Using Large Language Models to Enhance Information Retention From Data Visualization *

Project Proposal

September 13, 2024

1 Basic Info

1.1 Title

Using Large Language Models to Enhance Information Retention From Data Visualization

1.2 Group Members

1. Jake Wagoner; u6048387; jakew@sci.utah.edu
2. Sudhanva Manjunath Athreya; u1529299; u1529299@utah.edu
3. Tin Vo; u1289616; u1289616@utah.edu

2 Background and Motivation

Our motivation for choosing this project is rooted in our diverse backgrounds and shared interest in exploring the potential of Large Language Models (LLMs) to enhance visual analysis of complex datasets.

While working on VDL's UpSet BLV (blind/low vision) Accessibility project, Jake worked on a sub-study about LLM generation of alternative text with various prompt conditions. This involved a preliminary study, ablation study, and best-effort study. This work is directly applicable to this project; however instead of generating alternative text, the project would aim to generate meaningful data analysis through complex prompt engineering (pre-prompting) and user interactions.

Additionally, Jake was inspired by a famous visualization in pop-culture, the ridgeline plot of pulsar data used as an album cover by Joy Division. This album is widely known, and millions of people have seen the data visualization; however, it is likely that most of the people never knew what the underlying data was, or what the visualization was trying to accomplish. Jake thinks it would be very interesting to see if users without a base knowledge of cosmology/pulsars would be able to interact with a similar visualization and come away with an improved understanding of not only the visualization and how it presents the data, but knowledge of the actual context and usage of the data.

Sudhanva, who has a background in computer vision and statistics, is deeply interested in improving data analysis techniques. He believes that visualizations are an important aspect of every

*DS-4630 / CS-5630 / CS-6630 Visualization for Data Science; Fall 2024 Instructor: Paul Rosen, University of Utah

stage of machine learning as they help communicate results and insights to stakeholders effectively. His experience working on large deep learning models made him realize the challenges of explaining and interpreting the models, and how most people just treat it as a black-box without having the ability to explain the underlying logic and math. This motivated him to focus on developing visualizations that improve the interpretability and explainability of these models.

Currently studying the field of Artificial Intelligence, this project relates back to many things Tin has practiced and learned. This includes taking an NLP course that focuses on the process and background of LLMs, then researching and generating results that involve answering certain questions using the technology behind LLMs. Combining AI and Data Visualization can also just allow for better interaction with the data which is always an interesting topic when considering human computer interaction.

By combining our skills in these distinct but related areas, we aim to explore the novel application of LLMs in visual data analysis. Our goal is to investigate how these models can supplement traditional analysis methods by providing contextual information, on-the-fly data analysis of the user selections, and unique insights to data trends.

3 Project Objectives

We want to see if LLMs are able to assist in analysis of data and visualizations from user input. Given the output, did it provide good context and is it accurate enough to be of use? Will users be able to interpret the data behind the visualizations and improve their understanding after given enough context and information? Is the process of interacting with the LLM intuitive for users who may have limited technical expertise? These are some of the questions we want to answer through this project, which would definitely help build onto future questions about human interaction with LLMs and how they would analyze data given to them.

4 Data

1. **Data Source**: `https://earthquake.usgs.gov/earthquakes/map/`

2. **Data Description**: The seismic data is from USGS (United States Geological Survey). The seismic data is part of the Earthquake Hazards Program. The USGS Earthquake Hazards Program is under the National Earthquake Hazards Reduction Program (NEHRP) led by the National Institute of Standards and Technology (NIST).

3. **Data Size and Format**: The website allows us to fetch the data in CSV, GeoJSON, QuakeML, KML formats. The data can be scraped from the API with relevant queries.

5 Data Processing

1. **Data Sampling and Aggregation**: Based on the year slider, the earthquake data over the years is sampled, while maintaining the data distribution. The sampled data is then aggregated where multiple records are then merged and displayed.

