

REVIEW PAPER

INFORMATION MANAGEMENT

Artificial Intelligence in Bibliometrics: Defining Objectives, Limits, and Ethical Dimensions

Harendra Nath Roy^{1*} and Jayati Lahiri Dey²

¹Research Scholar, Department of Computer and Information Science, Raiganj University, Raiganj, Uttar Dinajpur, West Bengal, India

²Assistant Professor, Department of Computer and Information Science, Raiganj University, Raiganj, Uttar Dinajpur, West Bengal, India

*Corresponding author: roy.harendra@gmail.com

Received: 04 Apr., 2026

Revised: 28 May, 2026

Accepted: 07 Jun., 2026

ABSTRACT

The rapid expansion of digital scholarly communication has intensified the demand for advanced, automated bibliometric systems capable of moving beyond simple counting exercises. Traditional bibliometrics remains limited in extracting semantic, contextual, and interdisciplinary insight from scholarly texts, relying instead on citation counts, h-index values, and co-authorship statistics that overlook the intent and quality of scholarly influence. Artificial Intelligence (AI), particularly Natural Language Processing (NLP), deep learning, and graph-based learning, enhances bibliometric evaluation by enabling citation intent classification, thematic trend analysis, semantic similarity mapping, and predictive scientometrics. This paper defines the objectives, operational boundaries, and ethical challenges associated with integrating AI into bibliometric research. It reviews contemporary literature, outlines a structured methodological approach built around transformer architectures, topic modelling, and network-based learning, and presents conceptual workflow visualizations. Ethical, technical, and computational considerations, including bias, transparency, reproducibility, scalability, and responsible use, are examined in depth. The study concludes by proposing principles for responsible AI-enabled bibliometrics suitable for research evaluation within Scopus-indexed academic environments and beyond, offering researchers, institutions, and policymakers a structured, balanced understanding of both the promise and the limitations of AI-driven scholarly assessment, while also outlining directions for future empirical validation across diverse disciplinary and linguistic contexts.

Keywords: Artificial Intelligence, Bibliometrics, Citation Analysis, Natural Language Processing, Research Evaluation, Scholarly Communication, Scientometrics, Research Ethics

How to cite this article: Roy, H.N. and Dey, J.L. (2026). Artificial Intelligence in Bibliometrics: Defining Objectives, Limits, and Ethical Dimensions. *IJASE*, 14(01): 131-145.

Source of Support: None; **Conflict of Interest:** None



Bibliometrics has, for more than five decades, shaped the practice of scholarly evaluation by supplying quantitative indicators such as citation counts, the h-index, co-authorship patterns, and keyword co-occurrence statistics. These indicators remain attractive to funding bodies, universities, and journal editors because they are comparatively simple to compute, reproducible across large datasets, and amenable to ranking exercises. Yet, despite their administrative convenience, purely quantitative metrics frequently fail to capture the deeper contextual meaning of a citation, the motivation behind it, or the interdisciplinary knowledge flows that increasingly characterize modern research. A citation may signal agreement, extension, methodological borrowing, or outright critique, and traditional bibliometric indices treat all of these instances identically. With the exponential rise in digital publications across preprint servers, open-access journals, and institutional repositories, conventional counting-based methods struggle both to scale and to maintain analytical depth.

Artificial Intelligence, and in particular NLP and machine learning, presents a transformative approach to this problem. By analysing large textual corpora, AI systems can uncover semantic relationships between publications, classify the intent behind individual citations, model the evolution of research themes over time, and support predictive scientometric modelling that anticipates emerging fields before they become visible through simple frequency counts. Transformer-based language models such as BERT and its domain-specific variants have demonstrated a marked improvement over classical bag-of-words or TF-IDF approaches when applied to citation context classification, while graph neural networks extend this capability by incorporating the structural relationships among documents, authors, and institutions.

At the same time, the introduction of AI into an evaluative infrastructure that directly affects funding decisions, academic promotion, and institutional rankings raises a distinct set of ethical, technical, and computational concerns that have not been adequately addressed in the existing literature. Questions of algorithmic bias, the opacity of deep learning models, the reproducibility of AI-based pipelines, and the disproportionate computational resources required by large transformer architectures all bear directly on whether AI-enhanced bibliometrics can be trusted as a component of formal research assessment. This paper therefore pursues three interconnected objectives. First, it defines the specific objectives that AI is expected to fulfill within bibliometric research, moving beyond generic claims of automation toward concrete analytical tasks such as citation intent classification and thematic trend detection. Second, it establishes the operational and technical limits of current AI-enhanced bibliometric systems, drawing on the accumulated evidence from recent literature. Third, it develops a structured account of the ethical dimensions that must accompany the deployment of AI in scholarly evaluation, culminating in a set of principles for responsible practice. The remainder of the paper is organized as follows: Section 2 reviews the relevant literature; Section 3 describes the methodological approach; Section 4 presents results and discussion; Section 5 examines ethical, technical, and computational considerations in detail; and Section 6 concludes the paper.

