to fulfil their business objectives.

In tandem with its content processing and query answering capabilities, the system places a significant

emphasis on data visualization, leveraging the advanced functionalities of LIDA for automated visualization. This empowers users to derive actionable insights and uncover hidden patterns within their data, thereby facilitating informed decision-making and strategic planning across various business domains.

Furthermore, the system incorporates sophisticated fine-tuning techniques on the Mistral 7b model to optimize its performance and adaptability to specific business contexts. This iterative process involves training the model on domain-specific data and fine-tuning its parameters to enhance its accuracy and relevance in generating insights tailored to the target domain.

To fortify the integrity and confidentiality of the data processed within the system, robust security measures, including data encryption and user authentication, are diligently implemented. These measures ensure that sensitive information remains safeguarded throughout the processing pipeline, bolstering user confidence in the system's reliability and compliance with stringent data protection standards.

In essence, this comprehensive integration of LangChain, OpenAI models, Faiss, Streamlit, LIDA, and fine-tuned Mistral 7b models culminates in the creation of a powerful and user-centric platform for content processing, query answering, and data visualization, meticulously tailored to address the multifaceted needs of modern businesses.

5. IMPLEMENTATION :

5.1 PDF Processing:

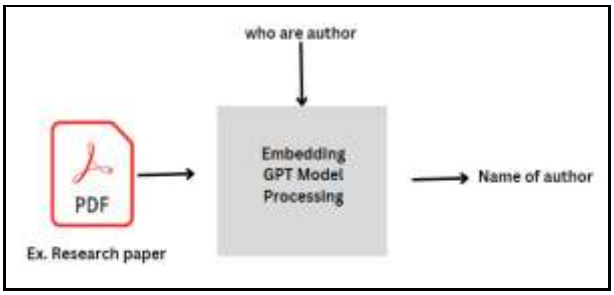
The PDF content and query processing system takes a PDF document as input and breaks it into text chunks, each of approximately 2000 characters with a 100-character overlap factor. These chunks are processed and then transformed into numerical vectors using a language model - Generative Pre-trained Transformer. The resulting embeddings are stored in a directory. These embeddings are stored in pkl(pickle) format. The pickle module is used to arrange Python objects periodically and save them to a file.

To efficiently extract data an indexing tool like Faiss is used in our system. It facilitates the storage and rapid retrieval of embedded vectors, The Langchain framework is used for whole system building which enhances the search process by providing a semantic understanding of the content.

When a user submits a query, it is embedded using the same language model used for the PDF content. Faiss is employed for similarity-based search, quickly identifying chunks that match the query. The Langchain, in conjunction with a language model (LLM), further refines the search by comprehending the query's context and intent. The final output is generated by presenting the relevant text chunks based on their relevance and context, offering users an efficient means to access information within the PDF document. This integrated approach optimizes content retrieval, ensuring accurate and context-aware results.

Facebook AI Similarity Search (FAISS) - It is a library used for quickly searching relevant data in a given document. It provides a semantic search for searching through documents.

The Embedding Process begins with creating a directory named "embeddings" if it doesn't already exist, serving as the storage location for the embedding’s vectors.



Then system takes a file and its original filename as input and stores the document embeddings. It first writes the file to a temporary location and then determines its file extension. Depending on the file type (CSV, PDF, or TXT), it loads the data using the appropriate loader (CSVLoader, PyPDFLoader, or TextLoader) and splits the text into chunks. It then uses the OpenAI Embeddings from OpenAI API to obtain embeddings and stores them in FAISS.

Then retrieval of document embeddings is performed with Langchain. Finally, it loads the vectors from the pickle file and returns them.

