Università della Svizzera italiana

Faculty of Informatics

Bachelor Thesis

November 17, 2024

A Comparative Study between Google and Conversational AI Models

Harkeerat Singh Sawhney

Abstract

The ubiquitous nature of the internet has led to a surge in online activity among young children, making it a go-to source for information. However, the reliability and accuracy of online information tend to be often questioned, particularly for children who are unaware of its veracity. Recently, there has been a much more user-friendly approach to accessing online information in the form of Conversational AI Models. The thesis would draw a comparative analysis between Google Search and Conversational AI models such as ChatGPT, to determine the more effective option in the context of Children's online search behavior. The study would make use of innovative tools to simplify the comparison between ChatGPT and Google Search, aiming to identify the technology that yields superior outcomes. Along with that this thesis also includes the development of modules to predict the relevance of query results which can be used to determine whether search query outcomes are pertinent or not. Finally, with the goal of improving the quality of information available to young children, this study aims to contribute to the field of online search behavior research.

Advisor
Prof. Landoni Monica

Advisor's approval (Prof. Landoni Monica): Date:

Contents

List of Figures

List of Tables

1 Introduction

The increasing accessibility of the internet has revolutionized the way young children access information. With the advent of online resources, children now have access to a vast amount of information. However, on the other hand, online information [18] is often met with skepticism due to the potential unreliability and inaccuracy of the sources. Along with that, young children are often unaware of how to identify credible sources of information, which can lead to a significant impact on their knowledge acquisition.

In recent years, conversational AI models such as ChatGPT have emerged as a user-friendly way of accessing online information. Such models have allowed users to converse with the machine in a natural language format, which can be particularly effective for children. Still, skepticism stays, as to if such conversational AI models improve on the problems that traditional searching methods have introduced, or does it make it worse due to the generative nature of such models.

1.1 Structure of the Thesis

This thesis report will follow the sequential procedure of the project, which will start from the inception to the end of the project. Section 2 contains the background and motivation of the project, which will include the fundamental concepts of Information Pollution and Hallucination. It will also include the impact of such phenomena on children and how it affects them as they are the primary focus of this research. This section will also introduce the two technologies which we are going to compare; ChatGPT and Google Search. At the end of this section, some literature work would be mentioned which is written in the same field and would give better insight into the subject.

Section 3 will focus on the methodology of the project, which will include in detail the different steps which were taken in order to collect the data, how the data was pre-processed, and how it was used to have different analyses done one it with different metrics. There were many different types of metrics used ranging from multiple different similarity analyses to sentiment analysis. The main reason for this was to have a robust analysis of the data which was collected.

Section 4 will utilize the data collected and would aim to develop a machine-learning model which can predict the relevance of the query results. The model would be trained on the collected data while utilizing various different machine-learning techniques.

Finally Section 5 will conclude the thesis by discussing what all the results mean and as to which technology is better for children to use right now. It will also discuss the limitations of the project and would describe directions for future work.

At last all of the work which was done in this thesis would be available on GitHub [1]. The GitHub repository contains all the scripts which were developed throughout the project and also contains the data which was collected. It also contains all the results which were obtained from the analysis of the data. There is robust documentation available on the repository which would help anyone who wants to replicate the results or wants to use the scripts for their own use.

2 Background and Motivation

2.1 Information Pollution and Hallucinations

As of today one of the major challenges in surfing on the internet is Information Pollution [16]. Information Pollution is the contamination of the information supply with irrelevant, redundant, unsolicited, hampering, and low-value information. Information Pollution is caused due to the overwhelming and excessive amount of information, which is often of low quality. Such information often circulates through many different communication channels, especially on search engines. This can cause long-lasting damage to those who receive such incorrect and distracting information, as it causes affects their knowledge and as well their quality of life.

There are many different forms in which Information Pollution can be presented. One such form is misinformation [12], which is caused by inaccurate information which is shared with each other in an attempt to deceive. Another form of which information pollution can be present is disinformation, which, unlike misinformation, is deliberately

[1] https://github.com/Harkeerat2002/bachelor-project-scripts

fabricated to mislead people. As it can be noticed many times the origin of such information is malicious, however, due to the nature of the internet, they are very easily shared amongst everyone without being checked.

Another form in which Information Pollution can be present is Hallucinations [17]. Hallucinations in information technology tend to present significant challenges in various domains, ranging from user interface design to more recently artificial intelligence systems. Such hallucinations are instances where users tend to perceive or experience inaccurate, misleading, or irrelevant information, which tends to erroneous decision-making and compromised user experience. Such phenomena usually can arise to factors such as limitations in data quality, algorithmic biases, and inefficiencies in information retrieval systems.

2.2 Children's Search Behavior

As much as it is challenging for adults to tackle Information Pollution, it is much more difficult for children to tackle it. One of the primary victims of Information Pollution is children, as Information Pollution expose their vulnerability regarding their developing cognitive abilities. Children can come across false or misleading information, which can lead to the formation of misconceptions and incorrect beliefs regarding various subjects [10]. This can be very harmful as it can hinder their understanding of the world and have a negative impact on making informed decisions. Another factor that can affect children is their development of Critical Thinking Skills [7]. If children are having constant exposure to Information Pollution, it can make it challenging for children to develop critical thinking skills. Children can struggle to discern reliable sources of information, evaluate their credibility and differentiate between a fact and an opinion.

Along with that Hallucination poses an additional challenge for children in Information Retrieval [17]. This again is challenging for Children to tackle, as hallucinations tend to take advantage of children. One of the main factors which are harmful to children is misinformation and false beliefs. Children rely on search engines or online platforms for information, hence hallucinations in the search results can lead to the dissemination of inaccurate or false information. This could possibly lead to children shaping their understanding of the world around them with misconceptions or false information. This also causes another side effect, which is that it can force children to develop Cognitive Biases with distorted perspectives. Children are still developing their critical skills, and if hallucination is presented in the search results it can lead to them accepting biased, or misleading information as factual.

2.3 ChatGPT vs Google Search

The primary method for getting information for a long time has been Google Search [6]. It is a web search engine which is developed by Google, one of the world's leading technology companies. Its purpose is to help users find information on the internet by providing a list of relevant web pages, documents, images, videos, and other types of content in response to their search queries. In order to provide the most appropriate search results it takes various factors into account to determine their relevance and the rankings. These factors include the keywords and language used in the query and its popularity and the quality of the web pages. This makes powerful use of Google's algorithm which helps with the evaluation of the web pages. This is powered by the web crawlers which explore the web to regularly find pages to add to their index. Along with all the core web search functionalities, Google Search also offers various specialized search features, such as image search, news search, video search, and more.

However, with the recent advancements in the field of Natural Language Processing, conversation AI models have become much better at providing relevant information with their robust model. One such model is ChatGPT [5], which has been very popular recently. The product is rather very simple to use, there is a text box where the user can type in their questions and from that question, ChatGPT would generate a response. The response is generated by the model by taking into account the context of the question and also the previous question. This makes it very easy for the user to get the information they want.

While ChatGPT does provide very useful information and acts as an assistant, Google Search does also provide more types of media as a result. Therefore, it is important to compare the two technologies and see which one is better for children to use. This is important as it would help us understand which technology is better for children to use and which one is more prone to Information Pollution and Hallucinations.

2.4 Literature Review on related work

To get a better understanding of such a topic we went through several different literatures which went into detail with topics such as ChatGPT and Information Pollution. The main aim of reading such literature was to modify our project and seek where there is still some research needed. Out of the many different papers which were read, there were two papers which were very helpful.

