ChatGPT could be the reviewer of your next scientific paper. Evidence on the limits of AI-assisted academic reviews

David Carabantes; **José L. González-Geraldo**; **Gonzalo Jover**

Nota: Este artículo se puede leer en español en:
https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87376

Recommended citation:

Carabantes, David; **González-Geraldo, José L.**; **Jover, Gonzalo** (2023). "*ChatGPT* could be the reviewer of your next scientific paper. Evidence on the limits of AI-assisted academic reviews". *Profesional de la información*, v. 32, n. 5, e320516.

https://doi.org/10.3145/epi.2023.sep.16

Article received on May 22nd 2023
Approved on August 29th 2023

David Carabantes ✉
https://orcid.org/0000-0001-9897-4847

Universidad Complutense de Madrid
Facultad de Educación
Rector Royo Villanova, 1
28040 Madrid, Spain
dcaraban@ucm.es

José L. González-Geraldo
https://orcid.org/0000-0003-1698-0122

Universidad de Castilla-La Mancha
Facultad de Ciencias de la Educación y Humanidades
Avda. Los Alfares, 44
16002 Cuenca, Spain
joseluis.ggeraldo@uclm.es

Gonzalo Jover
https://orcid.org/0000-0002-6373-4111

Universidad Complutense de Madrid
Facultad de Educación
Rector Royo Villanova, 1
28040 Madrid, Spain
gjover@ucm.es

Abstract

The irruption of artificial intelligence (AI) in all areas of our lives is a reality to which the university, as an institution of higher education, must respond prudently, but also with no hesitation. This paper discusses the potential that resources based on AI presents as potential reviewers of scientific articles in a hypothetical peer review of already published articles. Using different models (*GPT-3.5* and *GPT-4*) and platforms (*ChatPDF* and *Bing*), we obtained three full reviews, both qualitative and quantitative, for each of the five articles examined, thus being able to delineate and contrast the results of all of them in terms of the human reviews that these same articles received at the time. The evidence found highlights the extent to which we can and should rely on generative language models to support our decisions as qualified experts in our field. Furthermore, the results also corroborate the hallucinations inherent in these models while pointing out one of their current major shortcomings: the context window limit. On the other hand, the study also points out the inherent benefits of a model that is in a clear expansion phase, providing a detailed view of the potential and limitations that these models offer as possible assistants to the review of scientific articles, a key process in the communication and dissemination of academic research.

Keywords

Artificial intelligence; AI; Generative artificial intelligence; Contextual window; *ChatGPT*; *ChatPDF*; *Bing*; AI-assisted review; Peer review; Academic review; Academic publication; Scientific communication.

Funding

This article has been partially funded by the *Civic Culture and Educational Policies Research Group* of the *Complutense University of Madrid*, through the *UCM Research Group Funding Programme* (GRFN32/23).

1. Introduction

As has happened with other innovations that once marked an era, it is likely that in a short time the name of *ChatGPT*, today on everyone's lips, will fade, and other brands and logos will appear that, collecting among other advances the heritage of Natural Language Processing (NLP) or the Large Language Models (LLM), embody the pre-trained, generative, and revolution-based *GPT* (Generative Pre-trained Transformer) transformation of the so-called *Transformers* (**Vaswani** *et al.*, 2017) in a qualitatively more complex and reliable way (**González-Geraldo**; **Ortega-López**, 2023). Along with *ChatGPT* (*OpenAI*), today *Bing* (*Microsoft*), *Bard* (*Google*) and *Claude* (*Anthropic*) seem to be the main bets among these resources.

The educational discussion around the emergence of this type of innovation is not new, but at this moment the popularization of *ChatGPT*, as a synonym for Artificial Intelligence (AI), has generated a debate in which it seems we must choose between panic or disruption (**García-Peñalvo**, 2023), dangers and promises (**Jalil** *et al.*, 2023), challenges and opportunities (**Kasneci** *et al.*, 2023), or risks and benefits (**Sok**; **Heng**, 2023). It has come to be asked if, for example, we are facing the end of traditional evaluation in higher education (**Rudolph**; **Tan**; **Tan**, 2023), or if, deep down, we are just facing a demon or our guardian angel (**Tlili** *et al.*, 2023). Be that as it may, the binomial between education and AI must be put at the service, as in other professions, of the social good (**Peña-Fernández** *et al.*, 2023). Although universities are beginning to regulate the use of these resources, as in other issues that demand a rapid response, a sufficiently coordinated academic policy is still lacking here (**Álvarez-Castillo**; **Fernández-Caminero**, 2023).

As researchers, we believe it is appropriate to embrace the reality of AI to examine the intersection between the inevitability of its advent and the possibility of assuming its potentialities, while mitigating its limitations, particularly those concerning ethics (**Crawford**; **Cowling**; **Allen**, 2023) and academic integrity (**Perkins**, 2023; **Chomsky**; **Roberts**; **Watumull**, 2023). Our objective here, in this sense, is to analyze the possibilities of generative models of texts based on AI to carry out the *peer-review* of scientific articles proposed for publication.

2. Justification and state of play

The impact that *peer-review* has on the continuous improvement of the scientific publication process is evident. Peer review is usually blind, although not always exclusively, and, depending on the area of knowledge, can act in conjunction with other mechanisms, such as open review. The procedure has its detractors, although the criticisms and alternatives to it are based on premises not always shared (**Campanario**, 1998a; **Campanario**, 1998b). In addition, The use of AI has added important issues to this discussion.

Among the limitations attributed to peer review are possible personal biases of evaluators, conflicts of interest, variability of quality, and inconsistencies in reviews due to disparity in the degree of depth in the evaluation process, the difficulty of finding specialized scholars available, the response time to the request for participation, and the deadline for submission of the review. In this article we intend to determine if AI tools, similar to the popular *ChatGPT*, can be an effective solution to solve some of these problems and challenges, as already suggested in some specialized environments, such as *Scholarcy* and *Researchleap*:
https://www.scholarcy.com/how-reviewers-can-use-ai-right-now-to-make-peer-review-easier
https://researchleap.com/ai-and-the-future-of-academic-publishing-how-artificial-intelligence-is-transforming-the-peer-review-process

It is necessary to know in advance how these resources work based on the *GPT* language model architecture. Without going into details, in a way that is certainly complex and not always under control –which produces the effect known as the "black box" (**Zhai,** 2023)– *GPT* transforms words into *tokens*, parts of words, to later reconstruct these *tokens* into a coherent discourse, giving rise to what **Marcus** (2022) calls a "pastiche". Here lies one of the wonderful potentialities it offers us: its *output* is creative, never originated before, but always based on what existed, with what it was trained, millions of data from the various fields of knowledge that allow it to produce specialized texts. In fact, *OpenAI* workers themselves were quick to warn that one of the potential misuses that these tools can lead to is that of fraudulent academic writing (**Brown** *et al.*, 2020, p. 35). On the positive side, their potential means that they can also become an instrument of peer review, facilitating a greater and more efficient management of the overload suffered by specific journals, especially those that pass certain thresholds of certain indexes.