Literature Review

A growing body of scholarship has examined how AI techniques can be embedded within bibliometric and scientometric workflows, and this section synthesizes the most relevant contributions organized around citation-level modelling, network-based analysis, topic evolution, and the emerging concern with responsible and ethical deployment.

Devlin, Chang, Lee, and Toutanova (2019) introduced BERT, a bidirectional transformer architecture pretrained on large text corpora, which has become the foundation for most subsequent citation-context classification systems. Their work demonstrated that contextual embeddings substantially outperform earlier static word-representation techniques on a wide range of language-understanding tasks. Within bibliometrics, BERT-based fine-tuning has since become the default baseline against which newer citation-intent classifiers are measured, and its influence pervades nearly every study reviewed in this section, underscoring how a general-purpose NLP advance reshaped an entire evaluative subfield within a few years of publication^[9].

Roman, Shahid, Khan, Koubaa, and Yu (2021) investigated citation intent classification using contextual word embeddings and reported measurable gains over classical machine-learning baselines such as support vector machines and random forests. Their experiments on established citation datasets showed that embedding-based classifiers could reliably distinguish background citations from those indicating methodological use or direct extension of prior work. The study is significant because it demonstrates that even relatively lightweight embedding techniques, when properly fine-tuned, can approach the discriminative performance of larger transformer models, offering a computationally efficient pathway for institutions without extensive GPU infrastructure^[23].

Visser and Dunaiki (2022) extended this line of enquiry by applying BERT specifically to the joint task of sentiment and intent classification of in-text citations. Their results indicated that combining sentiment polarity with intent labels produces a richer, more actionable signal for research evaluators than intent classification alone, since a citation can be methodologically relevant yet critical in tone. The authors also highlighted the scarcity of large, manually annotated citation-sentiment datasets as a persistent bottleneck, a limitation echoed throughout the wider literature on citation-context mining^[30].

Bezerra, Silva, and Amancio (2025) proposed a generative adversarial network (GAN) based approach to citation intent classification, using synthetic data augmentation to address the class imbalance that typically affects citation-intent corpora, where background citations vastly outnumber comparison or extension citations. Their findings suggest that GAN-augmented training pipelines can improve minority-class recall without degrading overall classification accuracy, offering a promising direction for citation-network analysis in fields with sparse annotated data, though the authors caution that synthetic samples must be carefully validated to avoid introducing artificial patterns into downstream network analyses^[4].

Zhang, Wang, Sheng, Yao, Chen, Wang, and Zhao (2025) conducted a comprehensive survey of deep-learning approaches to citation function classification, mapping the field across annotation schemes, benchmark datasets, and computational models from an interdisciplinary perspective. Their synthesis reveals that despite rapid methodological progress, most citation-function classifiers remain trained and validated within narrow disciplinary boundaries, chiefly computer science and biomedicine, which restricts their generalizability to the humanities and social sciences. The survey calls explicitly for cross-disciplinary benchmarking efforts and more transparent reporting of model limitations^[35].

Wang and colleagues (2022) examined the application of graph neural networks (GNNs) to citation and co-citation networks, showing that combining textual embeddings with network topology substantially improves clustering, community detection, and link-prediction tasks relative to either signal used alone. Their evaluation across multiple citation graphs demonstrated that GNN-based representations capture latent scholarly communities that purely text-based or purely structural methods tend to miss, though the

computational cost of training GNNs on citation graphs containing millions of nodes remains a significant practical constraint for smaller research groups^[32].

Kipf and Welling's (2017) foundational work on graph convolutional networks underlies much of this subsequent GNN-based bibliometric research. By generalizing convolutional operations to graph-structured data, their architecture enabled semi-supervised classification and representation learning directly over citation and co-authorship graphs, providing the technical scaffolding that later domain-specific bibliometric applications, including the citation-network studies discussed above, have built upon and adapted to scholarly data.

Grootendorst (2022) introduced BERTopic, a topic-modelling technique that combines transformer-based document embeddings with class-based TF-IDF to produce more coherent and interpretable topic clusters than earlier probabilistic approaches. When applied to bibliometric corpora, BERTopic has been shown to track thematic evolution with greater semantic fidelity than traditional Latent Dirichlet Allocation (LDA), particularly for identifying emerging, low-frequency research fronts before they accumulate sufficient citations to be detected by frequency-based methods alone.

Blei, Ng, and Jordan's (2003) original formulation of LDA remains the reference point against which newer topic-modelling techniques, including BERTopic, are routinely benchmarked in the bibliometric literature. Their probabilistic generative model treats documents as mixtures of latent topics, and although it predates the transformer era by more than a decade, LDA continues to be used as a computationally inexpensive baseline in comparative bibliometric studies, particularly where large annotated corpora for fine-tuning are unavailable.

