Quantifying Similarity: Text-Mining Approaches to Evaluate ChatGPT and Google Bard Content in Relation to BioMedical Literature

Jakub Klimczak[1,2] and Ahmed Abdeen Hamed[2]

[1] Faculty of Computer Science, AGH, Kawiory 31, 30-055 Cracow, ["jklimczak@student.agh.edu.pl"]
[2] Clinical Data Science – Network Medicine and AI, Sano Centre for Computational Medicine , Nawojki 11, 30-072 Cracow, ["a.hamed@sanoscience.org"]

Abstract. The emergence of generative AI tools, empowered by Large Language Models (LLMs), has shown power in generating content. The assessment of the usefulness of such content has become an interesting research question. Using prompt engineering, we assess the similarity of such contents to real literature produced by scientists. In this exploratory analysis, we prompt-engineer ChatGPT and Google Bard to generate clinical content to be compared with medical literature, and we assess the similarities of the generated contents by comparing them with biomedical literature. Our approach is to use text-mining methods to compare documents and bigrams and to use network analysis to check the centrality. The experiments demonstrated that ChatGPT outperformed Google Bard in different similarity and term network centrality methods, but both tools achieved good results compared to the baseline.

Keywords: Generative AI · LLM · Content Assessment · Google Bard · ChatGPT · Text Mining · Provoking Questions

1 Introduction and Related Work

In 2022, our world witnessed an epic event an OpenAI launching pre-trained, and transformer-based Large Language Model. The tool is known to be conversational, generative, pre-trained, transformer-based, and hence its name [2]. The creation of ChatGPT started a new phenomenon in the IT world and implies the appearance of a lot of new models, with various architectures, like Google Bard with PaLM [1]. ChatGPT is known for its capability to receive prompts in natural languages. It also provides the human language responses [11]. LLMs can process extensive amounts of text for tasks such as translation [12], question answering, and content generation [4]. LLMs are used in health and biomedicine. Thirunavukarasu et al. describe usages of LLMs in medicine [16], such tools are very popular as chatbots in the biomedical domain, but with mixed results. A different case is dental medicine [5]. Eggmann et al. describe LLMs as a tool for finding and extracting information from giant amounts of medical data and

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_18

structuring medical Electronic Health Records (EHR). One promising direction is the Clinical Decision Support System. Scientists experiment with using LLM in such systems. Singhal et al. introduced MultiMedQA [14], the benchmark of medical questions to evaluate LLMs, as a tool that could be used in CDSS.

Before the launch of LLMs, and the generation of massive data, Real-World Data (RWD) played a recognizable role in CDSS and diagnostic applications [3]. The Food and Drug Administration defined RWD as "*data regarding the usage, or the potential benefits or risks, of a drug derived from a variety of sources other than traditional clinical trials*" [17]. PubMed articles can also be RWD. We think that LLMs can be a source of RWD and that checking their capabilities to generate such data is important. It is reasonable to infer that the natural language capabilities provided by generative AI could be used to build the next CDSS. As a result, researchers have explored their potential integration with such systems [10]. LLMs tools are associated with a lack of credibility [7]. Scientists point out a need for the detection of potential harms [8] of the generated data. The users of those tools confuse such responses for truth without questioning the harm. There are a lot of new AI projects, focused on decision-making. Scientists want to publish their system as fast as possible. Shortliffe claims, that despite the decision making we should focus on the evaluation [13]. Here we deal with the problem of initial verification. We are performing the exploratory analysis to evaluate the similarity to real data. In the following sections, we describe in detail the process for generating content that can be compared with articles. Comparing real and generated data is a popular topic. Researchers are looking for methods to detect and check the similarity of generated data [6].

2 Data

In this work, we use different datasets related to the prostate cancer topic:

(1) 10,000 **biomedical abstracts**, extracted from The PubMed web portal using the search keyword: "prostate cancer treatment" which is our baseline of comparison. PubMed contains science documents, but also clinical trials and reviews, and all of these types of documents have summary abstracts. We obtained the first 10000 abstracts of the most frequently cited.

