Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts

de Winter, J.C.F.

DOI
10.1007/s11192-024-04939-y

Publication date
2024

Document Version
Final published version

Published in
Scientometrics

Scientometrics
https://doi.org/10.1007/s11192-024-04939-y

Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts

Joost de Winter[1]

Received: 21 April 2023 / Accepted: 11 January 2024

Abstract
This study explores the potential of ChatGPT, a large language model, in scientometrics by assessing its ability to predict citation counts, Mendeley readers, and social media engagement. In this study, 2222 abstracts from PLOS ONE articles published during the initial months of 2022 were analyzed using ChatGPT-4, which used a set of 60 criteria to assess each abstract. Using a principal component analysis, three components were identified: Quality and Reliability, Accessibility and Understandability, and Novelty and Engagement. The Accessibility and Understandability of the abstracts correlated with higher Mendeley readership, while Novelty and Engagement and Accessibility and Understandability were linked to citation counts (Dimensions, Scopus, Google Scholar) and social media attention. Quality and Reliability showed minimal correlation with citation and altmetrics outcomes. Finally, it was found that the predictive correlations of ChatGPT-based assessments surpassed traditional readability metrics. The findings highlight the potential of large language models in scientometrics and possibly pave the way for AI-assisted peer review.

Keywords Citation prediction · Scientometrics · Altmetrics · ChatGPT · GPT-4 · Scientific abstracts · Artificial intelligence

Introduction

In scientific research, predicting a publication's number of citations and its alternative impact metrics (altmetrics) such as Twitter and blog mentions, is becoming increasingly important. Citations are often used as an indicator of the impact of research contributions, allowing researchers, institutions, funding agencies, and policymakers to make informed decisions (e.g., Aksnes et al., 2019; Caon et al., 2020). Altmetrics, on the other hand, can serve as a valuable index of societal impact, indicating public engagement and interest in research (Bornmann, 2014; Pulido et al., 2018). Nevertheless, their correlation with scientific impact in the form of citations may not be strong, as altmetrics scores could be

✉ Joost de Winter
j.c.f.dewinter@tudelft.nl

[1] Department of Cognitive Robotics, Faculty of Mechanical Engineering, Delft University of Technology, Delft, Netherlands

Published online: 15 February 2024

primarily driven by factors such as public appeal (De Winter, 2015; Hassan et al., 2017; Pandey Akella et al., 2021; Warren et al., 2017).

Mendeley, a reference management tool that enables researchers to organize literature, provides another valuable index of readership which can reflect the degree of interest in a publication (Thelwall, 2018). A higher number of Mendeley readers suggests that a publication is frequently accessed by fellow researchers, which could lead to collaboration and citation opportunities (Bornmann, 2015; Haustein et al., 2014).

The factors influencing citation rates may include the validity of the research methods and conclusions (e.g., Antonakis et al., 2014) and the clarity of the researcher's writing (Dowling et al., 2018). Estimating these metrics before submission could benefit researchers by allowing them to optimize their work and maximize its impact within both the scientific community and society at large. This predictive ability could also prove valuable for educators and writing instructors who play a role in shaping their students' research output (e.g., Baldwin & Chandler, 2002; Murray et al., 2008).

The advent of large language models, especially ChatGPT, has enabled a wealth of new research opportunities, which may be of benefit to scientometrics research. ChatGPT has demonstrated that it can successfully tackle a range of tasks, such as summarizing text (Yang et al., 2023), stance detection (Huang et al., 2023; Zhang et al., 2022), and sentiment analysis (Tabone & De Winter, 2023; Zhong et al., 2023). The latest version, ChatGPT-4, has been shown to outperform the average student in a variety of verbally-oriented tests and exams (De Winter, 2023; Katz et al., 2023; Nori et al., 2023; OpenAI, 2023a). The current study used ChatGPT-4 to rate scientific abstracts across multiple characteristics. The ratings generated by ChatGPT were then used to investigate their predictive validity in determining the number of citations and altmetrics scores received by the corresponding papers.

In scholarly literature, various approaches have been investigated for predicting the impact of scientific publications. These include basic textual features, such as paper metadata (Ma et al., 2021) and paper length (Haustein et al., 2015; Xie et al., 2019). Additionally, more refined textual analyses have been proposed as citation predictors, encompassing linguistic complexity and readability metrics (Ante, 2022; Lu et al., 2019; Sienkiewicz & Altmann, 2016), stylometry of titles and abstracts (Jimenez et al., 2020), and sentiment analysis (Liu & Zhu, 2023). In a review regarding the application of artificial intelligence for predicting citations and other research assessment metrics, Kousha and Thelwall (2022) observed that the existing predictors, such as readability scores and text-based analyses, have inconsistent predictive power. They posited that "*a risk with text mining to predict citation counts is that it is likely to work best by identifying highly cited topics, predicting higher citation counts for all articles on these topics. A successful prediction model for one year might be invalid for the next one due to topic changes, so text mining may need rebuilding each year to identify the new hot topics*" (p. 63). Consequently, there may be a need for more generalizable predictors that do not rely on specific textual elements.

Presently, a research gap exists in determining the extent to which large language models, such as ChatGPT-4, can contribute to the prediction of citation counts and altmetrics scores. Our study aims to address this gap by using ChatGPT-4 to evaluate scientific abstracts across various qualities, and subsequently investigating the predictive value of these evaluations concerning the attention received by the respective papers.

For this investigation, it was decided to concentrate on the open-access mega-journal PLOS ONE. Articles were selected within a specific time frame (January and February 2022) to ensure an approximately equal age and minimize temporal confounding factors when evaluating cumulative citations and altmetrics scores. Early 2022 was deliberately selected to circumvent ChatGPT-4's knowledge cut-off of September 2021, thereby mitigating the risk of 'contamination'. This term denotes the potential that the training modules of the base model may have been exposed to the abstracts, citation numbers, and associated data. The concern of contamination in benchmark analyses has also been recognized by other researchers (Aiyappa et al., 2023; Bubeck et al., 2023; De Winter, 2023; OpenAI, 2023a). In the present study, the ChatGPT-4 model was used to evaluate PLOS ONE paper abstracts across a diverse range of qualities (which themselves were proposed by ChatGPT), which were subsequently correlated with citation counts and altmetrics scores.

