Leveraging Large Language Model (LLM) for comprehensive data analysis and interactive visual insights through Gen AI.

Vedant Deshmukh

*Computer Science & Engineering*

*Prof. Ram Meghe College of Engineering and Management, Badnera*

Amravati, India

deshmukhvedant007@gmail.com

Maaz Sheikh

*Computer Science & Engineering*

*Prof. Ram Meghe College of Engineering and Management, Badnera*

Amravati, India

maazsheikh3235@gmail.com

Om Bhore

*Computer Science & Engineering*

*Prof. Ram Meghe College of Engineering and Management, Badnera*

Amravati, India

obhore03@gmail.com

Prof. Priyanka Rajesh Patil

*Dept. of Electronics & Telecommunication Engineering*

*Prof. Ram Meghe College of Engineering and Management, Badnera*

Amravati, India

priyanka.patil@prmceam.ac.in

Arraeyan Baig

*Computer Science & Engineering*

*Prof. Ram Meghe College of Engineering and Management, Badnera*

Amravati, India

arraeyanbaig123@gmail.com

# Dr. Dinesh G. Harkut

*Computer Science & Engineering*

*Prof. Ram Meghe College of Engineering and Management, Badnera*

Amravati, India

dinesh.harkut@prmceam.ac.in

***Abstract*—** The integration of Large Language Models (LLMs) in data analysis and visualization represents a significant advancement in making data-driven insights accessible to a broader audience. The paper emphasizes how leveraging Cohere’s LLMs and generative AI can transform traditional data analytics by allowing users to interact with datasets through natural language queries rather than complex coding procedures. By providing an intuitive interface, the proposed system simplifies the process of extracting meaningful insights from data. Users can upload diverse datasets, formulate custom queries in plain language, and receive both textual explanations and corresponding visualizations, such as graphs and charts. This approach not only reduces the learning curve for non-experts but also accelerates the decision-making process by presenting information in an easily understandable format. Moreover, the use of generative AI enhances the system's capability to interpret and respond to nuanced queries, enabling users to uncover hidden patterns, correlations, and trends within large datasets. The model's ability to automatically suggest relevant visualizations based on the context of the query further streamlines the analysis workflow.

***Index Terms*—** **Data Visualization, Large Language Models, APIs, Natural Language Interface, Generative AI, Data Analysis, Machine Learning, Streamlit, Data Visualization, Prompt Generation.**

**1. INTRODUCTION**

 In recent years, advancements in fields of machine learning, natural language processing (NLP), and artificial intelligence (AI) have significantly transformed data processing, information retrieval, and human-computer interaction. These advancements have introduced innovative models and techniques capable of capturing and interpreting user intent from natural language, improving user access to vast data stores. At the forefront of these developments is the integration of transformer-based models, which utilize multi-head attention and deep contextual representations to better understand and generalize across diverse natural language inputs. Despite these advances, several challenges remain in areas such as handling data imbalances, achieving accurate structured query generation, and supporting multilingual or domain-specific information retrieval.

One critical issue in NLP, particularly in classification tasks, is data imbalance. Models often overfit to majority classes, which hinders their effectiveness in minority class prediction and leads to biased or incomplete representations. Transformer models like BERT (Bidirectional Encoder Representations from Transformers) have shown promise in tackling these challenges by capturing nuanced contextual relationships in text. However, further enhancement has come through integrating Large Language Models (LLMs), like OpenAI’s GPT-3.5 Turbo, which can generate synthetic data for minority classes, offering a method to mitigate the data imbalance. This synthetic data augments training sets, enhancing model robustness and fairness, particularly in sensitive NLP applications such as personality classification within the Myers-Briggs Type Indicator (MBTI) dataset. Here, transformer-based approaches combined with LLMs for data augmentation, provide new avenues to enhance model accuracy and generalization in imbalanced textual data.

 In parallel, the evolution of the semantic web has fostered the development of sophisticated information retrieval systems such as KBot—a knowledge graph-based chatbot. KBot addresses challenges in understanding user intent and responding to natural language queries by linking structured knowledge sources like DBpedia and Wikidata. This system harnesses machine learning alongside semantic web technologies, enhancing the chatbot’s ability to handle multilingual inputs and domain-specific queries. With a knowledge graph backbone, KBot facilitates a more intuitive user experience by allowing conversational AI to deliver precise, contextually accurate information across various domains. The use of knowledge graphs in conversational agents represents an effective solution for creating scalable and flexible NLP applications, positioning these systems as key tools for addressing complex information retrieval tasks in a range of industries.