2. **Chunking and Lazy Loading**: Since the earthquake dataset consists of data points across the years, the data is chunked upon request and the remaining data is stored in a separate data store in chunks, and retrieved upon request.

3. **ETL pipelines**: Extract, Transform, and Load (ETL) pipelines allows us to automate and preprocess newly available data to streamline the integration of the new data into the dashboard. Having such a pipeline ensures efficient data ingestion and consistent updates.

6 Visualization Design

6.1 General Idea

The general concept is to visualize time-series data via a line/scatter plot. There will be a method to select a dataset, if there are multiple, and the user will be able to interact with the plot by clicking on data points to select individuals, clicking and dragging to select a series, and hovering to show a tool-tip or a visual change to indicate the hover state. Alongside the visualization will be an LLM output which will provide the user with contextual information about the plot shown as well as the user interactions.

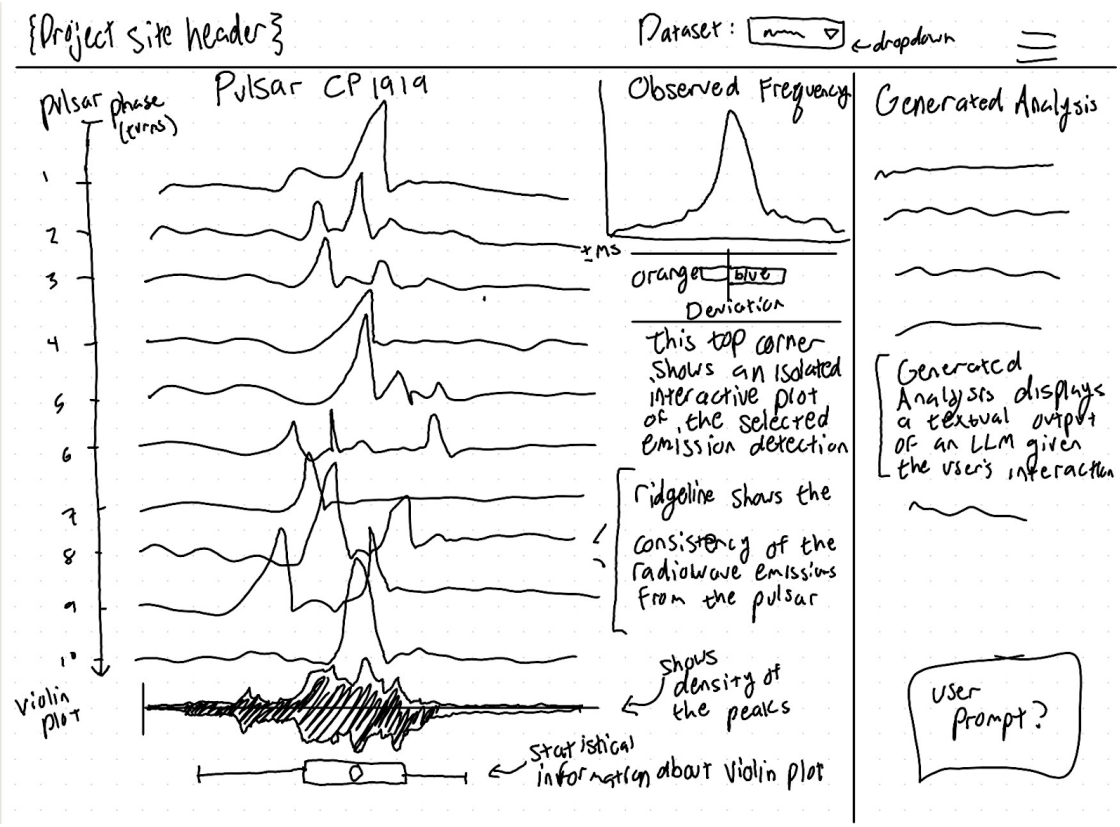

Figure 1: Pulsar dataset example, ridge-line plot + supplemental

For Figure 1, there will be a ridge-line plot which showcases a large number of observed values over time. The goal of this visualization is to provide a visual method to see the distribution of the emissions data, as well as the consistency. Below the ridge-line will be a violin plot, which shows the distribution of the data points. This will provide a method to quickly determine if the distribution is flat, concentrated, etc. Within this violin plot there will be encoded statistical information in the form of a box+whiskers plot. This will show mean, standard deviation, and quantile ranges.

