Concetta Press
Journal of Biomedical Informatics and AI

Detecting ChatGPT in Scientific Text:

A Historical Perspective on Chatbot Catchphrases

Edward J. Ciaccio PhD

Department of Medicine Columbia University College of Physicians and Surgeons, New York, NY, USA

Article Type	Historical Perspective
Volume / Issue	1 / 1:1 (2026)
DOI	10.5281/zenodo.18983792
Correspondence	Edward J. Ciaccio, PhD Columbia University, HP 9-956, 180 Fort Washington Avenue, New York, NY 10032
Disclosure	Funding: None. Conflicts of Interest: None.

An earlier version of this manuscript briefly appeared online in 2024 but was subsequently withdrawn by the publisher. The present article represents a revised and expanded version.

Abstract

Large language model (LLM)-generated text, beyond grammar and spelling improvements, may be present in scientific documents even without attribution. Hence, development of an accurate manual screening paradigm would be helpful to detect whether LLM, or chatbot, content is included. In this study, examples are provided to suggest how peculiar catchphrases detected in scientific text can be checked for chatbot generation. Catchphrases were first selected manually, and then Google Search was utilized in seeking articles with each catchphrase up to 2024 to show a historical perspective. Thereafter, it was determined whether paragraphs with the catchphrase were of chatbot origin using the GPTZero detector. The number of articles published with each catchphrase in recent literature, citations per article, the publishing journal Impact Factor, and the document section in which the chatbot phrase appeared were compiled to characterize chatbot phrasing. In this investigation, it was found that most suspected peculiar phrasings were indeed chatbot-associated. Based on a statistical analysis, the number of published biomedical and bioengineering articles with chatbot phrasings has increased substantially in recent years, particularly after the onset of ChatGPT. Moreover, most of the chatbot-containing articles studied were published in journals with a substantial Impact Factor. Chatbot-generated text is most commonly found in the Abstracts and Introduction sections, but also in the Methods, Results and Discussion, Limitations, and Conclusions. It is therefore inferred that Chatbot content has peculiar phrasing that can be detected manually, and that it ubiquitously appears in scientific documents in the peer-reviewed literature. Chatbot-generated content, over and beyond simple grammar and spell correction, is present and has been increasing even in top-ranked journals.

Keywords: artificial intelligence, chatbot, ChatGPT, detection, GPTZero

Introduction

“Chatbots” or “chatterbots” refer to computer programs that can be integrated into various platforms, such as messaging apps, websites, and virtual assistants, to simulate human-like conversations, and to help automate tasks [1]. Productivity is an important motivation for using them, and they are also employed for entertainment and socialization [2]. In business, chatbots can be implemented as a cost-cutting measure, such as by enabling the addressing of many customers simultaneously. The term “chatbot” is not new; early on it mostly referred to machine interactions with humans in “chatrooms”. Such early automated methods were maliciously useful for social

engineering, intelligence gathering, mounting phishing attacks, spamming and malware, and for reducing the usability and security of collaborative communication platforms [3]. The design of potential interrogation strategies to assist chatroom administrators in rooting out rogue chatbots was of high importance [4, 5]. Then and now, there is a growing demand for methodologies, tools, and strategies for flagging chatbot text [3]. Quantitative analysis of communication patterns can be helpful in detecting the "bots", with human text being distinguishable from chatbot based on communication pattern differences [3]. Recently, there has also been a high degree of concern for the problem of automatically detecting chatbot text on livestreaming platforms, since it can negatively influence recommendations, harm streamer and viewer trust in the platform, and reduce monetization for honest streamers [6]. For dynamic systems, chatbots can be distinguished from human interaction based on increased site activity volume, and a longer and more consistent delay in messaging that originates from the chatbot.

Lately, chatbots have become employed - asked, one might say - to write scientific text for journal articles, beyond the copyediting functions of grammar and spell correction. The onset of this era can be pinned to the introduction of Generative Pre-trained Transformers (GPT) by the OpenAI organization (https://chat.openai.com/), initially as GPT-1 in June 2018, GPT-2 in February 2019, GPT-3 in June 2020, and the ChatGPT conversational agent in November 2022. A GPT is a large-learning model or LLM, which refers to a deep learning neural network containing many network layers that are trained on vast amounts of data. Such deep learning models may include 1000 or more layers [7]. Although ChatGPT is the most well-known GPT, other companies besides OpenAI, and other countries besides the USA, have since developed similar forms of artificial intelligence, with names such as Bard and Claude [8]. Unlike the chatbots encountered in chatrooms and streaming services, which are dynamic, when an interactive chatbot is utilized to write an article, the information generated is forever contained within the article. Hence, detecting chatbot in written articles can be strategized based on text validity and on communication pattern differences between machine-based and human-based text, and specifically, in any peculiar phrasing that may be inimitable to the chatbot.

Chatbots utilized in biomedical and bioengineering research investigations are often employed as intended, i.e., as an assistive device to provide information, which can thereupon be utilized in scientific studies. They can increase work productivity by informing the scientist much more rapidly than a traditional online or paperback book search. However, academia is now witnessing an abuse of authorship in academic papers, with text generated by LLMs without attribution [9]. Authors who use chatbots in such ways may state that they are being employed to enhance their writing. However, even if original text was written by a human author, the alterations to their text may go well beyond copyediting for spelling and grammar correction, to instead formulate content that is unrecognizable from the original and thought by the machine, not by the human user. Such text requires attribution, just as when any paper-writing service is utilized [10]. Chatbots are thereby testing the limits of traditionally established publishing ethics [10]. Detection of chatbot text in academia can thus be deemed essential, and this is primarily the responsibility of editors and journals/publishers [9], although all have a responsibility to prevent unattributed usage in published documents. Hence, it is important to understand chatbot phrasing and its characteristics. Complicating the matter however, chatbot style will depend in part on the human-chatbot interaction, including the quality of the query posed to the chatbot, and the type of editing or paper writing requested. Ideally, human-chatbot interaction should be done with care and etiquette [10].