Methods

On April 2, 2023, a total of 2222 records of articles published in the journal PLOS ONE during January and February 2022, including their respective abstracts, were downloaded using Scopus (Scopus, 2023). The current study solely focused on document types marked as ''article', excluding other document types such as review articles, errata, and retractions from our analysis. On January 4, 2024, the number of citations, altmetrics scores, dimensions citations, and Google Scholar citations for the same 2222 papers were extracted using Scopus (2023), Altmetric (2023), Dimensions (2023), and Publish or Perish software (Harzing, 2023), respectively. The altmetrics scores included a variety of counts (e.g., news mentions, Twitter mentions, as well as Mendeley readers). Dimensions is a platform that extracts citations from various sources, including journal items, books, conference proceedings, pre-prints, and patents. Because Altmetric only includes the number of Mendeley readers when other altmetrics scores are available, the missing numbers of Mendeley readers were manually inserted based on the Mendeley database (Mendeley, 2023).

Between April 5 and 17, 2023, a custom script in MATLAB R2021b was used to submit each abstract individually to OpenAI's Application Programming Interface (API), together with a prompt that asked for scores on 60 different items. For each abstract, ChatGPT-4 was instructed to provide numerical ratings between 0 and 100. The abstract records characteristically included a copyright declaration, commencing with the term 'Copyright' or the copyright symbol. This segment of the abstract was automatically removed prior to the submission of the prompt to the API. The temperature setting of ChatGPT-4, which indicates the level of randomness in its responses, was set to 0 by the user to maximize the repeatability of results.

An example prompt is provided below, where (…) refers to omitted text for brevity.

Please rate the following abstract on each of the 60 items from 0 = Not at all to 100 = Very much. Only provide the numbers. For example:

1. 65
2. 50
3. 5
4. 95
5. …

This is the abstract:

Single nucleotide polymorphisms (SNPs) in the (…) variants in disease genesis

These are the items:

1. Original
2. Accessible
3. Circumlocutory
4. Nontechnical

(…)

60. Uninsightful

The items were generated using the following query in the ChatGPT-4 web interface: "*I have to rate a short scientific text according to 25 variables. For example 'Engaging', 'Controversial',* etc. *Could you list 25 of such variables? The variables should be as diverse and orthogonal as possible. Only provide the words; do not define their meaning.*" Five additional items were manually incorporated: *Difficult to understand*, *Exciting*, *Not well written*, *Theoretical*, and *To the point*, based on prior explorations with abstracts from another journal.

Furthermore, 30 antonyms corresponding to the 30 items were generated by using the following prompt in the ChatGPT-4 web interface: "*For these 30 items, provide the antonyms in the same order.*" The objective of the inclusion of the antonyms was to examine if negatively worded items would result in negative correlations with their counterparts: If the model exhibits consistent responses to both positive and negative wordings of a comparable concept, it serves as an indication that ChatGPT-4 is capturing the intended components. Moreover, the combination of 30 items with their respective 30 antonyms helps control for acquiescence bias. Analogous approaches have been applied in questionnaires designed for human respondents (Weijters & Baumgartner, 2012). For each prompt, the 60 items were presented in a random order, to minimize order effects that might influence the performance of ChatGPT. All 60 items are shown in Table 1.

For every individual abstract, the prompt yielded one corresponding ChatGPT-4 output. Using our MATLAB script, the numerical responses were extracted from the ChatGPT-4 outputs, resulting in a 2222 × 60 matrix of numbers between 0 and 100. Next, considering that some items might be anticipated to be correlated, a principal component analysis (PCA) was conducted on the standardized scores, i.e., a 2222 × 60 matrix with means of 0 and standard deviations of 1. The number of components to retain was based on a scree plot (Cattell, 1966). To improve interpretability of the components loadings, the loadings were subjected to orthogonal Varimax rotation. Principal component scores (i.e., a 2222 × 3 matrix) were computed using the standard procedure of multiplying the standardized scores with the pseudoinverse of the transpose of the 60 × 3 Varimax-rotated loading matrix, i.e., $x(\lambda^{T})^{+}$. Because of the Varimax rotation, the three component scores were uncorrelated ($r = 0.00$).

Next, the predictive value of the component scores was evaluated in relation to the following dependent variables: (1) the abstract length in number of characters, the number of mentions in (2) blogs, (3) news items, (4) Twitter, and (5) Reddit, (6) Mendeley readers, as well as the available citation counts obtained from (7) Dimensions, (8) Scopus, and (9) Google Scholar. Although additional altmetrics scores were available (e.g., the number of

Table 1 Items that were included with each prompt

No	Item	No	Item (antonym)
1	Engaging	31	Disengaging
2	Controversial	32	Uncontroversial
3	Rigorous	33	Lax
4	Innovative	34	Conventional
5	Accessible	35	Inaccessible
6	Methodical	36	Haphazard
7	Concise	37	Verbose
8	Persuasive	38	Unconvincing
9	Comprehensive	39	Superficial
10	Insightful	40	Uninsightful
11	Relevant	41	Irrelevant
12	Objective	42	Subjective
13	Replicable	43	Non-replicable
14	Structured	44	Unstructured
15	Coherent	45	Incoherent
16	Original	46	Derivative
17	Balanced	47	Unbalanced
18	Authoritative	48	Unreliable
19	Impactful	49	Inconsequential
20	Interdisciplinary	50	Narrow
21	Well-sourced	51	Poorly-sourced
22	Technical	52	Nontechnical
23	Provocative	53	Unprovocative
24	Hypothesis-driven	54	Speculation-driven
25	Ethical	55	Unethical
26	Difficult to understand	56	Easy to understand
27	Exciting	57	Dull
28	Not well written	58	Well written
29	Theoretical	59	Empirical
30	To the point	60	Circumlocutory

patent, Facebook, and Weibo mentions), these were deemed too infrequent (with a mean across articles of less than 0.053) to warrant inclusion as dependent variables.

Furthermore, based on research that demonstrated a correlation between the number of authors and citation count (Kousha & Thelwall, 2022; Sommer & Wohlrabe, 2017; Tahamtan et al., 2016), the number of authors was determined by counting the semicolons present in the 'Authors' record and subsequently adding one.

