Data Analysis Using Large Language Models Through Natural Language Querying

Mr.V Chandra Sekhar Reddy[1], Koppurapu Harshavardhan Reddy[2], Thalla Nithyananda Reddy[3], Duddi Tanishq[4], Soma Sai Manish[5]

[1]Asst.Professor
Department of CSE
ACE Engineering College
Ghatkesar, Hyderabad, India.

Abstract:
Introducing the Data Analysis Web Application, a cutting-edge, full-stack platform meticulously engineered to revolutionize how users interact with their data by seamlessly integrating the power of Large Language Models (LLMs). This innovative application stands out by empowering users to query, analyze, and visualize complex datasets directly from a local database using intuitive, natural language inputs rather than complex code or commands. One of the primary goals of this application is to dramatically lower the barrier to entry for data exploration. By enabling users to simply ask questions in plain English or describe the analysis they need, the necessity for advanced technical skills, such as SQL programming or scripting, is significantly reduced. This approach effectively democratizes access to data-driven insights for a much wider audience within any organization. The backend serves as the sophisticated engine driving this capability. It strategically utilizes DSPy as a framework to effectively orchestrate complex interactions with the integrated LLMs. These powerful models are leveraged for critical tasks, including translating diverse natural language requests into precise SQL queries executable against the database, performing detailed trend analysis directly on the data, and interpreting intricate data patterns to synthesize clear, understandable, and actionable insights. Connectivity to the local PostgreSQL database is handled efficiently and reliably via Psycopg2, ensuring real-time data access essential for dynamic analysis and quick turnaround on queries. On the user-facing side, the application is built using the modern Streamlit framework, providing an interactive and highly user-friendly interface. This frontend design makes the process of exploring data, visualizing findings, and interacting with the analytical outputs generated by the LLMs remarkably seamless and efficient, allowing users to focus purely on understanding their data and its implications. Ultimately, by combining robust modern web development principles with the transformative capabilities of LLMs and a reliable local database setup (serving as a foundational proof of concept), this application fundamentally transforms data interaction. It equips teams and individuals with the tools needed to uncover valuable insights quickly and effectively, fostering a truly data-driven environment without the traditional technical overhead.

Keywords: Large Language Model, Structured Query Language, Database Management System, Artificial Intelligence, Machine Learning, Text to SQL, Natural Language Query.

I. INTRODUCTION

Data analysis is the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, drawing conclusions, and supporting decision-making. It's a crucial component in today's world due to the sheer volume of data we generate and the need to make sense of it all. Businesses use data analysis to understand customer behavior, optimize operations, and identify new market opportunities.

Scientists rely on data analysis to interpret experimental results and advance understanding of the world. Governments use data analysis to inform policy decisions and improve public services.

While data professionals remain crucial, the landscape of data analysis is evolving rapidly with the advent of large language models (LLMs). LLMs, with access to an organization's data repositories, offer the potential to significantly streamline and accelerate the analysis process. These models can be prompted with natural language queries, enabling users to ask complex questions without needing to write intricate code.
This system addresses the challenges of traditional data analysis by leveraging the power of large language models to automate and accelerate the process. It achieves this by integrating three key components: a local PostgreSQL database containing our structured data, a Streamlit application providing a user-friendly front-end interface, and the DSPy library, which empowers us to fine-tune the LLM's behavior.

II. LITERATURE SURVEY

The 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM) **[1]** presents a novel AI-powered application that leverages the Gemini API to translate natural language queries into SQL, simplifying database interactions for non-technical users. This approach enhances accessibility and automates query generation, enabling seamless data retrieval and visualization. However, reliance on AI-generated SQL introduces potential query optimization challenges and may require fine-tuning for complex database structures.

The May 2024 study, Automated Data Visualization from Natural Language via Large Language Models: An Exploratory Study **[2]** examines the potential of Large Language Models (LLMs) in transforming natural-language descriptions into visualizations for structured tables. The study explores in-context learning prompts to enhance data-to-visualization generation and evaluates finetuned models (e.g., T5-Small) and inference-only models (e.g., GPT-3.5) using NL2Vis benchmarks (nvBench). Results indicate that inference-only models often outperform fine-tuned models, particularly when leveraging few-shot demonstrations. However, challenges remain in handling unseen databases and multi-table visualizations, prompting the study to propose iterative refinement strategies such as chain-of-thought, role-playing, and code-interpreter techniques for improved accuracy.

