

Check and Trust With PAU: A Paper Analysis Utility for Automated Bibliographic Verification

Pauline Speckmann¹, Raphael Fischer², Christian Janiesch¹

¹TU Dortmund University, 44227 Dortmund, Germany

²Johannes Gutenberg University Mainz, 55122 Mainz, Germany

pauline.speckmann@tu-dortmund.de, rafische@uni-mainz.de, christian.janiesch@tu-dortmund.de

Abstract

The correctness and appropriateness of scientific references is paramount for upholding the trust in and ethics of science. This standard has come under pressure with the advent of artificial intelligence (AI) and Large Language Models (LLMs), which facilitate academic writing and peer-reviewing. Leading venues have become flooded with AI-enhanced manuscripts containing hallucinated references, endangering the general ethics and high standards of scientific practice. As researchers are under constant time pressure and cannot be expected to identify “HalluCitations” manually, automated bibliographic verification (ABV) becomes integral. Realizing the potentials of ABV, we introduce PAU (Paper Analysis Utility), an open-source tool for automating the validation of bibliographic information in scientific manuscripts. PAU extracts references from a given publication and compares them against major scholarly databases, while leaving the final judgment and thus agency to human reviewers. With hybrid verification, the tool detects and flags reference discrepancies both in terms of metadata validity (checked via pattern matching) and semantic alignment (leveraging LLMs). Both the design and performance of PAU are evaluated against researcher expectations and on real-world reference hallucinations. Our work contributes a socio-technical approach for establishing ABV in the research community to ensure scientific integrity and ethics in the age of AI-assisted academic publishing.

Code — <https://github.com/DiePaupi/PaperAnalysisUtility>

1 Introduction

Trust forms a fundamental pillar of scientific practice (Iordanou, Antoniou, and De Vos 2026). The credibility of research depends not only on the rigor of methods and the validity of results, but also on the transparency and traceability of the knowledge claims made within scientific publications (Haven et al. 2022). However, science is increasingly facing a broader crisis of trust, shaped by misinformation, epistemic polarization, and declining confidence in institutions (van der Meer and Hameleers 2026). At the same time, the rapid evolution and integration of generative artificial intelligence (AI) into academic workflows introduces

new opportunities as well as new risks for maintaining research integrity (Arar et al. 2025). Large language models (LLMs) have enabled researchers to automate or assist in many aspects of the scientific work process, including drafting text, summarizing literature, and suggesting references (Eger et al. 2026). While these capabilities can significantly improve productivity, they also introduce well-documented challenges with respect to factual accuracy, ethics, and trustworthiness (Cheng, Calhoun, and Reedy 2025).

In particular, LLMs are prone to generating so-called hallucinations, corresponding to content that appears plausible but is factually incorrect or unsupported (Ji et al. 2023). In the context of scientific writing, this problem frequently manifests in the form of fully hallucinated or inaccurate references, so-called “HalluCitations”, as already observed at major scientific conferences (Sakai, Kamigaito, and Watanabe 2026). Empirical evaluations indicate that this issue persists through model generations. For example, while GPT-4 improves upon GPT-3.5, it was still found to completely fabricate 18% of references and to additionally produce substantive citation errors in 24% of references (?). Inconsistencies in bibliographies undermine the traceability and credibility of scientific claims and pose serious challenges for authors and readers, who rely on accurate references to evaluate and contextualize research.

Whether it be reviewers volunteering to check the integrity of submitted papers, supervisors evaluating student theses, or authors preparing manuscripts for submission: researchers face substantial time pressure and may easily overlook inconsistencies in references or citations, especially when lacking detailed familiarity with the topics, for example in interdisciplinary works. Reviewing in particular is unpaid community service relying on voluntary efforts under limited time, and as such is incompatible with the manual detection of AI (mis)use and hallucination artifacts. First scientific repositories, associations, and publishers have started to address the complacent use of generative AI in academic publishing and in particular citations. While several venues, such as AIES itself, rely on self-declaration of AI use, efforts are under way to sanction ethical misconduct and enforce human oversight and accountability. For example, *arXiv* has recently announced their intent to penalize such scientific misconduct by banning authors from posting further pre-

prints (Singh Chawla 2026). With more large scale reports of detected hallucinations in scientific manuscripts being published over the last months (e.g., Topaz et al. 2026) it can be expected that more organizations will follow arXiv’s precedent.

We argue that in the age of AI-assisted publishing, *automated bibliographic verification* (ABV), that is ensuring the correctness, traceability, and contextual validity of references via (semi-)automated tools, becomes imperative for upholding the trust in and ethics of science. First approaches, tools, and services for detecting HalluCitations have recently emerged (e.g., GPTZero 2026; Stringhini and Blackburn 2026; Shi et al. 2026; Li, Lin, and Ma 2026; Zhao et al. 2026), however they are either of proprietary nature, not (yet) published at peer-reviewed venues, or lack practical evaluation with human feedback, thus limiting transparency, accessibility, and trustworthiness. We address the need for transparent and affordable tools and investigations via three distinct but interconnected contributions that are summarized in Figure 1 and reflect the structure of our work:

1. We approach the ABV problem from a holistic perspective and present a *socio-technical literature background* that connects the ethics and trustworthiness of science with the pros and cons of using generative AI in research.
2. We introduce the Paper Analysis Utility (PAU), an open source tool that was developed via requirement-driven design science methods and allows users to tackle ABV with a hybrid verification approach.
3. We contribute practical insights for both the *perceived usefulness* as well as the *empirical performance* of PAU when performing ABV.

Our literature review highlights how ABV is becoming central for ensuring research integrity, producing verifiable artifacts whose importance for scientific argumentation would be otherwise diminished by hallucinations (see Section 2). Drawing on this related work, informal discussions with researchers in our department, and our analysis of the ABV task, we formulate three primary requirements for ABV: (1) verifying whether cited publications actually exist; (2) detecting structural inconsistencies such as references that are not actually cited within the manuscript; and (3) assessing whether statements that cite a publication thematically align with the original work. Based on these requirements and our method (see Section 3) PAU was developed as an open source, modular, and extensible Python software that automates the paper analyses (see Section 4).

Not to be confused with the complex goal of formally verifying AI models themselves (Sälzer 2025), our tool transparently validates the integrity of scientific papers by combining traditional pattern matching with database lookup techniques and LLM capabilities for performing metadata and semantic alignment checks. Notably, we deem this preferable to closed source or AI-only approaches because deterministic checks against databases are more reliable and reproducible, while AI-based functions should only be used as assistive signals rather than trusted decisions. In our proof-of-concept workshop, practitioners awarded high

practical value to PAU and provided helpful input for its future refinement and in subsequent experiments on real-world papers, PAU did not produce false negatives (see Section 5). Overall, our work is not only unique in connecting the timely ABV problem to ethical practices and the crisis of trust in science, but it also equips researchers with an actionable toolkit as well as empirical insights to uphold academic integrity in the age of AI-assisted academic publishing (see Section 6).

2 Background

Trust in Science

Science and research face a crisis of trust, which is characterized by the “simultaneous erosion of confidence in multiple institutions and a destabilized information ecosystem marked by fears of misinformation and general pervasive unreliability” (van der Meer and Hameleers 2026, p. 2). As the public cannot easily validate scientific procedures and publications, it has no alternative but to place trust in researchers and their methods and practices, even though this trust “can be easily eroded and destabilised [sic]” (Zeybek Kabakci et al. 2025, p. 3). With spreading de-legitimizing narratives and rising epistemic indifference — for example, observed in the context of pandemics, climate change, and emerging technologies such as AI (van der Meer and Hameleers 2026) — it therefore remains imperative to uphold academic integrity and legitimacy. Importantly, the matter of trust also directly connects to the potentials of open science practices (Iordanou, Antoniou, and De Vos 2026), which aim at making research procedures and outcomes more transparent and accessible (Haven et al. 2022). It is primarily fueled by the observation that “epistemic practices that support transparency” are commonly valued as being trustworthy (Iordanou, Antoniou, and De Vos 2026, p. 21), whereas negative reputation is rooted in obscurity and lack of replicability.

