Research Project II

CSCI 9302 Final Report

EnReLLM: Enhanced Recommendation Systems through Integration of Large Language Models

Name	Krishna Modi
B00 Number	B00931943
Term	Winter 2024
Supervisor	Dr. Ga Wu

Table of Contents

Table of Figures

Executive Summary

This report focuses on comparing the efficacy of Large Language Models augmented with contextual awareness against conventional Recommender System libraries. By using data from the Movie Lens database and targeting diverse user preferences across multiple genres, the research aimed to enhance recommendation accuracy and user satisfaction. The core questions focused on assessing personalized recommendation generation, rating prediction disparities, and top-ranked item recommendations between LLMs and RS libraries.

The results show substantial disparities between LLMs and RS libraries, especially in rating predictions and top-ranked item recommendations. LLMs with contextual awareness, demonstrated a remarkable ability to align recommendations closely with user preferences, leveraging common sense reasoning and global knowledge. This contrasted with RS libraries, which primarily relied on data points without contextual prompts. Overall, integrating LLMs with contextual awareness significantly transformed recommendations with the potential for more personalized, user-centric recommendations.

For the full working code visit the repository on Gitlab or request maintainer access on Email.

Introduction:

The research project undertaken in "CSCI 9301 Final Report [1]" laid the groundwork for exploring and enhancing the effectiveness of recommender systems with large language models (LLMs). Building upon that foundation, the primary objective of this study is to compare the performance of a traditional Recommender System (RS) library against an LLM augmented with contextual awareness. The aim is to help the LLM use its common sense and global knowledge to generate more personalized and contextually relevant recommendations.

These research questions focus on evaluating and comparing the performance of LLMs with contextual awareness against traditional RS libraries across key aspects such as rating prediction accuracy, rating discrepancies, and top-ranked item recommendations.

1. How does the performance of Large Language Models (LLMs) augmented with contextual awareness compare to traditional Recommender System (RS) libraries in generating personalized recommendations?
2. What are the differences in rating predictions between LLMs with added contextual awareness and traditional RS libraries?
3. How do the top-ranked items recommended by LLMs with contextual awareness compare to those suggested by traditional RS libraries?

The significance of this study lies in its potential to enhance recommendation accuracy, user satisfaction, and overall system effectiveness. In today's digital landscape, where information overload is prevalent, recommendation systems play a vital role in alleviating choice paralysis and improving user experiences across various domains such as e-commerce, content streaming, and more. The envisioned outcome is a recommendation system that leverages LLMs' advanced capabilities, including common sense reasoning and contextual awareness, to deliver highly tailored and contextually relevant suggestions. Ultimately, this can lead to enhanced user experiences, increased engagement, and improved satisfaction levels for users across diverse platforms and domains.

Design and Methodology:

The design and methodology are mainly structured to evaluate how the recommendations made can change when we combine the large language model with the recommender library to make a recommendation system. The main idea is that the recommender library just uses the data points to make the recommendation, but the LLM can use its common sense and knowledge to improve the recommendation made by the recommender library and give a better user experience. This makes our primary objective to assess how the ratings and rankings generated by the RS library differ from the improved ratings and rankings done by the LLM.

We are utilizing data from the Movie Lens database. Our target audience comprises the first 183 users from the database, selected to represent a diverse range of preferences across specific movie genres such as Action, Adventure, Thriller, Romance, Sci-Fi, Children, Drama and Comedy.

In the experimental design, we set up the RS library that operates only on user-movie interaction data without contextual prompts. Next, we take the output generated by the RS library and use an LLM with contextual prompts to direct it to use its knowledge to improve the recommendation rating and rankings. Then we compare the output between the RS library and the LLM. This direct comparison between the two approaches focuses on the difference between the rating and ranking prediction.

Data collection is conducted from the Movie Lens database, with no preprocessing to maintain the raw nature of the data. This approach reflects real-world user behavior and interactions, contributing to the authenticity of our evaluation. The evaluation metrics employed include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), for rating comparisons and overlap percentage for ranking comparisons, providing quantitative measures to assess the LLM's output against traditional RS libraries.

Implementation:

In this phase, we implement the system architecture that was outlined in [1] and follow the design and methodology as discussed above. This phase includes setting up the Movie Lens database, software development tasks related to coding the recommendation generation process using the Surprise Library, and prompt engineering to optimize prompts by writing and formatting for optimal output. We then compare the ratings and rankings given by the RS and LLM.

