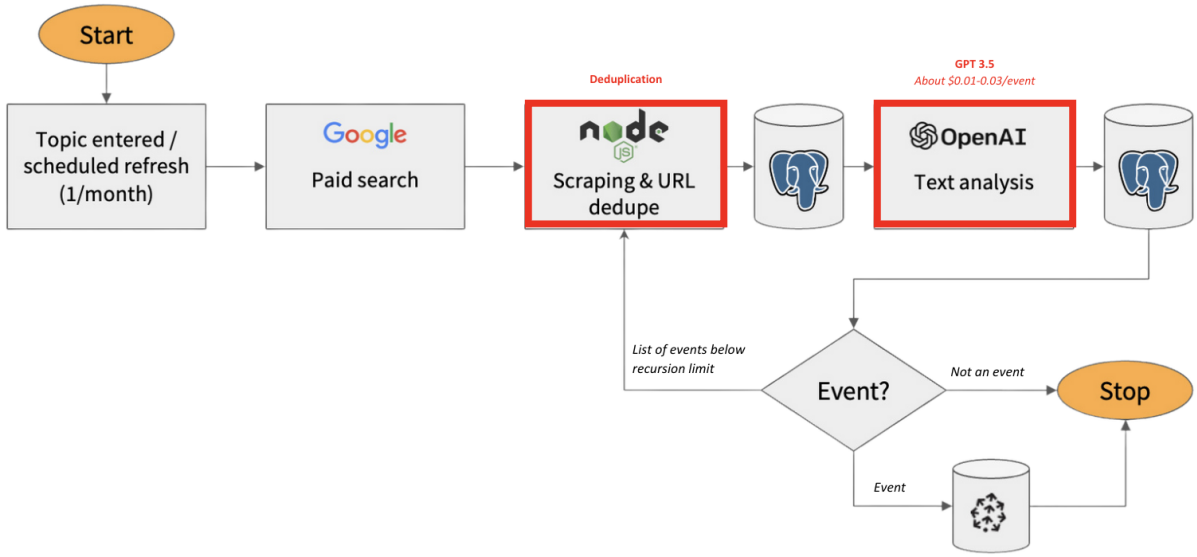
Sponsor Motion Summary Status Report
Scalable and Cost-Effective Deduplication: Leveraging Algorithms and LLMs
Group Members: Rohan Chaudhary, Sarmad Kahut, Valentina Torres

Business problem

SponsorMotion is a consulting and data company, founded by experienced professionals in the industry, that leverages artificial intelligence to create a comprehensive database of healthcare-related events in the United States. Their mission is to make all events easily discoverable for both sponsors and attendees, streamlining the process of connecting the right sponsors with the right events.

The goal of the project is to optimize the data ingestion process of SponsorMotion. Although the current pipeline is functional, it needs improvement to enhance the company's scalability and achieve the goal of making all US events searchable. To elaborate, the specific challenges to address are to establish a fully automated data quality control process that identifies problematic and duplicate records and to optimize the costs of post-scrape filtering and processing of event records. In the visualization below, the steps in red are the ones we will be working on.



Company Process Workflow Visualization with nodes of improvement highlighted

Dataset

The company leverages LLM by giving a prompt to analyze the scraped text from individual URLs and transforms it into a comprehensive dataset with 21 distinct columns. This dataset is a repository of an array of US healthcare event details, ranging from event names and locations to brief summaries generated from the scraped content, relevant event dates, organized in rows to facilitate efficient access and analysis. Events described in the dataset can be classified as medical conferences, seminars, trade shows and exhibitions, webinars, online events, marathons, walks related to a medical cause, etc. We were given a part of the dataset that contains 48k rows of events located in all 50 states in the US. After the initial discussion, the primary columns to focus on were determined to be the "start date" and "end date" of the events, "name", "state," and "summary" columns, as they were deemed most helpful for the analysis. Initially, we examined rows containing incomplete or missing event names and locations and observed that these missing records lacked other crucial data essential for event identification. Since such

records were redundant for analysis, we opted to exclude rows lacking event names, summaries, state information, and start dates from our analysis. The start date column required standardization to ensure consistency across all rows and hence the given timestamp was transformed to a standard format of YYYY/MM/DD. Finally, we decided to arrange the data by sorting based on location and start date to simplify analysis as each state and date can be considered as a bucket.

