Quantifying Similarities: Oncology Documents from Google Bard and ChatGPT

Muhammad Shumail Naveed[1]
[1]Department of Computer Science & Information Technology, University of Baluchistan, Quetta, Pakistan
*Correspondence: Muhammad Shumail Naveed; mshumailn@gmail.com

Citation | Naveed. M. S, "Quantifying Similarities: Oncology Documents from Google Bard and ChatGPT", IJIST, Vol. 5 Issue. 4 pp 773-786, Dec 2023
Received | Dec 04, 2023, **Revised |** Dec 08, 2023, **Accepted |** Dec 22, 2023, **Published |** Dec 28, 2023.

Large language models hold immense promise for the future of text generation. Google Bard and ChatGPT, two prominent large language models originating from different research laboratories, have been subjects of various studies since their introduction. Despite numerous perspectives explored in the studies, none has specifically delved into the analysis of the similarity between texts generated by these models within the same category. This study addresses this gap by comparing the document generation capabilities of Google Bard and ChatGPT. The analysis focuses on topic-wise comparable documents related to oncology. In this study, 50 oncology-related documents generated by Google Bard are juxtaposed with equivalent topic-wise documents produced by ChatGPT, utilizing both cosine similarity and Jaccard similarity for comparison. The analysis employed statistical tests including the Kolmogorov-Smirnov test, Shapiro-Wilk test, and the one-sample Wilcoxon signed-rank test. The findings revealed a significant level of resemblance among the documents generated by both models: cosine similarity (mean = 0.66, std. dev. = 0.11, min = 0.23, max = 0.80) and Jaccard similarity (mean = 0.88, std. dev. = 0.06, min = 0.7, max = 1.0). This suggests a probable commonality in their training datasets or sources of oncology-related information. The study also posited that the observed similarity could be attributed to the probabilistic nature of language models and the potential for overfitting during their training processes. This study stands out for offering a unique direction and outcomes that pave the way for further exploration in the domain of large language models.

Keywords: Large Language Models; ChatGPT; Google Bard; Cosine Similarity; Jaccard Similarity.

Acknowledgment
The author expresses gratitude to Muhammad Tahaam for offering substantial support in topic selection and text collection for the study. Special thanks are extended to Muhammad Aayaan for helping with data analysis and the write-up.

Author's Contribution
The complete article is written by a single author.

Conflict of Interest
No conflict of Interest.

Project details: Nill.

Introduction:

Artificial Intelligence (AI) is a broad term that denotes the utilization of computers to emulate intelligent behavior with minimal human intervention [1]. Various domains exist within artificial intelligence, and one notable field is natural language processing.

Natural Language Processing (NLP) encompasses techniques for handling human language, utilizing a combination of semantic and statistical approaches. It involves the understanding of both spoken and written language, encompassing technologies related to voice recognition and text processing. Research fields within natural language processing include the identification of named entities, document filtering, and classification, automatic abstract generation, information extraction, voice recognition, sentiment analysis, opinion mining, monitoring social media reputation, orthographic and grammatical correction, text-to-voice systems, intelligent and optimized search, automatic response systems, personal assistants, automatic translation, and dialogue systems [2].

The swift advancements in natural language processing (NLP) have given rise to large language models capable of generating text at a level comparable to human quality and engaging in seamless, natural conversations. These models hold immense potential for diverse applications, including customer service interactions, the generation of creative content, and the facilitation of personal assistance.

Large language models (LLMs) represent artificial intelligence (AI) models grounded in deep learning, specifically neural networks, designed for text generation [3][4]. These models possess intricate underlying architectures and an extensive array of parameters, honed through training on vast volumes of existing documents. In contrast to older natural language processing approaches that rely on supervised learning for particular tasks, the majority of LLMs adopt semi-supervised approaches. Large language models [5] can function as an initial method for acquiring exploratory insights into a particular subject.

Large language models are constructed on the Transformer, a state-of-the-art neural network architecture with numerous parameters. The principal innovation of the Transformer lies in its self-attention mechanisms, allowing the model to better comprehend the relationships between different elements of the input. Large language models utilize a two-stage training pipeline for efficient data learning. In the initial pretraining stage, these models employ a self-supervised learning approach, enabling them to learn from extensive amounts of unannotated data without the need for manual annotation. This capability provides a significant advantage over traditional fully supervised deep learning models, as it eliminates the requirement for extensive manual annotation and enhances scalability. During the subsequent fine-tuning stage, large language models are trained on small, task-specific, annotated datasets to leverage the knowledge acquired during the pretraining stage. This allows them to perform specific tasks as intended by end users. Consequently, large language models achieve high accuracy on various tasks with minimal human-provided labels.

Large language models find application in diverse analytical steps throughout the development of new diagnostic assays. They prove valuable in tasks such as formulating growth media recipes for rare pathogens, troubleshooting unsuccessful polymerase chain reaction (PCR) assays, crafting PCR primers, and generating programming code. Additionally, these models play a crucial role in aiding the interpretation of data, particularly in scenarios involving rarely encountered pathogens [6].

The emergence of large language models (LLMs) has marked a significant turning point in recent years, with programs like ChatGPT and Google Bard making their debut in the public domain. The absence of financial barriers to entry for the average user, coupled with remarkable ease of use, has unveiled the vast potential of artificial intelligence to permeate various aspects of our lives [7]. Google Bard and ChatGPT have emerged as groundbreaking interactive

chatbots [8], and their notable studies have been conducted from diverse perspectives since their inception.

