Annals of Library and Information Studies
Vol. 72, March 2025, pp. 32-41
DOI: 10.56042/alis.v72i1.11845

Scientific Productivity on ChatGPT: A Bibliometric Analysis

Sontu Nandi[a], Dipanjali Chakraborty[b], Amit Kumar Das[c] and Sabita Mandal[d]

[a]Senior Library Information Assistant, Central Library, Indian Institute of Technology Kharagpur, Paschim Medinipur, Pin- 721302, India. Email: nandisontu89@gmail.com

[b]Associate Engineer – Proccess Data SSW, Central library, Shell Business Operations Chennai, Chennai, Pin- 600100, India. Email: dipanjali2902@gmail.com

[c]Librarian, Central library, Bhatter College, Dantan, Paschim Medinipur, Pin 721426, India. Email: amitkumardas19@yahoo.in

[d]Independent Researcher, Paschim Medinipur, Pin-721451, India. Email: sabita773@gmail.com

Received: 15 June 2024; Accepted: 19 December 2024

Introduction: The discipline of Natural Language Processing (NLP) has experienced unprecedented advancements in recent years. Among these, OpenAI's ChatGPT has emerged as a frontrunner, captivating students, researchers and enthusiasts alike. As ChatGPT advances further, a need has arisen to judge, assess and understand the pattern and trajectory of scholarly contributions in the form of a 'Bibliometric Method'.

Motive: This article takes a bibliometric approach on 2302 scholarly publications related to ChatGPT from its inceptions year to 2023. It performed author productivity, citation analysis, keyword co-occurrence and productivity of journals and authors. It also performed various collaborative measures as well as Lotka's law of scientific productivity.

Methodology: Quantitative bibliometric analysis was chosen as the methodology for this research. Scopus was picked out to be the database to collect data. 2302 documents fulfilled the search query and thus, were chosen as the dataset for this research. Data refining and all the related works were performed in MS Excel. Vos-viewer and biblioshiny ware used to visualize the data.

Findings: after the analysis, it was found that most of the documents written over ChatGPT were articles, authors preferred collaboration over individual works, keywords i.e., artificial intelligence, large language models and ChatBot co-occur distinctively with ChatGPT, USA is the top productive country whereas Journal of Biomedical Engineering published most work over ChatGPT. It was also observed that the collaborative pattern of authors does fulfil 'Lotka's law of scientific productivity'.

Originality: As ChatGPT is comparatively a recently emerging concept, not a lot of bibliometric research has been performed on it. Thus, this research is one of the pioneers in ChatGPT-related bibliometric analysis and wishes to pave the way for future research.

Keywords: ChatGPT, Bibliometric Analysis, Author Productivity, Authorship Pattern, Keyword Co-occurrence

Introduction

The discipline of 'Natural Language Processing' (NLP) has experienced unprecedented advancements in recent years. Advanced language models have played a pivotal role in revolutionizing how humans interact with machines. Among these, OpenAI's ChatGPT has emerged as a frontrunner, captivating students, researchers and enthusiasts alike. As ChatGPT advances further, a need has arisen to judge, assess and understand the pattern and trajectory of scholarly contributions in the form of a bibliometric analysis.

Due to ChatGPT's tremendous development and influence, scholarly curiosity in its potential applications, constraints and capabilities has surged dramatically amongst scholars from an extensive spectrum of fields. As we traverse the dynamically emerging field of conversational AI, it is necessary to comprehend the philosophical foundations driving ChatGPT research for both researchers and professionals. This work, therefore, seeks to offer an in-depth perspective of ChatGPT-related research in 2022-23, by evaluating the bibliometric environment.

This work is a bibliometric analysis intending to find out the different types of documents written on ChatGPT since its emergence, along with the author distribution, keyword distribution, geographic landscape, etc. It also performs citation analysis along with finding the top 10 journals with the most ChatGPT-related content. Lastly, it conducts various collaborative measures over the documents concerning ChatGPT.

