The New Editorial Gatekeepers: Understanding LLM-based Interfaces, Their Benefits, Risks and Design

Luciano Floridi [1,2]

[1] Yale Digital Ethics Center, Yale University, 85 Trumbull Street, New Haven, CT 06511, US

[2] Department of Legal Studies, University of Bologna, Via Zamboni 22, Bologna, 40100, IT

Abstract

The article analyses the integration of Large Language Model (LLM)-based interfaces (editorial LLMs or eLLMs) in scholarly publishing workflows, focusing specifically on their growing role in editorial screening, manuscript preparation, and peer-review processes. It assesses the benefits eLLMs offer, including efficiency gains, improved compliance with journal guidelines, enhanced objectivity, and reduced editorial workload; and the risks, especially algorithmic biases, false positives and negatives, data privacy concerns, and potential opacity in automated decision-making. The article then offers some design recommendations for eLLMs that prioritise transparency, fairness, and user-centredness, ensuring human oversight remains integral to the editorial process. It concludes by encouraging a proactive and thoughtful engagement with these technologies to enhance scholarly publishing rather than undermine its values.

Keywords

Algorithmic Bias, Editorial Automation, Human–AI Collaboration, Large Language Models (LLMs), Scholarly Publishing Ethics.

1. Introduction: Framing the Debate

The emergence of large language models (LLMs) like OpenAI's ChatGPT has sparked intense debate within academic publishing circles. Headlines in early 2023 raised alarms about journals supposedly being overwhelmed by AI-generated papers, stirring significant concerns regarding scientific integrity (Stokel-Walker 2023, 2024). Although quantifying this phenomenon remains challenging and somewhat contentious (Crotty 2024), leading publishers and ethics organisations, e.g. *Nature* (Nature 2023) and COPE (COPE 2023), promptly declared that AI tools cannot receive author credit, and authors must disclose any LLM usage. This is reasonable: AI systems cannot assume responsibility for a publication's accuracy or provide meaningful consent to authorship. The scholarly community swiftly condemned early incidents where ChatGPT was erroneously listed as a co-author (Stokel-Walker 2023, Dwivedi et al. 2023). A broad consensus has thus emerged that transparency and accountability must be paramount when LLMs contribute to the writing process (Carobene et al. 2024, Tang et al. 2024). However, it remains an open question whether "distant writing" (or wrAIting) a research paper (Floridi 2025) will become increasingly accepted and normalised in the future. And I would argue that we are witnessing a shift from the author as maker of the text to the author as designer or architect. Many authors – especially those for whom English is a second language – have already embraced LLMs as writing aides to improve grammar, refine wording, or generate plain-language summaries of their findings (Katsnelson 2022). Since its public release, ChatGPT has been adopted as a "valuable writing assistant" for drafting and polishing manuscripts (Imran and Almusharraf 2023, Liu et al. 2024). Your philosopher likes to use it to ensure that the final text is slightly more idiomatic, not just grammatically correct (e.g. replacing "precedes" with "comes before").

This trend has led to a race: just as authors are leveraging LLM to enhance their writing, journals are leveraging LLM to detect LLM-generated texts. Editors worry that unscrupulous authors might submit ghostwritten text produced entirely by LLMs without disclosure, with low or no scientific or scholarly quality that may be difficult to detect easily (Májovský et al. 2023). In response, journals increasingly run manuscripts through LLM-detection software (or ask for author attestations of "no AI usage") as part of their screening. Tools like GPTZero, Turnitin's AI detector, and

others have been trialled to flag submissions with a high probability of being machine-written (Liu et al. 2024, Reinhart et al. 2025). The efficacy of these AI content detectors is debated. A growing body of evidence suggests current detectors are far from foolproof (Wu et al. 2025), not least because the same LLMs can help to hide their traces. Thus, detectors tend to perform reasonably well on unedited ChatGPT text but often falter on human-written prose or text that has been revised, sometimes even lightly. In one study, several popular detectors misclassified over half of genuine essays by non-native English writers as "AI-generated" (Liang et al. 2023). The detectors exhibited a strong bias, falsely flagging writing that uses simpler vocabulary or non-native phrasing. Even detectors with high overall accuracy can yield *false positives* that jeopardise "innocent" authors. For instance, a study found that a state-of-the-art classifier from OpenAI still incorrectly labelled about 9% of human-written passages as AI (Nelson 2023). OpenAI's guidance conceded that such tools should not be relied on in isolation for important decisions (Kirchner et al. 2023). Additionally, LLMs themselves are continually improving: the text produced by GPT-4.5 is remarkably more fluid and harder to distinguish from human writing than that of GPT-3.5, which means detection will only get more challenging. Finally, the shame surrounding any use of LLMs in the writing process is leading people to hide rather than stop or disclose the practice and its extent (Zhang et al. 2025). It may even generate a "supply-side prohibition paradox".[1]

Overall, the current debate centres on how to maintain trust in the scholarly record in the face of generative AI. Should a paper be rejected if there is a suspicion it was machine-written, even if the content is of high quality? How can journals detect LLM assistance without unfairly undermining legitimate work or disadvantaging non-native English writers (Liang et al. 2023)? These questions have no easy answers yet. Luckily, in the rest of this article, I wish to focus on a different side of the general issue: the use of LLM-driven tools *by journals themselves* to support and improve the publication process. LLM technology has the potential to assist in managing and reviewing content (Leung et al. 2023) and in the rest of the article I explore how LLMs

[1] In economics this happens when legal prohibitions create black markets that operate outside regulatory frameworks.

might be embedded in editorial workflows not to catch cheaters *after the fact*, so to speak, but to help authors, editors, and reviewers, to get things right *from the start.* A final disclosure: what follows is based on my fifteen years of experience as Editor-in-Chief of *Philosophy & Technology*, but I hope my remarks and suggestions may be extendable to the field of publishing in general.

