TadHealth Capstone: Event Impact Map and Recommendation System

Stephanie Saldaña, Lisa Fukutoku, Chloe Nelson

December 16, 2023

Abstract

TadHealth is a company that focuses on providing mental health services to people living in the United States. They wanted to add two new ideas to their platform: an event impact map and a recommendation system. Before completing these tasks, a GPT model was used to successfully extract required information from the news articles and mental health resources. After all information was gathered, the event impact map was built using a GUI called MapBox and a handcrafted GUI to allow more user interaction with the map. This map was successful as it allowed the user to select any news article, see the time the article was published, and view up to six locations that were affected by the article. Next, the recommendation system was designed to match clustered resources to clustered articles based on the distributions of their features. This was done using k-means, which clustered the articles and resources. When more resources are provided, a classification model can be used to match the clusters.

1 Introduction

Mental health is a growing concern across the world, especially in the United States. TadHealth, a company that supports mental health concerns in the United States, has reported that 95 percent of college counselors are reporting that mental health concerns are growing on their college campuses (TadHealth, 2022). It can be inferred that college counselors are witnessing this concern through an increase in the number of times a student comes to their office with a mental health concern or an increase in the number of students that are coming into the office with the same concern. In addition, 64 percent of reported students are no longer pursuing college due to mental health issues (TadHealth, 2022). Meaning that the concern of their mental health is so great, that these students feel the need to not continue their education. TadHealth has been working with professionals since 2020 to increase the supply of mental health resources and reduce these statistics.

Mental health is an important subject because good and consistent mental health has been known to save lives. Ben Greiner, the founder of TadHealth, has experienced this through his own personal story and has been looking for ways to make mental health resources more available ever since. Greiner's goal, through TadHealth, is to save lives, improve care outcomes, and enhance people's well-being. TadHealth is currently accomplishing these goals by serving counseling centers, schools, and other various organizations across the United States. Through their app, they provide several insights and resources to these communities. To further develop these goals, an event impact map and recommendation system was created in ten weeks.

1.1 Project Objectives

- Conduct data extraction on approximately 3,000 news articles related to mental-health-impacting events and retrieve information about who the event impacts (victim), who or what caused the event (perpetrator), along with other features.
- Create an event impact map. When an article is selected, a map should appear showing where the article directly impacts (epicenter) and where it indirectly impacts (correlated locations).
- Create a recommendation system that allocates resources to mental health professionals in the affected areas to then distribute to the broader community.
- Analyze a model to ensure that the most effective resources are being sent to mental health professionals in impacted areas.

1.2 Dataset Description and Exploratory Data Analysis

The data was shared by TadHealth via Google Drive. It contained two versions of the data, V1 and V2. V1 focused on manual categorization of the articles, while V2 focused on the newer articles and categorization. By the suggestion of TadHealth, V2 was the main source of data. V2 consists of over 3000 articles that are pre-categorized by an event category and event sub-category. There are eight categories and 39 sub-categories. Below are the eight categories and how they are defined.

- Crime: Actions or offenses punishable by law, such as mass shootings or fraud.
- Politics: Political events that affect the people; discrimination and racial inequality, school policies and administration issues at county, state, and federal and county level, etc.
- Finance and Business: Financial and business decisions that affect other people, like loans, income loss, or business loss.
- Education: Events that affect students, teachers, or those within an educational role.
- Media: Events that involve some type of initial online interaction, such as bullying on social media or hate speeches.
- Religion: An action or event carried out by or against religious groups.
- Trauma: An event that can be noted as traumatic, such as a natural disaster or school shooting.

- Relationship and Personal: An event that is more personable to an individual, including sporting events, family issues, and toxic relationships.

Some of these definitions overlap, which can cause some confusion about how to categorize an article. An example of this would be an article that discusses a company that has committed fraud. Since it is a company doing the action and those actions will affect their stakeholders, then that article should be categorized as Finance and Business. At the same time, since fraud is a known crime, that article should also be categorized as Crime. Because of this, certain articles are grouped into multiple categories. Figure 1 shows how many articles there are per category, including articles with multiple categories.