ChatGPT, an artificial intelligence conversational system developed by OpenAI, a company dedicated to AI research and deployment, operates on the GPT-3.5 [9], one of the largest language models (LLMs) with over 175 billion parameters, making it among the most advanced to date. Sharing numerous capabilities with its predecessor, ChatGPT was trained on a diverse corpus of internet texts, encompassing approximately 570 GB of content from books, articles, and websites, spanning various subjects such as news, wikis, and fiction. Specifically tailored for conversational tasks, ChatGPT underwent fine-tuning using reinforcement learning from human feedback. Through this methodology, ChatGPT dynamically adjusts its behavior, making it highly adept at comprehending user intentions, generating text that closely mimics human language, and maintaining coherence throughout a conversation.

Google Bard AI is a text-based chatbot powered by natural language processing (NLP) and machine learning (ML) that generates real-time answers to questions. It is built upon the Pathways Language Model 2 (PaLM 2), a language model that is part of the lineage of Google's Language Model for Dialogue Applications (LaMDA) technology [10].

Google Bard and ChatGPT have been compared and analyzed from different perspectives. However, no comprehensive study has delved into the similarity of text generated by these two large language models. This area of study holds substantial importance, as the precise details of the data used to train these models are not publicly available. Therefore, a study that examines the similarity of comparable text generated by these models would provide valuable insights into their capabilities. The primary objective of this study is to analyze the topic-wise similarity of text generated by these models. To achieve this goal, the study will focus on oncology, a medical subspecialty dedicated to the investigation, diagnosis, and treatment of individuals with cancer or suspected cancer.

The novelty of this study lies in its capacity to provide valuable insights and enhance our understanding of the capabilities and nuances of two popular language models. Consequently, it may also unveil potential biases embedded within these models.

Literature Review:

Large language models represent the forefront of current studies and research, prompting numerous investigations into their functionalities and applications. Cheong et al. [11] conducted an evaluation of patient education materials on obstructive sleep apnoea generated by ChatGPT and Google Bard. The research involved the extraction of fifty frequently asked questions in English from the patient information webpages of four major sleep organizations, categorizing them as input prompts. Responses from ChatGPT and Google Bard underwent independent rating by two otolaryngologists with a Fellowship of the Royal College of Surgeons and a specialized interest in sleep medicine and surgery, utilizing the Patient Education Materials Assessment Tool–Printable Auto-Scoring Form. As a secondary outcome, responses were subjectively screened for any incorrect or potentially harmful information. The Flesch-Kincaid Calculator was employed to assess the readability of responses from both ChatGPT and Google Bard. The study's findings indicated that ChatGPT offers superior patient education materials for obstructive sleep apnoea compared to Google Bard.

An electrocardiogram (ECG) is a crucial medical test for diagnosing heart conditions, though interpreting its results can pose a challenge. In a study conducted by Fijačko et al. [12], the accuracy of Bing Chat Enterprise, ChatGPT-4 Pro, and Google Bard in interpreting ECG images from the American Heart Association (AHA) Advanced Cardiovascular Life Support (ACLS) multiple-choice question exams were examined. Each chatbot provided three separate interpretations of the same ECG image with identical prompts, and these interpretations were then compared to the AHA ACLS exam answers. In addition to assessing overall accuracy, the

chatbots were prompted to estimate their "Level Of Correctness" for each interpretation of the ECG image. Throughout the study, a total of 81 interpretations were conducted using 9 different ECG images. The results revealed that ChatGPT-4 Plus emerged as the most successful chatbot, accurately interpreting nearly two-thirds of the ECG images. Following closely was Google Bard, correctly interpreting ECG images almost half of the time, while Bing Chat Enterprise demonstrated accurate interpretations less than a quarter of the time.

In a notable study by Patil et al. [13], the accuracy, response length, and response time of the latest versions of ChatGPT (ChatGPT-4) and Bard were compared in their ability to answer radiology board examination practice questions. Text-based questions from the 2017-2021 American College of Radiology's Diagnostic Radiology In-Training (DXIT) examinations were used to evaluate the two models. Analyzing 318 questions, the study found that ChatGPT responded significantly more accurately than Bard, but with shorter responses and longer response times. ChatGPT also demonstrated superior performance in neuroradiology, general & physics, nuclear medicine, pediatric radiology, and ultrasound. Overall, the study concluded that ChatGPT displayed superior radiology knowledge compared to Bard. However, both chatbots revealed limitations and fallibility, providing incorrect or illogical explanations and sometimes failing to address the educational content of the questions. This highlights the need to use such models with caution and awareness of their limitations.

Al-Ashwal et al. [14], conducted a study to examine the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard were evaluated regarding their predictive capabilities for drug-drug interactions (DDIs). The assessment involved comparing their abilities to detect clinically relevant DDIs for 255 drug pairs. The results revealed that Bing AI demonstrated the highest accuracy and specificity, surpassing the performance of Google's Bard, ChatGPT-3.5, and ChatGPT-4. These findings underscore the substantial potential of these AI tools in reshaping patient care. Although the current AI platforms analyzed are not exempt from limitations, their capacity to swiftly analyze potentially significant interactions with good sensitivity suggests a promising advancement toward enhanced patient safety.

Gan et al. [15], conducted a comparative study to assess the performance of the Google Bard and medical students in mass casualty incident (MCI) triage, utilizing the Simple Triage and Rapid Treatment (START) method. The study employed a validated questionnaire featuring 15 diverse MCI scenarios to evaluate triage accuracy through content analysis. The results revealed that Google Bard exhibited significantly higher accuracy at 60%, whereas ChatGPT achieved a lower accuracy of 26.67%. In comparison, medical students demonstrated an accuracy rate of 64.3% in a previous study. However, no significant difference was observed between the performance of Google Bard and medical students. The overall results indicated that Google Bard outperformed ChatGPT.