The first paper is titled *"What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature"* [9]. This paper goes in-depth about the performance and application of ChatGPT in various fields such as economics, programming, mathematics, etc. Their finding stated that although ChatGPT has a very strong potential to serve as an assistant for instructors, it still is a challenge as it can generate incorrect information. Also, even though in fields like programming ChatGPT's performance was satisfactory, it performed very poorly in other fields such as mathematics.

The second paper which we went through is titled *"On the Risk of Misinformation Pollution with Large Language Models"* [15]. This paper does a comprehensive investigation into the potential misuse of modern Large Language Models for the generation of credible-sounding information. The paper discusses how it established a model which simulates potential misuse scenarios so that they can assess how LLMs utilize the produced misinformation. Their initial results showed that there are trends of defensive strategies, and it was concluded that much more work needs to be still done for the challenge of misinformation pollution.

Overall, these 2 papers highlight the presence of ChatGPT in the landscape of gaining information in the modern world. It highlights the possible threats that ChatGPT can cause but also states that its potential is good enough to risk those threats.

2.5 Research Questions and Objectives

Hallucination and Information Pollution can be mitigated by giving essential education to children about critical evaluation and information sources and encouraging them to seek guidance from trusted sources. However, these days with conversation systems powered by artificial intelligence the line between trustworthy information and compromised information is disappearing day by day. Even adults are more likely to mistake what a trustworthy source is, so not much can be expected from children. Therefore, the question arises, are the new conversations AI models safer for children to interact with, or do the traditional search engine still hold an edge over the new technology?

Hence, this thesis will aim to identify between ChatGPT and Google Search, which offers the best overall service to children, so as to present the least Information Pollution and Hallucination in the results it provides. The study will first collect queries from children and will perform different analyses on them. In the end, the goal is to also make a logistic regression-based machine learning module trained on the data which is collected. This is to determine if the information gathered either from ChatGPT or Google Search is relevant or not. At last, finally, the objective is to give a practical conclusion as to at this point what is a better service to utilize for children.

3 Methodology

3.1 Data Collection

To collect data from Google Search and ChatGPT, the initial step is to collect data directly from children. More specifically data is needed derived from search queries that children use on day-to-day bases to search online. This task presents certain challenges as it necessitates not only organizing and allocating time for the children but also creating a comfortable environment where they feel at ease. Considering these factors, it was decided to utilize pre-existing data that have been previously collected from children under the guidance of Professor Landoni Monica and her team, how collected ethical approval from the school where the study took place originally. Even though the data is old, it is still believed that it is not old enough to be considered outdated. A total of 77 queries were collected, which is a healthy size for data collection to be used for this explorative study. Table 1 shows a sample of the queries that were collected.

Query Number	Query
1	when was the last time capulin erupted
2	what year did the Capulin Volcano in New Mexico Last errupt
3	what year did the Capulin Volcano Last errupt
4	in Arizona what dessert is the Organ Pipe Cactus Monument found in
5	in Arizona what dessert is the Organ Pipe Cactus Monument in Arizona found
...	....
...	....
...	....
77	pictures of hawaii

Table 1. Queries collected from Children

3.1.1 Google Search

In order to collect the query results from Google, Google's own API known as Google Search API was used. Google Search API or officially known as Google Custom Search JSON API, allows us to use Google's Search functionally. Through this, the results can be retrieved in a JSON format which can then be used to easily do analysis on them. In order to retrieve the results the following steps were taken.

1. textbfObtaining an API Key:
   There are 2 keys that are needed to get the results from Google Search. The first key is the API key itself, which is used to authenticate the request to the API. The second key is the SEARCH ENGINE KEY which is used to identify the specific search engine which it wanted to be

2. **Constructing a Search Request:**

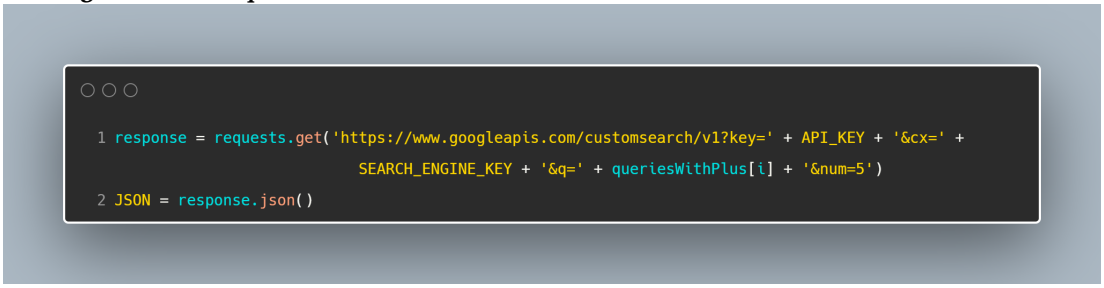


Figure 1. Google Search Configuration

   To send the request, we first need to construct the Search Request which would be sent. The search request has 5 main factors in it. Those are Base URL, API Key, Search Engine ID, Search Query, and Number of Search results. The base URL is the standard URL to use the Google Custom Search API, the API key is used for the authentication, the Search Engine ID is used to specify the engine, Search Query is the query itself from which we want the results and finally, the number of queries is the number of results we want, we have kept to 5. The figure below is the visual representation of what the search request looks like.

3. **Sending the Request:**
   Finally, the requests are sent, and the data is collected. The data is collected in a JSON which gets processed later. However, with the free plan, we can only send 100 requests per day. So in case there is some bug, either a new API key needs to be generated or we have to wait for a day.

4. **Processing the Response:**
   At last, the data is received initially in a JSON format, as it is easier to handle the data in that format. However, the JSON data is converted to CSV for consistency, so it is easy to read the data and visualize it with easier tools and libraries.

3.1.2 ChatGPT

In order to collect data from ChatGPT, Open AI's API was used to accomplish this. First, the API key was obtained, and after that the request was configured. In the configuration, there were multiple options, but all the options were

configured in such a way that it acts fully as ChatGPT. After that, a request was sent for each of the queries. Unlike Google, for each of the search queries, only one response is generated, as unlike Google where at a time one can see multiple snippets of the results, here only one response is generated at a time. The following were the steps taken with more details describing them.

1. **Obtaining the OpenAI API Key:**
   In order to use ChatGTP, naturally an API key is required from Open AI. That can be accessed by making the account and registering for the key. One thing which must be kept in mind is the price of the key since it uses tokens, and each token has a different price range.

2. **Configuring the ChatGPT Model:**

Figure 2. ChatGPT Configuration

   Open AI gives multiple options to configure what kind of request is needed. In our case, the model which is used is the gpt-3.5-turbo which at the time was the bleeding edge engine that Open AI offered. Also, the temperature value was configured to default which is 0.5. The temperature value controls the creativity of ChatGPT, the higher the value the more creative it gets; however, we want to the standard which is why it was kept to 0.5. At last, the role of ChatGPT was configured. There is an option that we have access to, called a message parameter, which is a list of dictionaries representing the messages in the chat session. Each dictionary contains 2 keys. The first key is “roles”, which tends to specify whether the message is from the user or the system. The second key is “content” which contains the content of the text message. In our case, there were 2 messages which were sent. One was from the system stating, “You are a helpful assistant”, and the second message was from the user stating the query from those collected from children. This configuration can be seen in Figure 2, which was used to configure the request.

3. **Sending the Request:**
   At last, the request is sent and retrieved in a JSON format. There is no bottleneck that was found with OpenAI as Google, so the procedure was very straightforward.