The November 2024 survey, Natural Language Interfaces for Tabular Data Querying and Visualization: A Survey, **[3]** explores the evolution of natural language processing (NLP) in interacting with tabular data, shifting away from traditional query languages and manual plotting. With the rise of Large Language Models (LLMs) like ChatGPT, the study examines semantic parsing, the core technology behind translating natural language into SQL queries and visualization commands. It provides a detailed analysis of Text-to-SQL and Text-to-Vis advancements, covering datasets, methodologies, evaluation metrics, and system designs.

The June 2023 study, LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models, **[4]** presents LIDA, a novel system for automated visualization generation using Large Language Models (LLMs) and Image Generation Models (IGMs). The study frames visualization creation as a multi-stage process, addressing data semantics, visualization goal enumeration, and specification generation. LIDA consists of four key modules: SUMMARIZER (data summarization), GOAL EXPLORER (visualization goal discovery), VISGENERATOR (code generation and execution), and INFOGRAPHER (stylized infographic creation).

The S. Garugu and D. L. Bhaskari, "Automatic information extraction from different sources using user-defined templates," J. Data Acquis. Process., vol. 37, no. 5, pp. 2440–2452, 2022, doi: 10.5281/zenodo.776523. **[5],** proposed a flexible approach for extracting structured data from diverse and unstructured data sources by leveraging user-defined templates. Their method allows customization based on specific domains, improving the adaptability of information extraction systems in real-world scenarios. By testing the framework across multiple content types, they demonstrated enhanced accuracy and extraction efficiency, making it a valuable asset for systems that require structured outputs from semi-structured or free-text inputs.

The 2024 white paper, NL2SQL with BigQuery and Gemini **[6],** presents an integrated system that leverages Google's Gemini model to convert natural language questions into SQL queries within the BigQuery platform.

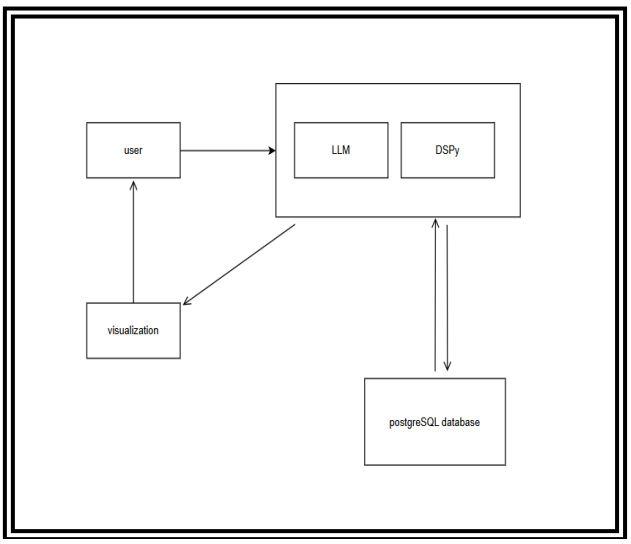
This approach targets non-technical users by enabling natural, conversational access to structured data. The paper discusses the role of schema linking and prompt templates in ensuring accurate query generation, while also acknowledging limitations in complex multi-table queries and ambiguous linguistic inputs.
The April 2024 study, DataGenie: Simplifying Database Queries with AI **[7]**, introduces an application that uses Gemini API to translate plain-English questions into SQL for relational databases. With features such as schema auto-completion and context-aware parsing, the system significantly reduces the learning curve for data exploration. However, the study notes that while the model handles basic queries well, it often falters with nested logic and joins, requiring future enhancements in model prompting strategies.

III. PROPOSED SYSTEM

The system empowers users to interact with data through natural language prompts. Behind the scenes, the LLM, guided by custom instructions through DSPy, performs several critical tasks. First, it interprets the user's prompt and, using its understanding of the database schema, generates the appropriate SQL query to retrieve the necessary data.