To develop this investigation, it was hypothesized that to the human reader, ChatGPT has a distinctive and peculiar writing style. If ChatGPT's phrasing is indeed peculiar to the human reader, it might therefore be possible to manually detect its presence in article text with high accuracy. Thus herein, examples of ChatGPT-associated phraseology are provided, and a possible means to detect its origin is shown. The study suggests how manual screening for chatbot is possible, and how it may have practical utility to confirm and validate machine-based chatbot detection [11].

Methods

Chatbot phrasing was investigated as follows. Using a series of four sample documents, portions of which contained some chatbot-generated text, places in the text where a chatbot origin was surmised to be present were manually

highlighted for peculiar phrasings. In the four exemplar documents, N=93 such phrasings were found. Google Scholar was then used to search for articles with the precise phrasings, utilizing double quotation marks to do so. With an initial assessment period of January 2023- March 2024, for articles appearing in Google Search, paragraphs of the searched articles containing the catchphrase were checked for chatbot, and the catchphrase was marked as having a chatbot origin if any of the paragraphs were flagged. For this study, GPTZero (https://gptzero.me/) was utilized to determine whether a phrase thought to be of chatbot origin was actually chatbot, and was employed for its ease-of-use, popularity, low cost, and relative accuracy compared to many other chatbot detectors [12, 13].

Text Processing

To preprocess before checking with GPTZero, some extraneous content was first removed from the text. This was done by converting the paragraph with a chatbot-suspected phrase into plain text, and removing all line breaks and tabbing and replacing them with single spaces, all of which was done semi-automatically in Microsoft Word 365 Apps for Enterprise. Said text was then presented to GPTZero. If GPTZero noted that it was more likely to be of GPT origin (displayed as a yellow screen and yellow highlighted text), it was recorded as such in the analysis. Sometimes GPTZero found portions but not all phrasing in a paragraph to be of chatbot origin. In such instances, the paragraph was still marked as being of chatbot origin.

If GPTZero was unsure of a paragraph's origin, additional preprocessing measures were taken; specifically, when the text included citations, all citation components were then removed by hand. This was an important step for analyzing some texts. For example, a 176-word paragraph, which included a possible catchphrase, had no citations except for this one embedded in it, shown in generic form:

(Xxxx, et al., 20xx, Xxxx, et. al., 20xx, Xxxx, et. al., 20xx)

where Xxxx represents author surnames and 20xx represent the year. With this single citation contained in the paragraph, GPTZero found the probability of it being AI generated was 43%, and it was uncertain how to classify it, but more likely to be of human origin. Removing the single reference above, GPTZero increased the likelihood of AI generation to 97%, and gave it a yellow flag, and the system was highly confident of it being AI generated. Hence, this very important step of removing referencing was taken if the paragraph was initially classified by GPTZero to be of human origin or if it was uncertain its origin.

An additional step, if needed, was to add additional text to the chatbot paragraph text (add the prior or following paragraph from the original manuscript). If GPTZero still did not detect chatbot, or was unsure, the phrasing was marked as unlikely to have a chatbot affiliation, and it was not further analyzed. It was surmised that any such text in which GPTZero was unsure of the origin, even after complete preprocessing, might for example still have been originally generated by chatbot, but could have been thereafter heavily edited by a human.

Further Statistical Analysis

A subset of the chatbot-flagged catchphrases was then selected for further example statistical analysis. For N=50 such chatbot-associated catchphrases in which catchphrase paragraphs were flagged in at least 3 articles, the following metrics were recorded: the number of articles containing each catchphrase that were published for the time intervals 2012-2014, 2015-2017, 2020-2022 (after GPT introduction), and 2023-2024 (after ChatGPT introduction), the total citations per article, the publishing journal impact factor, and the document section in which the chatbot paragraph appeared. It was hypothesized that from GPT introduction in 2018, and particularly since conversational chatbot ChatGPT was introduced and had attained over 100 million users worldwide in 2023 [14], an excess of articles would contain chatbot-associated phrasings. The endpoint, mid-March 2024, marked the end of the quantitative analysis period for this study and was used for historical perspective.

Catchphrases

The catchphrases investigated in the test documents were ones that appeared to be peculiar and were therefore suspected of having a chatbot origin. However, here is an example in which subsequent confirmation of a chatbot origin to be used in additional statistical analysis failed:

whose prowess in live fluorescence imaging

"Prowess" can be considered a peculiar word, rarely used by humans in scientific text. However, using the entire 6-word phrase, no articles appear in Google Search for any time period. Hence, shorter, root components were tested. The phrases "prowess in live", and separately "whose prowess in", which included the peculiar word, were checked in Google Search. The phrase "prowess in live" resulted in only one entry for all time on Google Search, year 2022, so it could not be compared for the four stated time intervals. The phrase "whose prowess in" appeared in articles on several pages of Google Search for each of the four time intervals aforementioned. It is therefore a candidate chatbot catchphrase. However, many of the articles associated with this catchphrase were unsearchable due to the presence of a paywall, or due to the lack of a searchable text version (for example, it being shown in image form rather than text). None of the searchable entries were flagged by GPTZero, and hence, the phrase was not further investigated.

