Preprints.org

Article

Improving Detection of ChatGPT-Generated Fake Science Using Real Publication Text: Introducing xFakeBibs a Supervised-Learning Network Algorithm

Ahmed Abdeen Hamed [*]

Article

Improving Detection of ChatGPT-Generated Fake Science Using Real Publication Text: Introducing xFakeBibs a Supervised-Learning Network Algorithm

Ahmed Abdeen Hamed [1,*,†] and Xindong Wu [2,*,†]

[1]   Sano Centre for Computational Medicine, Clinical Data Science, Cracow, 30-072, Poland
[2]   Zhejiang Lab, Research Center for Knowledge Engineering, Hangzhou, 311121, P.R. China
[*]   Correspondence: xwu@zhejianglab.com (A.A.H.); a.hamed@sanoscience.org (X.W.)
[†]   These authors contributed equally to this work.

Abstract: ChatGPT is becoming a new reality. In this paper, we show how to distinguish ChatGPT-generated publications from counterparts produced by scientists. Using a newly designed supervised Machine Learning algorithm, we demonstrate how to detect machine-generated publications from those produced by scientists. The algorithm was trained using 100 real publication abstracts, followed by a 10-fold calibration approach to establish a lower-upper bound range of acceptance. In the comparison with ChatGPT content, it was evident that ChatGPT contributed merely 23% of the bigram content, which is less than 50% of any of the other 10 calibrating folds. This analysis highlights a significant disparity in technical terms where ChatGPT fell short of matching real science. When categorizing the individual articles, the xFakeBibs algorithm accurately identified 98 out of 100 publications as fake, with 2 articles incorrectly classified as real publications. Though this work introduced an algorithmic approach that detected the ChatGPT-generated fake science with a high degree of accuracy, it remains challenging to detect all fake records. This work is indeed a step in the right direction to counter fake science and misinformation.

Keywords: ChatGPT; Generative AI; Fake Publication; Human-Generated Publications; ML Algorithm; Fake Science

---

Introduction

With ChatGPT [1] being a new reality, our world is in a controversial state. On the one hand, there is a camp of optimists that sees potential and seeks to utilize it. On the other hand, there are doubters who remain skeptical and search for validation and further assessments to decide how this new paradigm affects our lives. This split provided a strong motivation for this work and triggered the effort of providing an assessment tool for fake publications generated by ChatGPT. Without any doubt, real science (and publications) is one of the most sacred sources of knowledge. That is because it is an invaluable source for any future discovery [2–4]. The spread of various predatory journals has contributed to fake science and caused it to be on the rise [5–8]. With the influence of social media, the impact is far-reaching [9,10]. Particularly, during the Coronavirus global pandemic, the propagation of misinformation that surrounded the significance of vaccination had led people to reject it, putting others at harm [11–13]. Another disturbing example that occurred also during the pandemic was the propagation of an article that reported fake results about how the deficiency of vitamin D had led to the death of 99 percent of the population studied. Though the article was eventually withdrawn, the damage of misinformation was magnified by a DailyMail news article and made it reach globally [14,15]. It is crucial to protect the authenticity of the scientific work recorded in publications from fraud or any influential factors that make it an untrusted source of knowledge.

Here, we demonstrate how the emergence of ChatGPT (and many other Generative AI tools) has, in many ways, impacted our society today: (1) the launch of many special issues and themes to study, assess, analyze, and test the impact and the potential of ChatGPT [16–20], (2) the adoption of new policies by journals regarding ChatGPT authorship [21–25], (3) the development of ChatGPT

plugins, and inclusion in professional services such as Expedia and Slack [26], (4) the development of educational tools (e.g., Wolfram); and the potential of developing learning and educational support tools (e.g., Medical Licensing Examination [27].)

Contribution

The main contribution of this article two items: (1) the introduction of a bigram network analysis approach that distinguishes between contents of real science coming from real publications and simulated publications created by ChatGPT, (2) The introduction of a network-based supervised machine learning algorithm to detect fake articles from real articles conducted by real scientists.

Results

The results of this paper were derived from two distinct types of experiments: (1) Analyzing the dataset of ChatGPT-generated publications in bulk, and (2) classifying the articles to detect each individual one using the xFakeBibs algorithm and confirming it as fake. In the first type of experiments, we examined both the content and the network structures generated from the article bigrams.

*The Outcome of ChatGPT Bigram Analysis*

Table 1 presents a comprehensive overview of summary statistics obtained from comparing the quantities of bigrams produced by the 10-folds dataset against ChatGPT-generated bigrams. The "Data Source" column specifies which of the 10-folds from the ChatGPT data is being referred to. The subsequent column records the count of overlapping bigrams in comparison with the training dataset. The subsequent column gauges the overlap's percentage relative to the size of the dataset it pertains to. The final column compiles the overlap percentage, offering a depiction of similarity between all 10-folds and ChatGPT.

Table 1. Comparison of bigram overlaps and similarity to training dataset.

Data Source	Number of Bigram Overlaps	Percent to Self %	Similarity to Training%
Fold-1	202	0.21	0.22
Fold-2	183	0.19	0.20
Fold-3	178	0.22	0.19
Fold-4	178	0.20	0.19
Fold-5	180	0.21	0.20
Fold-6	180	0.19	0.20
Fold-7	179	0.19	0.19
Fold-8	181	0.20	0.20
Fold-9	184	0.20	0.20
Fold-10	201	0.23	0.22
GPT-Test	81	0.16	0.09

The count of overlapping bigrams across all 10-folds varied from 178 to 202 bigrams, whereas ChatGPT exhibited 81 bigram overlaps against the training data. This is succinctly expressed as a similarity percentage of 19% to 22% for all 10 folds, while ChatGPT's similarity amounted to only 9%.