2. The Future of LLM-based Editorial Interfaces

Imagine submitting your manuscript to a journal and receiving, within minutes, an interactive report detailing any missing disclosures, irregular formatting, or sections that need clarification, all generated by an LLM. Shortly after, the handling editor sees a concise summary of your paper with an initial assessment of its writing quality and fit for the journal, courtesy of an LLM assistant. Such scenarios are very realistic. In the coming years, all main journals will likely incorporate LLM-based editorial interfaces (henceforth editorial LLMs or eLLMs) into their submission systems to automate and streamline the editorial screening process.

Leading manuscript management platforms have been progressing in this direction for quite some time. Editorial Manager (EM) offers a compelling illustration since thousands of journals use it, including *Philosophy & Technology*. In 2023, EM partnered with Paperpal, an LLM tool, to introduce 'Preflight' for authors (George 2023). This integration enables authors to run comprehensive automated checks on their manuscript before formal submission. Trained on millions of published papers, the machine learning system scrutinises text for linguistic issues, verifies compliance with journal formatting requirements, and identifies common omissions. Authors promptly receive a detailed report highlighting problems such as inconsistent citation styles, grammatical errors, excessively long abstracts, or missing funding acknowledgements. They can subsequently make necessary corrections before submission, thus enhancing their manuscript's prospects of clearing initial checks. After submission, the refined manuscript enters the standard editorial workflow in EM. This LLM-assisted pre-check effectively frontloads quality control, benefiting authors by minimising delays or desk rejections for trivial matters, while reducing less intellectually demanding work for editors and reviewers. If this sounds futuristic, it is worth noting that it is already a reality for more than 400 journals. Building on the

success of author-side Preflight, in 2025, the same technology was extended to the editorial desk side of EM (Aries Systems Corporation 2025). Upon submission, manuscripts can now be automatically analysed by an LLM-driven service (Paperpal's algorithms in this case) which evaluates three key aspects: (a) formal compliance, (b) writing quality, and (c) research integrity. The results are transmitted to the editor's interface as a detailed report. For example, the eLLM will check whether all required sections (abstract, keywords, references, etc.) are present and properly formatted, whether the language is fluent and academic in tone, and whether there are potential integrity red flags like plagiarism or suspect citations. In essence, the eLLM pre-assesses the submission. Papers with serious deficiencies can be flagged, while those that pass the automated checks can move faster to peer review. Importantly, this is done *without* slowing down the process, since the analysis is completed in the background shortly after submission, before an editor even opens the file. The eLLM highlights problematic sections for the editors, enabling them to concentrate their time on content evaluation rather than clerical checks. Such LLM-driven screening significantly speeds up decision-making at the initial review stage.

Other major submission systems are following suit. Clarivate's *ScholarOne*, another widely used platform, began experimenting with AI integrations as early as 2018 (Clarivate Analytics 2018). In a pilot project, Clarivate partnered with the Denmark-based AI firm UNSILO to embed manuscript evaluation algorithms into ScholarOne. The goal was to provide decision support for editors by screening incoming submissions automatically – identifying papers that fall outside the journal's scope, have poor language quality, or lack key elements – to "save millions of hours in peer review time" (Clarivate Analytics 2018). Over the next few years, this evolved from pilot to production. By 2020, UNSILO's AI-powered *Technical Checks* tool (now part of Cactus Communications) was fully integrated into ScholarOne, offering automated checks for manuscript completeness and adherence to guidelines (Razack et al. 2021). This tool can verify, for instance, that ethics statements or conflict-of-interest declarations are present where required, that all cited references are included in the bibliography, and even that figure file sizes meet the specifications. Much like the EM integration, the aim is to catch *formatting and policy issues* at submission, freeing up editors to focus on evaluating the research content itself.

Even newer and smaller platforms are embracing these innovations. Scholastica, a modern cloud-based journal management system popular with independent and society journals, is actively exploring features across its workflow. In an interview, Scholastica's co-founder Brian Cody noted that recent advances in LLMs have created a "paradigm shift" in how their team thinks about solving problems (Meadows 2024). Instead of manually building every validation rule or feature, they now ask, "is there a solution that AI could, or should, be part of?" (Meadows 2024). This mindset suggests that Scholastica will likely integrate LLM-based checks or assistants soon. Given Scholastica's emphasis on user-friendly design, one can imagine an AI assistant built into their submission form, guiding authors in real time or assisting editors by summarising each submission and checking it against the journal's scope. While specific features from Scholastica are not publicly announced, the company's leadership predicts that "increasingly capable AI tools" will become a regular part of editorial workflows, operating reliably in the background to handle complicated or repetitive tasks. As Cody describes, the focus is on using eLLMs to augment human editors rather than replace them, strengthening the infrastructure so that the community can better deal with new challenges, like the influx of submissions, or "paper mill" generated content.

