Applied Intelligence (2024) 54:4172–4188
https://doi.org/10.1007/s10489-024-05298-0

# Analysing the impact of ChatGPT in research

Pablo Picazo-Sanchez[1] · Lara Ortiz-Martin[1]

Accepted: 28 January 2024 / Published online: 21 March 2024
© The Author(s) 2024

**Abstract**
Large Language Models (LLMs) are a type of machine learning that handles a wide range of Natural Language Processing (NLP) scenarios. Recently, in December 2022, a company called OpenAI released ChatGPT, a tool that, within a few months, became the most representative example of LLMs, automatically generating unique and coherent text on many topics, summarising and rewriting it, or even translating it to other languages. ChatGPT originated some controversy in academia since students can generate unique text for writing assessments being sometimes extremely difficult to distinguish whether it comes from ChatGPT or a person. In research, some journals specifically banned ChatGPT in scientific papers. However, when used correctly, it becomes a powerful tool to rewrite, for instance, scientific papers and, thus, deliver researchers' messages in a better way. In this paper, we conduct an empirical study of the impact of ChatGPT in research. We downloaded the abstract of over 45,000 papers from over 300 journals from Dec 2022 and Feb 2023 belonging to different research editorials. We use four of the most known ChatGPT detection tools and conclude that ChatGPT played a role in around 10% of the papers published in every editorial, showing that authors from different fields have rapidly adopted such a tool in their research.

**Keywords** Natural language processing (NLP) · Large language models (LLMs) · ChatGPT · Research

## 1 Introduction

In recent years, Natural Language Processing (NLP) has experienced significant advances thanks to the dedicated efforts of researchers and the advancements in other areas of Machine Learning (ML), such as deep learning. NLP is a field of Artificial Intelligence (AI) that focuses on the interaction between humans and computers, being the primary goal of making computers understand and generate text like persons. There are many scenarios where NLP can help us. For instance, NLP can generate text automatically in a wide variety of fields, can create bots that converse with people, can automatically summarise long and complex documents, explain difficult concepts using vocabulary and sentences accessible for all people, classify text based on different criteria, or help people to rewrite content for other audience.

The term Language Model (LM) refers to systems trained on string prediction tasks [1]. They are systems based on statistical models that assign a probability to a sequence of words based on the preceding context or the surrounding context to predict the likelihood of a word, string, or sentence. When these systems are trained with a large amount of data, we say they are Large Language Model (LLM), like OpenAI's GPT-3 [2], Google's GShard [3], and Switch-C [4]. By the end of February 2023, Meta publicly released LLaMA [5], a collection of foundation language models exclusively trained on publicly available data, ranging from 7B to 65B parameters and outperforms other models like GPT-3 trained with 175B parameters.

In December 2022, OpenAI released a web tool called ChatGPT based on GPT3. Such a tool interacts with users conversationally, one of the main characteristics that every generated text is unique. Since then, other important companies have also released similar tools like Bard [6], proposed by Google or directly releasing the LLM like the aforementioned LLaMA released to researchers by Meta. This boom has already impacted the way authors write books, and a recent study states that over 200 books on Amazon have

Pablo Picazo-Sanchez and Lara Ortiz-Martin contributed equally to this work

✉ Pablo Picazo-Sanchez
ppicazo@hh.se

Lara Ortiz-Martin
lara.ortiz.martin@gmail.com

[1] School of Information Technology, Halmstad University, Halmstad, Sweden

ChatGPT as part of the authors [7] (that number just an estimation since the real one is nearly impossible to determine as the same report claims). However, ChatGPT should not be an author, at least in research [8].

Recently, researchers used ChatGPT to generate 50 abstracts based on other papers published in JAMA, The New England Journal of Medicine, The BMJ, The Lancet and Nature Medicine. Later, they submitted such abstracts to academic reviewers, and they detected 63% of the AI-abstracts [9]. With such a tool generating "believable scientific abstracts", some authors claim that AI-abstracts will soon find their way into the literature [8]. One of the first consequences is that journals like *Nature* modified (in Jan 2023) the license and editorial policies to specify that text generated by ChatGPT cannot be used in the papers [8].

Before ChatGPT was released in April 2022, the impact of LLMs in research was already under discussion [10]. Now that ChatGPT is publicly accessible, it also opened the door for new issues. For instance, detecting when a piece of text comes from any AI-tool or humans is one of the most challenging ones.

Before ChatGPT was released in April 2022, the impact of LLMs in research was already under discussion [10]. Now that ChatGPT is publicly accessible, it also opened the door for new issues. For instance, detecting when a piece of text comes from any AI-tool or humans is one of the most challenging ones.

Inspired by the experiment recently carried out using ChatGPT to generate research abstracts [9], in this paper, we evaluate the impact of ChatGPT in research in the first quarter of 2023. We split our methodology into two main blocks: 1) ground truth, and; 2) impact in research. We first use two ground truth datasets, one with text produced by ChatGPT and humans and another composed of over 2,000 papers published in 2010, where LLMs were underdeveloped, to analyse the reliability of four of the most known ChatGPT detection tools. Later, we download over 45,000 abstracts from 317 different journals published between Dec 2022 and Feb 2023 and conclude that ChatGPT played a role in over 10% of the papers published in every editorial during that period, thus empirically demonstrating how some of the predictions recently made [8] became a reality within a few weeks after ChatGPT was released.

The rest of the paper is organised as follows. Section 2 introduces the reader to the topic of NLP and presents some definitions and concepts about statistics. Section 3 explains how we measure the accuracy of the four detectors using two ground-truth datasets. Section 4 is the core of our analysis. We discuss, in Section 5, the main authorship policies that editorials adopted as a consequence of the impact of ChatGPT in research. Finally, Section 6 concludes the paper.

## 2 Background

In the following, we delve into NLP, tracing its historical development, exploring the fundamental metrics used to evaluate the quality and performance of LMs, and addressing the vital components of Null Hypothesis Significance Tests (NHST) and AI-content detection tools. By examining these aspects in detail, we aim to provide a comprehensive overview of the field's evolution, assessment methodologies, and the challenges and solutions associated with AI-assisted content detection tools.

### 2.1 History

Back in 1900, Ferdinand de Saussure, a Swiss linguist, set the first stones of NLP, proposing the concept of "Language as a Science" in his *Cours de linguistique générale* book [11], published by two of his students (Albert Sechehaye and Charles Bally) after he passed away.

Since then it was not until 1950 that Alan Turing stated that if a machine could participate in a conversation as if it were a human with no noticeable differences, then we can assume it can think. Over the 60's, Noam Chomsky—one of the fathers of NLP, proposed a mathematical theory of syntax and semantic structures to model the language [12], i.e., a set of rules to generate and interpret a language (generative model). Since then, two main lines of research in NLP, symbolic and stochastic. The symbolic research focused on formal languages to create rules to handle those languages, leading to a rule-based system. On the other hand, in stochastic research, the focus was on statistical and probabilistic methods of NLP, e.g., given the latest letter ($x$), what is the next one ($x+1$)?

Until mid 80's there was a relatively little advance on NLP. However, with the statistical revolution and more powerful computers, probabilistic and statistical methods gained popularity among NLP models. Some examples that are still useful nowadays are *N-Grams* [13] and Long Short-Term Memory (LSTM) [14].

N-grams are contiguous sequences of n symbols, usually words in NLP context, where every word has a probability based on the previous one.

Traditional Neuronal Network (NN) suffer from lack of memory problem, i.e., their learning model is based on the information at some point in time and not on the previous knowledge. To overcome such a constraint, researchers proposed Recurrent Neuronal Network (RNN), which are, essentially, NNs with inner loops in the model so the information can persist. However, RNN suffer from both vanishing gradient problem [15] and long term dependency of words.

To solve such a problem, LSTM [14] modifies the RNN such that they can operate as RNNs (retaining information for some period of time) as well as keeping the information longer periods and thus solving the long-term dependency problem.

## 2.2 Metrics

To measure the quality of a LMs, there are two main metrics: *perplexity* and *burstiness*. As one of the detectors uses these two values to determine whether an input text does (not) come from a LLM like ChatGPT, in the following, we explain each one in more detail.

**Perplexity** The perplexity measures how well a probability distribution predicts a sample. A lower perplexity indicates a higher predictive accuracy of the distribution for the sample. Alternatively, lower perplexity corresponds to reduced randomness in the text. Since LLMs are made to maximise the text probability, the consequence is that, therefore, they minimise the perplexity.