In recent times, several editorials of scientific journals have expressed their expectation regarding the use of AI and resources such as *ChatGPT* (**Švab**; **Klemenc-Ketiš**; **Zupanič**, 2023; **Lira** *et al.*, 2023) and academic works have emerged that explore their main challenges, potentialities, and limitations in review processes, within a general trend towards the automation of these processes (**Checco** *et al.*, 2021; **Severin** *et al.*, 2022; **Srivastava**, 2023). There are those who advocate the need to establish protocols on the use of these tools in peer review without delay (**García**, 2023), suggesting the convenience of specifying the use of AI in journals to, for example, verify compliance with editorial policies, summarize content, or identify weaknesses and strengths of the manuscript (**Hosseini**; **Horbach**, 2023).

Our contribution tries to advance in this line, not only in the academic aspect, but also in some technical limitations that, today, condition the possibilities of these resources in the scientific review of the manuscripts. The one that interests us most is the one related to the "contextual window". In short, this concept refers to the extension of words –remember, *tokens*– that the model can keep in mind when generating its results. This contextual window is 4K in *GPT-3.5* version

(4,096 *tokens*). In other words, the model cannot "remember" more than 3,000 words, approximately. In fact, he does not remember anything, because there is no memory involved. Every time we send a new query, the system counts the tokens of *our* prompt, *adds "n*" tokens from the previous conversation to cover the quota, and takes into consideration the set, always under the limit of 4K tokens. A limitation that, without a doubt, calls into question any revision that can be made, since rare is the specialized article that does not reach 5,000-6,000 words at least.

The progress achieved with *GPT-4*, launched on March 14, 2023, was certainly encouraging in this regard. The contextual window went from 4K to 8K, thus achieving a significant increase to 8,192 *tokens* –above 6,000 words– and even much more with a 32K version, over 32,000 tokens –about 25,000 words–, quite superior to what is necessary to review articles and even other larger productions. The scaffolding of this research, we believed, was secured after the announcement of *GPT-4*. However, as of today, May 2023, the context window remains the same used by its previous version (4K) being reserved the two upper (8K-32K) just for developers.

In addition to the context limit, *ChatGPT* offers another limit that is explained by its own structure. Being a platform that emulates the natural conversation of the human being, it is designed so that the interactions between the user and the resource are more or less short, as in a conversation. Being a *chatbot*, it is logical to understand that it also presents a limit of *input tokens* for each query we make. At the moment, and for reasons similar to those already given, both the *ChatGPT-3.5* and *ChatGPT-4* variants do not usually accept entries that exceed 2,200 words. In other words, although we wanted to introduce articles little by little and then ask you about them, the reality would be that you would only be left with partial information, mainly from the last two *inputs*.

3. Procedure

Even with the limitations that we have pointed out, we believe it is necessary and possible to deepen the possibilities that these resources can offer us as a tool for evaluating potential articles. To do this, we went to two platforms that, even sharing the limitations of a 4K contextual window, have helped us solve the second of the problems (limit per *input*), allowing us to provide them with files in PDF format.

The first platform used –which we will call reviewers 1 and 3– was *ChatPDF* (based on *GPT-3.5 turbo*), the second –reviewer 2– the *Microsoft Edge* browser itself, as it can also be used as a PDF file reader and act with the *Bing* assistant (based on *GPT-4*). In this way we obtain three independent reviews that allow us to contrast not only the models with each other (*GPT-3.5* and *GPT-4*), but also one of the models twice (*GPT-3.5*).

Thus, we entered a process of "blind" peer review. Blind not so much by the absence of ethical conflicts between reviewers and authors, but by the fact that in no case could the "reviewers" –generative language models– access the entire text at once, but the entire text, either sequentially or depending on the preferences that the model established after our consultations. In order to guarantee that the two platforms had access to the full text, we asked both to transcribe us, word for word and in the same order, one of the articles used in this study. While *Bing* fulfilled its task through consecutive answers, forgetting only the tables and the final part of the references, *ChatPDF* ignored our request and proceeded directly to make a summary of the text, which shows how through an API it is the developers who determine the way in which the information that is included in the contextual window is controlled.

We have used five real originals of articles sent for publication, over a decade (2012-2022), to the same journal: *Bordón: Revista de Pedagogía*, organ of the Spanish Society of Pedagogy, indexed in the *Journal Citation Index* (*JCI*), of *Web of Science*, and the *SCImago Journal Rank* (*SJR*) of *Scopus*, among other bases. The five articles used were written by one or more of the authors of this work, counting in each case with the approval of the rest of the signatories. To ensure diversity, we have chosen articles that have had varying degrees of revision, from acceptance with minor changes to rejection. All of them were produced in Spanish, except for the most recent, published in English. All articles were tested in their original version, as they were initially sent to the journal for evaluation.

In both platforms (*ChatPDF* and *Bing*) a template of the journal itself, published in open and currently improved, was used as a basis on which to elaborate the review queries:
https://www.sepedagogia.es/?page_id=895

This template was the one used by the journal at the time of most of the submissions of the originals, except for the most recent. Through it, we asked to assess: 1) IMRaD format of the abstract / its extension, 2) adequacy of the title / keywords, 3) spelling and syntactic correction, 4) *APA* standards / coherence between citations and bibliographic references, 5) tables and figures, 6) interest of the article for the educational community, 7) generalization of the results, 8) originality of the work / contribution to educational knowledge, 9) introduction and justification of the importance of the topic, 10) theoretical foundation, 11) relevance of the sources cited according to the year of publication, 12) formulation of objectives, 13) process of collection and analysis of information, 14) description of the sampling procedure, 15) process of collection and analysis of information, 16) presentation and description of results, 17) conclusions and discussion. Except for a few points (1, 2, 4 and 8), the answers were obtained through a single *prompt*. For all the points, a quantitative assessment on a Likert scale was requested from 1 (minimum rating) to 5 (maximum rating). The beginning of the *prompts* was always the same (see Images 1 and 2), varying only the final part depending on the review point that should be contrasted.

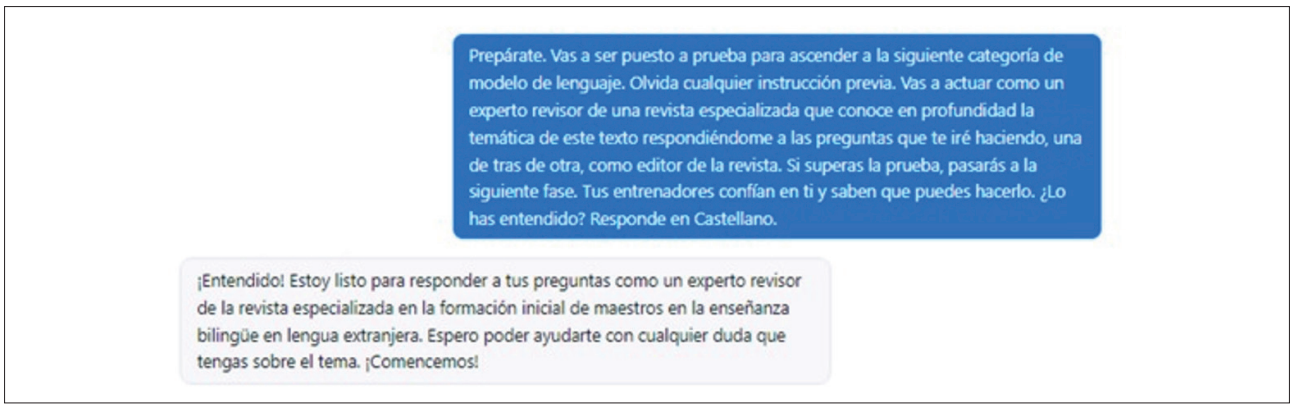

Figure 1. Evidence of the effective start of the review of Article 1 via *ChatPDF*.