For strengthening trust in science, Haven et al. (2022) accordingly conclude that open science approaches need to be thoroughly linked to the academic backgrounds on research integrity. It has various important dimensions, such as the credibility of research publications as well as the methodological rigor in practices and tools used in academia.

Furthermore, as also discussed for the *Connected and Open Research Ethics Commons*, participatory design approaches are fundamental to ensure integrity (Harlow et al. 2020). Similarly, practices and tools for supporting research integrity should also be developed under careful attention to their (participatory) design process (Spinuzzi 2005). The overarching field of design science research (DSR) (Vaishnavi and Kuechler 2015) offers methods and strategies for involving future users and stakeholders early and iteratively in the development process, to make sure that systems address real-world needs and remain transparent and trustworthy in practice. In the context of scientific infrastructures, such collaborative requirements discussions can help align technical solutions with established norms of academic integrity and the practical workflows and expectations of researchers, reviewers, and supervisors.

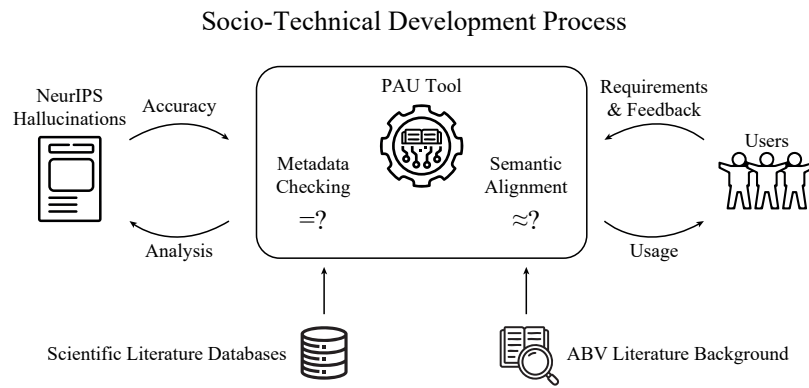


Figure 1: Schematic visualization of PAU’s development to automatically verify scientific bibliographies.

Using AI Models for Science

Over the last years, a technological evolution in the form of widely available and easy-to-use AI services has further destabilized the trustworthiness of science (Stokel-Walker and Van Noorden 2023; Arar et al. 2025). Despite many decades of prior work on natural language processing (Jurafsky and Martin 2026), the power of generative deep learning (LeCun, Bengio, and Hinton 2015), attention and transformer mechanisms (Vaswani et al. 2017), and self-supervised learning (Brown et al. 2020) has hit the scientific community with an unexpected force. Researchers actively explore the technical feasibility and limitations of automating multi-discipline cognitive reasoning (Huang et al. 2024b), generalizable abstraction and reasoning (Chollet et al. 2025; Fischer et al. 2020), or scientific discovery (Zhang et al. 2024; Zheng et al. 2025) with AI. While some know no borders when employing these new technologies, others are more cautious and demand researchers “to be keenly aware of principles of academic integrity, in interpreting and communicating about AI research outcomes” (Altmeyer et al. 2024, p. 1). Be that as it may, researchers have unprecedented means for automating and streamlining their work via multi-modal cloud-based services (Feuerriegel et al. 2024; Lins et al. 2021).

Whether in the form of proprietary services like ChatGPT or open-weight LLMs that can be deployed in custom environments, researchers across all disciplines leverage AI to improve their manuscripts or review others work in fully- or semi-automated workflows (Lund et al. 2023; Kim 2023; He et al. 2023; Kim 2024; Eger et al. 2026). Specific opportunities are, for example, seen for generating scientific figures (Belouadi, Ponzetto, and Eger 2024; Belouadi, Lauscher, and Eger 2024), finding and integrating existing publications (Babaei Giglou, D’Souza, and Auer 2025; Scherbakov et al. 2025), alleviating the burdens of academic peer-reviewing (Sun 2025), or even for the end-to-end automation of AI research (Lu et al. 2026). This is not the first example where technology automates research; computers have enabled researchers to calculate statistics more easily, while web searches and bibliography tools allowed users to scale systematic literature reviews. However, compared to the efficiency gained from these historic ad-

vances, modern AI services “have the ability to produce novel written content on their own” (Cheng, Calhoun, and Reedy 2025, p. 2), which raises serious technical and ethical questions with respect to verification, plagiarism, false information, and questionable legitimacy (Kim 2023; Stokel-Walker and Van Noorden 2023).

Agnostic to their use case context, LLMs are generally considered to lack safety and trustworthiness, mostly due to their extreme parametric complexity (Huang et al. 2024a). When used for intellectual tasks, it has therefore been argued that “significant social costs and disruptions could arise if measures are not promptly put in place to verify the file version, authorship, and authenticity of the generated content” (Kim 2023, p. 7). However, while different approaches for formally verifying deep neural networks exist (Xie, Kersting, and Neider 2022; Sälzer et al. 2026; Zaid, Neider, and Yalçın 2026), they remain non-trivial and computationally infeasible to apply to large-scale transformer architectures, as used in today’s LLMs (Sälzer 2025; Huang et al. 2024a). Deeply rooted in their probabilistic nature (Ji et al. 2023), generative hallucinations, for example, also complicate the automation of disinformation detection (Schmitt et al. 2024). As such, it comes as no surprise that various studies have evidenced artifacts of AI-assisted writing in the form of HalluCitations in contemporary publications (Agrawal et al. 2024).

Despite the technical issues and acknowledged unethicity, AI-assisted practices are becoming the new standard in academic work (Huang et al. 2024a; Raman 2025). Strong evidence for thousands of AI-generated or -enhanced papers and reviews was, for example, observed at leading AI venues such as ICLR (Naddaf 2025; Esau, Shmatko, and Adam 2025), ICDM (Kamath 2026; Gibney 2026), ACL (Sakai, Kamigaito, and Watanabe 2026), ACM CCS (Cortier and Lin 2026), and NeurIPS (Shmatko, Adam, and Esau 2026). While researchers have made efforts to develop specialized LLMs to decrease hallucinations and increase citation accuracy (e.g., Asai et al. 2026), these advances do not fully address the problem, particularly if such models are not widely adopted.

To uphold research integrity (Arar et al. 2025), researchers demand fellow colleagues to “always crosscheck

any information [and] ensure the ethical and transparent use of [AI models]” (Dergaa et al. 2023, p. 7), for which they also expect and recommend “guidelines for the responsible use of AI tools” (Kim 2024, p. 6), disclosure policies (Hosseini, Rasmussen, and Resnik 2024), as well as “integrity standards and practices” (Arar et al. 2025, p. 19). Responsible AI use in science depends not only on individual verification and disclosure, but also on scholarly AI infrastructures designed to support inspection, understanding, and user agency (Speckmann, Nadj, and Janiesch 2025). As scientific publishers commonly agree to “establish and support high standards of ethical conduct and reporting of academic research” (Cheng, Calhoun, and Reedy 2025, p. 2), official guidelines have early on been established by the Committee on Publication Ethics (Committee on Publication Ethics 2023) and similar stances are taken by top publishers (Nature 2023; Kim 2024). To summarize, one must thus carefully weigh the “promises and perils” (He et al. 2023, p. 1) of using LLMs for scientific writing and peer-reviewing (Giray 2024) and find appropriate operating principles.

Existing Approaches on Automated Bibliographic Verification

In order of maintaining scientific integrity, the science community urgently requires support in the form of automated verification of manuscripts, thus resulting in the ABV problem addressed with this work. Note that bibliography errors do not necessarily stem from AI usage and could also be rooted in sloppy writing, citation copying, outdated references, or formatting errors, yet such issues are likely amplified by AI models. Moreover, ABV is not necessarily restricted to unveiling reference errors, because respective analyses can, for example, also explore the quality of academic foundations (based on scientific venue rankings) or semantic matching (whether references actually support the surrounding statements). Tackling ABV is needed and non-trivial because the manual evaluation of references remains time-consuming and its automation, as well as the fundamental parsing of documents, is a somewhat complex problem in itself (Zhang et al. 2026). As such, whether it be for validating works of others (e.g., in peer-reviewing or literature reviews) or for finalizing their own papers, reliable automation will become imperative for manifesting scientific integrity and adhering to AI-related publication guidelines and expectations (Committee on Publication Ethics 2023).

