**Title Page**

* Title of the Paper
* Authors’ Names and Affiliations
* Contact Information

**1. Introduction**

**Background**

* Understanding data is difficult, resource intensive and require technical knowledge. Data scientists use innovative tables and figures and write the description to interpret them to the general audience. Despite the summarized data remain underutilized due to
* Data scientists often use data from different sources or receive some data anew which they do not have any background knowledge about the data or metadata. Data scientist spends significant amount to time to understand the data generating protocol and meta. After that they become capable of interpreting data for the audience. Often this understanding domain knowledge and data generating process is not straightforward and exploratory. The learning about data progress over time. Which often documented with less care. Consequently, if the team changes, the new person starts to learn the process from the scratch. It is crucial to document the domain knowledge, understanding about the data generating process, data, and metadata documented and aligned so that anybody get benefitted from the data both technical and non-technical user.
* The integrated data documentation process is not new however it requires plan, and well qualified team and infrastructure to maintain the total data asset. This is almost impossible for the small data analytic projects. Analysists usually get the data, get metadata for their analytic perspective, and prepare summary report based on analytic, after that, all background materials of the report left behind in the analysts compute of some sharepoint which often makes the report impossible to re-produce. The situation becomes worse when the database is complex.
* A system that would integrate the data, metadata, and background information and learning to the analytic report would be a solution as done in some dashboards. This has a potential difficulty of interpreting results to the audience.
* Large Language Models (LLMs) are a type of artificial intelligence that can understand, generate, and process human-like text and other content. They are trained on vast amounts of data, which allows them to recognize patterns, understand context, and produce coherent and relevant responses. LLMs enhance user engagement by providing more natural, personalized, and efficient interactions. Instead of following rigid, pre-programmed rules, they can respond to complex queries in a more conversational manner, making the experience feel more human and intuitive. This leads to higher user satisfaction and deeper connections with a product or service. LLM is a new tool that can understand human language and able to interpret the analytic results with high accuracy. Even AI can analyze data, perform difficult query and return the necessary codes and results if properly prompted. Integration of LLM in data analytic output like data dash or web page could be a smart solution for data.
* The integration of chatbots with traditional data dashboards represents a significant evolution in how users interact with business intelligence. Historically, dashboards have been static, visually rich interfaces that present data in a predefined manner through charts and graphs. While effective for monitoring key performance indicators (KPIs), they often require users to manually filter, pivot, and drill down to find specific insights. The introduction of an AI chatbot transforms this passive experience into an active, conversational one. This novel approach essentially creates a "conversational dashboard" where a chatbot acts as a virtual data analyst, allowing users to pose questions and receive tailored, on-demand insights without navigating complex menus or needing technical expertise.
* This integration facilitates a more dynamic and personalized user experience by enabling real-time data interaction through natural language. Instead of a user having to find and apply the correct filters to see regional sales data, they can simply type or speak a request such as, "What were the sales figures for the Northeast region last quarter?" The chatbot then processes this natural language query, translates it into a structured data query, and presents the answer in a conversational format or, even more powerfully, by automatically updating the dashboard's visualizations to show the requested data. This not only speeds up the time to insight but also democratizes data access, empowering non-technical users across an organization to ask complex, ad-hoc questions and receive immediate, data-driven answers.
* While promising, existing solutions in this space still face several limitations, highlighting the need for further innovation. Many current implementations are often an add-on to a dashboard platform rather than a deeply integrated core feature, which can lead to a fragmented user experience. Furthermore, many of these systems have challenges with contextual awareness and a lack of personalization. For example, a chatbot might struggle to understand a follow-up question that refers to a previous query, or it may not be able to tailor its responses based on a user's role or access permissions. Companies like Microsoft with Power BI Copilot and Snowflake with their conversational analytics tools are beginning to address these issues, but the field is still ripe for advancements in natural language understanding, sentiment analysis, and the ability to proactively highlight anomalies and trends without an explicit user prompt.

**2. Objectives**

* A streamlined system of analysis and representation can reduce statisticians’ coding burden and make results reusable over time. It would help researchers and statistician access and build on previous work easily, saving time, improving consistency, and enhancing collaboration. Integrating large language models (LLM) into the system would enable researchers to easily get information about the data and interpret the data.
* Therefore, we aimed to develop a data agent that integrates LLMs in analytics. The smart data agent contains the basic analytics of data which are interpreted in the pages. LLM’s behaviors is programmed so that it can respond all queries that might appear from the data.
* This would be an initiative to bridge the gap between Data scientist and user. The primary audience would be the researchers who are involved with n
* PRO-ACT is a large database which contains longitudinal data from multiple completed Phase II/III clinical trials for ALS. This includes multiple data tables and >13K ALS subjects. The data collection protocol is a mix of different study protocols, and the study population is diverse [1]. Therefore, A better understanding about data and metadata is required to get useful information from the data.
* Therefore, the project intends to develop such LLM supported R/Python Shiny app for PRO-ACT database, the largest database for ALS trial data to support TAK-138 in early and late phase development.

**3. Methodology**

**Paragraph 1: Design of the Chatbot**

Python/R shiny as will be developed integrating data, codebook, metadata. A dynamic log file was integrated with the app so that, the team can write the journal of learning about the data. Summary and analytic outcomes such as table and figures and their general description and interpretation are incorporated in pages and tabs in each pages of the dashboard.