To measure the perplexity, we first need a test set, i.e., a set of values coming from a probability distribution not used during the training phase. More formally, let $P$ be a probability distribution, $q$ be a probability model, $x_1, \ldots, x_n$ be a test sample derived from $P$, and $N$ the number of elements in the test sample, the perplexity of the model $q$ is defined as:

$$\left(\prod_{i=1}^{n} q(x_i)\right)^{1/N}$$

When used in NLP, perplexity is the inverse probability of the test set normalised by the number of words $(w_1, \ldots, w_n)$:

$$\left(\prod_{i=1}^{n} \frac{1}{P(w_i|w_1 \ldots w_{n-1})}\right)^{1/N}$$

**Burstiness** In NLP, the term burstiness describes the tendency of word recurrence where a word is more likely to occur if it has already appeared in the text. This implies that after the first appearance of a term, it becomes less significant. Also, there is a positive correlation between the burstiness of a word and its semantic content, meaning that more informative words are also more bursty [16, 17].

## 2.3 Null hypothesis significance tests

Null Hypothesis Significance Tests (NHST) is a statistical inference method by which an experimental factor is tested against a hypothesis of no effect or relationship based on a given observation. NHST is the combination of the *significance testing* [18] and the *hypothesis testing* [19]. The final decision of NHST are based on three main concepts (summarised in Table 1): 1) $\alpha$, which is the probability of making a type I error, also known as statistical significance. It is usually 0.05 although in the original document there is no mathematical justification for selecting a particular p-value [20]; 2) $\beta$, which is the probability of making a type II error, and $1 - \beta$ the power of a test, and; 3) *p-value*, which is the probability of obtaining a new sample far from the null hypothesis data distribution.

In significance testing, Fisher [18] relied on the null hypothesis ($H_0$) and the exact p-value to reject or accept $H_0$, i.e., the lower the p-value is the more chances we have to reject $H_0$ [21]. However, the hypothesis testing proposed by Neyman and Pearson [19] denied the interpretation of a p-value as a measure of evidence. Instead, they introduced the error rate, where type I error means false positive errors and type II error false negatives. They also proposed two competing hypotheses $H_0$ and $H_1$ (see Table 1) and used the p-value to assess the likelihood of observing the data if $H_0$ were true. According to the authors: "*the "best" test is the one that minimises false negatives subject to abound on false positives, the latter being the significance level of the test*" [22].

## 2.4 AI-content detection tools

With the recent release of ChatGPT, detection tools are becoming popular: Content at scale AI detector, GPTZero, Writer AI content detector, and zeroGPT are some of the most popular ones (in alphabetical order).

Content at scale offers a free AI-detection tool "trained on billions of pages of data, and can accurately forecast the most probable word choices that lead to a higher AI detection probability" [23]. In particular, they evaluate the input

**Table 1** Standard confusion matrix

| | | Predicted | |
|---|---|---|---|
| | | Yes<br>(Null Hypothesis $H_0$) | No<br>(Alternative Hypothesis $H_1$) |
| Actual | Yes<br>(Null Hypothesis $H_0$) | TP (hit)<br>$(1 - \alpha)$ | FN (miss)<br>(Type II error)($\beta$) |
| | No<br>(Alternative Hypothesis $H_1$) | FP (false alarm)<br>(Type I error)($\alpha$) | TN (hit)<br>$(1 - \beta)$ |

**Table 2** HC3 dataset pre-processing, where Total is the sentences in the dataset, Duplicates is the sentences without duplicates, len(text)≥420 are sentences longer than 420 characters, and Sample is the number of sentences analysed

| HC3 | Total | Duplicates | len(text)≥420 | Sample (30%) |
|---|---|---|---|---|
| ChatGPT | 23,867 | 23,363 | 22,747 | 6,824 |
| Human | 23,867 | 23,363 | 13,287 | 3,986 |

text according to three main parameters: 1. predictability; 2. probability, and; 3. pattern. In addition to these scores, they also provide a final score "Human Content Score" that predicts how likely the input text comes from a human or a LM. This tool needs at least 25 words to work and up to 25,000.

One of the most popular AI-content detector tools is GPTZero [24], claiming to be the most important AI-detector tool in the world with over 1 million users and to achieve an Area Under the Curve (AUC) score of 0.98 under certain circumstances[1]. It outputs three scores: perplexity, burstiness, and a final one that predicts how likely a human has generated the input text. This tool needs at least 250 characters.

Writer provides an AI content detector [25] that evaluates how predictable an input text is, i.e., how similar it is to other similar texts. They show examples of how to generate text unpredictably by its detector. As an output, the tool shows a detection score. This tool accepts text up to 1,500 characters.

Finally, ZeroGPT [26] analysed over 10M articles and text coming from both LMs and humans and proposed an algorithm with an accuracy rate of text detection higher than 98%. The output is composed of a message and a score, being the score the probability that the text comes from a human. Although the tool has no minimum length, it alerts users that results are only reliable if the text has at least 420 characters.

## 3 Ground truth

To measure the impact of ChatGPT in research, first, we need to measure how accurate the four ChatGPT-detection tools are. To do so, we use three repositories, one based on a public label dataset called HC3 database [27], another one composed of 776 research papers we downloaded from 48 journals from Elsevier published in December 2010, and a final one of 1,932 research papers from 86 journals from IEEE published in the same period.

For both datasets, we obtained the confusion matrices and use them to compute the Confidence Interval (CI). Since i) the experiments are independent; ii) the number is larger than 30, and; iii) there are only two possible outputs (Human and ChatGPT), we can use the Wilson score to compute the binomial proportion CI [28].

We calculated the 95% CI for the accuracy of every detector for HC3, Elsevier, and IEEE ground truth datasets. With such an interval, if we rerun our evaluation on another random dataset, we should expect, with 95% confidence, that the accuracy will fall within this CI.

In all the experiments we present in this paper, to fulfil the constraints some detectors impose about the length of the text and thus get more reliable results, we specifically focus on text with more than 420 characters.

In the following, we explain in more detail the methodology we followed as well as present the results of this experiment.

### 3.1 HC3

Recently, researchers released a dataset called HC3 [27] publicly available[2] composed of over 23k questions answered by humans and ChatGPT. Given the dataset's amount of data, we randomly select a sample of 30%, after filtering the sentences with more than 420 characters. We analysed 10,891 sentences (6,824 generated by ChatGPT and 3,986 by humans). We include a summary of the numbers of the analysed sentences in Table 2.

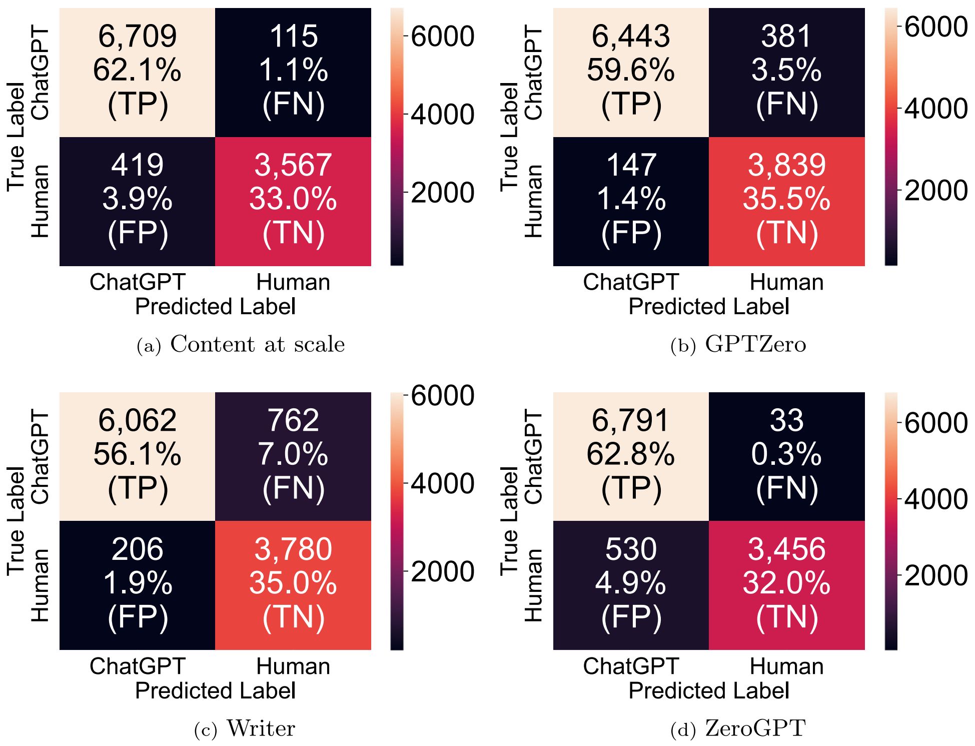
Once we have our ground truth dataset, we input every one of the detectors with the labelled sentences. In Fig. 1, we can see the confusion matrix that every detector produced, while Table 3 summarises the metrics associated with every detector. In particular, "Right" stands for texts correctly classified (TP+TN) and "Wrong" otherwise (FP+FN). We calculate the accuracy as usual, i.e., sentences correctly classified divided by the total amount of sentences x100.