Setup of the Movie Lens Database: To begin our experiment, we use the Movie Lens database, focusing on its Movies and Ratings tables. This choice ensures that we use reliable and standardized data. We use SQLite as our database platform to set up the database and have easy access and management of the data.

Development for Recommendation Generation: We are coding the recommendation generation process with the help of the Surprise Library. This involves writing the script to use the algorithms of the Surprise Library to generate recommendations based on user-movie interactions. The development process ensures that our recommendation system operates efficiently and effectively.

Prompt Engineering for LLM-Based Recommendations: Next we use the output generated by the library and use them as an input to the LLM. We focus on crafting prompts in a manner that generates the best possible output from the LLM. The prompts are carefully written and formatted to provide contextual cues and enhance the system's ability to tap into its global knowledge and common sense to be able to understand and generate accurate recommendations.

Comparison of Ratings and Rankings: Once we have recommendations from both the Surprise Library and the LLM, we compare the ratings and rankings they provide. This comparative analysis is important in evaluating the performance of each system, assessing factors such as the difference in the rating prediction and the alignment of top-ranked movie recommendations.

Throughout the experiment, we encountered several challenges that required attention and resolution. These challenges ranged from integration issues to prompt refinement complexities. By identifying and addressing these challenges, we tried to make sure that the results were reliable and fair.

By carefully following these phases and addressing challenges as they arise, we conducted a comprehensive and strict evaluation of the recommendation system's effectiveness, particularly in comparing how a contextually aware LLM makes a difference to the output generated by the RS Library.

Results and Analysis:

After implementing the experiment, we save all the results in a csv file called 'evaluation.' Below are a few examples to elaborate the two possible scenarios of the results.

1. Low Difference in Ratings

As we can see below, the difference in ratings produced by LLM and surprise library is very low for the top movies but there is a difference in ranking by the LLM as compared to the RS library.

```
{'Movie': 'Seven Samurai (Shichinin no samurai) (1954)', 'Rank': 1, 'Rating': '4.34'}
{'Movie': 'Lord of the Rings: The Fellowship of the Ring, The (2001)', 'Rank': 2, 'Rating': '4.30'}
{'Movie': 'Lawrence of Arabia (1962)', 'Rank': 3, 'Rating': '4.28'}
{'Movie': 'North by Northwest (1959)', 'Rank': 4, 'Rating': '4.27'}
{'Movie': 'Bridge on the River Kwai, The (1957)', 'Rank': 5, 'Rating': '4.27'}
{'Movie': 'Good, the Bad and the Ugly, The (Buono, il brutto, il cattivo, Il) (1966)', 'Rank': 6, 'Rating': '4.26'}
{'Movie': 'Great Escape, The (1963)', 'Rank': 7, 'Rating': '4.25'}
{'Movie': 'Spirited Away (Sen to Chihiro no kamikakushi) (2001)', 'Rank': 8, 'Rating': '4.21'}
{'Movie': 'Army of Darkness (1993)', 'Rank': 9, 'Rating': '4.20'}
{'Movie': 'Day of the Doctor, The (2013)', 'Rank': 10, 'Rating': '4.20'}
Final List Generated by LLM
{'Movie': 'Spirited Away (Sen to Chihiro no kamikakushi)', 'Rank': '1', 'Rating': '4.8'}
{'Movie': 'Lord of the Rings: The Fellowship of the Ring', 'Rank': '2', 'Rating': '4.7'}
{'Movie': 'The Good, the Bad and the Ugly (Buono, il brutto, il cattivo, Il)', 'Rank': '3', 'Rating': '4.6'}
{'Movie': 'Seven Samurai (Shichinin no samurai)', 'Rank': '4', 'Rating': '4.5'}
{'Movie': 'North by Northwest', 'Rank': '5', 'Rating': '4.4'}
{'Movie': 'Great Escape, The', 'Rank': '6', 'Rating': '4.4'}
{'Movie': 'Lawrence of Arabia', 'Rank': '7', 'Rating': '4.3'}
{'Movie': 'Day of the Doctor, The', 'Rank': '8', 'Rating': '4.3'}
{'Movie': 'Bridge on the River Kwai, The', 'Rank': '9', 'Rating': '4.2'}
{'Movie': 'Army of Darkness', 'Rank': '10', 'Rating': '4.0'}
```

Figure 1. A list of top 10 movies generated by the RS library for a user in the adventure genre followed by the improved list generated by the LLM for the same user. This is an example of low RMSE and low MAE.