Literature review

(Chawla and Goyal, 2021)[1] undertook a bibliometric analysis of Digital Transformation research trends by analyzing 234 research articles and highlighted important findings from co-citation network analysis along with the general progressive tendency of publications year over year, authors' performance, publication journals, affiliated institutions and research-driving nations. (**Su, Peng and Li, 2021)[2]** conducted a bibliometric analysis over 3057 peer-reviewed papers distributed over some time from 2000-2019 and revealed that the United States was the top productive country with 23.73% of all publications and 32.25% of all citations. MIT was the top productive organization with 41 submissions and 1079 citations, the highest citations in the MLE discipline were found in the 'Journal of Machine Learning Research'. Based on their research, we found that the knowledge bases of MLE were developed by the following research issues i.e. random forests, support vector machines, extreme learning machines, deep learning, statistical learning theory and Python machine learning. (**El-Alfy and Mohammed, 2020)[3]** conducted a bibliometric study on machine learning for big data analytics using various kinds of bibliometric measures and visualizing strategies. The results showed that journal-to-conference publications had a one-to-two ratio, indicating overwhelming productivity and widespread applications in several multidisciplinary domains. Over two-thirds of the overall productivity in research was produced by three countries: the United States, China and India. (**Ho and Wang, 2020)[4]** performed a bibliometric study on 'Artificial Intelligence' research over 13,251 articles spanning from 1991-2018. The study revealed that 'Out of 119 countries, the USA ranked top in terms of citations per publication, total citations, international collaboration and single author papers. The three most prolific universities were Massachusetts Institute of Technology (MIT) in the United States, Islamic Azad University in Iran and Chinese Academy of Sciences in China'. (**Song and Wang, 2020)[5]** provided a bibliometric evaluation of approximately 8660 publications from 2000 to 2019 on worldwide educational artificial intelligence research development revealing trends in publication outputs, national collaboration, cluster analysis and research evolution. (**Guo et al., 2020**)[6] conducted a bibliometric investigation on the use of AI in healthcare and discovered that, while publication output has grown annually on average by 17.02% since 1995, but the rate of growth of research articles has increased dramatically, jumping to 45.15% between 2014 and 2019. Cancer, depression, Alzheimer's illness, heart failure and diabetes are among the most common health issues investigated by AI researchers. Also artificial neural networks, support vector machines and convolutional neural networks have the most influence on healthcare. (**Li et al., 2020)[7]** examined 5772 publications on deep learning publication with the help of bibliometric investigation and offered rudimentary knowledge of deep learning to scholars who are interested in this subject, as well as a decisive and meticulous study of deep learning for those who wish to conduct additional research in this area. (**Lopez-Martinez and Sierra, 2020)[8]** performed a bibliometric investigation in 'Natural Language Processing' (NLP) related publications over 2010-2019 and revealed that Research in the field and subfields has risen steadily throughout the investigation period; the proceedings of conferences are the primary medium for communicating results; and biomedical informatics is one relevant sector of application of NLP. They concluded with a synchronic and a diachronic characterization of research topics conducted globally on natural language processing and related topics, demonstrating the close relationship that has recently developed between several artificial intelligence subfields and natural language processing. (**Bhattacharya, 2019)[9]** tried a bibliometric investigation of key areas in machine learning research, identifying research trends and the field's intellectual structure using quantitative and statistical analysis of research papers. (**Gupta and Dhawan, 2018)[10]** examined total 1,52,655 publications published worldwide in the field of artificial intelligence research between 2007 and 2016, which came from the Scopus database. The study also looked at papers on artificial intelligence research produced in India. Over the course of ten years, from 2007 to 2016, India amassed 9730 publications overall, with an average annual growth rate of 27.35%. The average citation impact per paper was 2.76 and the country's share of international collaborative publications during this time was 10.34%. (**Barakhnin et al., 2018)[11]** determined areas of high interest (Grammar checking, information extraction, machine translation and question answering) and low interest (information retrieval, opinion mining and text categorization) by analysing the bibliometric indicators of a rapidly emerging field

of study as automatic text processing. (**Chen et al., 2017)**[12] carried out a bibliometric analysis on 1405 medical research papers using NLP that were published over a ten-year period. They found that 18.39% is an average yearly growth rate where the top 10 publication sources were responsible for over 50% of all publications. They also found Computational biology, terminology mining, information extraction, text categorization, social media as a data source, information retrieval, etc. were the ten major subject areas where the largest number of publications came from the USA.

Objective

This study aims to:

1. Study the different kinds of documents written over ChatGPT
2. Analyse the collaboration, i.e., authorship pattern and country collaboration
3. Perform a citation analysis on the documents chosen for this research
4. Scrutinize the occurrences of keywords indexed
5. Explore the productivity of countries associated with the research
6. Explore the productivity of journals
7. Find the ratio of documents that come with DOI and without DOI
8. To analyse the different collaborative measures
9. Find out whether the dataset follows the 'Lotka's law of scientific productivity'.

Methodology

Quantitative bibliometric analysis was chosen as the methodology for this research as per fig. 1. Scopus was picked out to be the database to collect data from. In the 'Search within' section, 'Keywords' field was chosen. After that, in the 'Search Documents' section, 'ChatGPT' OR 'CHATGPT' OR 'chatgpt' phrase was used as the search query. 2302 documents fulfilled the search query and thus, chosen as the dataset for this research. Data was then exported in the excel file. Data refining and all the related works were performed in MS-Excel. Vos-viewer and Biblioshiny were used to visualise the data.

OpenAI launched ChatGPT on 30 November 2022, since this is the field which have quickest rate of growth on worldwide. As a result, Scopus retrieved a total of 2302 papers dispersed across the years 2022 and 2023, each with 4 and 2298 records, respectively.

Data analysis

Different types of documents authored over ChatGPT

To perform this task, only the 'Document type' column was separated from the Scopus file and put in a separate excel file. Later, the 'countif' and 'sort' function were used to find out the different types of documents and their occurrences.

Figure 2 displays the frequency of different document types. Articles are most common (979), followed by Letters (386) and Conference papers (308). Books and Book chapters have fewer occurrences (2 and 14, respectively), while other types like editorials, reviews and notes also contribute to the collection.

Analysis of collaboration

At first the data were analysed based on single and multiple authors. It was found that out of 2302 documents, 568 documents were written by single authors whereas a total of 1734 documents were written by multiple authors.

