Majestic Multidisciplinary Science Insight
Volume 1, 2026, **pp.** 1-12, **Article ID:** CM2601107003
DOI: 10.65080/inprocess

CONTENT MAJESTIC PUBLISHER
OPEN ACCESS

Systematic Review Article

Generative AI in Academic Libraries: A Systematic Review of AI-Mediated Reference Services, Librarian Roles, and User Trust

Arsalan Sheikh[1, *]

[1]*COMSATS University, Islamabad, Pakistan*

Article History

Received: 11 January, 2026 Revised: 29 February, 2026 Accepted: 03 March, 2026 Published: 28 March, 2026

Abstract:

Generative artificial intelligence (AI), and in particular large language models in the form of ChatGPT, is gradually getting integrated into the systems and workflows of academic libraries. Consequently, they are changing the nature of reference services, the work of librarians, and the information-seeking behaviour of users. This systematic review aimed to explore the impact of generative AI on the quality of reference services, restructuring of librarian tasks, and trust of users in academic libraries. A systematic search in Scopus, and Science Direct was conducted to find the relevant literature. A total of 14 major empirical studies, published between 2020 to 2025 were selected and analysed that included experimental assessments, cross sectional survey, qualitative interviews, mixed-methods research and quasi experimental audit studies across different geographic locations. The findings revealed a significant performance of AI-mediated reference services but generative AI can only be applied to low-complexity questions. However, it demonstrated severe limitations in responding to complex research questions, contextual questions, and citation verification, as indicated by high rates of hallucinations and bibliographic errors. However, the user-trust was also conditional, depending on the perceived utility, transparency, and the institutionalisation of AI services. Thus, generative AI should be understood as a socio-technical system, in which the efficacy is determined by human control, governance systems, and correlation with the fundamental values, which are the foundation of academic libraries. Thus, this study emphasize instead of replacing librarians, generative AI restructured professional practice in favour of mediation, verification, ethical oversight, and the delivery of AI literacy education.

Keywords: Generative artificial intelligence, academic libraries, reference services, librarian roles, user trust, information integrity, chatgpt.

1. INTRODUCTION

Chatbots based on generative artificial intelligence (GenAI), specifically, large language models (LLMs), are swiftly transforming academic communities. It helps to discover information, evaluate evidence, and obtain help with conducting research and composing scholarly writing. Student uptake is now high enough to affect the demand profile for academic library services. The UK Higher Education Policy Institute (HEPI) (2025) student Generative AI survey 2025 reports that 88% of students reported using generative AI tools for assignments (up from 53% the previous year). Although GenAI is becoming part of the daily routines of academic research, academic libraries are under ever-increasing pressure to act across three interconnected areas: maintaining the quality of reference services, recalibrating labour roles and skills, and preserving users' trust. Early experimental findings indicate that GenAI can offer daily advice but not the highest-quality reference results. In chronologically organised testing of ChatGPT across different types and complexity levels of reference queries, Lai (2023) found that the system performed poorly on advanced research questions and complex inquiries, including precisely the kinds of queries that characterise academic library reference work. One of the main weaknesses is the model's tendency to produce plausible but incorrect statements, such as bibliographic errors. A socio technological perspective in library practise acknowledges that AI results are governed by organisational settings, staff processes, professionalism, and policy alignment within an institution. Empirical studies across disciplines by Mugaanyi et al. (2024) were cross-disciplinary and average recall had correct citations of outputs only approximately 60‘%. In the case of libraries, these findings present a threat whereby, provided the use of generative AI to answer reference queries, recommend sources, or teach citation practise, there is a possibility of such uncorrected inaccuracies. These inaccuracies may lead to poor service provision and loss of confidence in library directions as the users make mistakes.

Simultaneously, there is also more effort to apply generative AI to the organisational level. According to the evidence presented by academic libraries, generative AI is being implemented in service delivery, user care, and internal processes. Indicatively, Gmiterek and Kotuwl (2025) used a

*Address correspondence to this author at COMSATS University, Islamabad, Pakistan; E-mail: arslan_sheikh@comsats.edu.pk

survey and content analysis to identify the degree of generative-AI adoption by academic libraries at Polish state universities (a mixed-methods study). The workforce implication is also evident as Cox (2025) finds out that the survey on AI and the library profession in UK across the country, that the current use and perception of AI by librarians and information professionals is listed on the current positions of issues of concern rather than being an issue of the future the survey on AI and the library profession UK wide. User trust is often underestimated but is the third variable indicating the connexion between the quality of references and the workforce change. As Deschênes and McMahon (2024) reported on the use of generative-AI chatbots in academic research, students use it, but they do not consider the information produced by generative AI helpful according to the data of the surveys. As Grammes (2024) notes, perceived usability advantages are in a better position than human reference interactions. However, the comfort factor will also tend to affect trust and preference where there are problems of credibility.