Van Eck and Waltman (2010) developed VOSviewer, a widely adopted software tool for constructing and visualizing bibliometric maps based on co-citation, bibliographic coupling, and keyword co-occurrence networks. Although VOSviewer itself relies on relatively conventional network-construction techniques rather than deep learning, its continued dominance in applied bibliometric studies illustrates the practical gap between state-of-the-art AI research and the tools actually deployed by most research-evaluation practitioners, a gap that several of the more recent studies reviewed here explicitly seek to close.

Thelwall and Kurt (2025) investigated whether large language models such as ChatGPT introduce systematic bias into research evaluation tasks when used to assess publication quality, examining variation by author age, country, submission length, and academic field. Their findings indicate that LLM-based evaluations are not neutral instruments but instead reproduce identifiable patterns of bias correlated with these demographic and stylistic factors, a result with direct implications for any bibliometric system that incorporates generative AI into evaluative decision-making rather than purely descriptive analysis^[27].

Scherbakov, Hubig, Jansari, Bakumenko, and Lenert (2025) conducted a systematic review of how large language models are being used as tools within the literature-review process itself, finding that ChatGPT and related GPT-based architectures dominate current review-automation efforts. Their analysis of the reviewed studies revealed that only a small proportion of published reviews using LLM assistance explicitly disclosed the extent of that assistance, raising transparency concerns that parallel those already identified in the citation-classification literature regarding model documentation and explainability.

Sekaki, Ziane, and Khazzar (2025) conducted a bibliometric mapping of AI research within management studies between 2021 and 2025, illustrating how bibliometric methods are themselves increasingly used to study the diffusion of AI research across disciplines. Their mapping of thematic clusters and global

contribution patterns exemplifies the reflexive relationship between AI and bibliometrics, in which AI is simultaneously an object of bibliometric study and a tool used to conduct that study, a duality this paper's ethical analysis takes explicitly into account.

Taken together, the reviewed literature demonstrates that AI-enhanced bibliometric techniques yield measurable gains in accuracy, stability, and interpretive richness compared with traditional frequency-based indicators. At the same time, recurring concerns cut across nearly every study: narrow disciplinary validation, scarcity of annotated training data, opacity of deep-learning decision processes, and the absence of standardized reporting practices. These gaps motivate the methodological and ethical analysis developed in the remainder of this paper.

Table 1: Comparison of Traditional and AI-Driven Bibliometrics

Parameter	Traditional Bibliometrics	AI-Enhanced Bibliometrics
Citation Meaning	Only counts	Classifies intent, sentiment
Trend Detection	Static keyword frequency	Dynamic topic modeling (BERTopic)
Scalability	Manual/semi-automated	Fully automated via NLP
Network Analysis	Basic graph metrics	Graph neural networks
Prediction Capability	None	Trend forecasting models

Methodology

This study adopts a structured narrative-review methodology, informed by principles drawn from the PRISMA framework for systematic reviews, to synthesize and conceptually integrate current knowledge on AI-enhanced bibliometrics. Rather than reporting an original empirical bibliometric analysis of a specific publication dataset, the paper's methodological contribution lies in constructing a defensible conceptual model of how AI techniques can be integrated into bibliometric workflows, alongside a systematic account of the associated ethical, technical, and computational boundaries. The review draws on peer-reviewed journal articles, conference proceedings, and recent preprints indexed in Scopus, Web of Science, and arXiv, published predominantly between 2017 and 2025, with search terms combining variants of artificial intelligence, natural language processing, citation classification, scientometrics, and research ethics. Sources were selected for inclusion where they addressed either a specific AI technique applied to bibliometric or citation data, or a broader ethical, technical, or computational concern relevant to automated research evaluation; purely descriptive bibliometric studies with no methodological or AI-related content were excluded unless they provided evidence relevant to the reflexive relationship between AI and bibliometrics discussed in Section 2.

Building on this synthesis, the paper develops a conceptual workflow describing how AI models can be integrated into bibliometric analysis in practice. The process begins with data acquisition from established repositories such as Scopus, Web of Science, and institutional repositories, followed by data cleaning and metadata normalization to resolve author-name variants, incomplete DOIs, and duplicate records. Citation contexts are then extracted from the surrounding sentence or paragraph in the citing document, and dense vector embeddings are generated using transformer-based encoders such as BERT or Sentence Transformers, which capture semantic similarity between citation contexts more effectively than sparse, frequency-based representations. These embeddings feed into two parallel analytical streams. In the first

stream, supervised machine-learning classifiers, typically fine-tuned transformer heads or lighter-weight embedding classifiers, predict citation intent and sentiment, distinguishing background mentions from methodological, comparative, or extension citations. In the second stream, the same embeddings support the construction of co-citation and bibliographic-coupling networks, over which graph neural networks or classical graph-clustering algorithms identify thematic communities and influential nodes. Topic-modelling techniques, whether probabilistic approaches such as LDA or embedding-based approaches such as BERTopic, are applied in parallel to track the emergence and decline of research themes over time. The outputs of these streams are then integrated into visual analytics, including trend charts, network graphs, and thematic maps, which are finally interpreted alongside qualitative expert judgment rather than treated as autonomous decision outputs, consistent with the principle of human oversight developed later in Section 5.