Ali et al. [16], evaluated the performance of GPT-3.5, GPT-4, and Google Bard using a question bank designed specifically for neurosurgery oral board examination preparation. The study utilized the 149-question Self-Assessment Neurosurgery Examination, which employed a single best-answer, multiple-choice format to assess LLM accuracy. Fisher exact and univariable logistic regression tests were employed to examine differences in performance based on question characteristics. The results revealed that GPT-4 outperformed ChatGPT and Google Bard, achieving a score of 82.6%.

Dhanvijay et al. [17] conducted a comprehensive cross-sectional study to assess the effectiveness of three large language models (LLMs): ChatGPT 3.5, Google Bard, and Microsoft Bing, in providing responses to case vignettes in Physiology. A total of seventy-seven case vignettes in physiology were meticulously curated by two physiologists and validated by two additional content experts. Subsequently, each LLM was presented with these cases, and their respective responses were systematically gathered. Two physiologists independently scrutinized

the accuracy of the answers furnished by the LLMs. Ratings were assigned on a scale ranging from 0 to 4, aligning with the structure of the observed learning outcome (pre-structural = 0, uni-structural = 1, multi-structural = 2, relational = 3, extended-abstract). The scores across the various LLMs were subjected to comparison through Friedman's test, and interobserver agreement was established using the intraclass correlation coefficient. The findings of the study revealed that ChatGPT 3.5 achieved the highest score, Bing exhibited the lowest score, and Bard occupied an intermediary position between the two in terms of performance. Consequently, ChatGPT demonstrated superior performance compared to both Bard and Bing in generating responses to case vignettes in the field of physiology.

O'Leary [18], conducted a study using questions from the Watson Jeopardy! The challenge is to compare the performance of BARD, ChatGPT, and Watson. Through the utilization of these Jeopardy! questions, the analysis reveals that, for high-confidence Watson questions, all three systems demonstrate similar accuracy to Watson. Additionally, both BARD and ChatGPT exhibit accuracy comparable to that of a human expert, and their sets of correct answers display a high degree of similarity, as indicated by a Tanimoto similarity score. The study also observed that both BARD and ChatGPT have the capability to modify their solutions when presented with the same input information on subsequent occasions. In instances where the same Jeopardy! category and question are repeated, both systems may generate distinct and conflicting answers.

Plevris et al. [19] conducted an assessment of the capabilities of ChatGPT-3.5, ChatGPT-4, and Google Bard in tackling mathematical and logical problems. The study employed a set of 30 questions, divided into two categories of 15 questions each. The first set comprised problems not readily available online, while the second set included problems commonly found online, often accompanied by solutions. Each question was presented to each chatbot three times, and their responses were recorded and analyzed. The findings revealed that chatbots demonstrate accuracy in providing solutions for straightforward arithmetic, algebraic expressions, and basic logic puzzles. However, this accuracy is not consistent across all attempts. For more intricate mathematical problems or advanced logic tasks, the chatbots' answers, while appearing convincing, may lack reliability. Moreover, there is a notable inconsistency issue, as chatbots often yield conflicting answers when presented with the same question multiple times. The results indicated that ChatGPT-4 outperforms ChatGPT-3.5 in both sets of questions. Google Bard ranks third in the original questions of the first set, trailing behind the other two chatbots. However, Bard exhibits the best performance, securing first place in the published questions of the second set. This discrepancy is likely attributed to Bard's direct internet access, unlike the ChatGPT chatbots, which, due to their designs, lack external communication capabilities.

Koga et al. [20], conducted an analysis and comparison of the predictive performance between ChatGPT and Google Bard in determining neuropathologic diagnoses based on clinical summaries. The study involved the examination of 25 cases of neurodegenerative disorders presented at the Mayo Clinic Brain Bank Clinico-Pathological Conferences. The models generated multiple pathologic diagnoses along with their respective rationales, which were then juxtaposed against the conclusive clinical diagnoses provided by physicians. In specific figures, ChatGPT-3.5, ChatGPT-4, and Google Bard accurately identified primary diagnoses in 32%, 52%, and 40% of cases, respectively. Furthermore, correct diagnoses were encompassed within the generated results for 76%, 84%, and 76% of cases, respectively. These results shed light on the varying predictive capabilities of the models and their alignment with the final clinical assessments.

Al-Ashwa et al. [14] evaluated the predictive performance of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard in forecasting drug-drug interactions, considering sensitivity, specificity,

and accuracy. The study involved a comparison of these large language models in the detection of clinically relevant DDIs across 255 drug pairs. Notably, Bing AI emerged with the highest accuracy and specificity, surpassing the performance of Google's Bard, ChatGPT-3.5, and ChatGPT-4. These findings underscore the considerable potential of these AI tools in revolutionizing patient care.

Seth et al. [21] conducted a comprehensive assessment to evaluate the efficacy of Google BARD, Bing AI, and ChatGPT-3.5 in delivering accurate and secure medical information related to rhinoplasty. The study involved presenting six specific questions about rhinoplasty to ChatGPT, BARD, and Bing AI. An expert panel comprising Specialist Plastic and Reconstructive Surgeons, possessing extensive experience in rhinoplasty, employed a Likert scale to assess the responses. Reliability was measured using the Flesch Reading Ease Score, Flesch–Kincaid Grade Level, and Coleman–Liau Index. The modified DISCERN score served as the criterion for evaluating suitability and reliability. The study identified that, in terms of reliability, both BARD and ChatGPT exhibited significantly higher reliability compared to Bing AI. Regarding suitability, BARD achieved a significantly higher DISCERN score than both ChatGPT and Bing AI. In terms of the Likert score, ChatGPT and BARD demonstrated similar scores, both surpassing those of Bing AI. In conclusion, the study determined that BARD provided the most concise and comprehensible information, followed by ChatGPT and Bing AI.

Material and Methods:

The primary objective of the study is to compare and examine Google Bard and ChatGPT by assessing the similarity of topic-wise compared documents generated by these models. To achieve this, the study followed the research methodology outlined in Figure 1.