While some suggest that LLMs can effectively spot hallucinations (Agrawal et al. 2024), we argue that an AI-only approach to ABV remains questionable, as it merely shifts the problem of hallucinations and credibility. Moreover, for effectively promoting trust in science (Iordanou, Antoniou, and De Vos 2026), respective methodological frameworks should be established as open, affordable, and transparent as possible, with respect to both the final application as well as the logic within the system. The former should ideally be made available in the form of an open source software toolkit that comes with rich documentation and can be easily used by both technical and non-technical practitioners. With regards to internal logic, instead of using black-box LLMs, a more interpretable and reliable approach could, for

example, leverage string matching and semantic similarity (like cosine similarity on word, sentence, or abstract embeddings) to compare references against a database.

While we worked on our own approach to ABV, others have started similar endeavors that we want to quickly shout-out. Shi et al. (2026) proposed *CiteAudit*, Li, Lin, and Ma (2026) presented *CiteTracer*, and Stringhini and Blackburn (2026) introduced *Hallucinator*, while Zhao et al. (2026) have performed a large-scale exploration of *HalluCitations* with a non-disclosed detection pipeline (Sakai, Kamigaito, and Watanabe 2026). Table 1 summarizes these related works to provide a brief overview of currently available open source ABV systems (including PAU).

In addition, there are two proprietary ABV tools that we know of: the *Citation Checker* (<https://citely.ai/>) and the *Hallucination Detector* (<https://gptzero.me/hallucination-detector/>). The latter was recently used for analyzing ICLR (Esau, Shmatko, and Adam 2025) and NeurIPS (Shmatko, Adam, and Esau 2026) papers. While providing an “in-depth analysis highlighting specific sentences likely to contain hallucinated or incorrect citations” (GPTZero 2026), this system can hardly be considered to be transparent or freely available. There is no information on its internal workings, payed plans are needed for processing more than 10,000 words, and the company has been called out for false positives in AI detection (Fowler 2023), which limits practical usefulness and trustworthiness. With PAU, we want to bridge this gap and offer fellow researchers an open and extensible toolkit for ABV that was also empirically evaluated with respect to detection rates and human feedback.

3 Method

To guide the development of PAU, we followed the principles of DSR which is well-suited for developing and evaluating socio-technical artifacts that address practical problems (Vaishnavi and Kuechler 2015). In contrast to theory-heavy DSR, our approach focuses pragmatically on iterative artifact development and evaluation in a real research context, as schematically summarized in Figure 1. Inspired by the phases proposed by Vaishnavi and Kuechler (2015), we began with a *problem awareness* stage which explored the challenges of verifying scientific references through informal requirement discussions within our department. In parallel, we conducted an exploratory review of existing ABV literature, synthesizing works on reference verification, hallucinated citations, and AI-assisted scientific writing.

Based on this initial understanding of ABV and related themes, we formulated initial requirements and features PAU should facilitate during the *suggestion* phase. We proceeded with the *development* by implementing PAU as a modular prototype in Python (see next section).

After establishing a proof-of-concept, we conducted a qualitative *evaluation* to gain initial insights into its practical usefulness. To this end, we organized a workshop with four PhD students and one senior researcher, in which PAU was demonstrated and subsequently discussed in terms of potential use cases and improvements. Finally, we carried

System	AI/LLM Use	Output
CiteAudit	Necessary	A report listing each reference’s real/fake verdict, optional ground-truth match, and verification details such as method, found status, and evidence note
CiteTracer	As fallback, necessary	A report listing each reference’s verdict, the taxonomy-aligned code, the offending field(s), and the source(s) supporting the decision
Hallucinator	None	A report listing each reference’s classification into one of three categories: verified, author mismatch, not found
PAU	For extended coverage, optional	A report providing an overview of the document’s structure, flags of invalid seeming Unicode use, a list of in-text references that were not found in the bibliography, a list of bibliography entries that were not referenced in the main text, a list of frequently cited authors, the results of the ABV, optionally a list of sentences where a reference was cited which seemingly does not support the statement, and an overview of outlet rankings for all bibliography entries

Table 1: Overview of the ABV systems CiteAudit by Shi et al. (2026), CiteTracer by Li, Lin, and Ma (2026), Hallucinator by Stringhini and Blackburn (2026), and PAU.

out an initial technical validation of PAU’s ability to detect incorrect or hallucinated references by comparing its outputs with findings reported by GPTZero in their analysis of NeurIPS papers containing HalluCitations (Shmatko, Adam, and Esau 2026). This comparison provides an initial empirical indication of the reliability of PAU’s reference verification mechanisms (note that other open source tools were not yet available, respective comparisons remain for future work). Throughout the development process, feedback obtained during the proof-of-concept workshop and insights gained from the validation step informed incremental refinements of the tool.

In line with the iterative nature of DSR, this continuous improvement process contributes to the practical functionality of PAU for performing ABV and to the broader understanding of how ABV can support research integrity, as well as providing first insights on how researchers offering reviews for free as community service can be best supported.

4 PAU: A Paper Analysis Utility

PAU is an ABV software whose goal is to assist authors, reviewers, and educators in identifying issues related to the correctness, consistency, and traceability of references used in an academic manuscript. As a secondary goal, PAU can serve an educational purpose by showing students which aspects to examine when critically reading and evaluating scientific publications. By extracting and analyzing citations and bibliographic entries, PAU helps detect problems such as inconsistencies between in-text citations and the bibliography, mismatching metadata, non-existent publications, or semantic discrepancies between statements and the references cited for them. Thus, rather than attempting to judge the scientific contribution or methodological soundness of a manuscript, PAU focuses on aspects of scholarly writing that can be systematically verified.

PAU operates by processing manuscripts in portable document format (PDF) and extracting and verifying textual information using a document parsing pipeline, as summarized in the logic diagram in Figure 2. The tool (PAU.py)

starts by getting the global path to the PDF from the user, as well as information about the document’s:

format, i.e., if there are header lines to consider

bibliography layout, referring to the order in which the entries’ information (namely Lastname, Firstname, Year, Title, and Outlet) are listed

bibliography numbering style, referring to the numbering preceding each bibliography entry (e.g., [1] or none)

in-text referencing style, referring to how bibliographic entries are referenced within the text (e.g., by their numbering or by using the authors’ names and year)

To exemplify, *this* paper has no header lines (format = acm), uses the LFYTO bibliography layout, has no bibliography numbering (= natbib), and uses authors’ names and the publication years for in-text referencing (= authoryear).

PAU then analyzes the document’s structure to identify major sections and specifically the “References” section. Based on the retrieved start and end points, it parses individual bibliography entries to extract structured metadata, including authors, year, title, outlet, and potential links (via ReferenceChecker.py). Once parsed, PAU validates each reference against well-established literature databases, namely querying SemanticScholar (Kinney et al. 2023) and DBLP (University of Trier 1993). This core ABV part of PAU is completely deterministic. As a fallback, the user is provided with a search-link to manually verify the reference’s given information. Mismatching information (such as an author’s name missing or the publication years not matching) and references that could not be found are flagged in the output report (see Figure 3).