From the fig.3, it is clear that documents with single author amounts to only 25% of the total dataset, whereas, documents with multiple authors

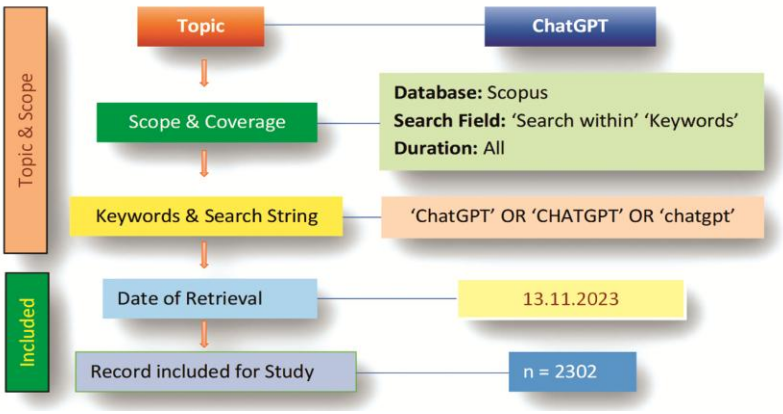

Fig 1 — Topic, Scope and Population of Study

adds up to an extensive amount of 75% of the total dataset. This emphasizes the collaborative nature of research, highlighting the significance of teamwork and diverse contributions in the creation of scholarly works within the represented field.

Authorship pattern

This section examines how publications are distributed and how authorship is distributed in the documents that made available on ChatGPT. The cardinality of co-authors has been used to indicate the authorship distribution of works.

From fig. 4, it is clear that most of the documents are authored by single authors (24.67%), followed by

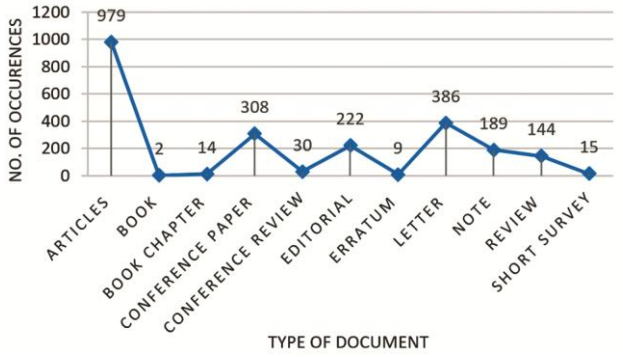

Fig 2 — Different types of documents and their occurrences on ChatGPT

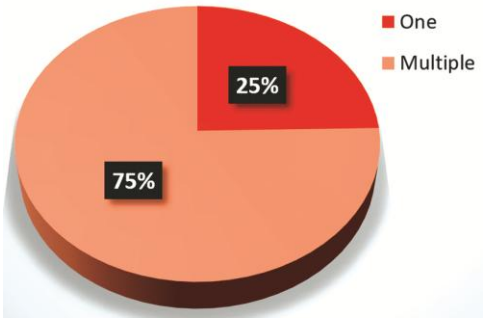

Fig 3 — Graphical representation of author collaboration on ChatGPT

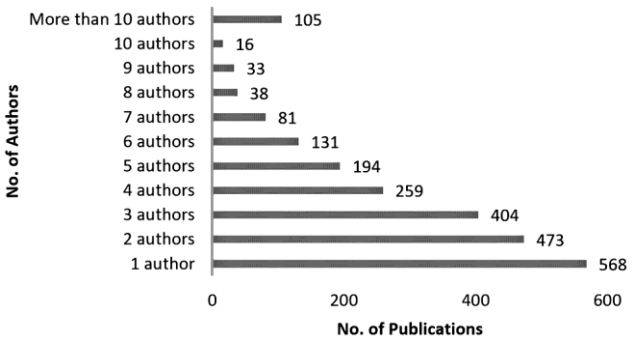

Fig 4 — Graphical representation of authorship pattern

2 (473) and 3 (404) authors, whereas a least number of documents are authored by ten authors (0.69%). Following section (table 1) finds out the authorship distribution over the span of two years i.e., 2022-2023, using the agglomeration of data.

Table 1 shows, in 2023, collaborative efforts increased, with 1, 2 and more than 10 authors showing growth. Notably, 2022 had only 4 publications, all with 1 to 3 authors. The shift implies a trend towards more extensive teamwork and possibly interdisciplinary research in 2023 compared to the previous year.

Country Collaboration

The map (fig. 5) provides insights into international research collaborations. Notable patterns include Australia engaging in diverse partnerships, especially with Hong Kong (6), France (5) and Singapore (5). Brazil collaborates with Belgium, Chile, and Egypt. Austria has collaborations with Belgium and Sweden. These findings reflect a globalized landscape of scientific cooperation, fostering knowledge exchange and diverse perspectives. The data underscores the interconnected nature of scientific endeavours, transcending geographical boundaries for the advancement of knowledge and innovation on a multinational scale.

Citation analysis

The citation analysis was performed with the help of MS-Excel. Here, 'C' represents the no. of citations received by an article.