The study commenced with the exploration and selection of oncology-related topics. Initially, 73 topics were identified, and, following consultation with two medical practitioners, 50 were chosen for the study. Using Google Bard and ChatGPT, text generation occurred from November 12, 2023, to November 14, 2023, based on the selected topics. The texts produced by these models for each topic were nearly identical in size, typically ranging from 380 to 400 characters. The selected oncology topics for study are listed in Appendix A.

Before assessing the similarity between the topic-wise comparable texts, thorough preprocessing was conducted on the collected data. This involved converting all text to lowercase and eliminating accented characters using the Uni-decode package in Python. Additionally, regular expressions were employed to remove non-ASCII characters from the data. To further reduce the dimensionality of the data, contractions were expanded, and punctuations were removed. The data underwent additional processing steps, including part-of-speech tagging, lemmatization, and the elimination of stop words, facilitated by spaCy. SpaCy is an open-source leading Python software library for advanced natural language processing [22][23] implemented in Python and Cython.

In evaluating the similarity between comparable documents, the study utilized both cosine similarity and Jaccard similarity. Typically, cosine similarity values are employed to assess the level of similarity between two sets of elements [24]. Cosine similarity serves as a widely adopted metric [25][26], for assessing text similarity. Essentially, it compares two non-zero vectors within an inner product space. This measure is solely reliant on the angle between the vectors rather than their magnitudes. The standard formula for computing cosine similarity is expressed as follows:

$$\text{Cosine Similarity (A, B)} = \frac{A.B}{||A||\,.\,||B||}$$

The Jaccard similarity, an algorithm based on words, is widely employed for its efficiency and widespread use across various applications. The Jaccard similarity algorithm, known for its efficiency and widespread use across diverse applications, relies on word-based comparisons

[27][28]. This method assesses the similarity between two sets of key phrases by scrutinizing the uniqueness and commonality of the data. Consequently, it compares all conceivable pairs of sets to ascertain their similarity [29][30]. The formula for Jaccard similarity is as follows:

$$\text{Jaccard Similarity (A, B)} = \frac{|A \cap B|}{|A \cup B|}$$

The findings from the similarity analysis undergo subsequent statistical analysis using SPSS (Ver: 25), and the resultant insights are further visualized using R (Ver: 4.2.3).

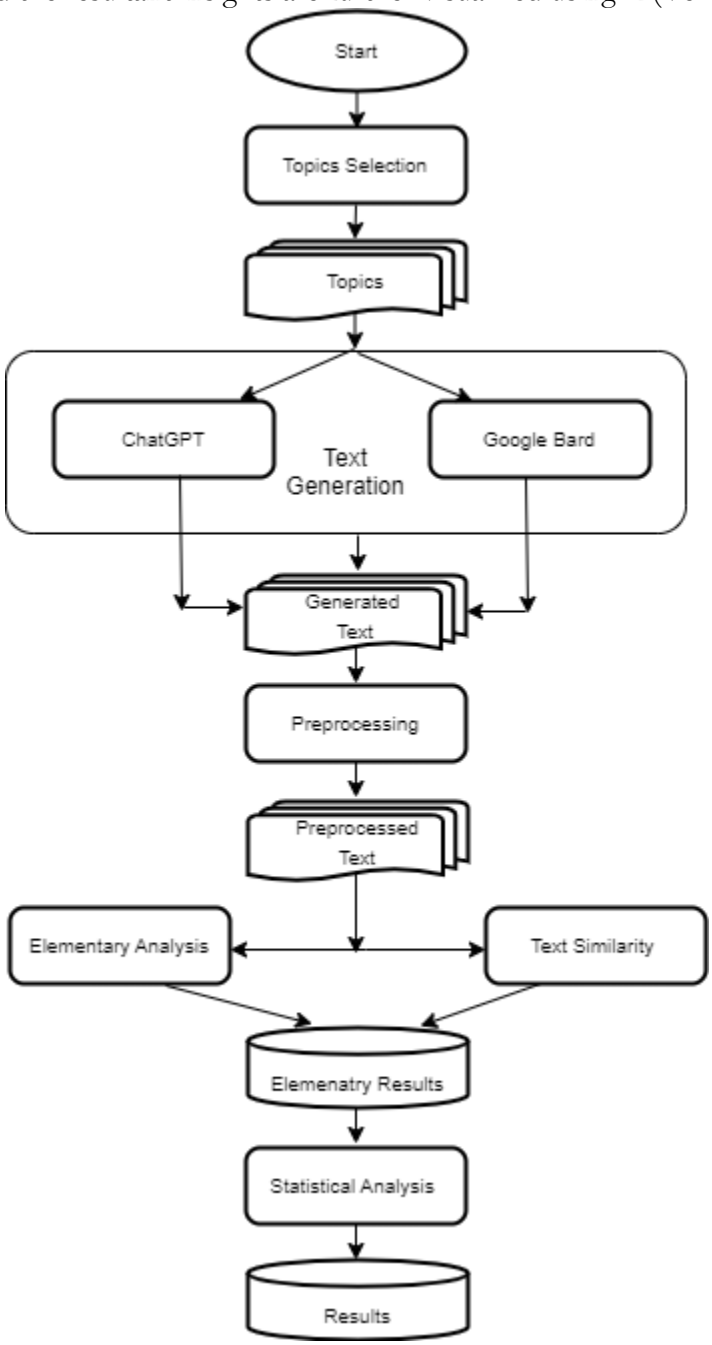

Figure 1. Research Methodology

Results:

Google Bard and ChatGPT are two prominent large language models trained on extensive datasets of text and code. This study examines the similarity between 50 documents related to oncology generated by Google Bard and 50 corresponding texts generated by ChatGPT. The analysis is carried out in two distinct stages. During the first stage, elementary

analysis is performed, while the actual similarity analysis takes place in the second stage. The outcomes obtained from the elementary analysis are presented in Table 1.

