Understanding the Impact of Artificial Intelligence in Academic Writing: Metadata to the Rescue

Javier Conde, **Pedro Reviriego**, and **Joaquín Salvachúa**,
Universidad Politécnica de Madrid

Gonzalo Martínez and **José Alberto Hernández**, Universidad Carlos III de Madrid

Fabrizio Lombardi, Northeastern University

This column advocates for including artificial intelligence (AI)–specific metadata on those academic papers that are written with the help of AI in an attempt to analyze the use of such tools for disseminating research.

The rapid development and exponential adoption of artificial intelligence (AI) tools that can generate text for different tasks are poised to have significant and far-reaching implications in many sectors.[1] These tools can, for example, summarize, translate, or paraphrase text but also can write text on any topic and provide relevant citations.[2] These capabilities are useful for academic writing, and many researchers are starting to use AI tools as assistants when writing papers.[3] The use of AI tools poses many potential issues, ranging from the accuracy of the text generated, which may contain false statements, to ethical concerns.[4] Therefore, the impact of AI tools should be carefully analyzed, not only in research datasets but also in real data as AI-generated content starts to become widespread and for academic writing in published papers.

There are many angles to be considered when analyzing the impact

of AI tools in academic writing. For example, will the use of AI tools have an impact on the citations of the papers? Which authors are more prone to use those tools? Which tools are more popular and for which tasks? Those are some of the basic questions, but more fundamental aspects have to be considered. Initial studies suggest that AI tools will have an impact on the language itself.[5,6] So, for example, will

> The questions are numerous, and soon academic papers written with the help of AI tools will become common.

linguistic features of papers written with the assistance of AI tools be different from those of human-written papers? Will the use of AI tools introduce biases in the vocabulary[7] used? The questions are numerous, and soon academic papers written with the help of AI tools will become common.

To analyze the impacts of these AI tools, the first step is to reliably identify AI-generated content with as many details as possible; for example, knowing the AI tool, its version, and the tasks for which it was used to assist in the writing. Unfortunately, such information is generally not available in academic papers. A potential solution is to use tools designed to detect AI-generated text.[8] This, however, has many limitations because these tools have limited accuracy and have to be constantly evolving to keep track of new AI-generative tools and features. Even if accurate, such tools can only provide limited information but not the details on the AI tool and version used. Another possible direction is the policies being implemented by many publishers, for example, IEEE, to request that authors disclose the use of AI tools in the acknowledgment section. However, this approach also has limitations because there is no standard form to report the use of AI tools and provide additional

details. Finally, for both approaches, even if the information could be extracted reliably (which is not the case), gathering even basic information, such as the percentage of papers that have used a given tool requires checking the full text of all papers. This is clearly not efficient.

A more scalable and future-proof solution is to add metadata describing the use of AI tools in each academic paper. This enables queries on the metadata on large numbers of documents but also the definition of common metadata formats that enable interoperability among different publishers. For example, the addition of the tool, version, and task makes answering the first set of simple questions discussed previously trivial. Interestingly, these metadata enable the generation of a large corpora of text written with the assistance of AI tools from which the answers to the second set of questions can be extracted. The use of metadata also enables tracking the evolution of the use of AI tools, thereby making possible comparisons among different tools or different versions of the same tool. Next, we discuss the metadata that should be added and illustrate the potential benefits.

METADATA FOR THE USE OF AI IN ACADEMIC WRITING

Information on the use of AI in an academic paper can be captured at different levels of detail, from a simple flag indicating the use of AI to assist the authors to detailed logs of the interactions with the AI tool. As it tends to happen, too little or too much detail is not good, so next, we try to list the main information that could be relevant to analyzing the impact of AI in academic writing:

› AI tool used
› version of the tool
› main parameters of the tool
› use of the tool (translation, summarization, writing, citations, etc.)
› parts of the paper on which the tool was used.

