From citation metrics to LLM judgement: a GAIDeT-based framework for structured AI disclosure

Serhii Nazarovets[*], Yana Suchikova[**], Natalia Tsybuliak[***] and Jaime A. Teixeira da Silva[****]

[*]*serhii.nazarovets@gmail.com*

https://orcid.org/0000-0002-5067-4498

Library, Borys Grinchenko Kyiv Metropolitan University, Ukraine

[**]*yanasuchikova@gmail.com*; [***]*nata.tsibulyak@gmail.com*; [****]*jaimetex@yahoo.com*

https://orcid.org/0000-0003-4537-966X; https://orcid.org/0000-0003-2474-8614; https://orcid.org/0000-0003-3299-2772

Berdyansk State Pedagogical University, Ukraine; Berdyansk State Pedagogical University, Ukraine; Independent researcher, Japan

Conference paper presented at the 30th Annual International Conference on Science and Technology Indicators (STI-ENID 2026), Antwerp, Belgium, 9–11 September 2026.

Abstract

Generative AI (GenAI) tools are increasingly used in research production, while large language models (LLMs) are being explored for research evaluation. However, most GenAI-use disclosures remain free-text statements that are not represented as structured metadata and therefore cannot be systematically aggregated or analysed. This paper argues that the next stage of development of the Generative AI Delegation Taxonomy (GAIDeT) is infrastructural rather than conceptual. We propose that GAIDeT can function as a semantic layer for machine-readable GenAI-use metadata, enabling interoperability across scholarly communication systems. Such metadata may support transparency and provide contextual information for future GenAI-aware evaluation systems.

1. Introduction

Research evaluation has long relied on peer review and citation-based indicators. Citation metrics show moderate correlations with expert assessments of research quality (Aksnes, Piro, & Fossum, 2023), although citation impact is widely recognised as context-dependent and multi-causal (Aksnes, Langfeldt, & Wouters, 2019; Waltman & Traag, 2021).

Recent studies suggest the emergence of an additional evaluative layer based on large language models (LLMs). Experimental evidence indicates that ChatGPT can generate research quality assessments that, under certain conditions, correlate with expert judgements or citation-based indicators (Thelwall, 2024; Thelwall, 2025a). LLM-based predictions derived from article abstracts have also been linked to later citation impact (Thelwall, Jiang, & Bath, 2025). Unlike citation indicators, however, LLM evaluations operate directly on textual input and are sensitive to prompting, scoring instructions, model versions, and input characteristics (Thelwall, 2025b; Thelwall, 2026; Thelwall & Mohammadi, 2026). Textual clarity and framing may also influence model-generated evaluations independently of methodological quality (de Winter, 2024).

As research production increasingly involves generative AI (GenAI) tools, the absence of structured metadata describing GenAI-assisted tasks creates a new interpretative challenge. Evaluation systems may analyse research outputs without access to information about how those outputs were produced. Building on the Generative AI Delegation Taxonomy (GAIDeT) (Suchikova et al., 2026), this paper argues that the next stage of AI disclosure should be infrastructural rather than conceptual. We examine how GAIDeT can function as a semantic layer for machine-readable GenAI-use metadata and discuss its potential role in future GenAI-aware evaluation systems.

2. The infrastructural problem of AI-use disclosure

In most journals, GenAI-use disclosure appears as a free-text statement within the manuscript. Although such statements improve transparency, they typically remain embedded in PDF or HTML text and are not encoded as structured metadata. Even when controlled vocabularies such as GAIDeT are used, declarations are usually presented as descriptive text rather than machine-readable metadata linked to persistent identifiers.

As a result, GenAI-use information is generally not deposited as dedicated metadata fields and is not indexed as a searchable element in major scholarly infrastructures. Unlike author identifiers or funding information, which are transmitted through standardized metadata workflows and harvested by bibliographic databases (Hendricks et al., 2020), GenAI-use disclosures remain difficult to retrieve, aggregate, and analyse across publications. The challenge is therefore not the absence of disclosure policies, but the absence of interoperable metadata integration.

3. GAIDeT as a semantic infrastructure

GAIDeT was originally developed as a structured taxonomy for declaring delegated GenAI tasks in research (Suchikova et al., 2026). Since then, it has evolved into a semantic infrastructure based on an ontology with persistent identifiers and a resolvable PURL (https://w3id.org/gaidet/gaidet.owl). This allows GenAI-assisted tasks to be linked to defined ontology classes rather than described solely as free text.