5.2 CSV Query Processing:

CSV content processing and query handling mirror PDF processing, where content is embedded and stored. In CSV processing, an additional step involves creating a Data Frame, facilitated by Langchain and Mistral 7B API. When users input queries via the GPT model, the queries undergo rephrasing and embedding. Subsequently, LLM keywords are extracted, and a data frame is generated using Mistral 7B API. Following this, the input content is processed in alignment with the data frame. This streamlined approach ensures uniformity and efficiency across content types, enabling seamless integration of CSV data into the processing pipeline. The utilization of advanced language models such as Mistral 7B enhances the accuracy and effectiveness of query handling and content extraction, facilitating robust data analysis and information retrieval.

5.3 Data Visualization:

Data visualization is the graphical representation of data to facilitate understanding and decision-making. Microsoft's LIDA library automates this process by leveraging large language models (LLMs) to generate accurate visualizations and infographics. Compatible with various programming languages and visualization libraries like Matplotlib and D3, LIDA enables users to create visually compelling representations of data, enhancing interpretability and communication of insights.

5.4 Fine Tunned Mistral 7B Model:

Mistral 7B, a 7-billion-parameter language model by Mistral AI, stands out for its blend of efficiency and high performance, facilitating real-world applications. Upon its debut, Mistral 7B surpassed the leading open-source 13B model, Llama 2, in performance. Leveraging AutoTrain on T4 GPUs, we fine-tuned Mistral 7B to optimize its capabilities further. AutoTrain, seamlessly integrated with the Hugging Face ecosystem, streamlines the process of training and deploying cutting-edge ML models. Comparing the fine-tuned Mistral 7B with Llama, Mistral demonstrated superior performance. Notably, Mistral 7B showcased exceptional speed in data extraction from datasets compared to Llama. This efficiency underscores Mistral 7B's suitability for various applications, emphasizing its role as a versatile and potent tool in the realm of language processing and beyond.

We use auto-train to fine-tune mistral 7B and upload it on hugging face to get ready for usage. Certain parameters were set while fine-tuning the Auto Train process for fine-tuning the LLM model. It specifies the project name, model, data path, and training parameters like learning rate, batch size, and epochs. Additionally, it employs performance-enhancing techniques like PEFT and INT4. The trainer used is SFT, targeting specific modules for training. Finally, it pushes the trained model to a Hugging face repository for further access and deployment. This process enhances the performance and efficiency of the Mistral 7B model for specific tasks.

6. RESULTS AND OUTPUTS :

Previous sections have focused on validating the AI models and components of the process. The system’s outcomes are remarkable, showcasing its effectiveness in processing content, conducting semantic searches, and presenting data visually. By integrating LangChain, LIDA, Faiss, and Fine Tune Mistral 7B models, we’ve established a robust knowledge base capable of accurately responding to user queries across various file formats like PDFs and CSVs. The user interface, designed with Streamlit, ensures a smooth and user-friendly experience, promoting accessibility. Security measures are in place to maintain data integrity, while containerization and continuous integration support scalability. Notably, the system can generate visual graphs based on queries, highlighting its comprehensive and innovative approach to solving problems. In essence, these results underscore the successful integration of cutting-edge technologies, delivering a holistic and user-centric solution.

The system adeptly processes PDF documents by employing a fine-tuned Mistral 7B model for user queries, offering a smooth interaction experience. This functionality enables users to extract pertinent information from PDFs through natural language inquiries, highlighting the efficacy of Mistral 7B in comprehending and responding to varied textual inputs, facilitating streamlined information retrieval and comprehension.

In the case of querying authors within a research paper, the Mistral 7B fine-tuned model efficiently delivers the names of the writers, simplifying the identification process. This demonstrates an efficient and accurate interaction with the document's content, enhancing user-friendliness.

For CSV query processing, the Mistral 7B model is utilized to systematically construct a data frame from CSV data based on specific inquiries, such as listing companies in Mumbai. Through query processing, the relevant information is extracted and organized into a structured data frame, correlating companies with the specified location. This method ensures precision and accuracy in data representation and subsequent analysis, showcasing the model's capability to comprehend and interpret user queries within the CSV context.