Although there is an increased amount of empirical studies of generative AI in academic libraries, the research remains sporadic, which is why a specific systematic review is necessary. The literature at hand analyses the independent factors of the AI usage. It, however, does not centre on the joint impact on library services. Indicatively, correctness of answers and validity of reference are the major measures of reference quality as applied by Lai (2023). However, these outcomes are not explicitly related to these measures of performance that are so critical to workforce redesign like models of supervision, timely mediation, workflows of quality-assurance retribution, or objective markers concerning user trust. Cox (2025) concentrated on professional expectations and views under a sample survey of librarians, however, the study lacked comprehensive data on workforce perception, measured performance of services, and patron confidence in academic libraries. Deschênes and McMahon (2024) documented patterns of using and attitude toward generative AI but did not compare differences in trust between AI use within library-branded services and independent use of external AI tools by students.

The given review fills these gaps by incorporating the available evidence on both sides of the service-workforce-trust spectrum in the academic library, and mainly focusing on generative AI-based reference and research support. This systematic review aims to summarise recent empirical evidence on the effects of generative AI on academic library services and professional practise. Notably, it explores implications on the quality of reference services such as accuracy, completeness, contextual relevance, and citation integrity, changes in librarian roles, competencies, and trends in user trust. The review will help illuminate evidence-based governance and the responsible incorporation of generative AI in academic libraries by providing comprehensive evidence.

2. METHODS

2.1. Research Design

The current study used a systematic literature review (SLR) design framework to analyse the effects of generative artificial intelligence (GenAI) on the practice of providing reference services in academic libraries and the workforce, in particular, the quality of reference services, the role of staff, and user trust. The systematic literature review provides a comprehensive understanding of the issue by synthesising available evidence (Lame, 2019). This was a review in compliance with the established PRISMA principles of evidence-based information science protocols employed to minimise selection bias and provide methodological transparency.

2.2. Search Strategy

The search strategy was aimed at ensuring that the best possible coverage of empirical evidence is given and at the same time ensure transparency, reproducibility and relevance to library and information science. The available evidence was found in two databases, one of which was Scopus and the other was ScienceDirect. These databases were chosen to provide balance between the depth of discipline and coverage of citation and the access to both the LIS-specific and interdisciplinary studies. The selection of these databases was based on their comprehensive coverage of library and information science, information systems, and higher education research, as well as their exceptionally high indexing of peer-reviewed research journals. The searches of publications were restricted to from 2020 to 2025 in the topical area, as large language models have advanced rapidly and require the inclusion of reputable publications in the search field. This temporal boundary ensures that findings reflect the post-ChatGPT phase of generative AI development. Appropriate keywords with Boolean operators were used to determine available evidence. Various search terms were used in all the databases that were chosen. Scopus included (ChatGPT OR ‘AI chatbot*’ OR ‘generative AI’) AND (‘reference librarian*’ OR ‘human reference’ OR ‘traditional reference’) AND (comparison OR comparative OR evaluation OR performance), leveraging Scopus’s structured indexing and quality control for comparative empirical studies. ScienceDirect included (‘generative AI’ OR ChatGPT OR ‘AI-assisted system’) AND (‘librarian’ OR ‘job role’). Results of the search of the databases are described in Appendix A. To guarantee a consistent interpretation and a similar methodological appraisal, only English publications were taken into account. It was found that duplicate records were being used and as a result, they were eliminated before the screening stage.

2.3. Study Selection Criteria

2.3.1 Inclusion Criteria

The studies that met a set of clearly specified eligibility conditions were considered in the review process. The studies had to be eligible and contain primary empirical results of either of the qualitative, mixed-methods, or quantitative research designs. Studies recorded that explicitly interacted with generative AI or large language model-based systems (ChatGPT or GenAI). The setting of the research also was to be a context that lay in academic libraries, academic information services or closely related, highly academic settings. Studies reported at least one of the core outcomes, reference or answer quality, citation accuracy, librarian or staff roles and competencies, or user trust and credibility were included. Only

studies published between 2020 and 2025 in peer-reviewed journals were included to ensure both rigour and relevance.