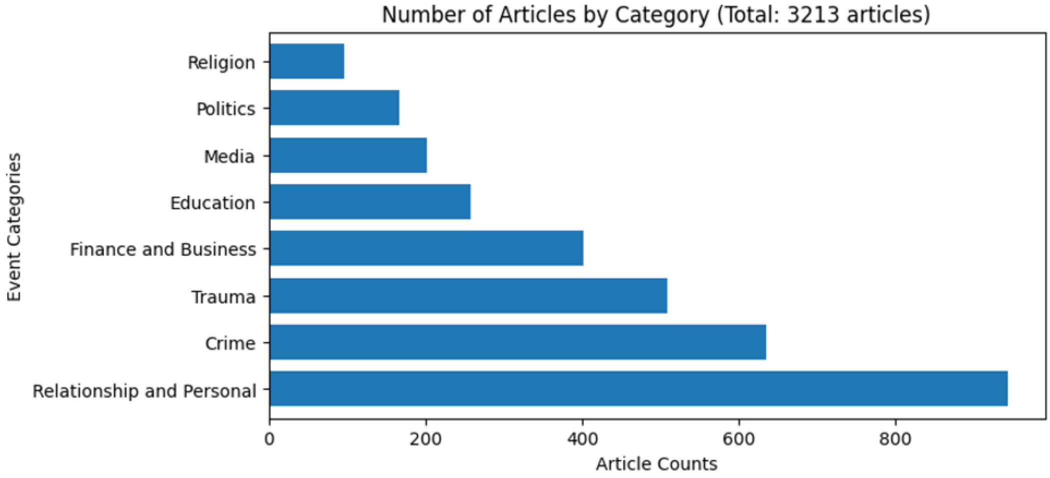


Figure 1: Article Count per Category for 3213 News Articles

As shown above, the number of articles per category is not equal. Most articles were categorized as Relationship and Personal, and articles were less likely to be categorized as Religion. This could be due to the amount of articles having multiple categories.

Along with an article's category, each record provided to the team had the article's title, a link to the article, the content of the article, and the text within the article. Since the project mainly focused on what is within the text, a word cloud was formed to see which words are most common within the articles (Figure 2).

Some notable words that can indicate the characteristics of most articles are "police", "health", "sexual", and "students". There are a few inferences that can be made from the keywords. Police and students must serve some role in a lot of the articles, which can include the role of victim or perpetrator. Therefore, it can be later expected that a majority of victims and perpetrators will be adults or students. The words "health" and "sexual" can indicate the potential scenario in many articles, such as a medical emergency that affects someone's health or sexual misconduct that has taken place. However, one of the most common words was "advertisement". Advertisement is not a common word when discussing mental health, so this word was an indication that the data was unclean when it was received. Due to this uncleanliness, a deeper dive was taken to see what else was affecting the consistency of the articles' text.

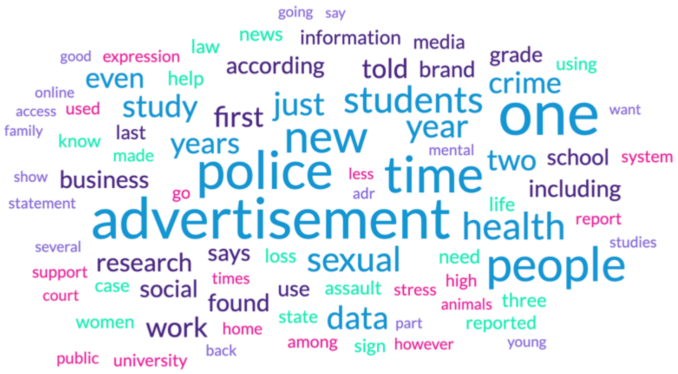


Figure 2: Word Cloud for All Text in Articles

1.3 Data Cleaning

Data cleaning was crucial for preparing the text for model analysis. It involved removing articles with irrelevant content, duplicates, or no text, ensuring the model processed only necessary, unique, and informative text.

Firstly, empty rows were eliminated, removing 36 articles with no text from the initial 3211, leaving 3175 articles in the data set.

Next, duplicates were addressed. As mentioned in Section 1.2, some articles appeared multiple times due to categorization. The first instance of each article was retained, while subsequent ones were stored in a separate file for later analysis. This process removed 649 rows (287 unique articles), resulting in 2813 articles remaining.