These statistics are visually represented in the barplot depicted in Figure 1. The X-axis denotes the data source (10-Folds and ChatGPT), while the Y-axis signifies the actual number of nodes contributed by each source. The barplot distinctly illustrates a significant disparity between the 10-folds of genuine publications and the fake publications generated by ChatGPT.
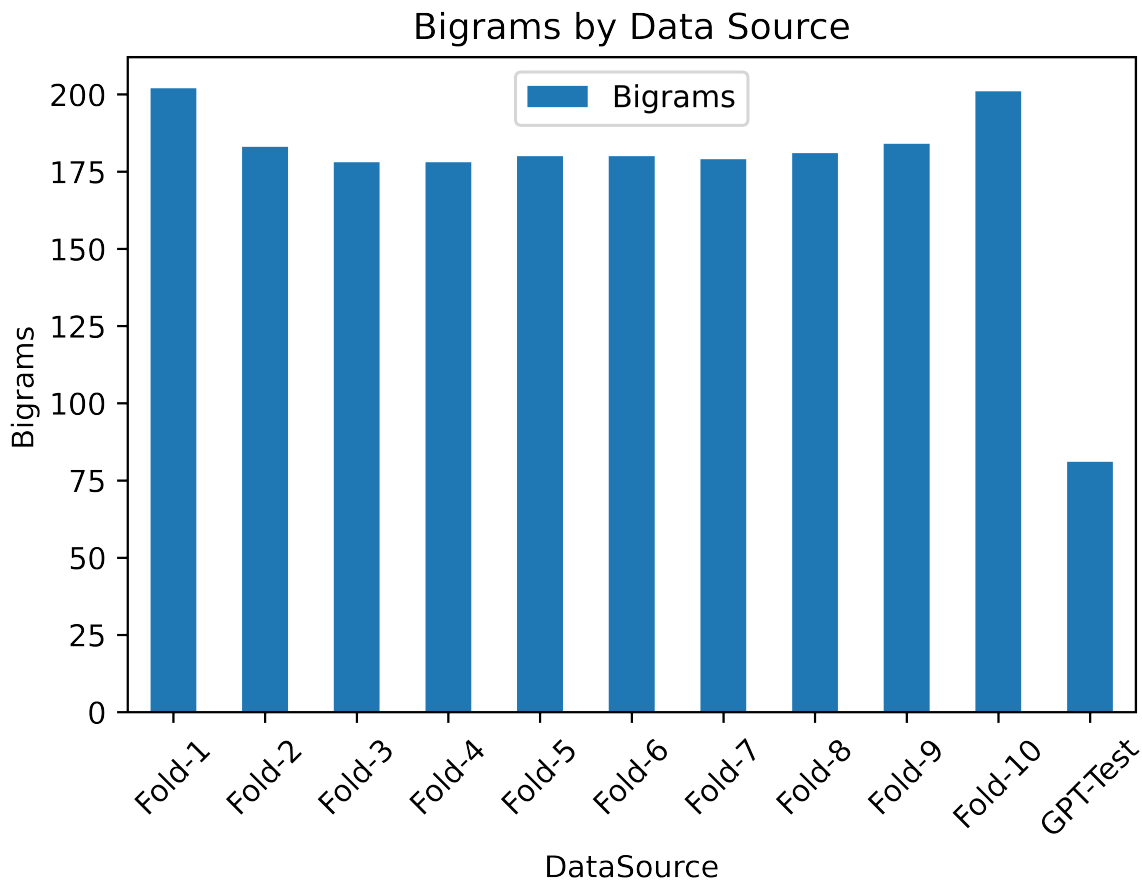
Figure 1.   The number of overlapping bigrams between 10-folds of real publications and ChatGPT-generated publications, in comparison with a training dataset from real publications.

The Structural Analysis of ChatGPT Bigram Network

The bigrams of each source offer a network model which can be further analyzed for their structure. In Table 2, we present the result of this analysis. The columns capture the Data Source of the bigrams that made up the networks (training, calibrating folds, and ChatGPT), The Number of Nodes of each network, Number of Edges of each source, size of the Largest Connected Components (s-LLC), and Percentage of new edges that contributed to the LLC without adding new nodes. The results indicate that the connected component percentages that were generated from the 10-folds lie between (51%-70%), and ChatGPT scored only 23%. Clearly, there is a serious deficit of the ChatGPT contribution to the structure of the training model by at least approximately 50%.

Table 2. Summary statistics of the network analysis.