In the domain of interactive information retrieval, systems that convert natural language queries into structured queries such as SQL or SPARQL have gained importance. These systems allow users to access and retrieve structured data stored in databases without requiring knowledge of query syntax. For example, IRNet, a deep learning-based Text-to-SQL model, uses Gated Graph Neural Networks (GGNNs) to encode database structures and relations, thereby enabling the model to understand the intricate schema of relational databases. By integrating GGNNs, IRNet can capture relationships between database entities, improving the accuracy of SQL query generation from natural language inputs. This innovation bridges the gap between unstructured user queries and structured database responses, providing a valuable tool for accessing data in applications like scientific repositories and news archives.

However, traditional web database interfaces still predominantly rely on keyword-based or Boolean logic, limiting their accessibility to users who lack specific database knowledge or expertise. These limitations are particularly noticeable in web-based information services, including scientific databases, code repositories, and data portals. Recognizing this challenge, the authors of the Enhanced Natural Language Interface for Web-Based Information Retrieval paper propose an advanced neural network model based on IRNet. This model interprets natural language queries by converting them into SQL, using GGNNs to encode the database structure and incorporating database values to improve query generation. This expanded IRNet model addresses previous limitations by enhancing entity recognition and relational mappings within databases, allowing the model to more accurately predict tables and columns based on user queries.

The proposed model builds on influential Text-to-SQL frameworks like Seq2SQL, SQLNet, and RAT-SQL, which have each contributed to advancements in converting NL to SQL. These frameworks use deep learning methods to interpret natural language and create structured queries, with each new version building on and addressing the limitations of earlier models.For instance, SQLNet improved upon Seq2SQL by eliminating the need for reinforcement learning, while RAT-SQL used relation-aware self-attention for database schema linking. In analyzing previous models, the authors identify a gap in effective database value representation, essential for accurate column identification and database field matching in NL queries. To address this, the enhanced IRNet model incorporates a representation layer that uses database values in the prediction process, enabling the system to better align NL expressions with relevant database attributes.

The efficacy of this improved model is demonstrated through experimentation on the Spider dataset, a benchmark dataset for Text-to-SQL tasks, which contains thousands of NL queries mapped to SQL statements across a variety of database schemas. The Spider dataset is particularly suited for evaluating complex Text-to-SQL systems because it includes multi-table databases and cross-domain applications. The enhanced IRNet model’s performance improvement is notable, particularly with simpler SQL queries often encountered in real-world web-based information retrieval. By refining table and column predictions, this model shows potential for wider applications in web information services, making data retrieval more accessible to non-expert users.

Across these diverse NLP applications, the combined challenges of data imbalance, intent recognition, and structured query generation require continued methodological innovation to enhance model robustness and usability. The convergence of LLM-based data augmentation, graph-based knowledge representations, and deep learning-based semantic parsing provides a powerful toolkit for addressing these challenges, paving the way for more resilient, accessible AI systems across sectors. Future research will likely focus on further refining accuracy, supporting underrepresented data classes, and expanding multilingual capabilities in NLP applications. These advancements underscore the potential of NLP and AI to provide more equitable, efficient, and user-friendly access to information, with promising implications across fields such as scientific research, personalized user interaction, and large-scale data management.

In recent years, advancements in fields of machine learning, natural language processing (NLP), and artificial intelligence (AI) have significantly transformed data processing, information retrieval, and human-computer interaction. These advancements have introduced innovative models and techniques capable of capturing and interpreting user intent from natural language, improving user access to vast data stores. At the forefront of these developments is the integration of transformer-based models, which utilize multi-head attention and deep contextual representations to better understand and generalize across diverse natural language inputs. Despite these advances, several challenges remain in areas such as handling data imbalances, achieving accurate structured query generation, and supporting multilingual or domain-specific information retrieval.