4. **Processing the Response:**
   At last, the data is received initially in a JSON format, as it is easier to handle the data in that format. However, the JSON data is converted to CSV for consistency, so it is easy to read the data and visualize it with easier tools and libraries.

3.2 Data Preprocessing

This is a crucial step for the next sections Data Analysis and Machine Learning, as using the raw data could cause many inconsistencies. It is important to transform the data in a formal which can be then easily analyzed and be used with different algorithms. The main goal is to improve the quality of the data and make it more suitable for our project.

What we have done is, we have taken the results from Google and ChatGPT and processed them to create a list of dictionaries that contains the user queries and an array of results. For Google results, the array of results has 5 results in them (the first five returned), while for ChatGPT it has only one result in them. Also, in some cases of Google Results, there were characters that were not needed and had no use of them in the actual analysis. These results were converted from CSV to JSON, from which algorithms such as Sentiment Analysis, Similarity Analyses, and Machine Learning Model were used for their purpose.

3.3 Data Analysis

3.3.1 Query Length Analysis

The main purpose of the Query Length Analysis [1] was to compare the length of the responses from both Google Search and ChatGPT. The main purpose of this analysis is to see which one gives more text. Having more text could have several advantages such as much more detailed and comprehensive information related to the topic being searched on. However, they are also several disadvantages to this as well, such as the information can be overwhelming and reading and assessing it can be very time-consuming.

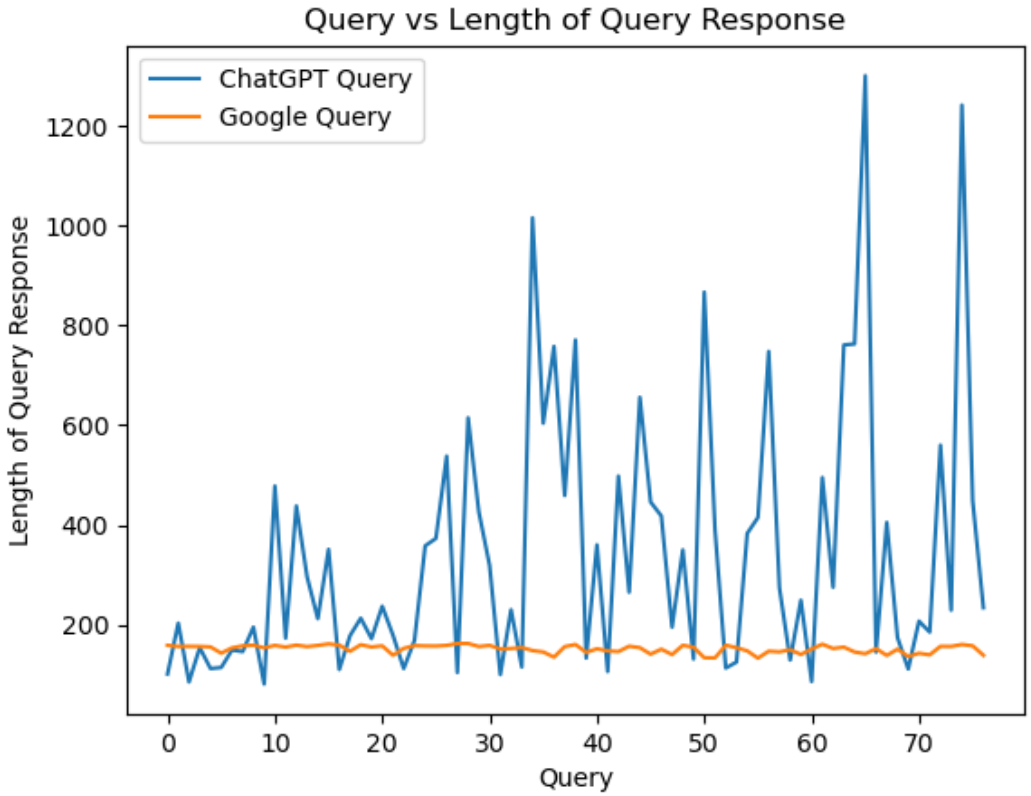


Figure 3. ChatGPT Query Length vs Google Search Query Length

From Figure 3 we can see that Google and ChatGPT have big differences in terms of the length of their results. Google results in average length is around 151 characters and as compared to ChatGPT its average results length is 343 characters, more than twice as compared to Google results. However, Google Results tends to be more consistent with its query length, as the max is 162 characters and the min is 132 characters. On the other hand, the max for ChatGPT is 1321 characters, while the min goes as low as 81 characters. Table 2 is representing the findings.

	Google Search	ChatGPT
Average (Characters)	151	343
Max (Characters)	162	1321
Min (Characters)	132	81

Table 2. Query Length Analysis

With this there are several findings, such as ChatGPT results length can vary depending on what question is being asked. This is a good experience as it can identify which topics need more explanation and which need less. However, it can also lead to more information pollution in the text if it tries to generate more. On the other hand, Google results tend to stay consistent with its length, which could mean that for some results it does not provide more details, however, it also can possibly mean that whatever it is providing is more likely relevant as compared to ChatGPT.

3.3.2 Named Entity Recognition (NER)

Named Entity Analysis [13] is a natural language processing technique that involved identifying and classifying named entities in text into different predefined categories. For example, such categories can be a person, names, organization names, locations, dates, and so on. This analysis can help us judge how much important information a response contains. The more it contains, we could use it as a factor to conclude the more relevant information it has. Table 3 is a representation of all the categories which is contained in the Named Entity Analysis

Categories	Description
Person	Refers to individuals, including their names, titles, and professions.
Organization	Refers to companies, institutions, and other organizations, including their names and types.
Location	Refers to places, including countries, cities, and other geographical locations.
Date	Refers to specific dates or date ranges, including days, months, years, and seasons.
Time	Refers to specific times or time ranges, including hours, minutes, and seconds.
Money	Refers to monetary values, including currencies, amounts, and units of measurement.
Percentage	Refers to percentages, including numerical values and units of measurement.
Product	Refers to products, including their names and types.
Event	Refers to specific events, including their names and types.
Miscellaneous	Refers to other named entities that do not fit into the above categories.

Table 3. Named Entity Analysis Categories

To start the analysis, we first need the en-core-web-sm model, which is a pre-trained statistical model for natural language processing. It is part of the spacy library which is a popular open-source library for NLP in Python. For each of the responses of ChatGPT and Google, we use the NLP object to perform the NER and to write the results to the output file. It does this by tokenizing the text into individual words and assigning part-of-speech tags to each word. We can then use the metadata for each word to associate NLP tasks, in which in our case we are naming each entity.