However, one aspect is worth discussing based on our results: type I and II errors. Although the four detectors perform similarly, slight differences might make us prefer one detector over the others, depending on how conservative we are. For instance, in the case of Nature, the editorial banned any LLMs from helping in the writing process. Editors should then minimise type II errors. Thus, they should use ZeroGPT, the detector that achieves the best results despite performing the worst in terms of false positives. On the other hand, if another editorial prefers minimising type I error, i.e., detecting whether a person produced the text, GPTZero should be

[1] https://gptzero.me/faq

[2] https://huggingface.co/datasets/Hello-SimpleAI/HC3

**Fig. 1** Confusion matrix of the four ChatGPT-content detectors evaluated with HC3 dataset



(a) Content at scale

(b) GPTZero

(c) Writer

(d) ZeroGPT

the option. However, the amount of false negatives, i.e., text generated by ChatGPT, is higher than others.

### 3.2 Ground truth in research

We carry out one more experiment to measure how reliable the results are when using a specific type of text, like scientific papers. To do so, we gathered 776 research papers we downloaded from 48 journals from Elsevier and 1,932 from 86 IEEE journals published in the same period and evaluated the output of the four detectors. Note that no LLMs were available by that year to help researchers write papers. The reason why we chose Elsevier and IEEE is arbitrary, and other editorials or journals could have been chosen instead.

Similar to our previous experiment, in Table 4, we include the statistics we got from both Elsevier and IEEE analysed papers. We also added in Appendix A the confusion matrices of IEEE (see Fig. 5) and Elsevier (see Fig. 6). Since ChatGPT did not exist as of 2010, we do not have false positives nor true negatives as in Fig. 1, having only two possible results, either true positives (text generated by humans) and false negatives (text generated by humans but classified by detectors as being generated by ChatGPT).

Interestingly, although we observe that the results of all the detectors but Writer are better than those from the previous experiment when we used the HC3 dataset, probably explained because of the number of experiments (input text), the order of the detectors (in terms of accuracy) remains the same. This is also the case when comparing the results among different editorials. In other words, no matter which editorial the analysed text comes from, the detector with the highest accuracy is GPTZero, later comes Content at scale, ZeroGPT and, Writer.

## 4 ChatGPT in research

According to some recent statistics [29], the world top-6 of the editorials where researchers publish their scientific work,

**Table 3** Accuracy of the 4 detectors using a subset of (30%) of the HC3 dataset

| | Tool | Right | Wrong | Total | Accuracy |
|---|---|---|---|---|---|
| HC3 | Contentatscale | 10,264 | 546 | 10,810 | 94.9% ± 0.4% |
| | GPTZero | 10,277 | 533 | 10,810 | 95.1% ± 0.4% |
| | Writer | 9,843 | 967 | 10,810 | 91.1% ± 0.6% |
| | ZeroGPT | 10,245 | 565 | 10,810 | 94.8% ± 0.4% |

Right stands for texts correctly classified and Wrong otherwise

**Table 4** Accuracy of the 4 detectors using Elsevier and IEEE papers from Dec 2010

| | Tool | Right | Wrong | Total | Accuracy |
|---|---|---|---|---|---|
| IEEE | Contentatscale | 1,888 | 44 | 1,932 | 97.7% ± 0.8% |
| | GPTZero | 1,924 | 8 | 1,932 | 99.6% ± 0.4% |
| | Writer | 1,705 | 227 | 1,932 | 88.3% ± 1.5% |
| | ZeroGPT | 1,853 | 79 | 1,932 | 95.9% ± 1.0% |
| Elsevier | Content at scale | 752 | 24 | 776 | 96.9% ± 1.5% |
| | GPTZero | 771 | 5 | 776 | 99.4% ± 0.9% |
| | Writer | 672 | 104 | 776 | 86.6% ± 2.6% |
| | ZeroGPT | 722 | 54 | 776 | 93.0% ± 2.0% |

Right stands for texts correctly classified and Wrong otherwise

in terms of publications, is: 1) Elsevier; 2) Springer; 3) Wiley; 4) MDPI; 5) Taylor & Francis, and; 6) IEEE.

This paper evaluates 4 of the most important editorials regarding their volume, i.e., Elsevier, IEEE, MDPI and Springer. In addition, we also analyse two more journals, given their importance in research: Science, and The Lancet. For every one of the journals, we crawled their web pages and focused on those with *impact factor*.

For some editorials, like MDPI, we analyse all the journals, i.e., 38 with impact factor as of Feb 2023. Others, like Elsevier, given the number of journals they have, we used their web filtering option to restrict our analysis to journals that mainly belong to "Computer Science".

In more detail, for every paper we get i) the abstract; ii) author(s); iii) affiliation(s), and; iv) date of publication. In Table 5, we include a summary of the editorials, the number of journals, as well as those papers we finally analyse (i.e., those that have more than 420 characters in their abstracts).

We follow the same methodology as for the ground truth with the obvious difference that the text is not labelled. In total, to measure the impact of ChatGPT in research, we crawled 317 journals from 6 editorials totalling 48,415 papers and analysed the abstract of 45,180, those with more than 420 characters. In the following, we explain in more detail our findings, summarised in Table 6.

**Table 5** Editorials (alphabetical order) with the number of journals, papers we downloaded, and those we finally analysed from December 2022 to February 2023

| Editorial | Journals | Papers | Analysed |
|---|---|---|---|
| Elsevier | 57 | 7,827 | 7,479 |
| IEEE | 138 | 14,387 | 14,387 |
| MDPI | 38 | 23,900 | 21,273 |
| Science | 1 | 372 | 240 |
| Springer | 82 | 1,401 | 1,123 |
| The Lancet | 1 | 528 | 528 |
| TOTAL | 317 | 48,415 | 45,030 |

## 4.1 Editorials

The first thing we noticed is the unbalanced dataset in terms of published papers (see Fig. 2 and Table 5). For instance, if we take the papers we analysed within 3 months from MDPI as a reference, we got almost 2 times more papers than IEEE, almost 3 times more papers than Elsevier, and; 20 times more papers than Springer. This is a particular case since MDPI is the editorial with fewer journals (38). However, MDPI is not the only example of the unbalanced dataset. If we use IEEE, we got almost 2 times more papers with respect to Elsevier and 10 times more with respect to Springer.

Intuitively, we might think that the more papers in a journal, the more positives the detectors mark, i.e., there is a linear correlation between these two factors. This is the case of Writer (see Fig. 3c). However, such a correlation does not hold in the other detectors, as we can see in Fig. 3a, b and d. There are many positives (abstracts detected to be written with the help of ChatGPT) in journals where the total published papers are around 500 within 3 months. This also holds after applying the corresponding error margin previously computed in Section 3.

All this also led us to wrongly think that there might be a correlation between detectors and editorials (see Table 6). However, we are mixing different topics in research, e.g., The Lancet is focused on Medicine, Science is on "all fields of science", and so are the journals we analysed from IEEE, MDPI, and Springer. Elsevier might be the only editorial that might be controversial because we tried to filter journals by "Computer Science" area of knowledge. We manually checked the "Subject and Category" from Scimago of all the journals we used from Elsevier. While the majority of the journals have Computer Science as the main subject, they also include areas like Engineering (29 out of 57), Mathematics (24 out of 57), and Social Sciences (15 out of 57), among others.