The information can be divided into three categories: which tool was used, how it was used, and where it was used. The knowledge of the tool enables a comparative analysis of different tools but also the study of the evolution of a given tool or its parameters. Similarly, knowing for what the tool was used provides information for analyzing the impact of AI on different aspects of academic writing. Finally, knowing where it was used enables the extraction of the relevant sections of the paper for further analysis. The three categories are illustrated in Figure 1.

Which AI tools and parameters
This group of metadata must capture the tool, version, and configuration. The data can be coded in JavaScript Object Notation, XML, or any other convenient format to automate its retrieval and processing. The following list is an example of how the data could look:

› *Name*: GPT-4
› *Description*: GPT-4 is a large multimodal model (accepting text inputs and emitting text outputs today, with image inputs coming in the future) that can solve difficult problems with greater accuracy than any of our previous models due to its broader general knowledge and advanced reasoning capabilities.
› *Version*: gpt-4-0613
› *URL*: https://platform.openai.com/docs/models/gpt-4
› *Author*: OpenAI
› *Web*: https://chat.openai.com/
› *Size*: not disclosed
› *Window*: 8,192 tokens

› *Temperature*: zero
› *Context*: default.

The metadata include information on the AI model itself but also on the specific parameters' values selected when using it to write the paper. This enables the evaluation of the impact of the model parameters on the written text.

Where AI was used

The information on where AI tools were used can be described with the parts of the papers—for example, the abstract, each of the sections, the figures, the tables, or the references. This enables the identification of the text for which AI assistance has been used.

How AI was used

AI tools can be used for many different tasks: summarizing, translation, paraphrasing, finding related work and citations, etc. Therefore, it is important to have information on how AI tools were used in the paper. For example, we can encode in the metadata that GPT-4 (the "which") was used to summarize (the "how") and write the abstract (the "where").

UNDERSTANDING THE IMPACT OF AI IN ACADEMIC WRITING

Let us consider now that we have a large corpus of papers, and we want to know how many of them have used AI to summarize the abstract. Without metadata, all papers look the same [Figure 2(a)], so we have to extract the text and either try to detect the use of AI in the abstract or find a disclosure from the authors that states the use of AI in the abstract. Instead, if the proposed metadata have been added to the paper,

---

A more scalable and future-proof solution is to add metadata describing the use of AI tools in each academic paper.

---

we can just look at the how (summarizing) and where (abstract) to find the papers. The papers are now marked and can be easily identified [Figure 2(b)].

The metadata can be used to analyze many aspects of the use of AI in academic writing; for example, we can analyze the following:

› the adoption of the different AI tools and their variations over time

› the tasks for which AI tools are more frequently used
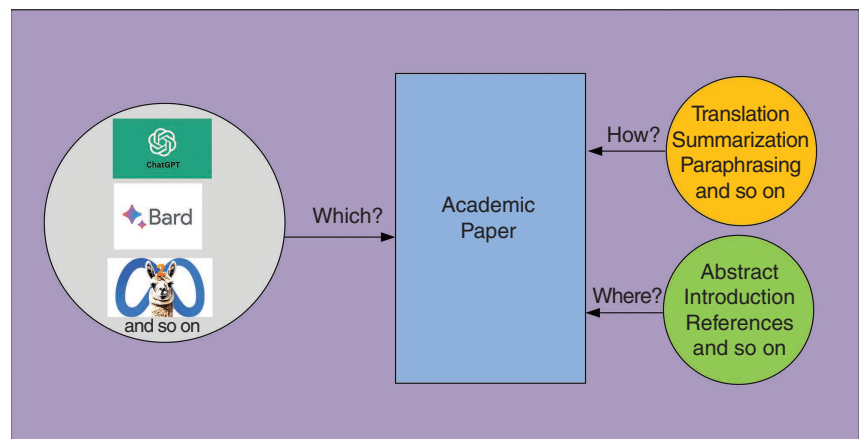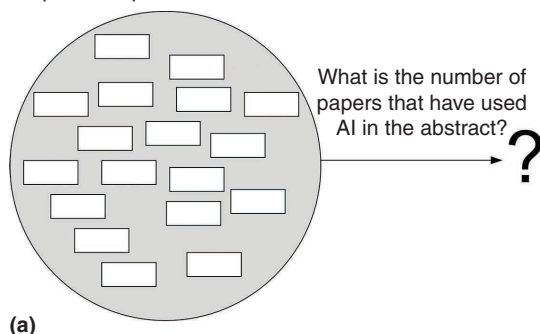› the parts of the papers for which AI tools are more widely used