Index	Data Source	No. Nodes	No. Edges	No. Connected Components	Connected Components Percent
0	F-1	1086	1624	855	0.67
1	F-2	1078	1625	867	0.69
2	F-3	1019	1494	777	0.51
3	F-4	1058	1570	827	0.61
4	F-5	1068	1533	825	0.61
5	F-6	1102	1624	860	0.68
6	F-7	1077	1593	871	0.70
7	F-8	1108	1580	846	0.65
8	F-9	1108	1580	846	0.65
9	F-10	1075	1540	848	0.65
10	ChatGPT	801	1312	632	0.23

In Figure 2 we summarized the full statistics represented by the number of nodes, the number of edges, and the size of connected components. This figure shows an interesting behavior of how the sizes of the connected components were very closely clustered together, while the size of ChatGPT connected components (represented by a blue dot at the bottom left of each plot) looked isolated and appeared to be an insignificant anomaly. Figure 3 shows a piechart that demonstrates the different sizes of the LLC of each data source. The ChatGPT contributes a much smaller slice when it is compared with all other 10-folds.
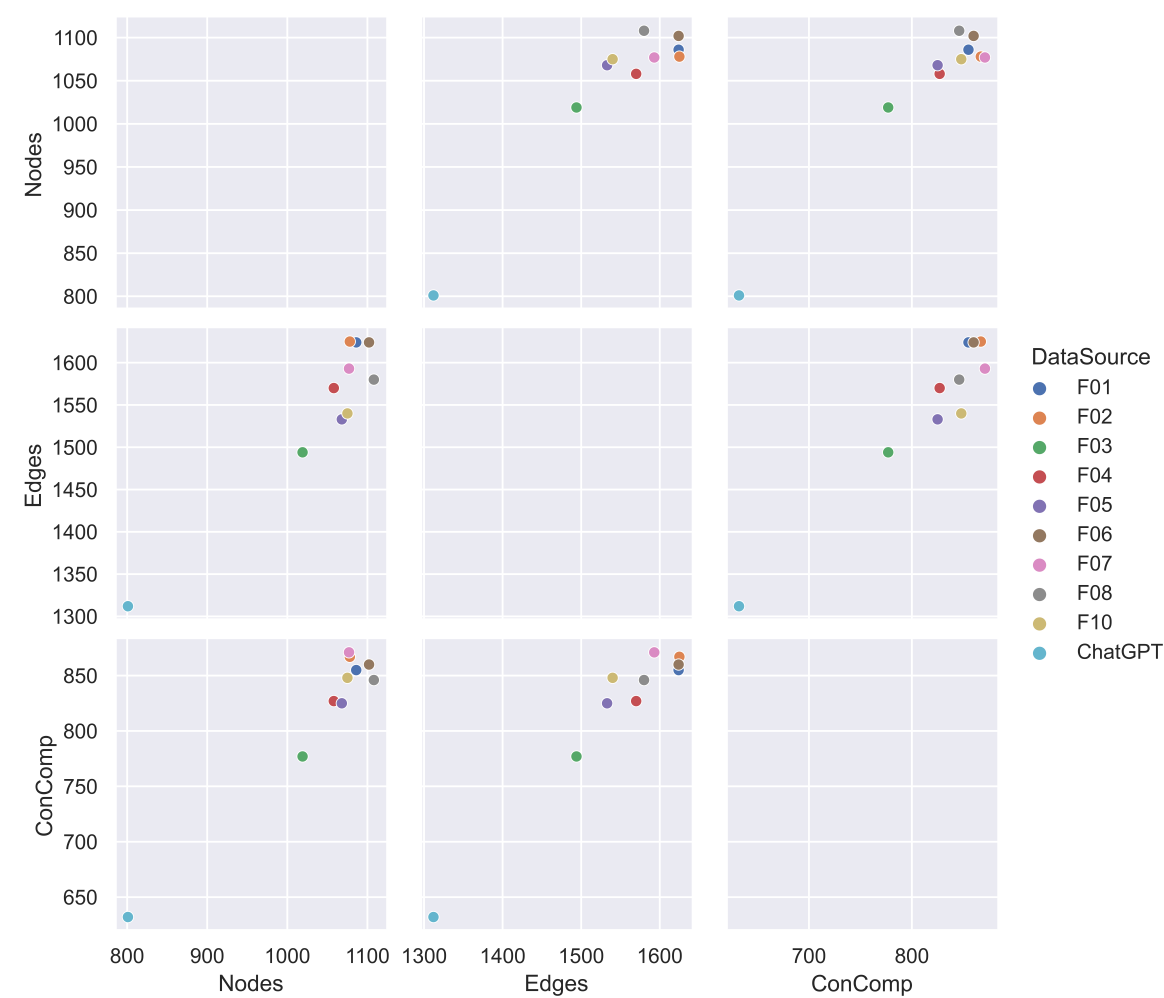
Figure 2. The summary statistics are generated from a pairplot that displays the relationships among any pair of variables in the plot, including Nodes, Edges, and Connected Components.

