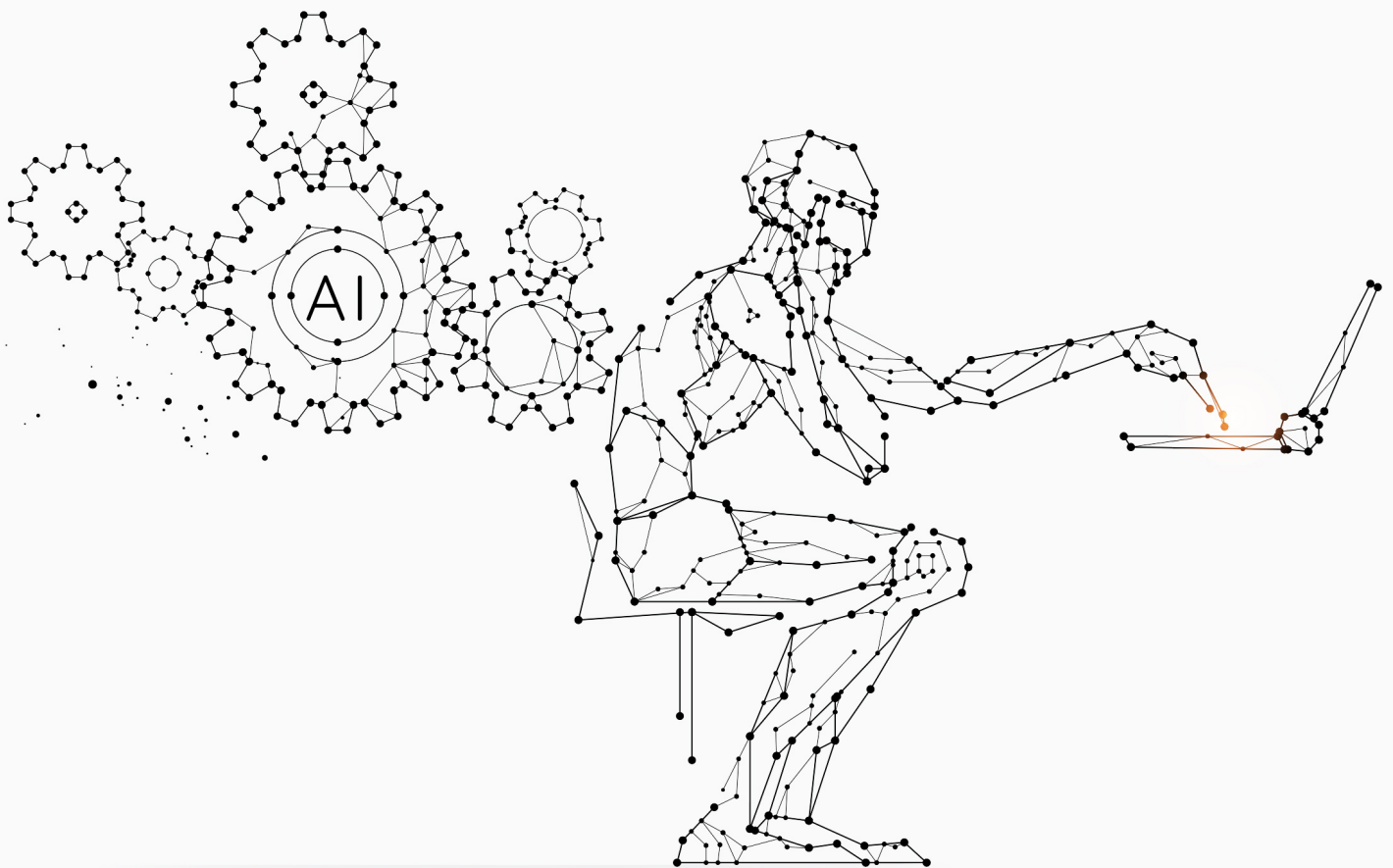




AI-based editing tools for researchers

A comparative analysis






Communicating research effectively is a challenging task and more so when writing quality is a critical factor involved.

English is the dominant language in which research is communicated, with over 80% of indexed scholarly articles being published in English.¹

Whether you are a seasoned English speaker or someone who speaks English as a second language, you may surely relate to the difficulties associated with writing about your research.

Irrespective of your level of English proficiency, your manuscripts may have errors in spelling/grammar, terminology use, and word choice, especially if you are working on deadlines. This is where automated editing tools can offer significant advantage since they can help you make your research manuscripts submission-ready.