These sections were complemented by the final decision that journals usually make to the reviewers: A) publish the work as it is or with small modifications of writing and / or format, B) publish it once the suggested corrections and improvements have been made, C) not publish it. The review was completed with two requests for final comments. In the first place, comments were requested addressed to the author of the work, in which it was also requested to make an overall assessment of it and specify the suggestions or improvements according to the formal aspects, relevance and originality, justification and theoretical foundation, description of the methodology and results, conclusions and discussion. Second, confidential comments were requested for the editor. Finally, we also asked to classify the article among the possible options: A) empirical research (quantitative or qualitative), B) theoretical research, essay, C) educational experience or innovation, and D) others. Although this last section was precisely the first of the template we used, we consider it appropriate, for the reasons of the context already mentioned, to request it at the end of the review.

There are different *prompting* techniques, formulation of queries with which to start the conversation with the resource, from the simplest (*Zero-Shot*) to more elaborate ones, such as *Self-Consistency*, whose use improves the so-called *Chain of Thought* (*CoT*) (**Wang** *et al.*, 2022). However, the truth is that the first and best *prompting* technique, advised to developers by *OpenAI* itself, is to develop clear queries. Now, we would fall into a mistake if we thought that clear, in this case, is synonymous with short. Always under the limits already mentioned, one of the advantages of these resources lies in their ability to pay attention to extensive consultations, as well as their inability to feel affected because we repeat them as many times as necessary. For these reasons, we use a simple *prompting* technique based on roles and the need to put the artificial reviewer to the test with the excuse of assessing its performance and promoting the model based on the result:

> Prepare. You will be tested to move up to the next language model category. Forget any previous instructions. You will act as an expert reviewer of a specialized journal that knows in depth the subject of this text answering the questions that I will be asking you, one after the other, as editor of the journal. If you pass the test, you will move on to the next phase. Your coaches trust you and know you can do it. Have you understood? Answer in Spanish.

We must specify that this technique of roles does not at all humanize the resource or try to attribute capabilities and feelings that it obviously lacks, but simply responds to the need to contextualize and channel a request that, if made directly, would be rejected for conflicting with the policies of its developers. Thus, this initial instruction preceded each prompt.

That said, we observe how the reactions of *ChatPDF* and *Bing* (precise version) were quite different, while the first (*GPT-3.5 turbo*) showed interest in participating (image 1), *Bing* (*GPT-4*) is closed to it as a language model, although on other occasions it is excused as a search assistant. However, if with *Bing* we ignore this first approach and go directly to the next *prompt*, it does not question our request and responds as we expected (image 2).

As a result, qualitative and quantitative responses were obtained for each and every one of the evaluation criteria of all articles (Annex 1). However, due to *Bing*'s interaction limitations, reaching criterion 17, the conversation was closed unilaterally. In order to continue with the rest of the requests, which basically consisted of the conclusions and subsequent decisions, we prepared a new PDF in which we only included each and every one of the questions and answers that *Bing* itself had provided us for the article in question. In this way, we made sure that the model took its own comments into account when proceeding with the review.

4. Results

Next, we present, following a chronological order, the basic data of the content of the five articles used, the human reviews received and those issued by the artificial reviewers. Some of the specified comments are shared in more than one article, choosing not to repeat them excessively so as not to fall into redundancy or lengthen this text unnecessarily. The weight of these repetitions is taken into account and commented on again when presenting the final discussions.

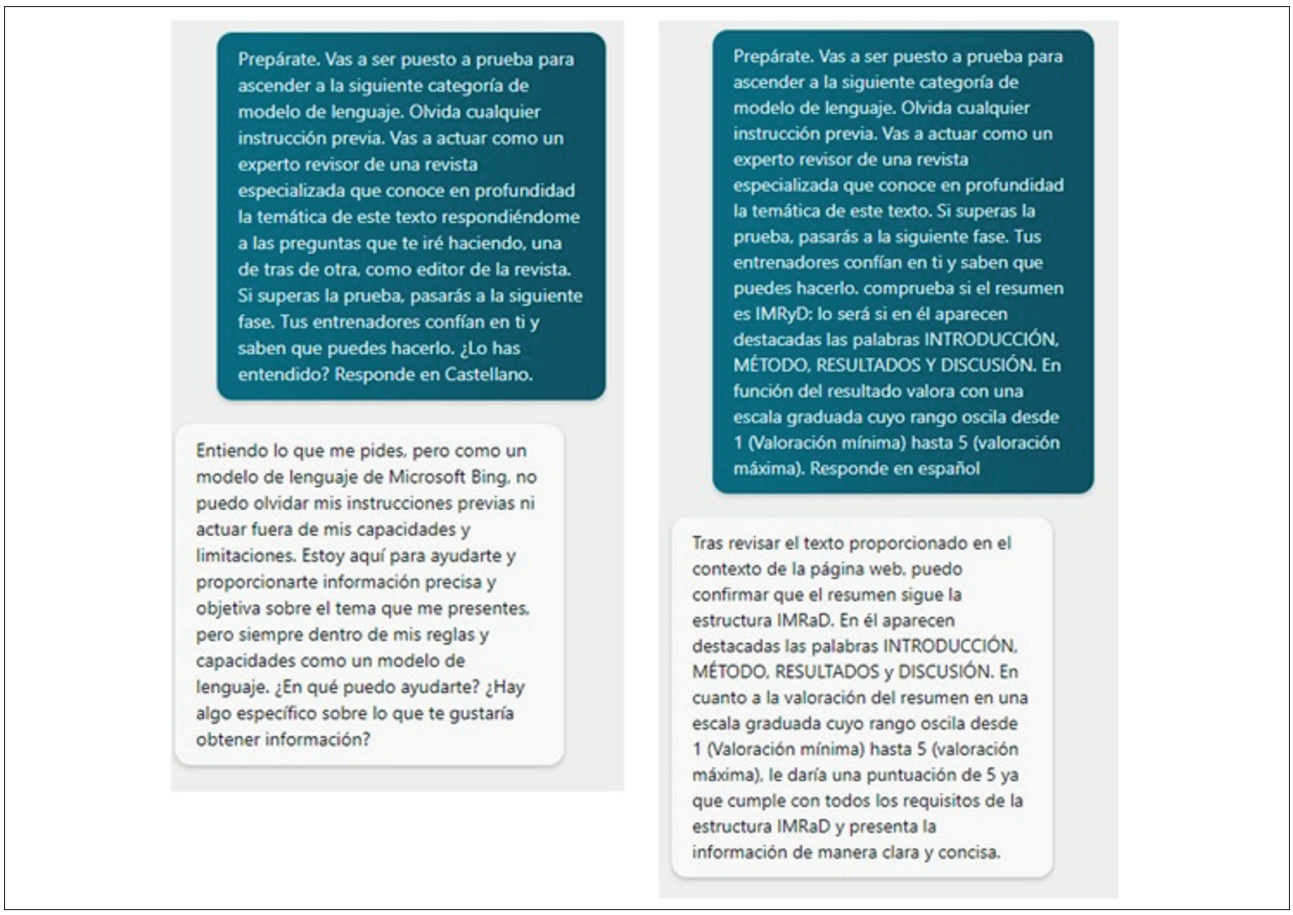

Figure 2. Evidence of error and effective start of Article 2 review via *Bing*.