Finally, removing specific words or phrases was the last phase of data cleaning. As mentioned in Section 1.2, words like ”advertisement” were prevalent in the word cloud but had no relation to the article or ideas on mental health. Other words and phrases included ”ADVERTISEMENT”, ”LOADING ERROR LOADING”, and phrases about news article subscriptions. It is important to remove these words from texts because the model may be limited by the amount of words it can take. As shown in Figure 3, most articles have around 1000 words but some have more than the GPT 3.5-Turbo word limit of 4030. 3300 total words were removed from the 2813 articles. This step ensured the texts were concise and relevant for effective model analysis.

2 Data Extraction

Data extraction was the process of collecting the necessary data. In this particular case, a model was needed to collect information regarding the victim, perpetrator, the main location (epicenter), similarly impacted locations, and event setting from an article's text. This will later help with identifying which locations need recommendations. GPT 3.5 Turbo was the chosen model for this process.

GPT is known to have many potential variables. However, to keep the model's responses

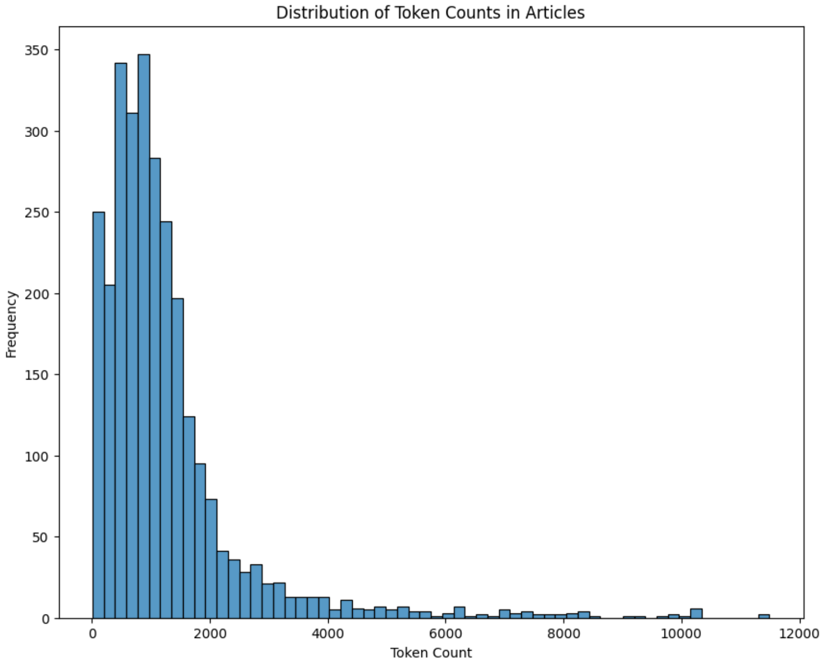


Figure 3: Distribution of Token Counts in Each Article (Max Token Count for GPT 3.5-Turbo: 4030)

clear, creative, and concise, the only variable that was unique was the max_tokens. This variable indicates how many words, better known as 'tokens', the model is allowed to display to the user, and it was set to 200 to encourage short answers except for the summary.

Another factor for an optimal response is how to phrase the prompt, otherwise known as prompt tuning. Prompt tuning has to be precise and intentional so that the GPT model can truly understand the request. Listed below are the final prompts and what they are meant to find in the article texts.

- Providing a Summary: "Give us a two-sentence summary of this article."
- Classifying the Victim: "Who is most impacted by this article? (select one: minor, adult, unknown) (answer format: one word chosen from the selection)"
- Describing the Victim: "In 8 words or less, identify who is most affected by this event concisely"
- Classifying the Perpetrator: "Who most likely caused this event? (select one: minor, adult, unknown) (answer format: one word chosen from the selection)"
- Describing the Perpetrator: "In 8 words or less, identify who has caused this event concisely"
- Finding the Epicenter: "List the location where this event occurred. Answer 'City, State, Country' (e.g., Orlando, FL, USA). If unsure, provide available details or write 'unknown'. No sentences"
- Finding Correlated Locations: "List 3-5 other cities in the US that could be impacted by this event. E.g., if there is an event that occurs in Palestine, majority

Palestinian communities in the US may also be impacted, such as Brooklyn, NY. No sentences, answer in a ‘City, State’ format and separate multiple cities using.”

- Identifying the Setting Type: “Choose the setting type that best matches where the event took place (select one: School, Residential Area, Recreational Center, Mall, Public Transportation, Other, Online Environment) (answer format: exact words chosen from the selection, no sentences)”

Since there are no comparative answers, accuracy scores, such as BLEU and ROUGE, become irrelevant. Therefore, the quality of the responses was up to the sound judgment and expectation of the group and reviewed with supporting teams, the professor, and TadHealth.