The trajectory is clear: eLLMs are poised to become standard infrastructure across scholarly publishing. The result should be faster turnaround times and fewer administrative hurdles. As AI developments affect content production and editorial management (Floridi 2024), platforms are evolving from passive workflow websites to more interactive decision-support systems that eLLMs use to evaluate content pre- and post-submission. Clearly, integrating eLLMs into journal submission workflows brings both potential benefits – I have already highlighted some of them here – but also notable risks. Both are the focus of the next section.

3. Benefits and Risks of eLLMs

In this section, I will be schematic, but the points I list below are intertwined and interact with each other. Let me start with the benefits, since we have already seen some of them in the previous sections.

1) Faster screening and resource optimisation. We saw that an eLLM can handle the time-consuming initial review of submissions much faster than humans (I speak from experience), swiftly checking whether manuscripts meet basic requirements and flagging obvious problems. This process reduces editorial backlogs and considerably shortens the time needed to make desk decisions. By automating technical checks and weeding out non-compliant or low-quality submissions early, eLLMs save editors and reviewers from examining papers destined for rejection, allowing them to concentrate on more value-added tasks like author support, content curation, and substantive evaluation rather than policing formatting or grammar. Editorial staff and volunteers can focus their expertise where it matters most, with less concern about misconduct detection. In the long run, these efficiencies should contribute to better reviewing processes and faster publication times.

2) Improved compliance and quality. An eLLM can ensure that submissions adhere to journal guidelines for format and style. Editors and reviewers often feel they are doing the job that a machine could do better and more quickly. They are right. An eLLM can also mean fewer submissions get bounced back later to authors for trivial issues like missing keywords or improper reference format. Over time, as authors anticipate these checks, the average quality of submitted manuscripts may improve. The promise is a higher baseline quality of submissions – clearer writing, complete data, correct formatting – which makes peer review more effective and fairer.

3) Consistency and objectivity. Unlike a human editor, an eLLM can uniformly apply the journal's checklist to every submission. For instance, it will not overlook a missing funding statement due to fatigue. This consistency helps maintain editorial standards. It can also mitigate individual biases at the screening stage (but see below for the introduction of new biases): the eLLM evaluates based on the set criteria (language

clarity, presence of sections, etc.) without being influenced by author reputation or institutional prestige. This could lead to a more level playing field, where every paper gets the same initial scrutiny.

4) Enhanced fraud detection. Depending on the implementation and editorial strategies, eLLMs can be trained to spot specific forms of academic misconduct, including plagiarism, self-plagiarism, or "paper mill" outputs that humans might miss. For instance, they can cross-check citations against databases to see whether references are real, fabricated, relevant, or excessively self-referential, or scan the text for hallmarks of GPT-style composition. While detection is imperfect (as discussed earlier), coupling simple plagiarism checks (like iThenticate) with more nuanced LLM analysis could catch blatant cases of copied text or AI-generated gibberish *before* they enter peer review. This proactive filtering saves reviewers' time and efforts (Hosseini and Horbach 2023), strengthens research integrity and could prevent embarrassing later retractions. Even if the eLLM catches only the most egregious problems, e.g. nonsensical references, it is still a useful net to have in place.

5) Author empowerment. When authors are given access to the pre-check eLLM tools, they can empower themselves to improve their submissions. Rather than receiving negative feedback or even a rejection for technical reasons, sometimes after a long delay, authors get immediate feedback and can fix issues on their own. This makes the submission process less unpleasant and more fruitful. It is especially helpful for researchers with limited access to professional editing resources, as the LLM interface can act like a virtual editorial assistant, guiding them to meet the journal's expectations. In effect, it can democratise some aspects of the publishing process by providing all authors with a baseline level of editorial support.

6) Benefits throughout the editorial workflow. Beyond initial submission screening, eLLMs can enhance multiple stages of the publication process. They can assist in identifying appropriate reviewers by analysing the manuscript's content and matching it with potential reviewers' expertise while flagging possible conflicts of interest or previous collaborations between authors and reviewers. After reviews are submitted,

eLLMs can evaluate the quality and thoroughness of reviewer comments, ensuring they meet journal standards before being sent to authors. When revised manuscripts are resubmitted, eLLMs can systematically compare versions to verify whether and how satisfactorily authors have addressed reviewer concerns. In cases of contradictory reviewer assessments, an eLLM can provide information about points of agreement, helping editors arbitrate between conflicting recommendations. Finally, eLLMs can flag inconsistencies in the review process itself, such as when a paper receives positive initial reviews but is later recommended for rejection without clear justification, thus promoting fairness and transparency throughout the editorial journey.

These promises are compelling but come with significant risks that must be acknowledged and addressed. The following analysis provides a list of the most pressing.