Off to the side, there will be an isolated plot. As the user hovers over individual ridgeline entries,

the isolated plot will update to show the hovered entry. This will give the user the ability to look at certain plots in more depth. The axes will be the same as any individual datapoint in the ridgeline.

Below or included in this isolated visualization, there will be a "deviation" bar which indicates that specific data point's deviation from the mean.

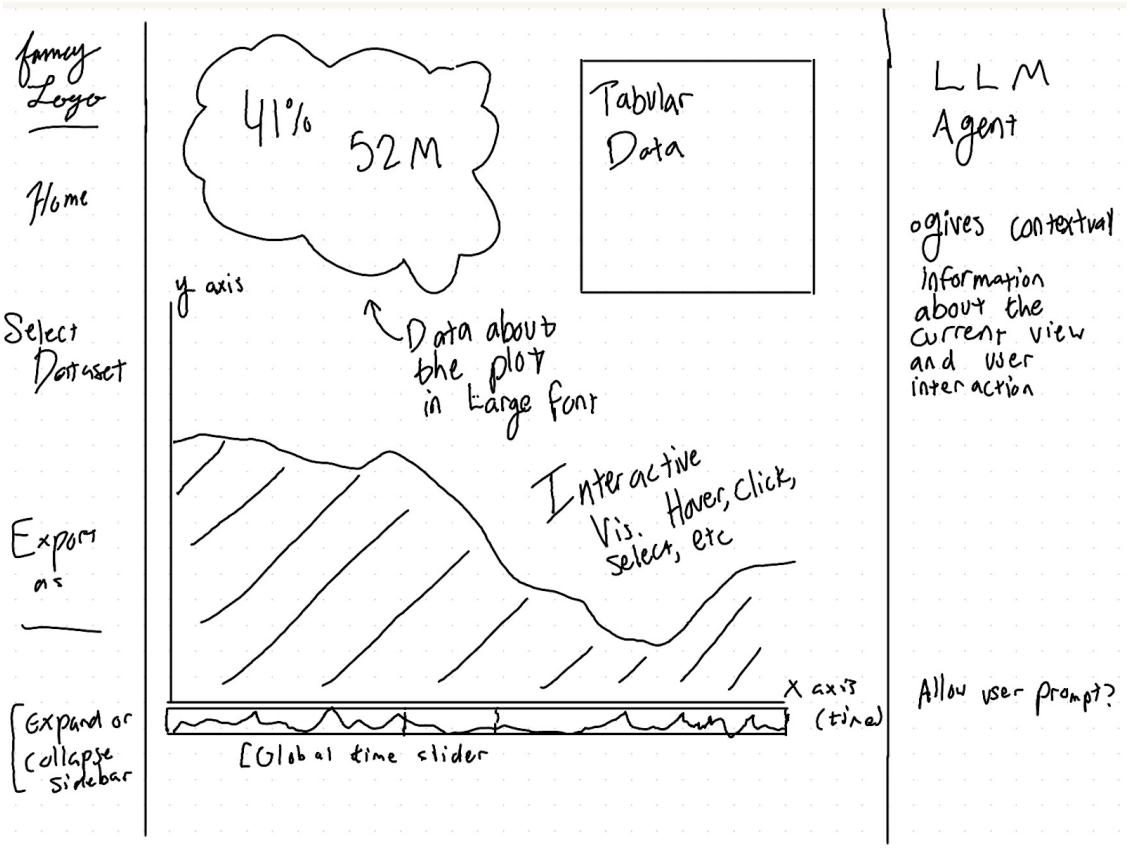

Figure 2: Any time-series 2D data, line plot with temporal slider

Figure 2 features a line plot which contains time-series data. The current visualization is a subset of the global dataset, where the time range can be selected using the temporal slider below the visualization.