Largest Connected Component Percentages Per DataSource
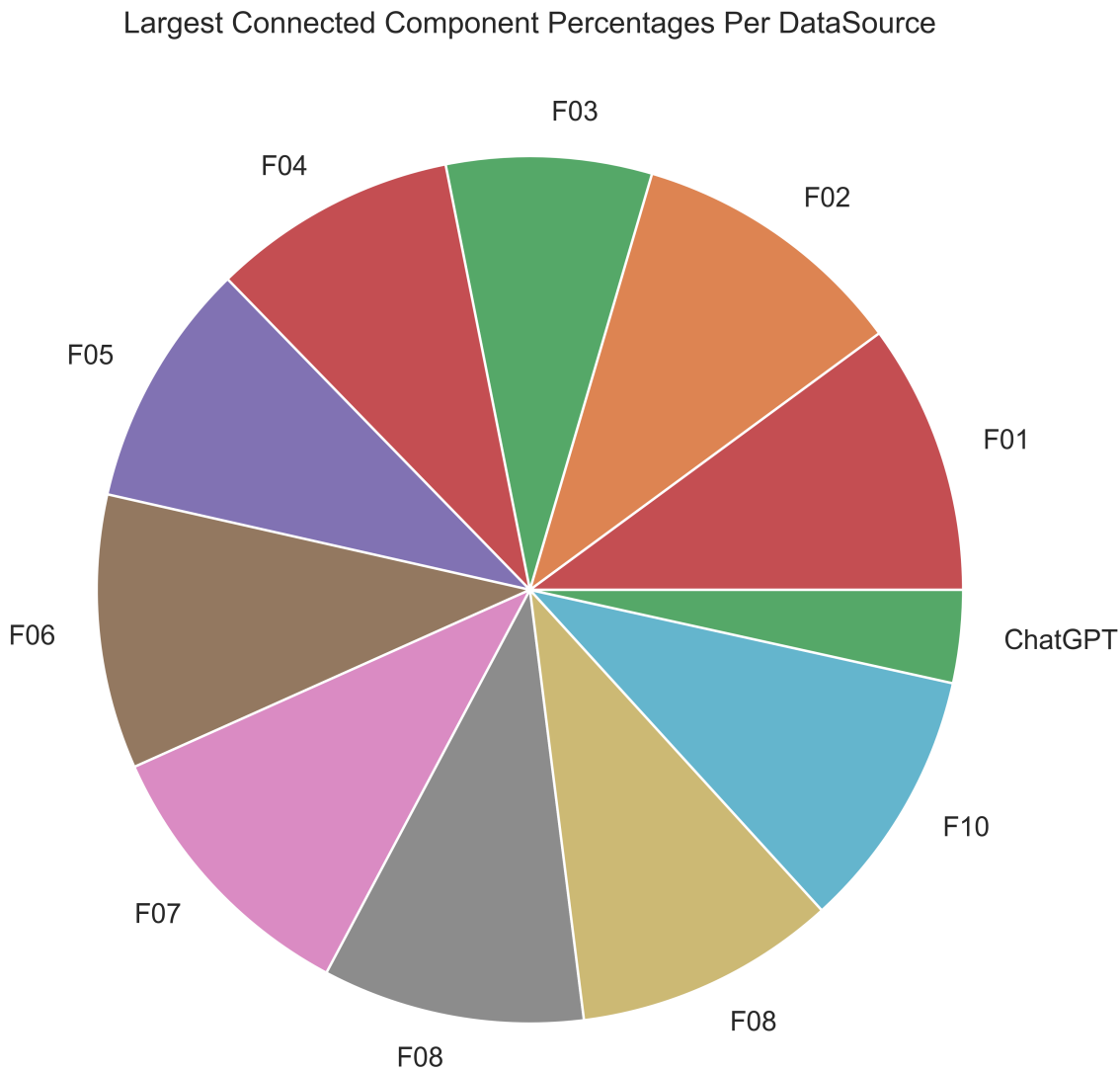
Figure 3. The size of the largest connected component in a network generated from ChatGPT fake articles is compared with the 10-folds. It is evident that ChatGPT has the smallest size among all the other slices, and the least contributions.

Classification of ChatGPT Fake Articles

On average, each article from the calibrating folds introduced around 21.96 to 25.12 new bigrams that impacted the training model. This established a threshold range for identifying genuine articles, with the lower bound set at 21.96 and the upper bound at 25.12. The scores for each of the calibrating folds are displayed in Table 3, offering insight into the computation of these lower and upper bounds for the classification process.

Table 3. Average number of bigrams contributing edges to the LCC model for each fold.

Fold	F-1	F-2	F-3	F-4	F-5	F-6	F-7	F-8	F-9	F-10
Avg. Bigrams	24.93	25.12	21.57	23.80	22.49	22.91	23.52	22.65	23.73	21.96

The xFakeBibs algorithm successfully identified 98 ChatGPT-generated articles that exceeded these bounds, indicating they were likely fake. Interestingly, 2 ChatGPT articles managed to remain