Therefore, a comparison of the capabilities of such tools is relevant and useful and is the purpose of this white paper.



Artificial intelligence is helping academic authors achieve their most important goals

Sophisticated AI-driven language solutions are transforming scholarly communications.² These solutions include AI tools that can improve the quality of submitted manuscripts in a few minutes.

These tools are based on natural language processing (NLP), which is a subset of AI that helps computers understand, interpret, and use language just like human beings can. With “training” on data related to how humans write or edit academic text, an AI-powered editing tool “learns” to recognize and understand the nuances of academic language and keeps improving over time.

These tools automatically detect and correct errors, and flag complex issues too. Now, you can benefit from this cutting-edge technology to improve the grammar, punctuation, tone, style, and formatting of your manuscripts, which, in turn, can increase their chances of acceptance.

Why do you need this white paper?

As researchers, some of you may be optimistic about the advantages of AI-based editing programs. And some of you may be skeptical about them, wondering how useful they are in correcting errors, if they have any inherent biases, and how reliable they are given the lack of “human” judgment.

Whatever your level of confidence in these tools, you need a way to reliably evaluate and compare their performance against human editing standards.

This white paper will help you judge the effectiveness and accuracy of different AI tools from different angles.

What is the objective of this white paper?

The standard methods used in the industry for such evaluation are based on metrics such as recall, precision, and F score.

We use these metrics to give you a comprehensive overview of the performance of 5 AI-powered tools used in the industry:



Paperpal, AJE Digital, and Writefull-FE cater to academic editing, while Grammarly and Instatext support general writing. Although Grammarly offers the option to choose “academic” as a domain, it has not been designed to exclusively serve academic purposes.

In this paper, we

analyze the language-editing performance of the 5 tools;

explain the evaluation approaches and results in a simple and jargon-free manner so that you are equipped to judge tool performance yourself even if you are unfamiliar with AI-related terminologies; and

assess and compare the performance of various tools in editing actual scholarly content from different subject areas.

Analyses

We compared the performance of Paperpal with that of Grammarly (with the “academic” setting), AJE Digital, Instatext, and Writefull.

We selected six academic writing samples, one each from the following subject areas: medicine, social sciences, engineering and technology, materials science, business and finance, and life sciences. Each sample was between 400 and 650 words long, and each was run through the 5 AI tools, giving us 30 tool-edited samples.

For an all-round comparison of the tools, we performed a three-fold assessment using these AI-edited samples to answer the following questions:

- 01 How well do the tools perform compared with a human expert?
- 02 How well do the tools perform against a broad repertoire of multiple human editors?
- 03 How well can the tools assist human editing?

01

How well do the tools perform compared with a human expert?

Flow of the analysis

An academic editor first reviewed each of the six original writing samples and marked out all errors (instances where language-related corrections were necessary). The editor was then asked to review the tool-edited versions (while being blinded to the tools used) and assess editing performance in terms of the following outputs and measures:

Definitions of editing outputs

Gold standard

All corrections made by the human expert

Total edits

The total of all tool-proposed edits of any nature

Correct edits

Tool edits that matched with the human (gold) edits

Incorrect edits

Incorrect edits: Errors introduced by the tool [tool edits that didn't match the human (gold) edits and were incorrect]

Improvements

Tool edits that didn't match the human (gold) edits but were enhancements/improvements over the gold edits

Neutral edits

Changes that do not fall in any of the above categories because they are optional or stylistic [tool edits that didn't match the human (gold) edits and were neutral (neither incorrect nor an improvement)]

Missed edits

Missed edits: Instances where a human edit was present, but the tool missed marking an edit

Performance measures

We compared the performance of the 5 tools by calculating the percentages of correct, incorrect, and missed edits.