Research and Initial Approach

To gain deeper insights into the business process, we started by understanding the workflow involved in a user searching for specific healthcare event information and the process includes utilizing LLM (GPT-3.5) to analyze scraped text from individual URLs and store the information in a database in the form of 21 attributes.

However, with the increasing size of data and advancements in OpenAI models, the company is striving for further cost optimization. This actually served as the initial point of research, as we started by understanding the functioning of LLMs (Language Model Models) by focusing more on the analytical capabilities, prompt evaluation, and data generation parts but also keeping in mind the technical engineering aspect of LLM development. Online articles and websites helped us in getting acquainted with the architecture of LLMs and utilize LLM models for various applications in Natural Language Processing. Research papers and video tutorials helped us to dive further into evaluating the generative capabilities and efficiencies of LLM models and how tools like LangChain can be used to fine-tune and customize these models for user-specific purposes.

Since the scope of our project involves the cost structure, we reviewed the pricing information provided by OpenAI on their website. The problem of cost optimization led us to an informative research paper about FrugalGPT. The paper explores and evaluates three distinct strategies on how to lower the computational cost associated with utilizing LLMs. These strategies include prompt adaptation, LLM approximation, and LLM cascade. To illustrate the LLM cascade, FrugalGPT effectively determines the appropriate combinations of LLMs to utilize for different queries, aiming to enhance accuracy while minimizing expenses.

Our initial focus was to tackle the duplicate records in the data which result from multiple URLs referring to the same event and hence coming up as multiple records. Our research on several similarity detection measures and distance metrics to calculate the similarity between texts, led us to learn about the use of cosine similarity, Levenshtein distance, and TF-IDF (Term Frequency-Inverse Document Frequency). Before implementing these techniques, we realized that textual columns like “summary” and “name” of events are the most useful to identify duplication. On preprocessing the summary column, we identified patterns and common errors in a sizable number of rows such as values like “Unable to parse summary” and removed such records.

In our preliminary approach, we employed word2vec and TF-IDF vectorization techniques that involve preprocessing and tokenizing the text data in the summary column. These algorithms were selected based on their ability to capture the semantic meaning and contextual nuances of the text. Tokenization of the summary column involves initial cleanup, removal of stop words, converting all text to lowercase, and eliminating the most frequent words in the descriptions, which included words like conference, health, annual, and medical. This step aimed to decrease the likelihood of false duplicates due to commonly used words in the healthcare industry. The deployed techniques used cosine distance to find similarity and flag duplicates. The results were generated using different similarity distance thresholds and as a result word2vec yielded a notably low count of flagged duplicates but TF-IDF exhibited

better performance in identifying pairs of duplicates. However, upon manual verification, significant inconsistencies were found in the outcomes of TF-IDF, which raised concerns about the reliability of the results. This showed a constraint in heavily relying on the summary column as the sole indicator for detecting duplicate records.