2.3.2 Exclusion Criteria

Review articles, systematic reviews, conceptual papers, editorials, opinion pieces, and policy commentaries were excluded. Pure technical AI studies that did not have a direct relationship with the services of an academic library or the information work in general were excluded. Research studies that did not exhibit methodological transparency, empirical evidence, or connection to library services or workforce implications were excluded. To maintain a contextually and analytically consistent research process, studies based on K-12 education (education from kindergarten till 12th grade) alone and those based on corporate knowledge management or non-academic public library environments were excluded.

2.4. Data Extraction

A data-extraction framework was created to enable comparability between the studies that were included. The information extracted included bibliographic information (author, year, country), study design, sample characteristics, library service context, and outcome variables. Special care was given to reference quality (*e.g.,* the accuracy, completeness, citation reliability), workforce (*e.g.,* redistribution of work, skills demanded, professional identity), and user trust (*e.g.,* perceived credibility, reliance, willingness to trust AI-mediated services) indicators. Critical findings were documented in a systematic way.

2.5. Data Synthesis and Analysis

The diversity of the study designs and outcome measures made it necessary to use the narrative and thematic synthesis methodologies. The primary categorization of studies was on the focus of analysis: quality of references, assessments on the workforce, or reliability to the users. The identification of cross-cutting themes later, allowed the analysis of cross-domain interactions (e.g., mediation of workforce and trust, or reference accuracy and professional role redesign). The quality of methodology was evaluated to determine the robustness of the evidence base and to establish research gaps that are long term.

2.6. Ethical Considerations

The proposed study is limited to publicly available secondary data. Therefore, no ethical approval was necessary. The credibility of the included studies was ensured by providing correct citations, documenting all procedures, and reporting the true findings. There was no interpretive bias, as the inclusion criteria were adhered to and analytical procedures were followed systematically throughout the review.

3. RESULTS

3.1. Data Screening

The selection of studies was conducted in accordance with the PRISMA framework to ensure transparency, consistency, and methodological rigour across all screening steps (Fig **1**), PRISMA. A preliminary search of two major bibliographic databases, Scopus (n = 257) and ScienceDirect (n = 320), yielded 577 results. Before formal screening, 40 duplicate records were removed. After this, 537 records were screened based on title and abstract relevancy. At this point, 12 non-English language publications and 120 articles not specifically relevant to the topic were excluded. Afterwards, 285 peer-reviewed journal articles, including books, editorial papers, conference papers, and records without a particular focus on the use of generative AI in academic library services, were excluded. 60 doctoral dissertations, secondary review sources, and case studies were excluded. Afterwards, 9 records that were unavailable because of the absence of open access or institutional archives restrictions were also excluded. Unavailability refers to the unavailable full-text articles that either were not available through the open access or institutional repositories at the time of screening. This limitation can lead to biasness on available journals which can philtre away relevant but restricted studies. In the end, the final systematic review included fourteen studies that met all the eligibility criteria, which represents a highly focused and strictly elaborated evidence base. Full-text retrieval outcomes and exclusion reasons are briefly listed in Appendix **A**.

3.2. Quality Assessment

Overall, the quality appraisal indicates that the included studies are of moderate to high methodological quality, with strengths such as clarity of objectives, suitability of the design, and reporting of results, and recurrent limitations. The research questions of the cross-sectional and descriptive studies by Choudhury and Shamszare (2023), Deschênes and McMahon (2024), Haris et al. (2025), Deschênes and McMahon (2024), Ismail et al. (2024), and the descriptive research studies of Del Castillo and Kelly (2024), as well as Elsayed and Abusharhah (2025), were appraised in the Critical App All of these studies were deemed to generate sound and clear findings, thus making them relevant to academic library environments (see Appendix C). Nevertheless, recruitment approaches were often rated as being unclear, which was also due to the prevalence of convenience, volunteer, or snowball sampling methods and thereby representativeness. Chigwada and Pasipamire's (2024) study also showed that there was a deficiency in the size of the sample and the depth of its analysis, which also means that the risks of sampling bias are high.

The appraised qualitative studies by Chen (2025) and Kim (2025) that were evaluated with the help of the CASP Qualitative Checklist had high levels of coherence in their methods, had a clear purpose, proper qualitative designs, thorough thematic analysis, and clear statements of findings. The main weakness of these studies was that there was no explicit discussion on the issue of researcher-participant relationships or reflexivity, which limited the overall evaluation of the possible interpretive bias. Most of the quality criteria for the qualitative, quantitative, and integrative elements were met in the mixed-methods study by Saeidnia et al. (2024), as assessed using the MMAT (2018). However, the representativeness of the participants and the control of confounding factors were limitations that reduced generalisability. The quasi-experimental study by Wang et al.