(2) two generated datasets of 100 **abstracts**, received by prompt-engineering ChatGPT and Google Bard. We used LLMs to produce documents similar to PubMed abstracts (a random ID, a title, an abstract). LLMs use everything on the internet, so these documents are in abstract format but have sources on the whole internet. We have limited the amount of generated articles because, at the time of the data collection, we did not have access to the APIs.

2.1 Large Language Models

The data generation process was performed in May and June of 2023. During this time there were available versions of large language models: **(a) ChatGPT-3.5 turbo** - this is an upgraded version of the 3.5 version, with faster processing and

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_18

response time. At the moment there is a newer version available: ChatGPT-4.0. **(b) Google Bard** - experimental version of Google chatbot, the precursor of the modern model: Google Gemini. The information about specifications and specific subversion of the model is not accessible by the website, and it is impossible to receive this information directly. The API provides such information and allows one to choose one from a list of specific sub-models.

2.2 Prompt Engineering

This work aims to perform prompt engineering in ChatGPT and Google Bard to generate content related to prostate cancer treatment. Algorithm 1 shows the steps of the prompt engineering process to generate what we call real-world data.

Algorithm 1 Prompt-engineering for generating abstract-like documents

Require: The number n of simulated articles.
Require: The number w of words in each article.
[Content:] Generate a list of n real-world data reports with titles and abstracts.
[Specs:] For each abstract that contains three fields – GPT-ID, Title, and Abstract – make it to m words.
[Specs:] Make the GPT-ID random, containing at most five letters and numbers.
[Format:] A valid JSON format returned as an array of valid JSON records.
[Topic:] Investigating prostate cancer treatment.

3 Methods

The abstracts are analyzed using text mining and network analysis. We have performed two text mining methods (1) document similarity using the Cosine and Jaccard similarity[15], (2) bigram frequency comparison with Term Frequency-Inverse Term Frequency (TF-IDF) [9]. The network analysis methods are derived from the bigrams forming networks that can be analyzed to compare modularity and term centrality [18].

3.1 Text Mining Similarity Analysis

Following the step of generating the reports, we further compare the content of both tools using traditional text mining. This includes: (1) comparing the documents against other documents, (2) extracting and comparing bigrams of words, and (3) constructing networks of bigrams with identifying novel links. We conducted a comparison using random samples of documents against the PubMed corpus. for the **Document Similarity Analysis** – we count the similarity between real medical abstracts and the reports that were generated using LLMs. We are looking for the most matching pairs of generated-real articles. For this task, we used two metrics: (1) the Cosine, and the Jaccard. On the other hand, in **Bigram Analysis** we use bigrams. Bigram is a sequence of two next elements in the text, usually composed of letters, syllables, or words. For

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_18

example, the following words are bigrams: ('prostate cancer', 'cancer cells'). This measures the frequency similarity of bigrams extracted from literature and documents generated by generative AI tools. Bigrams can be used for the creation of graphs, that offer a model that can be used to explore topology and structural property. Here, we use the TF-IDF method with the different datasets of bigrams to count the importance of bigrams within the documents. In each case we are comparing also PubMed to PubMed articles, to use it as a baseline to interpret results. These Cosine and Jaccard methods are syntactic, they compare the structure of the text, and the number of words. But they can be used in other methods, to compare the semantic similarity. TF-IDF uses Cosine, but it is semantic similarity. This statistical-based method counts the importance of words in the text.

3.2 Networks Analysis

Bigrams can construct interesting networks that can be analyzed for their topology properties. The common words act as a linking node to connect more than one bigram. These approaches enable us to dissect the structural relationships in different terms in the generated and real data. Bigram networks are a popular approach for text analysis [7]. Such networks can be used in various tasks, including text classification, sentiment analysis, pattern recognition in text, and topic modeling. We select the top 50 most frequent bigrams from the entire corpus of documents from PubMed, ChatGPT, and Google Bard and build the bigram networks as rigorous models of comparison. The type of analysis we present is degree and closeness centrality as common measures that demonstrate the differences. The degree presents how many connections a specific unigram (word) has in the graph [19]. The closeness presents how close a given word is to all others in the same graph [19]. By comparing centrality metrics across different data sources, we can get valuable knowledge about similarities and differences in the texts. Centrality comparison allows to comparison of semantic relations between words in the text.