Moreover, leveraging the LIDA library, a visualization tool is developed to render CSV data based on user queries. Consisting of summarization, goal exploration, visualization code generation, and infographic creation modules, LIDA facilitates the generation of diverse output graphs. Queries are seamlessly transformed into data frames, enabling efficient code generation for visualization purposes. This comprehensive approach not only enhances data organization but also underscores the models' effectiveness in

Handling various data manipulation tasks, ultimately improving decision-making and insights extraction processes. The integration of Mistral 7B and LIDA showcases a robust methodology for data processing and visualization, offering valuable insights for research and practical applications alike.

7. PERFORMANCE OF FINE-TUNED MISTRAL-7B MODEL :

Results comparing the output of the Fine-tuned Mistral 7B model and Llama for data extraction reveal Mistral 7B's significantly faster performance. The fine-tuning process notably enhances data extraction speed. Mistral 7B outpaces Llama in efficiency, showcasing its superiority in processing data. This improvement underscores the effectiveness of fine-tuning in optimizing performance, offering expedited data extraction capabilities for enhanced efficiency and productivity.

The fine-tuned Mistral 7B model required 7.06 seconds to extract data from a CSV file based on user input, while Llama took 27 seconds for the same task. This substantial difference in extraction time highlights the superior efficiency of Mistral 7B compared to Llama.

The performance disparity between the Fine-tuned Mistral 7B model and Llama for data extraction underscores the tangible benefits of employing advanced AI technologies in streamlining tasks. Mistral 7B's accelerated processing speed signifies a leap forward in efficiency, revolutionizing data extraction processes. This enhanced agility translates into significant time savings, crucial in today's fast-paced business landscape where every second counts.

The fine-tuning process acts as a catalyst for Mistral 7B's remarkable performance, refining its capabilities to meet specific requirements with precision. By tailoring the model to extract data efficiently from CSV files based on user input, organizations can experience a marked improvement in workflow efficiency and productivity. This optimization empowers users to handle large datasets with ease, facilitating quicker decision-making and enhancing overall operational effectiveness.

The contrast in extraction times between Mistral 7B and Llama illustrates the transformative impact of leveraging cutting-edge AI solutions. While Llama remains a competent tool, Mistral 7B's supremacy in speed highlights the evolution of AI-driven technologies in pushing the boundaries of what's possible. This competitive edge positions Mistral 7B as a formidable asset for businesses seeking to streamline data-related processes and gain a competitive advantage in their respective industries.

Furthermore, the significance of fine-tuning cannot be overstated. It not only optimizes performance but also ensures that the model is finely attuned to the intricacies of the task at hand. This tailored approach enhances accuracy and reliability, minimizing errors and maximizing output quality. As such, fine-tuning emerges as a pivotal strategy in unlocking the full potential of AI models, enabling organizations to extract actionable insights from data more effectively than ever before.

In conclusion, the performance comparison between the Fine-tuned Mistral 7B model and Llama underscores the transformative power of AI-driven solutions in revolutionizing data extraction processes. Mistral 7B's superior efficiency, fuelled by fine-tuning, heralds a new era of expedited data processing, offering unparalleled benefits in terms of productivity, efficiency, and competitive advantage.

8. FUTURE SCOPE :

LangChain LLM-based project is expansive, with opportunities to further enhance and diversify its applications. Firstly, we can explore the integration of advanced machine learning techniques to improve the precision and efficiency of data analysis. This could involve incorporating contextual understanding and semantic relationships, allowing the system to provide more nuanced responses to user queries.

In the realm of CSV analysis, encompasses leveraging advanced machine learning algorithms for pattern recognition, anomaly detection, and predictive modelling within tabular data. Integration with natural language processing (NLP) techniques can facilitate deeper insights extraction. Additionally, enhancing scalability and compatibility with big data frameworks will enable efficient processing of large-scale datasets, fostering data-driven decision-making across industries.