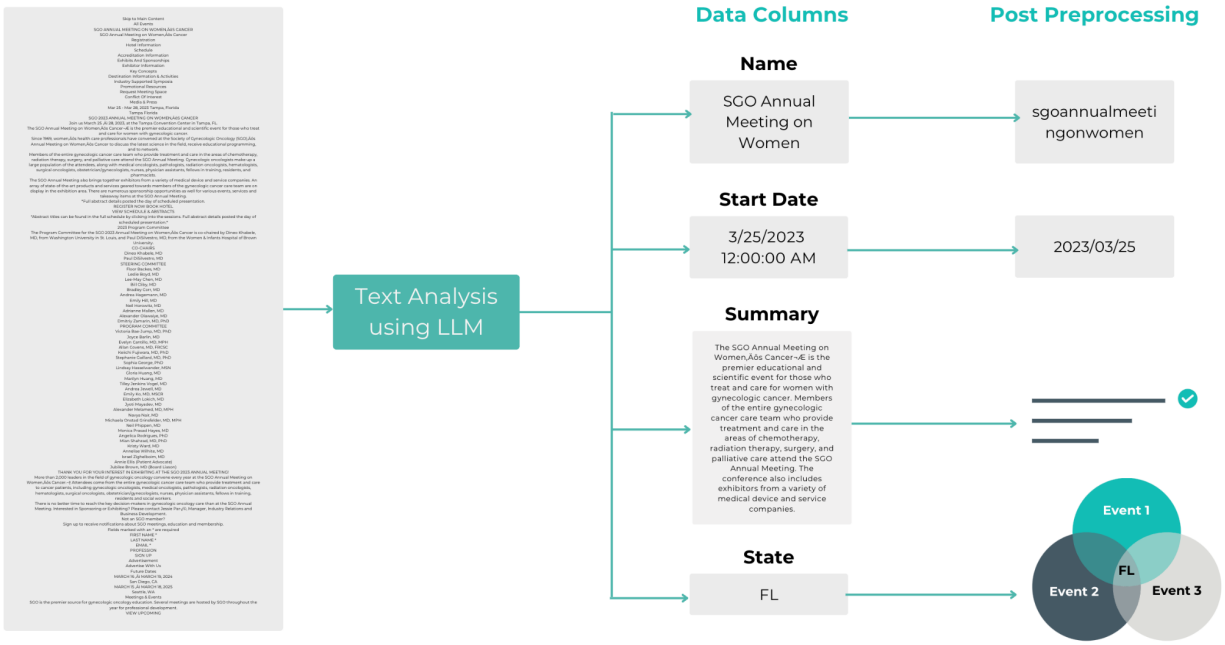
Word cloud of the most frequent and generic words in the text (summary) column

Shifting our focus to the "name" column, we found various errors and instances to understand the level of similarity between names of identical events and it was found that the "name" column exhibited fewer discrepancies in data, with minor variations such as typos, spelling differences, punctuation, and spacing.

To address the issue of similar names with small variations, we explored fuzzy matching, a powerful string-matching technique that allows for approximate comparisons between text strings. Fuzzy matching employs algorithms like Levenshtein distance to calculate the similarity score, which quantifies how closely two strings resemble each other despite typos, misspellings, or slight differences. This technique has the potential to effectively identify and group together duplicate records with small textual variations.

Final Approach and Method

To tackle the deduplication process, we followed several steps to identify and flag duplicate records effectively. We used Python as a tool to clean the dataset by dropping records that lacked essential information, specifically those with missing values in the name, start date, and state columns. This ensured that we focused only on meaningful records for the deduplication process. Additionally, in order to properly sort the data and facilitate comparisons, we standardized the format of the start date column. On the other hand, the name column contained valuable information but was prone to inconsistencies due to variations in capitalization, spaces, and special characters. We performed clean-up operations on the name column, including converting all names to lowercase, removing unnecessary spaces, and stripping special characters, resulting in a more standardized and consistent name format.