Additionally, for each abstract, a set of readability indexes were computed based on textual features, such as the number of sentences, words, characters, syllables, and complex words (those with three or more syllables). These features served as building blocks for calculating various classical readability indexes, including the Automated Readability Index (Senter & Smith, 1967), SMOG grade (McLaughlin, 1969), Flesch-Kincaid grade level (Kincaid et al., 1975), Flesch Reading Ease (Flesch, 1948), Coleman-Liau index (Coleman & Liau, 1975), and Gunning fog index (Gunning, 1952). These indexes, which take into account text characteristics such as the number of characters per word, syllables per word, words per sentence, and the number of polysyllabic words (words with 3 or more syllables), have been widely used in the field of scientometrics (e.g., Hartley, 2016; Wang et al., 2022). The predictive validity of these traditional text-based methods was compared with the component scores derived from ChatGPT. For computing the readability scores, an API provided by Ipeirotis (2023) was used.

Due to the potential presence of outliers (e.g., a small number of articles attracting a large number of tweets), Spearman rank-order correlation coefficients were used for determining associations with citations and altmetrics. The Spearman rank-order correlation is robust against the influence of outliers (Croux & Dehon, 2010; De Winter et al., 2016), making it a more suitable choice for the analysis (for similar approaches in scientometrics studies, see e.g., De Winter, 2015; Jimenez et al., 2020).

In interpreting correlation coefficients, Cohen (1988) once proposed that $r=0.10$, $r=0.30$, and $r=0.50$ could be considered as small, medium, and large effects, respectively. Later, Ferguson (2009) recommended a minimum correlation of $r=0.20$ to deem an effect practically significant. On the other hand, a meta-analysis by Gignac and Szodorai (2016) showed that correlations of 0.50 or stronger are rare in individual-differences research, and proposed a more realistic guideline of $r=0.10$, $r=0.20$, and $r=0.30$ representing small, medium, and large effects. This article adopts the same guideline for interpreting the strength of the correlation coefficients. A literature review on predictive effects in scientometrics by Kousha and Thelwall (2022) concurs that correlations $r>0.50$ are rare. Such large correlations may suggest a hidden causal relationship, variability due to a small sample size, or may be the result of aggregation prior to calculating the correlation. For comparison, in Ante (2022), readability statistics of 135,502 abstracts showed statistically significant correlations with citation counts, but the Spearman correlation coefficients were not higher than 0.1, while a meta-analysis of 24 effect sizes showed a Pearson correlation of 0.31 between paper length and the number of citations (Xie et al., 2019).

Results

Statistical reliability checks

The script developed for this study generated 60 responses for all 2222 abstracts. The duration for ChatGPT-4 to process a single abstract was around 43 s, with fluctuations observed throughout the day (ranging from approximately 35 to 55 s), likely attributable to varying server load at OpenAI. Overall, obtaining responses for all 2222 abstracts required roughly 27 h, and incurred a cost of USD 91.

To verify the reliability of the ChatGPT-4 output, the script was run three additional times. One repetition had the items in reverse order compared to the original run, whereas two other repetitions had the items in random order, with the one repetition inadvertently including a few words from another abstract ("Coaching has been described, in part,") for all abstracts. During this process, an interesting finding was uncovered: the item means were highly consistent ($r>0.998$ for all combinations of runs). For example, averaged across all 2222 abstracts, the item 'Unethical' consistently received low scores (5.85, 5.86, 6.03, and 5.98 for the four runs), while 'Relevant' received high scores (87.70, 87.21, 87.07, 87.52 for the four runs). However, when examining scores assigned to individual abstracts ($n=2222$), reliability was relatively low, averaging at $r=0.27$ (average of 60 correlation coefficients), ranging from $r = 0.00$ for ‘Ethical’ to $r = 0.57$ for ‘Easy to understand’.

Thus, ChatGPT-4 generated scores that are reliable at the population level (i.e., when averaged across all abstracts) yet unreliable for individual abstracts. Some variability can be attributed to the inability of ChatGPT to provide entirely non-random outputs, as OpenAI noted: “*Setting temperature to 0 will make the outputs mostly deterministic, but a small amount of variability may remain*” (OpenAI, 2023b). Moreover, it seems that the ChatGPT-4 output was highly sensitive to the order in which the 60 items were presented. Additionally, it is probable that certain items, like ‘Ethical’, are challenging to score since none of the abstracts would be expected to be unethical, resulting in minimal variance. The above repetition process highlights the importance of considering multiple items and not relying on a single ChatGPT-4 output score. Due to these reliability concerns, the average of the four scores per item per abstract was calculated before proceeding with further analysis.

Data reduction

Figure 1 presents the scree plot derived from the 60 × 60 correlation matrix of item scores. Based on the scree plot, the decision was made to retain three components. Although retaining five components would seem appropriate as well, the loadings for the three-component outcome were deemed more interpretable. Specifically, when extracting five components, one component was found to cluster along the 30 antonyms, and another component had low and uninterpretable loadings. Table 2 provides the means, standard deviations, and Varimax-rotated component loadings for all 60 items.

In general, items with high mean scores (> 50) are associated with positive qualities, such as Engaging, Rigorous, and Accessible, while their antonyms with lower mean scores (< 50) represent negative qualities, such as Disengaging, Lax, and Inaccessible. Furthermore,

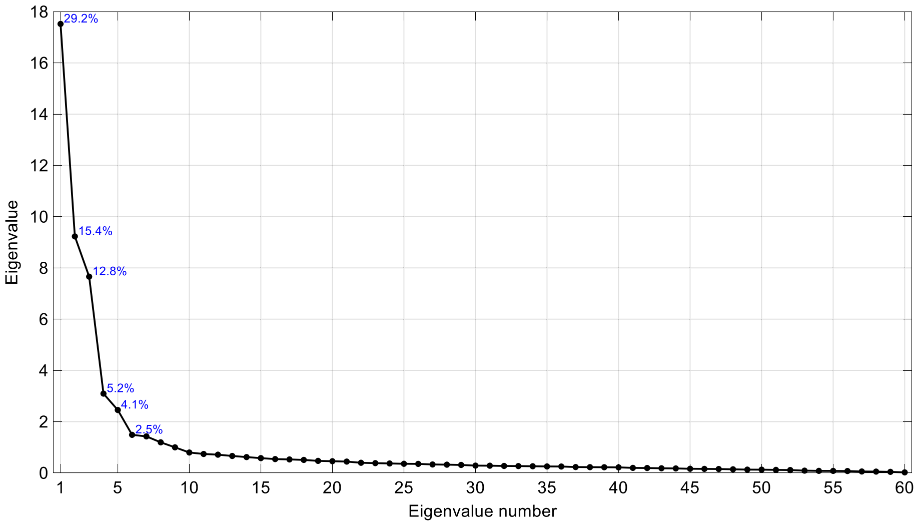

Fig. 1 Eigenvalues of the correlation matrix in descending order, also called the ‘scree plot’. Listed in blue is the percentage of variance explained