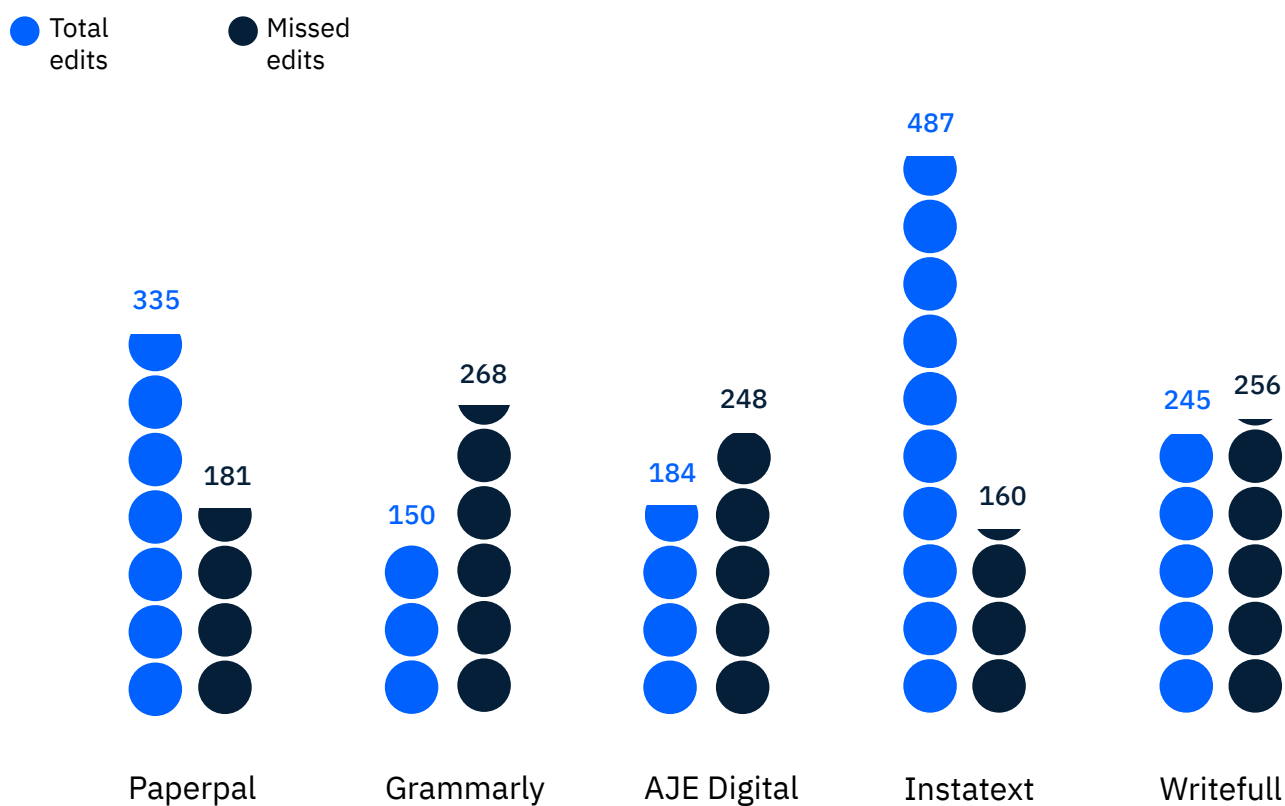
What we found

- Grammarly had the lowest number of total edits and the highest number of missed edits. This indicates that while Grammarly is a highly popular tool, it might be a poor choice for academic editing even though it allows users to choose “academic” as a domain (Fig. 1a).
- AJE Digital and Paperpal had the highest percentage of correct edits and the lowest percentage of incorrect edits (Fig. 1b). However, AJE Digital had a higher number of missed edits than did Paperpal.
- While Instatext had the highest number of total edits and fewest missed edits, it also had the highest percentage of neutral edits and a much lower percentage of correct edits than did Paperpal and AJE. This suggests that merely the total number of tool-proposed edits is not a meaningful measure of how effective an AI editing tool is.

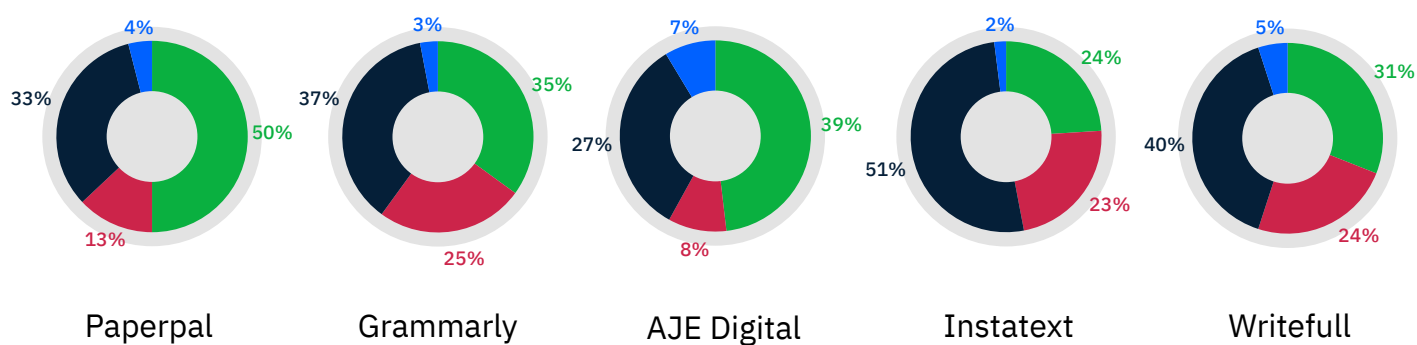
In summary, Paperpal and AJE Digital outperformed Grammarly, Instatext, and Writefull in terms of accuracy of editing.