4 Results

We performed similarity experiments to measure the similarity of LLMs data and a sample of PubMed documents with PubMed article abstracts. We performed experiments on different dataset sizes (10, 25, 50, 75, and 100).

4.1 Document Similarity Analysis

We use two different methods to compare document similarity : (1) Cosine similarity, and (2) Jaccard similarity. Table 1 shows the result of the scores.

Results in table 1 show that the average cosine similarity scores for this method are very high, between 32-38 % for both solutions. ChatGPT shows more similarity than Bard. ChatGPT's similarity is around 35-38 %, and Bard's

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_18

Table 1. Combined Analysis Results

		Document Similarity		Bigram Analysis	Centrality	
Size	**Source**	**Cosine**	**Jaccard**	**TF-IDF**	**Degree**	**Closeness**
10	Bard	0.3435	0.1954	0.3999	0.0349	0.1618
	GPT	0.3802	0.2286	0.4182	0.0256	0.0471
	Pubmed	0.7683	0.1711	0.3524	0.0264	0.0590
25	Bard	0.3389	0.1914	0.40473	0.0392	0.1640
	GPT	0.3604	0.2139	0.44100	0.0273	0.0633
	Pubmed	0.8801	0.1914	0.3512	0.0219	0.0608
50	Bard	0.3336	0.1873	0.37391	0.0483	0.2223
	GPT	0.3612	0.2157	0.46699	0.0250	0.0456
	Pubmed	0.8804	0.1978	0.3636	0.0219	0.0461
75	Bard	0.3205	0.1775	0.3814	0.0505	0.2415
	GPT	0.3595	0.2147	0.4556	0.0250	0.0484
	Pubmed	0.8681	0.1956	0.3441	0.0256	0.0744
100	Bard	0.3202	0.1775	0.36766	0.0425	0.1821
	GPT	0.3531	0.2093	0.42217	0.0286	0.0829
	Pubmed	0.8421	0.1833	0.3217	0.0264	0.0722

average similarity is around 32-34 %. In the plot 1) we can see the trend, that for every sample, the average similarity score is higher for ChatGPT.

The Jaccard analysis is performed with a word-bag representation of text, that counts the number of common words. The similarity is not so high, but the generated texts are shorter. The average scores (table 1 are hovering around 18-23 %. This result indicates a good connection between the generated and real data. ChatGPT shows more similarity in average scores, than Bard. ChatGPT's similarity is around 21-23 %, and Bard's average similarity is 17-19 %. Plot 2) demonstrates the advantage of ChatGPT. The similarity between a random sample of PubMed articles with the whole dataset is at a similar level, between ChatGPT and Google Bard.

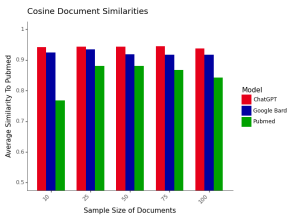

Fig. 1. Cosine similarity of documents

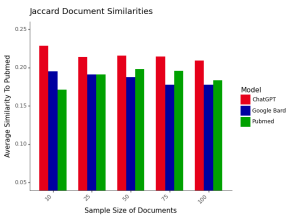

Fig. 2. Jaccard similarity of documents

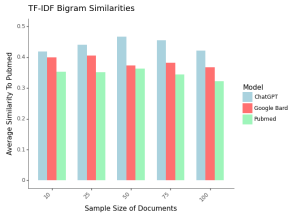

Fig. 3. TF-IDF similarity of bigrams

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_18

4.2 Bigram Similarity – TF-IDF Bigram Frequency Analysis

Both ChatGPT and Bard, show a meaningful level of average similarity (image 3), hovering around the 37-47 % mark. This shows a visible connection between the RWD generated by these models and existing medical research from sources like PubMed. The 3 plot shows that the difference is bigger with a bigger sample of documents. With this method, values of similarity between the PubMed sample and the whole PubMed corpus are a little lower in comparison to ChatGPT and Google Bard. That also speaks for the good quality of generated data.