Table 2 Means and standard deviations of scores assigned to abstracts ($n = 2222$), and Varimax-rotated component loadings

No	Item	Mean	*SD*	Quality and Reliability	Accessibility and Understandability	Novelty and Engagement
1	Engaging	58.46	7.58	− 0.14	**0.54**	**0.63**
2	Controversial	23.49	9.90	− **0.47**	0.10	**0.44**
3	Rigorous	82.54	3.85	**0.77**	− 0.11	0.13
4	Innovative	62.62	9.08	0.15	− **0.41**	**0.71**
5	Accessible	63.25	11.37	− 0.11	**0.92**	− 0.01
6	Methodical	83.04	3.50	**0.74**	0.00	0.03
7	Concise	66.47	7.88	0.20	**0.76**	− 0.05
8	Persuasive	66.38	5.23	0.38	0.29	**0.60**
9	Comprehensive	72.25	5.09	**0.42**	− 0.05	0.25
10	Insightful	70.51	5.18	0.20	0.14	**0.79**
11	Relevant	85.91	4.12	**0.47**	0.31	0.38
12	Objective	82.21	4.78	**0.77**	0.02	− 0.22
13	Replicable	75.14	5.41	**0.68**	0.05	− 0.05
14	Structured	81.97	3.77	**0.64**	**0.50**	0.07
15	Coherent	82.60	4.82	**0.47**	**0.70**	0.12
16	Original	63.38	7.42	0.19	− 0.38	**0.68**
17	Balanced	76.16	4.16	**0.62**	**0.41**	0.03
18	Authoritative	74.03	4.11	**0.65**	0.06	0.23
19	Impactful	69.19	6.24	0.34	0.07	**0.70**
20	Interdisciplinary	50.04	9.74	− 0.19	− 0.07	**0.65**
21	Well-sourced	80.78	4.23	**0.72**	0.13	0.12
22	Technical	65.04	13.51	0.39	− **0.78**	0.03
23	Provocative	38.39	9.78	− 0.30	− 0.02	**0.72**
24	Hypothesis-driven	73.56	6.32	**0.41**	− 0.14	0.09
25	Ethical	87.38	6.29	0.22	0.18	− 0.13
26	Difficult to understand	37.71	11.74	0.08	− **0.93**	0.01
27	Exciting	41.22	9.78	− 0.08	− 0.06	**0.78**
28	Not well written	23.22	6.61	− 0.36	− **0.69**	− 0.24
29	Theoretical	44.90	12.04	− 0.32	− **0.48**	0.29
30	To the point	69.04	7.23	0.24	**0.76**	− 0.05
31	Disengaging	29.71	6.07	− 0.24	− **0.52**	− **0.50**
32	Uncontroversial	67.53	11.96	0.32	0.00	− **0.54**
33	Lax	14.01	3.87	− **0.80**	− 0.03	− 0.18
34	Conventional	49.82	8.24	− 0.02	0.22	− **0.67**
35	Inaccessible	35.45	10.65	0.05	− **0.91**	0.00
36	Haphazard	12.98	4.12	− **0.76**	− 0.36	− 0.10
37	Verbose	24.73	8.13	− 0.31	− **0.76**	− 0.01
38	Unconvincing	24.97	5.19	− **0.71**	− 0.20	− 0.30
39	Superficial	20.62	4.76	− **0.71**	− 0.06	− 0.38
40	Uninsightful	21.32	6.08	− **0.43**	− 0.15	− **0.51**
41	Irrelevant	10.49	3.31	− **0.63**	− 0.19	− 0.28
42	Subjective	18.79	6.00	− **0.74**	0.18	0.22
43	Non-replicable	23.33	6.02	− **0.68**	− 0.07	− 0.03

Table 2 (continued)

No	Item	Mean	*SD*	Quality and Reliability	Accessibility and Understandability	Novelty and Engagement
44	Unstructured	15.30	4.50	– **0.67**	– **0.45**	– 0.06
45	Incoherent	12.25	4.84	– **0.55**	– **0.60**	– 0.10
46	Derivative	29.60	6.05	– **0.44**	0.09	– **0.53**
47	Unbalanced	21.71	5.34	– **0.64**	– 0.25	– 0.03
48	Unreliable	18.99	4.60	– **0.75**	– 0.11	– 0.13
49	Inconsequential	20.50	5.81	– **0.51**	– 0.05	– **0.53**
50	Narrow	41.68	6.93	– 0.10	– 0.25	– **0.51**
51	Poorly-sourced	16.14	4.81	– **0.76**	– 0.14	– 0.10
52	Nontechnical	39.47	16.37	– 0.37	**0.78**	– 0.04
53	Unprovocative	40.66	12.21	– 0.09	– 0.02	– **0.62**
54	Speculation-driven	22.79	6.40	– **0.67**	– 0.15	0.26
55	Unethical	5.93	2.19	– **0.58**	0.03	– 0.01
56	Easy to understand	62.00	12.27	– 0.12	**0.93**	– 0.03
57	Dull	31.63	6.26	– 0.20	– 0.38	– **0.65**
58	Well written	73.75	6.00	0.31	**0.76**	0.27
59	Empirical	82.81	5.44	**0.54**	0.11	– 0.17
60	Circumlocutory	20.65	7.87	– 0.37	– **0.72**	– 0.02

Note Loadings exceeding 0.40 or falling below −0.40 are depicted in boldface (see Peterson, 2000, who identified that a common loading cutoff value is 0.40)

antonyms display inverse component loadings. Specifically, the loadings of Items 1–30 for each of the three components exhibited a strong correlation with the loadings of Items 31–60, that is, their corresponding antonyms ($r = -$ 0.91, – 0.92, and – 0.93, respectively).

Component 1 captures the *Quality and Reliability* of the abstract. High negative loadings correspond to negative aspects (e.g., Lax, Superficial, Unreliable), while high positive loadings signify positive aspects (e.g., Rigorous, Methodical, Objective). Component 2 represents the *Accessibility and Understandability* of the abstract. High negative loadings are associated with difficulty in understanding (e.g., Difficult to understand, Inaccessible, Technical), while high positive loadings indicate ease of understanding (e.g., Accessible, Easy to understand, Nontechnical). Component 3 captures *Novelty and Engagement*. High negative loadings are indicative of more conventional and uninteresting aspects (e.g., Unprovocative, Conventional, Dull), while high positive loadings correspond to engaging and innovative aspects (e.g., Engaging, Innovative, Exciting).