Fig. 1 Comparison of tool performance based on human expert assessment

(a) Number of total and missed edits



(b) Tool performance based on edit categories



02

How well do the 5 tools perform against a broad repertoire of multiple human editors?

Human editing output depends on an editor's skill, experience, and even personal style. To rule out any biases creeping into the analysis because of the editorial preferences/repertoire of a single human editor, we evaluated tool performance by comparing tool-proposed edits with the edits of three human editors.

This would enhance the set of gold-standard edits and bring it closer to the ground truth. Here, we focused on whether the tools were able to recognize the same errors that human editors would.

Definitions of editing outputs

Gold standard

A set of the best of the three edits for each sentence (see Appendix for details)

Total edits

The total of all tool-proposed edits of any nature

Correct edits

Tool edits that matched with the gold edits

Incorrect edits

Errors introduced by the tool (tool edits that didn't match the gold edits and were incorrect)

Performance measures

Recall

A measure of how many of the gold standard edits the tool makes/proposes. For example, if a document has 10 language errors and the tool corrects 5, the recall is 50%.

Precision

An indicator of how many of the tool edits made are correct. For example, if the tool makes or proposes 5 edits and all 5 are correct (i.e., match with the gold standard edits), the precision is 100%.

F-score

Harmonic mean of precision and recall

What we found

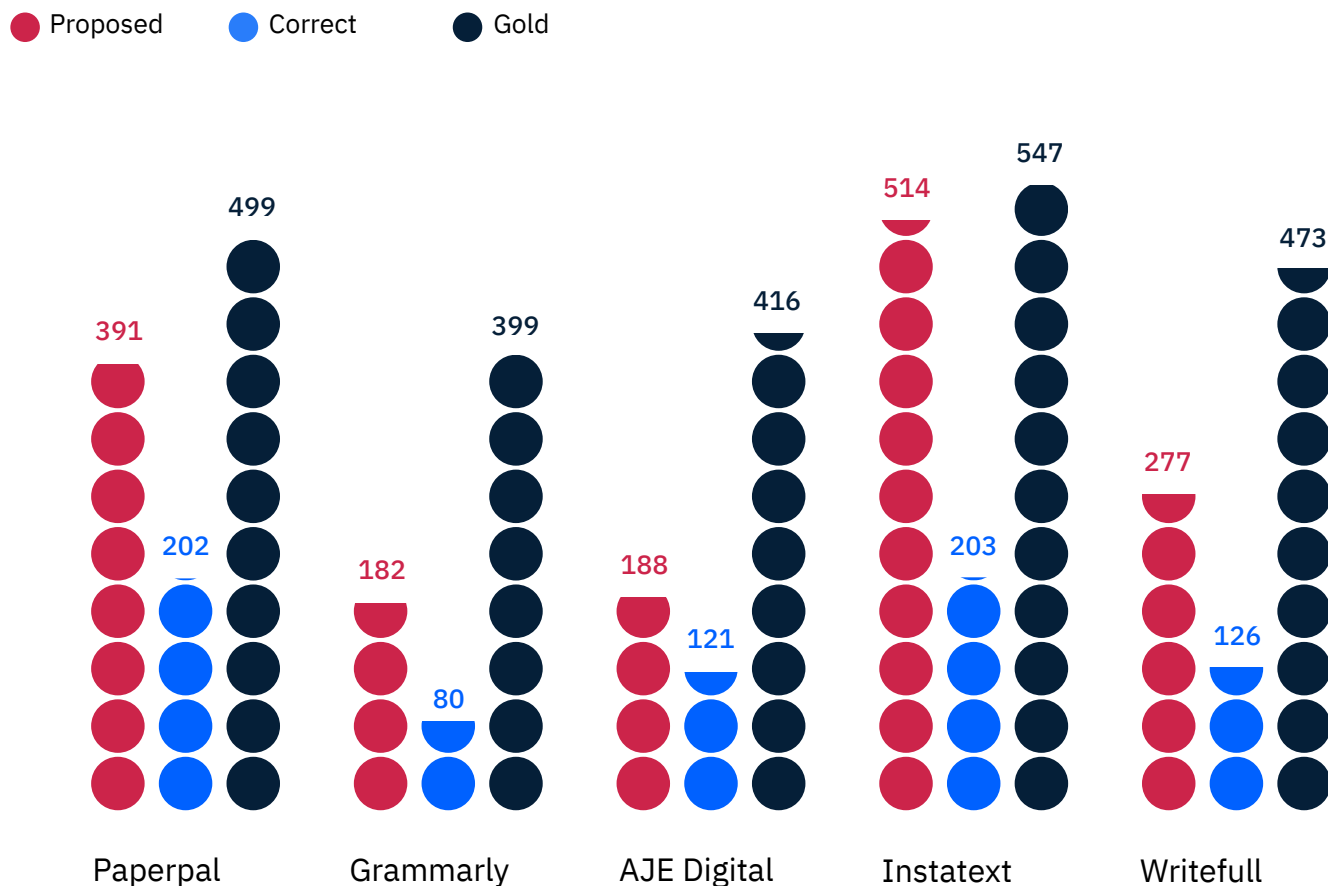
- Consistently across the multiple-editor panel, the top two performers for precision were AJE Digital and Paperpal. In terms of error coverage, Paperpal fared better, with higher recall (Fig. 2b). This could be because AJE Digital made fewer correct edits, almost half the number of those made by Paperpal (Fig. 2a).