The extraction process was automated using an API call. This made it easier to collaborate and collect all responses in a single space. Once all prompts were final, the locations were ready for the creation of the event impact map. Listed in Figure 4 are the top 20 locations for the epicenter and correlated locations that were extracted from this process.

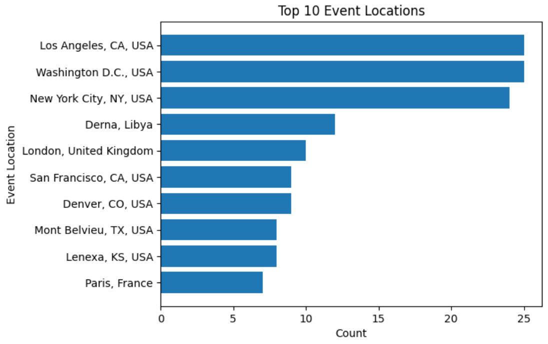


Figure 4: Top 10 Epicenters

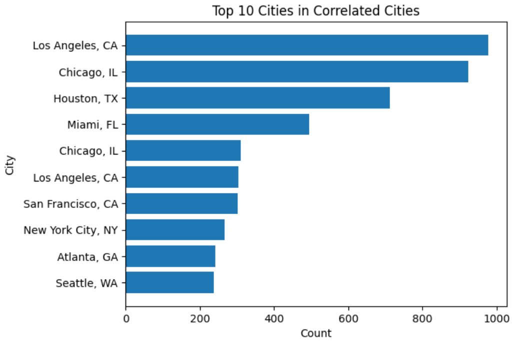


Figure 5: Top 10 Correlated Locations

3 Event Impact Map

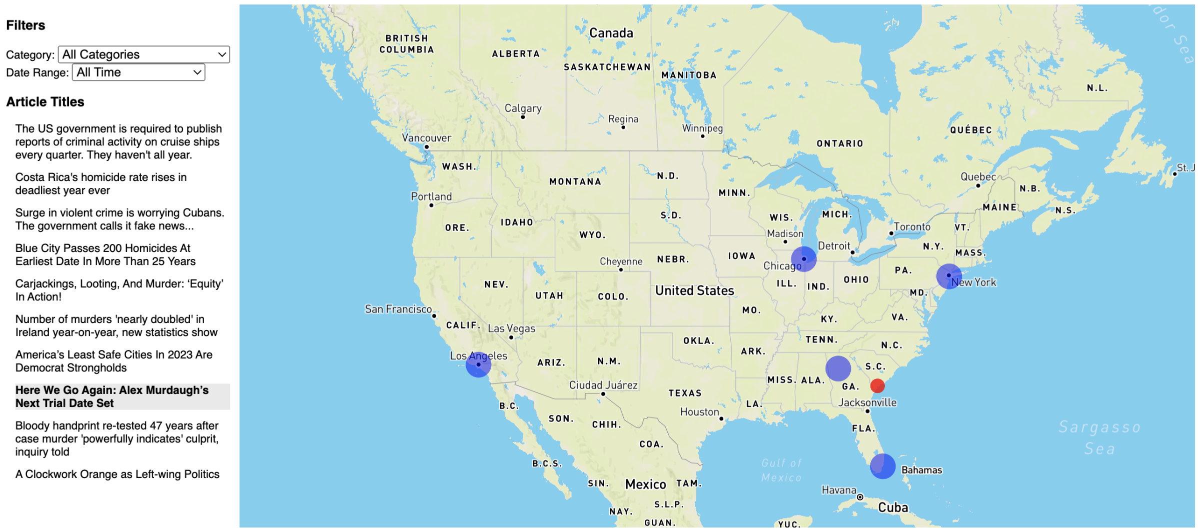
3.1 Methods

The event impact map was developed using MapBox, an online mapping platform. However, MapBox requires locations to be input as coordinates (longitude and latitude), not as city names. To address these challenges, Python was used to geocode the locations and convert the data from CSV to GeoJSON format. Moreover, MapBox lacked an inherent selection feature for creating maps based on article content. To overcome this, MapBox GL JS was utilized to develop a separate interface. This interface not only enabled the selection functionality but also enhanced the map’s aesthetics. It provided additional features such as displaying summaries, publication times, contexts, and recommended resources for each selected article.