(2025), which was assessed using the JBI checklist, demonstrated high internal validity, efficient outcome measurement, and appropriate statistical analysis. The unclear rating pertained only to the pre- and post-measurement and follow-up criteria, which were not the focus of the audit-based study. Hence, the quality analysis justifies the inclusion of all studies but emphasises the need for careful interpretation, especially regarding generalisation and long-term institutional impact.

Importantly, no studies were excluded based on quality appraisal criteria, as all met the lowest methodological acceptability standards established by CASP, MMAT, or JBI. In turn, quality assessment was used to give greater weight to the interpretation of findings rather than to establishing eligibility. Experimental and quasi-experimental studies were accorded greater evidentiary weight in assessing reference accuracy and bias than survey-based and qualitative studies, which were more cautious in extrapolating to workforce preparedness or user confidence when the study context was not replicated. The reference to limitations identified, especially the non-probability sampling method, representativeness, and the unavailability of longitudinal designs, helped to understand generalisability constraints. To that end, the synthesis of results focuses on convergent patterns across methods and geographical locations, rather than on single-study influences, so that the conclusion reflects the strengths and consistency of the relationships within the evidence base rather than a single high-impact report. Quality assessments for each study, as per their designated appraisal tool, along with an overall summary of quality assessments, are provided in Appendices B and C.

3.3. Study Characteristics

The articles included in this review have a diverse methodology and analyse the field of generative AI and its use in the environment of academic library service, workforce practise, and user trust. The study designs of the fourteen major researches include experiment studies, big surveys, mixed-method studies, and the case study of qualitative analysis, thus demonstrating the complexity of using AI in libraries (see Table 1). The quality of references and reliability are also the subject of a significant percentage of the studies, especially through comparative indicators of the performance of ChatGPT compared to its actual or simulated reference queries. Empirical evidence shows that simple, factual, and navigational queries can be sufficiently met with generative AI, but systematic weaknesses with complex academic sources (advanced research queries, citation generations) are only identified by Lai (2023) and Mugaanyi et al. (2024).

The articles incorporated show the high professional interest but show a gap in the preparedness of librarians. In various parts of the world, the studies by Gmiterek and Kotuawa (2025), Elsayed (2025), Abusharhah (2025), Hussain (2024), and Kim (2025) through the case studies and surveys point at significant optimism on the prospects of AI. However, there are still long-term obstacles, such as the lack of skills, the shortage of governance, and the need to be ready to handle ethical issues. Together, such works highlight generative AI as a disruptive phenomenon, which alters the role of a librarian towards mediation, training AI literacy, quality control, and ethical accountability instead of eliminating it. The paradox of usage-trust observed in user-related studies that Deschênes and McMahon (2024) and Choudhury and Shamszare (2023) indicate may be due to high adoption rates, but low trust in AI-generated information on the quality of the output, transparency, and source credibility is also relevant. Interestingly, the experimental results used by Wang et al. (2025) are quite enticing because the findings can be generalised to the idea that contemporary large-scale language models can provide consistent reference answers to various demographic groups, which implies a means to reduce prejudice in the course of time through proper controls and model selection. The characteristics of the present study support the idea that the impact of generative AI in academic libraries is not a categorical phenomenon and depends on the human control versus the ability to provide services and the services based on the trust.

4. FINDINGS

The results section is explicitly structured around a socio-technical integration model in which reference service quality outcomes shape user trust and, in turn, necessitate workforce mediation and governance responses. This model provides the analytical framework through which empirical findings are organised, interpreted, and synthesised across studies, ensuring that quality, labour, and trust are examined as interdependent rather than isolated dimensions.

4.1. Scope and Distribution of the Evidence Base

The review is anchored on fifteen main empirical studies, published in 2023-2025, which implies that the field of generative AI is rapidly becoming a mature and scholarly topic in academic libraries. The level of methodological difference is obvious: Lai (2023), Mugaanyi et al. (2024), and Wang et al. (2025) evaluate AI reference through experiments; large-scale quantitative surveys are represented by Choudhury and Shamszare (2023), Haris et al. (2025), and Elsayed et al. (2025); qualitative and interview-based Robertly by Chen (2025) and Kim (2025); and mixed-method designs are offered by combination These have been found on North America, Europe, Asia, Middle East and the Global South, which appears to make them have a global applicability rather than region-specific experimentation. Thematic areas of the research include (1) the quality and accuracy of references, (2) the roles and competencies required in the workforce, and (3) the trust of users and their adoption behaviour. It is worth noting that the volume of publications surged afterwards, after the widespread introduction of ChatGPT, thus creating an interdependent socio-technical evidence base, where experimental research identifies performance constraints, survey research identifies expectations and attitudes, and qualitative research clarifies how workforce mediation is generated under the threat of trust-sensitive reference risk.