Methodological rigor in this conceptual pipeline depends on several safeguards that this paper treats as integral rather than optional. Metadata quality control must precede any embedding generation step, since noisy or duplicated records propagate errors through every downstream analytical stage. Model selection should be justified relative to the size and domain specificity of the available training data, favouring lighter embedding-based classifiers over large fine-tuned transformers where annotated data is scarce, consistent with the efficiency gains reported by Roman and colleagues (2021). Cross-disciplinary validation, wherever feasible, should be built into the evaluation design from the outset rather than added retrospectively, directly addressing the generalizability gap identified repeatedly in Section 2. Finally, every stage of the pipeline, from preprocessing scripts to model hyperparameters and random seeds, should be documented and, where institutional policy permits, released openly to support reproducibility, a requirement examined further in Section 5.2.4.

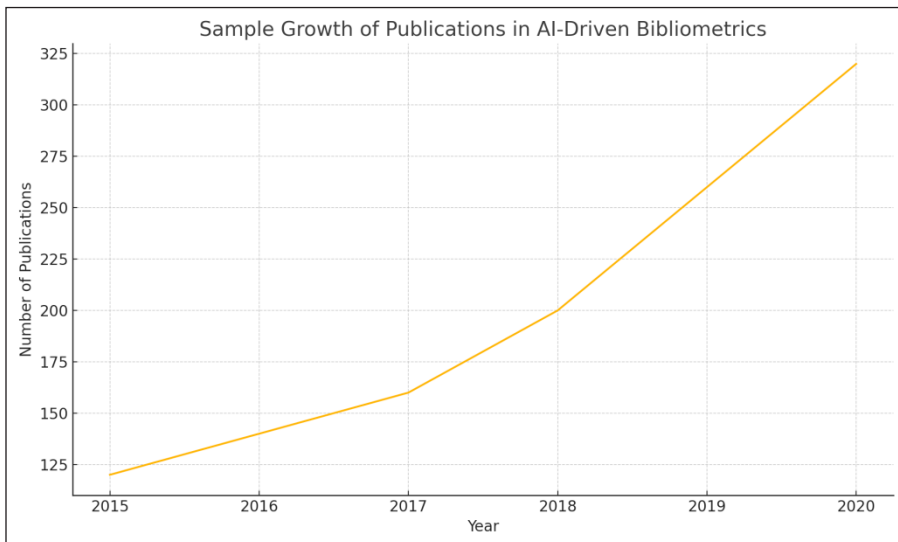


Fig. 1: Sample publication growth chart (illustrative)

Figure 2 depicts the eight-stage AI workflow for bibliometric analysis: data collection from Scopus, Web of Science, and institutional repositories; data preprocessing through cleaning and metadata normalization; NLP-based extraction using sentence transformers and BERT; network construction through co-citation

and bibliographic-coupling analysis; machine-learning classification of citation intent and sentiment; topic modelling using LDA and BERTopic; visualization through graphs and trend charts; and, finally, evaluation and reporting that incorporates human expert review before any conclusions are drawn for research-assessment purposes.