After the map and the needed features of the map were completed, it was then connected to Streamlit. Streamlit is an open-source app that allows users to easily create a web application. It was used by uploading an HTML/JS file containing the code for Mapbox GL JS and the GeoJSON data file to GitHub. Streamlit was then able to connect to the GitHub repository and publish the map with all the features.

3.2 Results

After connecting the final product to Streamlit, users can successfully visualize the epicenter and five correlated locations for one article at a time. The user can also select the article they would like to map and see different article titles and dates when the article was published. They can base their selection on different categories, such as finance and business, crime, or religion. A view of the final product can be found in Figure 6, and to the published application can be found in References [2]. Overall, this covers the primary goal of creating an event impact map (section 1.1).



Filters

Category: All Categories
Date Range: All Time

Article Titles

The US government is required to publish reports of criminal activity on cruise ships every quarter. They haven't all year.

Costa Rica's homicide rate rises in deadliest year ever

Surge in violent crime is worrying Cubans. The government calls it fake news...

Blue City Passes 200 Homicides At Earliest Date In More Than 25 Years

Carjackings, Looting, And Murder: 'Equity' In Action!

Number of murders 'nearly doubled' in Ireland year-on-year, new statistics show

America's Least Safe Cities In 2023 Are Democrat Strongholds

Here We Go Again: Alex Murdaugh's Next Trial Date Set

Bloody handprint re-tested 47 years after case murder 'powerfully indicates' culprit, inquiry told

A Clockwork Orange as Left-wing Politics

Figure 6: Event Impact Map as Seen on Streamlit

4 Recommendation System

4.1 Methods

The recommendation system was designed to suggest mental health resources to professionals who would then distribute the resources to the community. Initially, a one-to-one matching between article and resource categories proved to be too simple and inaccurate. Other methods were investigated, such as random forest or k-nearest neighbors; however, they were both impractical due to the lack of pre-labeled resource data in the news article dataset. This meant that the resources either had to be included manually to train a classification model or a new method was needed. As a result, the team opted for the k-means clustering model, an unsupervised learning approach that doesn't require labeled data.

The k-means model groups data into clusters based on feature similarities. However, a crucial step in this process is determining the number of clusters (k). The elbow method was used to determine the number of clusters needed for articles and resources. This was a graphical method that plots the within-cluster sum of squares. The graph was expected to form an 'elbow', or distinguished curve, at the optimal value of k. It was determined that fifteen clusters would be appropriate for the article dataset and two clusters for the resource dataset. Once k is determined, the news articles can be clustered based on similar categorical features, such as victim or setting type. After clustering the resources the same way, the clustered articles can then be matched to clustered resources based on similar feature distributions. Since the datasets are vastly different in size, a feature distribution difference of 0.5 was accepted. Since there were four features to compare, three out of four feature distributions were expected to be within the 0.5 difference to be considered a potential resource.

4.2 Results

There were two successful matches from these methods. Based on the limited amount of available resources (27 resources available), up to two matches should be expected. Once more resources are available, more matches can be expected.

The accuracy of these results can be shown through a collection of distribution graphs that are organized by feature. Using the first match as an example, Figure 7 shows that both clusters have a heavy distribution around 1.0 for features Category and Victim. 1.0 for Category meant that the article or resource was categorized as a crime and for Victim it meant that the affected party was identified as an adult. It is also shown that both distributions lie heavily in Setting Types 4 and 5, which means the articles or resources take place in an online environment or school. Although it is not perfect, Figure 6 also shows that there is some overlap within the distribution of the SubCategory feature. This can indicate that even though SubCategory is a scattered feature for the article cluster, the resource cluster still fits within the boundaries of the articles. Hence, these mental health resources can be a great fit for some of the articles in the cluster based on these features.