We manually inspected some of the results and got that most of the positives come from journals whose reviewing time is less than 4 weeks. As an example, The Lancet aims at publishing papers within 4 weeks under its *Swift*+ and

**Table 6** The number of publications and percentage (over the overall analysed abstracts) that detectors marked as being written with the help of ChatGPT

| Editorial | Analysed | Content at scale | GPTZero | Writer | ZeroGPT |
|---|---|---|---|---|---|
| Elsevier | 7,479 | 212 (2.8%) | 48 (0.6%) | 872 (11.7%) | 546 (7.3%) |
| IEEE | 14,387 | 223 (1.6%) | 33 (0.2%) | 1,470 (10.2%) | 671 (4.7%) |
| MDPI | 21,423 | 2,537 (11.8%) | 984 (4.6%) | 4,237 (19.8%) | 5,620 (26.2%) |
| Science | 240 | 37 (15.4%) | 17 (7.1%) | 15 (6.2%) | 118 (49.2%) |
| Springer | 1,123 | 105 (9.3%) | 38 (3.4%) | 187 (16.7%) | 183 (16.3%) |
| The Lancet | 528 | 114 (21.6%) | 49 (9.3%) | 24 (4.5%) | 307 (58.1%) |
| TOTAL | 45,180 | 3,228 (7.1%) | 1,169 (2.6%) | 6,805 (15.1%) | 7,445 (16.5%) |

*fast-publication* modes [30]; some IEEE journals like IEEE Access[3] (the outlier of Fig. 2), MDPI[4], and Science[5] have similar time frames, and; some Springer journals like Expert Systems with Applications[6] (the outlier of Fig. 2) even less. This is a reasonable explanation since our study analyses papers published within the last 3 months.

As a summary, having shown in Table 4 that the accuracy of the four detectors does not depend on the (two) editorials we used for the ground truth (see Section 3), we assume that the probability of detecting ChatGPT in one abstract is the same independently of the editorial. Note that this differs from assuming that papers are not dependent on journals, which is not usually true.

In Table 6, we include the overall results after analysing all the abstracts. If we analyse the results in more detail and focus on ZeroGPT, we get that ChatGPT was involved in 58% of the abstracts published in The Lancet between Dec 2022 to Feb 2023. Also, ChatGPT was part of over half of the abstracts published in Science within that period. MDPI follows the top-3 rank with more than 25% abstracts. It is worth remembering that ZeroGPT is the detector with fewer False Negatives (0.3% in the HC3 ground truth experiment). Also, this detector has the highest FP rate, i.e., it marks more human papers than being written by ChatGPT. In conclusion, assuming that the accuracy is over 95% (extracted from the intersection between all the ground truth datasets—HC3, IEEE, and Elsevier), the final numbers do not statistically deviate from the presented ones significantly.

On the other hand, it is also interesting to see that GPTZero, the one with the highest True Positive rate and lowest type I error, i.e., the one that predicts with the highest accuracy whether a text comes from ChatGPT, corroborates the order in which editorials published papers where ChatGPT is involved. The main difference lies in the final numbers. This time, given that GPTZero minimises type I error (i.e., FP) instead of type II error (FN) like ZeroGPT. As a result, it marks over 9% of the papers in The Lancet, over 7% of papers in Science, and 4.60% in MDPI. Note that the order of the editorials is the same as before, but this detector is more conservative, as we expected, given the ground truth experiment.

## 4.2 Affiliations

To demographically evaluate the usage of ChatGPT, we first tried to infer the nationality of the authors. However, guessing the nationality from both the name and the surname is an active line of research [31]. Instead, we gather the country of the authors' affiliation, whenever they declared it, and plot the top-10 per editorial in Fig. 4. Note that we count every country only once, i.e., if there are 4 authors in a paper whose affiliations belong to the same country, we count that country once.

Interestingly, we can see how authors from China dominate the editorials. Science and The Lancet are the exceptions, where authors from US and UK are the most common. Looking at the top-10 ranking, and more concretely to the Chinese-dominated editorials, papers published with affiliations from Asia (e.g., China, India, Iran, Japan, and South Korea) usually double or even triple (like in IEEE) those from other continents.

We can argue whether these results are far from those affiliations marked as being human-written. To answer that, in Table 7 (see Appendix A), we include the top-10 countries in terms of the number of publications. We can observe how, in general, these countries' positions do not match those presented in Fig. 4. Using Elsevier as an example (first one in alphabetical order), China is the country that publishes the most and matches with the country that uses ChatGPT the most. However, the other countries do not occupy the same positions in both ranks. Turkey, Iran, and South Korea are in the top-10 countries that use ChatGPT the most; however, they are not in top-10 countries that publish the most. The same conclusions arise when analysing authors' affiliations in the other editorials and Science and The Lancet.

[3] https://ieeeaccess.ieee.org/about-ieee-access/frequently-asked-questions/

[4] https://blog.mdpi.com/2021/10/27/submission-questions-answered/

[5] https://www.science.org/content/page/science-information-authors

[6] https://journalinsights.elsevier.com/journals/0957-4174/oapt

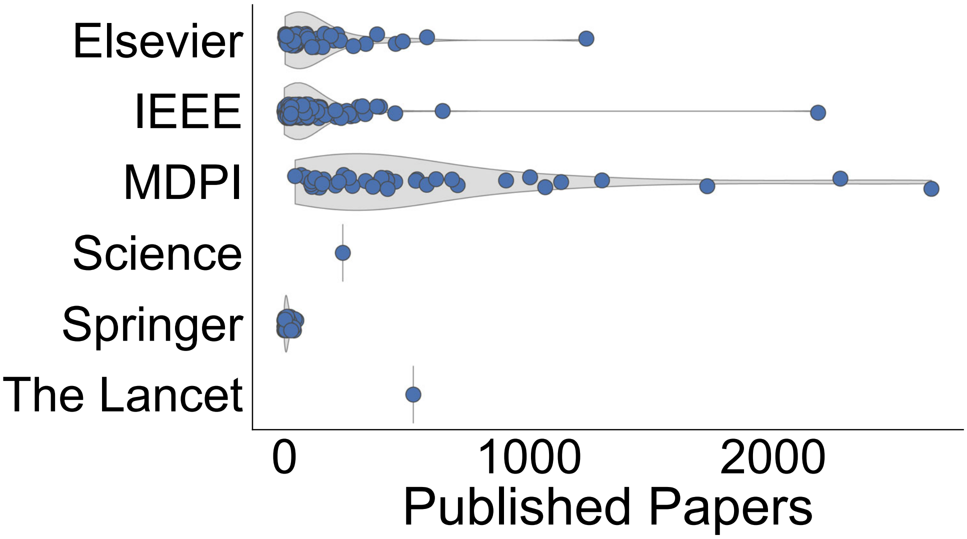


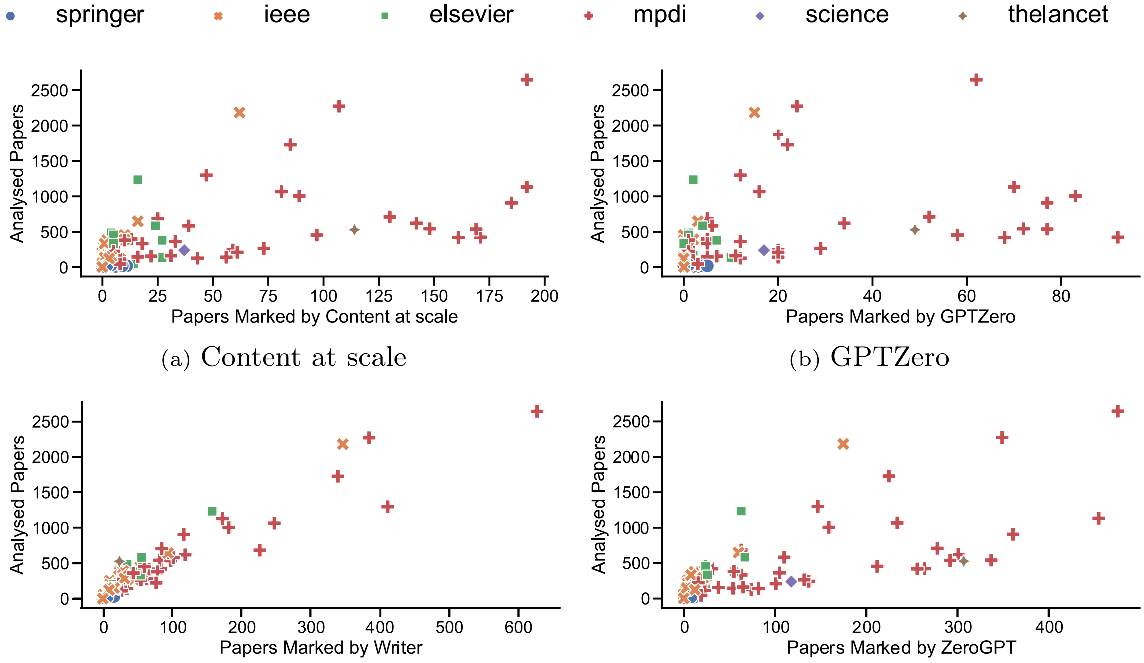
**Fig. 2** Papers with abstracts larger than 420 characters published from Dec 2022 to Feb 2023. Every point in the violin plot belongs to a journal of that editorial

We also analysed the relation of published papers versus those written with the help of ChatGPT (see Table 8 in Appendix A). It is interesting to see how the countries in this list do not usually have English as their official language-concluding that LLMs might be beneficial for researchers to deliver their message better and more clearly.