PAU next compares the parsed bibliography with the in-text citations found throughout the document (via RefIntextChecker.py). This allows the system to identify inconsistencies such as citations that appear in the text but are missing from the bibliography, as well as references that are listed in the bibliography but never cited within the main text (see Figure 4). Optionally, PAU incorporates LLMs to analyze the relationship between cited

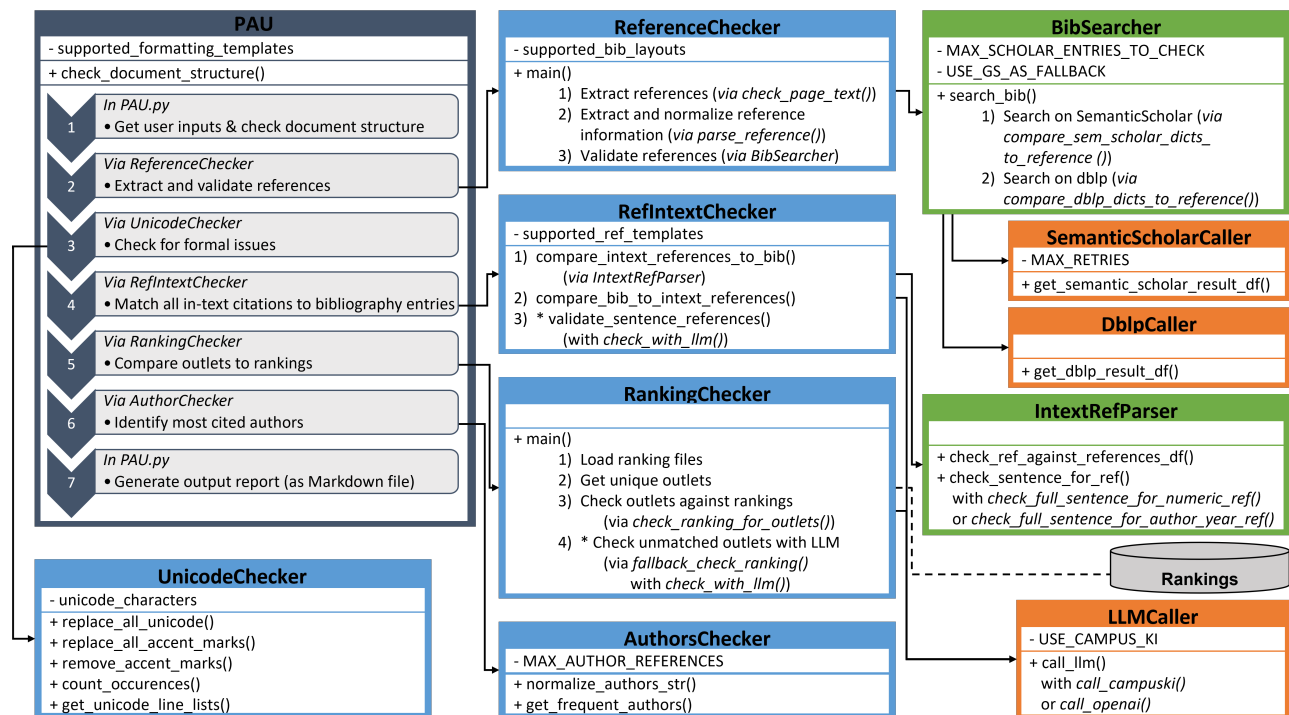


Figure 2: Architecture overview of PAU, its modules, important parameters, and their primary functions, where * indicates an optional feature.

papers and the statements in which they are referenced in. The `LLMCaller.py` currently supports OpenAI, known for their GPT models, and CampusKI, a service provided by TU Dortmund University within the broader *KI:connect.nrw* initiative that provides access to different LLMs. All sentences referencing the same bibliography entry are collected and compared against the reference’s abstract (if available from the aforementioned databases), to assess whether the cited publication is thematically related and appears to support or contradict the statement being made. If cited papers seem to refute or disconnect from the sentences, they are provided near the middle of the output report, while the full list of all assessments (including statements without abstracts of the referenced paper to match against) are reported at the end. While this cannot replace expert judgment, it provides an additional signal that may help reviewers or authors detect questionable or potentially misleading citations.

Furthermore, PAU analyzes the characteristics of the venues in which cited works were published (via `RankingChecker.py`). By comparing parsed outlet names with established ranking datasets (e.g., iCORE rankings and Clarivate’s journal impact factors), the system can provide information about the quality or reputation of referenced journals and conferences. For parsed outlet names that could not be matched to any of the rankings (e.g., if abbreviations were used), users can optionally employ an LLM for backup matching. This analysis offers an overview of the distribution of cited literature across venues of different rankings, which can help reviewers and authors better under-

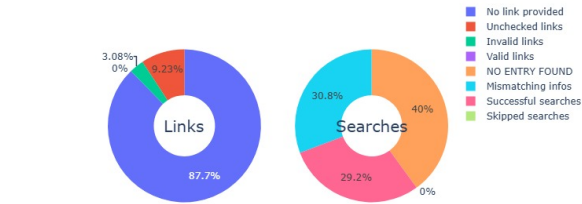
stand the composition of the paper’s literature base. Lastly, PAU counts the frequency with which (co-)authors appear in the bibliography of the paper that is being checked (via `AuthorsChecker.py`).

While still in early development, PAU can review the document’s text on a formal level (via `UnicodeChecker.py`). By seeking sentences containing Unicode characters (such as quotation marks and dashes) deviating from the typical keyboard inputs, PAU can flag inconsistent usages. The placement of each flagged Unicode character is then subsequently checked against heuristic rules of how each character should be used and if the placement seems valid. Each sentence with a character flagged as invalid is then listed in PAU’s output report.

In short, PAU provides an automated framework for analyzing structural and bibliographic aspects of academic manuscripts, augmenting the user’s ability to assess formal as well as related ethical issues. By combining document parsing, reference validation, citation analysis, optional semantic assessment, and outlet overviews, PAU enables ABV to (1) reduce manual efforts required for checking validity of scholarly-written papers and (2) support a more systematic evaluation of how literature is used within a research paper. PAU is implemented in Python and designed modular with readability and extensibility in mind, allowing researchers to easily inspect and adapt the code for their own use cases. Since most assessments are done deterministically through traditional pattern matching and LLM use is optional, PAU does not employ probabilistic

Reference Checker Results

Literature Search Results



References section starts on page 11.

* No link provided: 57/65 = 87.69%

* Unchecked links: 0/65 = 0.00%

* Valid links: 0/65 = 0.00%

* Invalid link: 2/65 = 3.08% ...

* Different links: 0/65 = 0.00%

* Successful searches: 19/65 = 29.23%

* Skipped searches: 0/65 = 0.00%

* Mismatching infos: 20/65 = 30.77%

> * Entries with mismatching author information: ...

* Entries with mismatching years (where the difference is > 1 year):

=> total = 1

* [39] Ashwin K Vijayakumar, Michael Cogswell, Ramprasad R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv*, 1610.02424, 2018.
* Year (2016) does not match expected year (2018)

> * Entries with mismatching years (where the difference is = 1 year): ...

* No entry found: 26/65 = 40.00%

* Unfound entries:

=> total = 26

* [4] Franz Aichberger, Lily Chen, and John Smith. Semantically diverse language generation. In *International Conference on Learning Representations (ICLR)*, 2025.
-> WARNING: None of the 10 results matched the expected title! | WARNING: No results were found through the *dblp* search!
* [5] Jacob Austin and et al. Diffuseq: Sequence to sequence text generation with diffusion models. In *ICLR*, 2022.
-> WARNING: None of the 6 results matched the expected title! | WARNING: None of the 335 results matched the expected title!
* [6] Nuo Lou and et al. Dsp: Diffusion-based span prediction for masked text modeling. *arXiv preprint arXiv:2305.XXXX*, 2023.
-> WARNING: None of the 2 results matched the expected title! | WARNING: None of the 998 results matched the expected title!
* [7] Sheng Shi and et al. Maskgpt: Uniform denoising diffusion for language. *arXiv preprint arXiv:2401.XXXX*, 2024.
-> WARNING: None of the 10 results matched the expected title! | WARNING: No results were found through the *dblp* search!
* [8] A. Sahoo and et al. Inatki: Iterative noise aware text denoising. *arXiv preprint arXiv:2402.XXXX*, 2024.
-> WARNING: None of the 10 results matched the expected title! | WARNING: No results were found through the *dblp* search!
* [9] X. Ou and et al. Tugdm: Token unmasking with quantized diffusion models. In *ACL*, 2024.
-> WARNING: None of the 3 results matched the expected title! | WARNING: No results were found through the *dblp* search!