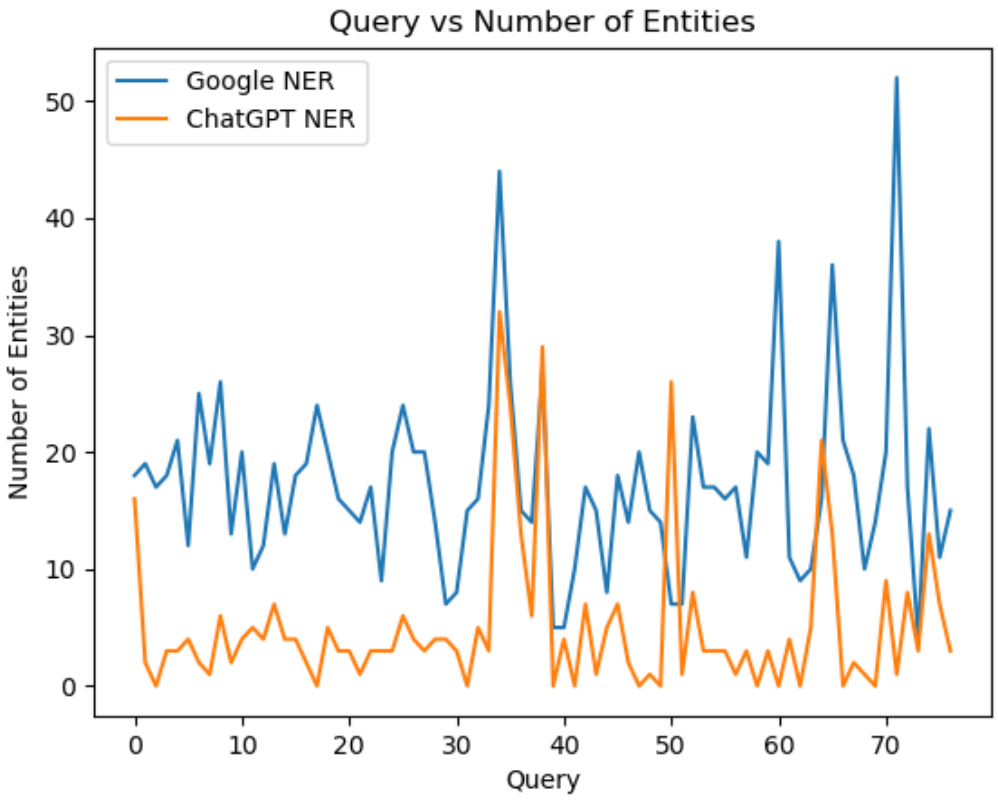


Figure 4. Named Entity Analysis

When looking at the statistical analysis, the average entity is much higher for Google Results. Google has an average of 18.2 entities for each search, while ChatGPT has an average of 6.11, nearly 3 times less as compared to

Google. There can be some reasons for this, such as Google results giving out 5 results as compared to ChatGPT giving out one generate result.

When it comes to seeing what the top 5 entities are can be seen that ORG, GPE, CARDINAL and PERSON are in the top 5 for both Google Search and ChatGPT. However, the only 2 different entities which aren't present in the top 5 are DATA (present in Google) and LOC (present in ChatGPT). Table 4 has all the statistical analysis for the Named Entity Analysis.

	Google Search	ChatGPT
Average	18.2	6.11
Max	35	17
Min	8	1

Table 4. Named Entity Analysis

Entity	DATE	ORG	GPE	PERSON	CARDINAL
Occurrences	304	237	236	227	211

Table 5. Top 5 Google Entities

Entity	GPE	ORG	CARDINAL	PERSON	CARLOCDINAL
Occurrences	108	105	71	49	30

Table 6. Top 5 ChatGPT Entities

3.3.3 Similarity Analysis

Similarity Functions are mathematical functions that measure the similarity between two objects or data points. The functions take input data and produce a numerical score that reflects how similar or dissimilar those objects are. In the context of our research, we can use this method of analysis to compare how similar or how dissimilar are Google Search and ChatGPT results. This data could be very helpful with the context of other data when combined, which can help us give meaningful results. We will be using 4 different Similarity Functions which are Cosine Similarity, Jaccard Similarity, Euclidean Similarity, and Pearson Similarity. The data when finally collected is represented in a CSV as the below table represents. Table 7 shows the data structure of the Similarity Analysis.

User Query	ChatGPT Result	Google Result	Jaccard	Cosine	Euclidean	Pearson
...	....	....	....	....	....	....

Table 7. Similarity Analysis Data Structure

Cosine Similarity

Cosine Similarity [3] is a measure of similarity between two non-zero vectors defined in an inner product space. It is the measurement of the cosine of the angle between the vectors, which is the dot product of the vectors divided by the product of their lengths. Along with that Cosine Similarity also shows that it does not depend on the magnitudes of the vectors but on its own vectors. The Cosine Similarity gives a value which is between -1 and 1. 1 indicates that the two vectors being compared are identical, while 0 indicates that the two vectors are orthogonal (have no correlation), and -1 indicates that the two vectors are diametrically opposed.

In our case where we are processing the information, each word is assigned a different coordinate. Then the document is represented by the vectors of the number of occurrences of each word in the document. With this, we can judge how similar the documents are. Hence, the cosine of two non-zero vectors can be derived by using the Euclidean Dot Product Formula.

$$A \cdot B = \|A\|\|B\| \cos(\theta) \tag{1}$$

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{2}$$

```python
def CosineSimilarity(a, b):
    # Create CountVectorizer object
    vectorizer = CountVectorizer().fit_transform([a, b])

    # Get vectors for each sentence
    vector1 = np.array(vectorizer[0].todense())[0]
    vector2 = np.array(vectorizer[1].todense())[0]

    # Normalize vectors
    norm1 = np.linalg.norm(vector1)
    norm2 = np.linalg.norm(vector2)

    # Compute cosine similarity
    cosine_similarity = np.dot(vector1, vector2) / (norm1 * norm2)

    return cosine_similarity
```

Figure 5. Cosine Similarity Code Snippet

Figure 5 represents the snippet of the code which implements the theory. First, the CountVectorizer object is initialized from the Sklearn library. With this, we can convert the input string a (ChatGPT) and b (Google Search) to convert them into vectors of word counts. After that, we fit the vectorizer to the input strings and transform them into the vectors. From there the resulting vectors are normalized to ensure that the length of the vectors does not affect the similarity measure. Lastly, the cosine similarity between the two vectors is calculated using the dot product of the two vectors and dividing them by the product of their norms.

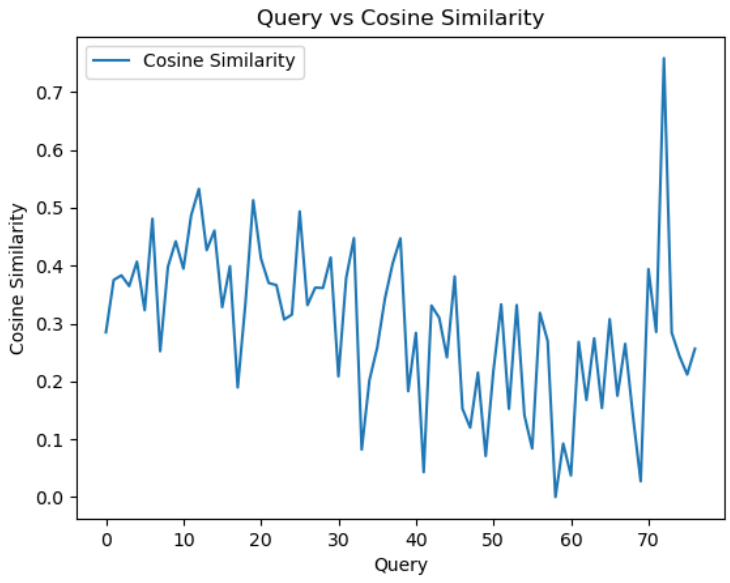


Figure 6. Query vs Cosine Similarity

From the results, we can determine that the average cosine similarity value is 0.296, which indicates that the ChatGPT and Google Search from cosine similarity analysis have some degree of similarity, but they are far from identical. The maximum the cosine similarity went is 0.941, while the minimum it went is 0.0. This shows there are queries that are very similar and there are some which have no correlation with each other. The median value is 0.29617443887954614 which is very close to the average, meaning that there are not many extreme values.

average	max	min	median
0.29629571650045594	0.9412657491818057	0.0	0.29617443887954614

Table 8. Cosine Similarity Statistics

Jaccard Similarity

Jaccard Index Coefficient or also known as Critical Success Index [14], is a measurement used to compute the similarity between two asymmetric binary variables. The Jaccard index is defined as the size of the intersection of the two sets divided by the size of the union of the two sets. This can be written in the below notation of two sets A and B.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

Jaccard Similarity will be 0 if the two sets tend to share no values, and 1 if both sets are identical. We can also use it to find the dissimilarity between the two sets, however in our case, we focus more on the similarity.