In the dash board, A separate page was allocated to chat with the LLM. LLM was accessed through Application Programming Interface (API). We have integrated Gemini developed by Google and Generative Pre-trained Transformer (GPT) developed by OpenAI. User can select the model to get the advantage of consulting the result.

We developed a generic prompt that guides the behavior of the of AI in responding users’ query. A schema for the integrated data base is also included in the prompt. A custom function generated the schema from the data and codebook and include the schema in the generic prompt to guide LLM about the content of the database. Analytic results were also amended to in the prompting.

The chat page has a query box, and respond on the screen when submitted. The conversations are stored in a log file to be used by the analysts and developers to improved the users chat experience.

The code was written in Python 3 and can be published through a dedicated server.

To enhance functionality, conversational dashboards often rely on a variety of external APIs including OpenAI and Google and a variety of Python libraries. We used … libraries to develop analyzing and visualize the critical components, understand users query. The LLM returns. The Additionally, backend services for data querying and retrieval, such as those offered by DuckDB. The seamless integration of these tools is crucial for delivering a robust and responsive user experience.

We used Pooled Resource Open-Access Amyotrophic Lateral Sclerosis Clinical Trials database (PRO-ACT) Database [ref]. Pro-ACT is a large database which contains longitudinal data from multiple completed Phase II/III clinical trials for ALS. This includes multiple data tables and >13K ALS subjects. The data collection protocol is a mix of different study protocols, and the study population is diverse. Therefore, A better understanding about data and metadata is required to get useful information from the data. The public health researchers need that additional background to interpret the data therefore we consider this data to demonstrate the application. The data tables were reorganized for the ease of integration. We performed descriptive analytic to demonstrate clinical tail outcomes, patients survivals in the dashboard.

A update option was also integrated to the system. An inclusion in published data will update the data automatically.

This is relatively a novel idea and we have limited knowledge and experience of standards. Understanding the limitation, we keep our approach flexible so that any required improvement can be done to improve the user experience.

Regarding the data confidentiality, we have the approach in the app where we do not need to share the data to LLM rather we are sharing the schema only. Therefore, we do not have any risk of data breach through the LLM.

**5. Results**

The dashboard display summary results in tabs and pages. The first few pages include the data details and summary statistics. The summary statistics includes the patient demographic and clinical characteristics, genetic mutations, and the major endpoints that measure the severity of disease including ALSFRS score, and longitudinal distribution of the Score for both treated and control subjects in different trials.

The LLM chat option is included in the last page. User can ask any question related to this data will be responded in clear and concise manner. When asked any information about the data or the analytic results displayed in the dash, the response will appear in the response area of the page. Users analytic question will be responded in analytic result, the R code and if necessary, the interpretation. For example, “Show the dash results for female patients only” the LLM will return the query and analytics in R/Python. The code will be compiled by local R/Python interpreter and return the results and codes. The user can come up with a followup question in relation to the previous response.

**6. Discussion**

We found the app an effective tool to explore the data by the non-technical researchers like clinicians who are not familiar with database management or coding. Converse with LLM is very simple and intuitive therefore, they can extract information from data and come up with more meaningful research questions to the data scientists to answer from the data.

Analyzing new data is resource-intensive. Clinical researchers ask questions, and statisticians analyze the data to get this answered. Initially, neither fully understands the data contents nor its collection process. Once clarified, statisticians provide insights, prompting further questions from researchers with better understanding. This creates an ongoing cycle of analysis and requests, keeping statisticians occupied with coding and data interpretation.

Data and analysis results often left disorganized leads repetition of analysis for later needs. Codes and results are often left scattered on platforms like SharePoint which are often difficult to reproduce, as seen with the PRO-ACT [1]. The data analytic process repeats as we do for new data, especially with different researchers and statisticians at different times and needs. This repetition wastes time and leads to underutilized outcomes. A better system is needed to preserve and reuse data and analysis outputs effectively.

A streamlined system of analysis and representation can reduce statisticians’ coding burden and make results reusable over time. It would help researchers and statistician access and build on previous work easily, saving time, improving consistency, and enhancing collaboration. Integrating large language models (LLM) into the system would enable researchers to easily get information about the data and interpret the data.

The app model is in its developing stage; we have collected users experience mostly unstructured manner among researchers and found its utility and popularity.

We strongly believe that, this kind of smart data dash could benefit the monitoring and evaluation, following stocks prices and so on. And we have the tools to make data more useful the users. This will promote the data driven decision making.

**Paragraph 4: Limitations of the Study**

The app is reproducible for any database that includes data tables. However, limited used of unstructured data such as images, sound are included. Large amount of unstructured data inclusion and interpretation by LLM could be expensive. Despite, we strongly believe, this app would be a great support to the data users in different research industries and industries, business and NGOs that require data base monitoring system. A further development of the app is possible if we can incorporate the machine learning model which could learn form the chat and develop a recommendation system for the data deiven decision making.

**7. Conclusion**

* This is a novel approach of presenting and sharing data that is sustainable, re-usable and easy to understand by the non-technical person. It provides the opportunity to non-technical person to extract information from data with the help of LLM.

**8. References**

1 Pooled Resource Open-Access ALS Clinical Trials Consortium. (2024), <https://ncri1.partners.org/proact/document/displaylatest/6>, Access date: 08/23/2025