Table 2 and figure 6 show the distribution of documents based on the number of citations received. The majority (1301) have no citations. As the citation count increases, the number of documents decreases,

Table 1 — Tabular representation of authorship distribution on ChatGPT over the year 2022-23

Serial no.	No. of authors	Year 2023	Year 2022
1	1 author	565	3
2	2 authors	473	0
3	3 authors	403	1
4	4 authors	259	0
5	5 authors	194	0
6	6 authors	131	0
7	7 authors	81	0
8	8 authors	38	0
9	9 authors	33	0
10	10 authors	16	0
11	More than 10 authors	105	0
	Total	2298	4

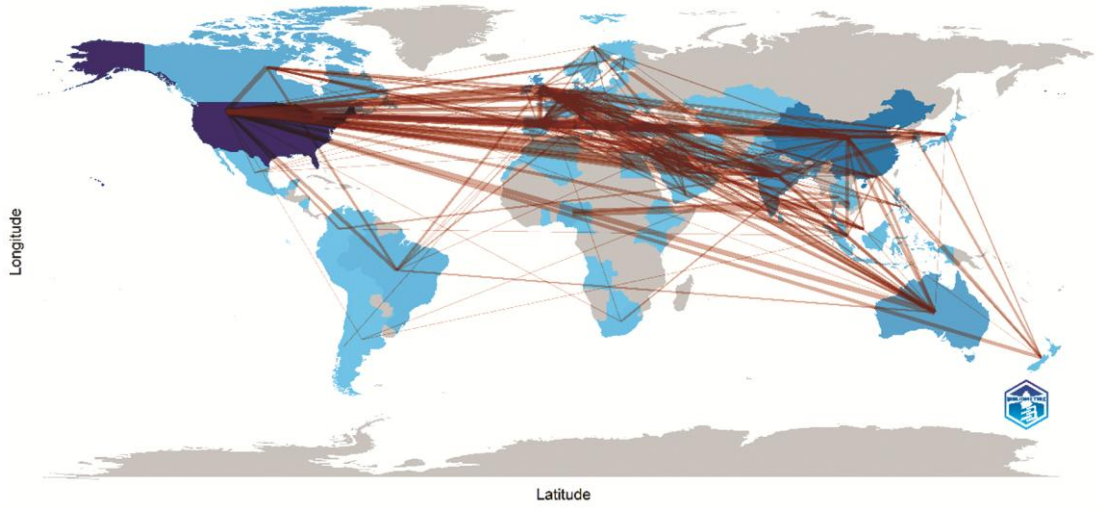

Fig 5 — Country collaboration map

Table 2 — Tabular representation of citation analysis of documents on ChatGPT in 2022-23

Serial no.	No. of citation(s) received	No. of documents
1	C=0	1301
2	10 ≥C ≥1	812
3	20 ≥C ≥11	101
4	30 ≥C ≥21	30
5	40 ≥C ≥31	18
6	50 ≥C ≥41	7
7	100 ≥C ≥51	17
8	300 ≥C ≥101	16

indicating a skewed distribution. Only a few documents have higher citation counts, suggesting that a small percentage of works receive significant attention, while a large portion remains less cited, reflecting the common pattern in scholarly literature.

From the above analysis, it is evident that most (56.5%) of the publications has received 0 citation. Whereas, very few documents (0.3%) have received more than 41 but less than 50 citations.

Occurrence of keywords

To perform this analysis, VOSviewer was used. VOSviewer puts special emphasis on constructing visual illustrations of these bibliometric networks, assisting scholars to understand the relation within a set of scientific publication.

Figure 7 presents word occurrences in a corpus. "ChatGPT" dominates with 1729 mentions, followed by "artificial intelligence" (792) and "large language models" (203). Notably, variations like "generative AI" and "chatbot" appear, emphasizing the diversity and prevalence of topics, including education, ethics, and AI ethics.

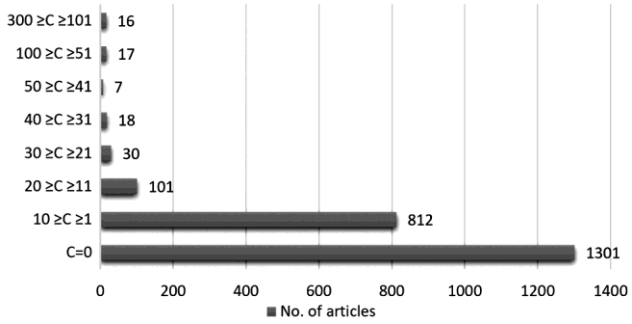

Fig 6 — Citation accumulation pattern

Country wise productivity

For this analysis as well, VOSviewer was used. Data were simply put in the VOSviewer from which it extracted relevant data to perform the following analysis. Result of the analysis is shown below.

Figure 8 indicates publication frequency by country. The USA leads with 2189, followed by China (921) and India (427). This reflects the global distribution of research output, showcasing significant contributions from diverse regions such as Europe, Asia, and the Middle East. The data underscores the international collaboration and varied research landscapes across countries, contributing to a comprehensive and inclusive knowledge ecosystem.

Journal productivity

In order to perform this task, Biblioshiny software was used. Finally, top twenty journals with maximum number of documents published over ChatGPT is represented below.