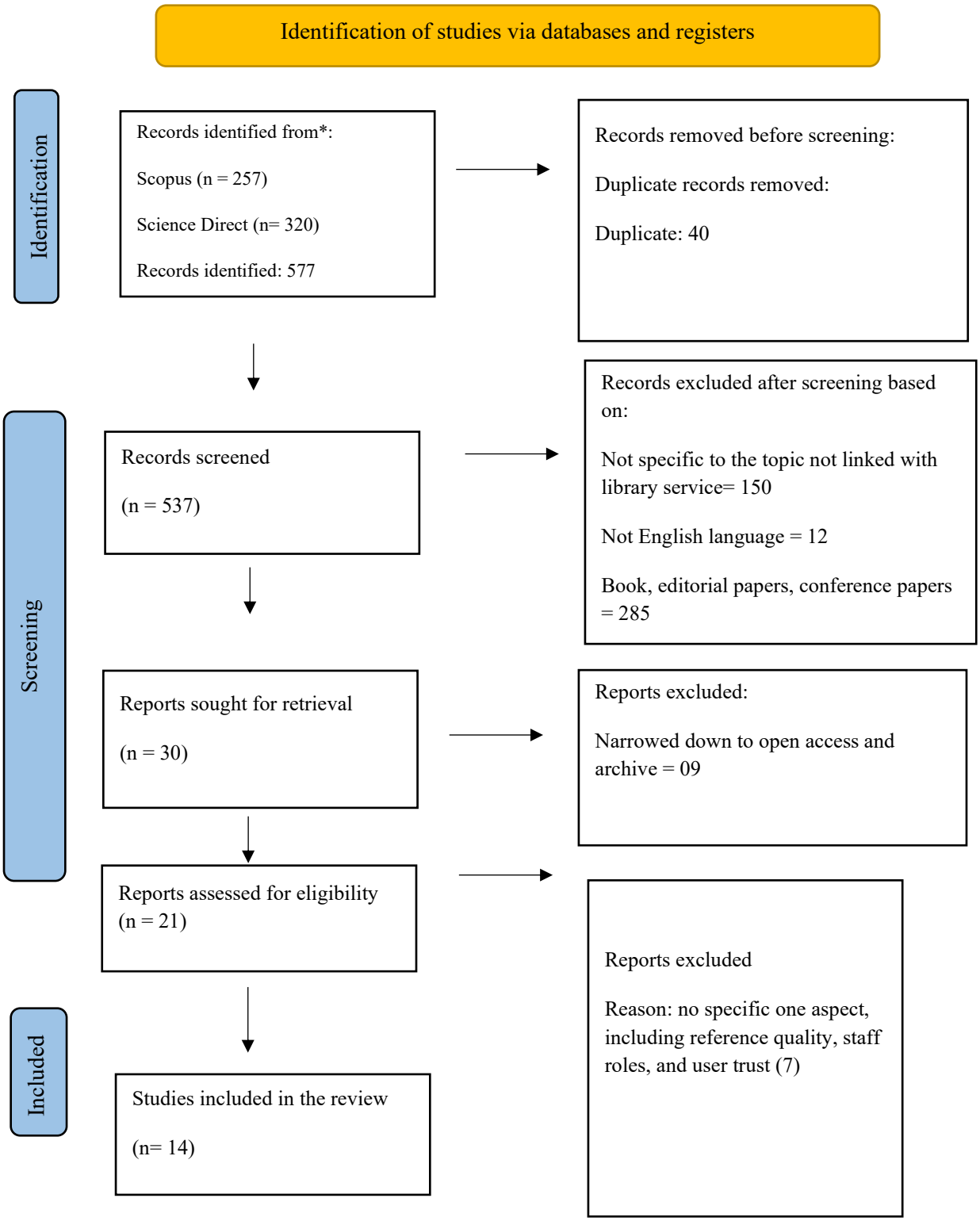

Fig. (1): PRISMA flowchart.