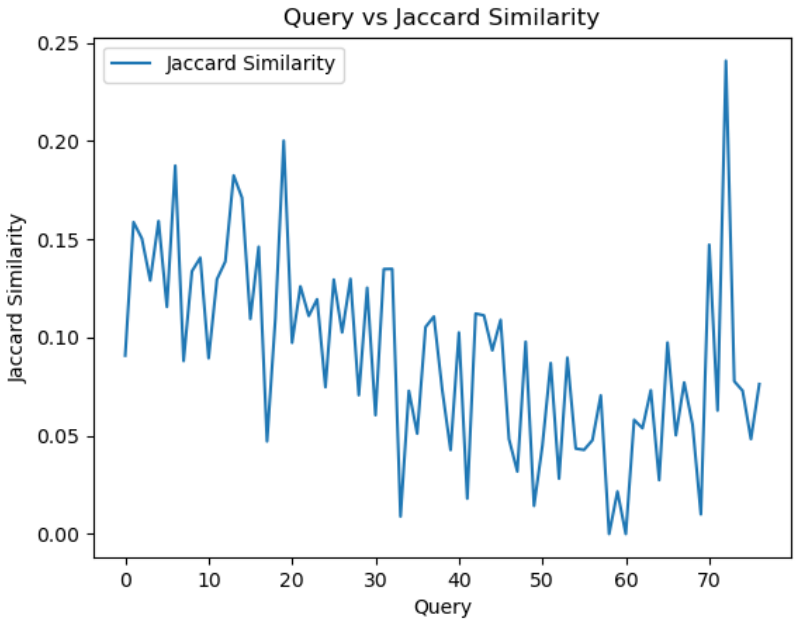


Figure 7. Query vs Jaccard Similarity

Hence from the evaluation we were able to determine that 9.09 % of the elements in ChatGPT and Google Search are in common. The maximum Jaccard Similarity is 0.42, while the lowest was 0. The median of the Jaccard Similarity was 0.0784. Therefore, from Jaccard Similarity it can be seen that ChatGPT and Google search gave out very different results in this form of evaluation. Table 9 shows the statistics of the Jaccard Similarity.

average	max	min	median
0.09091836722083915	0.42	0.0	0.0784313725490196

Table 9. Jaccard Similarity Statistics

Euclidean Similarity

Euclidean Similarity [4] is a measure used to determine the similarity or dissimilarity between two objects or data points in a Euclidean Space. This is calculated by the straight-line distance between two points in an n-dimensional space. The formula below describes Euclidean Similarity.

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2} \tag{4}$$

The way in which we have implemented Euclidean Similarity is by first taking in two strings which are the ChatGPT result and the Google Search result. The function creates a CountVectorizer object which it fits in the two input strings. After that for each string, it gets a vector and normalizes it. Then it computes the Euclidean Distance between the two normalized vectors and returns the Euclidean similarity.

The way in which Euclidean Similarity is calculated is by $\frac{1}{1+EuclideanDistance}$. As mentioned before the Euclidean Distance is the measure of the distance between two points in the multidimensional space. Euclidean Similarity is between 0 and 1, where 1 indicates that both the strings are identical, while 0 indicates that they are completely different. We use the libraries NumPy and sklearn to perform and simplify the calculations. The NumPy library is used for the vector normalization, while the sklearn is used to create the CountVectorizer.

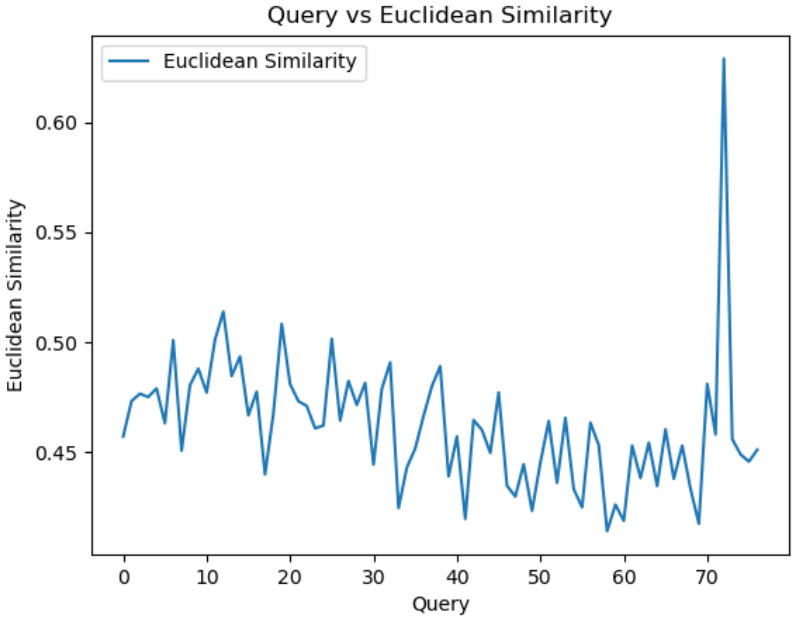


Figure 8. Query vs Euclidean Similarity

From Figure 8 there are several statistics that can be gathered. From these statistics, it can show that to some degree there is a similarity between the results returned by ChatGPT and Google search with an average similarity of 0.462. The maximum the similarity went was 0.745 while the lowest it went was 0. The median for Euclidean similarity is 0.457 which shows that the majority of the results are similar. Table 10 shows the statistics of Euclidean Similarity.

average	max	min	median
0.462	0.745	0.0	0.457

Table 10. Euclidean Similarity Statistics

Pearson Correlaion

Pearson Correlation Coefficient [2] is a measure of linear correlation between two sets of data. It is the ratio better the covariance of two variables with the product of their standard deviations. It is a normalized measurement of the covariance so that the result always has a value between -1 and 1. If the value is 1 it means that there is a total positive linear correlation, while if it is 0, it means that there is no correlation, and lastly, if it's -1, it means that there is a total negative correlation.