It is worth noting here that Paperpal performed well in terms of both recall and precision.

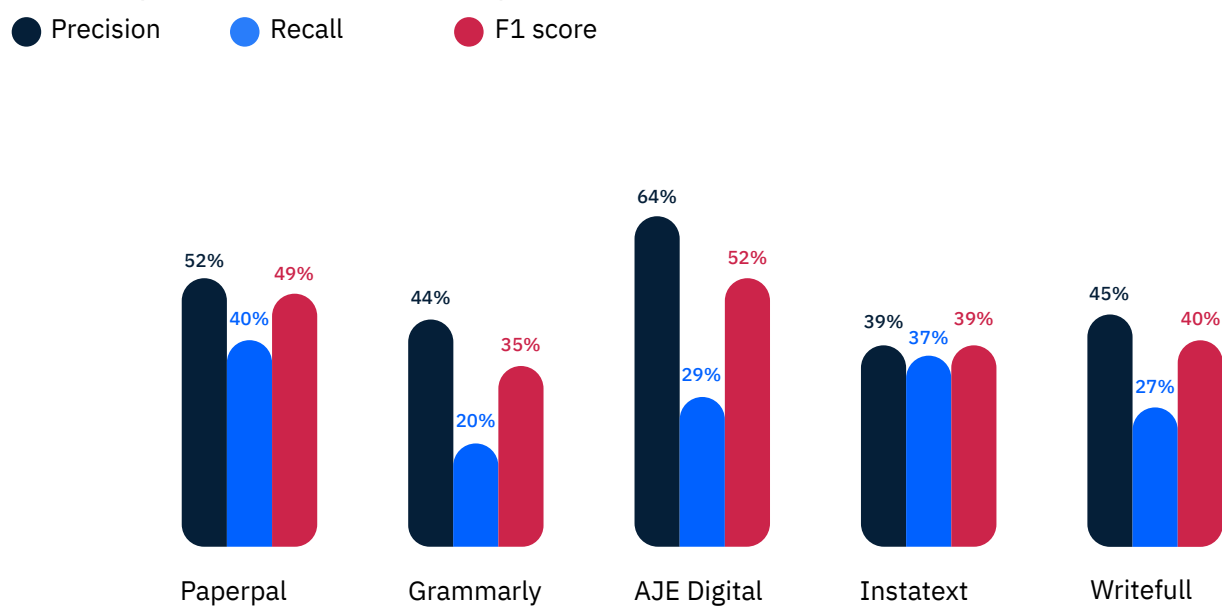
- The F score varied considerably. It conflates both recall and precision, which tend to be inversely related to each other. Thus, a moderately high F-score might point to an imbalance between precision and recall.

Fig. 2 Comparison of tools against multi-annotator data

(a) Number of gold, correct, and proposed edits



(b) Tool performance based on precision, recall, and F1 score



03

How well can the 5 tools assist human editing?