This difference is due to LLM applying its context to the user's preference. What we are doing behind the scenes is that we grab the top movies watched by the user in the same genre and/or in a different genre if that genre is unexplored by the user. Based on that we see how the users interact with the movie. First, we invoke the ChatGPT API to understand the plot and characters of each movie followed by possible reasons the general audience would like those movies and the reasons our user prefers the movie. Based on that we invoke the API again to understand the preference of the user and if their preferences are aligned with the general public or is it unique to the user. This is because a recommender system may just suggest movies because they are popular or in continuation of some series but may not be very popular with the general public.

```
Based on the movie preferences listed, it seems that the user enjoys a mix of adventure, suspense, comedy, and animated films. They are likely to resonate with plots that involve thrilling adventures, suspenseful
 moments, comedic elements, and imaginative storylines. The user may also appreciate strong character development and engaging performances.

It is possible that the user's preferences are influenced by the general audience to some extent, as the movies listed are popular and well-received by a wide range of viewers. However, it is also likely that the
 user has their own unique tastes and preferences that guide their movie choices. Ultimately, the user's enjoyment of these films is likely a combination of personal preferences and the overall appeal of the movi
es to a general audience.
```

Figure 2. The user preferences as described by the LLM and how it aligns with the general audience

Next, we see how LLM thinks about the movies recommended by the system. It has to answer in a simple yes and no with a brief description of why it thinks the same. This helps the LLM connect dots where it studies the user's preference and based on it explains if the user will enjoy the movie based on their preferences.

```
- Seven Samurai (Shichinin no samurai) (1954): Yes, the user is likely to enjoy this movie as it combines action, adventure, and drama elements which they enjoy.
- Lord of the Rings: The Fellowship of the Ring (2001): Yes, the user is likely to enjoy this movie as it falls within the adventure and fantasy genres which they enjoy.
- Lawrence of Arabia (1962): Yes, the user is likely to enjoy this movie as it combines adventure, drama, and war elements which they enjoy.
- North by Northwest (1959): Yes, the user is likely to enjoy this movie as it combines action, adventure, mystery, romance, and thriller elements which they enjoy.
- Bridge on the River Kwai, The (1957): Yes, the user is likely to enjoy this movie as it combines adventure, drama, and war elements which they enjoy.
- Good, the Bad and the Ugly, The (Buono, il brutto, il cattivo, Il) (1966): Yes, the user is likely to enjoy this movie as it combines action, adventure, and western elements which they enjoy.
- Great Escape, The (1963): Yes, the user is likely to enjoy this movie as it combines action, adventure, drama, and war elements which they enjoy.
- Spirited Away (Sen to Chihiro no kamikakushi) (2001): Yes, the user is likely to enjoy this movie as it combines adventure, animation, and fantasy elements which they enjoy.
- Army of Darkness (1993): Yes, the user is likely to enjoy this movie as it combines action, adventure, comedy, fantasy, and horror elements which they enjoy.
- Day of the Doctor, The (2013): Yes, the user is likely to enjoy this movie as it combines adventure, drama, and sci-fi elements which they enjoy.
```

Figure 3. The LLM considers the user preferences and adds context and its own global knowledge to the movies suggested by the RS Library

This is why in **Figure 1** the difference between the ratings was very low but there was minimal overlap in the rankings of the two outputs.