1) False positives and false negatives. We have already encountered this problem. Pre-screening with AI content detectors requires acknowledging their inherent fallibility. These tools may incorrectly flag acceptable text as problematic (false positive) or miss genuine issues (false negative). Non-native English speakers producing well-written but straightforward papers are particularly vulnerable to being misidentified as using AI-generated content simply due to stylistic simplicity. When editors place excessive trust in eLLMs, they risk unjustly rejecting valid work. Simultaneously, sophisticated LLM-generated or plagiarised content may escape detection by avoiding known patterns. Automation reliance thus introduces both type I and II errors into editorial judgement. Even sophisticated detectors with impressive accuracy rates (80–90%) still misclassify significant numbers of submissions (Elkhatat, Elsaid, and Almeer 2023). A particularly concerning risk is that eLLMs may hallucinate non-existent issues in manuscripts, e.g., fabricating violations of journal guidelines that do not exist, inventing formatting requirements, or falsely claiming problems with references such as incorrect titles or volume numbers. Conversely, they might hallucinate compliance, erroneously confirming the presence of missing required elements like funding statements or ethics declarations. These phantom findings could trigger unwarranted rejections or falsely reassure editors about a manuscript's completeness. Human

editors must double-check any LLM-driven flags. Admittedly, this could reduce the efficiency gains, if not carefully managed, but "human intelligence inside" remains indispensable.

2) Bias and fairness issues. As noted, AI detectors have exhibited biases, particularly against authors writing in a second language. Suppose an eLLM language check is not calibrated correctly. In that case, it might systematically give lower clarity scores to manuscripts from specific linguistic communities simply because the writing style differs from the training data norm. This raises concerns about equity. eLLMs could introduce new biases whereby some groups' submissions face harsher automated scrutiny. Of course, this is not a novelty: human editors and reviewers have similar biases. But in this case, it could reach an industrial level, both in terms of quantity and in terms of scale (imagine the same eLLMs used by many journals all affected by the same biases), could acquire a status of pseudo-objectivity that would mask its negative nature ("if the LLM says so, it must be so"), and could further reinforce the already present, human biases, solidifying them. Ensuring that the eLLMs are trained on diverse data and regularly audited for bias is essential; otherwise, the "streamlining" could disproportionately harm some authors. It is also worth noting that eLLMs themselves can easily reproduce societal biases present in their training data, which might manifest in how they evaluate content. For instance, one could imagine an eLLM being more likely to flag content that does not match some culture-based research norms as low quality. Such pitfalls must be proactively guarded against. Beyond linguistic and cultural biases, eLLMs risk amplifying resource and skill disparities in academia, exacerbating the already severe digital divide. Authors from well-funded institutions with premium access to advanced LLM tools and training in effective prompting techniques will navigate eLLM-mediated submission systems more successfully than those without such privileges. Researchers at elite universities may receive institutional support specifically designed to optimise manuscripts for automated screening, while those from less-resourced institutions or regions struggle with the basic requirements. Although these inequities already exist in traditional publishing workflows, analogous to how article processing charges create barriers, algorithmic gatekeeping could further entrench and systematise these advantages,

making the playing field even less level. The scholarly community could collectively help by developing open-source models specifically for academic publishing needs (see below recommendation 5).

3) Opacity and trust. Many LLM tools operate as "black boxes, " providing conclusions without transparent reasoning. For authors and editors to trust an LLM's recommendation, they need some explanation. If a submission is rejected after AI screening with no apparent reason beyond "the system flagged it", this will breed resentment and confusion, besides running into all the risks already seen above. Editors might be unable to defend a decision based only on an opaque algorithm. Likewise, authors would be frustrated by vague feedback and decisions based solely on opaque algorithms. Lack of transparency can erode trust in the editorial process among the research community. This risk requires that any LLM integration include interpretable outputs or detailed justifications when it flags something. And just in case you thought this is mere theory, as an author, I have received such "boilerplate" feedback that was not only useless but also frustrating (Naddaf 2025).

4) Over-reliance and desk-rejection culture. If editors lean too heavily on eLLM screening, there is a risk of creating an overly hasty desk-rejection culture. Automated pre-submission assessment might encourage some editors to reject a paper without a thorough human read, simply because the eLLM gave an initial poor evaluation. This could especially affect borderline cases. For example, a truly novel paper that challenges conventional wisdom might be written in a way that the eLLM deems "low quality" or off-topic, leading to an early rejection that a more nuanced human reading might avoid. In other words, *innovation could suffer* if algorithms geared towards typical "good papers" patterns can veto the atypical ones. This would further reinforce an already problematic trend in human assessment, which privileges the safe and boring over the exciting but unsafe. Evaluations, both human and artificial, tend to favour the top of the Gaussian, because to avoid the bottom left (extremely low quality), they also tend to sacrifice the bottom right (extremely high quality). LLMs could easily reinforce such a trend. The risk here is substituting algorithmic consistency for human judgment in cases that warrant the latter. It will be important that any eLLM remains a support

tool – a first pair of eyes – rather than an arbiter of fate for submissions. Unfortunately, the history of brilliant papers rejected is embarrassing. Just *Nature* managed to reject nine submissions that led to Nobel prizes in Chemistry, Physics, and Physiology or Medicine, as well as revolutionary contributions in evolutionary biology and chaos theory (Nature 2003). It also retracted approximately fifty papers over the same period. The hope is that human editors supported by eLLMs might be better than either individually.