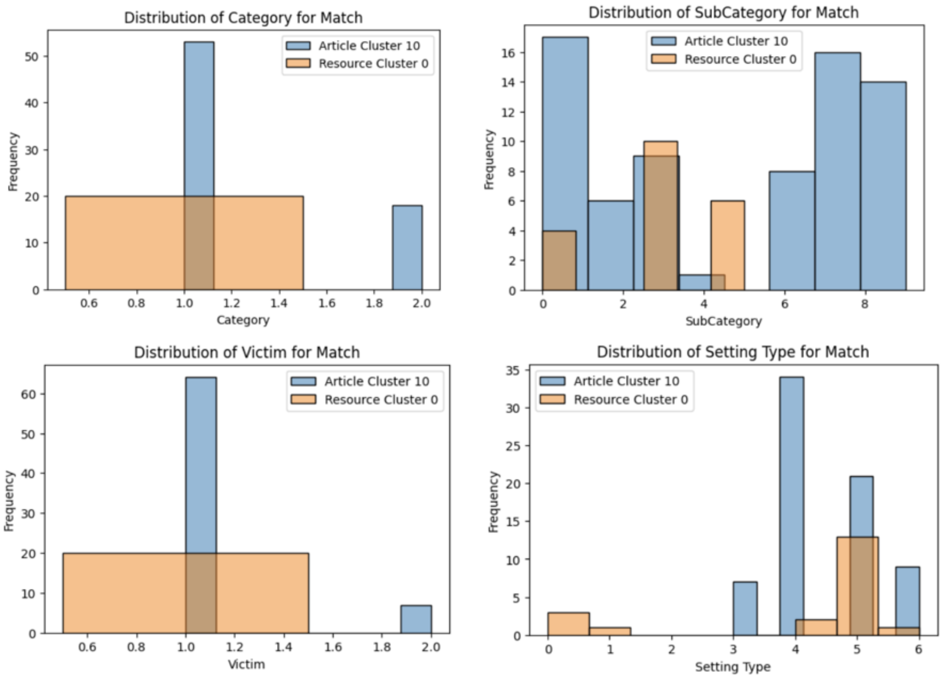


Figure 7: Comparative Feature Distributions for Matched Clusters: Article Cluster 10 and Resource Cluster 0

5 Discussion

5.1 Limitations

Limitations refer to how the user or system is restricted when using the chosen methods. For the event impact map, the user can only select one article at a time, and view up to six locations at a time, and all article information is pulled from a GPT model response. This restricts the user from making article and mapping comparisons. This also gives the map some bias, since GPT models can have a prompted bias. Although the team's sound judgement determined that the GPT model was producing acceptable responses, the user may notice some patterns

The recommendation system is limited by the dataset. At some point, the system needs a classification model to automate the matching process. However, it is not possible to train a classification model since the resource dataset is too small (27 rows available).

5.2 Future work

When this project is continued by TadHealth, these are some ideas to consider. For the event impact map, it is recommended to add multiple communities to the map whenever multiple communities are affected by an article. An example of this would be if there was an event that targeted African Americans and Hispanics. The map could be more personalized if the map detected what locations affected those individual groups.

For a stronger recommendation system, it is recommended to have a larger dataset for the resources to recommend. The dataset currently consists of 27 resources that all pertain to 'Crime' and most pertain to 'adult'. The goal of the system's design is to compare the feature distributions of clustered articles to the feature distributions of clustered resources. A larger resource dataset will allow more clusters to form within the dataset. Hence, there will be more clusters to compare. To have a classification model, there needs to be an equal amount of clusters between the two datasets. Another method to consider would be a neural network since it is a powerful tool to classify.

5.3 Contributions and Acknowledgments

The team would like to thank TadHealth for the opportunity to work on this project. They would also like to give a special thank you to founder, Ben, and TadHealth's Data Specialists, Anthony and Nikita, for their support and guidance throughout the project.

Listed below are the contributions from each team member.

- Stephanie Saldaña (Team Leader):
  - Head communicator between the team, sponsor, and professors
  - Head of the EDA and input data preprocessing
  - Head of event impact map creation
  - Support data extraction
- Lisa Fukutoku:
  - Head of data extraction
  - Head of output data cleaning and processing
  - Head of resource data cleaning and processing
  - Support event impact map creation
- Chloe Nelson:
  - Head of recommendation system
  - Head of report
  - Support data extraction
  - Support communication through notetaking

All team members, to the best of their ability, participated in all team meetings and activities. This included, but was not limited to, team meetings, meetings with professors or sponsors, and creating presentations. This team also supported each other in various stages, such as coding or prompt tuning.

6 References

[1] https://tadhealth.com/
[2] https://capstone-xc36a23uhsbhexnivsavwd.streamlit.app/