One critical issue in NLP, particularly in classification tasks, is data imbalance. Models often overfit to majority classes, which hinders their effectiveness in minority class prediction and leads to biased or incomplete representations. Transformer models like BERT (Bidirectional Encoder Representations from Transformers) have shown promise in tackling these challenges by capturing nuanced contextual relationships in text. However, further enhancement has come through integrating Large Language Models (LLMs), like OpenAI’s GPT-3.5 Turbo, which can generate synthetic data for minority classes, offering a method to mitigate the data imbalance. This synthetic data augments training sets, enhancing model robustness and fairness, particularly in sensitive NLP applications such as personality classification within the Myers-Briggs Type Indicator (MBTI) dataset. Here, transformer-based approaches combined with LLMs for data augmentation, provide new avenues to enhance model accuracy and generalization in imbalanced textual data.

In parallel, the evolution of the semantic web has fostered the development of sophisticated information retrieval systems such as KBot—a knowledge graph-based chatbot. KBot addresses challenges in understanding user intent and responding to natural language queries by linking structured knowledge sources like DBpedia and Wikidata. This system harnesses machine learning alongside semantic web technologies, enhancing the chatbot’s ability to handle multilingual inputs and domain-specific queries. With a knowledge graph backbone, KBot facilitates a more intuitive user experience by allowing conversational AI to deliver precise, contextually accurate information across various domains. The use of knowledge graphs in conversational agents represents an effective solution for creating scalable and flexible NLP applications, positioning these systems as key tools for addressing complex information retrieval tasks in a range of industries.

In the domain of interactive information retrieval, systems that convert natural language queries into structured queries such as SQL or SPARQL have gained importance. These systems allow users to access and retrieve structured data stored in databases without requiring knowledge of query syntax. For example, IRNet, a deep learning-based Text-to-SQL model, uses Gated Graph Neural Networks (GGNNs) to encode database structures and relations, thereby enabling the model to understand the intricate schema of relational databases. By integrating GGNNs, IRNet can capture relationships between database entities, improving the accuracy of SQL query generation from natural language inputs. This innovation bridges the gap between unstructured user queries and structured database responses, providing a valuable tool for accessing data in applications like scientific repositories and news archives.

However, traditional web database interfaces still predominantly rely on keyword-based or Boolean logic, limiting their accessibility to users who lack specific database knowledge or expertise. These limitations are particularly noticeable in web-based information services, including scientific databases, code repositories, and data portals. Recognizing this challenge, the authors of the Enhanced Natural Language Interface for Web-Based Information Retrieval paper propose an advanced neural network model based on IRNet. This model interprets natural language queries by converting them into SQL, using GGNNs to encode the database structure and incorporating database values to improve query generation. This expanded IRNet model addresses previous limitations by enhancing entity recognition and relational mappings within databases, allowing the model to more accurately predict tables and columns based on user queries.

The proposed model builds on influential Text-to-SQL frameworks like Seq2SQL, SQLNet, and RAT-SQL, which have each contributed to advancements in converting NL to SQL. These frameworks use deep learning methods to interpret natural language and create structured queries, with each new version building on and addressing the limitations of earlier models.For instance, SQLNet improved upon Seq2SQL by eliminating the need for reinforcement learning, while RAT-SQL used relation-aware self-attention for database schema linking. In analyzing previous models, the authors identify a gap in effective database value representation, essential for accurate column identification and database field matching in NL queries. To address this, the enhanced IRNet model incorporates a representation layer that uses database values in the prediction process, enabling the system to better align NL expressions with relevant database attributes.

The efficacy of this improved model is demonstrated through experimentation on the Spider dataset, a benchmark dataset for Text-to-SQL tasks, which contains thousands of NL queries mapped to SQL statements across a variety of database schemas. The Spider dataset is particularly suited for evaluating complex Text-to-SQL systems because it includes multi-table databases and cross-domain applications. The enhanced IRNet model’s performance improvement is notable, particularly with simpler SQL queries often encountered in real-world web-based information retrieval. By refining table and column predictions, this model shows potential for wider applications in web information services, making data retrieval more accessible to non-expert users.