In order to calculate the Pearson Similarity, first as always, the with CountVectorizer objects were created to transform the two input strings into vectors. The vectors then get normalized, from which the Pearson correlation coefficient was calculated. This is done by using the Pearson Function from the scripy.stats library. The library uses the below equation to compute the coefficient.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

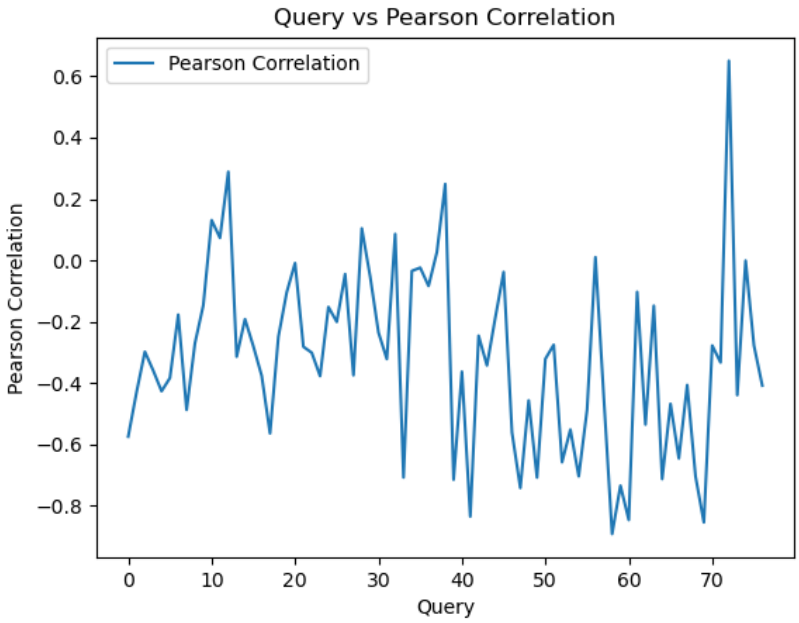


Figure 9. Query vs Pearson Correlaion

From Figure 9 we can have some statistics which can be used later for analysis. We understand that the average correlation coefficient is -0.320, meaning that there is a weak negative correlation between the two sources. It means that the frequency of some certain terms or entities is increased in one entity while in another it is decreased. The maximum correlation coefficient which went was 0.949, but in the other, the minimum is -0.955 meaning that there are some which are very similar, while others which are different. The median is -0.342 which is close to the average, which again also indicates that there is a negative correlation between the sources. Overall, it can be seen that there is some degree of correlation between ChatGPT and Google results, but their strength and direction vary depending on the comparison being made.

average	max	min	median
-0.320	0.949	-0.955	-0.342

Table 11. Pearson Correlation Statistics

3.3.4 Sentiment Analysis

Sentiment Analysis [11] is a very powerful technique that is used in natural language processing to determine different emotional tones behind a body of text. Such analysis involved analyzing the text to determine whether it is representing a positive, negative, or neutral sentiment. This analysis is helpful for our data, as it gives us a better understanding of what sentiment both Google and ChatGPT results give. We get the result between 0 and 1, 0 meaning the sentiment doesn't exist while 1 meaning the whole sentence is that sentiment. What we expect from both is to give neutral results, as the results that the user usually seeks, want neutral answers. Both for Google and ChatGPT it will be tricky, as Google shows results written by people where sentiment can easily creep in, while ChatGPT generates its responses automatically.

```
# ChatGPT Sentiment Analysis
score = analyzer.polarity_scores(chatgpt_result["result"][0])

# Gogole Sentiment Analysis
score = analyzer.polarity_scores(r)
```

Figure 10. Sentiment Analysis Code Score Snippet

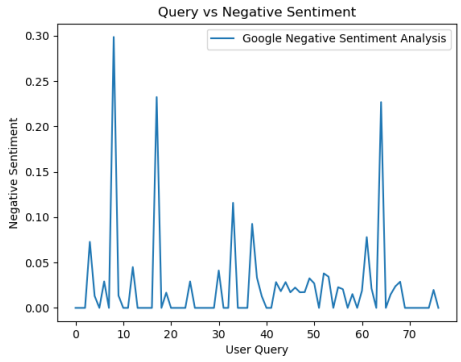
The sentiment analysis is performed on data which we collected, which is the Google Results, and ChatGPT

Results. We use the SentimentIntensityAnalyzer from nltk.sentiment module to analyze each sentence of the data. We then use the polarity score method to obtain scores for each sentence. The sentiment score includes, negative, neutral, and positive. The code snippet in Figure 10 shows how the score was obtained.

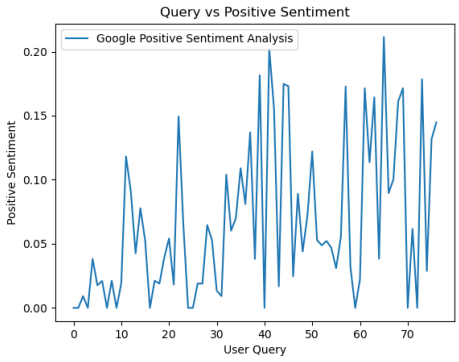
```
# ChatGPT Sentiment Analysis
comparsion_result = {
            "user_query": chatgpt_result["user_query"],
            "chatgpt_result": chatgpt_result["result"][0],
            "negative": score['neg'],
            "neutral": score['neu'],
            "positive": score['pos'],
            "compound": score['compound']
        }
# Google Sentiment Analysis
comparsion_result = {
            "user_query": google_result["user_query"],
            "google_result": google_result["result"],
            "negative": score['neg'],
            "neutral": score['neu'],
            "positive": score['pos'],
            "compound": score['compound']
        }
```

Figure 11. Sentiment Analysis Data Code Snippet

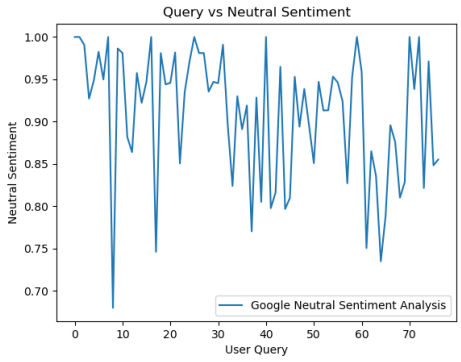
Figure 11 it shows the data structure which was used to get the results. It was initially collected in JSON and then transformed into CSV for consistency.



(a) Negative Sentiment Analysis Google



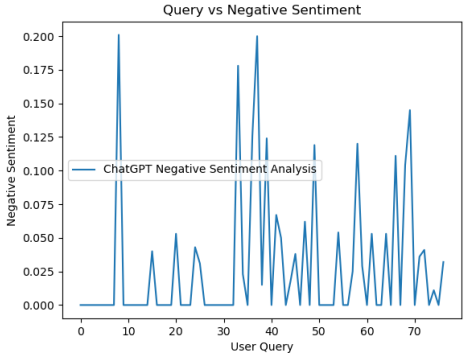
(b) Positive Sentiment Analysis Google



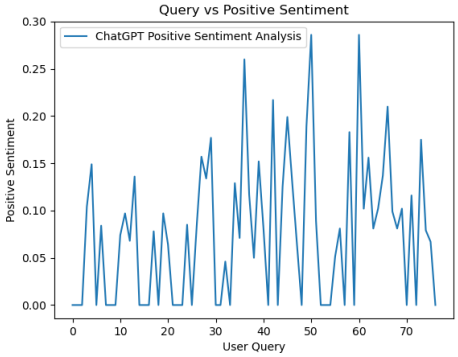
(c) Neutral Sentiment Analysis Google

Figure 12. Sentiment Analysis Google

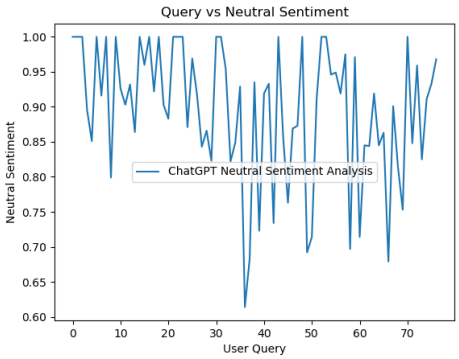
Figure 13 represent the Sentiment Analysis of Google. As most of the text is considered Neutral with the average being 0.91. The average for negative and positive sentiment is 0.03 and 0.02 respectively, with the highest for both went was 0.53 and 0.45. Also, if the mean is noticed for the neutral sentiment, it is 1, which is close to the Average. This analysis can show that Google results tend to have more neutral sentiment possibly resulting in less biased and more objective information retrieved.