Additionally, a supplementary reliability check was performed using a split-half approach. In this method, instead of aggregating the item scores from all four runs as in the primary analysis, component scores were calculated by aggregating the item scores from runs 1 and 2, as well as from runs 3 and 4. The three component scores exhibited inter-correlations of $r = 0.56$, 0.81, and 0.70.

Predicting altmetrics scores and citation counts from ChatGPT-4 component scores

The correlations in Table 3 indicate the extent to which the three component scores predicted altmetrics and citation scores. The *Quality and Reliability* component displayed weak positive correlations with the number of citations (ρ = 0.05 to 0.10 for the three citation measures). Furthermore, weak correlations were observed between *Quality and Reliability* and the altmetrics.

The *Accessibility and Understandability* component demonstrated weak to moderate positive correlations with several altmetrics scores, including the number of news mentions (ρ = 0.10), blog mentions (ρ = 0.07), Twitter mentions (ρ = 0.18), and Reddit mentions (ρ = 0.06), as well as a large correlation with the number of Mendeley readers (ρ = 0.40). This component also showed weak to moderate positive correlations with the number of citations (ρ = 0.08 to 0.17). In other words, more accessible and understandable abstracts tend to receive more attention in terms of citations and altmetrics.

The *Novelty and Engagement* component exhibited weak to moderate positive correlations with altmetrics such as the number of news mentions (ρ = 0.09), blog mentions (ρ = 0.15), Twitter mentions (ρ = 0.23), Reddit mentions (ρ = 0.14), and Mendeley readers (ρ = 0.12). This component also displayed a moderate positive correlation with the number of citations (ρ = 0.18).

In summary, the *Accessibility and Understandability* and *Novelty and Engagement* components were found to be positively correlated with various altmetrics and citation scores, including Mendeley readers, while the *Quality and Reliability* component showed weak correlations with these metrics.

Predicting altmetrics scores and citation counts from classical readability scores

The correlation coefficients presented in Table 4 reveal a distinct pattern, specifically that there is only a very weak correlation between the readability indexes and altmetrics scores or citations. Among the relationships between readability indexes and outcome measures,

Table 3 Means, standard deviations (*SD*), and Spearman rank-order correlations between component scores and altmetrics scores & citations

	Mean	*SD*	ChatGPT-4: Quality and Reliability	ChatGPT-4: Accessibility and Understandability	ChatGPT-4: Novelty and Engagement
Number of news mentions	1.69	11.05	– 0.01	0.10	0.09
Number of blog mentions	0.12	0.95	– 0.03	0.07	0.15
Number of Twitter mentions	11.94	136.5	– 0.01	0.18	0.23
Number of Reddit mentions	0.07	0.64	– 0.03	0.06	0.14
Number of Mendeley readers	29.04	30.26	– 0.05	0.40	0.12
Number of Dimensions citations	6.30	8.31	0.10	0.11	0.18
Number of Scopus citations	5.23	6.92	0.10	0.08	0.18
Number of Google Scholar citations	8.65	11.94	0.05	0.17	0.18

Note. Correlations of 0.05 or stronger are statistically significantly different from 0, $p < 0.05$

Table 4 Means, standard deviations (*SD*), and Spearman rank-order correlations between altmetrics scores and readability scores

	Mean	*SD*	News mentions	Blog mentions	Twitter mentions	Reddit mentions	Number of Mendeley readers	Number of Dimensions citations	Number of Scopus citations	Number of Google Scholar citations
Abstract attributes										
Number of authors	6.63	4.53	− 0.05	− 0.04	0.08	− 0.04	0.07	0.03	0.04	0.00
Number of sentences	10.91	3.31	− 0.04	− 0.04	0.02	− 0.07	0.26	0.05	0.04	0.05
Number of words	251.0	63.43	− 0.03	− 0.01	0.06	− 0.03	0.26	0.05	0.04	0.04
Number of characters	1408.3	360.79	− 0.03	− 0.01	0.07	− 0.04	0.29	0.06	0.05	0.06
Number of syllables	485.3	125.07	− 0.04	− 0.01	0.07	− 0.04	0.27	0.07	0.06	0.07
Number of polysyllables	67.24	20.45	− 0.04	0.00	0.06	− 0.05	0.27	0.08	0.07	0.08
Readability scores										
Automated readability index (ARI)	16.92	2.78	0.02	0.06	0.07	0.05	0.01	0.02	0.02	0.02
SMOG grade	17.43	1.83	0.01	0.05	0.06	0.03	0.00	0.03	0.03	0.03
Flesch-Kincaid grade level	16.53	2.34	0.01	0.05	0.05	0.03	− 0.02	0.03	0.03	0.02
Flesch Reading Ease	18.89	11.86	0.00	− 0.02	− 0.03	0.03	− 0.04	− 0.05	− 0.06	− 0.05

Table 4 (continued)

Readability scores										
Coleman-Liau index	16.38	2.02	0.02	0.01	0.04	− 0.04	0.13	0.06	0.05	0.08
Gunning fog index	19.88	2.65	0.01	0.04	0.04	0.02	0.00	0.02	0.02	0.02

Note For the readability measures, higher scores suggest more complexity, except for the Flesch Reading Ease score, where higher means easier to read the strongest Spearman correlation was only 0.13. Furthermore, it is worth highlighting that the text attributes themselves (such as the number of sentences, words, and characters) exhibited a stronger prediction with the number of Mendeley readers (up to ρ = 0.29) in comparison to the readability scores.

Upon further investigation, it was observed that the readability scores displayed weak to moderate correlations with the three ChatGPT-4 scores (see Table 5). More specifically, abstracts that received higher scores on the *Accessibility and Understandability* component were associated with higher readability scores (ρ up to -0.21), while abstracts that achieved higher scores in *Novelty and Engagement* received lower readability scores (ρ up to 0.28). Moreover, articles with a higher number of authors had higher *Quality and Reliability* scores (ρ = 0.20).