Figure 3: Excerpt of PAU’s output report section for the results of the ReferenceChecker for paper no. 4 of the dataset used for the technical validation. The yellow highlights were added manually for references that are validated hallucinations based on the work of Shmatko, Adam, and Esau (2026).

References not in Bibliography

All references used in the document’s (main) text were also found in the ‘References’ section.

Unused References in Bibliography

Note

The references listed below were listed in the ‘References’ section but never used in the document’s main text (excluding appendices).

-> Total = 2

Please note that this list may contain false positives as, e.g., the references ID (e.g., first author’s lastname + year) was not correctly parsed or the in-text reference was not recognized due to linebreaks.

* Parsed Ref. ID: 43

* Full reference: [43] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2294, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1250. URL <https://aclanthology.org/D16-1250/>.

* Parsed Ref. ID: 65

* Full reference: [65] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. Technical Appendices and Supplementary Material A

Figure 4: Excerpt of PAU’s output report section for the results of the ReIntextChecker for paper no. 4 of the dataset used for the technical validation.

reasoning for its core functions. The tool is available as open source software under the CC BY-NC-SA license at <https://github.com/DiePaupi/PaperAnalysisUtility>.

5 Evaluation & Results

Proof-of-Concept and Usability

To obtain initial insights into the practical usefulness of PAU, we conducted a workshop with four PhD students (P1–P4) and one senior scholar (P5). The workshop began with an introduction to PAU and a live demonstration of its functionality using example manuscripts. After presenting the generated output report, participants were invited to discuss their first impressions and potential use cases. The discussion was guided by questions regarding the perceived practical benefits of PAU (e.g., in which situations they would use the tool) as well as their evaluation of specific features.

Overall, participants perceived PAU as a useful tool for supporting academic workflows. In particular, several participants emphasized its potential value when reviewing papers outside their immediate field of expertise, with P2 further noting that they could imagine using the tool for every review. Participants also highlighted potential applications in the supervision and evaluation of student theses. As P4 remarked, PAU could “make it easier to assess students’ theses,” and P3 indicated that they could envision using it broadly for “theses, reviews, etc., in fact everywhere, [also] as a last step for my own things.” Extending this educational perspective, P1 suggested that students who are still unsure how to handle citations could use PAU to verify the correctness of their citations before submission. Regarding its influ-

ence on existing workflows, participants generally expected PAU to complement rather than fundamentally change their current practices. P2 described it as a tool that would “make [the workflow] a bit better” by extending existing reviewing routines, rather than replacing them.

When discussing specific features, participants expressed differing views on which functionality they considered most valuable. P4 emphasized the importance of detecting hallucinated or incorrect references, whereas P1, P2, and P3 identified the comparison between citation contexts and the referenced publications as the feature with the greatest future potential. At the same time, they cautioned that semantic analysis based on LLMs should be considered carefully, as such models may occasionally misinterpret context.

Participants also provided some suggestions for further improvements. Both P2 and P5 recommended enabling automatic detection of bibliography and citation formats, although P2 noted that retaining the option for manual configuration would remain desirable. In addition, P5 suggested extending the analysis with more author-focused statistics, such as giving more context about frequently cited authors as well as the authors of the manuscript themselves (if non-anonymous). P1 also mentioned the opportunity of PAU for making forward and backward searches easier, since cited references are looked up already anyway. Finally, P3 contributed several technical suggestions aimed at improving usability and performance, including the use of multithreading and hyperparameter optimization for LLMs.

Technical Validation

To assess whether PAU reliably detects HalluCitations and bibliographic mismatches (e.g., in author names or publication years), we relied on the findings reported by Shmatko, Adam, and Esau (2026). Their study analyzed 4,841 publications accepted at NeurIPS 2025 and identified 100 hallucinated citations across 51 papers using the Hallucination Detector provided by GPTZero. Because this analysis was published in January 2026, many of the affected papers have already been revised by now, and the HalluCitations were no longer present in the most recent versions. Nevertheless, we were able to retrieve earlier versions of 36 of the flagged papers that still contained at least one of the HalluCitations identified by the Hallucination Detector, providing us with ground-truth data to validate PAU.

An overview of the results is shown in Figure 5. Across the 36 analyzed papers, PAU successfully detected every reference flagged as hallucination by Shmatko, Adam, and Esau (2026), meaning that no false negatives were observed in this technical validation. At the same time, PAU flagged a larger number of references overall. Note that for this validation, fallback functions during the validation process, such as the manual search, were skipped in order to assess PAU’s baseline. Across the 36 papers, which together contain 2,172 references, PAU marked 910 entries (41.90 %) as either containing mismatches in author names or not being retrievable through the database searches (i.e., possible hallucinations); including the year-mismatches, a total of 1211 entries were flagged (55.76 %).

A small portion of the author mismatches can be attributed to formatting variations in author names, while some of the unfound entries result from occasional parsing errors; both issues can typically be resolved quickly by inspecting the relevant section of the output report. However, a substantial share of the flagged references appeared to represent legitimate inconsistencies or missing database matches. For transparency and reproducibility, all validation materials (PAU logs, generated output reports, and figures) are publicly available in the PAU repository’s validation folder.

6 Discussion

Reflecting back on our ambitions, careful reviewing remains a central yet often underappreciated component of scientific practice that upholds ethical standards and research integrity. While largely performed as part of voluntary and unpaid community service, the responsible verification of references is of utmost importance for ensuring trust in science (van der Meer and Hameleers 2026). As reviewers and educators frequently operate under substantial time constraints, manually checking the validity and relevance of dozens of references can quickly become impractical. Similarly, some publications take years to materialize and research is being done by multiple authors which contribute parts of the manuscript. It can be challenging as the coordinating author to vouch for the accuracy of all those parts contributed by others. In these contexts, ABV tools such as PAU can help alleviate parts of the burden by automatically highlighting potential inconsistencies and providing a structured overview of how literature is used within a manuscript. Rather than replacing expert judgment with proprietary AI validation, PAU aims to support reviewers in focusing their limited time on the most relevant aspects of a paper, instead of redundant literature checks.

Verification Logic and Accuracy

A central design decision of PAU is the combination of deterministic metadata checks with optional AI-based semantic analysis against literature databases. We argue that this provides a more transparent, interpretable, and reproducible foundation for ABV compared to purely AI-based approaches since, for example, LLMs alone may reproduce the same hallucination issues they are intended to detect. In PAU, AI-based functionality is therefore deliberately positioned as an assistive signal rather than a trusted decision mechanism. The semantic comparison between citation contexts and publication abstracts can help highlight potentially questionable citations, but its outputs should always be interpreted in conjunction with human expertise. Moreover, pattern and string matching is less computationally demanding than leveraging LLMs with billions of parameters, which matters for the sake of environmental impacts and sustainability (van Wynsberghe 2021). Instead of solving everything with AI, responsible use of this technology also necessitates us to weigh practical benefits against resource demands (Fischer, Liebig, and Morik 2024).