4.2 Impact of Generative AI on Reference Quality and Information Integrity

In both the experimental and evaluative literature, a clear performance gradient is evident in generative AI-mediated reference services, and effectiveness depends heavily on question complexity. Experimental research using controlled testing with absolute reference questions shows that ChatGPT performs relatively well on low-complexity queries (directional, facilities-related, policy, and simple factual). In an experimental assessment of 58 real academic library reference queries (Lai 2023), the data indicate that accuracy was highest for low-READ-scale queries (See Table 2). In contrast, in a qualitative study by Chen (2025), all librarians agreed that low-READ-scale queries could be effectively handled by AI, thereby reducing the paperwork they had to manage.

In comparison, high-complexity reference tasks cause a steep decline in performance. Lai (2023) notes that ChatGPT was rated lowest in higher research queries, known-item searches, and access to e-resources, particularly when a specific institutional context or disciplinary complexity was required. This fact is supported by Saeidnia et al. (2024), who found that ChatGPT had a higher overall score for non-specialised information needs (Mean 3.77/5) than for specialised or expert-level enquiries (Mean 3.13/5) as epistemic requirements increased. The librarians interviewed in Chen (2025) also reported being reluctant to use AI in high-stakes reference interactions due to the risk of misinterpretation and partial contextual reasoning. The experimental methods prove that retrieval accuracy decreases in proportion to the complexity of a query, and the qualitative data of practising librarians supports this notion, indicating that the compared performance deficiencies are due to structural limitations and not temporary conditions. These quality limitations directly increase reliance on librarian mediation and shape users' scepticism toward AI-mediated reference outputs.

Alongside the superficial quality of the answer, information integrity failures pose the most significant threat to AI generation in the academic reference domain. There is also quantitative experimental evidence of high percentages of citation hallucination and bibliographic inaccuracy. Mugaanyi et al. (2024) found that 32.7% of DOIs in the natural sciences and 8.5% in the humanities were correct citations, and that hallucinations in humanities outputs reached up to 89.4% (See Table 3). Such a difference in disciplines shows that generative AI is weakest in areas where citation conventions are heterogeneous and context-specific. These integrity failures provide a causal bridge between experimental evidence of citation inaccuracy and survey-based findings of user scepticism, linking reference quality directly to trust erosion and governance risk.

These results build upon previous issues raised in reference-service assessments. Lai (2023) also identified fabricated or misleading references in responses to research-oriented questions, but Chen (2025) states that librarians cited hallucinations as one of the primary reasons they should not use advanced AI in reference services. Combined, these studies demonstrate that plausibility and not verifiability are used to control the work of AI- another epistemic problem that the standards of academic libraries cannot accept. In short, experimental studies reveal that gains in speed and accessibility are achieved at the cost of epistemic reliability, a trade-off that survey-based optimism alone fails to capture. Despite the fact that generative AI helps to optimise the process and make it more user-friendly, it also threatens the principles of verification according to which scientific librarianship is established. Deschênes and McMahon (2024) disagree with the idea of unsupervised AI-based reference services, especially those managed by users who do not possess the necessary experience to identify the flaws. As a result, the necessity of the human validation layers can be found in the literature. Instead of doing away with reference services, Chen (2025) and Kim (2025) suggest that AI might operate as a first-line or auxiliary service, but that librarians will not only have to maintain the quality, contextual relevance, and moral accountability.

4.3 Transformation of Librarian Roles and Workforce Competencies

Generative AI redeploys librarian labour not as a proactive innovation strategy, but as a compensatory response to documented weaknesses in reference quality and information integrity. According to the interview-based results provided by Chen (2025), 100% of the reference librarians' respondents (n=25) stressed the importance of human validation when using ChatGPT, especially for complex or high-stakes queries, which would make librarians central filters of quality rather than dispassionate intermediaries. This job increase can be further supported by the embedded case study by Kim (2025), which reports that staff reported an overt redistribution of the labour force away from repetitive question answering to knowledge curation, AI oversight, and support for digital literacy, with professional scope enhancements apparent rather than deskilling. But the workforce's preparedness is skewed. According to Gmiterek and Kotuwa (2025), only 46.4% of Polish academic libraries today use generative AI but only 7.1 percent have official guidance on AI, and 77.8% cite the lack of employee competence as the most significant challenge. While survey studies report strong professional optimism, they simultaneously expose governance and skills gaps that limit safe operationalisation, revealing a disconnect between aspiration and institutional readiness. Elsayed and Abusharhah (2025) found that 81% of Arab academic librarians self-reported ethical risks, yet 12% reported direct or indirect experience or encounters with AI-related ethical dilemmas, indicating a disconnect between Self-perceived responsibility and operational readiness (See Figure. 2). The signs of misalignment can be seen at an early stage in the career ladder; according to Chigwada and Pasipamire (2024), over 70% of LIS students mention repeat misinformation, bias, and ethical uncertainty despite the heavy use of ChatGPT, which indicates that early adoption moves up the career ladder faster than formal teaching on AI literacy can do. These workforce pressures emerge directly from trust-sensitive reference risks, illustrating how declining epistemic reliability necessitates increased human mediation rather than automation.