To support practical implementation, a GAIDeT Declaration Generator (https://w3id.org/gaidet/) has been developed to produce structured GenAI-use statements aligned with the ontology. Together, these resources provide a foundation for machine-readable disclosure.

The transition from descriptive statements to interoperable metadata does not require a new taxonomy, rather the integration of existing semantic resources into scholarly communication workflows. The main challenge is therefore no longer conceptual design, but infrastructural adoption by publishers, repositories, and indexing systems.

4. Implications for GenAI-aware evaluation

Structured GenAI-use metadata is not intended to function as a new performance indicator. Rather than assigning value to GenAI involvement, GAIDeT-based disclosure can provide contextual information for evaluation systems that increasingly rely on text-analytic methods. In the same way that field, document type, or funding information can support interpretation of citation indicators, information about delegated GenAI tasks may serve as a filtering or stratification variable in future GenAI-aware evaluation environments.

The availability of structured metadata would enable analyses of how GenAI-assisted research production interacts with GenAI-assisted assessment. For example, LLM-generated evaluations could be compared across outputs with different forms of declared GenAI involvement, while disciplinary patterns of GenAI use could be examined over time. The purpose is not normative judgement but interpretive transparency.

At present, however, such applications remain largely conceptual. Although GAIDeT provides a semantic structure for machine-readable disclosure, its practical implementation depends on adoption within publisher, repository, and indexing workflows. The main challenge is therefore not technical design, but systematic infrastructural uptake across scholarly communication systems.

5. Conclusion

Current GenAI-use disclosures are increasingly common but remain largely invisible to scholarly metadata infrastructures. GAIDeT provides a semantic foundation for representing delegated GenAI tasks in a machine-readable form and may support future GenAI-aware evaluation systems. The next stage of development is therefore not conceptual but infrastructural: integrating structured GenAI-use metadata into scholarly communication workflows.

Acknowledgments

The authors, SN, YS, and NT, would like to express their deepest gratitude to all Ukrainian defenders, whose courage and resilience made it possible to complete and publish this work.

Author contributions
All authors contributed equally to the conception, design, writing, and revision of this manuscript. Each author has reviewed and approved the final version of the manuscript and agrees to be accountable for the content and accuracy of the work.

Competing interests
The authors declare no relevant conflicts of interest.

Funding information
This research was not funded.

References
Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, 9(1), 215824401982957. https://doi.org/10.1177/2158244019829575

Aksnes, D. W., Piro, F. N., & Fossum, L. W. (2023). Citation metrics covary with researchers' assessments of the quality of their works. *Quantitative Science Studies*, 4(1), 105–126. https://doi.org/10.1162/qss_a_00241

de Winter, J. (2024). Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. *Scientometrics*, 129(4), 2469–2487. https://doi.org/10.1007/s11192-024-04939-y

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427. https://doi.org/10.1162/qss_a_00022

Suchikova, Y., Tsybuliak, N., Teixeira da Silva, J. A., & Nazarovets, S. (2026). GAIDeT (Generative AI Delegation Taxonomy): A taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. *Accountability in Research*, 33(3), 2544331. https://doi.org/10.1080/08989621.2025.2544331

Thelwall, M. (2024). Can ChatGPT evaluate research quality? *Journal of Data and Information Science*, 9(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Thelwall, M. (2025a). Evaluating research quality with Large Language Models: An analysis of ChatGPT's effectiveness with different settings and inputs. *Journal of Data and Information Science*, 10(1), 7–25. https://doi.org/10.2478/jdis-2025-0011

Thelwall, M. (2025b). Research quality evaluation by AI in the era of large language models: Advantages, disadvantages, and systemic effects – An opinion paper. *Scientometrics*, 130, 5309–5321. https://doi.org/10.1007/s11192-025-05361-8

Thelwall, M. (2026). Large language models and responsible research evaluation: an extension of the Leiden Manifesto. *Scientometrics*. https://doi.org/10.1007/s11192-026-05552-x

Thelwall, M., Jiang, X., & Bath, P. A. (2025). Estimating the quality of published medical research with ChatGPT. *Information Processing and Management*, 62(4), 104123. https://doi.org/10.1016/j.ipm.2025.104123

Thelwall, M., Mohammadi, E. (2026). Can small and reasoning large language models score journal articles for research quality and do averaging and few-shot help? *Scientometrics*. https://doi.org/10.1007/s11192-026-05585-2

Waltman, L., & Traag, V. A. (2021). Use of the journal impact factor for assessing individual articles: Statistically flawed or not? *F1000Research*, 9, 366. https://doi.org/10.12688/f1000research.23418.2