Another peculiar, possibly chatbot-originating phrase that was noticed is: "captures a broader", in which subsequent confirmation of a chatbot origin succeeded. Articles with this catchphrase occurred in 14 pages' worth of Google Search results for each time interval searched, where each page contains 10 articles, for a total of 140 such articles. 140 articles is the maximum results displayed in Google Search, according to the standard settings, and there were likely many more such articles actually published with this catchphrase during the search intervals. Therefore, "captures a broader" is a common phrasing in the scientific literature, both pre- and post-GPT. Since 3+ articles were found positive by GPTZero for containing chatbot in the time interval 2023-2024, this phrase was utilized for additional quantitative analysis.

Overall paradigm

Figure 1, the graphical Abstract, exhibits the overall paradigm for the example analysis. N = 93 catchphrases which seemed odd and unusual were studied. A Google search was done for each phrase. When found, the paragraph with the catchphrase was checked in GPTZero. If chatbot was not detected, the paragraph was preprocessed and rechecked with GPTZero (loop 1). This continued until all preprocessing was done. If GPTZero still did not detect chatbot, the phrase was marked as not being chatbot (step 2). If chatbot was detected in the catchphrase paragraph by GPTZero, the catchphrase was marked as having a chatbot association. Furthermore, the Google Search results were then

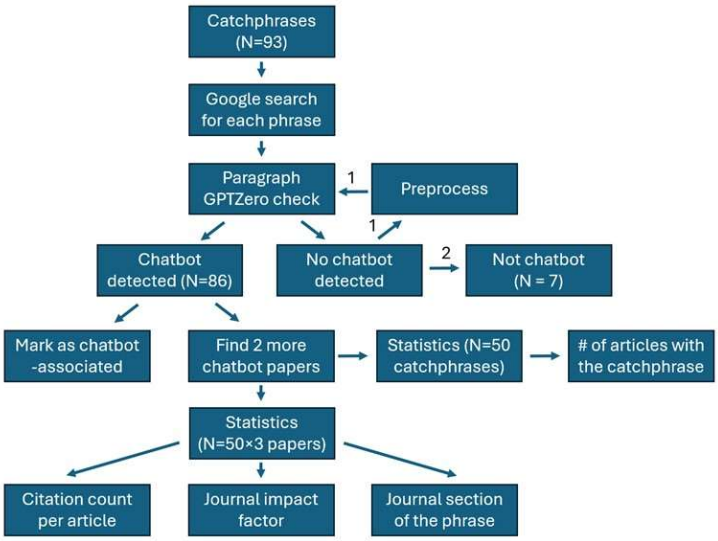

Figure 1 Schematic of the paradigm used for quantitative analysis of suspected chatbot catchphrases. A Google search was done for each of 93 odd phrasings found in example documents with some sections having chatbot origin. Paragraphs from the search results with the phrasing were checked with GPTZero. If a paragraph with the phrasing was flagged, the phrasing was marked as being a chatbot catchphrase. If two other such instances were found among the Google search results, the statistics of the three instances was determined (total of N = 50 × 3 = 150).

examined to seek two more instances in which the catchphrase paragraph was found positive for chatbot by GPTZero. If this criterion was met, the catchphrase was utilized as 1 of N = 50 such catchphrases in which further statistical analysis was done. The number of articles with the catchphrase, and for the three chatbot associated articles, interval January 2023 - March 2024, the mean citations per article, Journal impact factor, and journal section with the catchphrase were determined to show example characteristics for this report.

Results

Table 1 provides a list of chatbot-associated (N = 86) articles of the N = 93 suspicious catchphrases analyzed. The list is alphabetized. An association was made if at least one Google Search entry in which the catchphrase paragraph was flagged by GPTZero was found. All of the initial phrasing studied is shown. The bold text conveys the final phrasing searched in which GPTZero subsequently flagged the catchphrase paragraph for the N = 86 that were chatbot-associated. The text in brackets are additional words from the original document added to the search, not originally included, which became part of a chatbot-associated catchphrase that were used for the search. Table 1 demonstrates the peculiar dialect of GPT. Non-associated phrasing (N = 7) consisted of: amidst this brilliance, an innovative approach, assuming the mantle, at its helm, case further attests, sage guidance of, whose prowess in live fluorescence imaging.

Subsequently, N=50 of the 86 chatbot-associated phrasings in which two more GPT-affiliated articles were found were further analyzed, and then ordered based on the total number of Google Search pages delivered for each phrasing, all time periods. Table 2 provides information for articles with ≳ 10 total Google Search page results (100+ articles) for all time periods combined. The chatbot phrase is noted in the left-hand column. The subsequent columns show how many published documents were found by Google Scholar to utilize the phrase during the years 2012-2014, 2015-2017, 2020-2022, and January 2023-March 2024. The right-hand column depicts the number of citations received so far in total for the three articles initially flagged as being of chatbot origin from January 2023 March 2024. Thus, the cites per article is calculated as Cites divided by a factor of 3. Table 3 provides information for articles with a lesser number of Google Search results pages. The column headings are the same as for Table 2. As in Table 2, there is often a rapid increase in Google Search results after ChatGPT introduction, beginning with the 2020-2022 time period. This suggests a relationship between peculiar catchphrases and their increased publishing in recent years, presumably because they are chatbot text.