Table 1. Results of Elementary Analysis

Descriptive Statistics	ChatGPT			Google Bard		
	Words	Difficult Words	Total Sentences	Words	Difficult Words	Total Sentences
Minimum	382	92	14	380	43	12
Maximum	496	159	25	495	136	28
Mean	394.80	119.40	20.76	394.60	100.68	19.16
Median	392.00	119.00	21.00	393.00	102.00	18.00
Std. Deviation	15.35	12.92	2.38	15.58	23.79	3.53

The documents or texts within each pair were meticulously matched in size. Findings from the elementary analysis demonstrated a significant likeness in the number of words, difficult word usage, and total sentences across both document groups. This implies a notable resemblance in the texts generated by Bard and ChatGPT in the domain of oncology, particularly concerning word count, difficult words, and total sentences. To vividly present the results of the elementary analysis, line charts have been employed, as depicted in Figure 2.

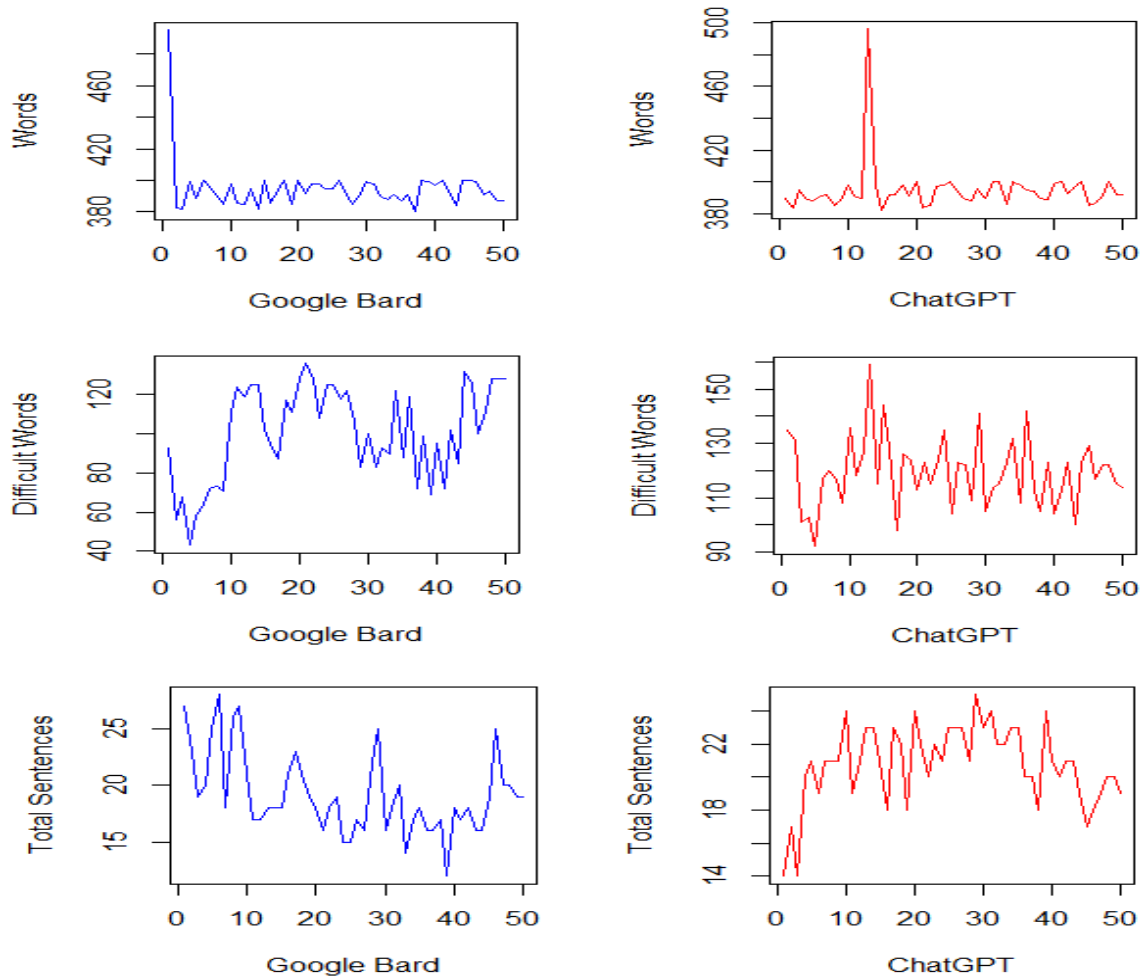

Figure 2. Line Chats of Elements in Generated Documents

The line charts clearly illustrate the results of elementary analysis. These findings, acquired through the elementary analysis, are then employed as features to evaluate the similarity between the text generated by Bard and ChatGPT. The results of this similarity analysis are presented in Table 2.

Table 2. Result of Similarity Analysis

Metric	Mean	Median	Std. Dev.	Range	Skewness	Kurtosis
Cosine Similarity	0.66	0.69	0.11	0.57	-1.99	5.66
Jaccard Similarity	0.88	0.89	0.06	0.3	-0.84	0.99

The analysis of the text from Google Bard and ChatGPT using cosine similarity suggests a reasonably strong alignment and resemblance among the compared oncology-related documents. However, it does not indicate perfect similarity. The analysis of analyzed documents performed with Jaccard Similarity identified a substantial level of similarity between the documents. Similarly, the analysis of the examined documents using Jaccard Similarity revealed a considerable degree of similarity among the documents. For a detailed illustration of the results from similarity analyses, the extreme values are itemized and presented in Table 3.