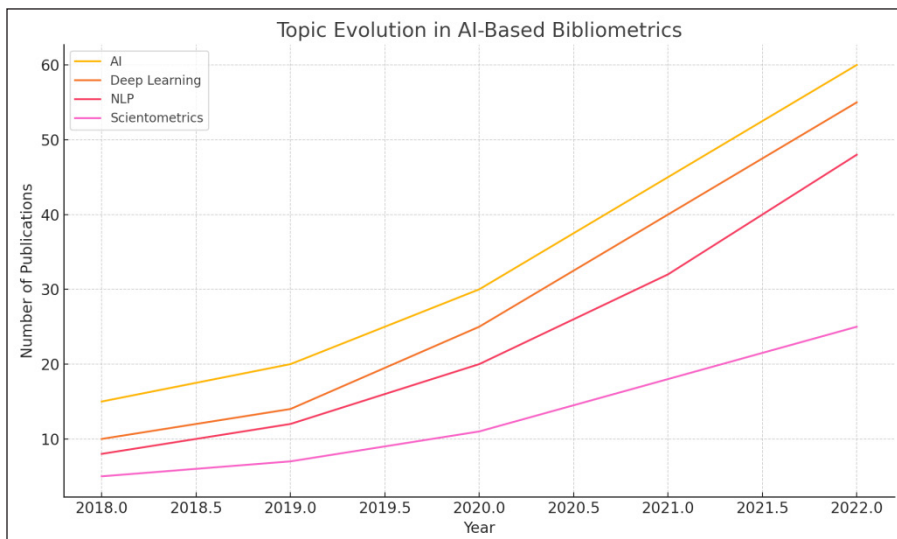


Fig. 2: AI Workflow Diagram for Bibliometric Analysis

RESULTS AND DISCUSSION

The synthesis presented in Sections 2 and 3 indicates that AI meaningfully enhances bibliometric analysis by uncovering latent structures in scholarly text and citation networks that traditional frequency-based indicators cannot detect. Transformer-based NLP models consistently improve classification accuracy for citation intent and sentiment relative to earlier bag-of-words or TF-IDF baselines, while graph-based learning methods uncover hidden scholarly communities that purely textual analysis would miss. Topic-modelling approaches, particularly embedding-based methods such as BERTopic, provide a more semantically coherent picture of thematic evolution than earlier probabilistic techniques, enabling earlier detection of emerging research fronts.

At the same time, the reviewed evidence shows that these gains are neither uniform nor unconditional. Results vary considerably by disciplinary domain, reflecting differences in writing style, citation norms, and the density and quality of available metadata; models trained predominantly on computer science and biomedical corpora, as several reviewed studies note, do not transfer reliably to the humanities or regional-language scholarship. The interpretability of deep-learning systems remains a persistent challenge, limiting their adoption within formal research-evaluation frameworks where evaluators and administrators require justifiable, auditable reasoning behind any classification decision that could influence funding, promotion, or ranking outcomes. The finding by Thelwall and Kurt (2025) that large language models exhibit demographic and stylistic biases when used for evaluative tasks reinforces this concern, suggesting that the risk of embedding systemic bias into automated research assessment is not merely theoretical but has already been empirically documented in adjacent evaluative contexts.

A further pattern emerging from the discussion concerns the gap between state-of-the-art research and practical deployment. While GNN-based and transformer-based methods dominate the recent methodological literature, tools such as VOS viewer, which rely on comparatively simple network-construction techniques, remain the *de facto* standard among practising bibliometricians. This gap suggests that the translation of AI advances into everyday bibliometric practice is uneven, constrained by computational cost, data-access barriers, and the technical expertise required to deploy and interpret transformer-based or graph-based systems responsibly. Institutions with limited computing infrastructure, particularly in the Global South, risk being further marginalized in this transition unless lighter-weight, well-documented alternatives are prioritized alongside cutting-edge architectures.

The discussion also surfaces a reflexive dimension unique to this domain: AI is increasingly both an object of bibliometric study, as illustrated by mapping exercises such as that of Sekaki, Ziane, and Khazzar (2025), and a methodological tool used to conduct bibliometric analysis itself. This duality has practical implications for how the field defines objectivity and validity, since the same techniques used to evaluate AI research output are increasingly the techniques through which that evaluation is performed, creating a feedback loop that has not yet been fully theorized in the existing literature. Taken together, these results support the central argument of this paper: that AI offers substantial analytical benefits for bibliometric research, but that realizing these benefits responsibly requires the explicit ethical, technical, and computational safeguards developed in the following section, rather than uncritical adoption of automated indicators as replacements for expert judgment.

Ethical, Technical, and Computational Considerations

Ethical Considerations

1. Algorithmic Bias and Skewed Knowledge Representation

AI models trained on historical citation datasets risk inheriting and amplifying existing systemic biases present in the underlying scholarly record. Disciplines with high publication volume, such as biomedical sciences and computer science, tend to dominate the training corpora used to build citation-intent and topic-modelling systems, producing biased predictions that privilege these fields over comparatively smaller or slower-publishing domains such as the humanities or regional area studies. This imbalance can distort derived metrics including citation intent, topic prominence, and estimated author influence, since a model calibrated on high-volume disciplines may systematically misclassify citation patterns that are entirely normal within a smaller field but statistically rare in the training distribution. The empirical finding that ChatGPT-based evaluations vary by author demographic and field, reported by Thelwall and Kurt (2025), lends direct support to the concern that such bias is not a hypothetical risk but a documented feature of contemporary AI systems applied to evaluative tasks.

2. Reinforcement of Dominant Scholarly Voices

Traditional bibliometrics already favours highly cited, English-language scholarship produced by well-resourced Western institutions, and AI-driven systems, particularly those relying on large language models trained predominantly on English-language corpora, may inadvertently reinforce these existing

hierarchies by overfitting to well-indexed journals and widely cited authors. As a consequence, emerging researchers, scholars working in regional languages, and non-mainstream epistemological traditions risk further marginalization within automated evaluation pipelines that were not designed with linguistic or institutional diversity in mind. This dynamic raises substantive concerns about epistemic justice and equitable knowledge dissemination, since an evaluation infrastructure that systematically underrepresents non-dominant scholarly voices does not merely fail to correct existing inequities but risks actively deepening them through the appearance of objective, algorithmically generated assessment^[18].