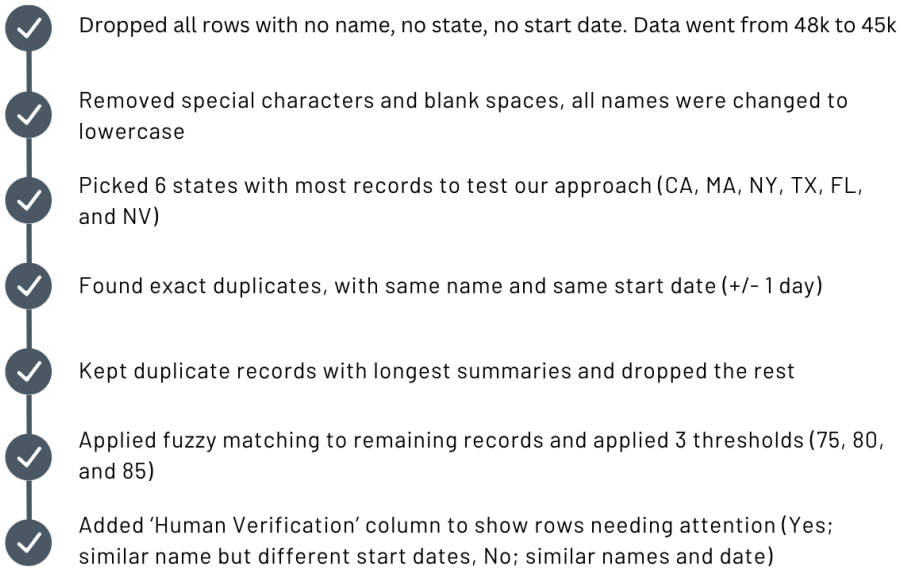


Overview of Data and Preprocessing

When we had a standardized and clearer dataset, we filtered it to include only the six states with the highest number of records. This step allowed us to narrow down the dataset and focus on the regions with the most significant portion of event data that would also be a good indicator for accuracy. To identify exact duplicates within each of the selected states, we compared records based on the combination of name and start date. If two records had the same name and start date, they were considered exact duplicates. Among the exact duplicates identified, we decided to retain the record with the longest summary. This decision was based on the assumption that longer summaries contained more comprehensive information about the event, making them more valuable for SponsorMotion's database.

Finally, after handling exact duplicates, some potential duplicates with similar names and different start dates remained. To address this, we implemented a combination of fuzzy matching on the name column and date comparison between event records. The data is filtered by each of the 6 states and then it is grouped and ordered by the start date of the event. This allowed us to organize the data and focus on potential duplicates within each state. The algorithm iterates through the data to identify pairs of potential duplicate events in each group. After sorting the events by the start date, the algorithm compares events within the group and checks if the events have similar names (using fuzzy matching) and occur on the same date or differ by just one day. The reason for including a difference of 1-day in the comparison of start dates is due to instances where the dataset comprises events that are essentially identical, yet their start dates vary by just one day. This one day difference could also be the result of recording dates with different timezone stamps.

If the calculated similarity using the cosine distance is above a certain predefined threshold, the events are flagged as potential duplicates and each pair or set of duplicates identified is given a unique number. However, the events that fall below the threshold are all pooled up as unique events with unique names and different start dates in each state. An extra column called "human verification" indicates whether the potential duplicates with similar names but different start dates require further human verification or not. By following this systematic approach and utilizing Python, we can successfully identify duplicates in the dataset.



Deduplication Algorithm: Python Notebook Workflow

Results and Key Findings

To assess the performance of our deduplication system, we tested different threshold levels to determine the percentage of actual duplicates identified for the six states with the highest number of records (California, Massachusetts, New York, Florida, Texas, and Nevada). Using a threshold of 75 in fuzzy matching means that when comparing two pieces of text to check if they are identical, we consider them a match if they share at least 75% similarity. In other words, if the data elements are 75% or more alike, we flag them as potential duplicates. Below are the results of the three tested thresholds:

- At a threshold of 75%, the deduplication system demonstrated remarkable performance, accurately identifying 81.15% of the actual duplicates for the six states with the highest number of records. This high accuracy rate ensured a substantial reduction in data redundancy and improved data quality for SponsorMotion's event database.
- Even at a more stringent threshold of 80%, our deduplication system remained effective, accurately identifying 67.69% of the duplicates. This reinforced the system's ability to maintain a significant level of accuracy while ensuring data integrity.
- When the threshold was set to 85%, the system continued to perform well, with an accuracy rate of 57.88%. Although the accuracy slightly decreased compared to the lower thresholds, it still provided valuable deduplication results, allowing us to maintain a reliable and accurate event database.

These findings validate the effectiveness of our deduplication system and its vital contribution to optimizing data quality and reducing data redundancy in SponsorMotion's data ingestion process.

Table 1: A summary of the results of the accuracy obtained from the three thresholds in fuzzy matching

Accuracy Results (NY, TX, MA, CA, FL, NV)					
Similarity Threshhold	**Total Records**	**Unique Events (from code)**	**Duplicates Identified (from code)**	**Actual Duplicates (Code + Manual)**	**% of Actual Duplicates Identified**
75	5646	4033	680	838	81.15%
80	5646	4183	530	783	67.69%
85	5646	4287	426	736	57.88%

The detailed accuracy report can be found at the following link: Final Accuracy Report

Accuracy Results

In our discussion with the business advisor, we concluded that higher thresholds lead to more conservative duplicate detection, hence higher false negatives, which is not desirable in the context of the given data due to the increased risk of missing actual duplicates and the associated costs and inefficiencies. Hence, we set up a cost function to estimate the cost associated with each threshold by giving higher weightage to false negatives and lower weightage to false positives.

