https://doi.org/10.26069/greynet-2025-000.769-gn

PISA 2025

Challenges in Evaluating the Quality and Credibility of AI-Generated Grey Literature

Dobrica Savić
IAEA Consultant in Nuclear Knowledge Management

In the rapidly evolving landscape of scientific communication, grey literature (GL) stands as an important resource. Defined as recorded information not controlled by commercial publishers and made publicly available without the traditional peer-review process, GL includes a wide range of materials such as reports, theses, conference papers, and white papers. The advent of Artificial Intelligence (AI) has revolutionized content creation, offering unprecedented speed and volume. However, this advancement presents unique challenges, particularly in managing, evaluating the integrity and quality of AI-generated content. This paper examines these challenges, highlighting the limitations of traditional evaluation methods for AI-generated outputs. It discusses the rise of AI in GL production, the inadequacies of conventional assessment approaches, and the specific issues associated with AI-generated content. The paper concludes with proposed solutions and best practices to ensure the integrity of scientific communication in the AI-driven era.

Introduction: Understanding Grey Literature

GL represents any recorded, referable and sustainable data or information resource of current or future value, made publicly available without a traditional peer-review process (Savić, 2017).

This definition considers all major elements of the GL concept. Namely, long term preservation, sustainability, usability, and value, while acknowledging the lack of a traditional peer-review process for regular 'white' literature.

As Figure 1 shows, there are many types or forms of GL, although this paper mainly deals with grey data sets. The GreyNet website (greynet.org) lists over 150 document types including databases, data sets, data sheets, data papers, satellite data, and product data.

Bibliographies	Rejected manuscripts	Publications from NGOs and consulting firms
Discussion papers	Un-submitted manuscripts	Videos
Newsletters	Conference abstracts	Wiki articles
PowerPoint presentations	Book chapters	Emails
Program evaluation reports	Personal correspondence	Blogs and social media
Technical notes	Newsletters	Data sets
Publications from governmental agencies	Informal communications	Committee reports
Reports to funding agencies	Census data	Working papers
Unpublished reports	Pre-prints	Company reports
Dissertations	Standards	Catalogues
Policy documents	Patents	Speeches
	Webinars	Reports on websites

Figure 1: Grey literature types

https://doi.org/10.26069/greynet-2025-000.769-gn

The Proliferation of AI Outputs in Grey Literature Production

Quantifying the exact amount of GL produced daily or the total available is challenging due to its diverse and decentralized nature. Many types of GL are often not tracked in the same way as formally published literature, making precise estimates difficult.

However, some general insights can be provided. For example the daily production of GL is substantial, especially given the number of institutions, organizations, and individuals generating reports, white papers, and similar documents. With the rise of digital platforms and the ease of online publication, the rate of GL creation has increased significantly in recent years.

The total volume of GL collected and managed is immense, encompassing millions of documents globally. For example, repositories like AGRIS (12M records), INIS (5M records), the OpenGrey (700,000 EU GL), or ProQuest's Dissertations & Theses (3M records), provide access to large collections, but they still represent only a fraction of what exists.

Regarding emails, as of January 2025, it's estimated that approximately 376.4 billion emails are sent and received daily worldwide. This figure has been steadily increasing over the years and is projected to reach 408.2 billion daily emails by 2027.

In terms of email users, there ware about 4.48 billion individuals utilizing email services globally in 2024. This number is expected to grow to well over 5 billion by 2027 (oberlo.com).

These statistics highlight the enduring significance of email as a primary mode of communication, both personally and professionally, despite the proliferation of alternative messaging platforms.

Impact of AI on GL Creation

AI has revolutionized the production of GL through its ability to automate and enhance various aspects of content creation. Key types of AI-generated GL include:

- **Research Summaries**: AI tools like GPT-4 can condense complex studies into succinct summaries.
- **Preprints**: AI accelerates the drafting of preliminary research findings, facilitating quicker dissemination.
- **Technical Reports**: By synthesizing data, AI produces detailed reports on various subjects, enhancing the accessibility of technical information.
- **Search:** AI-driven platforms, such as Semantic Scholar (semanticscholar.org), have demonstrated the potential for AI to organize and present GL more effectively, enabling researchers to navigate vast amounts of data with greater ease. It allows users to explore over 200 million academic publications and more than 120 million patents. This includes many non-peer-reviewed documents (Karolinska Institutet, 2023). Although BASE (Bielefeld Academic Search Engine) (base-search.net) does not explicitly advertise the use of AI in its core search functionalities, it hosts an immense collection of over 420 million records, many of them GL.

https://doi.org/10.26069/greynet-2025-000.769-gn

AI's role in GL creation is marked by several advantages:

- **Speed**: AI dramatically reduces the time needed to produce GL, as it can generate summaries, reports, and metadata rapidly.
- **Volume**: The scalability of AI allows for the processing of large datasets, which contributes to a significant increase in the availability of GL.
- **Customization**: AI can tailor content to specific audiences, adjust to various contexts, and produce multilingual documents, thereby broadening the reach and applicability of GL.

A study by Mayer-Schönberger and Cukier (2013) highlighted how big data, enabled by AI, is transforming information dissemination by making vast amounts of data manageable and actionable.

Challenges Posed by AI-Generated GL

Despite its advantages, AI-generated GL is not without challenges:

- **Hallucinations and Factual Inaccuracies**: AI models, while advanced, can still produce content that is factually incorrect or fabricated, posing risks to the credibility of GL (Flanagin et al., 2023).
- **Evaluation Challenges**: The rapid pace of AI content generation outstrips the capabilities of traditional peer review processes, necessitating new evaluation methods.
- **Transparency Issues**: The opaque nature of AI algorithms ("black box" problem) complicates the verification of sources and reasoning behind AI-generated content.