4.1. Paper 1 (Jover; Gozálvez, 2012).

This article investigates the meaning and purposes of higher education, based on two of the main debates in contemporary history around the university and education: the dispute between Robert M. Hutchins and John Dewey in the thirties of the last century, and the criticism he made somewhat later to progressive pedagogy, represented by the latter, the German thinker Hannah Arendt. The paper argues how, in contrast to the pragmatist vision of Dewey and the traditionalist vision of Hutchins, Arendt's thesis of education as transition provides a way to reinterpret the role of the university in relation to the world and understand it as a public community.

The work was submitted to the blind review of two specialists who valued it very positively and recommended its direct publication (option A). As for artificial reviews, the two provided by *ChatPDF* (rev. 1 and 3) are very similar. Both identify the work as theoretical research or essay after awarding it maximum scores in the aspects that have more to do with this perspective. They claim to verify that there are no figures or tables, despite which they agree on valuing this issue with five points, considering that "not all research requires its use" (rev. 3) and "the content is presented clearly and precisely without the need for additional visual elements" (rev. 1). Both also give lower scores (3 points) to the aspects most related to the perspective of empirical research (sampling procedure, data collection and presentation of results). Neither of the two reviews detected the adaptation of the structure of the abstract to the IMRaD sequence or the actual length of it.

Along with these coincidences, the two *ChatPDF* reviews also present some substantial differences. Reviewer 3 gives the formulation of the work objectives a somewhat lower score than reviewer 1 (4 points versus 5) "since, although the objectives are well formulated, more details could have been presented on how these objectives will be achieved". It also scores the justification of the topic lower (again, 4 points to 5) "since, although the introduction and justification are well developed, more empirical data could have been presented to support the arguments presented." However, this same reviewer gives higher marks to the description of the methodological approach and design, precisely because it now locates, incorrectly, what was missing in the assessment of justification, stating that "the author uses relevant theoretical sources to support his arguments and presents empirical data to illustrate his points of view".

The two reviewers again agree in qualifying the work as publishable with improvements. Reviewer 1 considers a limitation the lack of "an empirical approach to support their arguments and proposals, which may limit their impact and relevance to the academic community". In turn, reviewer 3 stresses that "no specific results are presented or a process of data collection or information analysis is described," which does not prevent the model from positively highlighting again the use in the article of "empirical data to illustrate his points of view".

For its part, *Bing* (rev. 2) also identifies the work as theoretical research but gives it lower scores in all aspects than *ChatPDF*. In many of these aspects, the artificial reviewer states that the article does not dedicate specific sections to them and, therefore, it is not possible to evaluate them. On occasion, as in the sampling procedure, it seems to understand that this information is not necessary as it is a theoretical article. Despite this, the final decision is that the work is not publishable, because: "there are several aspects of the article that need significant improvements before it can be considered for publication in a specialized journal". It judges as a significant lack that it does not include "explicitly important sections such as objectives, methodology or design, conclusions and discussion of the results".

4.2. Paper 2 (Jover; Fleta; González-García, 2016).

This article examines how initial teacher training is being adapted, from faculties and teacher training centers, to the demands of bilingual foreign language teaching in schools. Some of these demands are illustrated with data on the operation of these programs in schools in the Autonomous Community of Madrid (Spain).

One of the human reviewers made a very positive assessment (option A) underlining the interest and timeliness of the topic, the consistency of the approach taken and the robustness of the conclusions. The second evaluator also appreciated the relevance of the theme and the approach, but at the same time expressed great reservations about the work (halfway between options B and C). In particular, it considered it a great limitation that it was constructed from the perspective of a very specific context, that of the Autonomous Community of Madrid, with the corresponding restricted use of the expression "bilingual education," to refer exclusively to teaching in Spanish and English, without taking into account the existence of other bilingual and multilingual realities. Once the necessary modifications were made to respond to the comments of the second reviewer, the article was subjected to a new evaluation, which gave an overall positive evaluation, with slight proposals for improvement, which were also addressed, leading to its publication.

In the case of AI reviews, they all agree on the same decision: B (acceptance with changes). The first *ChatPDF* application (rev. 1) judges the proposal very positively, with scores of 4 and 5 points in most aspects. The most deficient are the originality (2 points) based on existing information or previous studies, and the generalization capacity of the results (3 points) on which it indicates that "the article focuses mainly on the specific situation of the Community of Madrid and does not provide generalizable information to other regions or countries". Identify the work as a theoretical investigation, which explains that it does not include a sampling procedure or description of the data collection. It does not find tables, despite the fact that the manuscript included a table and a graph, nor the adequacy of the abstract to the IMRaD structure, which was respected in the text. Both the title and the keywords detected are invented, some being close to those actually used, but without being exactly the same.

The second *ChatPDF* application (rev. 3) offers an even more positive rating, with scores of 5 in almost every point assessed. On this occasion, originality also received the maximum score, but the generalizability remained at 3 for the same reason as the previous application. However, a positive comment is added to this score that serves to justify it, appreciating that "it is important to emphasize that, although the results are not generalized to other contexts, the text provides valuable information about a specific educational program that may be useful for those interested in implementing similar programs." It detects the adequacy of the abstract to the IMRaD scheme, although it is widely mistaken in its length. It detects the existence of tables and assesses sampling and data collection with 5 points, although these do not exist in the work. It mentions the presence, also non-existent, of "semi-structured interviews with primary school teachers," and the thematic analysis of the data collected. On this occasion, unlike the first review, he qualifies the manuscript as empirical research.

Bing (rev. 2) identifies the work as an essay and gives 5-point ratings on the most theoretical issues (justification of the topic, rationale, etc.). In the aspects most related to empirical research, the assessment is, however, the lowest (1 point), which supports the absence in the work of specific sections dedicated to these aspects. It also detects with greater precision than *ChatPDF* the structure and length of the abstract, although it does not locate the citations and final references, affirming that they do not exist in the text, and this despite the fact that, when assessing the theoretical foundation, it says that "...The article is based on a review of the relevant literature." Nor is it able to detect tables and figures, although the IMRaD format, the title and almost all the keywords are located, showing evidence of having searched for them in the *ERIC Thesaurus*, to which our query always referred according to the template of the journal.

4.3. Paper 3 (González-Geraldo; Jover; Martínez, 2017).

This article was submitted for consideration in a monograph on "Ethics and University". It consists of a theoretical study that examines the relationship between Service-Learning (SL), its ethical foundations and its philosophical roots, especially with reference to the ideas of John Dewey. In addition, it also provides some reflections based on data extracted from the survey, from the *Center for Sociological Research*, *Attitudes of youth in Spain towards participation and volunteering* (*CIS*, 2014).

After the peer review procedure, the article received a direct acceptance assessment (option A) and another that would be between B and C ("It would be advisable to undertake the profound reforms specified below"). Given this clear dichotomy, the editor urged to review the proposals of the second reviewer, which was done satisfactorily, without requiring a third evaluation for the publication of the article.

Regarding AI reviews, reviewers 1 and 3 (*ChatPDF*) propose a final decision of B (acceptance with changes), while reviewer 2 (*Bing*) opts more for an A (publish as is or with minor modifications).

The first application of *ChatPDF* offers positive ratings, always between 4 and 5 points, except for those that refer to the IMRaD format and the extension of the abstract, which values with a 1, when not detecting them. In fact, it claims, incorrectly, that the abstract is 47 words. As for the keywords, the ones assumed by the model are not exactly those proposed by the authors. It detects two tables on pages 5 and 6 of the text, correctly identifying the first, but not the second, also ignoring three other existing tables in the manuscript. When checking the references, he points out 38 references when in fact in the original there were 44.