One weakness identified is the lack of formal AI governance structures. Gmiterek and Kotuwa (2025) also note that there are few libraries that document AI policies, which leads to disjointed, uncoordinated, and spontaneous experiments. This policy vacuum undermines the continuity of services, professional trust, and accountability, particularly when AI products are incorporated into academic scholarship.

4.4 User Trust, Perceived Credibility, and Adoption Behaviour

Empirical research involving end-users reveals a strong use-trust paradox rooted in uneven reference quality and limited transparency of AI outputs. Deschênes and McMahon (2024) state that about 64-65% of students had viewed or planned to use generative AI in academic work. Still, an 65% had found AI-generated work untrustworthy enough to use in academia. The fear of information integrity was particularly high, as 88.6 % of respondents reported fear of fake information, and 83.1 % reported unclear or unreliable sources. Thus, cross-validation with references to trusted data sources or human oversight was a regular practice among many users. This behaviour reflects pragmatic dependency driven by convenience, rather than epistemic confidence grounded in verified accuracy. However, trust is a critical factor in the level of adoption. The factor of trust is a decisive level of adoption. Equation-modelling, as conducted by Choudhury and Shamszare (2023) on 607 users indicates that the trust produces a significant positive effect on intention to utilise AI (β = 0.711) and a direct effect on actual utilisation (β = 0.302); the actual utilisation mediates an intent. These models show that trust is mediating the connexion between the observed reference performance and the maintenance in the reliance used by a user instead of existing independently as an attitudinal variable. Although audit-based studies mitigate concerns about demographic bias, they do not address the dominant trust barrier identified in experimental studies: accuracy and verifiability. The diffusion-oriented evidence also provides further context in terms of the dynamic as Haris et al. (2025) in their survey of 383 users of academic libraries have found a high average score in optimism about the potential use of AI in enhancing library services (3.99/5) and no differences by gender or age, suggesting that the perception of usefulness, rather than demographics, drives the path of adoption (See Table 4 and Figure 3).

Table 1: Characteristics table of selected Studies

Author (Year)	Context	Method	What Was Tested	Key Quantified Findings	Core Outcome
Chen (2025)	Taiwan university libraries	25 librarian interviews	ChatGPT for reference services	100% required human oversight; effective only for simple queries	ChatGPT is assistive, not autonomous
Chigwada and Pasipamire (2024)	LIS students (Zimbabwe)	Survey (n=59)	ChatGPT for academic work	>70% demanded AI-literacy training	Users rely on AI despite ethical & accuracy risks
Choudhury and Shamszare (2023)	US ChatGPT users	SEM survey (n=607)	Trust → Use	Trust → Intention (β=.71); Trust → Use (β=.30)	Trust drives adoption
Del Castillo and Kelly (2024)	US library instructors	TAM survey	ChatGPT in IL teaching	Usefulness ↑ adoption; trust concerns ↓	ChatGPT needs pedagogical mediation
Deschênes and McMahon (2024)	Harvard students	Survey (n=360)	ChatGPT for research	64% use; 66% do not trust outputs	High use, low credibility
Elsayed and Abusharhah (2025)	Arab university libraries	Survey (n=272)	AI in library services	37.5% use AI; only 12% met ethical issues	Low readiness, weak governance
Gmiterek and Kotuła (2025)	Polish university libraries	Mixed-methods	GAI in libraries	46% use GAI; 7% have policies	Adoption without regulation
Hussain (2024)	Pakistan libraries	Survey (n=150)	AI literacy & readiness	AI interest = 4.61/5; weak infrastructure	Motivation > capacity
Lai (2023)	Academic music library	Experimental	ChatGPT accuracy	Performs poorly on complex queries	Unsafe for high-level reference
Mugaanyi et al. (2024)	Academic citations	Experimental	ChatGPT citations	Only 8–33% DOIs accurate	High hallucination risk
Saeidnia et al. (2024)	Info-seeking users	Mixed-methods	Trust & usefulness	Experts trust less (M=3.13)	Trust drops with expertise
Wang et al. (2025)	Virtual reference	Audit of 6 LLMs	Bias & fairness	No racial bias; minor gender bias	LLMs can be equitable with governance

Table 2. The Average Quality of ChatGPT's Answers Based on Question Complexity Using the READ Scale; Source: Lai (2023).