Human editors reviewed manuscripts edited by the different tools (without knowing which tool was used). They identified/corrected mistakes, retained tool-recommended changes that were correct, and made additional changes that the tool had missed.

The following parameters of editorial output and tool performance were compared:

Definitions of editorial outputs

Gold standard

The final number of changes in each manuscript (tool proposed + human-made) after the human editor completed the process

Total edits

The total of all tool-proposed edits

Correct edits

Tool-proposed revisions that the human editor retained

Inorrect edits

Tool-proposed edits that the editor rejected

Performance measures

Recall

An estimate of how many of the correct results are found, calculated as tool-proposed edits retained by human editor/gold edits

Precision

An indicator of how many of the edits made are correct. Calculated as tool edits retained by human editor/total tool-proposed edits

F-score

Harmonic mean of precision and recall

What we found

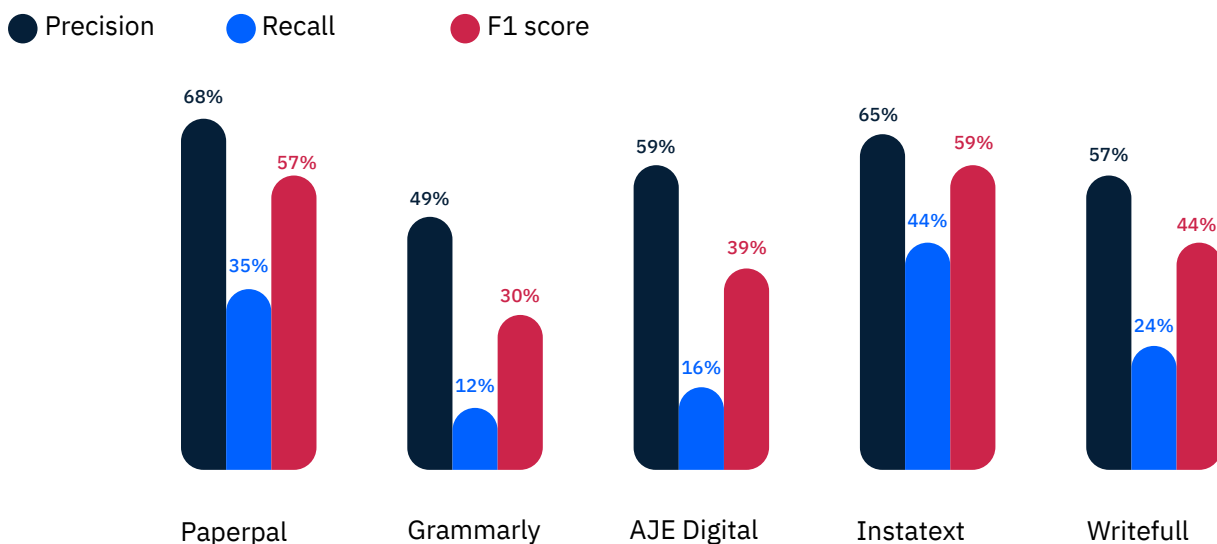
A high recall means that the tool flags errors at a high rate, and a high precision means that a human editor will spend less time undoing incorrect revisions made by the tool.

Paperpal outperformed all the 5 tools in terms of precision (Fig. 3a). For recall, it was second to Instatext. AJE Digital, which performed well in the previous analysis, did not seem to support human editing to the extent that Paperpal and Instatext do.

Paperpal and Instatext support real-world academic writing more efficiently than the other tools can.