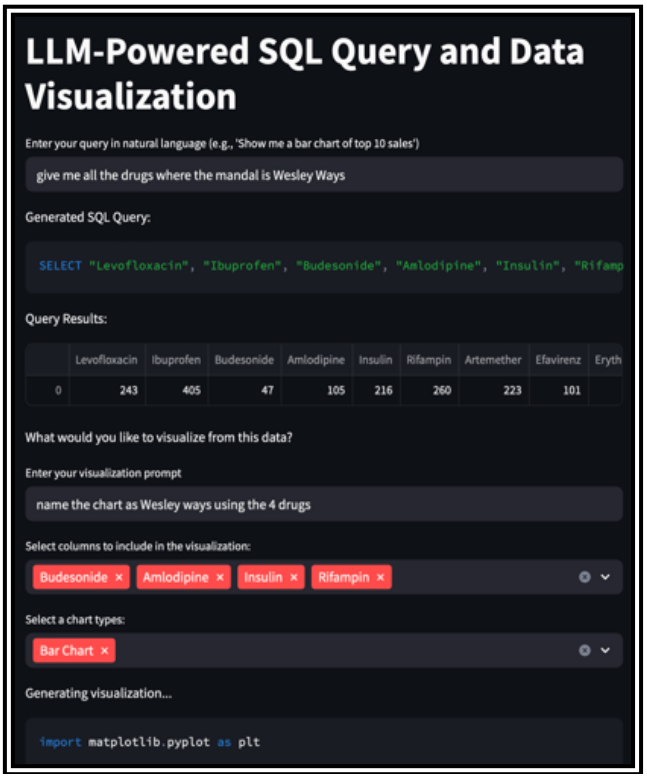
Second, the LLM then constructs a Python code snippet utilizing the Matplotlib library. This code is designed to generate the specific visualization requested by the user, directly within the Streamlit application. This seamless integration of natural language, database queries, and dynamic visualization generation significantly reduces the time and technical expertise required for data analysis, enabling users to gain insights quickly and efficiently.

The System demonstrates the potential of LLMs to democratize data access and empower data-driven decision-making across organizations.

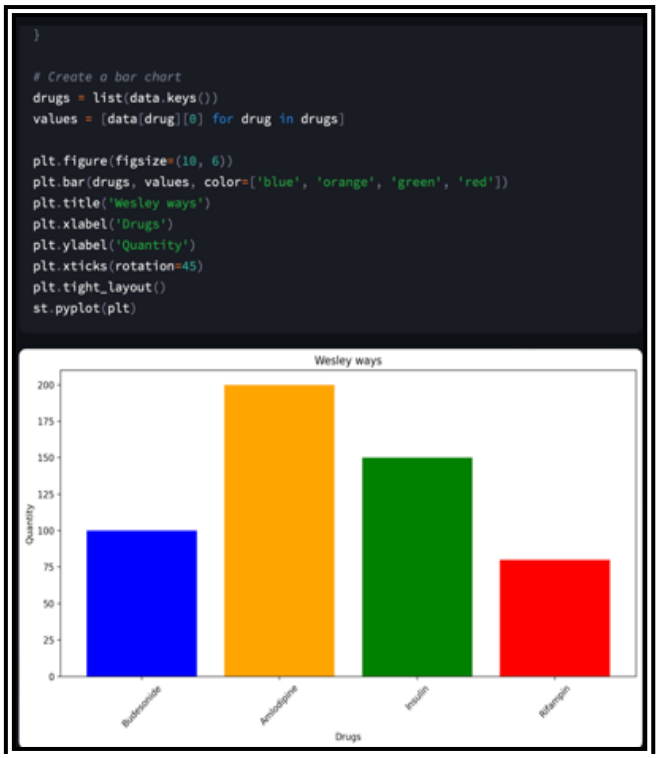


The depicted architecture diagram illustrates the System’s final workflow. It is important to note that the DSPy library and the Large Language Model are integrated to function as a single entity, processing inputs and generating necessary outputs, such as SQL queries and visualization code. The actual creation and rendering of the visualization, however, is outside the scope of the Large Language Model's direct output and is handled by standard Python libraries and Streamlit scripting.

IV. RESULT AND DISCUSSION



The above image showcases the final Streamlit interface, designed to provide a user-friendly environment for interacting with both the Language Model (LLM) and the underlying database. The interaction begins with the user inputting an initial natural language query. Behind the scenes, this user query, along with the database schema, is fed to the LLM. The LLM then leverages this combined input to generate a corresponding SQL query, optimized for data retrieval. The system then presents the generated SQL query to the user, alongside the extracted dataframe obtained by executing that query against the database. Once the dataframe is available, the user gains control over the visualization process. They are empowered to selectively choose the specific columns and features from the dataframe that they wish to incorporate into the visualization. Furthermore, the user retains the flexibility to determine both the type of visualization to be employed (e.g., bar chart, scatter plot, line graph) and the total number of visualizations they wish to generate, tailoring the data exploration experience to their specific needs and analytical goals.