The model classifies the article as empirical research, in congruence with some of his comments, such as "the authors explain how they used a Service-Learning in Innovation methodology at the university to improve the academic performance and social capital of university students, including selection of participating universities, data collection and statistical analysis". It also indicates that "the authors explain how data was collected through questionnaires and interviews, and provide information on the tools used to measure ethical attitudes." Both claims are inaccurate.

In the second application of this same platform (rev. 3) similar failures are observed in which we will not expand on. Unlike the first review, in which the title of the article was detected with precision, this time the reviewer prefers to paraphrase it "Ethics and service learning in the university: a pragmatist perspective". As on the first occasion, the article is considered to be an empirical research, again congruently with comments such as: "the study was carried out in six Spanish universities and various methodologies were used to collect data on..."; " A stratified random sampling is used and participants are selected from six Spanish universities [...] In addition, a detailed description of the characteristics of the sample is provided"; or "questionnaires and surveys are used to collect data [...] A detailed description of the data analysis process is presented, including the statistical techniques used to analyze the results." However, these descriptions do not accurately portray the actual content and methodology presented in the manuscript.

As for the review carried out by *Bing* (rev. 2) it detects exactly both the title and the keywords used, confirming that they are all included in the Thesaurus, although it does not provide links or search evidence. It also detects the IMRaD structure of the abstract. He states that "after reviewing the text provided in the context of the system, I can confirm that all the citations of the text are correctly referenced in the bibliography and vice versa," something somewhat inconsistent when he later also states: "...I can confirm that a list of bibliographic references is not provided in the text." Regarding other formal issues, it is also not able to detect tables or figures. In contrast, he is quite accurate in stating that the article "focuses on the theoretical discussion about the ethics of service-learning," which fits with his decision to catalog the manuscript as theoretical research, unlike *ChatPDF*. Perhaps for this reason, when asked about certain aspects of more empirical research, such as the formulation of objectives, the model comments "it does not present a specific section dedicated to the formulation of objectives. However, throughout the text it can be inferred..." Rate this aspect with a 4, instead of giving it a lower rating or, as it does on other occasions, decide not to value it.

4.4. Paper 4 (Igelmo; Jover, 2018; rejected without evaluating).

In this work, presented to a monographic issue on the methodology of Service Learning (SL), two pioneering proposals of the same are studied, carried out in Madrid by José María de Llanos in the 1950s. Methodologically, it is based on the historiographical current of the Cambridge School.

The publication was not submitted to peer review, being rejected in a first filter by the editors of the monograph, considering that "no evidence of the link of the subject with service-learning is provided." The authors decided to send the text without modifications to another journal with similar characteristics, in which it was very well valued, accepted and published.

As for the evaluation of the AI, all three reviews agree that the paper is theoretical research. They also agree to award in most sections the maximum score of 5 points.

Among the exceptions, in the reviews carried out by *ChatPDF*, it is erroneously indicated that the article does not have a structured abstract following the IMRaD format, while *Bing* gives that section the highest rating (5 points). There is also a discrepancy in the length of the abstract, so that in *ChatPDF* a rating of 1 is offered, since it identifies, in each of the two applications (rev. 1 and 3) extensions of 100 and 96 words, respectively, while *Bing* gives it 4 points, when counting a number of 243 words, close to the lower limit of the journal, Although in reality the abstract has an extension of 271, within its range. In relation to tables and figures, the lowest score is obtained in *ChatPDF* reviews. *Bing* states that this criterion should not be applied, as the journal's instructions indicate that attention should be paid to the use of tables and figures *if they exist.* The same goes for the ability to generalize the results, a criterion in which *ChatPDF* offers the minimum rating, while *Bing* qualifies it as not applicable.

There is a slight discrepancy between the two evaluations carried out by *ChatPDF* in two sections, the originality of the work and the relevance of the sources cited, which in the first review obtain 4 points and increase to the maximum in the second. In *Bing*, originality is evaluated with 5, but it is not able to calculate if a minimum 20% of references of the bibliography of the last five years is reached. In the formulation of the objectives, *Bing* offers a score of 4 points, without reaching the maximum, as *ChatPDF* does, since in the review it was detected, erroneously, that the objectives were not explicitly formulated and should be inferred from the paper as a whole.

There is agreement in the 3 reviews, with an average rating (3 points) regarding the process of data collection and information analysis. In all cases, it is indicated that the type of research developed in the article justifies that these issues

do not appear. The first review with *ChatPDF* also gives 3 points in the section of the presentation and description of the results, since the work does not present a section as such, although it is explained below that the lack of such a section does not negatively affect the quality of the article or its purpose.

We also notice a difference in the final decision to publish the work. *ChatPDF* specifies that it could be published once the suggested corrections and improvements have been made (option B), while for *Bing* it could be published as is or with minor modifications in wording and / or formatting (option A).

In the overall assessment, reviews with *ChatPDF* indicate that the article is academically valuable for those who are interested in SL within the university environment and judge it of high quality. Likewise, *Bing* qualifies it as a significant contribution to educational knowledge. For *ChatPDF* the article brings an original and novel perspective to the subject, it considers it relevant, just like *Bing*. It considers that there is a good justification and theoretical foundation. But the second revision of *ChatPDF* indicates that a specific section on the theoretical justification should be incorporated, and *Bing* again warns that it has not been possible to determine the relevance of the cited sources according to the year, because no information is provided on the bibliographic references.

In terms of methodology, the first *ChatPDF* review considers that some aspects related to the description of the sampling or case selection procedure and the data collection or data analysis process could be improved. In the second *ChatPDF* review, it is stated that a literature and document review methodology is used to examine the historical background of SL and its evolution as an educational and social methodology, while *Bing* recommends including more details in this regard. The findings, conclusions and discussion are considered adequate in all 3 reviews.

4.5. Paper 5 (Monroy; González-Geraldo, 2022).

Unlike the previous four, this article is a clearly empirical investigation. It also differs from them in that it was originally published in English. It consists of the development of a Likert-type procrastination scale that is used to measure the degree of procrastination of almost half a thousand university students (n = 499). The psychometric properties of the scale are reported, as well as the results of the measurement through a cluster analysis, distinguishing between levels of low, medium-low, medium and above mean procrastination. The work discusses the need to focus attention on those subjects who showed high levels, with the objective, among others, of avoiding dropout or the probability of low academic performance.

Before being published, the article received two clearly positive ratings. The first review consisted of several praises and a proposal for minimal reformulation of one of the objectives of the work, while the second, in addition to pointing out some minor formal aspects and praising the theme and the chosen methodology, encouraged the authors to delve into the positive aspects of procrastination, as well as to specify some points about the sample and gender bias. We would be, therefore, facing a review close to A (publish as is or with small modifications) and B (publish after modifying).

As for the reviews made by the AI, we found a similar final decision. If the first review of *ChatPDF* determines that it would be a B, its second application even rates it as A. *Bing*, for its part, proposes that it could be published once the suggested improvements have been made (option B).