4.3 Bigram Networks Analysis

The **degree centrality** (plot 4), shows a structural similarity between the PubMed and ChatGPT. Bard exhibits much higher values. This suggests that ChatGPT's values are closer to PubMed's. ChatGPT and PubMed centrality have prevalent degree centrality, between 0.02 and 0.03, and Bard is between 0.035 and 0.05. The **closeness centrality** (plot 5), shows a correlation between PubMed and ChatGPT. Bard has higher values, and ChatGPT's values are close to PubMed, showing that the networks are more similar. PubMed closeness is between 0.05 and 0.08, and for ChatGPT is between 0.05 and 0.89, for Google Bard it is between 0.16 and 0.24. The numbers support ChatGPT's better fit.

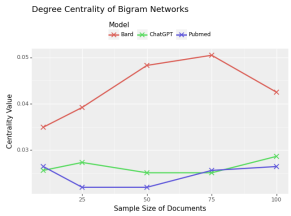

Fig. 4. Degree centrality of networks of bigrams

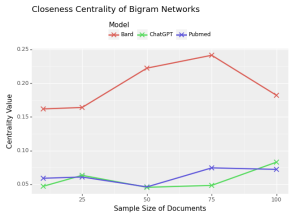

Fig. 5. Closeness centrality of networks of bigrams

5 Summary and Discussion

We presented a text-mining and network analysis approach to count the similarity between generated and real biomedical data. With random samples of documents, we observed that ChatGPT scored a closer similarity than Bard. This analysis in three different measures favors ChatGPT over Google Bard. The network analysis offers us another field to compare the similarities. The results of the experiments show that ChatGPT graphs exhibit closer similarities in structure and centrality. Table 2 shows top bigrams. We observed that ChatGPT offered 7, and Google Bard offered 3 bigrams that overlapped. While both LLMs have "prostate cancer" as the first bigram, they vary in the rest of the common bigrams. The "cancer patients" which was 7 in PubMed is 4 in ChatGPT. The

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_18

"quality life", which was 11 in PubMed is 5 in the ChatGPT dataset. This could indicate that ChatGPT was trained on data related to patient wellness, while PubMed data is more about the clinical aspects of the diseases.

Rank	PubMed Bigrams	GPT	Bard	Rank	PubMed Bigrams	GPT	Bard
1	prostate cancer	1	1	7	cancer patient	4	47
2	radiation therapy	8	–	8	specific antigen	–	–
3	radical prostatectomy	28	–	9	external beam	–	–
4	localized prostate	32	–	10	free survival	43	–
5	prostate specific	–	–	11	quality life	5	44
6	androgen deprivation	–	–	12	patient treated	–	–

Table 2. Bigram ranks in Pubmed and generated datasets

6 Conclusions and Future Directions

Gathering new datasets related to various diseases is one of the directions to further study. These results are within the scope of "prostate cancer", but there is a need to check other domains (e.g., diabetes, depression, cardiovascular) and newer models. For ChatGPT, we used version 3.5 (not 4.0), and for Google Bard, we used the version before Google Gemini, which now empowers Google Bard. Different benchmarks show that there is a huge difference between models and their successors. We are aware that such LLMs are black boxes and we do not know the details of them. Due to that future work will include testing such models in answering clinical questions, with MultiMedQA [14]. This popular benchmark is often used to test the quality and progress of new language models.

Acknowledgements This publication is partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement Sano No. 857533 and carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund. Additionally is partially created as part of the Ministry of Science and Higher Education's initiative to support the activities of Excellence Centers established in Poland under the Horizon 2020 program based on the agreement No MEiN/2023/DIR/3796.