FIGURE 1. The relevant information (which, how, and where) on the use of AI in an academic paper.
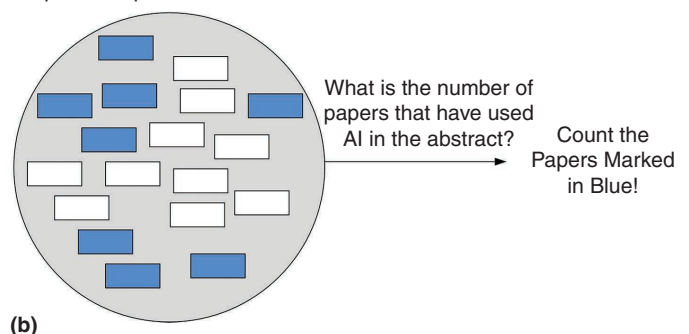

FIGURE 2. Finding papers with specific AI use features. (a) No metadata are used, so all papers look the same and have to be parsed to try to extract the information. (b) Metadata are used, and papers can be easily identified without having to access the full text.

- the correlation between the use of AI in a paper and its citation and popularity metrics
- the use of AI tools in different journals and conferences by the same publisher
- the use of AI tools by authors of different regions and institutions
- the use of AI tools by an author over time

These are just a few simple examples that show how adding simple metadata to academic publications makes it possible to gain different insights into the impact of AI tools in academic writing. These analyses can be easily run and automated once the metadata are integrated into the publisher's information systems, opening the possibility of periodic reporting and analysis.

The proposed metadata not only enable purely AI-related studies but also those that combine existing metadata, such as, for example, the number of citations and the use of AI tools. The combinations of parameters that can be studied are endless; one possibility is to feed the raw data into a machine learning system to extract patterns and relations.

In addition to the analysis of the metadata, it is also possible to use the metadata to perform analysis of the content of the main body of the papers. This is of interest when studying the impact of AI tools on the linguistic features of the text generated,[6] for example, lexical richness.[7] Using the metadata, the text in papers (or sections within those papers) in which AI tools have been used can be extracted to build a corpus of text generated, for example, with the assistance of a given tool. More generally, corpora of text for a given set of AI metadata values can be easily generated. This is very interesting because it enables the analysis of the impact of AI tools on real data generated by many different users. The availability of such corpora opens new possibilities, such as, for example:

- analyzing how AI tools affect the features of generated text depending on the tool and version
- analyzing how AI tools affect the features of generated text depending on the task
- comparing text generated by the same authors with and without the help of AI tools
- comparing text generated by native and nonnative speakers as authors with and without AI tools

- providing datasets of user-generated data to develop and validate tools that detect AI-generated text
- comparing text generated by authors when using AI tools with text generated directly by AI tools to understand how the assistance rather than the direct use of AI tools impacts the text.

The overall scenario is illustrated in Figure 3 with an example that shows how metadata can be used to understand the summarizing process with AI tools as used in different journals and also how metadata are combined with full text to analyze the lexical richness of AI-generated abstracts.