Tables 2 and 3 are shown in graphical form in Figure 2 for ease-of-comparison, and to better depict the relationships between the variables. The top panel in Figure 2 represents Table 2, and the lower panel represents Table 3. The ordinate axis depicts the sum total of Google Search pages for all four time intervals studied. The numbers on the abscissa correspond to the listing number of each catchphrase in Tables 1 and 2 (total N = 50). Colors depict the number of search pages procured for each time interval, with the key provided in the panel. The time intervals are color-coded as follows: 2012-2014 burgundy, 2015-2017 orange, 2020-2022 yellow, and 2023-March 2024 green. It is evident that there is a large transition to a greater number of search pages found for chatbot-affiliated phrasings, post-GPT, i.e. post-2020 onset. The slight increase from 2012-2014 to 2015-2017 might be expected from the increase in overall academic journal publishing occurring during that time.

For the nonchatbot-associated phrasings, Table 1, “an innovative approach”, “assuming the mantle”, and “at its helm” all had the maximum 14 pages of listings in Google Search for each time period. Hence, these phrasings are commonly used by humans and not by chatbots. Chatbot-associated catchphrases are shown in Table 4. The means more than triple over the ten-year interval, and particularly from 2020, suggesting the influence of chatbot usage. The significances of the page trends are shown in Table 5. There was no significance to the slight increase in search results delivered for 2012-2014 versus 2015-2017 time periods. There was a moderately significant increase in pages of search results for time interval 2015-2017 versus 2020-2022 ($p=0.01$). There was a highly significant increase in pages of search results for time interval 2020-2022 versus 2023-March 2024 ($p=0.004$), despite the time interval 2023- March 2024 being only 14 ½ months, as compared with the three years length for the other periods. Additionally, high significant difference is evident when comparing the early years to latest GPT years ($p<0.001$).

achieving a commendable	**incorporating innovative techniques**
adaptability and proactive approach	innovation lies in the **sophisticated integration of**
adaptability to diverse data structures	**intriguing presence of**
aiming to leverage	is its **strategic combination**
[a] meticulous evaluation	**iterative feature selection**
a **more comprehensive understanding**	it **notably enhanced performance**
and an impressive	**leverage the advantages**
beacon of guidance	making **a significant contribution**
beacon of promise	**meticulously addresses** ethical considerations
bedrock of innovation	meticulously **crafted roadmap**
broader and more diverse dataset	model **a promising solution for automated**
can be attributed to several key factors	**model innovatively [combines]**
chart a course towards **transformative discovery**	more **robust and adaptable** model
collectively, these innovations	obscured, **casting a pall**
computational **lightweight nature**	**pioneers the application**
contributes to the overall robustness	**pivotal research endeavor**
coupled with advanced	**poised to unravel** the mysteries shrouding
delve into the intricate mechanisms	**potentially capture diverse**
demonstrating a comprehensive strategy	**potential scalability to** larger
dual application **leverages the strengths**	presents **a pivotal advancement**
dynamically selects the most suitable pattern	**reflects a deliberate** choice
elevating the model	**relentless pursuit of excellence**
embodies a rich tapestry	rich, **multi-dimensional representation**
embodies **a symphony of innovation**	**robustness and generalizability across diverse** datasets and conditions
enables comprehensive feature extraction	**showcasing the model's adaptability**
enhances the model's ability	**solid cornerstone for** this
enhance the model's effectiveness	**streamlined approach accelerates** the deployment process
enhancing its ability to adapt	that offer **a rich, multi-dimensional** representation
enhancing its capability	**[their] academic prowess**
enhancing its versatility	the model **captures a broader** set
enhancing their growth and productivity	**the model's limitation**

ensures a robust	**these innovations enhance**
excels in extracting	the **turbulent seas of** scientific inquiry
exhibits promising findings	**this amalgamation [forms]**
extraction of meaningful features	This incongruity **introduces a potential constraint**
facilitates a **self-organized selection**	**this versatility, coupled** with its
fluctuation in the model's	to **harness existing knowledge**
fortifying the scientific rigor	to **the pressing issue**
future endeavors could benefit	transformative **discoveries in the realm** of
highly accurate and efficient [tool]	underscores its effectiveness **and suitability for diverse** applications
hints at their potential	**underscoring a limitation**
implies a level of autonomy	**unveiled a wealth**
incorporates a multifaceted approach	while the proposal **brims with potential**

Table 1. Chatbot-associated articles (N = 86)

Label	2012-2014	2015-2017	2020-2022	2023-2024	Cites
1 captures a broader	14	14	14	14	2
2 strategic combination	14	14	14	14	1
3 leverages the strengths	14	14	14	14	0
4 sophisticated integration of	14	14	14	14	0
5 a meticulous evaluation	13	14	14	14	0
6 future endeavors could	5	5	14	14	1
7 intriguing presence of	9	10	10	9	0
8 discoveries in the realm	6	7	7	9	2
9 a pivotal advancement	2	3	7	14	1
10 dynamically selects the most	6	6	8	10	0
11 enhances the model's ability	0.9	0.6	9	14	1
12 can be attributed to several key factors	2	3	5	14	0
13 potential scalability to	6	7	8	6	5
14 a rich, multi-dimensional	8	8	7	5	0
15 adaptability and proactive	0.9	3	8	10	15