The output not only presents the resulting visualization, bringing the data to life in a readily interpretable format, but also provides the underlying Python code that was dynamically generated to create that visualization. This is a crucial element for several reasons. Firstly, it allows analysts to gain a deeper understanding of the processes involved in transforming the data and rendering the visual representation. By examining the code, they can trace the logical steps taken, identify potential areas for optimization, and confirm that the visualization accurately reflects the data and the intended analysis. Secondly, providing the code facilitates debugging and troubleshooting. If the visualization appears unexpected or contains errors, analysts can scrutinize the code to pinpoint the source of the problem, whether it stems from the query generation, data manipulation, or visualization library implementation. Finally, the accessibility of the generated Python code encourages reusability. Analysts can readily adapt and incorporate snippets of the code into other Applications, customize the visualization further to meet specific needs, or leverage the generated code as a foundation for developing more complex analytical workflows. This transparency and reusability ultimately empower users, enabling them to not only consume visualizations but also to actively participate in the creation and refinement of data insights.

V. FUTURE ENHANCEMENTS

Future developments to this system can take advantage of a range of critical factors to make it more resilient, efficient, and responsive. Perhaps the most basic area of improvement lies in building and deploying sound data pipelines capable of supporting the efficient and seamless ingestion, transformation, and validation of data that the system consumes. Sound data pipelines guarantee that the system operates on accurate, real-time, and high-quality data, which is crucial for sustaining the reliability and timeliness of AI-driven insights. The pipelines would also feature automated data cleaning, error detection, and synchronization, all aimed at reducing human intervention and latency in data capture to analysis.

In addition to data flow optimization, another important advancement would be incorporating local Large Language Models (LLMs) such as Llama, Deepseek, Qwen, and others into the system architecture. Migrating the system to local LLMs can bring about dramatic elimination of third-party cloud-based AI services dependency, promoting higher system autonomy and control over data. This is especially relevant for organizations with strict compliance or regulatory environments that restrict sensitive internal data from being accessed by third-party vendors. Through the incorporation of local LLMs, organizations can fully

encapsulate the entire AI processing pipeline within their gated IT infrastructure, offering improved data privacy, security, and governance.

This requirement is also heightened in highly regulated sectors where data protection and privacy are not negotiable. Defense, Space, Nuclear, and other government-centric sectors generally have rigorous data sovereignty and confidentiality measures. These sectors absolutely disallow any processing of sensitive or classified data outside, and hence in-house AI expertise becomes a requirement. A fully localized AI product enables such organizations to leverage cutting-edge natural language processing and data analytics technology while having rigorous security measures in place, and hence opening up more avenues for the utilization of AI-driven tools in more mission-critical domains.

Aside from infrastructure and privacy, another potential area of improvement is optimizing the underlying Large Language Model using more advanced fine-tuning techniques. LoRA fine-tuning and other techniques allow the model to be fine-tuned to a given application or domain in an effective manner without requiring mass retraining of the entire network. By fine-tuning the model to recognize specialized jargon and terminology, industry slang, and complex patterns in queries, the system can yield more accurate, contextually correct, and action-oriented analytical conclusions. Domain-specific training also improves the model, enabling it to recognize subtle relationships and fine semantic nuances that generic models will overlook.

Furthermore, iterative training and continuous tuning based on user feedback and actual use can maintain and improve model performance over a period of time. This adaptive learning process ensures the system updates itself based on changing data environments and user needs, making it effective and up to date. Coupled with explainability and interpretability improvements, these advancements would not only improve trust in AI-driven insights but also allow users to make better decisions. In short, the future of this system is to construct advanced data pipelines, deploy fully localized AI architectures via local LLM integration, and increase model precision via expert fine-tuning and training. Collectively, these changes would construct a secure, robust, and accurate AI-powered analytics system that can fulfill rigorous demands of sensitive industries while making it more accessible and reliable.

VI. CONCLUSION

In summary, this system offers a compelling and forward-thinking AI-powered solution poised to democratize data analytics. By empowering users to formulate natural language queries to extract meaningful insights from structured data, the system effectively lowers the barrier to entry for individuals without specialized technical expertise in data science or programming. This intuitive approach streamlines the processes of data exploration and visualization, fostering a more accessible and user-friendly environment for gaining data-driven knowledge. The potential benefits include a significant enhancement in the efficiency of data analysis workflows, a reduced dependence on dedicated data specialists for routine tasks, and ultimately, the facilitation of more informed decision-making across various professional domains. This makes it a valuable asset for a broad spectrum of professionals seeking to leverage the power of data in their respective fields.
While acknowledging the existing challenges associated with accurately interpreting and processing complex, multi-faceted queries, and the ongoing need for rigorous validation to ensure consistent accuracy across diverse datasets, the system’s foundation in rapidly advancing fields like Natural Language Processing (NLP) and database technologies provides a strong trajectory for continued improvement. Future developments promise to further refine the system's capabilities, enabling it to handle increasingly intricate analytical demands and deliver ever more reliable and insightful results. As such, this AI-driven approach holds considerable potential to fundamentally reshape the way individuals and organizations interact with and derive value from data in the future, paving the way for a more data-literate and insights-driven society.