5) Gaming and new forms of misconduct. Once it is known that journals use eLLM, unscrupulous actors, e.g. paper mill operators or unethical authors, will adapt to game the system. We might see the emergence of eLLM-evasion techniques, such as using paraphrasing tools to trick detectors, a practice already observed (Liu et al. 2024), or intentionally inserting a few typos or human-like errors to avoid a "too perfect" LLM-written text. Similarly, if authors know what eLLMs flag as specific phrases or a lack of citations, they might stuff their text with superficial changes or references to appease the system without truly improving. This cat-and-mouse dynamic could lead to an arms race, with the journals having to update their eLLMs continuously. Paradoxically, it could also result in unintended consequences, such as more bloated or convoluted writing as authors try to "beat" the eLLMs.

6) Content homogenization and algorithmic optimisation. As authors become increasingly aware of which features eLLMs evaluate favourably, a more subtle risk emerges: the gradual homogenization of scholarly writing to satisfy algorithmic preferences rather than human readers. Academic writing is a distinct *genre* characterised by established rules, formal conventions, and disciplinary expectations. It often compels writers to adopt a particular style, typically objective, impersonal, and tightly structured, which may limit the expression of individual voice or alternative rhetorical approaches. While these constraints aim to promote clarity, rigour, and reproducibility, they can also marginalise diverse forms of knowledge, privileging specific modes of argumentation and epistemic authority over others. However, one may argue that such constraints are an acceptable and justified trade-off and that at least the academic community is aware of it. Instead, in a publish-or-perish, eLLM-

based environment, authors may feel pressured to optimise their manuscripts for eLLM approval, potentially prioritising stylistic patterns and structural elements that arbitrarily appeal to algorithms over innovative or field-challenging content, without any real justification apart from "algorithmic idiosyncrasy". This could lead to a form of "LLM-friendly" writing that lacks distinctive voice or creative expression and satisfies strange LLM-preferred features. Imagine, for example, the presence of "to delve" becoming a preferred feature.[2] Unlike deliberate gaming through deception, this subtler adaptation represents a potentially widespread shift in scholarly communication. Over time, we might see academic writing converge toward a standardised, algorithm-pleasing middle ground that systematically disadvantages breakthrough work presented in unconventional ways. This risk echoes concerns about search engine optimisation in other domains, where content creators prioritise algorithmic visibility over depth or originality. Scholarly literature could thus become more uniform, predictable and optimised for machine processing rather than advancing human knowledge.

7) Technical failures and integration challenges. As always with any technology, there is the risk of technical glitches. An eLLM integrated into a submission platform might occasionally malfunction – e.g., time out under heavy load, misreading a PDF with unusual formatting, or failing to parse math or tables correctly – incorrectly flagging things. If the editorial staff come to trust the system uncritically, such errors could slip through. Moreover, integrating cutting-edge eLLMs into legacy systems is complex: any system updates carry the risk of bugs that might temporarily disrupt the submission process. Journals will need robust IT support and contingency plans for when the LLM tool is down or misbehaving, to avoid grinding the workflow to a halt.

8) Privacy and ethical concerns. I shall return to this point in the conclusion. Here, let me stress that when manuscripts are analysed by third-party eLLMs, even if under contract, authors might worry about their unpublished work being exposed or used to train commercial models further. The content of a new submission is intellectual

[2] https://hesamsheikh.substack.com/p/why-does-chatgpt-use-delve-so-much

property that, if sent to an external API, could technically be stored or learned from by that system. Some detection tools have been criticised for possibly using submitted text to improve their algorithms.[3] This is a legitimate concern: feeding a manuscript into an eLLM, means an additional copy exists outside the journal's secure system. Without careful data agreements and perhaps on-premises solutions, journals risk losing the trust of authors who fear their work could leak or be exploited. Ensuring confidentiality and compliance with data protection norms is thus a non-negotiable aspect, and a severe risk if ignored.

In weighing these benefits and risks, it becomes clear that while eLLM integration can significantly enhance efficiency and quality in editorial workflows, it must be done thoughtfully and with safeguards. In the next section, I discuss how eLLMs may be designed to maximise trust and utility, harnessing the benefits while minimising, if not avoiding, the risks.

4. Designing eLLMs for Trust and Utility: Some Recommendations

eLLMs may gain acceptance simply because journals will impose them. It is not a welcome scenario, but I am afraid it is realistic. Journals already impose cumbersome and time-consuming procedures resented by any author (me included), it would not be surprising if they were to put in place more requirements. However, such tools could be implemented in a user-friendly, transparent, and responsible way. Preferably, authors, editors, and reviewers could appreciate these eLLMs as helpful rather than inscrutable gatekeepers. So here are some recommendations for designing eLLMs that could foster trust, improve acceptability, and deliver real utility.

1) Human oversight with integrated appeal mechanisms. An eLLM should function as a decision-support system, not an autonomous gatekeeper. To maintain accountability, eLLM-generated assessments must remain advisory, with final decisions—especially rejections—resting with human editors who use the eLLM's reports as one factor among many. Various organisations have emphasised this principle; OpenAI's

[3] https://citl.news.niu.edu/2024/12/12/ai-detectors-an-ethical-minefield/

guidance, for instance, explicitly stated that text classifier results should serve only as supplemental information (Kirchner et al. 2023). A robust implementation should include both preventative and corrective human oversight. Preventatively, systems should require human editors to review and confirm any automated 'reject' flags before action is taken. Correctively, platforms should incorporate straightforward mechanisms for authors to contest eLLM-generated warnings—for example, allowing authors who believe their paper was falsely flagged as AI-generated to provide justification or request manual review through an intuitive interface (perhaps a 'request editorial review' button next to flagged items). These dual oversight mechanisms ensure no submission is rejected solely by an algorithm and acknowledge the fallible nature of current technologies. Editors must retain the ability to override eLLM suggestions through the interface. The very existence of this human-centred failsafe system makes the entire process fairer and more transparent, establishing a safety net that bolsters trust while delineating the eLLM's limited role as a support tool rather than an arbiter. Over time, as technologies improve, the need for appeals should diminish, but their availability remains essential to prevent algorithmic control in academic gatekeeping.