The temporal slider shows the line plot of the entire dataset, condensed into the short but long view. In the slider is a range window which can be expanded or moved. This range window represents what data is being shown on the larger visualization.

At the same time, there is a tabular data panel open in the top right, which shows the tabular data for the current plot shown. If the user selects a region, the data will change to reflect this interaction. This can be switched to a density plot, scatter plot, or smaller line plot based on selection.

The LLM panel is a static sidebar to the right of the screen which outputs LLM text generation for the contextual information about the data and the user interactions.

The left menu is collapsible and includes options to toggle the dataset if possible, as well as export the current plot.

Figure 3 shows two visualizations overlaid, with encoded information within. The first plot is a scatterplot for the data points over the selected temporal region. This scatterplot also contains information in the form of trend lines. If multiple trends are detected within the data, both trend

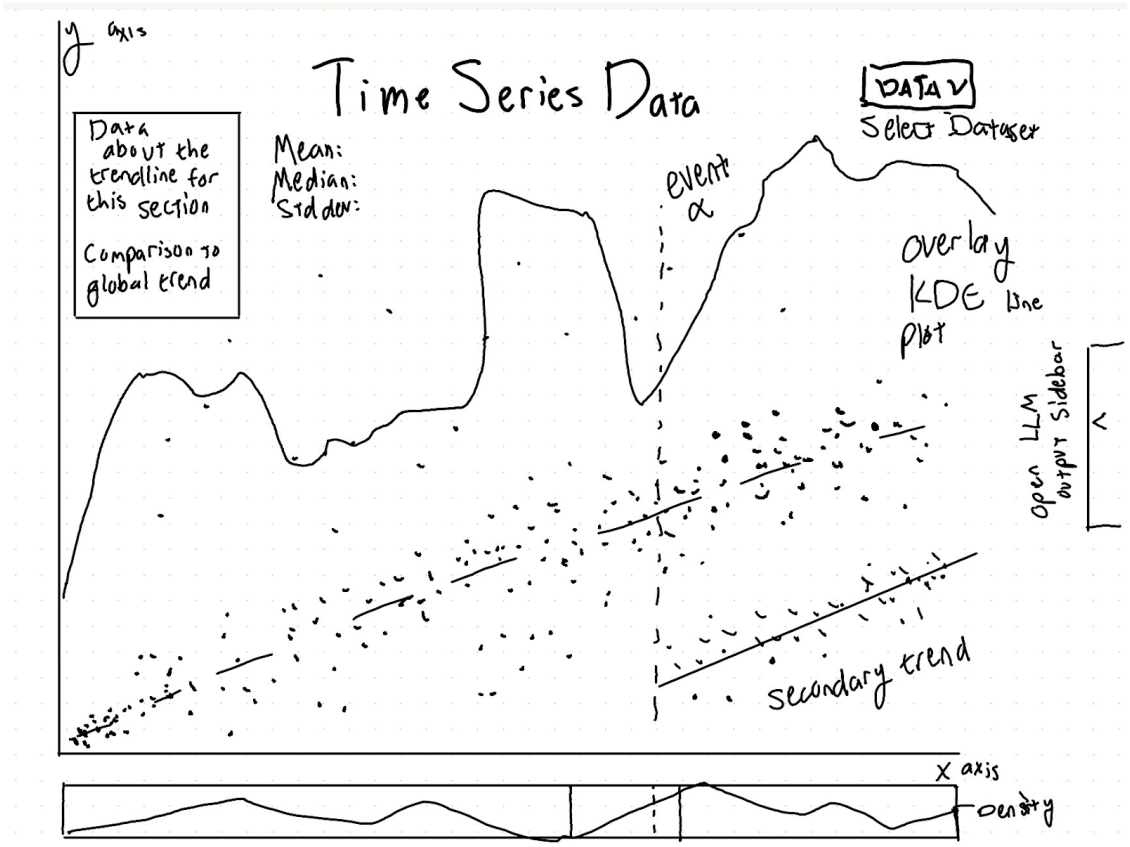

Figure 3: Any time series 2D data represented as a scatter-plot with encoded trend line and statistical data

lines will be displayed in different colors and the origin of the secondary trend will be marked with an "event line".