3. Transparency, Explainability, and Accountability

Deep-learning models, including BERT-based classifiers, graph neural networks, and larger transformer architectures, function largely as black boxes, making it difficult for research evaluators to understand precisely how a given classification decision, such as labelling a citation as positive, negative, or purely methodological, was reached. This lack of explainability undermines institutional trust, a concern that becomes especially acute in high-stakes contexts such as research funding allocation, institutional rankings, and faculty tenure decisions, where the consequences of a misclassification can materially affect a researcher's career^{[9][4][24]}. Ethical deployment of AI-enhanced bibliometrics therefore requires the systematic use of explainable AI techniques capable of surfacing the features or textual spans driving a given prediction, comprehensive and transparent model documentation covering data sources, hyperparameters, and validation metrics, and sustained human oversight at every interpretive step where automated outputs feed into consequential institutional decisions, rather than treating model outputs as self-evidently authoritative.

4. Data Privacy and Responsible Use of Scholarly Texts

Although the bulk of bibliometric data is nominally public, citation contexts, institutional affiliations, and author identifiers embedded within full-text scholarly documents can nonetheless raise legitimate privacy concerns, particularly when AI systems are trained on preprints, open-review platform comments, or student thesis repositories that were not necessarily produced with large-scale automated reuse in mind. AI systems operating within this space must therefore adhere to norms of responsible data reuse, including clear provenance tracking, respect for platform-specific terms of use, and caution around aggregating personally identifying information, such as author affiliation histories, in ways that could enable unintended profiling of individual researchers beyond the scope of legitimate bibliometric evaluation.

Technical Considerations

1. Inconsistent and Incomplete Metadata

Bibliometric datasets frequently contain errors such as missing digital object identifiers, inconsistent author-name formatting, duplicate records, and varied citation formats across source databases. These inconsistencies propagate through the analytical pipeline, producing unreliable embeddings and unstable clustering outcomes when passed into transformer-based models that assume relatively clean, well-structured input text. Common manifestations of this problem include variations in journal-title abbreviations across databases, incomplete abstract fields for older or non-English publications,

inconsistent disambiguation of authors sharing similar names, and missing reference lists in certain institutional repositories. Rigorous metadata normalization and cleaning protocols are therefore essential prerequisites for any AI-based bibliometric analysis, and the absence of such preprocessing rigor represents one of the more underappreciated sources of downstream analytical error identified throughout the reviewed literature.

2. Limitations of Pre-trained Models for Domain Semantics

General-purpose language models such as BERT and RoBERTa are not always optimal for highly specialized academic domains, since they are typically pretrained on broad web and encyclopaedic corpora rather than discipline-specific scholarly language. Such models may misinterpret technical terminology particular to a given field, fail to recognize domain-specific citation practices that differ markedly between, for instance, legal and biomedical scholarship, and misread field-specific rhetorical structures that carry evaluative meaning within a discipline but appear neutral outside it. For instance, a critical citation within humanities scholarship may be lexically similar to a negative citation within biomedical literature yet function very differently in each context, since critical engagement is often a marker of scholarly seriousness in the humanities rather than a rejection of the cited work. Fine-tuning with domain-specific corpora, following the SciBERT approach of pretraining directly on scientific text, becomes necessary to achieve acceptable levels of semantic accuracy across disciplinary boundaries^{[30][3]}.

3. Interdisciplinary Variability in Citation Behaviour

Citation intent differs significantly across disciplines: legal scholarship typically cites for precedent and authority, humanities scholarship frequently cites for theoretical dialogue and critique, and the natural sciences generally cite in support of experimental or methodological claims. AI models trained predominantly on one field's citation conventions may therefore produce invalid or misleading interpretations when applied uncritically to another field's scholarly output. Cross-disciplinary benchmarking, in which classifier performance is explicitly evaluated across multiple disciplinary corpora rather than a single dominant field, is consequently critical to avoid the overstated claims of generalizability that recur throughout the surveyed citation-classification literature.

4. Reproducibility Gaps in AI Pipelines

AI-based bibliometric workflows are highly sensitive to model versions, training hyperparameters, preprocessing pipeline configuration, random seed selection, and the availability of the underlying dataset over time, since many bibliometric databases are periodically updated or revised. Without detailed documentation and, wherever institutionally feasible, open-source code release, results produced by AI-enhanced bibliometric systems cannot be reliably reproduced or independently validated by other research groups, undermining confidence in findings intended to inform formal evaluation processes. Infrastructure diversity across different hardware configurations, software library versions, and cloud computing environments further complicates exact replicability, a concern that mirrors well-documented reproducibility challenges across empirical machine-learning research more broadly.