Moreover, in the domain of data visualization model entails advancements in interactivity, scalability, and compatibility, alongside integration with emerging data formats. Incorporating AI-driven insights for predictive analytics and real-time visualization, while also enhancing collaborative features, will broaden its applicability across industries and analytical tasks, empowering users to glean deeper insights from complex datasets with greater efficiency and precision.

For the fine-tuning module involves refining its algorithms for increased efficiency and adaptability across various domains. Incorporating advanced optimization techniques, automated hyperparameter tuning, and transfer learning capabilities can enhance its effectiveness in tailoring models to specific tasks, ultimately leading to improved performance and generalization across diverse datasets and scenarios.

Additionally, exploring the incorporation of multilingual support and cross-language understanding would open avenues for a global user base. Collaborations with educational institutions, content creators, and businesses could facilitate real-world testing and refinement, ensuring the project's adaptability and relevance in diverse scenarios. Overall, the future holds immense potential for our project to evolve into a versatile and indispensable tool in the realms of document analysis, data interpretation, and educational support.

9. CONCLUSIONS :

In summation, the fusion of cutting-edge Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies within the framework of LangchainIQ offers a groundbreaking solution to the significant challenges faced by both educators and students in the realms of knowledge acquisition and assessment. This innovative platform, on the cusp of realization through the delineated objectives, holds the promise of instigating a transformative shift in educational methodologies. \par

At its core, LangchainIQ seeks to revolutionize the educational landscape by seamlessly integrating various features such as data visualization with Microsoft's Lida, CSV analytics, and fine-tuning of the Mistral 7b model. This multifaceted approach aims to provide a comprehensive and all-encompassing solution to the diverse needs of both educators and learners. By intricately weaving together these functionalities, the application aspires to create a synergistic blend that addresses the nuances of learning in a technologically advanced era.

One of the pivotal aspects of LangchainIQ lies in its commitment to extracting knowledge from a myriad of formats. Whether it's deciphering information from PDF documents, conducting in-depth analysis on data stored in CSV files, or leveraging the power of video responses, the platform endeavours to cater to the diverse learning preferences and resources available to users. This adaptability underscores LangchainIQ's dedication to inclusivity and accessibility in the educational sphere.

Moreover, the automated data analysis capabilities of LangchainIQ represent a leap forward in streamlining the often time-consuming and labour-intensive process of evaluating student performance. Through the amalgamation of advanced AI algorithms, the platform can provide insightful assessments, enabling educators to focus more on personalized guidance and instructional strategies rather than the administrative burden of assessment.

By employing advanced NLP techniques, the platform strives to generate high-quality, contextually relevant questions that challenge and engage learners. This not only fosters a deeper understanding of the subject matter but also promotes critical thinking and problem-solving skills.

In essence, LangchainIQ stands as a testament to the potential synergy between engineering precision and educational efficacy. It heralds a new era in comprehensive, technology-driven learning solutions that are not only adaptive to the evolving needs of the educational landscape but also empower educators and students alike to navigate the complexities of knowledge acquisition and assessment with unparalleled efficiency and effectiveness.

REFERENCES :

[1] Oguzhan Topsakal1, and Tahir Cetin Akinci (2023). Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. Google Scholar↗

[2] Bagiya Lakshmi S, Sanjjushri Varshini R, Rohith Mahadevan, Raja CSP Raman, (2023). Comparative Study and Framework for Automated Summariser Evaluation: LangChain and Hybrid Algorithms. Google Scholar↗

[3] Rakha Asyrofi; Mutia Rahmi Dewi; Muhammad Irfan Lutfhi; Prasetyo Wibowo (2023). Systematic Literature Review Langchain Proposed. Google Scholar↗ CrossRef/DOI↗

[4] Pedro Neira-Maldonado, Diego Quisi-Peralta (2024). *Intelligent Educational Agent for Education Support Using Long Language Models Through Langchain*. Google Scholar↗

[5] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas (2023). Mistral 7B. Google Scholar↗