2) Transparency of operations. The eLLM should provide clear explanations for its findings. If the tool flags a section of a manuscript, it should indicate why – e.g. "the references section may be incomplete (citation [15] not listed)" or "unusual phrasing detected; resembles AI-generated text". Providing these details helps authors, editors, and reviewers understand the basis of the eLLM's evaluation. It converts a black-box verdict into actionable feedback. Some of this can be achieved through simple rules, like highlighting a missing element, while more complex evaluations, like "scope mismatch", might include a summary of what the eLLM has identified as the paper's topic compared to the journal's scope. The key is that the interface should not present a mystery score or a red light with no context. User trust grows when the system is interpretable. In practice, developers of these tools should incorporate features such as highlighting problematic sentences, listing specific guideline violations, or giving confidence levels for their predictions. An example to emulate is how plagiarism

checkers present a similarity report with sources; an eLLM content check could analogously show snippets that triggered suspicion.

3) Author-friendly and informative feedback. To be broadly accepted, eLLMs' checks should be framed as helpful feedback in tone and content rather than policing. The interface might, for instance, present issues in two categories: "errors to fix before acceptance" and "suggestions for improvement". This makes it clear that the goal is to help the author succeed. The design should avoid language that feels accusatory or absolute. Many authors appreciate direct guidance, but it should be delivered diplomatically. Providing links to resources (such as the journal's style guide or relevant articles on scientific writing) could increase utility. Essentially, the eLLM should mimic a diligent editorial assistant: thorough but supportive. Early experiments have shown that when authors receive eLLM-generated feedback before submission, they often incorporate the suggestions and resubmit successfully (George 2023). That outcome should be the aim.

4) Bias mitigation and fairness. Developers must actively work to identify and correct biases in eLLMs' outputs. This could involve using diverse training data, including manuscripts from different regions, disciplines, and writing styles. It also requires some critical vigilance. Periodic audits using test submissions – for example, a set of papers by native and non-native English writers – can reveal whether the eLLM is disproportionately flagging one group. If bias is found, recalibration is needed, perhaps by adjusting sensitivity or adding rules, e.g., do not penalise minor grammar quirks in otherwise sound text. Additionally, the interface could be designed to flag its uncertainty. If the eLLM is, say, only 60% confident a text is AI-written, it might display a caution like "possibly AI-generated content – review recommended" rather than a definitive statement. This communicates nuance to the editor, who can then make a more informed judgment. It also prevents borderline cases from being treated too harshly. In general, erring on the side of not rejecting (to avoid false positives) is a safer default in design, with the understanding that the human editor can still act on obvious problems.

5) Data security and privacy. To address the authors' concerns, the system's handling of manuscript data must be transparent and secure. Ideally, eLLM analyses should be performed locally or in a secure cloud where the data is not used to train unrelated models without consent. If a third-party service is used, there should be an explicit agreement that submitted content will be kept confidential and not retained beyond the analysis, or only retained in anonymised form for tool improvement, if necessary and with permission. The interface should include a notice to authors to this effect. Journals may also allow an opt-out for particularly sensitive submissions, even though opting out might slow down processing. One promising approach to mitigate these privacy concerns is for publishers to adopt open-source LLMs that can be deployed locally within their secure infrastructure. Open-source models like Llama, Falcon, or Mistral offer the advantage of more transparency in their architecture and operation while allowing higher degrees of data control. Publishers could fine-tune these models on domain-specific scholarly content without sharing sensitive unpublished manuscripts with commercial vendors. This approach would reduce dependence on proprietary black-box systems, potentially lowering costs while addressing confidentiality concerns. Additionally, the scholarly community could collectively contribute to improving these open models specifically for academic publishing needs. Building trust on this front is crucial; otherwise, authors might avoid journals that use eLLMs, especially in fields with competition or intellectual property concerns. In designing these systems, following best practices of data protection (compliance with GDPR, etc.), both legally sound and ethically preferable, is crucial.

6) Continuous learning and Human-AI collaboration. Designing for utility means allowing the eLLM to learn from human corrections. If editors consistently override specific flags as false alarms, the system should adjust its parameters or at least notify its developers. More generally, eLLMs should incorporate feedback loops. Additionally, the interface could enable editors to provide simple feedback on the eLLM's output. Such input can guide iterative improvements. The ultimate vision is a collaborative intelligent approach: the eLLM does the tedious work and highlights issues, the human makes the decision and provides feedback, and the eLLM uses that feedback to improve over time. This collaboration can be highlighted in training and

documentation so that users approach it with the right mindset, seeing the eLLM as a tool that can improve over time. When users feel they have control and the ability to improve the system, they are more likely to embrace it.