We propose the simple cost function below (first approximation) to estimate the cost from the given count of false positives and false negatives:

Cost (fn) = 3 x FN + 1 x FP

- Cost (Threshold: 75) = 3 x 223 + 1 x 65= 669 + 65 = 734
- Cost (Threshold: 80) = 3 x 288 + 1 x 35 = 864 + 35 = 899
- Cost (Threshold: 85) = 3 x 336 + 1 x 26 = 1008 + 26 = 1034

Where FN = Number of False Negatives
FP = Number of False Positives

The three thresholds for the healthcare events dataset were evaluated based on a cost function that places more importance on false negatives. Threshold 75 resulted in the lowest cost of 734, indicating the most balanced approach in identifying duplicates, while thresholds 80 and 85 resulted in higher costs of 899 and 1034, respectively. These higher costs correspond to increased false negatives, reflecting poorer performance in detecting duplicates. Based on this analysis and the selected cost function, the threshold of 75 appears to be the most effective option for this healthcare data deduplication problem.

The results highlight a trade-off between false negatives and false positives as the threshold increases. The threshold of 75 appears to provide the best balance based on the chosen cost function, emphasizing the importance of identifying false negatives in the context of the business. In addition, the given dataset while tested for 6 states was found to have 15% actual duplicate records which were to be identified and the results for threshold 75 have been efficient in identifying 12% duplicate records thus the chosen threshold empowers the deduplication process to be methodical, precise, and effective.

Cost Optimization

Estimate of Deduplication Cost

To evaluate the potential cost implications of the deduplication of 48,000 records for 50 states in the US, we conducted an estimation. For an average of 960 records for each state, during manual deduplication, a team member invested an average of 3 hours per state to flag duplicates ensuring accuracy.

Scaling this approach to include all 50 states and all 48,000 records, the total time required for deduplication amounts to 150 hours. Considering a compensation rate of $25 USD per hour, the projected cost incurred for this manual deduplication effort would have amounted to $3,750 USD.

On the other hand, the manual identification of duplicates for the same amount of data would require a complex examination process. This process would involve scrutinizing each state and date entry and conducting a careful comparison with the corresponding name and summary fields across all rows for that date. Based on our estimations, this exhaustive manual deduplication procedure would translate to approximately 16 hours of labor for each individual state and a total of 800 hours for all 50 states. With the same hourly wage of $25 USD, the estimated financial expenditure in the manual deduplication would amount to a substantial $20,000 USD.

In conclusion, if a very unsophisticated person wanted to perform this task it would simply be undoable because it would be very inaccurate and tiring and even if it had to be done, it would be very expensive and not cost-effective.

LLM Cost Optimization

In the context of the initial cost objective – optimizing data retrieval from URLs related to healthcare events – the concept of cost-effective Large Language Models (LLMs) is another concern. While the current utilization of GPT-3.5 for extracting information from healthcare event data has proven effective, we are conscious of the need to balance budget and quality as the data size increases. After some research, we present "Frugal LLMs" as our recommended strategy specifically for data retrieval needs.

While the conventional practice revolves around employing a singular high-cost LLM for queries, this approach introduces a dynamic model selection process. This method would allow SponsorMotion to have access to LLMs including GPT-J, GPT-4, and ChatGPT, each aligned with different query requirements. By taking into account this approach, they could effectively tackle expenses while upholding, or even amplifying, the accuracy of the retrieved data.

The current approach that uses GPT 3.5 presents some challenges where data quality might vary, leading to incomplete or inconsistent information for some of the records. On the other hand, if we wanted to increase the data accuracy, GPT-4 would be a good alternative but it is not very cost friendly. With the use of FrugalGPT, we would be able to implement multiple models that would suffice the data retrieval needs, ensuring optimal accuracy without straining our budget.

Finally, we recommend initiating a pilot project employing a subset of 100 healthcare event-related URLs. This would entail subjecting the URLs to both GPT-3.5 and the frugal LLMs, including GPT-J and ChatGPT. The approach would involve deep analysis of the accuracy, completeness, and relevance of the information provided by each model. This would require looking into the costs for each model, giving us a clearer picture of their financial impact. Through this side-by-side comparison, the company could get valuable insights that display how well more budget-friendly LLMs perform.