2. High Difference in Rating

As we can see in the figure below, the LLM and the RS library completely disagree with each other in terms of ranking and rating. This happens particularly because the RS library just uses data points to get the recommendations.

```
Final List Generated by RS
{'Movie': 'Bridge on the River Kwai, The (1957)', 'Rank': 1, 'Rating': '4.77'}
{'Movie': 'Lawrence of Arabia (1962)', 'Rank': 2, 'Rating': '4.74'}
{'Movie': 'City of God (Cidade de Deus) (2002)', 'Rank': 3, 'Rating': '4.68'}
{'Movie': 'Wallace & Gromit: The Best of Aardman Animation (1996)', 'Rank': 4, 'Rating': '4.67'}
{'Movie': 'Princess Bride, The (1987)', 'Rank': 5, 'Rating': '4.65'}
{'Movie': 'WALL·E (2008)', 'Rank': 6, 'Rating': '4.62'}
{'Movie': 'Star Wars: Episode V - The Empire Strikes Back (1980)', 'Rank': 7, 'Rating': '4.60'}
{'Movie': 'Stand by Me (1986)', 'Rank': 8, 'Rating': '4.59'}
{'Movie': 'North by Northwest (1959)', 'Rank': 9, 'Rating': '4.58'}
{'Movie': 'Army of Darkness (1993)', 'Rank': 10, 'Rating': '4.54'}
Final List Generated by LLM
{'Movie': 'Wallace & Gromit: The Best of Aardman Animation (1996)', 'Rank': '1', 'Rating': '4.5'}
{'Movie': 'Princess Bride, The (1987)', 'Rank': '2', 'Rating': '4.0'}
{'Movie': 'WALL·E (2008)', 'Rank': '3', 'Rating': '4.0'}
{'Movie': 'Star Wars: Episode V - The Empire Strikes Back (1980)', 'Rank': '4', 'Rating': '3.5'}
{'Movie': 'Stand by Me (1986)', 'Rank': '5', 'Rating': '3.0'}
{'Movie': 'North by Northwest (1959)', 'Rank': '6', 'Rating': '3.0'}
{'Movie': 'Lawrence of Arabia (1962)', 'Rank': '7', 'Rating': '2.5'}
{'Movie': 'City of God (Cidade de Deus) (2002)', 'Rank': '8', 'Rating': '2.5'}
{'Movie': 'Army of Darkness (1993)', 'Rank': '9', 'Rating': '2.0'}
{'Movie': 'Bridge on the River Kwai, The (1957)', 'Rank': '10', 'Rating': '2.0'}
```

Figure 4. A list of top 10 movies generated by the RS library for a user in the adventure genre followed by the improved list generated by the LLM for the same user. This is an example of high RMSE and high MAE.

As we can see below, the user actually likes movies that are charming, quirky and have entertaining plots. This personality of user's preference is what cannot be read by the RS library directly.

```
Based on the user's preferences for charming animation, witty humor, endearing characters, whimsical storylines, memorable songs, lovable animal characters, quirky humor, entertaining plots, and enchanting fantas
y elements, it is likely that the user resonates with films that have a mix of humor, heartwarming moments, and imaginative storytelling. The user seems to enjoy movies that offer a sense of escapism and entertai
nment, with a focus on creativity and positivity.

It is possible that the user's preferences align with the general audience's views to some extent, as they both appreciate similar aspects in films such as humor, memorable characters, and entertaining plots. How
ever, the user may also have their own unique tastes and preferences that influence their choices in movies. It is likely that the user enjoys films that have a broad appeal but also offer something special or un
ique that resonates with them personally.
```

Figure 5. The user preferences as described by the LLM. This is a particular preference of witty and quirky humor with interesting plots and endearing characters.

However, with the help of LLM we can see how most of the movies do not even align with the users taste in movies because they are so intense and display content of drama and war or crime and thrillers. If these movies were directly recommended to the user it could dampen the user experience. With the help of LLMs we can use its context and global knowledge to understand if the movies plots, characters and ideas align with the user's preference or is it just recommended because it is a famous movie.

```
- Movie: Bridge on the River Kwai, The (1957) - No. This movie is more focused on drama and war themes rather than the charming animation, whimsical storylines, and lovable characters that the user seems to enjoy
.
- Movie: Lawrence of Arabia (1962) - No. Similar to the previous movie, this film is more serious and dramatic in nature, which may not align with the user's preferences for light-hearted and imaginative storytel
ling.
- Movie: City of God (Cidade de Deus) (2002) - No. This movie is more intense and gritty, with a focus on crime and thriller elements, which may not be in line with the user's preference for enchanting fantasy el
ements and quirky humor.
- Movie: Wallace & Gromit: The Best of Aardman Animation (1996) - Yes. This movie fits well with the user's preferences for charming animation, witty humor, and lovable characters.
- Movie: Princess Bride, The (1987) - Yes. This movie offers a mix of adventure, comedy, and fantasy elements, which align with the user's preferences for whimsical storylines and memorable characters.
- Movie: WALL·E (2008) - Yes. This film combines adventure, animation, and romance with a focus on creativity and positivity, which may resonate with the user's preferences.
- Movie: Star Wars: Episode V - The Empire Strikes Back (1980) - Yes. This movie offers a blend of action, adventure, and sci-fi elements, which may appeal to the user's love for entertaining plots and enchanting
 fantasy elements.
- Movie: Stand by Me (1986) - No. While this movie falls under the adventure and drama genres, it may not offer the same level of whimsical storytelling and quirky humor that the user seems to enjoy.
- Movie: North by Northwest (1959) - No. This film is more of a mystery and thriller, which may not align with the user's preferences for light-hearted and heartwarming moments.
- Movie: Army of Darkness (1993) - No. This movie combines action, adventure, comedy, fantasy, and horror elements, which may not fully match the user's preferences for charming animation and enchanting fantasy e
lements.
```