A report by Bender et al. (2021) discussed the dangers of "stochastic parrots" in language models, where AI generates text that sounds plausible but lacks grounding in reality, further emphasizing the need for careful scrutiny of AI outputs (Bender et al., 2021).

Traditional Methods of Evaluating GL

Traditional methods such as peer review and bibliometric indicators have been the cornerstones of evaluating GL:

- **Peer Review**: Involves subject matter experts who assess the quality and relevance of content.
- **Bibliometric Indicators**: Metrics like citation analysis, impact factors, and altmetrics help gauge the influence and reach of GL.

However, the applicability of these methods to AI-generated content is limited, as they are designed to evaluate human-authored texts.

Limitations of Traditional Evaluation Methods for AI-Generated Content

Traditional evaluation methods encounter several limitations when applied to AI-generated content:

- **Inadequacy in Assessing AI's Outputs**: Human-centric evaluation metrics often fail to capture the nuances of AI-generated content, such as its authenticity and originality.
- **Speed of Content Generation**: AI can produce content at a pace that surpasses the capacity of traditional review systems, risking the dissemination of unchecked information.

https://doi.org/10.26069/greynet-2025-000.769-gn

- **Bias and Misinterpretation**: AI may incorporate biases from its training data or misinterpret complex scientific concepts, further complicating the evaluation process (MIT Sloan EdTech, 2024).

Vincent and Hecht (2018) have argued for the development of new evaluation frameworks that are specifically designed to address the unique challenges posed by AI-generated content (Vincent & Hecht, 2018).

SIFT and RADAR are such frameworks used for evaluating information sources, particularly in digital and academic contexts. **SIFT** stands for Stop, Investigate the source, Find better coverage, and Trace claims, quotes, and media to the original context. This method, developed by digital literacy expert Mike Caulfield, focuses on quickly assessing information before investing time in reading or viewing a source (Caulfield, 2024). **RADAR** is an acronym for Relevance, Authority, Date, Appearance, and Reason. This evaluation tool helps assess the quality, credibility, and relevance of information sources (Mandalios, 2013).

Proposed Solutions and Best Practices

To address the limitations of traditional evaluation methods, the following solutions are proposed:

- **AI-Specific Evaluation Criteria**: Developing new metrics, tools, and frameworks tailored to the characteristics of AI-generated content.
- **Enhanced Transparency**: Implementing clear labeling to indicate AI involvement in content creation, ensuring that readers are aware of the origins of the text.
- **AI Detection Tools**: Utilizing software to identify and flag AI-generated content, thus maintaining the integrity of GL.
- **Ethical Guidelines**: Establishing comprehensive standards for the responsible use of AI in scientific communication.
- **Clear Marking of AI-Generated Content:** Using a statement such as *"This article was partially generated using artificial intelligence tools"*.

A joint report by UNESCO and the OECD emphasizes the importance of fostering interdisciplinary collaboration to develop effective guidelines and frameworks for evaluating AI-generated content (UNESCO & OECD, 2022).

Conclusion

AI is revolutionizing GL production, offering unprecedented speed, volume, and customization capabilities. However, traditional evaluation methods are inadequate for assessing AI-generated content, necessitating new approaches. Key issues such as hallucinations, accuracy concerns, and ethical considerations pose significant challenges to the credibility of AI-generated GL. Proposed solutions, including AI-specific evaluation frameworks and enhanced transparency, are crucial for maintaining scientific integrity. Interdisciplinary collaboration between AI experts and domain specialists is essential for developing effective evaluation strategies.

https://doi.org/10.26069/greynet-2025-000.769-gn

The integration of AI into GL production offers unparalleled opportunities for innovation and efficiency. However, it also necessitates a fundamental shift in how we evaluate and ensure the integrity of this content. By adopting new evaluation methods, we can harness the potential of AI while safeguarding the credibility and utility of GL.

As we move forward, it is crucial to remain adaptive and vigilant. The rapid pace of technological advancement requires continuous refinement of our evaluation methods to ensure that grey literature continues to serve as a reliable and valuable resource in the scientific community.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

Caulfield, M. (2024). *Evaluating Sources*. Simon Fraser University. www.lib.sfu.ca/help/research-assistance/evaluating-sources

Flanagin, S., Bibbins-Domingo, K, Berkwits, M. & Christiansen, M.A. (2023). Nonhuman *"Authors" and Implications for the Integrity of Scientific Publication and Medical Knowledge.* JAMA. 2023;329(8):637-639. doi:10.1001/jama.2023.1344

Karolinska Institutet,University Library Web Team (2023) https://kib.ki.se/en/search-evaluate/grey-literature

Mandalios, J. (2013). *RADAR: An approach for helping students evaluate Internet sources*. Journal of Information Science, 39(4), 470-478.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt.

MIT Sloan EdTech (2024). *When AI Gets It Wrong Addressing AI Hallucinations and Bias*. https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/

Savić, D., 2017. *Rethinking the Role of Grey Literature in the Fourth Industrial Revolution*. 10th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology. TGJ (The Grey Journal) Special Winter Issue, Volume 14, 2018.

Vincent, N., & Hecht, B. (2018). *Data Strikes: Evaluating the Feasibility of Social Protest against Data-Driven Technologies*. *Proceedings of the ACM on Human-Computer Interaction*.

UNESCO & OECD. (2022). *Guidelines for the Responsible Use of AI in Scientific Communication*.