Business Impact and Implications

Deduplication

By incorporating filtering, sorting, and fuzzy matching techniques, we have not only achieved faster and more efficient outputs but have also paved the way for increased automation and scalability in the future.

- **Automation:** The combination of filtering out irrelevant records, standardizing the data format, and testing for multiple states has substantially reduced the manual effort required to identify potential duplicates. These preprocessing steps allowed us to focus only on relevant and consistent data, streamlining the entire deduplication process. Additionally, by employing fuzzy matching algorithms, we automated the identification of potential duplicates with similar names, allowing us to evidence an opportunity to save valuable time and resources, to focus on more strategic tasks rather than manual data cleaning.
- **Scalability:** As Sponsor Motion's database continues to expand with an increasing number of healthcare-related events, our strategy can effortlessly handle larger datasets without compromising on the deduplication accuracy. The filtering and sorting techniques, coupled with automated fuzzy matching, are designed to handle considerable volumes of data efficiently. This scalability ensures that the deduplication process remains effective, even as the dataset grows in size, enabling the team to accommodate more events in their database without facing bottlenecks or performance issues.
- **Cost Savings:** The implementation of the deduplication algorithm has a direct and significant impact on the cost incurred and the operational efficiency of the company. The calculated estimate of the deduplication cost for the company shows the number of manual deduplication efforts and the cost saved if the process was not automated. This is a big win-win situation for a growing company as beyond financial benefits the algorithm is effective in enhancing data integrity and eventually improving the business outcomes.

LLM Optimization

Optimizing the use of Large Language Models (LLMs) through techniques like LLM cascading can have significant business impacts and implications for budding businesses like SponsorMotion that are utilizing LLMs, such as GPT 3.5 Turbo, to analyze text and are paying per 1000 tokens.

- **Cost Savings and Efficiency:** The most impactful and immediate benefit of optimizing LLM usage is cost savings. Switching from traditional LLMs to more cost-efficient options allows small businesses to achieve better cost-effectiveness in their text analysis operations. The model of pay-per-1000-tokens becomes more economical, allowing more analysis within the same budget.
- **Enhanced Scalability:** The LLM cascade enables small businesses to scale their text analysis operations without significantly increasing costs. This scalability allows businesses to process large volumes of data and expand their analytics capabilities.
- **Faster Processing:** Frugal LLMs would possibly provide faster processing time as compared to bigger and more intensive models. This can result in a quicker turnaround for textual content analysis, improving overall operational efficiency.

Limitations and Avenues for Improvement

Upon rigorous examination and manual analysis of various data sets to find the accuracy of our results, we have concluded that there is substantial room for enhancing the outcome through improvements in data quality. The filters applied in our current methodology only yield a data set of acceptable quality for us to perform the necessary analysis and fuzzy matching. This highlights the need to reevaluate the data quality being delivered by the data-scraping tool utilized by SponsorMotion. It is notable that in the world of data scraping, a well-planned and robust filtering mechanism could considerably streamline the process and improve results.

One aspect of this would involve a focus on the main event URLs which in many cases in the healthcare industry end with ".org". Based on our research and understanding of such events, implementing a filter to specifically highlight these URLs as legitimate could drastically improve the precision of our data. Furthermore, optimizing the efficiency of the web scraping process also warrants attention. This can be achieved by employing sophisticated filters to eliminate null values in key columns such as event name, start date, state, and summary. Null values in these critical columns can significantly impair our analysis and lead to incomplete or misleading results. Implementing these changes in our data scraping and analysis procedures could lead to more accurate, efficient, and valuable insights. It would equip us to better understand the extent of our research field and make decisions that are more informed. However, there are webpages where dates are either missing or the information provided is incomplete, such as those containing "Stay tuned!" notices about events. In these situations, we record "null" values to accurately represent the information. These values should not be seen as a result of the scraping process, but rather as a reflection of the information that is publicly available at that specific time. While some instances might indeed be connected to the scraping process, others merely mirror the inherent nature of the raw materials with which we are working.