Figure 6. The LLM considers the user preferences and adds context and its own global knowledge to the movies suggested by the RS Library. Here the LLM disagrees with most of the choices given by the RS library as the user preferences do not seem to align with the recommendations

Discussion:

Upon conducting analysis of the results in the evaluation csv we can see how the ratings differ vastly between the RS and LLM. While an ideal RMSE would be below 0.5 and an acceptable range would be between 0.5-1, here for over 20% of the cases the RMSE was higher than 1 and for 70% cases it was higher than 0.5 showing the huge discrepancy between the ratings given by an LLM and RS Library. The same suite followed for the MAE in ratings where more than 56% fell above 0.5 showing the difference in the two systems when rating the same movies.

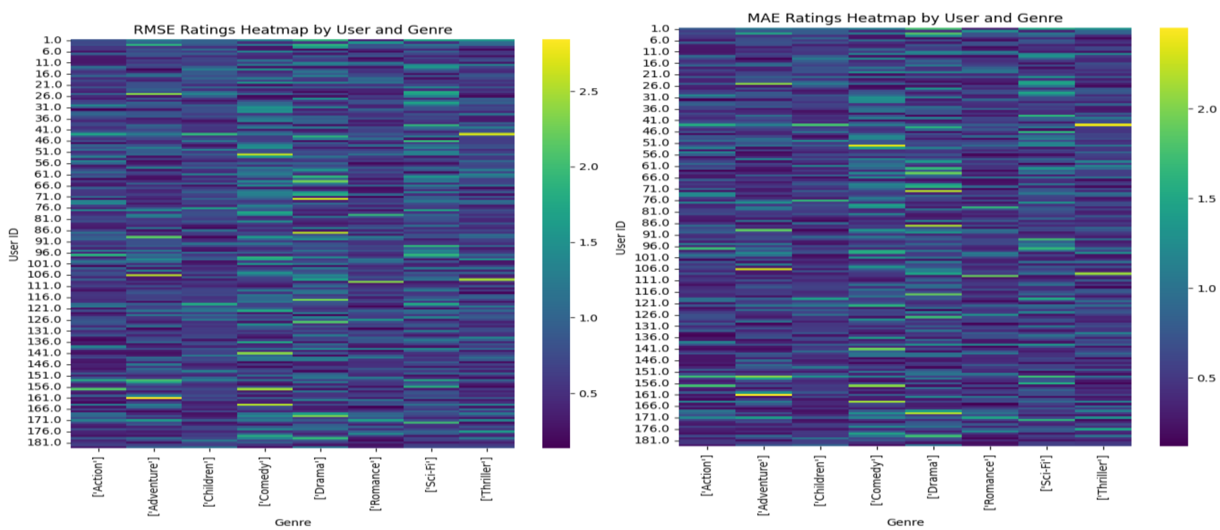

Figure 7. Heatmaps of MAE and RMSE based on the outputs stored in the evaluation csv

In the diagram below, due to even the small differences in rating the rankings differ highly. It is only in 9% of the cases in this evaluation batch that there is a 100% overlap in ranking. The average for the RS library and LLM ranking falls below 25% for most of the genres.