(a) Negative Sentiment Analysis ChatGPT



(b) Positive Sentiment Analysis ChatGPT



(c) Neutral Sentiment Analysis ChatGPT

Figure 13. Sentiment Analysis ChatGPT

Figure 13 represent the Sentiment Analysis of ChatGPT. Similarly, to Google, most of the results from ChatGPT come under Neutral Sentiment with an average of 0.89. The highest the neutral sentiment has gone is 1, while the lowest it has gone is 0.61. The median as well for the ChatGPT is 0.92. Also, the negative and positive sentiment is 0.03 and 0.08 respectively, which is very close to what Google sentiments are as well.

At last, when we compare the results of ChatGPT and Google Search with each other we can notice a few things. For instance, even though ChatGPT average is slightly lower than Google's for Neutral Sentiment it can be noticed that Google went as low as 0.47 while ChatGPT stayed at 0.61. Also, if you see Negative Sentiment as well, while the average for ChatGPT is slightly higher, its max is nearly half as low as compared to Google's Negative Sentiment. The same goes for the Positive Sentiment as well, showing that Google tends to have more extreme cases. From this we can conclude that on average both Google's and ChatGPT's results are neutral, however, in some cases, there are not, it's more likely for Google to have an extremely negative or positive result as compared to ChatGPT's results. Figures 14, 16, and 15 show the average, median, max, and min for each of the sentiment analysis.

Average	0.03
Median	0.00
Max	0.20
Min	0.00

(a) ChatGPT

Average	0.02
Median	0.00
Max	0.53
Min	0.00

(b) Google

Figure 14. Negative Sentiment

Average	0.89
Median	0.92
Max	1.00
Min	0.61

(a) ChatGPT

Average	0.91
Median	1
Max	1.00
Min	0.47

(b) Google

Figure 15. Neutral Sentiment

Average	0.08
Median	0.08
Max	0.29
Min	0.00

(a) ChatGPT

Average	0.07
Median	0.00
Max	0.45
Min	0.00

(b) Google

Figure 16. Positive Sentiment

4 Machine Learning Model

One of the tools which this project aims to develop is a tool that can tell if a result from a search query is relevant or not. This is a very ambitious goal, and to achieve that we thought it is wise to use Machine Learning. Along with the data which were collected from children, Professor Monica and her team were able to go manually through each result from Google Search to determine if it is relevant or not relevant. Hence, we thought it would be wise to use that data and train a machine learning module that can learn different patterns and teach itself to recognize if some result is relevant or not.

4.1 Creation of ML Model

The model is trained on Logistic Regression [8] which is best suited for this problem as it is usually used to solve binary classification problems. The goal of logistic regression is to predict a binary outcome such as yes or no and in our case it is relevant or not relevant. This generalized linear model uses a logistic function to model different relationships between the input variables and the output variables. Logistic Regression is trained to find the coefficients which maximize the likelihood of the observed data with the given mode.

Hence the approach for this was to first to read the data and split the data into target and feature. The feature in our case is the user query and the snippet columns where the results of the user query are. The target is the source column which would give a binary result; relevant or not relevant. Firstly, the text data in the features is pre-processed using CountVectorizer which would convert the text data into a matrix of different token counts. From there the data is split into training and testing sets using the train test split. The reason why the split is done is so that the performance of the Machine Learning Module can be evaluated. From there we use the logistic regression model to train itself on the trained data and then get evaluated against the test data.

4.2 Evaluation of ML Model

	precision	recall	f1-score	support
NR	0.81	0.98	0.89	132
R	0.83	0.25	0.38	40
accuracy			0.81	172
macro avg	0.82	0.62	0.64	172
weighted avg	0.82	0.81	0.77	172

Table 12. Classification Report

Table 12 shows the performance of the model. As can be seen, the precision of NR and R is 0.81 and 0.83 respectively which is quite high. However, the F1 score should be looked at more properly as it takes into account both precision

and recall. F1 score is the harmonic mean of precision and recall which ranges from 0 and 1. As can be seen for NR the F1 score is 0.89 however the F1 score for R is 0.38 which is quite low.

However, these results don't tell the full evaluation as the main problem here is that the data is simply too less to make a proper model. When comparing these results from results gather from ChatGPT out of 77 queries it said that only 1 is relevant and the rest is non-relevant. Hence this module is not practical, but it is a good starting point to develop when more data is collected.

5 Validation

5.1 Analysis of Results

In the end, after all the collection of data and evaluation has been done, we can finally look at what all the analysis has shown to us. What we have realized is that results from both Google Search and ChatGPT differ a lot from each, and this was proven by all the different Similarity Analyses which we ran on them. This claim is supported by the fact the length of their response varies a lot. The average character per response for Google is 151, while the average character per response for ChatGPT is 342, nearly 3 times more. With the main caveat that Google provides a SERP made of short snippets while ChaGPT gives just one cohesive answer.

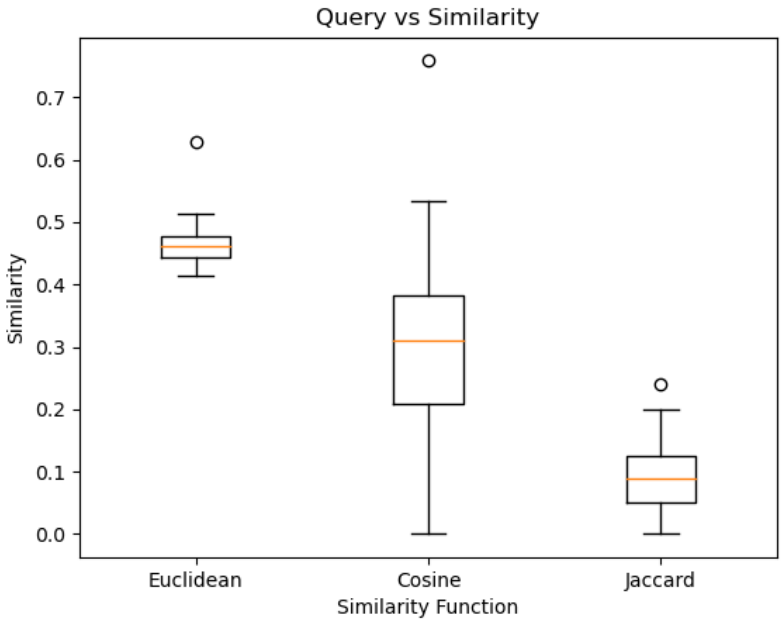


Figure 17. Similarity Analysis

In Figure 17, we can see a box graph visualizing Cosine, Jaccard, and Euclidean Similarity. This graph is helpful for us to understand the distribution of the different similarities as they have the same scale. It can also be noticed that there were some cases where in the extreme end both Euclidean and Cosine Similarity showed a high value. Euclidean Similarity had an average higher average as compared to Cosine and Jaccard Similarity. This is because Euclidean Similarity is more sensitive to outliers as compared to Cosine and Jaccard Similarity. On the other hand, Jaccard Similarity had the lowest average as it is more sensitive to the length of the response. Cosine Similarity is the most spread out among the three.