7) User-centred design and testing. Finally, these eLLMs should be developed with extensive input from the end-users: authors, editors, and reviewers. Early beta testing with diverse user groups can uncover usability issues. For instance, editors might say the report is too long and they only want an executive summary, leading to a redesign of how information is presented. Or authors might report feeling overwhelmed by a barrage of eLLM's suggestions, indicating the need to prioritise the most critical ones. By applying user-centred design principles, the interface can be made intuitive. Simple things like integrating it seamlessly into existing submission steps and ensuring it works reliably across devices will affect adoption. The technology might be cutting-edge, but it must meet users where they are, in terms of workflow and comfort level. A phased rollout would help build trust gradually. Journals might begin by using eLLMs optionally, for a trial period, sharing feedback with authors manually, before fully automating the process.

To summarise, designing eLLMs for trust and utility requires a blend of technical safeguards and thoughtful user experience design. When done right, the eLLMs support the publication process: always available, consistent, and beneficial to all stakeholders. Authors should feel that using the system improves their chances of a fair and timely assessment, and editors and reviewers should feel that it lightens their load without stealing their authority. Achieving this balance is challenging but feasible, as evidenced by some publishers' careful approaches.

5. Conclusion: the opportunities and risks of DIY

LLMs are swiftly becoming woven into the fabric of scholarly publishing, not merely as writing assistants but as powerful tools enhancing editorial workflows. Soon, submitting papers to journals will likely routinely involve eLLM-driven assessments of compliance, quality and integrity before human editors even begin their review. This development promises faster decisions, fairer screening and better-prepared

manuscripts, provided that crucial issues of trust, bias and transparency are thoughtfully addressed. Major platforms are already establishing the foundation, and these practices will likely become standard across the industry within the next few years. Eventually, researchers may find it difficult to imagine how scholarly publishing functioned without AI support. Nevertheless, at present, relatively few journals have implemented such systems. Many editors continue to rely exclusively on manual checks, whilst authors submit manuscripts without benefiting from automated pre-review processes. Until LLM-powered interfaces are widespread, authors can take proactive steps to emulate the "pre-check" that an LLM might perform. In fact, using the publicly available LLMs (like ChatGPT or Anthropic's Claude) as a personal submission assistant is an increasingly popular strategy. By leveraging these tools, authors can catch errors and improve their manuscripts before clicking the submit button. A good prompt can guide the LLM to act as a journal submission checker, performing tasks analogous to what I have described in this article. While current public LLMs may not have access to all the journal-specific rules or be able to verify references against databases, they can analyse text for clarity, structure, logical flow, and even for consistency with typical academic writing conventions. They can simulate the role of an editorial assistant. Of course, all the previous benefits and risks apply. Authors should use judgment and not treat LLM feedback as the last word. LLMs might "hallucinate" a problem that is not there, or give generic advice. But even then, the exercise can help review the work from an alternative perspective, and confront areas for improvement one might have overlooked. Here is a final list of warnings:

- Data retention: LLM providers usually store your inputs as part of their training and improvement processes. According to their data retention policies, these conversations may be retained for some time.
- Usage rights: when you input text, you typically grant the service provider a license to use that content for purposes like improving their models, but you generally retain ownership of your original content.
- Privacy considerations: the provider may have contractual obligations to maintain the confidentiality of your inputs, though there are usually exceptions for legal requirements.

- Training data: your inputs might be used to train future versions of the model unless you opt out, if that option is available.
- Human review: human reviewers may review some inputs for quality control and model improvement.

The specific legal details are covered in the provider's terms of service, which users agree to when using the service. Different providers and subscription tiers may have varying policies regarding data retention and usage. If you are concerned about the confidentiality of a text, you should review the privacy policy and terms of service of the specific LLM service you are using, as these legal frameworks govern exactly how your inputs are stored, used, and protected.

In closing, the relationship between LLMs and academic writing is often portrayed adversarially, with AI generically described as a threat to academic integrity. In this article, I highlighted a constructive path: eLLMs as supporting tools that can uphold and even elevate the quality of scholarly communications. Authors, editors and reviewers could all gain if the technology is implemented with care, and fully aligned with the values of academia: innovativeness, rigour, fairness, and openness. Until then, we as authors do not have to wait on the sidelines. We can experiment with non-specialised LLMs as part of our own writing and revision process.

Acknowledgements

Many thanks to Jessica Morley and Claudio Novelli for their insightful comments on a previous version of this article. They improve it significantly, as always.

References

Aries Systems Corporation. 2025. "Aries Systems and Cactus Communications Expand Partnership to Improve Research Quality with Automated Checks." *Report* https://www.ariessys.com/newsletter/march-2025/aries-and-cactus-expand-partnership-to-improve-research-quality-via-automated-checks/.

Carobene, Anna, Andrea Padoan, Federico Cabitza, Giuseppe Banfi, and Mario Plebani. 2024. "Rising Adoption of Artificial Intelligence in Scientific Publishing: Evaluating the Role, Risks, and Ethical Implications in Paper Drafting and Review Process." *Clinical Chemistry and Laboratory Medicine*

(CCLM) 62 (5):835-843.

Clarivate Analytics. 2018. "Clarivate Analytics and Unsilo Partner to Power Scholarone with Ai." *Report* https://www.stm-publishing.com/clarivate-analytics-and-unsilo-partner-to-power-scholarone-with-ai/.