within the acceptable of being real range and went undetected. In Figure 4, we visually represent the number of bigrams contributed by each article, which in turn influenced the training model for every fold. The figure also highlights the consistency of the calibrated folds, showing their similar behavior and alignment, while ChatGPT's contributions appear as an anomaly.
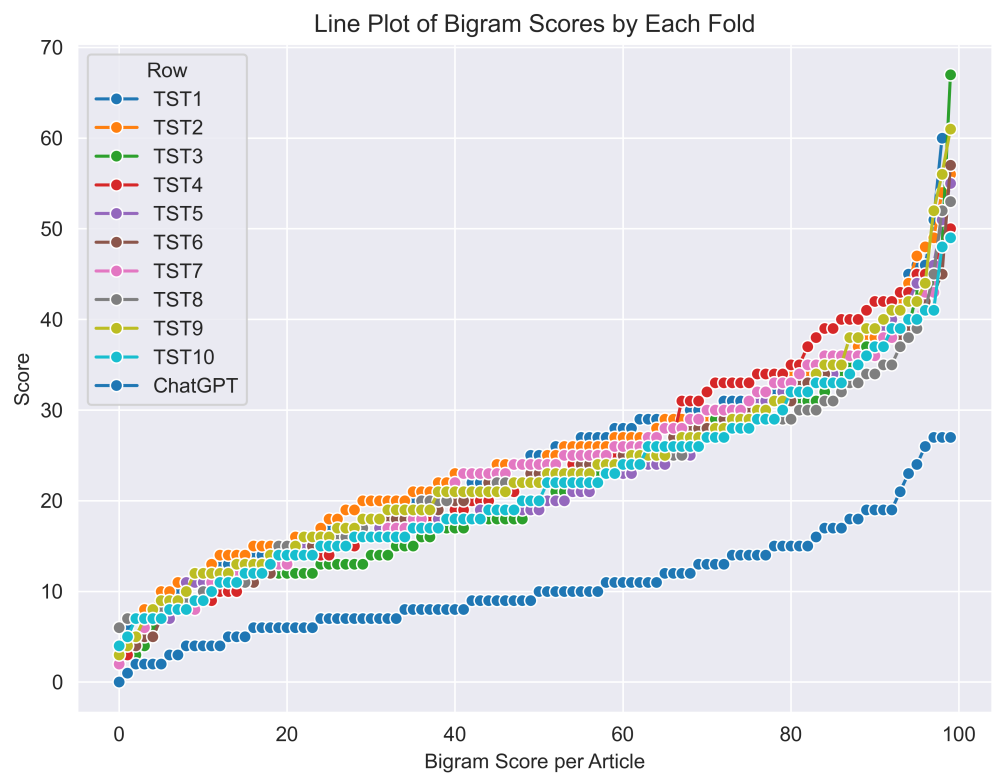
Figure 4. The bigram scoring behavior of the 10-folds of real publications shows a high degree of similarity, whereas ChatGPT exhibits a distinctly different behavior compared to all other folds, except for a few points that display similar behavior.

Methods

Data Collection

We utilized two distinct datasets: (1) for training and calibrating the algorithm, we queried PubMed for "Alzheimer's Disease and comorbidities" and gathered 1000 abstracts of publications that are human-produced; (2) for detecting ChatGPT articles, we instructed ChatGPT (version GPT-3.5) to generate 100 abstracts also related to "Alzheimer's disease and comorbidities." Both datasets underwent preprocessing using the same mechanism to handle data noise and stop words.

Code Availability

Both the datasets and the code required for executing the algorithm have been included as part of the supplementary material. The algorithms are fully implemented using the Python programming language, and we have provided a detailed step-by-step guide for execution using Jupyter Notebook.

Exploratory Analysis of ChatGPT Articles

To distinguish ChatGPT-generated content from real publications, we conducted two different types of analyses: (1) using an equal number of records, we compared the Term Frequency-Inverse Document Frequency (TFIDF) [28–33] of bigrams generated from the two sources (ChatGPT records, PubMed records), and (2) since bigrams provide a natural mechanism to form networks (where the words in the bigrams act as nodes and the edge represents the bigram relationship), we analyzed these networks from a structural point of view.

ChatGPT Bigram-Similarity Analysis

Bigrams are any two consecutive words that may prove significant in any given text-based dataset [34,35]. Using the Term Frequency-Inverse Term Frequency (TF-IDF) approach, we generated bigrams from three different types of resources: (1) a training dataset, which comprises 100 real publications, (2) a calibration using 10 folds, each containing 100 publications, and (3) the classification of 100 publications generated using ChatGPT. We compared the TF-IDF scores of the training dataset with all the calibrating 10 folds to establish lower/upper bounds for testing against the ChatGPT-generated fake articles. We computed the bigram similarities that overlapped between the training dataset and the calibrations. This process was repeated to test the bigram similarities generated from ChatGPT content against the 10-fold calibrations. The similarity ratios from the two comparisons revealed significant differences between the bigrams of real publications and the bigrams of ChatGPT-generated publications. Figure 5 presents a side-by-side WordCloud comparison of TF-IDF bigrams, where the left side represents real publications, while the right side displays bigrams generated from ChatGPT-generated fake records. Although phrases like "Alzheimer's disease" and "cognitive impairment" appear in both real publications and ChatGPT records, they are assigned different weights and rankings. Table 4 displays the scoring of the top 4 color-coded bigrams with their similarities.

Table 4. Comparison of Top-5 Bigrams in ChatGPT Publications and Real Publications, Color-Coded for Overlapping

Source	Bigram-1	Bigram-2	Bigram-3	Bigram-4	Bigram-5
ChatGPT	AD patients	Cognitive impairment	older adults	Alzheimer's disease	Risk factors
Real Bibs	Alzheimer's disease	Disease AD	AD patients	Cognitive impairment	Increased risk
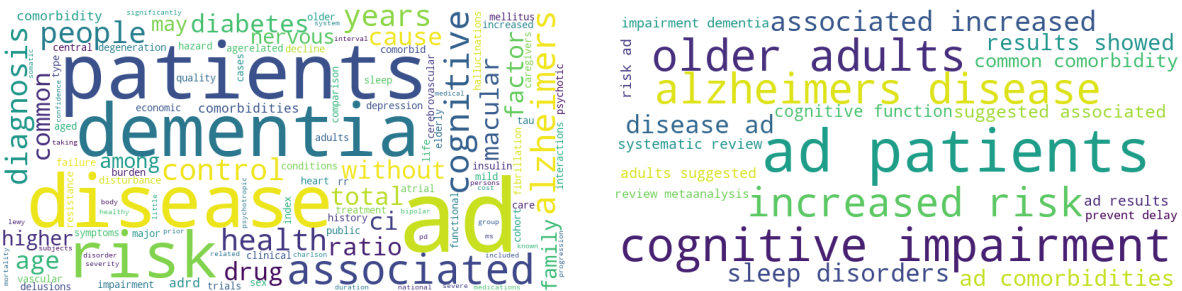
Figure 5. A side-by-side WordCloud comparison depicting the top-20 bigrams from real publications versus ChatGPT-generated content.