Discussion

In the current study, a large language model, ChatGPT-4, was used to perform a scientometrics analysis. Prior research has predominantly relied on techniques such as string searches, sentiment analysis, and linguistic complexity analysis (e.g., Hassan et al., 2020; Sienkiewicz & Altmann, 2016). For example, Hu et al. (2021) examined the association between citation counts of scientific articles and the writing styles of abstracts. Their analysis revealed that abstracts of highly cited papers had a more complex vocabulary, longer sentences, and more complex syntactic structures, and were less readable than uncited abstracts. On the other hand, Dowling et al. (2018) found that readability had a positive

Table 5 Spearman-rank order correlations between ChatGPT-4 component scores and readability scores

	ChatGPT-4: Quality and reliability	ChatGPT-4: Accessibility and understandability	ChatGPT-4: Novelty and engagement
Abstract attributes			
Number of authors	0.20	0.07	– 0.03
Number of sentences	0.12	0.12	– 0.14
Number of words	0.13	– 0.03	0.00
Number of characters	0.11	– 0.02	0.04
Number of syllables	0.13	– 0.05	0.03
Number of polysyllables	0.12	– 0.05	0.07
Readability scores			
Automated readability index (ARI)	– 0.06	– 0.18	0.28
SMOG grade	– 0.01	– 0.20	0.25
Flesch-Kincaid grade level	– 0.01	– 0.21	0.25
Flesch Reading Ease	0.01	0.13	– 0.21
Coleman-Liau index	– 0.08	0.02	0.19
Gunning fog index	– 0.01	– 0.19	0.23

Note. Correlations of 0.05 or stronger are statistically significantly different from 0, $p < .05$. For the readability measures, higher scores suggest more complexity, except for the Flesch Reading Ease score, where higher means easier to read impact on the citations that *Economics Letters* articles receive, particularly for methods and macroeconomic papers. The discrepancy in predictability observed in the two selected articles may be indicative of domain-specific variations, where the importance of readability may be more pronounced in Economics compared to other disciplines (Kousha & Thelwall, 2022). However, it may also imply that the readability index that we used, which relies on features such as mean sentence length and the number of syllables per word, might not offer an accurate portrayal of readability, as it does not assess the words in the text in context (Hartley, 2016).

The current study pioneers a language-based evaluation of article abstracts. In contrast to conventional scientometrics methods—which may be prone to overfitting and inconsistent results, as discussed in the Introduction—the present approach represents an innovative direction. By assessing abstracts across a range of characteristics using semantically diverse items, this method may produce more generalizable outcomes. Our findings demonstrated that ChatGPT-4 scores displayed stronger correlations with altmetrics and citation counts in comparison to conventional readability scores, which are calculated based on the number of sentences, words, syllables, and characters.

A set of 30 items, along with their respective antonyms, was used in this study. This approach was anticipated to be robust against potential response biases. For example, relying solely on positively-worded items could have led ChatGPT-4 to assess the 'overall positivity' of the abstract. Our approach appeared effective, as demonstrated by the strong negative correlations between the loadings of the 30 items and their 30 corresponding antonyms. A strategy of aggregating across four script runs, and aggregating across items by means of principal component analysis, with items presented in a random order, was used to mitigate the limited reliability of individual item scores. The present approach offers a framework for appraising abstracts and may encourage additional research to further refine this method of prompting large language models.

Our correlational analysis showed several noteworthy findings. The number of authors positively correlated with *Quality and Reliability*, as could be expected. However, *Quality and Reliability* did not display a substantial association with altmetrics scores, while only a weak association was observed with the number of citations. These results are disconcerting, considering the worldwide attention on aspects such as methodological rigor and replicability, which are facets that highly loaded on the *Quality and Reliability* component. Others have similarly remarked that rigorous research, including pre-registered replications, does not necessarily correlate with popularity in terms of citations (Akcan et al., 2013; Hardwicke et al., 2021).

Abstracts assessed by ChatGPT-4 as *Accessible and Understandable* attracted a higher number of tweets and, more prominently, Mendeley readers, while also exhibiting a weak increase in citations. The correlation with Mendeley readers was large, at $\rho = 0.40$. It was also found that abstracts judged by ChatGPT-4 to be more *Novel and Engaging* tend to receive more citations, with moderate effects. From the perspective of science communication, having exciting and objective abstracts could make it easier for readers to appreciate the work, which in turn may promote knowledge dissemination (Sand-Jensen, 2007). However, there is a risk that the focus on crafting exciting abstracts might encourage sensationalism, which could detract from the pursuit of rigorous research.

This study presented a total of 60 items to the ChatGPT-4 model, incorporating some redundancy. Undoubtedly, further aspects could have been investigated through ChatGPT, such as the sentiment and tonality of the abstracts. While the relationship between the three identified components and citation counts appears to be interpretable, the strength of the correlations was found to be only weak to moderate, with Spearman's ρ values ranging between 0.08 and 0.18. Various factors might have imposed an upper limit on the correlations. One such factor is the potential unrepresentativeness of an abstract for the entire paper. Future research could benefit from analyzing full-text articles, including their titles, instead of just abstracts. Specifically, it is unclear whether the abstract itself is a causal factor or if it acts as an epiphenomenon, reflecting the overall quality of the research paper, with the latter being the true driving force behind social and scientific impact. In the case of altmetrics, however, it is worth considering whether even the full paper is meaningful, as people may retweet a title without reading the paper (or abstract). This ambiguity extends to the aspect of the abstract that should be rated by ChatGPT-4. For example, our exploration of the data revealed that one abstract was rated as controversial, presumably because the studied topic (tax evasion) was controversial, but this does not imply that the abstract, its research methods, or its conclusions are controversial.

Upon conducting a series of additional manual assessments, it was noted that many of the ratings provided by ChatGPT were in fact meaningful. For example, abstracts that received the lowest rating in terms of being ‘well-written’ or the highest in ‘not well-written’ were, according to the present author, hard to read due to complex sentence structures, or a frequent use of acronyms/abbreviations. Another illustrative case was an abstract that exhibited the lowest score on ‘replicability’; the present author observed that this abstract lacked a methodology and more closely resembled an essay. Conversely, some scores assigned by ChatGPT appeared to be somewhat superficial. An example of this was observed for an abstract that received a high score on ‘originality’ seemingly because the authors of the abstract had incorporated the words “We present novel …”. Another abstract was given a low rating on ‘balance’ possibly due to the presence of a recommendation for a balanced approach within the text. Future research could explore different types of prompts, such as prompts about the topic of the abstract, the methods and conclusions of the abstract, and the potential impact of the abstract on readers, in order to provide more precise insights into cause-and-effect relationships.