### 4.3 Discussion

In this paper, we presented the conclusions we extracted after analysing more than 45,000 research abstracts from different journals and editorials. Although we are aware of the impact and consequences, we tried to corroborate the results by contacting a small subset of authors to check whether they used or not ChatGPT in research without success. The results we presented are an effort to measure the impact of ChatGPT in research and not to judge authors or the quality of the papers.

We know that the number we got is strongly influenced by the time every journal takes for the reviewing process. Since the goal of the paper is to measure the impact as of March 2023, we expect that the numbers we presented will grow significantly within the following months.



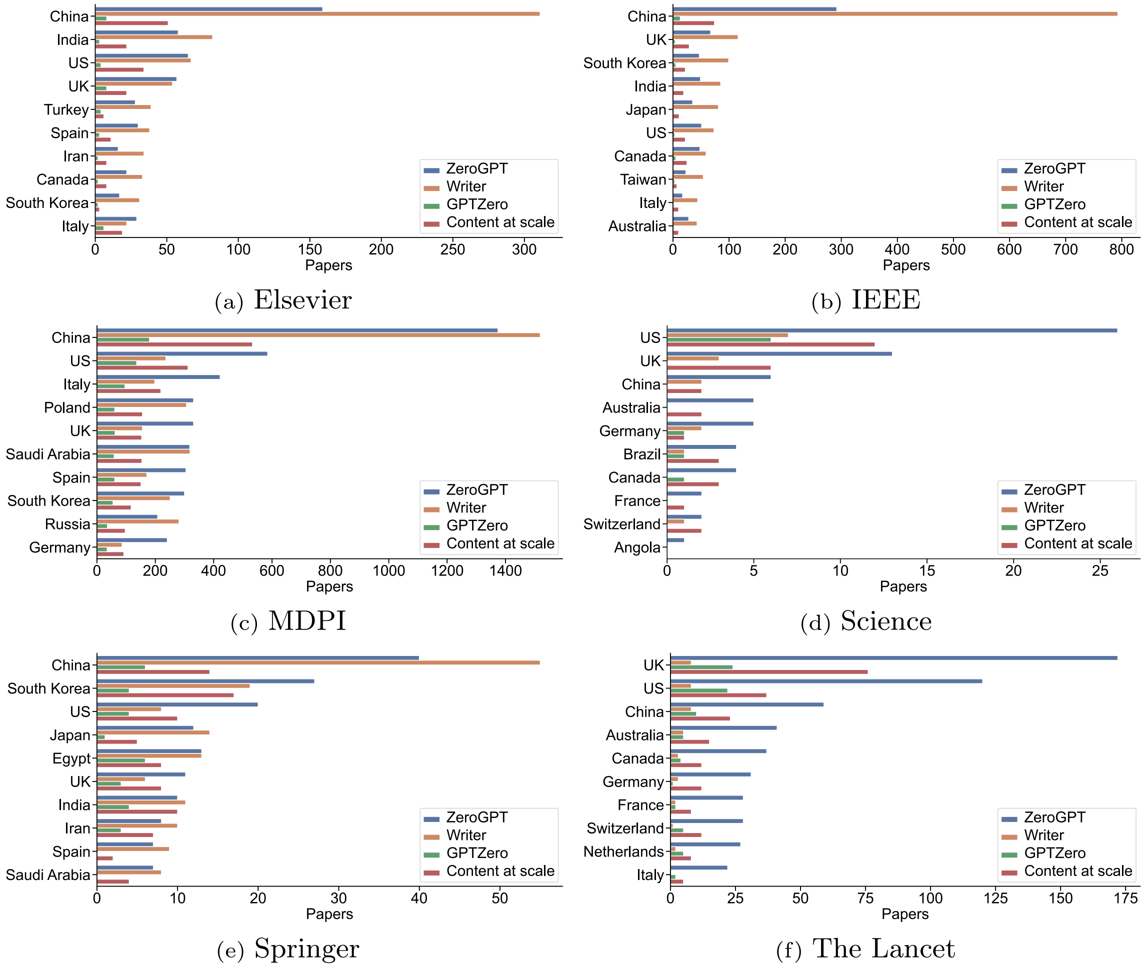
(a) Content at scale

(b) GPTZero

(c) Writer

(d) ZeroGPT

**Fig. 3** Number of detected abstracts by everyone of the detectors in relation to the number of published papers



(a) Elsevier

(b) IEEE

(c) MDPI

(d) Science

(e) Springer

(f) The Lancet

**Fig. 4** Top-10 of countries based on the affiliation of the authors that use LLMs in their abstracts per editorial

We analysed abstracts of papers published in journals with impact factor, and if the number of journals was large enough, then we restricted the analysis to a small subset of Computer Science, being all the journals and papers in English. It might have been interesting to perform a similar analysis by splitting the papers into different research areas. We leave this analysis for future work in this area.

For privacy reasons and to avoid any retaliations, we do not disclose the dataset we used for the analysis. However, we provided enough details along the paper to reproduce the same experiments. However, note that detectors might improve their detection mechanism and therefore the results might vary. It is also worth remembering that HC3 dataset is public and can be easily used to corroborate our results in the ground truth section.

The results show that authors have rapidly adopted such technology in their writing process. However, we need to find out to which extent authors use ChatGPT, e.g., producing the entire text from scratch or simply rewriting it for readability purposes. We carried out one more experiment and analysed all the detected abstracts with two commercial plagiarism tools, but unfortunately, neither output was conclusive.

We can clearly identify numerous advantages and drawbacks regarding the usage of LLMs in research. Researchers can benefit from LLMs enhancing their efficiency and speed in several ways, such as generating drafts, improving writing speed, and refining the clarity and coherence of written content. LLMs prove beneficial in aiding researchers to express complex ideas effectively. Furthermore, they serve as a valuable tool for non-native English speakers, offering language support and helping overcome language barriers, thus enhancing accessibility.

However, LLMs come with their set of challenges [32]. One notable concern is the potential lack of domain-specific knowledge. While proficient in general language use, these models require researchers to verify and add information related to their specific research. There is also a risk of unintentional plagiarism, as the suggestions made by LLMs may inadvertently resemble existing published content. It is crucial to emphasise the ethical use of LLMs, particularly in adhering to authorship policies outlined by journals and editorials. Authors remain responsible for the produced text.

## 5 Authorship policy

This section summarises editorials' main actions towards using LLMs in research articles. To provide a temporal perspective, we categorise this analysis into March 2023 and December 2023, allowing us to observe the evolution of editorials' perspectives on the usage of LLMs. Finally, we explore the influence of the Committee on Publication Ethics (COPE) in aligning the editorial policies of the journals and editorials studied, working towards achieving greater standardisation in practices.

March, 2023

**Elsevier** This editorial was the first one that published a Corrigendum [33] on a published research paper where ChatGPT was one of the authors [34]. Since then, Elsevier decided that ChatGPT cannot be part of the author list anymore as its Publishing Ethics document states [35]. Also, this editorial does not prevent authors from using LLMs but asks authors who use them to add a statement at the end of their manuscript entitled "Declaration of AI and AI-assisted technologies in the writing process" where authors have to indicate the tool they use as well as the reason [36].

**IEEE** We could not find any reference to ChatGPT nor LLMs in any of their policies.

**MDPI** Similar to IEEE, we could not find any reference to ChatGPT nor LLMs in any of their policies.

**Science** In the editorial policy, this journal has a subsection dedicated to AI [37]. In it, the journal states that *"text generated from AI, machine learning, or similar algorithmic tools cannot be used in papers published in Science journals, ...without explicit permission from the editors"*. Also, similar to other editorials, the policy mentions that AI cannot be listed as an author of any Science journal paper.