ChatGPT Network Structural Analysis

In addition to the bigram similarity analysis mentioned above, we also leveraged the natural mechanism for constructing networks using connected pairs of words [15,36]. When all the bigrams are linked together, each word can be connected to multiple other words. This forms a network where the nodes are words, and the edge semantics is the bigram predicate. The purpose of this analysis is to study a specific structural property of the ChatGPT bigram network and compare it with

bigram networks generated from real publications. The main idea is to measure the largest connected components (LCC) of each type of network and observe whether there is a significant difference between the networks generated from ChatGPT contents when compared against the networks of real publications. The LCC presents an admissible characteristic that represents the bigram networks. This is due to the availability of large degree nodes, which ensures the stability and robustness of the networks [37].

Here, we computed the LCC of the training model to establish a baseline of comparison. We then proceeded by computing the LCCs of all the 10-folds, one-fold at a time, to establish calibration. The final step was to compute the LCC of the ChatGPT network of bigrams to test against the calibration. The calibration process is outlined in a step-by-step manner in Algorithm 1 as follows:

Algorithm 1 The calibration steps involve computing using 10 individual folds for comparison against the training baseline.

1: Begin with the 10-fold datasets and compute the largest connected component for each fold one at a time.

2: Combine the largest component of each fold with the training components. Add them to the training network while considering each edge represented by a bigram.

3: Identify the bigrams that caused structural changes in the network, specifically those that introduced new edges to the original nodes in the training network. Exclude bigrams that merely added new nodes.

4: Evaluate the impact of bigrams from each fold on the training network by calculating the following ratio:

$$\text{Ratio} = \frac{\text{Number of Newly Introduced Edges}}{\text{Number of Original Edges}}$$

5: Extract the range that the ratio provides to discern how the inclusion of a specific publication affects the structure of the training network.

6: Compute the same ratio for the network generated from ChatGPT and compare it with the ratios obtained from the 10-fold datasets mentioned earlier.

Classification of ChatGPT Fake Articles

The above analysis demonstrated how fake ChatGPT publications are fundamentally different in content and structure when compared to real publications. Although the analysis succeeded in discriminating against those 100 generated publications, it does not solve the problem of detecting a single fake article. In this section, we address this issue by introducing our newly designed algorithm, which we call xFakeBibs. Previous work has attempted to address the problem of detecting fake science and news, especially during the Coronavirus pandemic and the associated infodemic. However, we believe it suffered from the unavailability of comprehensive fake publication datasets [15]. Another subsequent issue was the lack of model calibration and a proper data-driven mechanism to identify a lower-upper bound for classifying individual publication instances. Here, we present the xFakeBibs algorithm that is specialized in classifying fake publications, particularly those generated using ChatGPT. We train the algorithm using a partition of 100 publications selected from the Alzheimer's dataset based on the TF-IDF bigrams generated earlier. This step is followed by a calibration process to compute the lower-upper bound scores. Algorithm provides the is a detailed steps involved in this process:

Algorithm 2 A detailed description of the xFakeBibs algorithm

---

1: Construct a network of bigrams using the training data to represent the training bigram network.
2: Extract the largest connected component (LCC) from the network and compute its size.
3: **for** fold $\in$ Fold1-Fold-10 **do**
4:     **for** article in fold **do**
5:         Extract the TF-IDF bigrams from the article.
6:         Initialize counters: $count\_edges = 0$, $count\_matched = 0$.
7:         **for** bigram in TF-IDF bigrams **do**
8:             **if** bigram contributes an edge to the LCC **then**
9:                 Increment $count\_edges$ by 1.
10:                **if** bigram matches LCC edge **then**
11:                    Increment $count\_matched$ by 1.
12:                **end if**
13:            **end if**
14:        **end for**
15:        Reset the training model to the original state.
16:        Calculate the mean of the number of edges contributed by each fold.
17:        Designate the minimum value as the lower bound for labeling.
18:        Designate the maximum value as the upper bound for labeling.
19:        **if** number of bigrams contributing edges is within bounds **then**
20:            Classify article as POSITIVE (real publication).
21:        **else**
22:            Classify article as NEGATIVE (fake publication).
23:        **end if**
24:    **end for**
25: **end for**

In Figure 6, we present a screenshot showcasing the Python code responsible for executing the classification process on each of the ChatGPT-generated articles. This step involves utilizing the predefined lower and upper bounds to assess the count of bigrams that introduce new connections to the LCC model without introducing additional nodes. If this count falls within the established bounds, the article is classified as POSITIVE; conversely, if it lies outside the bounds, it is categorized as NEGATIVE. This approach effectively identifies each generated article based on its alignment with the structural properties of real publications.

```python
## classify ChatGPT articles one at a time

lower_bound = 21.96
upper_bound = 24.93

counter = 0
positive = 0
negative = 0

for fake_article in stopped_chatGPT_test[0:100]:
    edges_added = measure_chat_GPT_article(fake_article, giant_cc)
    counter += 1
    if (len(edges_added) >= lower_bound) and  (len(edges_added) <= upper_bound):
        positive += 1
        print('Article: ', counter, ', added edges: ', len(edges_added), 'POSITIVE')
    else:
        negative += 1
        print('Article: ', counter, ', added edges: ', len(edges_added), 'NEGATIVE')

print('\n\n--------------------------------------\n')
print('NeoNet Classifier TRUE POSITIVE :  ', format(positive/100), '.2f')
print('NeoNet Classifier FALSE NEGATIVE:  ', format(negative/100), '.2f')
print('\n\n--------------------------------------\n')
```