16 enhancing its versatility	2	3	3	14	0
17 beacon of guidance	3	4	6	8	0
18 and suitability for diverse	1.1	2	7	9	0
19 poised to unravel	3	4	4	9	4
20 elevating the model	1	1	3	13	0
21 hints at their potential	2	3	5	5	0
22 collectively, these innovations	2.2	3	4	5	1
23 adaptability to diverse data	0.1	0.4	0.5	12	1
24 unveiled a wealth	2.1	2	4	6	0
25 incorporating innovative techniques	1.2	2	3	7	3

Table 2. Frequently utilized chatbot-suspected phrases (≳ 10 Googe Search pages result)

Label	2012-2014	2015-2017	2020-2022	2023-2024	Cites
26 delve into the intricate mechanisms	0	0.1	0.5	10.0	2
27 notably enhanced performance	0.2	0.7	3.0	6.0	1
28 meticulously addresses	0.7	1.6	2.0	6.0	9
29 enhance the model's effectiveness	0.1	0	0.6	7	4
30 enhancing its ability to adapt	0.3	0.3	3	4	1
32 model innovatively combines	0.2	0.7	3	3	1
33 harness existing knowledge	1.4	1.6	3	2	0
31 brims with potential	0.7	0.5	2	3	1
34 a promising solution for automated	0.1	0.5	1	4	10
35 this versatility, coupled	0.5	1.5	1	3	0
36 fortifying the scientific rigor	0	0	0	4	1
37 incorporates a multifaceted approach	0.5	0.7	1	2	0
38 broader and more diverse dataset	0.1	0	0.2	3	6
39 underscoring a limitation	0.2	0.4	1	2	0
40 a symphony of innovation	0.1	0	0	3	0
41 this amalgamation forms	0.2	0	0.2	2	1

42 robustness and generalizability across diverse 0	0	0.2	2	0	
43 showcasing the model's adaptability	0.4	0.4	0	2	1
44 highly accurate and efficient tool	0.2	0.3	1	0.4	1
45 crafted roadmap	0.2	0.2	0.5	0.6	3
46 enhancing their growth and productivity	0.1	0.1	0.5	0.7	1
47 streamlined approach accelerates	0	0	0	1	1
48 pivotal research endeavor	0	0	0	1	0
49 fluctuation in the model's	0	0	0	0.9	0
50 introduces a potential constraint	0	0	0.1	0.6	4

Table 3. Less utilized chatbot-suspected phrases (≲ 10 Google Search pages result)

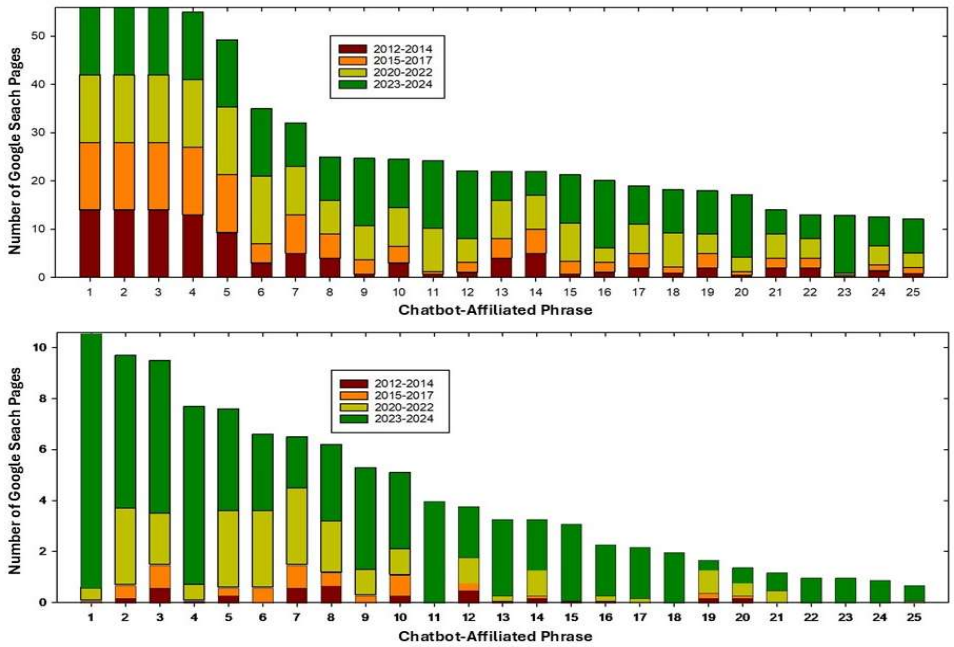

Figure 2. Transition in chatbot-affiliated phrasings over four time periods. For many of the phrasings, there is a large increase in Google Search page results from 2012-2014 and 2015-2017 (burgundy and orange) to 2020-2022 and 2023 – March 2024 (yellow and green). The upper panel shows more frequently found chatbot phrasings, while the lower panel shows less frequently found chatbot phrasings. They are separated for simplicity and clarity – so that differences in the heights of the colored bars become more apparent as a result. The year 2018, the inception of GPT, marked the transition to more Google Search page results for chatbot-associated phrasing.

Time interval	2012-2014	2015-2017	2020-2022	2023-March 2024
Number of pages	21.7±38.7	25.6±40.8	43.2±45.7	67.2±48.2

Table 4. Number of Google Search pages with chatbot-associated catchphrases

Time interval comparison	Significance
2012-2014 versus 2015-2017	not significant
2015-2017 versus 2020-2022	p = 0.01
2020-2022 versus 2023-March 2024	p=0.004
2012-2014 versus 2020-2022	p = 0.002
2012-2014 versus 2023-March 2024	p < 0.001
2015-2017 versus 2023-March 2024	p < 0.001

Table 5. Significance of the difference between Google Search page totals

Hence once again, an abrupt increase in chatbot usage is evident from the jump in chatbot catchphrasing.