The second plot (overlay) is a KDE density plot. This shows the data density across the currently shown/loaded data. This is useful for seeing where data is most concentrated. This overlay is togglable.

To the top left, there is a trendline data section. This will show data for n-trendlines through a paginated format. The data will be slope, intercept, and r^2.

On the bottom is a temporal slider which behaves similarly to the one described in Figure 2. . This one, however, shows the data point density rather than every point. This would be illegible at such a small scale and would also be very computationally expensive for large datasets.

This visualization takes up the majority of the screen space on the website. There is no menu or nav-bar. The LLM output is a collapsible menu that is opened/closed with a tab marker on the right of the screen. This will allow the user to perform data analysis with and without the LLM at their choice.

The title of the website will be overlaid across the visualization at the top.

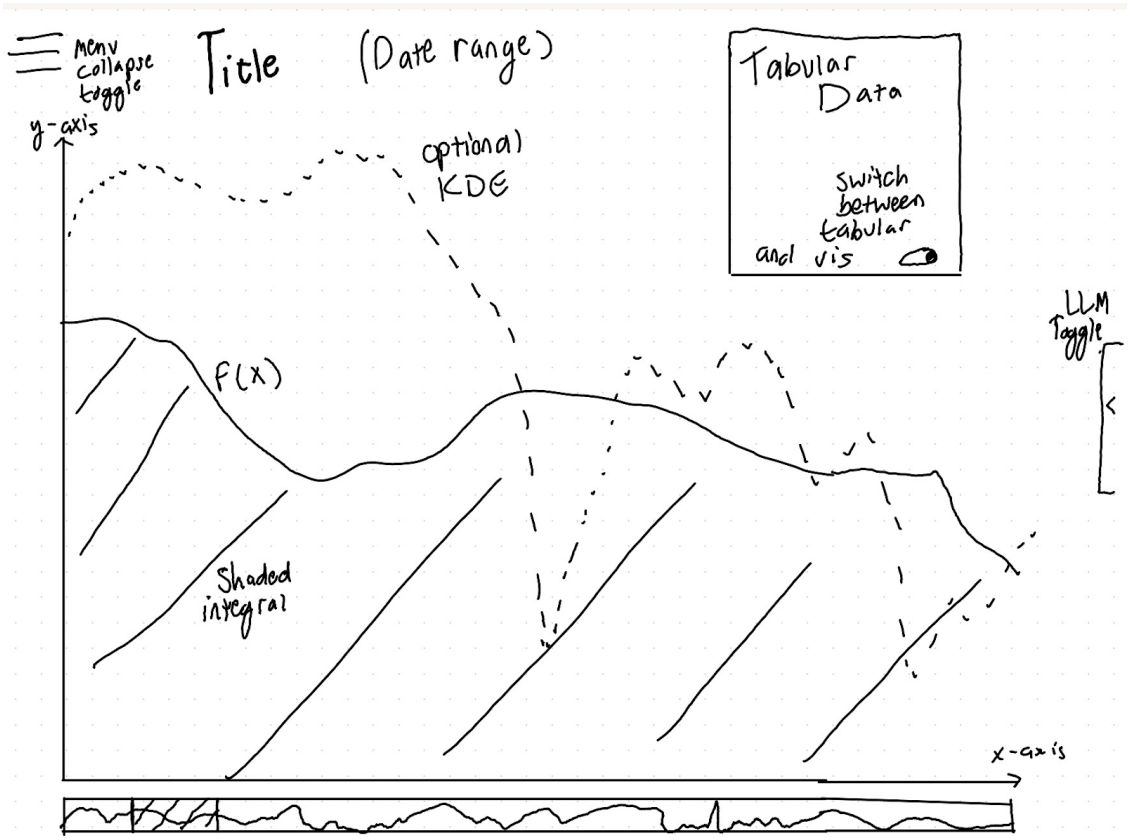

Figure 4: Line plot with temporal slider, optional KDE overlay, tabular data, and collapsable menus

The final design 4 is the cumulation of the previous 3 divergent designs. It broadly features a line plot with a shaded interior (integral). There is a temporal slider below. Tabular data is displayed for the entire plot, unless the user selects a range of data, in which case the table updates to display that data. The table can be switched to a visualization (KDE, scatterplot, or line plot).

The line plot shows the $f(x)$ output for a given point in time (X-axis). By using a solid line, this becomes a visual emphasis on the visualization. Lightly shading the interior of the line plot can give the user a sense of amplitude change over time.

The temporal slider contains a line plot over the entire range. This is condensed to fit within the X-axis length. The full width of the slider represents the entire data range of time. This allows users to get a global picture of trends across large regions of the dataset, without selecting the entire range. In the temporal slider is a range selector. This is lightly shaded to easily indicate the current selection. By clicking and dragging on either (left or right) edge of the internal slider, the user can expand the range. The user can also click anywhere on the range and drag to "slide" the current selection. The plot will update to show the current range selection. The x-axis will also update. The Y-axis will always show the global range. This allows for consistent comparison between temporal ranges.