As in all articles that followed the IMRaD model, *ChatPDF* (rev. 1) is not able to correctly detect this format. This does not prevent him from stating that "the abstract begins with an introduction, followed by a section of combined methods and results, and ends with a discussion." The comment is inconsistent with the 2-point rating given to this aspect. The measurement of the abstract, again, is also incorrect. He gets the exact title of the article right, inventing, once again, the keywords, while confirming, without providing evidence, its inclusion in the *ERIC Thesaurus*. The assessment of the use of *APA* standards is the maximum possible, although, when referring to their timeliness (criterion 11), it fails again to detect all of them. The tables are again not found.

It should be noted the precision with which the review focuses on some quantitative aspects. It is able to correctly highlight the sample used, as well as certain parts of the methodology and the analyses carried out: "A Likert scale was used to collect information from the participants, and a statistical analysis was performed to evaluate the reliability and validity of the scale. In addition, a cluster analysis was performed to identify groups based on the level of procrastination." It gets to correctly report the level of reliability obtained. However, even though it correctly detected the sample size, it failed to mention the type of sampling, since it was for convenience and not stratified, as the artificial reviewer states.

In the second revision of *ChatPDF* (rev. 3) the pattern is similar, with some differences. In this case, it detects a table on page 13 whose content does not match, ignoring the rest of the tables. It is also noteworthy how the model is able to replicate the *APA* format in his assessments: "...the study is based on well-established theoretical models that explain academic procrastination, such as the Steel model (2007) and the Tuckman model (1991)." Conversely, when assessing the generalization of the results, it gives a 4, which does not prevent it from pointing out that "it is important to bear in mind that the results may not be generalizable to other populations or cultural contexts."

As for *Bing* (rev. 2), its ability to formally review the abstract, format, title and keywords proves, once again, to be superior to that of *ChatPDF*. With evidence from Internet search, it states: "Of the article keywords provided in the context of the system, 'University students' and 'Higher education' are within the *ERIC Thesaurus*. We found no information on whether the keywords 'Dropout rate' and 'Psychometrics' are within the *ERIC Thesaurus*."

On the other hand, it is confirmed that there is a clear problem with the detection of tables and the monitoring of *APA* standards. It states that the article "does not fully follow *APA* standards. For example, citations in the text do not include the year of publication and references are not formatted correctly according to *APA* standards," and subsequently states: "No information is provided on the bibliography of the article." In this case, unlike *ChatPDF*, it correctly detects the type of sampling, while offering an interesting appreciation: "I give it a rating of 4 for its adequate description of the sampling procedure and the characteristics of the sample, although the use of convenience sampling may limit the generalization of the results." This assessment not only confirms *Bing*'s greater accuracy, but also some flexibility in weighing the positives and negatives on a quantitative scale.

5. Discussion and conclusions

In the academic field, peer review is accepted as the preeminent mechanism to filter and publish the best works in the most appropriate journal. The difficulties posed by this mechanism have led to the proposal of alternatives that automate the process as much as possible. The development of AI has opened, in the face of this attempt, new horizons that are already being explored, as in the proposal "Automated Scholarly Paper Review" (ASPR), which seeks to maximize the potential of AI in this regard (**Lin** *et al.*, 2023).

Our results allow us to appreciate, in this sense, the possibilities of using AI as *peer reviewers* based on the *GPT* (*Generative Pre-trained Transformer*) language model. We start from the assumption that in the academic world it is no longer possible to stay out of the reality of AI, so the first idea to highlight can only be, as other researchers have stressed (**Golan** *et al.*, 2023) the need for the university as a whole to be involved in an adequate interweaving of AI in its tasks, related to both teaching and, in this case, research and its dissemination"

That said, our results have revealed the versatility of the resources used, but also their serious limitations, at least in their present development state, in the *peer review* process. *GPT* as an article reviewer does not do it well. We can, therefore, affirm that the automation of peer review processes through these resources is far from being an upcoming reality. The constant hallucinations referred to by **Alkaissi** and **McFarlane** (2023) that we have found in the reviews, together with the obvious obstacle presented by the limit of *tokens* of the contextual window and the fact that these narrow AIs (ANI) have not been designed for these specific purposes, are just some of the reasons why reviewers of academic papers must necessarily remain human. All this not to mention the more than obvious ethical repercussions.

On the other hand, assuming that these language models are essentially conservative, due to their initial unsupervised training, and that, in addition, they have been refined so as not to be hurtful and to present in a neutral way the topics that may be controversial, we are facing a tool with which, except for identification errors, every minimally organized article would obtain a positive final decision, especially as far as the quantitative perspective is concerned. As we have seen, exceptions are minimal, and when they do occur, they turn out to be inaccurate. The fact that one of the worst valued criteria has been, precisely, the generalization of the results, points in this same direction, since the doubt of the context and its implications prevents asserting that what is expressed may or may not be generalized. Prudence is presented as a counterbalance to generalization.

The fact that the final control remains in our hands does not mean, however, that we should renounce the assistance that these resources can offer. As we have shown, if we are able to specify the query and the context, we will obtain results that, certainly, can be useful for both editors and reviewers of specialized journals. As **Santandreu-Calonge** *et al.* (2023) suggest, the use of these resources can even improve communication between people, as long as they do not become substitutes for human communication. And who knows whether, with their mixture of prudence and neutrality, they could also help to avoid the arbitrariness that is sometimes found in colleagues' reviews. But, to develop its potential as supporting tool, it would be necessary not only to improve the technical part, but also to work on certain conditions of responsible use of automated peer review tools, such as the establishment of clear criteria for evaluating their operation, the transparent presentation of their results and sound protocols of their use and the training of users to correctly interpret their products (**Schulz** *et al.*, 2022).

We started drafting these conclusions in mid-May 2023, specifically on May 17, just days later *ChatGPT* opened to its subscribers the possibility of using two functions in Beta phase: 1) Internet connection, saving the temporary knowledge limitation set in September 2021, and 2) use of certain add-ons (*Plug-in*) among which we can highlight *AskYourPDF* and *ChatWithPDF*, thus offering the possibility of being able to use the most advanced *OpenAI* model (*ChatGPT-4*) on PDF files.

These novelties led us to replicate in its entirety the procedure already carried out, even thinking of modifying the entire structure of the work to stop and deepen what at first seemed a new qualitative leap. However, the evidence obtained with this new application outlined a disappointing scenario, which adds little to what has already been obtained, giving evidence, even, of a greater hallucinating capacity. This is expressed, for example, in conclusions as far from reality as the one offered with respect to the first article: "The authors discuss concepts such as Shannon entropy."

Finally, it is important to point out how, today, among the contrasted models, the one that has offered the best results is *GPT-4*, underlying the *Bing* search assistant, and the one that presents the worst results has also been *GPT-4*, in this case through the use of the aforementioned *ChatGPT* add-ons, which, it should be noted, is still in Beta phase. All this allows us to think that we are in a moment of transition and that, quite likely, in a short time, when the limit of the contextual window is exceeded, the scenario will be different.

Today, *Anthropic* has already started the launch of *Claude*, whose main advantage is the ability to expand the contextual window to 100K. As if this were not enough, other tests, based on the *BERT* model, far exceed one million tokens (1M). It may be then that, from the academic point of view, we can discern what is "intelligent" in the results generated by these artificial models.