Computational Considerations

1. High Computational Cost of Large Models

Transformer architectures, including BERT, SciBERT, and Longformer variants designed for longer scholarly documents, require substantial graphics-processing-unit or tensor-processing-unit resources for both training and inference, particularly when processing millions of citation contexts drawn from large-scale bibliometric databases. This computational burden raises meaningful cost barriers for smaller institutions and creates a structural inequity in research capability between well-resourced universities that can operate large-scale AI infrastructure and those that cannot, echoing the broader equity concerns raised in Section 5.1.2 regarding the reinforcement of dominant scholarly institutions.

2. Scalability Challenges with Large Citation Networks

Modern citation databases contain tens of millions of articles and corresponding reference links, and running topic modelling, graph neural networks, or large-scale embedding generation over networks of this magnitude demands careful engineering, including efficient batching strategies for model inference, sparse matrix representations for graph storage, distributed computing frameworks capable of parallelizing training across multiple nodes, and efficient indexing systems for rapid retrieval of citation contexts. Without these optimizations, AI-enhanced bibliometric systems risk unacceptably slow performance or outright memory overload when applied at the scale required for institution-wide or national research assessment exercises.

3. Model Drift and Dataset Versioning

Citation networks evolve continuously as new articles are published, cited, and occasionally retracted, meaning that AI models trained on an earlier snapshot of a bibliometric dataset may become outdated relatively quickly, a phenomenon commonly described as model drift. Continuous model updating strategies, together with rigorous dataset version tracking that records precisely which snapshot of a citation database informed a given analysis, are therefore essential for maintaining the analytical relevance and comparability of AI-enhanced bibliometric outputs over time, particularly in longitudinal research-assessment exercises that span multiple evaluation cycles.

Principles for Responsible AI-Enabled Bibliometrics

To ensure ethical, transparent, and sustainable use of AI within scholarly evaluation, this paper proposes seven interrelated principles for responsible practice. Fairness requires that training corpora be balanced across disciplines and languages wherever feasible, with systematic bias evaluation built into model development rather than addressed only after deployment. Accountability requires sustained human-in-the-loop evaluation for interpretive outputs and, especially, for any high-stakes decision informed by AI-derived indicators, ensuring that a human evaluator retains final responsibility for consequential judgments. Transparency requires that institutions and researchers deploying AI-enhanced bibliometric tools provide model cards, comprehensive dataset documentation, and explainability mechanisms

sufficient for external scrutiny. Reproducibility requires sharing code, parameter settings, and, where licensing permits, representative data subsets, so that findings can be independently verified. Contextual interpretation requires that AI-derived metrics never be used in isolation but always be supplemented with qualitative expert judgment attentive to disciplinary norms. Cross-disciplinary validation requires that models be tested systematically across multiple fields before their outputs are generalized to institution-wide or national assessment exercises. Finally, avoiding over-reliance on automated metrics requires that AI-generated insights be treated consistently as decision-support tools rather than as authoritative indicators capable of standing alone in high-stakes evaluative contexts. Together, these seven principles offer a practical framework through which institutions can pursue the analytical benefits of AI-enhanced bibliometrics identified throughout this paper while guarding against the ethical, technical, and computational risks documented in the preceding sections.

CONCLUSION

This paper has examined the objectives, operational limits, and ethical dimensions of integrating Artificial Intelligence into bibliometric and scientometric research. The reviewed literature demonstrates that AI, and transformer-based NLP in particular, meaningfully extends the analytical reach of bibliometrics beyond simple citation counting, enabling citation intent and sentiment classification, semantic similarity mapping across large scholarly corpora, graph-based discovery of latent scholarly communities, and predictive modelling of thematic and disciplinary trends that would remain invisible to frequency-based indicators alone. At the same time, the analysis presented across Sections 2 through 5 makes clear that these analytical gains are neither automatic nor free of risk. Algorithmic bias rooted in disciplinary imbalance within training corpora, the opacity of deep-learning decision processes, persistent gaps in reproducibility, and the substantial computational resources required by state-of-the-art architectures collectively constrain how far AI-enhanced bibliometrics can be trusted as a standalone basis for consequential decisions in funding, ranking, and tenure evaluation. Establishing clear objectives alongside explicit ethical, technical, and computational boundaries, as this paper has attempted to do, is therefore a necessary precondition for responsible integration of AI into research evaluation systems rather than an optional afterthought. The seven principles for responsible AI-enabled bibliometrics proposed in Section 5.4, spanning fairness, accountability, transparency, reproducibility, contextual interpretation, cross-disciplinary validation, and resistance to over-reliance on automated metrics, offer a practical starting point for institutions, journal publishers, and policymakers seeking to operationalize this balance. As scholarly communication continues to grow in both volume and disciplinary complexity, AI-enhanced bibliometrics will inevitably play an increasing role in shaping evidence-based academic assessment; the central argument of this paper is that this role should be shaped deliberately, with sustained attention to fairness, transparency, and human oversight, rather than allowed to expand through the unexamined adoption of increasingly powerful but insufficiently understood computational tools. Future work should pursue empirical validation of the conceptual framework proposed here across multiple disciplinary corpora, alongside the development of open, well-documented benchmark datasets capable of supporting the cross-disciplinary and reproducible evaluation practices this paper has argued to be essential.