A collapsable menu can be opened using a hamburger menu in the top left. This menu contains export options for the current plot as well as settings/toggles for the plot. One setting is to display a density plot overlay. This will show as a different color and thickness line from the main line plot. There will also be an option to show this as a dotted line. The density plot is calculated using a KDE. An optional additional feature is to display the density plot as a simple strip plot just above the X-axis.

Finally, there is a toggle tab on the right side of the screen. Clicking this tab will open the LLM side panel. This side panel displays the output of an LLM. The LLM output will be in paragraph form similar to standard output from popular LLMs, such as chatgpt. There will be a close button both on the top right of the sidebar (an X), and on the tab which was used to open the sidebar.

7 Must-Have Features

1. Display a line/scatter plot for the data, should have interaction by selecting individual data points or clicking and dragging over multiple points to highlight more information.
2. LLM analysis text generation for when the user selects an individual data point or multiple by clicking and dragging over them on the line/scatter plot.
3. Tabular or visual data on the data in which the user selected, which could be switchable to show a selected subset or entire plot if no particular type is selected.
4. LLM text generation from a pre-provided prompt.
5. Collapsable menus to show full visualization.

8 Optional Features

1. Ability to switch between datasets if we use more than one dataset.
2. Switch between multiple visualizations for each dataset.
3. Temporal slider or selector when viewing 2D visualizations.
4. Export options for visualizations and data.
5. 3D visualization of data.
6. Additional prompt suggestions for LLM
7. The option to switch between a KDE or strip plot at the bottom (raincloud plot style), can be turned on and off.

9 Project Schedule

- **Week 1-2**:
  - Forming a group.
- **Week 3-4**:
  - Project Proposal.
- **Week 3-4**:
  - Project Proposal.
- **Week 5-6**:

- Project Review with Staff (as a group).

- **Week 7-8**:
  - Midterms + Fall Break.
- **Week 9-10**:
  - No Class (More time to work on the project).
  - Project Milestone (Code + Process Book Check).
- **Week 11-12**:
  - Review Peer Feedback.
- **Week 13-14**:
  - Record Screencast.
- **Week 15-16**:
  - Thanksgiving Break (More time to work on the project).
  - Final Submission and Peer Assessment.