References

1. Google bard. https://bard.google.com/, accessed: 2023-08-03
2. Openai chatgpt. https://chat.openai.com/, accessed: 2023-08-03
3. Baumfeld Andre, E., Carrington, N., Siami, F.S., Hiatt, J.C., McWilliams, C., Hiller, C., Surinach, A., Zamorano, A., Pashos, C.L., Schulz, W.L.: The current landscape and emerging applications for real-world data in diagnostics and clinical decision support and its impact on regulatory decision making. Clinical Pharmacology & Therapeutics **112**(6), 1172–1182 (2022)
4. Chung, J., Kamar, E., Amershi, S.: Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. pp. 575–593. Association for Computational Linguistics (ACL) (8 2023). https://doi.org/10.18653/v1/2023.acl-long.34

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_18

5. Eggmann, F., Weiger, R., Zitzmann, N.U., Blatz, M.B.: Implications of large language models such as chatgpt for dental medicine (2023). https://doi.org/10.1111/jerd.13046
6. Gao, C.A., Howard, F.M., Markov, N.S., Dyer, E.C., Ramesh, S., Luo, Y., Pearson, A.T.: Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. NPJ digital medicine **6**(1), 75 (2023)
7. Hamed, A.A., Wu, X.: Improving detection of chatgpt-generated fake science using real publication text: Introducing xfakebibs a supervised-learning network algorithm (2023)
8. Hamed, A.A., Zachara-Szymanska, M., Wu, X.: Safeguarding authenticity for mitigating the harms of generative ai: Issues, research agenda, and policies for detection, fact-checking, and ethical ai. IScience (2024)
9. Kim, S.W., Gil, J.M.: Research paper classification systems based on tf-idf and lda schemes. Human-centric Computing and Information Sciences **9** (12 2019). https://doi.org/10.1186/s13673-019-0192-7
10. Liao, Z., Wang, J., Shi, Z., Lu, L., Tabata, H.: Revolutionary potential of chatgpt in constructing intelligent clinical decision support systems (2023). https://doi.org/10.1007/s10439-023-03288-w
11. Moro, A., Greco, M., Cappa, S.F.: Large languages, impossible languages and human brains. Cortex **167**, 82–85 (10 2023). https://doi.org/10.1016/j.cortex.2023.07.003
12. Mu, Y., Reheman, A., Cao, Z., Fan, Y., Li, B., Li, Y., Xiao, T., Zhang, C., Zhu, J.: Augmenting large language model translators via translation memories. pp. 10287–10299. Association for Computational Linguistics (ACL) (8 2023). https://doi.org/10.18653/v1/2023.findings-acl.653
13. Shortliffe, E.H.: Role of evaluation throughout the life cycle of biomedical and health ai applications. BMJ health & care informatics **30**(1), e100925 (December 2023). https://doi.org/10.1136/bmjhci-2023-100925
14. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., y Arcas, B.A., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., Natarajan, V.: Large language models encode clinical knowledge. Nature **620**, 172–180 (8 2023). https://doi.org/10.1038/s41586-023-06291-2
15. Thada, V., Jaglan, V.: Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. International Journal of Innovations in Engineering and Technology **2**, 202–205 (2013), http://www.dknmu.org/uploads/file/6842.pdf
16. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine (8 2023). https://doi.org/10.1038/s41591-023-02448-8
17. U.S. Food and Drug Administration: Framework for fda's real-world evidence program (Year of Publication), https://www.fda.gov/media/120060/download, accessed on October 27, 2023
18. Wang, G., Shen, Y., Luan, E.: Measure of centrality based on modularity matrix. Progress in Natural Science **18** (2008). https://doi.org/10.1016/j.pnsc.2008.03.015
19. Zhang, J., Luo, Y.: Degree centrality, betweenness centrality, and closeness centrality in social network. Atlantis Press (2017). https://doi.org/10.2991/msam-17.2017.68

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_18