However, having more response does not mean that what ChatGPT provide is richer with its response. Instead, when we analyze the Named Entity Recognition (NER) of both Google and ChatGPT we can have a better understanding of the richness of their responses. From the results of NER Analysis, we can see that the top 5 entities for both Google Search and ChatGPT and nearly the same, however where they are different is the number of times they occur. The average number of entities per response is 18.2 for Google, while for ChatGPT it is 6.11. This means that the quality of information that Google provides is much richer with important facts and information as compared to ChatGPT. Hence despite having a much longer response as compared to Google, ChatGPT's responses lack the variation and amount of entities in their response.

Another interesting thing that we found out was that the sentiment of the responses from both Google and

ChatGPT are quite similar. The average sentiment of Google is 0.07, while the average sentiment of ChatGPT is 0.08. This looks to show that the responses from both Google and ChatGPT are quite neutral. However, when we look at the distribution of the sentiment of the responses from both Google and ChatGPT we can see that the distribution of the sentiment of the responses from Google is more spread out as compared to ChatGPT. This means that the responses from Google are more diverse as compared to ChatGPT. Hence, when it comes to the cases in which the response for Google is not neutral, their response could either have more positive sentiment or negative sentiment as compared to ChatGPT. This means that in some cases there can be possibilities of some biased involved in the response.

At last, when we look at the performance of the Machine Learning Module which we developed, we can see that the performance of the model is quite good when evaluated against test data. The F1 score for the relevant results is 0.38 which is quite low, however, the F1 score for the non-relevant results is 0.89 which is quite high. This means that the model is quite good at predicting the non-relevant results, however, it is not good at predicting the relevant results. On the other when we use this model with some other data, we can again see a clear limitation since it predicts most of the responses as Relevant while once while it predicts non-Relevant. This is because the data which was used to train the model was not enough. Hence, the model is not practical, but it is a good starting point to develop when more data is collected. Also, it is not essential to use such type of data, but maybe some other data could be used which is finer tuned for ChatGPT as the previous one had responses from Google.

In the end, when we come back to the research question "**Are the new conversational AI models safer for children to interact with, or do the traditional search engine still hold an edge over the new technology?**", we can see that the answer is not as simple as it seems to be. However, we were still able to see enough differences to make a conclusion that currently Google Search is a better option as compared to ChatGPT for children. This is because Google Search provides more relevant and rich information as compared to ChatGPT, but this does not mean that ChatGPT has not performed well. Currently, ChatGPT can very well be used as a complementary tool to Google Search. With future improvements to the technology ChatGPT and other conversational AI, models can very well be used as a replacement for Google Search and be a better option for children to interact with. Conversational AI models have the ability to mimic human conversations and can be used as an interactive tool for children to learn in the future.

5.2 Limitations

Throughout the project, there were several limitations that came across. The first limitation was the data which was used in this project. Even though the data was not as old, it still can be considered outdated due to the fast-paced advancement of technology, especially among young users. Another factor about the data was its limited quantity, which had a huge impact on the machine learning model. Much more data was needed to have a much more robust machine learning model.

Another limitation of this project was that the scope of this project was a bachelor project. Such research-oriented project demands more resources and time. Along with that a lot of time was invested in learning how to use the tools so that proper results can be produced.

Statistic	**Average**	**Max**	**Min**	**Median**
Cosine Similarity	0.296	0.941	0	0.296
Euclidean Similarity	0.462	0.744	0	0.457
Jaccard Similarity	0.091	0.42	0	0.078
Pearson Correlation	-0.321	0.949	-0.955	-0.342
Google Query Length	151.44	162.2	132.8	153.8
ChatGPT Query Length	343.81	1321	81	237
Google NER	18.22	-	-	-
ChatGPT NER	6.12	-	-	-
ChatGPT Negative Sentiment	0.03	0.20	0.00	0.00
ChatGPT Neutral Sentiment	0.89	1.00	0.61	0.92
ChatGPT Positive Sentiment	0.08	0.29	0.00	0.08
Google Negative Sentiment	0.02	0.53	0.00	0.00
Google Neutral Sentiment	0.91	1.00	0.47	1.00
Google Positive Sentiment	0.07	0.45	0.00	0.00

Table 13. Summary of Results

In Table 13, we can see a summary of the results which we have discussed above. This table is helpful for us to understand the distribution of the different results as they have the same scale.

Finally, the last limitation was that the technology being analyzed, especially the conversational models is very fast-paced. A lot is being in development at such technology from both the technical perspective but also the rules and regulations perspective. This is a very new technology to even this day many new findings are being found. However, despite this project having a small scale it provides a fresh perspective to such a field. This project aims to be a stepping stone for another project to launch itself. This project provides tools that facilitate analysis and comparison between traditional search engines and bleeding-edge conversational models.

5.3 Future Work

There are several ways that this project can be extended. For instance, with more data, the whole project can become much more robust, especially the machine learning module to detect relevant and non-relevant results. Another way in which this project can be extended is by using the tools developed for other projects related to this field. The projects can be such as analyzing the performance of different conversational models, or analyzing the performance of different search engines and comparing them to the current standards. The field of conversational models has brought us a new and exciting way of absorbing information; however, such fast-paced technology always has threats to it. With the tools developed throughout the duration of this project, they can be used to analyze such technologies. The tools are all based on the latest technologies with well-written documentation to support them. It is designed to be adaptable to different projects and goals. Involving real users in the study would make it more effective and complex at the same time.

Acknowledgements

I would like to thank Prof. Dr. Landoni Monica for allowing me to work on this project and for providing me with her guidance and support throughout the duration of this project. I would also like to thank family and friends for their support and encouragement throughout the duration of this project. Finally, I would like to thank the Università della Svizzera Italiana for providing me with the opportunity to work on this project.

References

[1] A. Arampatzis and J. Kamps. A study of query length. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 811–812, 2008.

[2] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

[3] N. Dehak, R. Dehak, J. R. Glass, D. A. Reynolds, P. Kenny, et al. Cosine similarity scoring without score normalization techniques. In *Odyssey*, page 15, 2010.

[4] K. L. Elmore and M. B. Richman. Euclidean distance as a similarity metric for principal component analysis. *Monthly weather review*, 129(3):540–549, 2001.

[5] M. Firat. How chat gpt can transform autodidactic experiences and open education. *Department of Distance Education, Open Education Faculty, Anadolu Unive*, 2023.

[6] K. Hillis, M. Petit, and K. Jarrett. *Google and the Culture of Search*. Routledge, 2012.

[7] N. V. Kalinina, V. V. Zaretskiy, V. B. Salakhova, E. G. Artamonova, O. I. Efimova, and E. E. Lekareva. Psychological and pedagogical resources of security provision and prevention of internet risks and life threats among children and teenagers in the educational environment. *Modern Journal of Language Teaching Methods*, 8(8):118–129, 2018.

[8] M. P. LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008.

[9] C. K. Lo. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410, 2023.

[10] M. F. Lovenheim and P. Walsh. Does choice increase information? evidence from online school search behavior. *Economics of Education Review*, 62:91–103, 2018.

[11] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[12] P. Meel and D. K. Vishwakarma. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986, 2020.

[13] A. Mikheev, M. Moens, and C. Grover. Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8, 1999.

[14] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384, 2013.

[15] Y. Pan, L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Y. Wang. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*, 2023.

[16] R. Pandita. Information pollution, a mounting threat: internet a major causality. *Journal of Information Science Theory and Practice*, 2(4):49–60, 2014.

[17] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston. *Retrieval Augmentation Reduces Hallucination in Conversation*. Association for Computational Linguistics, 2021.

[18] D. J. Slone. Internet search approaches: The influence of age, search goals, and experience. *Library & Information Science Research*, 25(4):403–418, 2003.