Table 3. Result of Extreme Values of Similarity Analysis Results

Cosine Similarity				Jaccard Similarity			
Case No.	**Highest**	**Case No.**	**Lowest**	**Case No.**	**Highest**	**Case No.**	**Lowest**
44	0.80	50	0.23	28	1.00	40	0.70
27	0.79	25	0.29	9	0.97	30	0.74
45	0.79	39	0.49	13	0.97	29	0.78
17	0.77	46	0.52	24	0.97	19	0.78
3	0.76	36	0.56	8	0.94	41	0.8

The results of the similarity analysis are depicted in scatter plots, as shown in Figure 3.

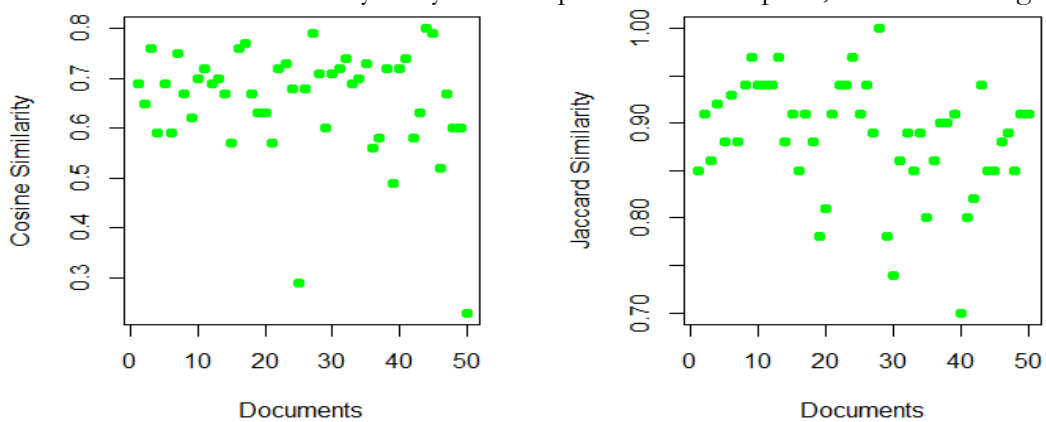

Figure 3. Scatter Plots of Similarity Analysis

To assess the normality of cosine similarity and Jaccard similarity measures, the Kolmogorov-Smirnov and Shapiro-Wilk tests were conducted. The Kolmogorov-Smirnov test revealed a test statistic of 0.170 and a p-value of 0.001 for cosine similarity. As the p-value is less than the significance level of 0.05, it can be concluded that the distribution of cosine similarity does not adhere to a normal distribution. This implies that cosine similarity values are not normally distributed. Similarly, the Shapiro-Wilk test yielded a test statistic of 0.829 and a p-value of 0.00 for cosine similarity, further supporting the non-normality of its distribution.

The Kolmogorov-Smirnov test revealed a test statistic of 0.130 and a p-value of 0.035 for Jaccard similarity, signifying a departure from normal distribution. This suggests that Jaccard similarity values deviate from a normal distribution. Similarly, the Shapiro-Wilk test, with a test statistic of 0.949 and a p-value of 0.030 for Jaccard similarity, supports the conclusion of its non-normal distribution. For further statistical analysis, a one-sample Wilcoxon signed-rank test, a non-parametric statistical method, is employed to assess the results of the similarity analysis.

The one-sample Wilcoxon signed-rank test was performed on cosine similarity, revealing a mean of 0.88 and a standard deviation of 0.060, with a corresponding p-value of 0.35. Consequently, the results were deemed statistically significant. Likewise, the one-sample

Wilcoxon signed-rank test was applied to Jaccard similarity, yielding a mean of 0.66 and a standard deviation of 0.109, along with a corresponding p-value of 0.01. Thus, the results were also determined to be statistically significant.

Discussion:

Bard and ChatGPT are two significant large language models, both trained on a vast amount of data and code. This article compares Google Bard and ChatGPT by analyzing topic-wise comparable documents on oncology generated with these models. During the study, 50 documents were generated by Google Bard, and the same topic-wise documents were generated with ChatGPT. On average, Google Bard generated documents with 394.60 words, of which 100.68 were identified as difficult. In contrast, documents generated with ChatGPT had a mean value of 394.80 words, with 119.40 identified as difficult. The 17.01% difference in difficult words suggests that Google Bard.

An examination of documents related to oncology generated by Google Bard and ChatGPT unveiled a notable degree of similarity. The analysis of cosine similarity yielded a score of 0.66, indicating a robust alignment between the documents. Further exploration using Jaccard similarity resulted in an even higher score of 0.88, signifying a substantial level of resemblance. These findings imply that, despite being developed by distinct research teams, both models were likely trained on analogous datasets or sources of oncology-related information, such as medical literature, research papers, or publicly available data.

The observed similarity can be ascribed to the probabilistic nature of language models, wherein predictions of the next word or sequence are made based on patterns learned during training. Without specific fine-tuning for uniqueness or novelty, these models have a tendency to generate similar content contingent on the provided input and context.

Furthermore, the architecture and design of language models may also play a role in the similarities observed in the generated content. Models sharing akin architectures or training methodologies are predisposed to producing similar outputs. Moreover, overfitting to specific patterns or information during training may contribute to similarities that do not necessarily reflect the broader diversity of oncology-related content. In its entirety, the study reached the conclusion that, with regard to the analyzed pages on oncology, Google Bard and ChatGPT exhibit a notable degree of similarity in terms of document generation.

Google Bard and ChatGPT have been analyzed from distinct perspectives in this study, representing a novel approach as it quantitatively analyzes and compares documents, an aspect that has not been studied before. The findings indicate that, despite being distinct language models, the content they generate on oncology exhibits noteworthy similarities. This observed similarity could be ascribed to the probabilistic nature of language models and the potential for overfitting during their training processes. Consequently, these results implicitly validate the originality and novelty of this work.