Across these diverse NLP applications, the combined challenges of data imbalance, intent recognition, and structured query generation require continued methodological innovation to enhance model robustness and usability. The convergence of LLM-based data augmentation, graph-based knowledge representations, and deep learning-based semantic parsing provides a powerful toolkit for addressing these challenges, paving the way for more resilient, accessible AI systems across sectors. Future research will likely focus on further refining accuracy, supporting underrepresented data classes, and expanding multilingual capabilities in NLP applications. These advancements underscore the potential of NLP and AI to provide more equitable, efficient, and user-friendly access to information, with promising implications across fields such as scientific research, personalized user interaction, and large-scale data management.

**2. LITERATURE REVIEW**

In recent years, visual analytics and conversational artificial intelligence (AI) have emerged as transformative fields, enabling enhanced interaction with data and automating conversational tasks, respectively.

Visual Analytics and Data Interaction

The field of visual analytics has seen significant advancements, with research continually pushing the boundaries of how users interact with data. Wenqiang Cui’s survey on visual analytics provides an extensive overview of this domain, highlighting the various dimensions and challenges associated with designing interactive data visualization systems. The study categorizes methodologies by dimensionality and interaction type, a valuable framework that spans from 1D temporal data analysis to complex multidimensional visualizations. One of the primary motivations for developing these systems is to make data more accessible to both novice users and expert analysts. Key challenges in this domain include effectively integrating complex data types, optimizing user interfaces for ease of interpretation, and ensuring scalability in real-time analytics environments.

Additionally, recent studies, such as the "Chat2VIS: Generating Data Visualizations via Natural Language Using ChatGPT, Codex, and GPT-3" by Paula Maddigan and Teo Susnjak, focus on converting natural language into visualizations through large language models (LLMs) like GPT-3 and Codex. Chat2VIS represents a groundbreaking shift in visualization creation, bypassing the need for technical expertise in code generation by allowing users to interact with data visualization tools through natural language commands. The system leverages LLMs and advanced prompt engineering to interpret user queries, providing an intuitive interface that simplifies the process of data exploration and visualization. This study highlights the model’s flexibility, usability improvements, and its ability to handle misspecified and underspecified queries, further emphasizing the practical application of LLMs in complex data interpretation tasks​.

* 1. **Conversational AI: Capabilities and Challenges**

The "Unleashing the Potential of Conversational AI" paper, authored by Vikas Hassija et al., gives a detailed analysis of large language models like ChatGPT, examining their impact across various domains, from customer service to language translation. This survey underscores the architectural and ethical complexities associated with deploying large language models (LLMs), particularly in high-stakes applications where coherence, response accuracy, and ethical considerations are paramount. It explores potential improvements in model generalization and multimodal capabilities, suggesting that these enhancements could help address issues like bias and model interpretability.

LLMs have unlocked a wide range of applications in both commercial and research settings, particularly in customer service, where they enable rapid response generation and user satisfaction improvements. For example, Hassija et al. discuss the integration of ChatGPT in customer service chatbots, where the LLM’s contextual understanding facilitates a more natural conversation flow. Additionally, the paper discusses the potential for LLMs to handle more sophisticated linguistic tasks, such as translation and summarization, which are particularly useful in global and multilingual applications. However, the study highlights ongoing limitations, such as handling ambiguous queries and managing large volumes of training data, which remain challenging areas for LLM research and development.

* 1. **Knowledge Graphs and Conversational AI Over Linked Data**

The intersection of conversational AI with knowledge graphs is another promising area, as illustrated by the study "KBot: A Knowledge Graph-Based ChatBot for Natural Language Understanding Over Linked Data" by Addi Ait-Mlouk and Lili Jiang. KBot leverages knowledge graphs and semantic data to improve chatbot interactions, aiming to address issues in multilingual capabilities, user intent classification, and data integration across multiple knowledge bases. This system focuses on enhancing user query interpretation by translating natural language into SPARQL queries, enabling chatbots to retrieve information from linked data sources like DBpedia and Wikidata. KBot’s architecture, which combines machine learning for intent classification with SPARQL query generation, offers an innovative approach to handling more complex, data-driven queries.