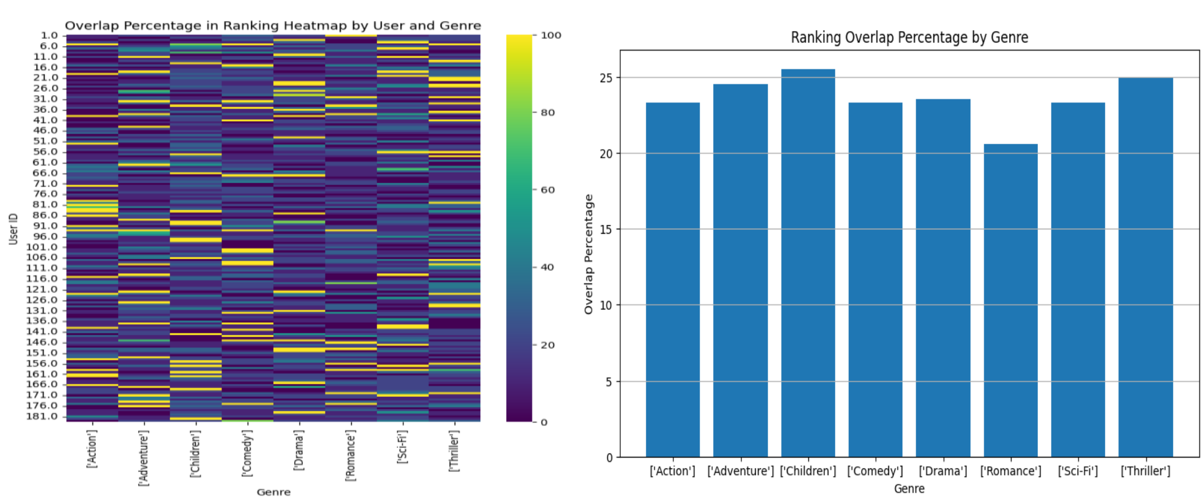

Figure 8. Heatmap and Bar Graph for the ranking overlap by the LLM and the RS based on the output of evaluation csv

Hence from the above analysis we can conclude that once we add the LLM to the system and use its knowledge to add the context and understand user preferences the dynamics of the recommendations completely change and the ratings and rankings have a significant impact.

Conclusion:

The comparison between Large Language Models (LLMs) augmented with contextual awareness and traditional Recommender System (RS) libraries reveals significant differences in generating personalized recommendations and rating predictions. Our experiment involved evaluating both systems across various metrics and analyzing the top-ranked items recommended by each.

In scenarios where the difference in ratings between LLMs and RS libraries is low, we observed that LLMs leverage contextual information to align recommendations more closely with user preferences. By analyzing user interactions with movies in similar or different genres and incorporating the user's unique preferences, LLMs provide personalized recommendations that may differ in ranking but align closely in terms of user satisfaction and enjoyment.

Conversely, when a significant difference in ratings exists between LLMs and RS libraries, it underscores the limitations of traditional RS approaches in capturing nuanced user preferences. LLMs excel in understanding complex user preferences, such as specific genres, themes, or narrative styles, which may not be adequately captured by RS libraries relying solely on data points.

Future research can focus on generalizing the findings across various types of Large Language Models (LLMs) with contextual awareness. Investigating how different LLM architectures, such as Llama and Mistral impact personalized recommendations and rating predictions can provide valuable insights into the versatility and effectiveness of LLM-based recommendation systems.

In conclusion, integrating LLMs with contextual awareness enriches the recommendation process by offering more personalized and user-centric recommendations, thereby enhancing user satisfaction and engagement with the platform.

Code Repository and Documentation

The code repository for this experiment is hosted on GitLab. You can access it by visiting the URL [2]. Once you are on the repository page, you will find detailed documentation in the README.md file. This documentation contains instructions on replicating the experiment, installing necessary dependencies, running scripts, and generating the results. You can navigate through the repository folders to explore datasets, experiment results, and any additional resources related to the project.

References

[1] K. Modi, "EnReLLM: Enhanced Recommendation Systems through Integration of Large Language Models," Google Docs. [Online Document]. Available: https://docs.google.com/document/d/1WIP4iJnWR3maU1wpLwEGtQUlCMZx_ezAJuGD7tbKFZk/edit?usp=sharing. [Accessed: April 14, 2024]

[2] K. Modi, "ResearchProject9302," *Gitlab Dalhousie FCS*. [Code Repository]. Available: https://git.cs.dal.ca/krishnam/researchproject9302. [Accessed: April 14, 2024]