The mean number of citations for articles with chatbot phrases published in 2023-2024 was 0.57±.0.97. The mean journal impact factor was 4.99±4.66 for the articles published in N = 112/150 journals which have an SCI impact factor (i.e., 75% of the total of 150 articles studied). Of the 38 articles that were not published in an impact factor journal (25%), these included: 1 archive, 4 book chapters, 6 proceedings, and 27 articles contained in journals not yet fully indexed. Of note, a chatbot-containing article was published in one journal with impact factor > 30, and 10 of the chatbot-containing articles were published in journals with an impact factor of 10-20. Hence chatbot usage is found even in the leading scientific journals.

Figure 3 illustrates the number of 150 chatbot-affiliated articles (N = 3 × 50) associated with each of several broad topics which are marked on the abscissa. Fifteen articles were published under the general field of medicine, 11 were in ecology and environmental science, 10 were related to any kind of imaging technology, 10 pertained to any sensor

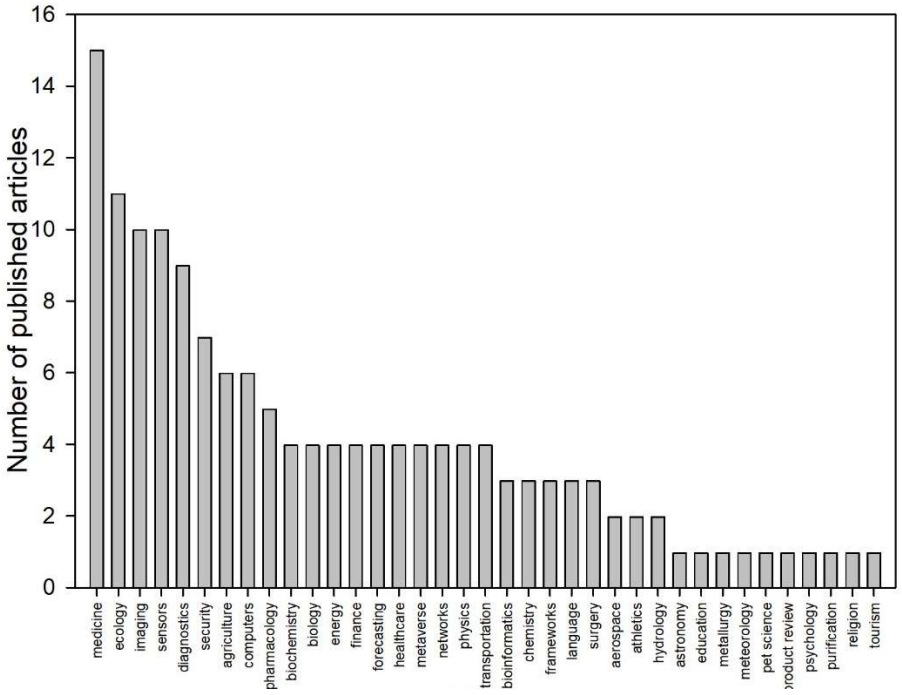

Figure 3. Type of article for each of N = 3×50 = 150 published articles analyzed for the study which had a chatbot affiliation. Many such articles were in the general field of medicine, ecology and environmental studies, imaging rendering and technology, sensors, and diagnostics, including medical diagnostics.

technology, including medical sensors, and so on. Hence, there are diverse topics in which chatbot-associated text has been published, but many are in medicine and the physical sciences. Figure 4 denotes the article section that each chatbot phrasing paragraph originated from. For a majority, 57 papers, this was the Abstract. Second most commonly, in 30 papers the appearance occurred in the Introduction. Authors using chatbot appear to be most comfortable incorporating chatbot in the Abstract and Introduction of their papers. In 7 papers, chatbot phrasing occurred in the Methods section. For 25 papers, phrasing occurred in the Results and/or Discussion sections. Only the section of the article in which the phrasing first appeared was included in the statistics shown. The question of whether the chatbot-affiliated phrasing was attributed in any way in the Acknowledgment section of the document was not investigated.

Discussion

GPT-generated content

Detecting AI-generated text is important for reasons of attribution, and also because such text may be incorrect [15]. However, humans may perform only slightly better than chance, i.e. with approximately 50% correct and 50% incorrect results, when classifying machine-generated versus human-written text [16]. This has led investigators to consider automated detection methods that may identify signatures difficult for humans to recognize. These include feature-based classifiers and statistical measures, such as finding variation in sentence lengths, and the frequency of