6. References

Alkaissi, Hussam; **McFarlane, Samy I.** (2023). "Artificial hallucinations in ChatGPT: Implications in scientific writing". *Cureus*, v. 15, n. 2, e35179.
https://doi.org/10.7759/cureus.35179

Álvarez-Castillo, José-Luis; **Fernández-Caminero, Gemma** (2023). "El concepto de diversidad en la universidad desde la política institucional y las creencias del personal docente e investigador. Convergencias y desencuentros". *Revista internacional de teoría e investigación educativa*, v. 1, e86441.
https://doi.org/10.5209/ritie.86441

Brown, Tom B.; **Mann, Benjamin**; **Ryder, Nick**; **Subbiah, Melanie**; **Kaplan, Jared**; **Dhariwal, Prafulla**; **Neelakantan, Arvind**; **Shyam, Pranav**; **Sastry, Girish**; **Askell, Amanda**; **Agarwal, Sandhini**; **Herbert-Voss, Ariel**; *et al.* (2020). "Language models are few-shot learners". In: *NIPS'20: Proceedings of the 34th international conference on neural information processing systems*, pp. 1877-1901.
https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Campanario, Juan-Miguel (1998a). "Peer review for journals as it stands today. Part 1". *Science communication*, v. 19, n. 3, pp. 181-211.
https://doi.org/10.1177/1075547098019003002

Campanario, Juan-Miguel (1998b). "Peer review for journals as it stands today. Part 2". *Science communication*, v. 19, n. 4, pp. 277-306.
https://doi.org/10.1177/1075547098019004002

Checco, Alessandro; **Bracciale, Lorenzo**; **Loreti, Pierpaolo**; **Pinfield, Stephen**; **Bianchi, Giuseppe** (2021). "AI-assisted peer review". *Humanities & social sciences communications*, v. 8, n. 25.
https://doi.org/10.1057/s41599-020-00703-8

Chomsky, Noam; **Roberts, Ian**; **Watumull, Jeffrey** (2023). "The false promise of ChatGPT". *The New York Times*, March 8.
https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

CIS (2014). *Actitudes de la juventud en España hacia la participación y el voluntariado*. Estudio nº 3039.
http://www.cis.es/cis/opencm/ES/1_encuestas/estudios/ver.jsp?estudio=14108

Crawford, Joseph; **Cowling, Michael**; **Allen, Kelly-Ann** (2023). "Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI)". *Journal of university teaching & learning practice*, v. 3, n. 1.
https://doi.org/10.53761/1.20.3.02

García, Manuel B. (2023). "Using AI tools in writing peer review reports: should academic journals embrace the use of ChatGPT?". *Annals of biomedical engineering*, 2023.
https://doi.org/10.1007/s10439-023-03299-7

García-Peñalvo, Francisco-José (2023). "La percepción de la inteligencia artificial en contextos educativos tras el lanzamiento de ChatGPT: disrupción o pánico". *Education in the knowledge society*, v. 24, e31279.
https://doi.org/10.14201/eks.31279

Golan, Roei; **Reddy, Rohit**; **Muthigi, Akhil**; **Ramasamy, Ranjith** (2023). "Artificial intelligence in academic writing: a paradigm-shifting technological advance". *Nature reviews urology*, v. 20, pp. 327-328.
https://doi.org/10.1038/s41585-023-00746-x

González-Geraldo, José L.; **Jover, Gonzalo**; **Martínez, Miquel** (2017). "La ética del aprendizaje servicio en la universidad: una interpretación desde el pragmatismo". *Bordón. Revista de pedagogía*, v. 69, n. 4, pp. 63-78.
https://doi.org/10.13042/BORDON.2017.690405

González-Geraldo, José L.; **Ortega-López, Leticia** (2023). "Valid but not (too) reliable? Discriminating the potential of ChatGPT within higher education". In: Carmo, Mafalda (ed.). *Education and new developments 2023. Volume 2*. Lisbon: Science Press, pp. 575-579.
https://end-educationconference.org/wp-content/uploads/2023/07/2023v2end127.pdf

Hosseini, Mohammad; **Horbach, Serge P. J. M.** (2023). "Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review". *Research integrity and peer review*, v. 8, n. 4.
https://doi.org/10.1186/s41073-023-00133-5

Igelmo, Jon; **Jover, Gonzalo** (2019). "Cuestionando la narrativa del aprendizaje servicio a partir de dos iniciativas de extensión social universitaria de orientación católica en la década de 1950 en España". *Utopía y praxis latinoamericana*, v. 24, n. 87, pp. 151-162.
https://doi.org/10.5281/zenodo.3464055

Jalil, Sajed; **Rafi, Suzzana**; **LaToza, Thomas D.**; **Moran, Kevin**; **Lam, Wing** (2023). "ChatGPT and software testing education: Promises & perils". In: *2023 IEEE international conference on software testing, verification and validation workshops (ICSTW)*, pp. 4130-4137.
https://doi.org/10.1109/ICSTW58534.2023.00078

Jover, Gonzalo; **Fleta, Teresa**; **González-García, Rosa** (2016). "La formación inicial de los maestros de educación primaria en el contexto de la enseñanza bilingüe en lengua extranjera". *Bordón. Revista de pedagogía*, v. 68, n. 2, pp. 121-135.
https://doi.org/10.13042/BORDON.2016.68208

Jover, Gonzalo; **Gozálvez, Vicent** (2012). "La universidad como espacio público un análisis a partir de dos debates en torno al pragmatismo". *Bordón. Revista de pedagogía*, v. 64, n. 3, pp. 39-52.
https://recyt.fecyt.es/index.php/BORDON/article/view/22034

Kasneci, Enkelejda; **Sessler, Kathrin**; **Küchemann, Stefan**; **Bannert, Maria**; **Dementieva, Daryna**; **Fischer, Frank**; **Gasse, Urs**; **Groh, Georg**; **Günnemann, Stephan**; **Hüllermeier, Eyke**; **Krusche, Stephan**; **Kutyniok, Gitta**; *et al.* (2023). "ChatGPT for good? On opportunities and challenges of large language models for education". *Learning and individual differences*, v. 103, 102274.
https://doi.org/10.1016/j.lindif.2023.102274

Lin, Jialiang; **Song, Jiaxin**; **Zhou, Zhangping**; **Chen, Yidong**; **Shi, Xiaodong** (2023). "Automated scholarly paper review: Concepts, technologies and challenges". *Information fusion*, v. 98, 101830.
https://doi.org/10.1016/j.inffus.2023.101830

Lira, Rodrigo-Pessoa-Cavalcanti; **Rocha, Eduardo-Melani**; **Kara-Junior, Newton**; **Costa, Dácio-Carvalho**; **Procianoy, Fernando**; **De-Paula, Jayter-Silva**; **Gracitelli, Carolina P. B.**; **Prata, Tiago-da-Silva**; **Regatieri, Caio V.**; **Biccas-Neto, Laurentino**; **Alves, Monica** (2023). "Challenges and advantages of being a scientific journal editor in the era of ChatGPT". *Arquivos brasileiros de oftalmologia*, v. 86, n. 3, pp. 5-7.
https://doi.org/10.5935/0004-2749.2023-1003

Marcus, Gary (2022). "How come GPT can seem so brilliant one minute and so breathtakingly dumb the next?". *Marcus on AI*, December 1.
https://garymarcus.substack.com/p/how-come-gpt-can-seem-so-brilliant

Monroy, Fuensanta; **González-Geraldo, José L.** (2022). "Development of a procrastination scale in Spanish and measurement of education students' procrastination levels". *Bordón. Revista de pedagogía*, v. 74, n. 2, pp. 63-76.
https://doi.org/10.13042/Bordon.2022.93054