COPE. 2023. "Committee on Publication Ethics - Authorship and AI Tools." https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools.

Crotty, David. 2024. "The Latest 'Crisis': Is the Research Literature Overrun with Chatgpt- and LLM-Generated Articles?" https://scholarlykitchen.sspnet.org/2024/03/20/the-latest-crisis-is-the-research-literature-overrun-with-chatgpt-and-llm-generated-articles/.

Dwivedi, Yogesh K, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, and Manju Ahuja. 2023. "Opinion Paper: "So What If Chatgpt Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy." *International journal of information management* 71:102642.

Elkhatat, Ahmed M, Khaled Elsaid, and Saeed Almeer. 2023. "Evaluating the Efficacy of Ai Content Detection Tools in Differentiating between Human and Ai-Generated Text." *International Journal for Educational Integrity* 19 (1):17.

Floridi, Luciano. 2024. "On the Future of Content in the Age of Artificial Intelligence: Some Implications and Directions." *Philosophy & Technology* 37 (3):112.

———. 2025. "Distant Writing: Literary Production in the Age of Artificial Intelligence." *SSRN* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5232088.

George, Elizabeth O. 2023. "Paperpal Partners with Aries Systems to Simplify Research Writing and Editing." *Paperpal* https://paperpal.com/blog/news-updates/press-releases/paperpal-partners-with-aries-systems-to-simplify-research-writing-and-editing.

Hosseini, Mohammad, and Serge PJM Horbach. 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 8 (1):4.

Imran, Muhammad, and Norah Almusharraf. 2023. "Analyzing the Role of Chatgpt as a Writing Assistant at Higher Education Level: A Systematic Review of the Literature." *Contemporary Educational Technology* 15 (4):ep464.

Katsnelson, Alla. 2022. "Poor English Skills? There's an AI for That." *Nature* 609.

Kirchner, Jan Hendrik, Lia Ahmad, Scott Aaronson, and Jan Leike. 2023. "New Ai Classifier for Indicating Ai-Written Text." *OpenAI Blog* https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/.

Leung, Tiffany I., Taiane de Azevedo Cardoso, Amaryllis Mavragani, and Gunther Eysenbach. 2023. "Best Practices for Using AI Tools as an Author, Peer Reviewer, or Editor." *Journal of Medical Internet Research* 25:e10525.

Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. "GPT Detectors Are Biased against Non-Native English Writers." *Patterns* 4 (7):100779.

Liu, Jae Q. J., Kelvin T. K. Hui, Fadi Al Zoubi, Zing Z. X. Zhou, Curtis C. H. Yu, Jeremy R. Chang, and Arnold Y. L. Wong. 2024. "The Great Detectives: Humans Versus AI Detectors in Catching Large Language Model-Generated Medical Writing." *International Journal for Educational Integrity* 20:8.

Májovský, Martin, Martin Černý, Matěj Kasal, Michal Komarc, and Dušan Netuka. 2023. "Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened." *Journal of Medical Internet Research* 25:e46924.

Meadows, Alice. 2024. "Kitchen Essentials: An Interview with Brian Cody of Scholastica." *The Scholarly Kitchen* https://scholarlykitchen.sspnet.org/2024/08/21/kitchen-essentials-an-interview-with-brian-cody-of-scholastica/.

Naddaf, Miryam. 2025. "AI Is Transforming Peer Review—and Many Scientists Are Worried." *Nature* 639 (8056):852-854.

Nature. 2003. "Coping with Peer Rejection." *Nature* 425 (6959):645-645.

———. 2023. "Tools Such as Chatgpt Threaten Transparent Science; Here Are Our Ground Rules for Their Use." *Nature* 613 (7945):612.

Nelson, J. 2023. "Openai Quietly Shuts Down Its AI Detection Tool." *Decrypt*

https://finance.yahoo.com/news/openai-quietly-shuts-down-ai-190818632.html:2023.

Razack, Habeeb Ibrahim Abdul, Sam T Mathew, Fathinul Fikri Ahmad Saad, and Saleh A Alqahtani. 2021. "Artificial Intelligence-Assisted Tools for Redefining the Communication Landscape of the Scholarly World." *Science Editing* 8 (2):134-144.

Reinhart, Alex, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. "Do Llms Write Like Humans? Variation in Grammatical and Rhetorical Styles." *Proceedings of the National Academy of Sciences* 122 (8):e2422455122.

Stokel-Walker, Chris. 2023. "Chatgpt Listed as Author on Research Papers: Many Scientists Disapprove." *Nature News*:620-621.

———. 2024. "AI Chatbots Have Thoroughly Infiltrated Scientific Publishing." *Scientific American*.

Tang, Arthur, Kin-Kit Li, Kin On Kwok, Liujiao Cao, Stanley Luong, and Wilson Tam. 2024. "The Importance of Transparency: Declaring the Use of Generative Artificial Intelligence (AI) in Academic Writing." *Journal of Nursing Scholarship* 56 (2):314-318.

Wu, Junchao, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. "A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions." *Computational Linguistics*:1-66.

Zhang, Zhiping, Chenxinran Shen, Bingsheng Yao, Dakuo Wang, and Tianshi Li. 2025. "Secret Use of Large Language Model (LLM)." *Proceedings of the ACM on Human-Computer Interaction* 9 (2):1-26.