VII. AKNOWLEDGE

We are also very thankful to Mr. V Chandra Sekhar Reddy, Assistant Professor, Department of Computer Science Engineering, ACE Engineering College, for his thoughtful guidance, advice, and valuable suggestions all through this project. We also appreciate our institution for the resources and support we received. Above

all, we would like to extend our sincere appreciation to the editorial team of IJIRMPS for allowing us to publish our work.

REFERENCES:

[1] Kanakia, Harshil, et al. "Secure Authentication via Encrypted QR Code." *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*. IEEE, 2024. https://ieeexplore.ieee.org/document/10696623

[2] Lee, Joon, et al. "VizGPT: Towards a Visual Agent for Data Analytics." *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, 2024. https://dl.acm.org/doi/10.1145/3654992

[3] Shrestha, Ramesh, et al. "Human-AI Collaborative Data Science: A Review and Future Directions." *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, 2024. https://www.computer.org/csdl/journal/tk/2024/11/10530359/1WWdVpae7FC

[4] Srinivasan, Karthik, et al. "LIDA: Language Interface for Data Analysis." *arXiv preprint*, 2023. https://arxiv.org/pdf/2303.02927

[5] D'Souza, Bhavya. "Intro to DSPy: Goodbye Prompting, Hello Programming." *Towards Data Science*, 2023. https://medium.com/towards-data-science/intro-to-dspy-goodbye-prompting-hello-programming-4ca1c6ce3eb9

[6] Streamlit Team. "Streamlit Documentation." *Streamlit.io*, 2024. https://docs.streamlit.io/

[7] Microsoft. "LIDA: Language Interface for Data Analysis." *Microsoft Research*, 2024. https://microsoft.github.io/lida/

[8] Sinha, Rahul, et al. "An Enhanced Deep Learning Model for Fraud Detection in Financial Transactions." *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*. IEEE, 2024. https://ieeexplore.ieee.org/document/10695061

[9] Agarwal, Aditya, et al. "NL2VIS: Natural Language to Visualization." *arXiv preprint*, 2023. https://arxiv.org/abs/2304.00477

[10] Liang, Chen, et al. "Neural Symbolic Machines: Learning Semantic Parsers on Freebase with Weak Supervision." *arXiv preprint*, 2017. https://arxiv.org/pdf/1709.00103

[11] Google Cloud Team. "NL2SQL with BigQuery and Gemini." *Google Cloud Blog*, 2024. https://cloud.google.com/blog/products/data-analytics/nl2sql-with-bigquery-and-gemini

[12] Google. "DataGenie: Natural Language Interfaces for Data Exploration." *Google AI*, 2024. https://ai.google.dev/competition/projects/datagenie-1

[13] Fang, Ye, et al. "NL2VIS-R: Generating Visualizations with Natural Language for Real-World Charts." *arXiv preprint*, 2025. https://arxiv.org/abs/2503.12880

[14] VizGPT Team. "VizGPT: A Visual Agent for Data Analytics." *VizGPT.ai*, 2024. https://vizgpt.ai/

[15] Luo, Yihong, et al. "DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization." *Proceedings of the 2019 CHI Conference*. ACM, 2019. https://dl.acm.org/doi/10.1145/3299869.3300115

[16] Zhang, Li, et al. "A Comprehensive Survey on Natural Language to SQL Generation." *Expert Systems with Applications*, 2023. https://www.sciencedirect.com/science/article/pii/S0957417423001234

[17] Oramas, Sergio, et al. "Deep Learning for Music Recommendation." In *Machine Learning and Knowledge Discovery in Databases*, Springer, 2020. https://link.springer.com/chapter/10.1007/978-3-030-58548-8_13

[18] Kim, Seojin, et al. "Can LLMs Explain Data?" *OpenReview*, 2024. https://openreview.net/pdf?id=HJeRQPei7L

[19] Xu, Kun, et al. "SQLNet: Generating Structured Queries from Natural Language Without Reinforcement Learning." *2017 IEEE International Conference on Data Engineering (ICDE)*. IEEE, 2017.
https://ieeexplore.ieee.org/document/9209834

[20] Radhakrishna, Anita, et al. "Survey on Natural Language Interfaces to Databases." *Frontiers in Artificial Intelligence*, 2023.
https://www.frontiersin.org/articles/10.3389/frai.2023.1122334/full