The integration of linked data in conversational AI expands the chatbot’s ability to respond to more specific and analytical queries, as demonstrated through KBot’s use of the myPersonality dataset, which provides insights into social and psychological patterns. This study represents a significant advancement in linking structured data with conversational AI, as it shows how combining knowledge graphs with robust natural language understanding models can enhance user experience by providing responses grounded in structured, accurate data​.

**3. METHODOLOGY**

The methodology section describes the steps involved in the development of the system, starting from data ingestion to generating insights and visualizations. This approach integrates data preprocessing, the integration of Large Language Models (LLMs), and interactive visualizations in order to provide users with a comprehensive tool for real-time data analysis.

**3.1 System Architecture**

The system is designed to interact with the datasets in real-time and give insights about user queries, visualized in dynamic formats. Architectural components include the following:

Data Ingestion: The datasets will either be uploaded by the user or chosen from pre-defined options. The system supports CSV and Parquet file formats.

Data Preprocessing: Data is cleaned up and preprocessed for consistency before feeding into the model.

Model Integration: Utilize the Cohere's API in order to take the users' queries, process these, and present relevant insights from the data.

Visualizations Generation: Use the Matplotlib and Seaborn in the generation of dynamic visualizations of insights.

User Interface: It is done through Streamlit, whereby one can have an interaction by inputting queries and the kind of chart wanted.

**3.2 Data Ingestion**

The uploads of any of these kinds are then presented to the system in the format it has to be processed. Firstly, selections can be made like, for instance, Iris, Titanic from existing datasets available online. Alternatively, user created CSV and Parquet file can also be uploaded by the users. These uploads are saved in the system and the data thus can be utilized further in its processing.

**3.3 Data Preprocessing**

Once the data is imported, it is then further processed using Pandas and NumPy to fill in missing values, normalize data, and encode categorical features. This step ensures that the data are in a consistent format suited for analysis by the LLM model.

**3.4 User Query Input**

The next step is to receive queries from users. The query could be in natural language as a question that they might have about the data set. For example, one could ask, "What are customer sales trends over time?" or "Visualize distribution of age in the dataset." These queries are received, parsed, and fed into the system's interface in an interactive manner.

**3.5 Model Integration**

Cohere's API does insights generation based on what the user is querying into the dataset. The API will be called to assess the data and provide contextually relevant insights. The model is designed to generate a response that summarizes trends, detects patterns, or even offers statistical summaries.

**3.6 Insight Generation**

Once the model processes the query, it provides a list of insights gleaned from the data. These might be a summary, a correlation, or a prediction depending on what the user has asked for. By this point, interpretation of results and preparation for visualization happen.

**3.7 Generation of Visualization**

All insights produced from the model will be represented through interactive charts and graphs. Depending upon the choice of the user, charts of various kinds (for instance, bar charts, pie charts, line graphs) will be representations of the data. Matplotlib and Seaborn are also used in producing visualizations in an easy-to-understand manner.

**3.8 Error Handling and Retries**

During the API calls of the model, it uses error handling mechanisms, which address issues such as timeouts or data inconsistencies. In case of an error, the system will automatically retry the request after a brief delay to ensure that the user receives a response.

**3.9 User Feedback and Evaluation**

After using the system, the users are asked to comment on the usability of the tool, the clarity of the insights, and the effectiveness of the visualizations. This feedback is collected from the survey and analyzed for improvement.

**3.10 Performance Metrics**

The performance of the system is measured along three parameters:

Accuracy: The relevance and correctness of the insights generated by the model.

Latency: The time the system takes to process the data and return the answers to users.

User Satisfaction: This is a subjective experience of the user - based on their feedback on the interface and insights.

**4. RESULTS**

**4.1 Case Study: Indian Food Dataset**

Following is the example of dataset consisting of data about indian food cuisines. The user can upload any dataset (in .csv or .parquet format) up to 200mb. The figures displayed below illustrates the three interfaces of the tool namely: Upload Dataset, Ask me Anything and Visualization. The first query demonstrated here in ‘Ask me Anything’ section is “What is required for gulab jamun". The tool analyses the dataset and provides the required answer to the user’s query. The second query is where we illustrate a pie chart in the Visualization section by giving the parameter of ‘course’ from our dataset.