Another limitation is that, while 2222 abstracts were analyzed, it would have been preferable to process a larger dataset. Our current method was not particularly cost-effective, amounting to approximately 0.04 USD per abstract. In contrast, other scientometrics studies have analyzed millions of abstracts (De Winter & Dodou, 2015; Ioannidis, 2019; Pei et al., 2023; Sienkiewicz & Altmann, 2016), which is currently unattainable through ChatGPT-4 for regular researchers. Our analysis was based on abstracts from January and February 2022, just after the knowledge cut-off of ChatGPT-4. Consequently, the papers have had only 22–24 months to accumulate citations and altmetrics scores. Despite this limitation, our findings still offer valuable insights and may set the stage for further investigations of ChatGPT-4 or similar methods. Future research could consider applying a larger time gap between the moment of publication and the collection of citation counts. The data underlying this research is available in a public data repository, enabling other researchers to recalculate the correlations in a few years when the articles have garnered more citations. However, it is conceivable that by then, the ChatGPT models have already undergone improvements, making it sensible to reassess the abstracts at that time.

This study demonstrated the potential application of large language models in scientometrics. Our study may also stimulate reflection from a broader perspective. For example, it has been argued that artificial intelligence should play a role in the peer-review process for scientific articles, such as in manuscript pre-screening (Hancock, in press). In this context, our findings hold important implications. In the current study, artificial intelligence evaluated abstracts as subjective, unwieldy, and so forth. It is important to recognize that these assessments arise from a model trained on previously generated human data. Additionally, the GPT model underwent fine-tuning through a process known as reinforcement learning from human feedback (OpenAI, 2022, 2023a). As scientists strive to maintain academic integrity and autonomy, understanding *who* determines the quality of a scientific work remains a critical topic. It is hoped that this article stimulates a meaningful dialogue on this important issue.

Funding No funding was received for conducting this study.

Data availability Raw data and MATLAB scripts used for the analyses are available at https://doi.org/https://doi.org/10.4121/710585da-ed2e-4d36-b8e4-ad02c3af1e65.

Declarations

Conflict of interests The author has no relevant financial or non-financial interests to disclose.

References

Aiyappa, R., An, J., Kwak, H., & Ahn, Y.-Y. (2023). Can we trust the evaluation on ChatGPT? *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing*, Toronto, Canada, 47–54. https://doi.org/10.18653/v1/2023.trustnlp-1.5

Akcan, D., Axelsson, S., Bergh, C., Davidson, T., & Rosén, M. (2013). Methodological quality in clinical trials and bibliometric indicators: No evidence of correlations. *Scientometrics, 96*, 297–303. https://doi.org/10.1007/s11192-013-0949-0

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*. https://doi.org/10.1177/2158244019829575

Altmetric. (2023). Altmetric. https://www.altmetric.com/explorer/outputs?scope=all

Ante, L. (2022). The relationship between readability and scientific impact: Evidence from emerging technology discourses. *Journal of Informetrics, 16*, 101252. https://doi.org/10.1016/j.joi.2022.101252

Antonakis, J., Bastardoz, N., Liu, Y., & Schriesheim, C. A. (2014). What makes articles highly cited? *The Leadership Quarterly, 25*, 152–179. https://doi.org/10.1016/j.leaqua.2013.10.014

Baldwin, C., & Chandler, G. E. (2002). Improving faculty publication output: The role of a writing coach. *Journal of Professional Nursing, 18*, 8–15. https://doi.org/10.1053/jpnu.2002.30896

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics, 8*, 895–903. https://doi.org/10.1016/j.joi.2014.09.005

Bornmann, L. (2015). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics, 103*, 1123–1144. https://doi.org/10.1007/s11192-015-1565-y

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv. https://doi.org/10.48550/arXiv.2303.12712

Caon, M., Trapp, J., & Baldock, C. (2020). Citations are a good way to determine the quality of research. *Physical and Engineering Sciences in Medicine, 43*, 1145–1148. https://doi.org/10.1007/s13246-020-00941-9

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Academic Press.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*, 283–284. https://doi.org/10.1037/h0076540

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications, 19*, 497–515. https://doi.org/10.1007/s10260-010-0142-z

De Winter, J. C. F. (2015). The relationship between tweets, citations, and article views for PLOS ONE articles. *Scientometrics, 102*, 1773–1779. https://doi.org/10.1007/s11192-014-1445-x

De Winter, J. C. F. (2023). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00372-z

De Winter, J. C. F., & Dodou, D. (2015). A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ, 3*, e733. https://doi.org/10.7717/peerj.733

De Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods, 21*, 273–290. https://doi.org/10.1037/met0000079

Dimensions. (2023). Dimensions. https://app.dimensions.ai/discover/publication

Dowling, M., Hammami, H., & Zreik, O. (2018). Easy to read, easy to cite? *Economics Letters, 173*, 100–103. https://doi.org/10.1016/j.econlet.2018.09.023

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*, 532–538. https://doi.org/10.1037/a0015808

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221–223. https://doi.org/10.1037/h0057532

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gunning, R. (1952). *The technique of clear writing* (pp. 36–37). McGraw-Hill.

Hancock, P. A. (in press). Science in peril: the crumbling pillar of peer review. *Theoretical Issues in Ergonomics Science*. https://doi.org/10.1080/1463922X.2022.2157066

Hardwicke, T. E., Szűcs, D., Thibault, R. T., Crüwell, S., van den Akker, O. R., Nuijten, M. B., & Ioannidis, J. P. A. (2021). Citation patterns following a strongly contradictory replication result: Four case studies from psychology. *Advances in Methods and Practices in Psychological Science, 4*, 25152459211040836. https://doi.org/10.1177/25152459211040837

Hartley, J. (2016). Is time up for the Flesch measure of reading ease? *Scientometrics, 107*, 1523–1526. https://doi.org/10.1007/s11192-016-1920-7

Harzing, A.-W. (2023). Publish or perish (Version 8) [Software]. http://www.harzing.com/pop.htm

Hassan, S.-U., Aljohani, N. R., Idrees, N., Sarwar, R., Nawaz, R., Martínez-Cámara, E., Ventura, S., & Herrera, F. (2020). Predicting literature's early impact with sentiment analysis in Twitter. *Knowledge-Based Systems, 192*, 105383. https://doi.org/10.1016/j.knosys.2019.105383

Hassan, S.-U., Imran, M., Gillani, U., Aljohani, N. R., Bowman, T. D., & Didegah, F. (2017). Measuring social media activity of scientific literature: An exhaustive comparison of scopus and novel altmetrics big data. *Scientometrics, 113*, 1037–1057. https://doi.org/10.1007/s11192-017-2512-x