**Springer** The authorship policies in this editorial are relatively transparent about the authorship and the usage of LLMs like ChatGPT. There is a document called "Springer Authorship Principles" where we can find that authors can use LLMs if they are not part of the author list. Also, authors must state how they use LLMs in the Methods section (or a suitable alternative part of the article) [38].

**The Lancet** It includes a subsection in the Authors Guidelines document [30] dedicated to using AI in scientific writing. It states that LLMs are not authors. However, they can use these tools to *"improve readability and language of the work and not used to replace researcher tasks such as producing scientific insights, analysing and interpreting data, or drawing scientific conclusions"*. Also, when used, authors are responsible and accountable for the entire content and explicitly mention it at the end of the article.

**ACM** This editorial initiated a process to update its authorship policy to accommodate LLMs like ChatGPT and *"provide clear guidelines to the community for the appropriate use of these tools in ACM Publications"*. In this draft, contrarily to other editorials, ACM sees some potential in LLMs and instead of banning papers written with (the help of) LLMs, it argues for responsible use of these models with four main considerations: 1. LLMs are not authors; 2. authors are responsible of all the content written in the article; 3. LLMs cannot be used to plagiarise, misrepresent, or falsify content, and; 4. if used to create content, authors should explicitly mention it in the paper.

This mention depends on how authors use LLMs ranging from a footnote (limited to phrases or sentences) and adding a general disclaimer in the Acknowledgements section. If LLMs produce larger text (paragraphs or subsections), authors should provide information like which tool they use and the text of the prompts provided as input in an Appendix or a Supplementary Material document.

Dec, 2023

**Elsevier** This editorial has updated and extended the guidelines document [39], incorporating several changes, including modifications to the reviewing process. Notably, the editorial prohibits reviewers from using any online tool to write or enhance the quality of reviews for authors. Additionally, Springer requires authors to include a statement at the end of the paper disclosing the usage of generative AI in the writing process. Such a statement should be placed in a new section entitled "Declaration of Generative AI and AI-assisted technologies in the writing process" and should contain: *"During the preparation of this work the author(s) used [NAME TOOL / SERVICE] in order to [REASON]. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication'*.

**IEEE** This editorial has included a new subsection titled "Guidelines for Artificial Intelligence (AI)-Generated Text" where it differentiates between two types of use: 1) Automatic content generation and 2) the edition and grammar enhancement. For the former, authors must disclose in the acknowledgements section the AI system used, specifying the sections of the article where it was used and a brief explanation of the AI's contribution. In the case of the latter, as it is acknowledged as a common practice, authors are not obligated to disclose the use of AI, as it generally falls outside the scope of the policy mentioned above. IEEE also remarks that AI-tool cannot be listed as a co-author of the research paper [40].

**MDPI** This editorial has recently updated the authorship policy [41], and similar to other editorials, it allows authors to use LLMs during the writing process. However, MDPI requires authors to be fully transparent and add in the "Acknowledgments" section which tools they use and how they use them in another section called "Materials and Methods".

**Science** In November 2023, Science implemented a revised policy concerning the usage of generative AI and LLMs [42]. Like other editorials guidelines, Science does not allow any AI-tool to be part of the authors nor part of the references. Authors using AI for writing assistance must explicitly disclose this in the cover letter, and the "Acknowledgments" section of the manuscript. This disclosure should include information like the full prompt they used, the AI tool, and its version. Authors are accountable for the accuracy of the work and for ensuring that there is no plagiarism. Editors may only proceed with manuscripts if AI is correctly used. Also, Science emphasises that reviewers should not use AI technology in generating or writing their reviews because this could breach the confidentiality of the manuscript.

**Springer** While the editorial's policy remains unchanged, it has introduced a dedicated AI-specific section. Also, this editorial has included a new "AI-powered scientific writing assistant" available for authors [43]. The editorial claims this new tool is trained explicitly on academic literature, covering over 447 areas of study, more than 2,000 field-specific topics and drawing from a vast database of over 1 million papers.

**The Lancet** This journal has maintained its authorship structure, aligning it with the Elsevier editorial guidelines [44].

**ACM** This editorial presents the definitive version of the authorship policy, tailored to include LLMs. ACM categorised and slightly expanded upon the initial draft, allowing authors to employ LLMs similar to Grammarly for enhancing text quality in spelling, grammar, punctuation, clarity, and engagement. In doing so, there is no requirement to disclose the usage of these tools in the paper.

**Committee on publication ethics** Committee on Publication Ethics (COPE) is a not-for-profit organisation to establish ethical best practices in scholarly publishing. It was established in 1997 by a group of medical journal editors to address concerns related to publication misconduct, including issues like plagiarism, redundant publication, fraudulent data, unethical research, and breaches of confidentiality [45].

COPE addresses the usage of AI in research explicitly [46], emphasising that authors are fully responsible for the content of their manuscript, including those parts generated by AI tools, and are accountable for any violations of publication ethics. Since AI tools lack legal entity status, authors must handle assertions regarding conflicts of interest and manage copyright and license agreements. COPE explicitly prohibits listing AI tools as co-authors.

Furthermore, authors must be transparent using AI tools in writing, images or graphical creation or in the collection and analysis of data. To do so, authors should disclose in the "Materials and Methods" section, or a similar one, of the paper how they use AI, which tool they used, and provide any necessary detailed information.

All the editorials we analysed in this work are members of COPE (e.g., Elsevier, MDPI, Science, Springer) or support and follow best practice guidance from COPE (e.g., ACM, IEEE, The Lancet).

## 6 Conclusions

Between December 2022 and February 2023, a substantial shift was observed in the academic publishing landscape, primarily driven by integrating AI language models like ChatGPT. Our analysis revealed that ChatGPT played a role in the creation of over 10% of the papers published during this time frame, confirming that AI tools are increasingly shaping scholarly content.

In this study, we examined the impact of ChatGPT on research during the first quarter of 2023. Our methodology was divided into two main blocks: "ground truth" and "impact in research". We assessed the reliability of four prominent ChatGPT detection tools using two ground truth datasets-one comprising text generated by ChatGPT and humans, and the other composed of papers from 2010 when LLMs were in their early stages of development.

By analysing more than 45,000 abstracts from 317 diverse journals published between December 2022 and February 2023, we empirically demonstrated that ChatGPT had a notable influence on over 10% of the papers published across various editorial platforms during this brief period.

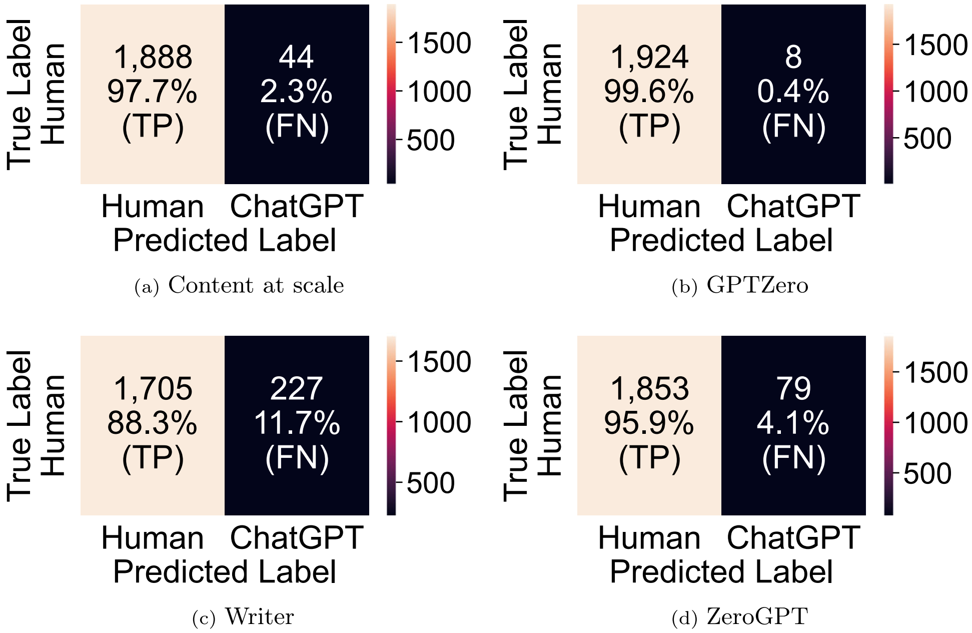
This conclusive evidence underscores the rapid transition toward AI-augmented scholarly content. As researchers continue to advance NLP and as AI tools like ChatGPT become more integrated into academia, it is imperative that we not only embrace tools like ChatGPT but also prepare for the emergence of similar models. We must adapt to coexist with LLMs and develop precise strategies, like co-authorship detection techniques [47], to tackle challenges like disseminating fake news and the potential for fake academic papers.