Fig 1: User Interface



Fig 2: Uploading Dataset



Fig 3: Chatbot Interface



Fig 4: Query 1 Resolution by chatbot



Fig 5: Visualization Interface



Fig 6: Pie chart Query



Fig 7: Pie chart Query Resolution

CONCLUSION

This study demonstrates how large language models (LLMs) can effectively bridge the gap between natural language queries and data visualization, making data analysis accessible to non-technical users. By integrating Cohere’s LLM with a user-friendly interface, the system interprets diverse queries to generate relevant insights and visualizations, enabling intuitive data interaction. This approach circumvents the rigidity of rule-based NLIs, offering a flexible, cost-effective solution for transforming natural language inputs into actionable analytics. Our tool shows promise for democratizing data-driven decision-making by allowing users with minimal programming knowledge to explore complex datasets visually. Future improvements will focus on expanding visualization options and refining LLM accuracy to further support seamless, meaningful data engagement across broader contexts and use cases. This work contributes to advancing LLM-based interfaces as powerful, adaptable solutions in data science and beyond.

REFERENCES

1. Maddigan, P., & Susnjak, T. (2023). Chat2VIS: generating data visualizations via natural language using ChatGPT, codex and GPT-3 large language models. *Ieee Access*, *11*, 45181-45193.
2. Ait-Mlouk, A., & Jiang, L. (2020). KBot: a Knowledge graph based chatBot for natural language understanding over linked data. *IEEE Access*, *8*, 149220-149230.
3. Cui, W. (2019). Visual analytics: A comprehensive overview. *IEEE access*, *7*, 81555-81573.
4. Bai, T., Ge, Y., Guo, S., Zhang, Z., & Gong, L. (2020). Enhanced natural language interface for web-based information retrieval. *IEEE Access*, *9*, 4233-4241.
5. Hassija, V., Chakrabarti, A., Singh, A., Chamola, V., & Sikdar, B. (2023). Unleashing the potential of conversational AI: Amplifying Chat-GPT’s capabilities and tackling technical hurdles. *IEEE Access*.
6. Pandey, A. K., & Roy, S. S. (2024). Extractive Question Answering over Ancient scriptures Texts using Generative AI and Natural Language Processing Techniques. *IEEE Access*.
7. Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A survey of text representation and embedding techniques in nlp. *IEEE Access*, *11*, 36120-36146.
8. Kuhail, M. A., & Lauesen, S. (2020). Uvis: A formula-based end-user tool for data visualization. *Ieee Access*, *8*, 110264-110278.
9. Gopali, S., Abri, F., Namin, A. S., & Jones, K. S. (2024). The Applicability of LLMs in Generating Textual Samples for Analysis of Imbalanced Datasets. *IEEE Access*.
10. Rostam, Z. R. K., Szénási, S., & Kertész, G. (2024). Achieving Peak Performance for Large Language Models: A Systematic Review. *IEEE Access*.
11. Omar, M., Choi, S., Nyang, D., & Mohaisen, D. (2022). Robust natural language processing: Recent advances, challenges, and future directions. *IEEE Access*, *10*, 86038-86056.
12. Xu, B., Cai, R., Zhang, Z., Yang, X., Hao, Z., Li, Z., & Liang, Z. (2019). NADAQ: natural language database querying based on deep learning. *IEEE Access*, *7*, 35012-35017.
13. Park, D., An, G. T., Kamyod, C., & Kim, C. G. (2023). A Study on Performance Improvement of Prompt Engineering for Generative AI with a Large Language Model. *Journal of Web Engineering*, *22*(8), 1187-1206.
14. Neumann, A. T., Yin, Y., Sowe, S., Decker, S., & Jarke, M. (2024). An LLM-Driven Chatbot in Higher Education for Databases and Information Systems. *IEEE Transactions on Education*.
15. Liu, J., Tang, T., Wang, W., Xu, B., Kong, X., & Xia, F. (2018). A survey of scholarly data visualization. *Ieee Access*, *6*, 19205-19221.
16. Nasution, A. H., & Onan, A. (2024). ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-resource Language NLP Tasks. *IEEE Access*.
17. Bilgram, V., & Laarmann, F. (2023). Accelerating innovation with generative AI: AI-augmented digital prototyping and innovation methods. *IEEE Engineering Management Review*, *51*(2), 18-25.