GENERATIVE AI DECLARATION

The author(s) confirm that there was no use of artificial intelligence (AI)-assisted technology for assisting

in the writing or editing of the manuscript, and no images were manipulated using AI. The authors developed and verified all scientific content, data analysis, interpretation of results, and conclusions. The authors take full responsibility for the accuracy, integrity, and originality of the work presented, and no AI tool was listed as an author.

REFERENCES

1. Agarwal, S. and Mittal, R. 2021. Machine learning in scientometric analysis. *Journal of Informatics*.
2. Becker, J., Wahle, J.P., Gipp, B. and Ruas, T. 2024. Text generation: A systematic literature review of tasks, evaluation, and challenges. *arXiv preprint arXiv:2405.15604*.
3. Beltagy, I., Lo, K. and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *Proceedings of EMNLP-IJCNLP*, pp. 3615–3620.
4. Bezerra, D.A., Silva, F.N. and Amancio, D.R. 2025. Leveraging GANs for citation intent classification and its impact on citation network analysis. *arXiv preprint arXiv:2505.21162*.
5. Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**: 993–1022.
6. Bornmann, L. and Leydesdorff, L. 2014. *Scientometrics in a changing research landscape*. Springer.
7. Chen, C. 2020. *Emerging trends in citation network analysis using AI*. Scientometrics.
8. Cockburn, A., Henderson, R. and Stern, S. 2018. The impact of artificial intelligence on innovation (NBER Working Paper No. 24449). *National Bureau of Economic Research*.
9. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pp. 4171–4186.
10. Floridi, L. and Cows, J. 2019. A unified framework of five principles for AI in society. *Harvard Data Science Review*, **1**(1).
11. Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
12. Hutson, M. 2018. Artificial intelligence faces reproducibility crisis. *Science*, **359**(6377): 725–726.
13. Ioannidis, J.P.A. 2005. Why most published research findings are false. *PLoS Medicine*, **2**(8): e124.
14. Jobin, A., Ienca, M. and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, **1**(9): 389–399.
15. Kipf, T.N. and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*.
16. Liu, Y., Zhao, L., Tu, B., Wang, J., He, Y., Jiang, R., Wu, X., Wen, W. and Liu, J. 2025. Application of artificial intelligence in echocardiography from 2009 to 2024: A bibliometric analysis. *Frontiers in Medicine*.

17. Mikolov, T., Chen, K., Corrado, G. and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
18. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, **3**(2).
19. O'Connor, A.M., Clark, J., Thomas, J., Spijker, R., Kusa, W., Walker, V.R. and Bond, M. 2024. Large language models, updates, and evaluation of automation tools for systematic reviews: A summary of significant discussions at the Eighth Meeting of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews*, **13**(1): 290.
20. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P. and Moher, D. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, **372**(71).
21. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. 2019. *Language models are unsupervised multitask learners*. OpenAI Blog.
22. Reimers, N. and Gurevych, I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992.
23. Roman, M., Shahid, A., Khan, S., Koubaa, A. and Yu, L. 2021. Citation intent classification using word embedding. *IEEE Access*, **9**: 9982–9995.
24. 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.
25. Sekaki, Y., Ziane, H. and Khazzar, A. 2025. Artificial intelligence in management studies (2021–2025): A bibliometric mapping of themes, trends, and global contributions. *International Journal of Accounting, Finance, Auditing, Management and Economics*, **6**(9): 62–80.
26. Tang, X., Duan, X. and Cai, Z.G. 2025. Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition. *arXiv preprint arXiv:2412.13612*.
27. Thelwall, M. and Kurt, Z. 2025. Research evaluation with ChatGPT: Is it age, country, length, or field biased? *Scientometrics*, **130**(10): 5323–5343.
28. Van Eck, N.J. and Waltman, L. 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, **84**(2): 523–538.
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, **30**.
30. Visser, R. and Dunaiski, M. 2022. Sentiment and intent classification of in-text citations using BERT. *Proceedings of the 43rd Conference of the South African Institute of Computer Scientists and Information Technologists*, **85**: 129–145.

31. Waltman, L. 2016. A review of the literature on citation impact indicators. *Journal of Informetrics*, **10**(2): 365–391.
32. Wang, X. *et al.* 2022. *Graph neural networks in scientometrics*. Information Processing & Management.
33. Wilkinson, M.D. *et al.* 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, **3**: 160018.
34. Yan, E. and Ding, Y. 2010. Applying centrality measures to impact analysis: A co-authorship network. *Journal of Informetrics*.
35. Zhang, Y., Wang, Y., Sheng, Q.Z., Yao, L., Chen, H., Wang, K., Deng, C. and Zhao, R. 2025. Deep learning meets bibliometrics: A survey of citation function classification. *Journal of Informetrics*, **19**(1): 101608.