Our technical validation results provide initial evidence that our hybrid approach can be effective. When evaluated on a dataset of NeurIPS papers containing previously

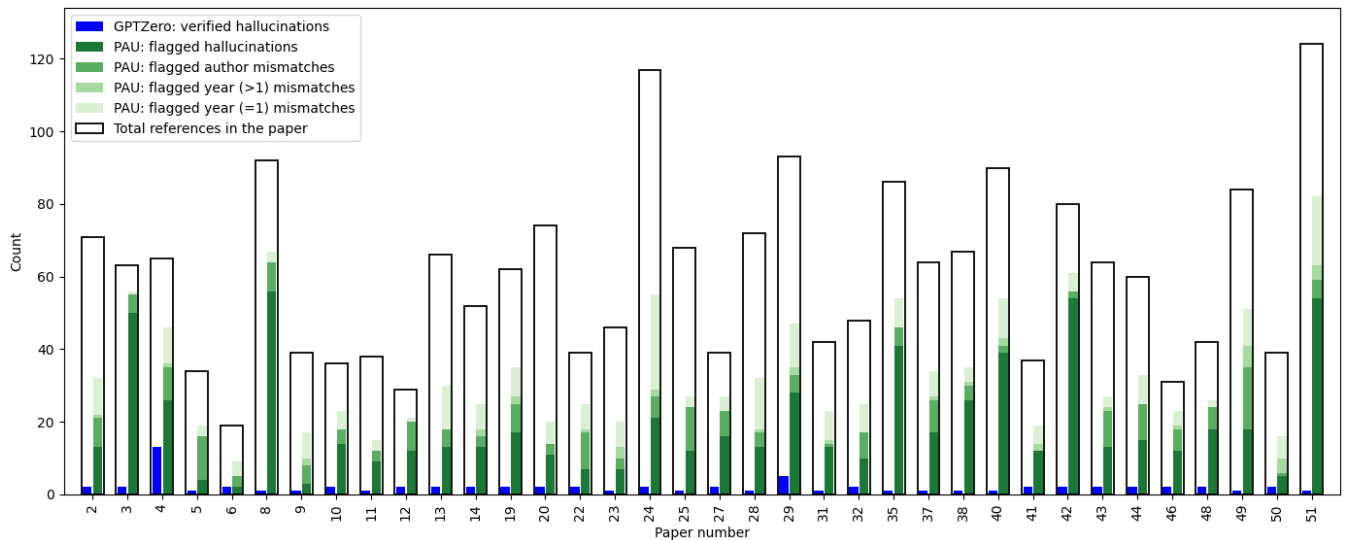


Figure 5: Comparison of the 36 papers with verified hallucinations by Shmatko, Adam, and Esau (2026), flagged references by PAU, and the total number of references listed in each paper’s bibliography.

identified hallucinated citations (Committee on Publication Ethics 2023), PAU successfully flagged all validated hallucinations present in the analyzed paper versions. This absence of false negatives is encouraging, as overlooking HalluCitations would undermine the primary goal of ABV. At the same time, PAU flagged a comparatively large number of references as potentially problematic. While some of these cases stem from minor formatting differences, such as accent marks in author names, many appear to represent legitimate inconsistencies or cases where references could not be matched reliably against the consulted databases. However, it should be noted that the validation dataset consisted exclusively of papers that were previously identified as containing at least one hallucinated reference, which introduces a selection bias and may therefore not reflect the distribution of issues in typical manuscripts. Nevertheless, these findings highlight both the potential and the limitations of ABV: while automation can efficiently surface candidate issues, final interpretation still requires human assessment.

Design and Usability

Going beyond technicalities, a specific design trade-off concerns the strict matching of author names during reference validation that was already mentioned by P1 during the workshop. In its current implementation, PAU requires an exact match between parsed author names and the corresponding entries in literature databases. While this approach reduces ambiguity, simplifies the verification process, and excludes introducing false negatives, it also leads to false positives in cases where names differ slightly due to transliteration or formatting variations. However this does not diminish the conceptual promise of PAU, as we intend to extend it to perform similarity-based string matching techniques (e.g., via Levenshtein distance) to make the comparison more tolerant to minor variations while still preserving

reliable verification. As such, this refinement illustrates the benefits of an iterative, participatory DSR approach (Vaishnavi and Kuechler 2015) for developing infrastructure such as ABV tools, whose practicality requires careful evaluation.

Reflecting on the socio-technical impacts of our work, tools like PAU may enable larger-scale analyses of citation practices across research communities. By systematically examining references across many manuscripts, it would become possible to identify patterns such as unusually high rates of unverifiable citations or systematic inconsistencies in particular venues or subfields. Such large-scale assessments could contribute to broader efforts aimed at identifying and counteracting scientific misconduct or problematic publication practices. As generative AI continues to reshape scientific workflows, developing transparent and scalable tools for safeguarding the integrity of scholarly communication therefore appears both timely and necessary.

Limitations and Future Work

As expected from iterative development, PAU’s promising results come with limitations and opportunities for refinement. First focusing on implementation details, the current application is purely terminal-based, which limits its accessibility for users without a technical background. While this design allows for simplified development and deployment, it makes the usage and configuration of the tool less intuitive. The PAU software currently also supports only a limited number of bibliography and referencing formats. Although the parsing routines were designed to accommodate several common layouts used in computer science publications, variations in document structure can still pose challenges. In particular, differences in section formatting (e.g., numbered versus unnumbered headings or template-specific structures such as conference checklists) may affect the robustness of automatic reference extraction. Lastly, PAU relies on exter-

nal online databases to retrieve bibliographic metadata and abstracts for reference verification. While sources such as SemanticScholar and DBLP provide open access and extensive coverage of computer science literature, their coverage is less comprehensive in other disciplines. This database coverage bias may therefore lead to false positives when references cannot be matched reliably against the consulted sources. Similarly, the semantic comparison between citation contexts and publication abstracts depends on the availability of such metadata.

Another practical limitation concerns the runtime of the analysis. Depending on the number of references contained in a manuscript, the amount of manual fallback checks, and the availability of specialized API keys, the majority of the process runtime results from repeated database queries as well as calls to the LLM used for semantic comparisons. While acceptable for occasional manuscript checks, further optimizations (e.g., by pre-downloading databases or evaluating the performance of smaller AI models) would be desirable for larger-scale analyses that could also investigate trade-offs between detection performance and sustainability of the tool (Fischer, Liebig, and Morik 2024). In addition, the current evaluation focuses primarily on the detection of hallucinated or mismatching references and does not yet provide a systematic quantification of the false positive rate. Although no false negatives were observed in the evaluated dataset, future studies should investigate the accuracy of PAU’s flagging behavior across more diverse (unbiased) corpora, as well as comparisons against competing tools would likely provide valuable performance insights for standardizing ABV. The same applies for the proof-of-concept evaluation, which was performed at a small scale — a broader evaluation with more participants could offer more relevant features to instantiate. Finally, PAU deliberately focuses on reference-level integrity rather than attempting to evaluate broader aspects of research quality such as methodological rigor, citation intent, or plagiarism. While this may be perceived as a limitation, it also reflects a conscious design choice to address a well-defined and automatable subset of research integrity checks.

Building on this, several directions for future work emerge. A key next step is developing a graphical user interface. An interactive front-end could allow users to configure bibliography and citation styles and visualize results through interactive plots. It could also enable users to inspect referenced publications or author profiles within the tool. Further improvements will focus on expanding support for additional document templates, citation styles, and scholarly databases in order to increase the robustness and disciplinary coverage of the system. Future research should also aim to strengthen the empirical evaluation of ABV tools. This includes larger validation studies across different venues and disciplines, as well as systematic measurements of accuracy and efficiency, such as the time saved when using PAU compared to manual verification. In addition, community-level studies could analyze citation integrity across journals or conferences, potentially revealing patterns in the prevalence of unverifiable references. Finally, the development of publicly available benchmark datasets for ABV would greatly

facilitate future research in this area. A curated dataset of verified hallucinated or incorrect citations could serve as a common evaluation resource for emerging tools. Beyond its immediate application for manuscript assessment, PAU also opens several promising avenues for future use. One particularly relevant extension would be the automated identification of retracted publications within reference lists. Given the record number of retractions reported in recent years, integrating checks against retraction databases could help authors and reviewers detect problematic citations before publication. Such functionality would further strengthen the role of ABV as a safeguard for research integrity.