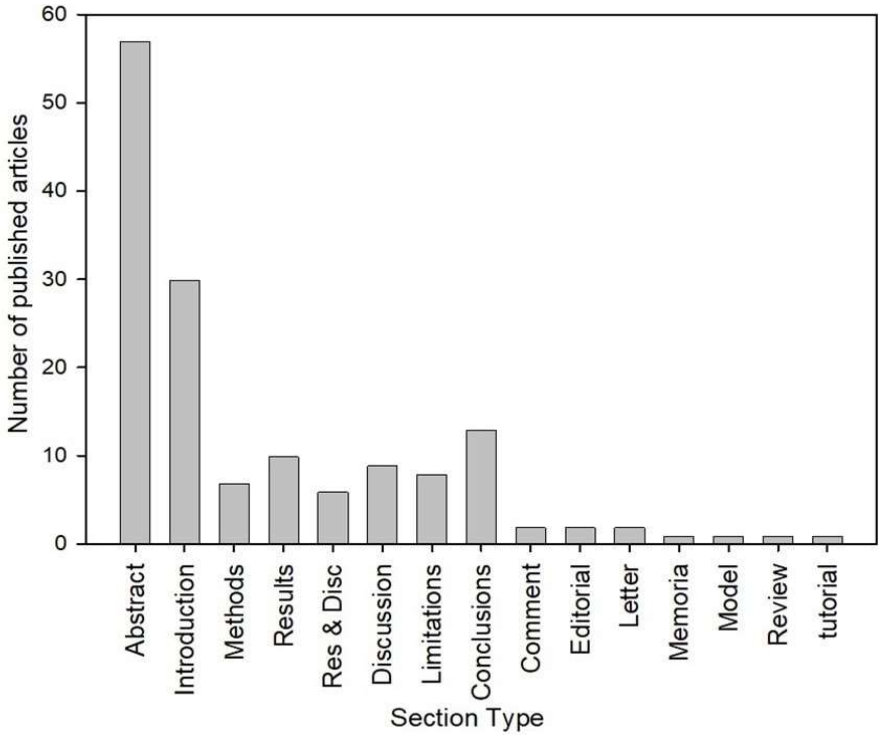

Figure 4. Article section in which the analyzed chatbot paragraph appeared. The Abstract, Introduction, and Conclusions sections were the most popular to utilize for chatbot phrasing.

certain words and punctuation marks [17]. A dataset has even been developed as a resource to evaluate and advance detection techniques [18,19].

Human or manual detection of GPT

Producing human-quality text has long been a main goal for the output of natural language generation systems, and

serves as an upper bound on performance [20]. Human evaluators are often asked to assess a text holistically, e.g., based on its overall quality, naturalness, or human-likeness [21, 22], where the exact evaluation criteria is left to the discretion of the evaluator. Yet, for state-of-the-art large language models, errors are often content based, not stylistic [23]. Research on GPT-3 detection suggests that human detection abilities decline as LLM complexity increases [24]. Without additional training to become familiar with chatbot writing derived from advanced neural networks, evaluators may distinguish between GPT and human-authored text only at the random chance level [20]. Training using LLM-generated examples, including detailed instructions, annotated examples, and paired examples, can marginally improve detection rates [20]. Requesting the GPT to self-edit can act to further reduce human detection rates. Exacerbating these difficulties, the results of human detectors are inconsistent across textual domains, and the reasons evaluators provide for their judgments are often diffuse [20].

To improve upon manual detection of chatbot, in this study, Google searched for a phrase before and after ChatGPT advent. The large jump in usage of the phrase in articles from early to later time intervals (Figure 3) and from 2020-2022 versus 2023-March 2024 (Results, $p = 0.004$), reflects an increased affiliation of the catchphrase to GPT in the latter two-year intervals ($N = 86$). Thus a combination of searching for peculiar phrases and then checking their frequency pre versus post ChatGPT advent may increase accuracy for manual chatbot detection. It can also be hypothesized that increased familiarity with chatbot-derived phrasing and increased practice at manual checking for chatbot will increase skill level to further improve detection rates. The human mind may be skillful at understanding chatbot phrasing, and in identifying such patterns, given sufficient practice in doing so. Hence, skilled human investigators might even rise to near machine-level accuracy in detecting chatbot, the latter of which is limited by the quality of the programming steps. Using feedback regarding phrase usage increases pre versus post ChatGPT, and GPTZero results can potentially improve manual skill level, and obviate the need for using machine-based chatbot detectors, which require preprocessing, can be inconvenient, often require a monthly fee, and are not always very reliable.

Advantages and Disadvantages of the Method

An advantage of using the manual detection paradigm outlined in Figure 1 is to not rely on a chatbot detection computer program, whose workings are unknow by the user. A combination of sudden increase in phrase usage pre versus post 2020 is a clue that chatbots have usurped it. Another clue is the GPTZero positivity for chatbot of the paragraph in which it is found by. Finally, if several such articles with the phrase are found to be chatbot-associated (a benchmark of three in total was used in this study) it suggests that chatbots have made the phrase a common buzzword in their lexicon. This study may provide helpful information as to how ChatGPT formulates sentence structure. It may assist editors who are uncertain how machine-driven text impinges upon a submitted work. Overall, developing an improved means to decode the phrasing associated with machine-driven text can be assistive in understanding artificial intelligence, its uses and limitations, which can benefit progress in the field.

Limitations

This investigation was developed to provide a historical suggestion as to how chatbot generated text might be detected and analyzed manually (to March 2024). However, only one chatbot detection program was used for the quantitative analysis, GPTZero. Adding other online detectors could provide complimentary or improved results, such as Turnitin's AI writing detection tool (https://www.turnitin.com/). A future research study should therefore use several high-quality online detectors to more rapidly and efficiently determine if a particular paragraph with odd phrasing is machine-constructed. Furthermore, in developing this study, the lack of identifiable chatbot in 7 phrasing instances, Table 1, may have been due in part to the lack of searchable texts, or to the minimal Google Search results for some of the phrasings.