Figure 9 displays publication sources and their associated article counts. Notable sources include

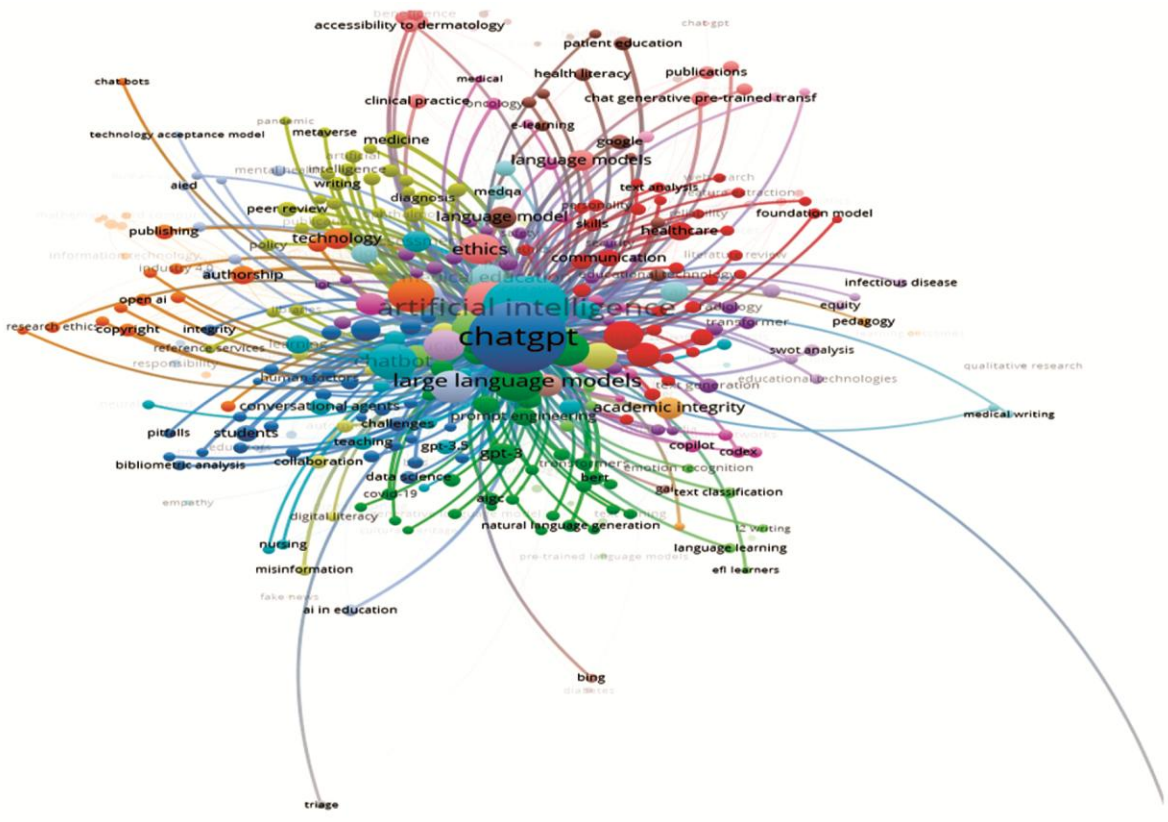

Fig 7 — Keyword Co-occurrence of publications

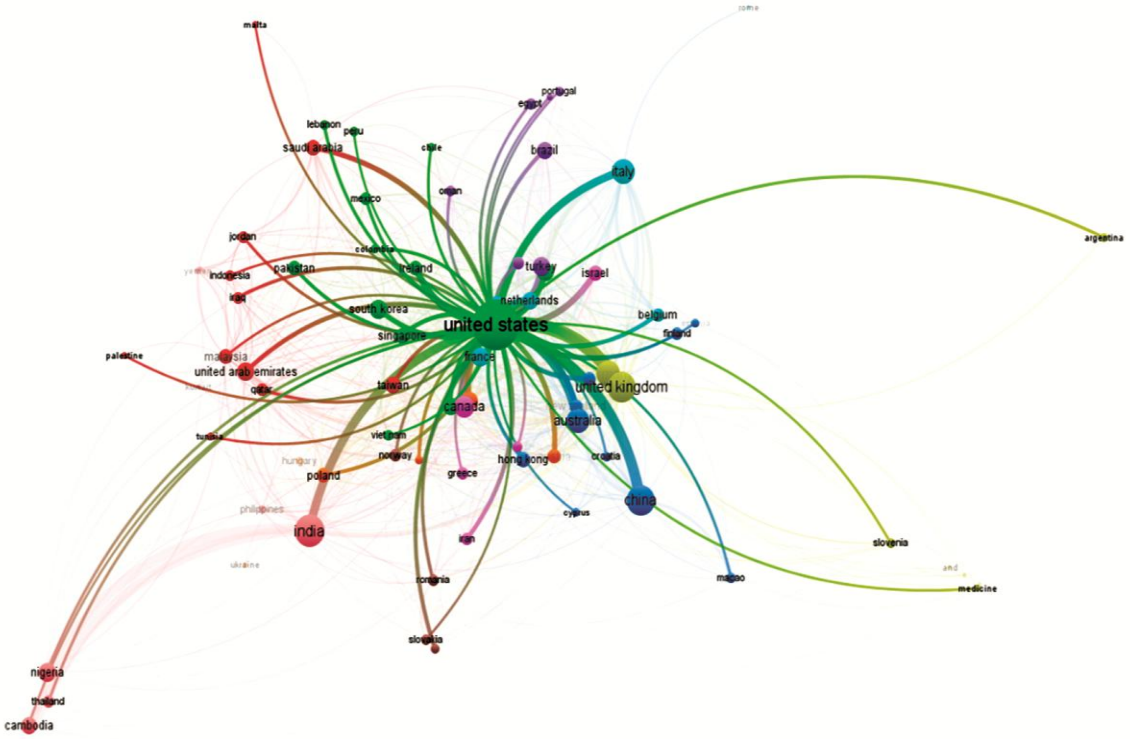

Fig 8 — Country wise productivity and collaboration

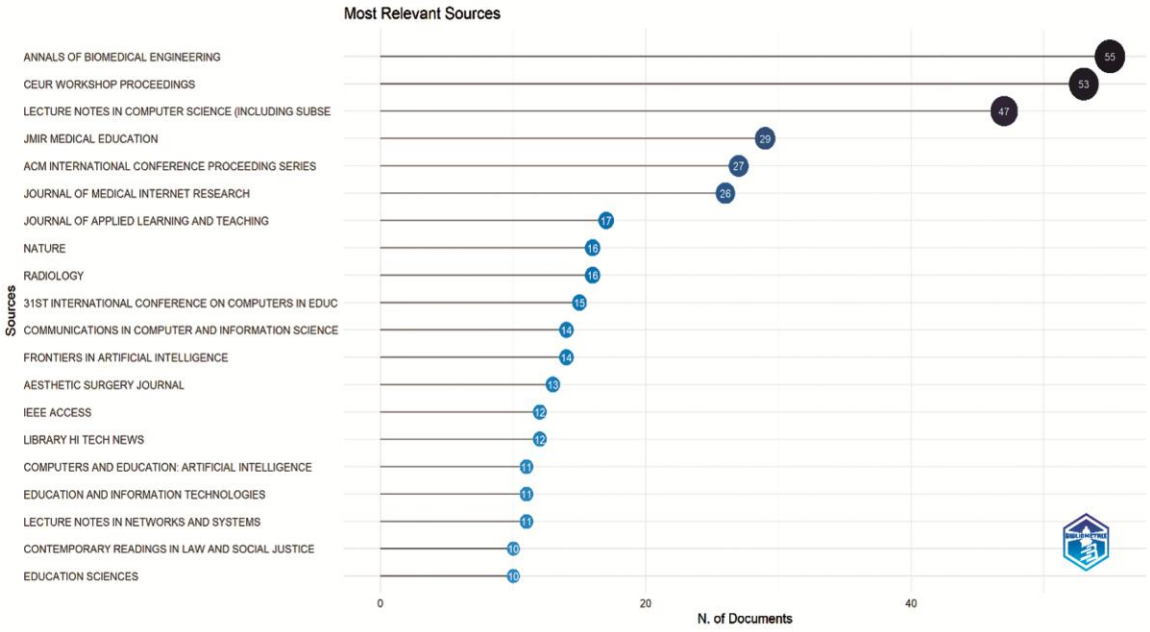

Fig 9 — Top 20 journals and their number of documents published over ChatGPT

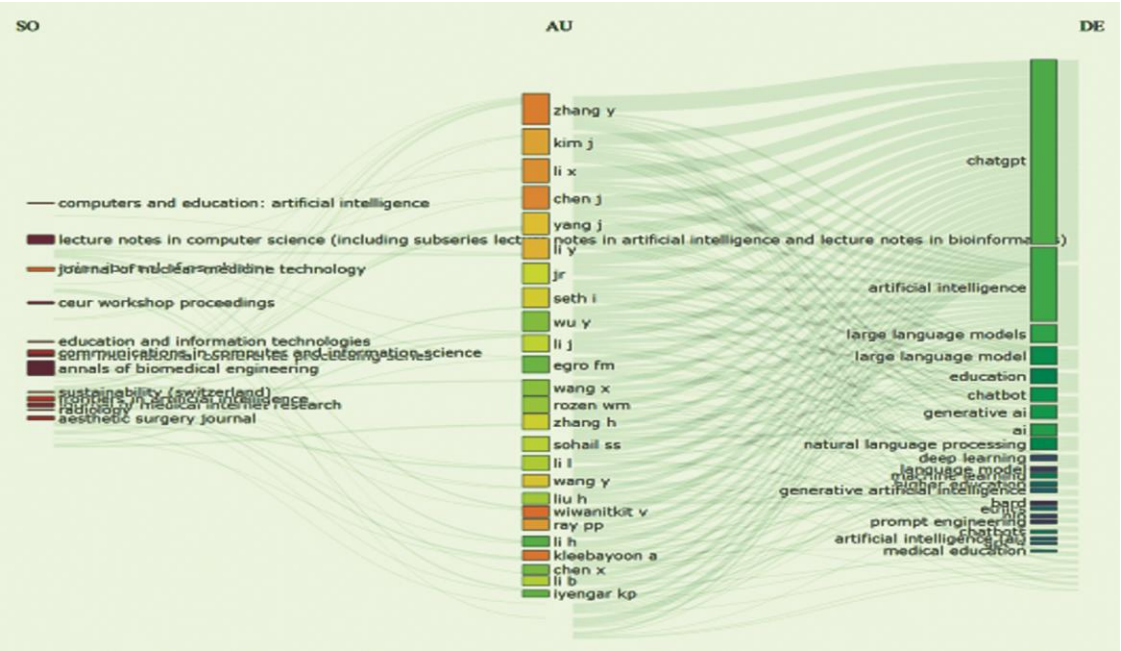

Fig 10 — Representation of Three-Field Plot (journals, authors & keywords) (*SO = Source; AU = Author; DE = Keyword)

"Annals of Biomedical Engineering" with 55 articles, "CEUR Workshop Proceedings" with 53 and "Lecture Notes in Computer Science" with 47 publications. The diversity of sources, ranging from medical journals to AI conferences, indicates a multidisciplinary approach in the research field, emphasizing collaboration and knowledge integration across various domains.

The top twenty “Three Field Plot” (fig. 10) visualisation technique also applied through R software in respect of source journals (SO), authors (AU) and keywords (DE) field and saw that, ‘Annals of biomedical engineering’ journal, author ‘Zhang, Y’ and the keyword ‘ChatGPT’ are most productive respectively. One thing is also prominent that all of the top authors in this field deal with artificial intelligence.