This column proposes adding metadata on the use of AI to scientific publications. The provision of having these metadata is critical for the analysis and understanding of AI and its impact on academic writing. It is important to note that the implementation of this solution will require changes to scientific journal and academic database systems by adding new fields to store the new metadata. ⬛

REFERENCES
1. T. Wu et al., "A brief overview of ChatGPT: The history, status quo and potential future development," *IEEE/CAA J. Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, May 2023, doi: 10.1109/JAS.2023.123618.
2. P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Physical Syst.*, vol. 3, pp. 121–154, Apr. 2023, doi: 10.1016/
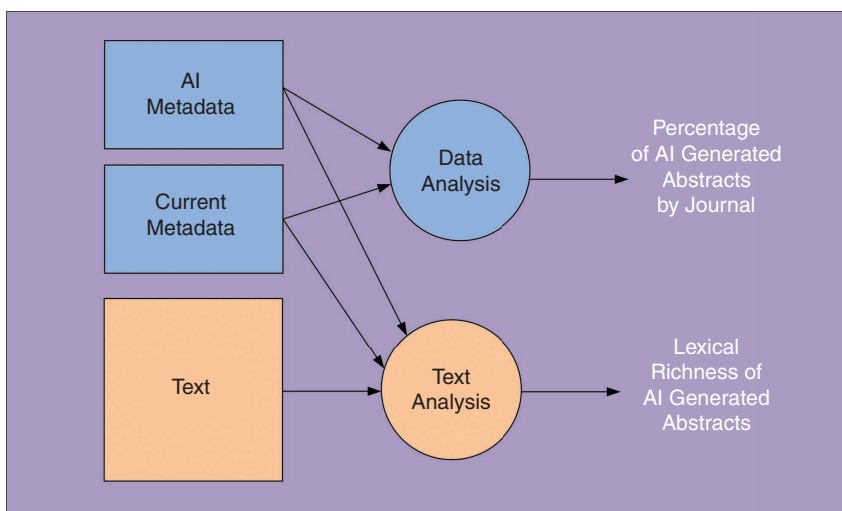
FIGURE 3. Use of AI metadata for data and text analysis.

j.iotcps.2023.04.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266734522300024X

3. N. Curtis and ChatGPT, "To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing," *Pediatric Infectious Dis. J.*, vol. 42, no. 4, p. 275, Apr. 2023, doi: 10.1097/INF.0000000000003852.

4. B. Marchandot, K. Matsushita, A. Carmona, A. Trimaille, and O. Morel, "ChatGPT: The next frontier in academic writing for cardiologists or a pandora's box of ethical dilemmas," *Eur. Heart J. Open*, vol. 3, no. 2, Feb. 2023, Art. no. oead007, doi: 10.1093/ehjopen/oead007.

5. Z. G. Cai, D. A. Haslett, X. Duan, S. Wang, and M. J. Pickering, "Does ChatGPT resemble humans in language use?" 2023, *arXiv:2303.08014*.

6. J. M. Toro, "Emergence of a phonological bias in ChatGPT," 2023, *arXiv:2305.15929*.

7. P. Reviriego, J. Conde, E. Merino-Gómez, G. Martínez, and J. A. Hernández, "Playing with words: Comparing the vocabulary and lexical richness of ChatGPT and humans," 2023, *arXiv:2308.07462*.

8. Z. Liu, Z. Yao, F. Li, and B. Luo, "Check me if you can: Detecting ChatGPT-generated academic writing using CheckGPT," 2023, *arXiv:2306.05524*.

JAVIER CONDE is an assistant professor at the Universidad Politécnica de Madrid, 28040 Madrid, Spain. Contact him at javier.conde.diaz@upm.es.

PEDRO REVIRIEGO is an associate professor at the Universidad Politécnica de Madrid, 28040 Madrid, Spain. Contact him at pedro.reviriego@upm.es.

JOAQUÍN SALVACHÚA is an associate professor at the Universidad Politécnica de Madrid, 28040 Madrid, Spain. Contact him at joaquin.salvachua@upm.es.

GONZALO MARTÍNEZ is a Ph.D. student at the Universidad Carlos III de Madrid, 28911 Madrid, Spain. Contact him at gonzmart@pa.uc3m.es.

JOSÉ ALBERTO HERNÁNDEZ is an associate professor at the Universidad Carlos III de Madrid, 28911 Madrid, Spain. Contact him at jahgutie@it.uc3m.es.

FABRIZIO LOMBARDI is a professor at Northeastern University, Boston, MA 02115 USA. Contact him at lombardi@coe.northeastern.edu.