However, this work is subject to several limitations and threats to validity: i) The study is based on a very small sample of analyzed documents, ii) Only two methods were employed to assess the similarity of documents, iii) The comparison of the two large language models focuses exclusively on oncology documents, iv) The reliance on similarity scores may naturally overlook qualitative differences in generated content, such as variations in writing style and depth of analysis, and v) The suggestion of potential overfitting lacks concrete evidence or a detailed analysis of model training dynamics, introducing uncertainty to the claim and necessitating further investigation. As a part of further study, a more detailed analysis will be performed by considering a large sample of documents on diverse topics with more methods for text similarity.

Conclusion:

Large language models have seen a remarkable surge in popularity in recent years, attributed to their exceptional ability to emulate human-like conversation and generate text. The

Google Bard and ChatGPT stand out as two prominent large language models, subject to analysis and comparison from diverse perspectives. This study compared the document generation capabilities of Google Bard and ChatGPT, two large language models, by analyzing topic-wise comparable documents on oncology. The findings revealed a notable degree of similarity between the documents generated by both models, suggesting that they were likely trained on analogous datasets or sources of oncology-related information. This similarity can be attributed to the probabilistic nature of language models and the potential for overfitting during training.

References:

[1] P. Hamet, and J. Trembla, "Artificial Intelligence in Medicine | Journal | ScienceDirect.com by Elsevier." Accessed: Dec. 16, 2023. [Online]. Available: https://www.sciencedirect.com/journal/artificial-intelligence-in-medicine

[2] R. E. Lopez-Martinez and G. Sierra, "Research trends in the international literature on natural language processing, 2000-2019 - A bibliometric study," J. Scientometr. Res., vol. 9, no. 3, pp. 310–318, Sep. 2020, doi: 10.5530/JSCIRES.9.3.38.

[3] Y. Shen et al., "ChatGPT and Other Large Language Models Are Double-edged Swords," Radiology, vol. 307, no. 2, p. 2023, Apr. 2023, doi: 10.1148/RADIOL.230163/ASSET/IMAGES/LARGE/RADIOL.230163.FIG1.JPEG.

[4] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland, "Risks and Benefits of Large Language Models for the Environment," Environ. Sci. Technol., vol. 57, no. 9, pp. 3464–3466, Mar. 2023, doi: 10.1021/ACS.EST.3C01106/ASSET/IMAGES/LARGE/ES3C01106_0004.JPEG.

[5] A. Bryant et al., "Qualitative Research Methods for Large Language Models: Conducting Semi-Structured Interviews with ChatGPT and BARD on Computer Science Education," Informatics 2023, Vol. 10, Page 78, vol. 10, no. 4, p. 78, Oct. 2023, doi: 10.3390/INFORMATICS10040078.

[6] A. Egli and A. Egli, "ChatGPT, GPT-4, and Other Large Language Models: The Next Revolution for Clinical Microbiology?," Clin. Infect. Dis., vol. 77, no. 9, pp. 1322–1328, Nov. 2023, doi: 10.1093/CID/CIAD407.

[7] C. Tan Yip Ming et al., "The Potential Role of Large Language Models in Uveitis Care: Perspectives After ChatGPT and Bard Launch," Ocul. Immunol. Inflamm., Aug. 2023, doi: 10.1080/09273948.2023.2242462.

[8] S. Thapa and S. Adhikari, "ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls," Ann. Biomed. Eng., vol. 51, no. 12, pp. 2647–2651, Dec. 2023, doi: 10.1007/S10439-023-03284-0/METRICS.

[9] T. B. Brown et al., "Language Models are Few-Shot Learners," Adv. Neural Inf. Process. Syst., vol. 2020-December, May 2020, Accessed: Dec. 16, 2023. [Online]. Available: https://arxiv.org/abs/2005.14165v4

[10] Ö. AYDIN, "Google Bard Generated Literature Review: Metaverse," J. AI, vol. 7, no. 1, pp. 1–14, Dec. 2023, doi: 10.61969/JAI.1311271.

[11] R. C. T. Cheong et al., "Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard," Eur. Arch. Oto-Rhino-Laryngology, pp. 1–9, Nov. 2023, doi: 10.1007/S00405-023-08319-9/METRICS.

[12] N. Fijačko, G. Prosen, B. S. Abella, S. Metličar, and G. Štiglic, "Can novel multimodal chatbots such as Bing Chat Enterprise, ChatGPT-4 Pro, and Google Bard correctly interpret electrocardiogram images?," Resuscitation, vol. 193, p. 110009, Dec. 2023, doi: 10.1016/j.resuscitation.2023.110009.

[13] N. S. Patil, R. S. Huang, C. B. van der Pol, and N. Larocque, "Comparative Performance

of ChatGPT and Bard in a Text-Based Radiology Knowledge Assessment," Can. Assoc. Radiol. J., Aug. 2023, doi: 10.1177/08465371231193716/ASSET/IMAGES/LARGE/10.1177_08465371231193716-FIG1.JPEG.

[14] F. Y. Al-Ashwal, M. Zawiah, L. Gharaibeh, R. Abu-Farha, and A. N. Bitar, "Evaluating the Sensitivity, Specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard Against Conventional Drug-Drug Interactions Clinical Tools," Drug. Healthc. Patient Saf., vol. 15, pp. 137–147, Sep. 2023, doi: 10.2147/DHPS.S425858.

[15] R. K. Gan, J. C. Ogbodo, Y. Z. Wee, A. Z. Gan, and P. A. González, "Performance of Google bard and ChatGPT in mass casualty incidents triage," Am. J. Emerg. Med., vol. 75, pp. 72–78, Jan. 2024, doi: 10.1016/J.AJEM.2023.10.034.