Conclusions

It is suggested that chatbot content in scientific papers can include peculiar phrasings that are identifiable by a human investigator. When chatbot-suspected phrasing is flagged by GPTZero or another detector, the same phrasing often

appears in other documents found in a Google Search as well. The list provided herein was an initial attempt to identify and characterize the chatbot lexicon. Moreover, it is apparent that chatbot-associated phraseology, as measured by Google Search, is increasing in published academic documents, and is sometimes present even in the top impact factor journals. Detection of chatbot phrasing has become more difficult over time.

A future study should investigate the detection of more current chatbot phrasing. Disclosures

Disclosures

None.

Conflicts of Interest

None.

Author Biography

Edward J. Ciaccio, PhD is a senior research scientist in Medicine at Columbia University Medical Center and an honorary principal research fellow at Imperial College London. He previously edited two journals, founding one, and brought them both to Q1 in rank. His interests include modeling heart arrhythmias and onset of celiac disease.

References

[1] Khalil M, Er E. Will ChatGPT Get You Caught? Rethinking of Plagiarism Detection. In International Conference onHuman-Computer Interaction 2023 Jun 9 (pp. 475-487). Cham: Springer Nature Switzerland.

[2] Adamopoulou E, Moussiades L. Chatbots: History, technology, and applications. Machine Learning with applications. 2020 Dec 15;2:100006.

[3] McIntire JP, McIntire LK, Having PR. Methods for chatbot detection in distributed text-based communications, 2010 International Symposium on Collaborative Technologies and Systems, Chicago, IL, USA, 2010, pp. 463-472, doi: 10.1109/CTS.2010.5478478.

[4] Fried I. “Warning sounded over “flirting robots”,” Beyond Binary, CNET News [web article], Dec. 7, 2007, Accessed April 5 2023: https://www.cnet.com/culture/warning-sounded-over-flirting-robots/

[5] Gianvecchio S, Xie M, Wu Z, Wang H. “Measurement and classification of humans and bots in Internet chat,” Proceedings ofthe 17th USENIX Security Symposium (Security”08), San Jose, CA, July 2008.

[6] Jain S, Niranjan D, Lamba H, Shah N, Kumaraguru P. Characterizing and detecting livestreaming chatbots. InProceedings ofthe 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019 Aug 27 (pp. 683-690).

[7] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference oncomputer vision and pattern recognition 2016 (pp. 770-778).

[8] Borji A, Mohammadian M. Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard. GPT-4, Claude, andBard (June 12, 2023). 2023 Jun 12.

[9] Kendall G, da Silva JA. Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatorypublishing, and paper mills. Learned Publishing. 2024 Jan 1;37(1):55-62.

[10] Ciaccio EJ. Use of artificial intelligence in scientific paper writing. Informatics in Medicine Unlocked. 2023 Apr 23:101253.

[11] Chaka C. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI contentdetection tools. Journal of Applied Learning and Teaching. 2023;6(2).

[12] Akram A. An Empirical Study of AI Generated Text Detection Tools. arXiv preprint arXiv:2310.01423. 2023 Sep 27.

[13] Perkins M, Roe J, Postma D, McGaughran J, Hickerson D. Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. Journal of Academic Ethics. 2023 Oct 31:1-25.

[14] Rasul T, Nair S, Kalendra D, Robin M, de Oliveira Santini F, Ladeira WJ, Sun M, Day I, Rather RA, Heathcote L. The role ofChatGPT in higher education: Benefits, challenges, and future research directions. Journal of Applied Learning and Teaching. 2023 May;6(1).

[15] Lin S, Hilton J, Evans O. TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th AnnualMeeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 32143252, Dublin, Ireland, May 2022.

[16] Gehrmann S, Strobelt H, Rush A. GLTR: Statistical detection and visualization of generated text. In Proceedings of the 57thAnnual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 111–116, Florence, Italy, July 2019.

[17] Bhattacharjee A, Liu H. Fighting fire with fire: can ChatGPT detect AI-generated text?. ACM SIGKDD ExplorationsNewsletter. 2024 Mar 28;25(2):14-21.

[18] Ma Y, Liu J, Yi F, Cheng Q, Huang Y, Lu W, Liu X. AI vs. Human--Differentiation Analysis of Scientific Content Generation. arXiv preprint arXiv:2301.10416. 2023 Jan 24.

[19] Qazi Z, Shiao W, Papalexakis EE. GPT-generated Text Detection: Benchmark Dataset and Tensor-based Detection Method.arXiv preprint arXiv:2403.07321. 2024 Mar 12.

[20] Clark E, August T, Serrano S, Haduong N, Gururangan S, Smith NA. All That's `Human" Is Not Gold: Evaluating HumanEvaluation of Generated Text. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 7282–7296, 2021.

[21] van der Lee C, Gatt A, van Miltenburg E, Krahmer E. 2021. Human evaluation of automatically generated text: Current trendsand best practice guidelines. Computer Speech & Language, 67:101151.

[22] Howcroft DM, Belz A, Clinciu M-A, Gkatzia D, Hasan SA, Mahamood S, Mille S, van Miltenburg E, Santhanam S, Rieser V. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In Proceedings of the 13th International Conference on Natural Language Generation, pages 169–182, 2021 Dublin, Ireland. Association for Computational Linguistics.

[23] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S,Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.

[24] Kumar R, Mindzak M, Eaton SE, Morrison R. (2022). AI & AI: Exploring the contemporary intersections of artificialintelligence and academic integrity. Canadian Society for the Study of Higher Education Annual Conference, Online. Werklund School of Education.