Fig. 3 Comparison of tools based on how well they support human editing

(a) Tool performance based on precision, recall, and F1 score

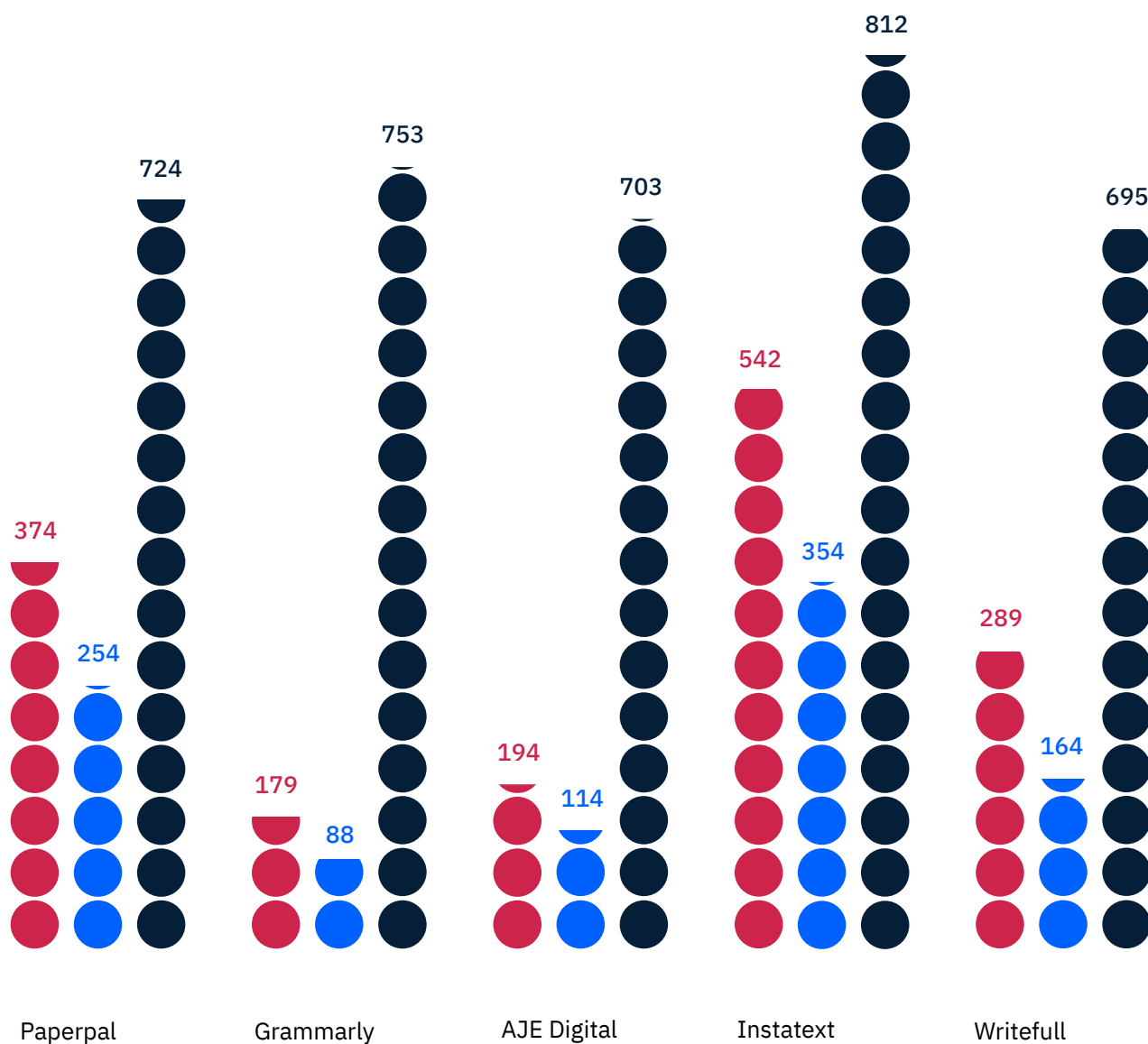


(b) Number of gold, correct, and proposed edits

● Proposed

● Correct

● Gold



Concluding remarks

This analysis offers unique and essential perspectives on how to evaluate AI-editing tools because of the following strengths:

- Transparent and unbiased assessments (lent by blinding)
- Good representation of currently used AI tools
- A wide coverage of academic fields (medicine, life sciences, physical sciences, humanities/social sciences)
- Use of industry-specific performance metrics that are robust indicators of overall accuracy and reliability

Our analysis shows that on the whole, Paperpal is the most efficient tool for researchers because it is designed specifically for academic texts, has high accuracy and precision, and can support real-world editing effectively.

From our analyses, the key points for you to consider when choosing an AI editing tool are as follows:

01

Prioritize tools meant for academic content

Tools trained on academic content help accurately edit and preserve technical aspects in scientific papers, such as subject-specific terminology and usage, units of measure, equations, and more.

For your research manuscripts, you would be well advised to choose one that has been developed specifically based on such content over general-purpose tools.

02

Choose tools trained to replicate professional human editing

Tools that have been trained merely on a database of published academic content may not be as sensitive to nuances of academic language and style as those that are trained on both original manuscripts and their human-edited versions.

It might be a good idea to check if the tool has been trained on pre- and post-edit versions of academic texts.

03

Don't rely on heavy edits alone

Considering only the number of changes a tool makes is not enough to judge whether it is reliable enough, especially if many of the changes are neither correcting errors nor enhancing the text.