Figure 6. provides a detailed insight into the implementation of the classification step within the algorithm. This screenshot illustrates how each individual article was assessed and categorized as either fake or real based on the algorithm's analysis. The process involves measuring the number of bigrams that contribute new edges to the largest connected component (LCC) model, without introducing new nodes. By applying the lower and upper bounds derived from the calibration process, articles that fall within the specified range are classified as POSITIVE (real publications), while those outside the bounds are classified as NEGATIVE (fake publications). This visualization offers a comprehensive view of the algorithm's decision-making process in distinguishing between real and generated content.

Limitations

We trained and calibrated our xFakeBibs algorithm using PubMed abstracts, not full text. The dataset was derived from a query related to "Alzheimer's Disease and Co-morbidities." The current ChatGPT portal was unable to generate a substantial dataset of full-text articles. As a result, we resorted to simplifying the use of abstracts as a workaround.

Discussion

In our study, we used two methods to analyze fake articles created by ChatGPT: (1) examining the content, and (2) studying the structure of networks formed by words. Both approaches revealed significant differences between articles from ChatGPT and those created by actual experts in the field. Although there were some common terms like "Alzheimer's Disease" or "AD disease," the way the topics were discussed in the ChatGPT articles was less scientific. As a result, the content of these articles, measured in pairs of words (bigrams), was much smaller compared to the real articles.

We also introduced a new algorithm called xFakeBibs, which includes a calibration step. To establish a benchmark for comparison, we used real article abstracts as they're trusted by the scientific community. We divided these abstracts into 10 sets, each with 100 abstracts, to ensure fair analysis. This gave us a range of values that real articles should fall within. The algorithm successfully identified 98 out of 100 ChatGPT-generated articles as NOT REAL (fake). While this result is promising, it's concerning that ChatGPT can still produce fake articles that pass these tests. Fake science is a serious issue that threatens our knowledge and safety.

The most significant results of this research are as follows: (1) the medical publications generated by ChatGPT exhibit significant behavior, whether considered individually or in bulk, when analyzed for content and network model structure. When dealing with a dataset of fake publications, such behavior can be identified using appropriate computational methods and machine learning algorithms, as presented in this study. (2) Despite the unusual behavior observed in ChatGPT-generated content, detecting a single publication as being generated by ChatGPT remains a challenging problem.

Conclusions

When we asked a high school student about their knowledge of ChatGPT, their response was, "Do you mean that thing that does your homework for you?" Indeed, ChatGPT is an incredibly intelligent tool with a wide range of impressive capabilities. It has introduced a valuable dataset for experimentation, which was previously missing. However, there's also a concerning aspect associated with it that could pose a threat to the future of science. If younger generations, who are the future pioneers, use ChatGPT to plagiarize, it could undermine the integrity of research and learning.

Although machine learning algorithms can help detect fake science, there's an ethical responsibility to use such tools responsibly and regulate their usage. It's noteworthy that certain countries like Italy have taken the extreme step of banning ChatGPT. While the authors believe that such measures might be too drastic, addressing ethical concerns is still not entirely clear. As ChatGPT itself states, "It is up to individuals and organizations to use technology like mine in ways that promote positive outcomes and to minimize any potential negative impacts."

Looking ahead, there are several avenues for future research based on our current work:

- Utilizing ChatGPT APIs to generate complete publications and comparing them with archived full-text articles.
- Testing the xFakeBibs algorithm with publications in various subject areas.
- Fact-checking ChatGPT's responses for well-known questions that require reasoning (ongoing) [38].
- Training ChatGPT to provide answers specific to certain domains, such as clinical, medical, chemical, and biological applications.

Author contributions statement: A.H. conceived the idea(s), A.A. and X.U. designed the experiment(s), A.A. and X.W. analyzed the results. All authors written reviewed the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. ChatGPT, 2023. Accessed August 15, 2023.
2. Synnestvedt, M.B.; Chen, C.; Holmes, J.H. CiteSpace II: visualization and knowledge discovery in bibliographic databases. AMIA annual symposium proceedings. American Medical Informatics Association, 2005, Vol. 2005, p. 724.
3. Holzinger, A.; Ofner, B.; Stocker, C.; Calero Valdez, A.; Schaar, A.K.; Ziefle, M.; Dehmer, M. On graph entropy measures for knowledge discovery from publication network data. Availability, Reliability, and Security in Information Systems and HCI: IFIP WG 8.4, 8.9, TC 5 International Cross-Domain Conference, CD-ARES 2013, Regensburg, Germany, September 2-6, 2013. Proceedings 8. Springer, 2013, pp. 354–362.
4. Usai, A.; Pironti, M.; Mital, M.; Aouina Mejri, C. Knowledge discovery out of text data: a systematic review via text mining. *Journal of knowledge management* **2018**, *22*, 1471–1488.