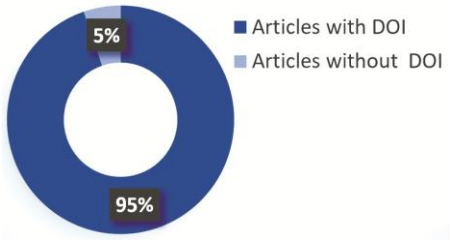

Fig 11 — Availability of documents with DOI

Document availability with DOI

Figure 11 shows that 2181 (95%) documents have DOIs, enhancing their accessibility and credibility. However, 121 (5%) documents lack DOIs, potentially affecting visibility and citation accuracy.

Various collaborative measures

Collaborative Index (CI)

This indicator is defined by Lawani. It represents mean number of authors per paper.

$$CI = \frac{\sum_{j=1}^{h} jf_j}{N}$$

Here, fj = the number of papers having j authors in the collection;
h = the maximal number of authors in a single paper;
N = the total number of papers;
n = total number of authors in the collection.

Using the above-mentioned formula, collaborative index of each year was calculated, which is shown below.

For year 2022:
(1x3 + 3x1)/ 4
= 6/4 = 1.5

For year 2023:
(1x565 + 2x473 + 3x403 + 4x259 + 5x194 + 6x131 + 7x81 + 8x38 + 9x33 + 10x16 + 11x105)/ 2298
= (565+946+1209+
1036+970+786+567+304+297+160+1155)/2298
=7995/2298 = 3.48

From the above calculations, it is evident that CI of 2023 is almost 7 times of the CI of 2022.

Degree of collaboration (DC)

Collaboration helps to assess the structure and trend of scientific research and collaborative activities. Degree of collaboration, (DC) helps us to quantify the multi-authored manuscripts. This index is mathematically represented as,