## Appendix A

In Fig. 5 we include the confusion matrices of the four detectors using as input 1,932 papers from IEEE, whereas in Fig. 6 we include the confusion matrices of the four detectors using as input 776 papers from Elsevier.

In Table 7 we include the top-10 of countries in volume of publications split into editorials, while in Table 8 we include the top-5 of countries with higher relation of publications marked by the detector vs total number of publications split into editorials and detectors.

**Fig. 5** Confusion matrix of the four ChatGPT-content detectors evaluated with 1,932 papers published in Dec 2010 in 86 journals from IEEE



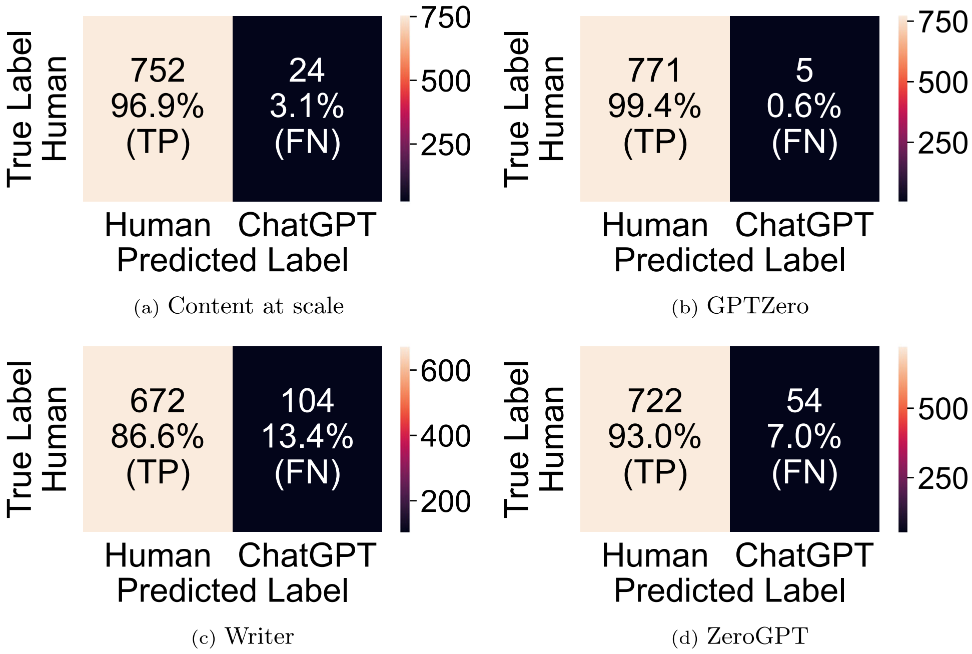
(a) Content at scale

(b) GPTZero

(c) Writer

(d) ZeroGPT

**Fig. 6** Confusion matrix of the four ChatGPT-content detectors evaluated with 776 papers published in Dec 2010 in 48 journals from Elsevier



(a) Content at scale

(b) GPTZero

(c) Writer

(d) ZeroGPT

**Table 7** Top-10 of countries in volume of publications split into editorials

| Rank | Elsevier | | IEEE | | MDPI | | Science | | Springer | | The Lancet | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Country | Publications | Country | Publications | Country | Publications | Country | Publications | Country | Publications | Country | Publications |
| 1 | China | 2,705 | China | 8,484 | China | 6,138 | US | 104 | China | 1,502 | UK | 274 |
| 2 | US | 771 | UK | 1,449 | US | 1,954 | UK | 59 | India | 416 | US | 208 |
| 3 | India | 659 | US | 1,211 | Italy | 1,571 | China | 35 | South Korea | 398 | China | 101 |
| 4 | UK | 602 | South Korea | 1,063 | Poland | 1,279 | Germany | 31 | US | 383 | Australia | 71 |
| 5 | Spain | 335 | India | 938 | UK | 1,210 | Canada | 18 | UK | 345 | Canada | 66 |
| 6 | Italy | 273 | Canada | 905 | Saudi Arabia | 1,117 | Switzerland | 16 | Japan | 250 | France | 60 |
| 7 | Canada | 259 | Japan | 635 | Spain | 1,094 | France | 15 | Italy | 202 | Germany | 57 |
| 8 | Germany | 250 | Germany | 547 | Russia | 1,083 | Australia | 13 | Saudi Arabia | 181 | Netherlands | 52 |
| 9 | France | 249 | Australia | 546 | South Korea | 1,038 | Denmark | 8 | Germany | 170 | Switzerland | 52 |
| 10 | Australia | 244 | Italy | 530 | Germany | 948 | Brazil | 8 | Canada | 161 | Italy | 48 |

These numbers are based on the affiliations and contain both, papers marked and not marked by detectors

**Table 8** Top-5 of countries with higher relation (publications marked by the detector vs total number of publications) split into editorials and detectors

| | Rank | Elsevier | | | IEEE | | | MDPI | | | Science | | | Springer | | | The Lancet | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Country | Total | Relation | Country | Total | Relation | Country | Total | Relation | Country | Total | Relation | Country | Total | Relation | Country | Total | Relation |
| Content at scale | 1 | Georgia | 2 | 50,0 | Saint Lucia | 1 | 100 | Saint Kitts and Nevis | 2 | 100 | Thailand | 1 | 100 | Ghana | 3 | 33,3 | Palestine | 1 | 100 |
| | 2 | Nigeria | 5 | 40,0 | Burundi | 1 | 100 | Bhutan | 1 | 100 | Saudi Arabia | 2 | 50,0 | Oman | 8 | 25,0 | Solomon Islands | 1 | 100 |
| | 3 | Ghana | 3 | 33,3 | Jersey | 1 | 100 | South Sudan | 1 | 100 | Austria | 2 | 50,0 | Georgia | 4 | 25,0 | Cook Islands | 1 | 100 |
| | 4 | Kuwait | 7 | 28,6 | Libya | 3 | 33,3 | Djibouti | 1 | 100 | Brazil | 8 | 37,5 | Sri Lanka | 5 | 20,0 | Nepal | 1 | 100 |
| | 5 | Cyprus | 8 | 25,0 | Venezuela | 6 | 16,7 | Monaco | 1 | 100 | Norway | 4 | 25,0 | Palestine | 6 | 16,7 | Kiribati | 1 | 100 |
| GPTZero | 1 | Cyprus | 8 | 25,0 | Libya | 3 | 33,3 | Bhutan | 1 | 100 | Iran | 1 | 100 | Sri Lanka | 5 | 20,0 | Ukraine | 1 | 100 |
| | 2 | Serbia | 16 | 6,3 | Kazakhstan | 12 | 8,3 | French Guiana | 3 | 66,6 | Hong Kong | 3 | 33,3 | Oman | 8 | 12,5 | Rwanda | 2 | 50,0 |
| | 3 | Qatar | 35 | 5,7 | Yemen | 18 | 5,6 | Gabon | 5 | 40,0 | Brazil | 8 | 12,5 | Serbia | 9 | 11,1 | Bulgaria | 2 | 50,0 |
| | 4 | United Arab Emirates | 38 | 5,3 | Seychelles | 20 | 5,0 | Moldova, of Republic | 7 | 28,6 | US | 104 | 5,8 | Nepal | 10 | 10,0 | Qatar | 3 | 33,3 |
| | 5 | Bangladesh | 21 | 4,8 | Romania | 26 | 3,8 | Guatemala | 4 | 25,0 | Canada | 18 | 5,6 | Slovakia | 12 | 8,3 | Ghana | 3 | 33,3 |
| Writer | 1 | Ukraine | 1 | 100 | Guinea | 1 | 100 | Turks and Caicos Islands | 1 | 100 | Iran | 1 | 100 | Gambia | 1 | 100 | Palestine | 1 | 100 |
| | 2 | Paraguay | 1 | 100 | Bosnia and Herzegovina | 3 | 66,7 | Faroe Islands | 1 | 100 | Thailand | 1 | 100 | Benin | 1 | 100 | Solomon Islands | 1 | 100 |