5. Thaler, A.D.; Shiffman, D. Fish tales: Combating fake science in popular media. *Ocean & Coastal Management* **2015**, *115*, 88–91.

6. Hopf, H.; Krief, A.; Mehta, G.; Matlin, S.A. Fake science and the knowledge crisis: ignorance can be fatal. *Royal Society open science* **2019**, *6*, 190161.

7. Ho, S.S.; Goh, T.J.; Leung, Y.W. Let's nab fake science news: Predicting scientists' support for interventions using the influence of presumed media influence model. *Journalism* **2022**, *23*, 910–928.

8. Frederickson, R.M.; Herzog, R.W. Addressing the big business of fake science. *Molecular Therapy* **2022**, *30*, 2390.

9. Rocha, Y.M.; de Moura, G.A.; Desidério, G.A.; de Oliveira, C.H.; Lourenço, F.D.; de Figueiredo Nicolete, L.D. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *Journal of Public Health* **2021**, pp. 1–10.

10. Walter, N.; Brooks, J.J.; Saucier, C.J.; Suresh, S. Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health communication* **2021**, *36*, 1776–1784.

11. Loomba, S.; de Figueiredo, A.; Piatek, S.J.; de Graaf, K.; Larson, H.J. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour* **2021**, *5*, 337–348.

12. Lewandowsky, S.; Ecker, U.K.; Seifert, C.M.; Schwarz, N.; Cook, J. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* **2012**, *13*, 106–131.

13. Myers, M.; Pineda, D. Misinformation about vaccines. *Vaccines for biodefense and emerging and neglected diseases* **2009**, pp. 255–270.

14. Matthews, S.; Spencer, B. Government orders review into vitamin D's role in Covid-19, 2020.

15. Abdeen, M.A.; Hamed, A.A.; Wu, X. Fighting the COVID-19 Infodemic in News Articles and False Publications: The NeoNet Text Classifier, a Supervised Machine Learning Algorithm. *Applied Sciences* **2021**, *11*, 7265.

16. Eysenbach, G.; others. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Medical Education* **2023**, *9*, e46885.

17. IEEE Special Issue on Education in the World of ChatGPT and other Generative AI, 2023. Accessed April 13, 2023.

18. Financial Innovation. Accessed April 13, 2023.

19. Special Issue "Language Generation with Pretrained Models", Year. Accessed April 13, 2023.

20. Call for Papers for the Special Focus Issue on ChatGPT and Large Language Models (LLMs) in Biomedicine and Health, Year. Accessed July 4, 2023.

21. Do you allow the use of ChatGPT or other generative language models and how should this be reported?, Year. Published March 16, 2023. Accessed April 13, 2023.

22. Null, N. The PNAS Journals Outline Their Policies for ChatGPT and Generative AI. *PNAS Updates* **2023**. Published online, doi:10.1073/pnas-updates.2023-02-21.

23. As scientists explore AI-written text, journals hammer out policies, Year. Accessed April 13, 2023.

24. Fuster, V.; Bozkurt, B.; Chandrashekhar, Y.; Grapsa, J.; Ky, B.; Mann, D.L.; Moliterno, D.J.; Shivkumar, K.; Silversides, C.K.; Turco, J.V.; others. JACC Journals' Pathway Forward With AI Tools: The Future Is Now, 2023.

25. Flanagin, A.; Bibbins-Domingo, K.; Berkwits, M.; Christiansen, S.L. Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. *Jama* **2023**, *329*, 637–639.

26. ChatGPT plugins, Year. Accessed April 13, 2023.

27. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D.; others. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education* **2023**, *9*, e45312.

28. Aizawa, A. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* **2003**, *39*, 45–65.

29. Qaiser, S.; Ali, R. Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications* **2018**, *181*, 25–29.

30. Ramos, J.; others. Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning. Citeseer, 2003, Vol. 242,1, pp. 29–48.

31. Trstenjak, B.; Mikac, S.; Donko, D. KNN with TF-IDF based framework for text categorization. *Procedia Engineering* **2014**, *69*, 1356–1364.

32. Wu, H.C.; Luk, R.W.P.; Wong, K.F.; Kwok, K.L. Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)* **2008**, *26*, 1–37.

33. Zhang, W.; Yoshida, T.; Tang, X. A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert systems with applications* **2011**, *38*, 2758–2765.

34. Tan, C.M.; Wang, Y.F.; Lee, C.D. The use of bigrams to enhance text categorization. *Information processing & management* **2002**, *38*, 529–546.

35. Hirst, G.; Feiguina, O. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing* **2007**, *22*, 405–417.

36. Hamed, A.A.; Ayer, A.A.; Clark, E.M.; Irons, E.A.; Taylor, G.T.; Zia, A. Measuring climate change on Twitter using Google's algorithm: Perception and events. *International Journal of Web Information Systems* **2015**, *11*, 527–544.

37. Kitsak, M.; Ganin, A.A.; Eisenberg, D.A.; Krapivsky, P.L.; Krioukov, D.; Alderson, D.L.; Linkov, I. Stability of a giant connected component in a complex network. *Physical Review E* **2018**, *97*, 012309.

38. Hamed, A.A.; Crimi, A.; Misiak, M.M.; Lee, B.S. Establishing Trust in ChatGPT BioMedical Generated Text: An Ontology-Based Knowledge Graph to Validate Disease-Symptom Links, 2023, [arXiv:cs.AI/2308.03929].