DC = Nm/Nm + Ns
where,
DC = Degree of collaboration;
Nm = Number of multiple-authored Papers;
Ns = Number of multiple-authored Papers.

Using the above-mentioned formula, degree of collaboration of each year was calculated, which is shown below.

For year 2022:
Nm / Nm + Ns
= 1/ 1+3 = ¼ = 0.25

From the above calculations, it is evident that DC of 2023 is almost 3 times of 2022.

For year 2023:

Total no. of multiple authors: 1733

Therefore, degree of collaboration (DC) = 1733/ 1733+ 565 = 1733/2298 = 0.75

Collaborative Co-efficient (CC)

It is defined by Ajiferuke et al. Advantages of both CI and DC can be observed in CC. This index possesses a value between 0 and 1. It usually equals to 0 since single-author papers dominate. This index is mathematically represented as follows:

$$cc = \frac{1 - \sum_{j=1}^{h} (1/j)_j^f}{N}$$

Using the above-mentioned formula, collaborative coefficient of each year was calculated, which is shown below.

For year 2022:
1- {3+ 1 x (1/3)}/4
=1- 0.83
=0.17

For year 2023:
= 1- {(565 + 473 x (1/2) + 403 x (1/3) + 259 x (1/4) + 194 x (1/5) + 131 x (1/6) + 81 x (1/7) + 38 x (1/8) + 33x (1/9) + 16 x (1/10) + 105 x (1/11)} / 2298
= 1- {565 + 236.5+ 134.3+ 64.75 + 38.8+ 21.83+ 11.57+ 4.75+ 3.67+ 1.6+ 9.54}/2298
= 1- 0.475
= 0.525