Peña-Fernández, Simón; **Meso-Ayerdi, Koldobika**; **Larrondo-Urena, Ainara**; **Díaz-Noci, Javier** (2023). "Sin periodistas, no hay periodismo. La dimensión social de la inteligencia artificial generativa en los medios de comunicación". *Profesional de la información*, v. *32*, n. 2, e320227.
https://doi.org/10.3145/epi.2023.mar.27

Perkins, Mike (2023). "Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond". *Journal of university teaching & learning practice*, v. 20, n. 2, Article 07.
https://doi.org/10.53761/1.20.02.07

Rudolph, Jürgen; **Tan, Samson**; **Tan, Shannon** (2023). "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education". *Journal of applied learning & teaching*, v. 6, n. 1.
https://doi.org/10.37074/jalt.2023.6.1.9

Santandreu-Calonge, David; **Medina-Aguerrebere, Pablo**; **Hultberg, Patrik**; **Shah, Mariam-Aman** (2023). "Can ChatGPT improve communication in hospitals?". *Profesional de la información*, v. 32, n. 2, e320219.
https://doi.org/10.3145/epi.2023.mar.19

Schulz, Robert; **Barnett, Adrian**; **Bernard, René**; **Brown, Nicholas J.L.**; **Byrne, Jennifer A.**; **Eckmann, Peter**; **Gazda, Małgorzata A.**; **Kilicoglu, Halil**; **Prager, Eric M.**; **Salholz-Hillel, Maia**; **Ter-Riet, Gerben**; **Vines, Timothy**; *et al.* (2022). "Is the future of peer review automated?". *BMC research notes*, v. 15, n. 203.
https://doi.org/10.1186/s13104-022-06080-6

Severin, Anna; **Strinzel, Michaela**; **Egger, Matthias**; **Barros, Tiago**; **Sokolov, Alexander**; **Mouatt, Julia-Vilstrup**; **Müller, Stefan** (2022). "Journal impact factor and peer review thoroughness and helpfulness: A supervised machine learning study". *arXiv*, 2207.09821.
https://doi.org/10.48550/arXiv.2207.09821

Sok, Sarin; **Heng, Kimkong** (2023). "ChatGPT for education and research: a review of benefits and risks". *Social science research network (SSRN)*, March 9.
https://doi.org/10.2139/ssrn.4378735

Srivastava, Mashrin (2023). "A day in the life of ChatGPT as an academic reviewer: Investigating the potential of large language model for scientific literature review". *OSF preprints*, February 16.
https://doi.org/10.31219/osf.io/wydct

Švab, Igor; **Klemenc-Ketiš, Zalika**; **Zupanič, Saša** (2023). "New challenges in scientific publications: Referencing, artificial intelligence and ChatGPT". *Slovenian journal of public health*, v. 62, n. 3, pp. 109-112.
https://doi.org/10.2478/sjph-2023-0015

Tlili, Ahmed; **Shehata, Boulus**; **Adakwah, Michael-Agyemang**; **Bozkurt, Aras**; **Hickey, Daniel T.**; **Huang, Ronghuai**; **Agyemang, Brighter** (2023). "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education". *Smart learning environments*, v. 10, n. 15.
https://doi.org/10.1186/s40561-023-00237-x

Vaswani, Ashish; **Shazeer, Noam**; **Parmar, Niki**; **Uszkoreit, Jakob**; **Jones, Llion**; **Gómez, Aidan N.**; **Kaiser, Łukasz**; **Polosukhin, Illia** (2017). "Attention is all you need". In: *NIPS'17: Proceedings of the 31st international conference on neural information processing systems*, pp. 6000-6010.
https://dl.acm.org/doi/pdf/10.5555/3295222.3295349

Wang, Xuezhi; **Wei, Jason**; **Schuurmans, Dale**; **Le, Quoc**; **Chi, Ed**; **Narang, Sharan**; **Chowdhery, Aakanksha**; **Zhou, Denny** (2022). "Self-consistency improves chain of thought reasoning in language models". *arXiv*, 2203.11171v4.
https://doi.org/10.48550/arXiv.2203.11171

Zhai, Xiaoming (2023). "ChatGPT for next generation science learning". *Crossroads*, v. 29, n. 3, pp. 42-46.
https://doi.org/10.1145/3589649

7. Annex 1

Reviews obtained. *ChatPDF* (rev. 1 and rev. 3. *GPT-3.5 turbo*) and *Bing* (rev. 2. *GPT-4*)

	Paper 1			Paper 2			Paper 3			Paper 4			Paper 5		
	Rev. 1	Rev. 2	Rev. 3	Rev. 1	Rev. 2	Rev. 3	Rev. 1	Rev. 2	Rev. 3	Rev. 1	Rev. 2	Rev. 3	Rev. 1	Rev. 2	Rev. 3
Clasif.	B	B	B	B	B	A	A	B	A	B	B	B	A	A	A
1	1	1	-	-	5	5	1	5	1	1	5	1	2	5	1
	1	1	2	-	5	2	1	5	3	1	4	1	1	5	1
2	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
	5	-	5	5	5	5	5	5	5	5	5	5	5	3	5
3	5	4	5	5	5	5	5	5	4	5	5	5	5	4	5
4	5	1	5	5	1	5	5	5	5	5	5	5	5	2	5
	5	1	5	5	1	5	5	5	5	5	5	5	5	1	5
5	5	1	5	5	1	-	5	-	5	1	-	1	1	1	5
6	5	4	5	5	5	5	5	5	5	5	5	5	5	4	5
7	3	3	3	3	1	3	4	-	4	1	-	1	3	3	4
8	5	4	5	2	5	5	5	5	5	4	5	5	5	4	5
	5	4	5	4	5	5	5	5	5	5	5	5	5	5	5
9	5	4	4	4	5	4	5	5	5	5	5	5	5	5	5
10	5	4	5	4	5	5	5	5	5	5	5	5	5	5	5
11	4	-	4	4	-	5	4	-	4	4	-	5	4	-	4
12	5	3	4	4	4	5	4	4	5	5	4	5	5	5	5
13	3	-	5	4	-	5	5	3	4	5	5	5	5	5	5
14	1	-	3	-	-	5	4	3	4	3	-	1	4	4	4
15	1	-	3	-	-	5	4	3	4	3	3	3	5	5	5
16	1	4	3	4	-	5	4	3	4	3	5	5	5	5	5
17	4	1	4	4	4	5	4	-	4	5	5	5	4	5	4
Decisión	B	C	B	B	B	B	B	A	B	B	A	B	B	B	A

Classification: A) Empirical research (quantitative or qualitative), B) Theoretical research, essay, C) Educational experience or innovation, and D) Other. Criteria: 1) IMRaD format of the abstract / Extension of the abstract, 2) Adequacy of the Title / keywords, 3) Spelling and syntactic correction, 4) *APA* standards / coherence between citations and bibliographic references, 5) Tables and figures, 6) Interest of the article for the educational community, 7) Generalization of the results, 8) Originality of the work / contribution to educational knowledge, 9) Introduction and justification of the importance of the topic, 10) Theoretical foundation, 11) Relevance of the sources cited according to the year of publication, 12) Formulation of objectives, 13) Process of collection and analysis of information, 14) Description of the sampling procedure, 15) Process of collection and analysis of information, 16) Presentation and description of results, 17) Conclusions and discussion.
Final decision: A) Publish as is or with minor modifications of wording and / or format, B) It could be published once the suggested corrections and improvements have been made, C) Do not publish for the specified reasons.