The usefulness of the edits is more important than the number of edits.

04

Balance precision and recall

Tools that have both high precision and high recall are ideal. However, precision and recall are inversely related, i.e., flagging more errors versus fewer missed edits.

Ideally, a tool that fares well on both counts would strike the best balance.

How Paperpal gives you an advantage

01

High-quality editorial coverage

- **Corrections beyond proofreading**

Closely mimics human editing; doesn't just correct grammar and spellings but “understands” context and addresses complex problems such as dangling modifiers, non-parallel constructions, and incorrect collocations

- **Suggestions to improve clarity and flow**

Provides in-depth language feedback, including suggestions to improve flow based on what it “learns” from the comparisons of pre- and post-human-edit files

- **Appropriate treatment of technical components**

Ensures proper use of abbreviations, terminologies, equations, SI units, non-English words, etc.

- **Recognition of UK style vs. US style**

Detects which language style is used in the document and adopts it

02

Beyond editing checks

- Performs tasks to improve other aspects of a manuscript required for submission to a journal, for example, structural and technical checks typically performed by a journal at the screening stage

03

Continuous improvement

- Is not stagnant; is constantly evolving and becoming more sophisticated with constant learning

04

Real-time availability

- MS Word Add-in available, providing real-time edits and suggestions as you write

The importance of rapid turnarounds along with accuracy cannot be stressed enough. Timely dissemination of research findings is crucial to scientific progress. Time-pressed authors will find automated solutions like Paperpal especially valuable because they can get the accuracy of human edits at a fraction of the cost of human editing and within a few minutes. In the long run, adopting such an approach decreases workload and increases productivity for authors.

Finally, we are still a long way from complete reliance on AI for “taking over” scholarly writing and editing tasks. Human oversight still plays a defining and decision-making role. However, this analysis shows that tools like Paperpal have reached a competitive level of sophistication and can be trusted to eliminate a considerable amount of the stress authors experience when trying to have a paper published.

Appendix: Details of the approach for the three comparative analyses

01

How well do the tools perform compared with a human expert?

To get an idea of how well these tools were able to perform tasks typically carried out by humans, we had an expert academic editor review each of the six samples and mark out all errors (instances where language-related corrections were necessary).

These marked-out instances represented the most reliable data on errors in the samples that could be obtained under reasonable conditions and, thus, served as the gold-standard data set for this analysis, i.e., data that would be as close as possible to the ground truth. The expert, blinded to the information about which tools were used on the manuscripts, then assessed the tool-edited versions to compare performance.

For this analysis, the editor recorded all the edits that each tool proposed and highlighted correct/overlapping edits, incorrect edits, missing edits, neutral (unnecessary but not wrong) edits, enhancements/improvements, and missed edits.

02

How well do the tools perform against a broad repertoire of multiple human editors?

Each of the six original samples was individually edited by a random set of 3 human editors. For each sentence of a sample, the best of the three edits was chosen, and this set of best edits served as the gold-standard edits. The “best” version was determined as follows:

- For each sentence, the F-score for the edits of each editor was considered. The sentence with the best F-score was chosen.
- If the F-scores for the three edits were not different, the recall values were considered next, and then the precision values.

We refer to the resulting data as the “multi-annotator dataset.” This process eliminates the limitation posed by having a single-editor perspective.

Next, we assessed the performance of all the tools by using two measures typically used to evaluate machine-learning models: recall and precision. Recall was calculated as tool edits matching gold edits/gold edits. Precision was calculated as correct edits/total tool-proposed edits.

03

How well can the 5 tools assist human editing?

Human editors were blinded to the tools used on the previously AI-edited manuscripts. They reviewed the tool-edited manuscripts and identified/corrected mistakes as needed. They retained any tool-recommended changes that were correct and made additional changes that the tool had missed. Accordingly, total, correct, and incorrect edits were determined. In this analysis, recall was calculated as tool-proposed edits retained by human editor/gold edits, and precision was calculated as tool edits retained by human editor/total tool-proposed edits.

References

- 1 McKinley, J., Rose, H. Conceptualizations of language errors, standards, norms and nativeness in English for research publication purposes: An analysis of journal submission guidelines. *Journal of Second Language Writing* 42, 1–11 (2018).
- 2 Singh, S. What you need to know about artificial intelligence in research and publishing <https://www.editage.com/insights/what-you-need-to-know-about-artificial-intelligence-in-research-and-publishing> Editage Insights (2022).