Lotka's law of scientific productivity

'Lotka's formula for scientific productivity' is as follows:

$$Y = \frac{C}{X^n}$$

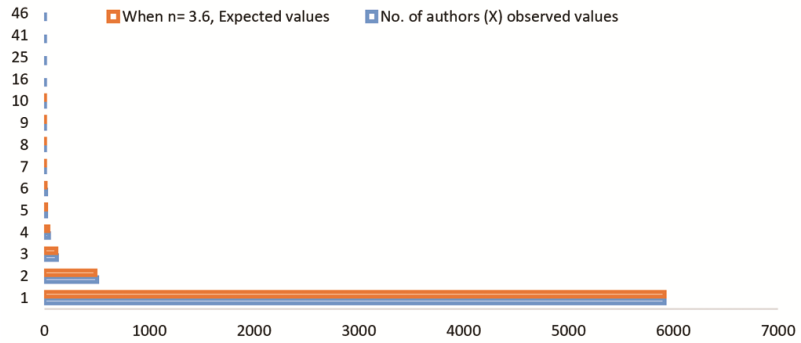

Fig 12 — Graphical representation of Observed and expected values of author as per 'Lotka's law of scientific productivity'

Table 3 — Observed and expected values of author as per Lotka's law of scientific productivity

No. of articles	No. of authors (X) observed values	When n = 3.6, Expected values
1	5916	5916
2	504	487
3	121	113
4	44	40
5	18	18
6	15	9
7	6	5
8	4	3
9	4	2
10	1	1
16	1	0
25	1	0
41	1	0
46	1	0

Where,
X = Number of Publications
Y = Relative Frequency of Authors with X publications
C = Constants depending on the specified field
Putting the value in above equation,
X=1, Y= 5916
We get, $5916 = C/ 1^n$

Therefore, C= 5916
Again, putting the value for X=2, Y= 504,
We get,
$504 = 5916/2^n$
$2^n = 5916/504$
$2^n = 11.74$
Therefore,
n log 2 = log 11.74
or, n = 3.6

It is clear from the analysis (table 3 & fig. 12) that the observed and expected values of the authors are nearly comparable. Therefore, it can be concluded that 'Lotka's law' is followed by the ChatGPT dataset.

Conclusion

In summary, the bibliometric analysis performed has shed valuable insight over the scholarly environment concerning ChatGPT. The exponential rise of research output, as reflected by the growing number of publications emphasises the significance and the widespread interest of this sophisticated language model. The evaluation of authorship trends, citation analysis and collaborative networks has shed light on the collaborative nature of the research community engaged with ChatGPT.

For researchers, industry professionals interested to stay up-to-date with the recent developments regarding ChatGPT, this paper intends to act as a roadmap as we navigate through the rapidly developing terrain of Artificial Intelligence (AI) and Natural Language Processing (NLP). As we advance, further research and collaboration will surely take the field to new heights, enabling ChatGPT to reach its full potential for a variety of uses and enhancing society as a whole.

References

1 Chawla R N and Goyal P, Emerging trends in digital transformation: a bibliometric analysis, *Benchmarking: An International Journal*, *29*(4) (2022) 1069-1112.
2 Su M, Peng H and Li S, A visualized bibliometric analysis of mapping research trends of machine learning in engineering (MLE), *Expert Systems with Applications*, *186* (2021) 115728.
3 El-Alfy E S M and Mohammed S A, A review of machine learning for big data analytics: bibliometric approach, *Technology Analysis & Strategic Management*, *32*(8) (2020) 984-1005.
4 Ho Y S and Wang M H, A bibliometric analysis of artificial intelligence publications from 1991 to 2018, *COLLNET Journal of Scientometrics and Information Management*, *14*(2) (2020) 369-392.
5 Song P and Wang X, A bibliometric analysis of worldwide educational artificial intelligence research development in recent twenty years, *Asia Pacific Education Review*, *21* (2020) 473-486.

6 Guo Y, Hao Z, Zhao S, Gong J and Yang F, Artificial intelligence in health care: bibliometric analysis, *Journal of Medical Internet Research*, *22*(7) (2020) e18228.

7 Li Y, Xu Z, Wang X and Wang X, A bibliometric analysis on deep learning during 2007–2019, *International Journal of Machine Learning and Cybernetics*, *11* (2020) 2807-2826.

8 Lopez Martinez R E and Sierra G, Research trends in the international literature on natural language processing, *Journal of Scientometric Research, 9(3) (2000) 310-318.*

9 Bhattacharya S, Some Salient Aspects of Machine Learning Research: A Bibliometric Analysis, *Journal of Scientometric Research*, *8*(2) (2019) 85-92

10 Gupta B M and Dhawan S M, Artificial Intelligence Research in India: A Scientometric Assessment of Publications Output during 2007-16, *DESIDOC Journal of Library & Information Technology*, *38*(6) (2018) *416-422*.

11 Barakhnin V B, Duisenbayeva A N, Kozhemyakina O Y, Yergaliyev Y N and Muhamedyev R I, The automatic processing of the texts in natural language Some bibliometric indicators of the current state of this research area, *Journal of Physics: Conference Series*, 1117 (1) (2018)

12 Chen X, Xie H, Wang F L, Liu Z, Xu J and Hao T, A bibliometric analysis of natural language processing in medical research, *BMC medical informatics and decision making*, *18*(1) (2018) 1-14.