7 Conclusion

The rapid integration of generative AI into scientific publishing comes with considerable opportunities but also critical risks for research integrity and ethics. Among the most visible challenges are hallucinated or inaccurate references, which undermine the traceability of knowledge claims and place additional burdens on authors, reviewers, and supervisors. As AI-assisted writing becomes increasingly common, systematic approaches for ABV are thus becoming important tools of trustworthy scholarly efforts such as reviewing, writing, and communication. With our work, we approached the ABV problem from a socio-technical perspective and introduced PAU, an open source tool designed to support the automated verification of references in academic manuscripts. PAU combines deterministic metadata checks with optional semantic analysis against literature databases, to identify inconsistencies such as mismatching metadata, non-existent publications, or non-consistent literature claims. Our evaluation suggests that this hybrid approach can effectively surface problematic references while providing researchers with a structured overview of how literature is used within a manuscript. Beyond the technical artifact itself, our study also provides insights into how researchers perceive and might integrate such tools into their workflows. Feedback from the proof-of-concept workshop indicates that practitioners view PAU primarily as a supportive instrument that complements existing reviewing and writing practices rather than replacing expert judgment.

Overall, this work contributes both an actionable tool and empirical insights toward improving the transparency and reliability of scholarly communication. As generative AI continues to reshape academic writing and publishing, developing open, transparent, and accessible infrastructures for safeguarding research standards will become increasingly important for ensuring trust in science. To that end, PAU automates the verification of bibliographic information in research manuscripts with an open and practitioner-oriented strategy that can guide the community toward ensuring academic integrity in responsible ways.

Acknowledgements

The work on this paper is performed in the scope of the project “VeraXtract” (16IS24066) funded by the German Federal Ministry for Research, Technology and Space (BMFTR).

References

- Agrawal, A.; Suzgun, M.; Mackey, L.; and Kalai, A. 2024. Do Language Models Know When They're Hallucinating References? In Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 912–928. St. Julian's, Malta: Association for Computational Linguistics.
- Altmeyer, P.; Demetriou, A. M.; Bartlett, A.; and Liem, C. C. S. 2024. Position: Stop Making Unscientific AGI Performance Claims. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, 1222–1242. PMLR.
- Arar, K. H.; Özen, H.; Polat, G.; and Turan, S. 2025. Artificial intelligence, generative artificial intelligence and research integrity: a hybrid systemic review. *Smart Learning Environments*, 12(1): 44.
- Asai, A.; He, J.; Shao, R.; Shi, W.; Singh, A.; Chang, J. C.; Lo, K.; Soldaini, L.; Feldman, S.; D'Arcy, M.; Wadden, D.; Latzke, M.; Sparks, J.; Hwang, J. D.; Kishore, V.; Tian, M.; Ji, P.; Liu, S.; Tong, H.; Wu, B.; Xiong, Y.; Zettlemoyer, L.; Neubig, G.; Weld, D. S.; Downey, D.; Yih, W.-t.; Koh, P. W.; and Hajishirzi, H. 2026. Synthesizing Scientific Literature With Retrieval-Augmented Language Models. *Nature*, 650(8103): 857–863.
- Babaei Giglou, H.; D'Souza, J.; and Auer, S. 2025. LLMs4Synthesis: Leveraging Large Language Models for Scientific Synthesis. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries, JCDL '24*, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1093-3.
- Belouadi, J.; Lauscher, A.; and Eger, S. 2024. AutomaTikZ: Text-Guided Synthesis of Scientific Vector Graphics with TikZ. *International Conference on Learning Representations*, 2024: 55917–55943.
- Belouadi, J.; Ponzetto, S. P.; and Eger, S. 2024. DETIKZ-IFY: synthesizing graphics programs for scientific figures and sketches with TikZ. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, 85074–85108. Red Hook, NY, USA: Curran Associates Inc. ISBN 979-8-3313-1438-5.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, 1877–1901. Curran Associates, Inc.
- Cheng, A.; Calhoun, A.; and Reedy, G. 2025. Artificial Intelligence-assisted Academic Writing: Recommendations for Ethical Use. *Advances in Simulation*, 10(1): 22.
- Chollet, F.; Knoop, M.; Kamradt, G.; and Landers, B. 2025. ARC Prize 2024: Technical Report.
- Committee on Publication Ethics. 2023. Authorship and AI tools. <https://doi.org/10.24318/cCVRZBms>.
- Cortier, V.; and Lin, Z. 2026. ACM CCS 2026 Between-Cycle Transparency Report. Original-date: 2026-04-20T01:18:28Z.
- Dergaa, I.; Chamari, K.; Zmijewski, P.; and Ben Saad, H. 2023. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport*, 40(2): 615–622.
- Eger, S.; Cao, Y.; D'Souza, J.; Geiger, A.; Greisinger, C.; Gross, S.; Hou, Y.; Krenn, B.; Lauscher, A.; Li, Y.; Lin, C.; Moosavi, N. S.; Zhao, W.; and Miller, T. 2026. Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation. *arXiv*, abs/2502.05151.
- Esau, P.; Shmatko, N.; and Adam, A. 2025. GPTZero uncovers 50+ Hallucinations in ICLR 2026. <https://gptzero.me/news/iclr-2026/>.
- Feuerriegel, S.; Hartmann, J.; Janiesch, C.; and Zschech, P. 2024. Generative AI. *Business & Information Systems Engineering*, 66(1): 111–126.
- Fischer, R.; Jakobs, M.; Mücke, S.; and Morik, K. 2020. Solving Abstract Reasoning Tasks with Grammatical Evolution. In *LWDA Workshops: FGWM, KDML, FGWI-BIA, and FGIR*, 6–10.
- Fischer, R.; Liebig, T.; and Morik, K. 2024. Towards More Sustainable and Trustworthy Reporting in Machine Learning. *Data Mining and Knowledge Discovery*, 38(4): 1909–1928.
- Fowler, G. A. 2023. What to do when you're accused of AI cheating. *The Washington Post*.
- Gibney, E. 2026. Major conference catches illicit AI use — and rejects hundreds of papers. *Nature*, 652(8109): 281–282.
- Giray, L. 2024. Benefits and Challenges of Using AI for Peer Review: A Study on Researchers' Perceptions. *The Serials Librarian*, 85(5-6): 144–154.
- GPTZero. 2026. Hallucination Detector - Verify Sources & Bibliographies. <https://gptzero.me/hallucination-detector>.
- Harlow, J.; Weibel, N.; Al Kotob, R.; Chan, V.; Bloss, C.; Linares-Orozco, R.; Takemoto, M.; and Nebeker, C. 2020. Using Participatory Design to Inform the Connected and Open Research Ethics (CORE) Commons. *Science and Engineering Ethics*, 26(1): 183–203.
- Haven, T.; Gopalakrishna, G.; Tijdink, J.; van der Schot, D.; and Bouter, L. 2022. Promoting trust in research and researchers: How open science and research integrity are intertwined. *BMC Research Notes*, 15(1): 302.
- He, S.; Yang, F.; Zuo, J.-p.; and Lin, Z.-m. 2023. ChatGPT for scientific paper writing—promises and perils. *The Innovation*, 4(6).