[16] R. Ali et al., "Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank," Neurosurgery, vol. 93, no. 5, pp. 1090–1098, Nov. 2023, doi: 10.1227/NEU.0000000000002551.

[17] A. K. D. Dhanvijay et al., "Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology," Cureus, vol. 15, no. 8, Aug. 2023, doi: 10.7759/CUREUS.42972.

[18] D. E. O and C. E. Daniel O, "An analysis of Watson vs. BARD vs. ChatGPT: The Jeopardy! Challenge," AI Mag., vol. 44, no. 3, pp. 282–295, Sep. 2023, doi: 10.1002/AAAI.12118.

[19] V. Plevris, G. Papazafeiropoulos, and A. Jiménez Rios, "Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard," AI, vol. 4, no. 4, pp. 949–969, Oct. 2023, doi: 10.3390/ai4040048.

[20] S. Koga, N. B. Martin, and D. W. Dickson, "Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders," Brain Pathol., p. e13207, 2023, doi: 10.1111/BPA.13207.

[21] I. Seth et al., "Comparing the Efficacy of Large Language Models ChatGPT, BARD, and Bing AI in Providing Information on Rhinoplasty: An Observational Study," Aesthetic Surg. J. Open Forum, vol. 5, Jan. 2023, doi: 10.1093/ASJOF/OJAD084.

[22] "Natural Language Processing with Python and spaCy: A Practical Introduction - Yuli Vasiliev - Google Books." Accessed: Dec. 16, 2023. [Online]. Available: https://books.google.com.pk/books?id=w_ZqywEACAAJ&printsec=copyright&redir_esc=y#v=onepage&q&f=false

[23] R. Spring and M. Johnson, "The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK and SpaCy tools," System, vol. 106, p. 102770, Jun. 2022, doi: 10.1016/J.SYSTEM.2022.102770.

[24] R. Verma and A. Mittal, "Multiple attribute group decision-making based on novel probabilistic ordered weighted cosine similarity operators with Pythagorean fuzzy information," Granul. Comput., vol. 8, no. 1, pp. 111–129, Jan. 2023, doi: 10.1007/S41066-022-00318-1/METRICS.

[25] R. Zhang, Z. Xu, and X. Gou, "ELECTRE II method based on the cosine similarity to evaluate the performance of financial logistics enterprises under double hierarchy hesitant fuzzy linguistic environment," Fuzzy Optim. Decis. Mak., vol. 22, no. 1, pp. 23–49, Mar. 2023, doi: 10.1007/S10700-022-09382-3/METRICS.

[26] D. Dede Şener, H. Ogul, and S. Basak, "Text-based experiment retrieval in genomic databases," J. Inf. Sci., Sep. 2022, doi: 10.1177/01655515221118670/ASSET/IMAGES/LARGE/10.1177_01655515221118

670-FIG4.JPEG.

[27] R. M. Suleman and I. Korkontzelos, "Extending latent semantic analysis to manage its syntactic blindness," Expert Syst. Appl., vol. 165, p. 114130, Mar. 2021, doi: 10.1016/J.ESWA.2020.114130.

[28] Y. Chen, S. Nan, Q. Tian, H. Cai, H. Duan, and X. Lu, "Automatic RadLex coding of Chinese structured radiology reports based on text similarity ensemble," BMC Med. Inform. Decis. Mak., vol. 21, no. 9, pp. 1–11, Nov. 2021, doi: 10.1186/S12911-021-01604-9/TABLES/3.

[29] T. Bin Sarwar, N. M. Noor, and M. S. U. Miah, "Evaluating keyphrase extraction algorithms for finding similar news articles using lexical similarity calculation and semantic relatedness measurement by word embedding," PeerJ Comput. Sci., vol. 8, p. e1024, Jul. 2022, doi: 10.7717/PEERJ-CS.1024/SUPP-1.

[30] D. Vogler, L. Udris, and M. Eisenegger, "Measuring Media Content Concentration at a Large Scale Using Automated Text Comparisons," Journal. Stud., vol. 21, no. 11, pp. 1459–1478, Aug. 2020, doi: 10.1080/1461670X.2020.1761865.

Appendix A. Selected Topics on Oncology

No.	Topic
1	Cancer Biology
2	Cancer Genetics
3	Tumor Immunology
4	Oncogenes and Tumor Suppressor Genes
5	Cellular Signaling in Cancer
6	Epigenetics in Cancer
7	Cancer Metabolism
8	Tumor Microenvironment
9	Cancer Stem Cells
10	Angiogenesis and Cancer
11	Tumor Invasion and Metastasis
12	Cancer Biomarkers
13	Liquid Biopsy
14	Genomic Instability in Cancer
15	Precision Oncology
16	Cancer Epidemiology
17	Cancer Risk Factors
18	Cancer Prevention and Screening
19	Early Detection of Cancer
20	Diagnostic Imaging in Oncology
21	Pathology in Cancer Diagnosis
22	Clinical Oncology
23	Medical Oncology
24	Surgical Oncology
25	Radiation Oncology
26	Pediatric Oncology
27	Gynecologic Oncology
28	Hematologic Malignancies
29	Breast Cancer
30	Lung Cancer
31	Colorectal Cancer
32	Prostate Cancer

33	Pancreatic Cancer
34	Liver Cancer
35	Ovarian Cancer
36	Head and Neck Cancers
37	Skin Cancer (Melanoma and Non-Melanoma)
38	Thyroid Cancer
39	Gastrointestinal Cancers
40	Brain Tumors
41	Bone and Soft Tissue Sarcomas
42	Lymphomas
43	Leukemias
44	Immunotherapy in Cancer
45	Targeted Therapies
46	Chemotherapy
47	Radiation Therapy Techniques
48	Supportive Care in Oncology
49	Cancer Survivorship
50	Palliative Care in Oncology