[6] Hiren Thakkar; A Manimaran (2023). Comprehensive Examination of Instruction-Based Language Models: A Comparative Analysis of Mistral-7B and Llama-2-7B. Google Scholar↗ CrossRef/DOI↗

[7] Taki, S.M. Abrar Mustakim Kar, Showmick Niloy, Soumik Deb Rakib, Mazharul Islam Biswas, Abdullah Al Nahid (2024). Mitigation of hallucination and interpretations of self attention of Mistral 7B AI to analyze and visualize context understanding ability of large language models. Google Scholar↗

[8] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, Yue Zhang (2024). A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. Google Scholar↗

[9] Lorenzo Mercolli, Thomas Pyka, George Prenosil, Kuangyu Shi (2023). Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?, *Volume 50, pages 1549–1552*. Google Scholar↗

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Google Scholar↗

[11] Yaru Hao, Li Dong, Furu Wei, Ke Xu (2019). Visualizing and Understanding the Effectiveness of BERT. Google Scholar↗

[12] S.L. Freeland & B.N. Handy (1998). Data Analysis with the SolarSoft System, Volume 182, pages 497–500. Google Scholar↗

[13] Steven G. Heeringa, Brady West, Steve G. Heeringa, Patricia A. Berglund, Patricia Berglund (2017). Applied Survey Data Analysis. Google Scholar↗

[14] Tatwadarshi P. Nagarhalli; Vinod Vaze; N. K. Rana (2020). A Review of Current Trends in the Development of Chatbot Systems. Google Scholar↗

[15] H. N. Io; C. B. Lee (2017). Chatbots and conversational agents: A bibliometric analysis. CrossRef/DOI↗

[16] Senay A. Gebreab; Khaled Salah; Raja Jayaraman (2024). LLM-Based Framework for Administrative Task Automation in Healthcare. Google Scholar↗

[17] Mathav Raj J, Kushala VM, Harikrishna Warrier, Yogesh Gupta (2024). Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations. Google Scholar↗

[18] Hemalatha Eed. (2022). A Customized Recommendation System using Streamlit . Google Scholar↗

[19] Saurabh Shukla; Arushi Maheshwari ; Prashant Johri (2021). Comparative Analysis of Ml Algorithms & Stream Lit Web Application. Google Scholar↗

[20] L.R. Bahl; P.F. Brown; P.V. de Souza; R.L. Mercer (1989). A tree-based statistical language model for natural language speech recognition. Google Scholar↗

[21] Hai-Son Le; Ilya Oparin; Alexandre Allauzen (2011). Structured Output Layer neural network language model. Google Scholar↗

[22] P.R. Clarkson; A.J. Robinson (1997). Language model adaptation using mixtures and an exponentially decaying cache. Google Scholar↗

[23] S. Issar (2002). Estimation of language models for new spoken language applications. Google Scholar↗

[24] Anh Tuan Nguyen; Tien N. Nguyen (2015). *Graph-Based Statistical Language Model for Code*. Google Scholar↗

[25] J.R. Bellegarda (2020). *Exploiting latent semantic information in statistical language modeling*. Google Scholar↗

[26] Guillaume Lample, Alexis Conneau (2019). *Cross-lingual Language Model Pretraining*. Google Scholar↗

[27] Godwin George; Rajeev Rajan (2023). A FAISS-based Search for Story Generation. Google Scholar↗ CrossRef/DOI↗

[28] Dimitrios Danopoulos (2019). Approximate Similarity Search with FAISS Framework Using FPGAs on the Cloud, page 373-386. Google Scholar↗

[29] Sanjay Chakraborty; Hrithik Paul (2022). An AI-Based Medical Chatbot Model for Infectious Disease Prediction. Google Scholar↗ CrossRef/DOI↗

[30] Ranci Ren; Mireya Zapata (2022). Experimentation for Chatbot Usability Evaluation: A Secondary Study. Google Scholar↗ CrossRef/DOI↗

****