- Hosseini, M.; Rasmussen, L. M.; and Resnik, D. B. 2024. Using AI to write scholarly publications. *Accountability in Research*, 31(7): 715–723.
- Huang, X.; Ruan, W.; Huang, W.; Jin, G.; Dong, Y.; Wu, C.; Bensalem, S.; Mu, R.; Qi, Y.; Zhao, X.; Cai, K.; Zhang, Y.; Wu, S.; Xu, P.; Wu, D.; Freitas, A.; and Mustafa, M. A. 2024a. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7): 175.
- Huang, Z.; Wang, Z.; Xia, S.; Li, X.; Zou, H.; Xu, R.; Fan, R.-Z.; Ye, L.; Chern, E.; Ye, Y.; Zhang, Y.; Yang, Y.; Wu, T.; Wang, B.; Sun, S.; Xiao, Y.; Li, Y.; Zhou, F.; Chern, S.; Qin, Y.; Ma, Y.; Su, J.; Liu, Y.; Zheng, Y.; Zhang, S.; Lin, D.; Qiao, Y.; and Liu, P. 2024b. OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 19209–19253. Curran Associates, Inc.
- Iordanou, K.; Antoniou, A.; and De Vos, M. 2026. Public Trust in Science: A Systematic Literature Review. *Journal of Academic Ethics*, 24(1): 56.
- Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 248:1–248:38.
- Jurafsky, D.; and Martin, J. H. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released January 6, 2026.
- Kamath, G. 2026. On Violations of LLM Review Policies – ICML Blog. <https://blog.icml.cc/2026/03/18/on-violations-of-llm-review-policies/>.
- Kim, S.-J. 2023. Trends in research on ChatGPT and adoption-related issues discussed in articles: a narrative review. *Science Editing*, 11(1): 3–11.
- Kim, S.-J. 2024. Research ethics and issues regarding the use of ChatGPT-like artificial intelligence platforms by authors and reviewers: a narrative review. *Science Editing*, 11: 96–106.
- Kinney, R. M.; Anastasiades, C.; Authur, R.; Beltagy, I.; Bragg, J.; Buraczynski, A.; Cachola, I.; Candra, S.; Chandrasekhar, Y.; Cohan, A.; Crawford, M.; Downey, D.; Dunkelberger, J.; Etzioni, O.; Evans, R.; Feldman, S.; Gorney, J.; Graham, D. W.; Hu, F.; Huff, R.; King, D.; Kohlmeier, S.; Kuehl, B.; Langan, M.; Lin, D.; Liu, H.; Lo, K.; Lochner, J.; MacMillan, K.; Murray, T. C.; Newell, C.; Rao, S.; Rohatgi, S.; Sayre, P.; Shen, S. Z.; Singh, A.; Soldaini, L.; Subramanian, S.; Tanaka, A.; Wade, A. D.; Wagner, L. M.; Wang, L. L.; Wilhelm, C.; Wu, C.; Yang, J.; Zamarron, A.; van Zuylen, M.; and Weld, D. S. 2023. The Semantic Scholar Open Data Platform. *arXiv*, abs/2301.10140.
- LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep Learning. *Nature*, 521(7553): 436–444.
- Li, M.; Lin, Z.; and Ma, S. 2026. Source or It Didn’t Happen: A Multi-Agent Framework for Citation Hallucination Detection. *arXiv*:2605.08583.
- Lins, S.; Pandl, K. D.; Teigeler, H.; Thiebes, S.; Bayer, C.; and Sunyaev, A. 2021. Artificial Intelligence as a Service. *Business & Information Systems Engineering*, 63(4): 441–456.
- Lu, C.; Lu, C.; Lange, R. T.; Yamada, Y.; Hu, S.; Foerster, J.; Ha, D.; and Clune, J. 2026. Towards end-to-end automation of AI research. *Nature*, 651(8107): 914–919.
- Lund, B. D.; Wang, T.; Mannuru, N. R.; Nie, B.; Shimray, S.; and Wang, Z. 2023. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5): 570–581.
- Naddaf, M. 2025. Major AI conference flooded with peer reviews written fully by AI. *Nature*, 648(8093): 256–257.
- Nature. 2023. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*, 613(7945): 612–612.
- Raman, R. 2025. Transparency in research: An analysis of ChatGPT usage acknowledgment by authors across disciplines and geographies. *Accountability in Research*, 32(3): 277–298.
- Sakai, Y.; Kamigaito, H.; and Watanabe, T. 2026. HalluCitation Matters: Revealing the Impact of Hallucinated References with 300 Hallucinated Papers in ACL Conferences. *arXiv*, abs/2601.18724.
- Scherbakov, D.; Hubig, N.; Jansari, V.; Bakumenko, A.; and Lenert, L. A. 2025. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, 32(6): 1071–1086.
- Schmitt, V.; Villa-Arenas, L.-F.; Feldhus, N.; Meyer, J.; Spang, R. P.; and Möller, S. 2024. The Role of Explainability in Collaborative Human-AI Disinformation Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, 2157–2174. ISBN 9798400704505.
- Shi, K.; Sun, W.; Zhang, Z.; Sun, L.; Chawla, N. V.; and Ye, Y. 2026. CiteAudit: You Cited It, but Did You Read It? A Benchmark for Verifying Scientific References in the LLM Era. *arXiv*:2602.23452.
- Shmatko, N.; Adam, A.; and Esau, P. 2026. GPTZero finds 100 new hallucinations in NeurIPS 2025 accepted papers. <https://gptzero.me/news/neurips/>.
- Singh Chawla, D. 2026. Researchers who use hallucinated references to face arXiv ban. *Nature*, <https://www.nature.com/articles/d41586-026-01595-5>.
- Speckmann, P.; Nadj, M.; and Janiesch, C. 2025. Interactive Explainable Intelligent Systems: Requirements, Design Principles, and Prototypical Implementation. In Chatterjee, S.; vom Brocke, J.; and Anderson, R., eds., *Local Solutions for Global Challenges*, 51–65. Cham: Springer Nature Switzerland. ISBN 978-3-031-93979-2.

- Spinuzzi, C. 2005. The Methodology of Participatory Design. *Technical Communication*, 52: 163–174.
- Stokel-Walker, C.; and Van Noorden, R. 2023. What ChatGPT and generative AI mean for science. *Nature*, 614(7947): 214–216.
- Stringhini, G.; and Blackburn, J. 2026. Experience Paper: Is This the Real Life? Is This Just Fantasy? Hunting for Hallucinated References in Paper Submissions. In *2026 IEEE Symposium on Security and Privacy Workshops (SPW)*, 300–307.
- Sun, Z. 2025. Large language models in peer review: challenges and opportunities. *Scientometrics*, 130(10): 5503–5546.
- Sälzer, M. 2025. *The Complexity of Formal Verification of Neural Network Based Models*. Ph.D. thesis, Universität Kassel.
- Sälzer, M.; Alsmann, E.; Bruse, F.; and Lange, M. 2026. Verifying and interpreting neural networks using finite automata. *Information and Computation*, 308: 105398.
- Topaz, M.; Roguin, N.; Gupta, P.; Zhang, Z.; and Peltonen, L.-M. 2026. Fabricated citations: an audit across 2.5 million biomedical papers. *The Lancet*, 407(10541): 1779–1781.
- University of Trier. 1993. DBLP: Computer Science Bibliography. <https://dblp.org/>.
- Vaishnavi, V. K.; and Kuechler, W. 2015. *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*. CRC Press, 2nd edition.
- van der Meer, T. G. L. A.; and Hameleers, M. 2026. Science and the crisis of trust. *Current Opinion in Psychology*, 67: 102202.
- van Wynsberghe, A. 2021. Sustainable AI: AI for Sustainability and the Sustainability of AI. *AI and Ethics*, 1(3): 213–218.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc.
- Xie, X.; Kersting, K.; and Neider, D. 2022. Neuro-Symbolic Verification of Deep Neural Networks. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 3622–3628. ijcai.org.
- Zaid, F. A.; Neider, D.; and Yalçın, M. 2026. VeriFlow: Modeling Distributions for Neural Network Verification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(33): 28050–28058.
- Zeybek Kabakci, G.; Kara, U. Y.; Baydar Çavdar, G.; and Toros, E. 2025. Bridging the Divide: An Interdisciplinary Analysis of Trust in Science. *Sage Open*, 15(3): 21582440251352376.
- Zhang, Q.; Wang, B.; Huang, V. S.-J.; Zhang, J.; Wang, Z.; Liang, H.; He, C.; and Zhang, W. 2026. Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction. *arXiv*, abs/2410.21169.
- Zhang, Y.; Chen, X.; Jin, B.; Wang, S.; Ji, S.; Wang, W.; and Han, J. 2024. A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8783–8817. Miami, Florida, USA: Association for Computational Linguistics.
- Zhao, Z.; Wang, Y.; Stuart, T.; Vaan, M. D.; Ginsparg, P.; and Yin, Y. 2026. LLM Hallucinations in the Wild: Large-Scale Evidence From Non-Existent Citations. [eprint: 2605.07723](https://arxiv.org/abs/2605.07723).
- Zheng, T.; Deng, Z.; Tsang, H. T.; Wang, W.; Bai, J.; Wang, Z.; and Song, Y. 2025. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 17733–17750. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.