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PloS one, 10*, e0120495. https://doi.org/10.1371/journal.pone.0127830

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics, 101*, 1145–1163. https://doi.org/10.1007/s11192-013-1221-3

Hu, H., Wang, D., & Deng, S. (2021). Analysis of the scientific literature's abstract writing style and citations. *Online Information Review, 45*, 1290–1305. https://doi.org/10.1108/OIR-05-2020-0188

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. *Companion Proceedings of the ACM Web Conference 2023*, Austin, TX, 294–297. https://doi.org/10.1145/3543873.3587368

Ioannidis, J. P. A. (2019). What have we (not) learnt from millions of scientific papers with *P* values? *The American Statistician, 73*, 20–25. https://doi.org/10.1080/00031305.2018.1447512

Ipeirotis, P. (2023). Readability metrics. https://rapidapi.com/ipeirotis/api/readability-metrics

Jimenez, S., Avila, Y., Dueñas, G., & Gelbukh, A. (2020). Automatic prediction of citability of scientific articles by stylometry of their titles and abstracts. *Scientometrics, 125*, 3187–3232. https://doi.org/10.1007/s11192-020-03526-1

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). GPT-4 passes the bar exam. *SSRN*. https://doi.org/10.2139/ssrn.4389233

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*

(Research Branch Repor No. 8–75). Naval Air Station Memphis – Millington, TN: Chief of Naval Technical Training.

Kousha, K., & Thelwall, M. (2022). *Artificial intelligence technologies to support research assessment: A review.* arXiv. https://doi.org/10.48550/arXiv.2212.06574

Liu, X., & Zhu, H. (2023). Linguistic positivity in soft and hard disciplines: Temporal dynamics, disciplinary variation, and the relationship with research impact. *Scientometrics*. https://doi.org/10.1007/s11192-023-04679-5

Lu, C., Bu, Y., Dong, X., Wang, J., Ding, Y., Larivière, V., Sugimoto, C. R., Paul, L., & Zhang, C. (2019). Analyzing linguistic complexity and scientific impact. *Journal of Informetrics, 13*, 817–829. https://doi.org/10.1016/j.joi.2019.07.004

Ma, A., Liu, Y., Xu, X., & Dong, T. (2021). A deep-learning based citation count prediction model with paper metadata semantic features. *Scientometrics, 126*, 6803–6823. https://doi.org/10.1007/s11192-021-04033-7

McLaughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading, 12*, 639–646.

Mendeley. (2023). Mendeley. https://www.mendeley.com/search

Murray, R., Thow, M., Moore, S., & Murphy, M. (2008). The writing consultation: Developing academic writing practices. *Journal of Further and Higher Education, 32*, 119–128. https://doi.org/10.1080/03098770701851854

Nori, H., King, N., Mayer McKinney, S., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on medical challenge problems.* arXiv. https://doi.org/10.48550/arXiv.2303.13375

OpenAI. (2022). Introducing ChatGPT. https://openai.com/blog/chatgpt

OpenAI. (2023a). GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf

OpenAI. (2023b). Text completion. https://platform.openai.com/docs/guides/completion/introduction

Pandey Akella, A., Alhoori, H., Kondamudi, P. R., Freeman, C., & Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics, 15*, 101128. https://doi.org/10.1016/j.joi.2020.101128

Pei, Z., Yin, J., Liaw, P. K., & Raabe, D. (2023). Toward the design of ultrahigh-entropy alloys via mining six million texts. *Nature Communications, 14*, 54. https://doi.org/10.1038/s41467-022-35766-5

Peterson, R. A. (2000). A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis. *Marketing Letters, 11*, 261–275. https://doi.org/10.1023/A:1008191211004

Pulido, C. M., Redondo-Sama, G., Sordé-Martí, T., & Flecha, R. (2018). Social impact in social media: A new method to evaluate the social impact of research. *PloS one, 13*, e0203117. https://doi.org/10.1371/journal.pone.0203117

Sand-Jensen, K. (2007). How to write consistently boring scientific literature. *Oikos, 116*, 723–727. https://doi.org/10.1111/j.0030-1299.2007.15674.x

Scopus. (2023). Scopus. https://www.scopus.com/search/form.uri?display=basic#basic

Senter, R. J., & Smith, E. A. (1967). *Automated readability index* (Report No. AMRL-TR-66–220). Wright-Patterson Air Force Base, OH: Aerospace Medical Research Laboratories.

Sienkiewicz, J., & Altmann, E. G. (2016). Impact of lexical and sentiment factors on the popularity of scientific papers. *Royal Society Open Science, 3*, 160140. https://doi.org/10.1098/rsos.160140

Sommer, V., & Wohlrabe, K. (2017). Citations, journal ranking and multiple authorships reconsidered: Evidence from almost one million articles. *Applied Economics Letters, 24*, 809–814. https://doi.org/10.1080/13504851.2016.1229410

Tabone, W., & De Winter, J. C. F. (2023). Using ChatGPT for human-computer interaction: A primer. *Royal Society Open Science, 10*, 231053. https://doi.org/10.1098/rsos.231053

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics, 107*, 1195–1225. https://doi.org/10.1007/s11192-016-1889-2

Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. *Scientometrics, 115*, 1231–1240. https://doi.org/10.1007/s11192-018-2715-9

Wang, S., Liu, X., & Zhou, J. (2022). Readability is decreasing in language and linguistics. *Scientometrics, 127*, 4697–4729. https://doi.org/10.1007/s11192-022-04427-1

Warren, H. R., Raison, N., & Dasgupta, P. (2017). The rise of altmetrics. *JAMA, 317*, 131–132. https://doi.org/10.1001/jama.2016.18346

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research, 49*, 737–747. https://doi.org/10.1509/jmr.11.0368

Xie, J., Gong, K., Cheng, Y., & Ke, Q. (2019). The correlation between paper length and citations: A meta-analysis. *Scientometrics, 118*, 763–786. https://doi.org/10.1007/s11192-019-03015-0

Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023). *Exploring the limits of ChatGPT for query or aspect-based text summarization.* arXiv. https://doi.org/10.48550/arXiv.2302.08081

Zhang, B., Ding, D., & Jing, L. (2022). *How would stance detection techniques evolve after the launch of ChatGPT?* arXiv. https://doi.org/10.48550/arXiv.2212.14548

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). *Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT.* arXiv. https://doi.org/10.48550/arXiv.2302.10198

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.