Moreover, achieving this level of automation would set the stage for tackling our next challenge - devising an intuitive and effective way to label taxonomy. By doing so, users could effortlessly search and filter events based on their respective categories, such as oncology, cardiology, and hematology, among others. Establishing a taxonomy labeling system would enhance the user experience by allowing for efficient navigation and quicker access to relevant information. This would further streamline the process of event look-up, creating a user-friendly interface that is easy to navigate and intuitive to use.

Concerning the proposed LLM Optimization techniques that offer several business benefits, while smaller models use limited resources to maintain performance, criteria like trade-offs in accuracy, response time, and comprehensiveness should always be considered. It is important to evaluate which model to choose based on the individual needs, which implies a thorough evaluation and comparison of results. Integrating the LLM cascading techniques would require adjustments to existing workflows and systems, which should not disrupt the current functionality. Hence, we need to assess task suitability by acknowledging that not all text analysis may be optimally served by frugal LLMs.

Conclusion

In conclusion, our capstone project has achieved significant milestones that have positively affected SponsorMotion's data ingestion process. We successfully addressed the initial challenges by enhancing the data pipeline and implementing an efficient deduplication system. Our efforts led to substantial improvements in data quality and consistency. Through filtering, sorting, and the application of fuzzy matching techniques, we were able to effectively identify and manage duplicate records, resulting in a more accurate and reliable dataset. Moreover, the streamlined processes we introduced have paved the way for automation and scalability, positioning SponsorMotion for future growth.

Our impact on SponsorMotion's business is evident through the reduction in data redundancy achieved by identifying 81% of actual duplicates, and the enhanced data quality resulting from our deduplication system. Additionally, by saving the substantial manual deduplication cost calculated above the company's operational efficiency and financial performance can be enhanced. Using automated deduplication techniques, the company can minimize resource-intensive manual labor costs and ensure accurate and reliable data quality. These outcomes align with SponsorMotion's mission of improving the accessibility and accuracy of healthcare-related event data. In summary, our capstone project has delivered tangible benefits to SponsorMotion, and we are optimistic about the potential for ongoing enhancements. We are proud of the progress we have made and are excited to see the lasting positive impact on SponsorMotion's operations.

Acknowledgements

We are deeply grateful for the guidance and support provided by our business advisor, Paolo De Marino, whose weekly meetings enriched our project. Additionally, we appreciate the continuous feedback and recommendations from our faculty advisor, Elgar Pichler, which greatly contributed to our project's success. We would also like to extend our gratitude to Professor Dokyun (DK) Lee for guiding us in the right direction and giving us his valuable time during the initial days.

References

- Fuzzy Matching:
  https://www.youtube.com/watch?v=1jNNde4k9Ng
  https://www.youtube.com/watch?v=y-EjAuWdZdI&t=729s
- Word Embedding:
  https://towardsdatascience.com/text-classification-with-nlp-tf-idf-vs-word2vec-vs-bert-41ff868d1794
  https://www.geeksforgeeks.org/python-word-embedding-using-word2vec/
- Web Scraping: https://blog.apify.com/chatgpt-web-scraping/
- Language Models:
  https://towardsdatascience.com/the-easiest-way-to-interact-with-language-models-4da158cfb5c5
- Generic LLM intro: https://en.wikipedia.org/wiki/Large_language_model
  OpenAI
  a. Cost structure: https://openai.com/pricing
  b. GPT-4 announcement: https://arxiv.org/abs/2303.08774
  c. Papers on GPT 3.5 vs. 4: https://arxiv.org/abs/2304.13714 (look beyond the "clinical" lens, at what the models did)
  d. API: https://platform.openai.com/docs/introduction (note difference between GPT-3.5 "chatty" format and GPT-4)
- BERT for dummies — Step by Step Tutorial: https://towardsdatascience.com/bert-for-dummies-step-by-step-tutorial-fb90890ffe03

- How to Train a BERT Model From Scratch: https://towardsdatascience.com/how-to-train-a-bert-model-from-scratch-72cfce554fc6
- Cost Optimization approach: https://arxiv.org/abs/2305.05176
- Frugal GPT explanation: New AI cascade of LLMs - FrugalGPT (Stanford)
- Sentence Transformation Approach: https://huggingface.co/sentence-transformers