**Table 8** continued

| | Rank | Elsevier | | | IEEE | | | MDPI | | | Science | | | Springer | | | The Lancet | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Country | Total | Relation | Country | Total | Relation | Country | Total | Relation | Country | Total | Relation | Country | Total | Relation | Country | Total | Relation |
| | 3 | French Polynesia | 1 | 100 | Antigua and Barbuda | 2 | 50,0 | Equatorial Guinea | 1 | 100 | Saudi Arabia | 2 | 50,0 | Philippines | 3 | 33,3 | Iran | 4 | 50,0 |
| | 4 | Bahrain | 5 | 40,0 | Latvia | 2 | 50,0 | Saint Lucia | 1 | 100 | Sweden | 6 | 16,7 | Ethiopia | 15 | 20,0 | Philippines | 7 | 28,6 |
| | 5 | Macao | 23 | 34,8 | Syria | 6 | 33,3 | French Guiana | 3 | 66,6 | Brazil | 8 | 12,5 | Kazakhstan | 5 | 20,0 | Chile | 7 | 28,6 |
| ZeroGPT | 1 | Fiji | 1 | 100 | Saint Lucia | 1 | 100 | French Guiana | 3 | 100 | Iran | 1 | 100 | Georgia | 4 | 50,0 | Philippines | 7 | 100 |
| | 2 | Armenia | 1 | 100 | Cuba | 2 | 50,0 | Malawi | 2 | 100 | Thailand | 1 | 100 | Oman | 8 | 25,0 | Indonesia | 6 | 100 |
| | 3 | Georgia | 2 | 50,0 | Venezuela | 6 | 33,3 | Central African Republic | 2 | 100 | Brazil | 8 | 50,0 | Nepal | 10 | 20,0 | Tanzania | 5 | 100 |
| | 4 | Nigeria | 5 | 40,0 | Libya | 3 | 33,3 | Trinidad and Tobago | 2 | 100 | Saudi Arabia | 2 | 50,0 | New Zealand | 10 | 20,0 | Qatar | 3 | 100 |
| | 5 | Ghana | 3 | 33,3 | Montenegro | 6 | 16,7 | Bhutan | 1 | 100 | Togo | 2 | 50,0 | Argentina | 10 | 20,0 | Peru | 3 | 100 |

**Funding** Open access funding provided by Halmstad University.

**Data availability and access** All data used in the paper are public and the experiments are explained with enough detail to be reproduced by future researchers.

## Declarations

**Competing Interests** The authors declare no interests.

**Ethical and informed consent for data used** All the data used in this paper are available and thus, we do not handle private information.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

1. Bender EM, Koller A (2020) Climbing towards NLU: On meaning, form, and understanding in the age of data. In: Annual meeting of the association for computational linguistics, pp 5185–5198
2. 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 (2020) Language models are few-shot learners. Advances in neural information processing systems 33:1877–1901
3. Lepikhin D, Lee H, Xu Y, Chen D, Firat O, Huang Y, Krikun M, Shazeer N, Chen Z (2021) Gshard: Scaling giant models with conditional computation and automatic sharding. In: International conference on learning representations
4. Fedus W, Zoph B, Shazeer N (2021) Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. J Machine Learn Res 23:1–40
5. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G (2023) LLaMA: Open and efficient foundation language models 1–27
6. Google: An important next step on our AI journey. https://blog.google/technology/ai/bard-google-ai-search-updates/
7. Reuters: ChatGPT launches boom in AI-written e-books on Amazon. https://www.reuters.com/technology/chatgpt-launches-boom-ai-written-e-books-amazon-2023-02-21/
8. Thorp HH (2023) ChatGPT is fun, but not an author. Science 379(6630):313–313
9. Else H (2023) Abstracts written by ChatGPT fool scientists. Nature 613(7944):423–423
10. Van Noorden R (2022) How language-generation AIs could transform science. Nature 605(7908):21
11. De Saussure F (1989) Cours de Linguistique Générale vol. 1. Otto Harrassowitz Verlag
12. Chomsky N (1957) Syntactic Structures. Mouton de Gruyter
13. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423
14. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9(8):1735–1780
15. Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncertainty, Fuzziness Knowl-Based Syst 6(02):107–116
16. Church KW, Gale WA (1995) Poisson mixtures. Natural Language Eng 1(2):163–190
17. Katz SM (1996) Distribution of content words and phrases in text and language modelling. Natural Language Eng 2(1):15–59
18. Fisher RA (1956) Statistical methods and scientific inference
19. Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. Biometrika 175–240
20. Szucs D, Ioannidis JP (2017) When null hypothesis significance testing is unsuitable for research: a reassessment. Front Human Neurosci 11:390
21. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB (2015) The fickle p value generates irreproducible results. Nature Methods 12(3):179–185
22. Lehmann EL (1993) The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? J American Statistical Association 88(424):1242–1249
23. Content at Scale: AI Detector. https://contentatscale.ai/ai-content-detector/
24. GPTZero. https://gptzero.me
25. Write AI: Content detector. https://writer.com/ai-content-detector/
26. ZeroGPT: The most Advanced and Reliable ChatGPT detector tool. https://www.zerogpt.com
27. Guo B, Zhang X, Wang Z, Jiang M, Nie J, Ding Y, Yue J, Wu Y (2023) How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. In: Symposium on large language models, colocated with the international joint conference on artificial intelligence
28. Wilson EB (1927) Probable inference, the law of succession, and statistical inference. J American Statistical Association 22(158):209–212
29. López-Cózar D, Martín-Martín A (2022) Detectando patrones anómalos de publicación científica en España: Más sobre el impacto del sistema de evaluación científica
30. The Lancet: Information for Authors. https://www.thelancet.com/pb/assets/raw/Lancet/authors/tl-info-for-authors-1676565160037.pdf
31. Ye J, Han S, Hu Y, Coskun B, Liu M, Qin H, Skiena S (2017) Nationality classification using name embeddings. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 1897–1906
32. Elsevier: To Err is Not Human: The Dangers of AI-assisted Academic Writing. https://scientific-publishing.webshop.elsevier.com/research-process/the-dangers-of-ai-assisted-academic-writing/
33. O'Connor S (2023) Corrigendum to "Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?" Nurse Educ Practice 67:103572
34. O'Connor S (2023) ChatGPT: Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? Nurse Educ Practice 66:103537
35. Elsevier: Publishing Ethics – The Use of AI and AI-assisted Technologies in Scientific Writing. https://www.elsevier.com/about/policies/publishing-ethics#Authors
36. Elsevier: The use of AI and AI-assisted writing technologies in scientific writing. https://www.elsevier.com/about/

policies/publishing-ethics/the-use-of-ai-and-ai-assisted-writing-technologies-in-scientific-writing
37. Science: Editorial Policies. https://www.science.org/content/page/science-journals-editorial-policies
38. Springer: Authorship Principles. https://www.springer.com/us/editorial-policies/authorship-principles
39. Elsevier: The use of generative AI and AI-assisted technologies in writing for Elsevier. https://www.elsevier.com/about/policies-and-standards/the-use-of-generative-ai-and-ai-assisted-technologies-in-writing-for-elsevier
40. IEEE: Submission and Peer Review Policies. https://www.elsevier.com/about/policies-and-standards/publishing-ethics#
41. MDPI: Research and Publication Ethics. https://www.mdpi.com/ethics
42. Science: Change to policy on the use of generative AI and large language models. https://www.science.org/content/blog-post/change-policy-use-generative-ai-and-large-language-models
43. Springer Nature Group: Springer Nature introduces Curie, its AI-powered scientific writing assistant. https://group.springernature.com/la/group/media/press-releases/ai-powered-scientific-writing-assitant-launched/26176230
44. Lancet T The use of AI and AI-assisted technologies in scientific writing. https://www.thelancet.com/publishing-excellence
45. COPE: COPE: Committee on Publication Ethics | Promoting integrity in scholarly research and its publication. https://publicationethics.org
46. COPE: Authorship and AI tools. https://publicationethics.org/cope-position-statements/ai-author
47. Chuan PM, Son LH, Ali M, Khang TD, Huong LT, Dey N (2018) Link prediction in co-authorship networks based on hybrid content similarity metric. Appl Intell 48:2470–2486

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

**Pablo Picazo-Sanchez** is an Assistant Professor at Halmstad University, Sweden. He has a MSc. in Computer Science (2013) and a Ph.D. in Computer Science (2016) from University Carlos III of Madrid. His current research interests include systems security, applied cryptography, web security, and applied Machine Learning. For additional information see: http://pica4x6.github.io/

**Lara Ortiz-Martin** is an independent researcher in systems security. She holds a PhD in Computer Science University Carlos III of Madrid. Her current research interests include applied cryptography, biometrics, and web security. She works for a private company.