The Average Quality of ChatGPT's Answers Based on Question Complexity Using the READ Scale

Question Complexity	Quality			Overall Average Quality
	Completeness	Accuracy	Further Assistance	
READ level 1 (n=3)	3.00	2.67	3.00	2.89
READ level 2 (n=25)	2.44	1.76	1.92	2.04
READ level 3 (n=16)	2.50	1.81	1.69	2.00
READ level 4 (n=12)	2.50	1.67	1.92	2.03
READ level 5 (n=2)	3.00	1.50	2.00	2.17
Overall	2.52	1.79	1.91	2.07

Table 3. Data analysis results; Source: Mugaanyi et al. (2024).

Variables	Natural sciences (n=55)	Humanities (n=47)	*P* value[a]
Citation exists, n (%)	40 (72.7)	36 (76.6)	.42
Citation accurate, n (%)	37 (67.3)	29 (61.7)	.35
Relevant, n (%)	39 (70.9)	35 (74.5)	.43
DOI[b] exists, n (%)	39 (70.9)	18 (38.3)	.001
DOI accurate, n (%)	18 (32.7)	4 (8.5)	.003
DOI hallucination, n (%)	34 (61.8)	42 (89.4)	.001
Levenshtein distance, mean (SD)	64.13 (42.26)	42.15 (40.23)	.009

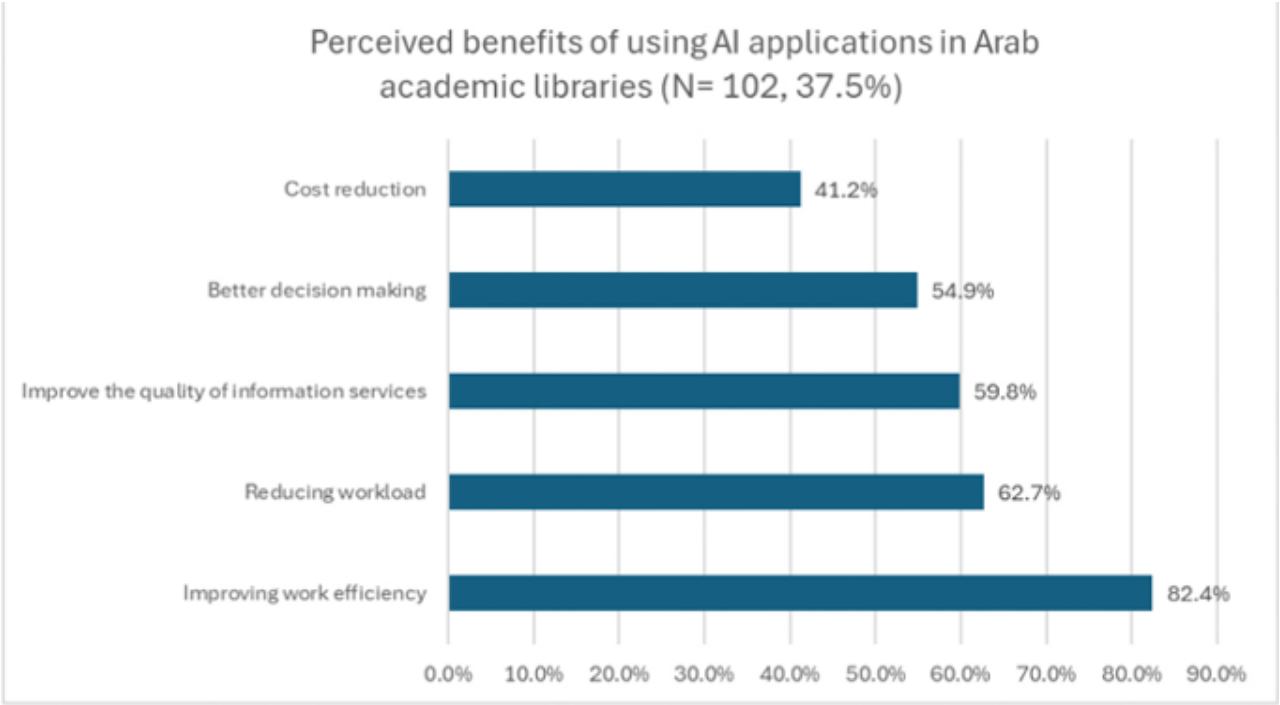

Fig. (2): Perceived benefits of using AI applications in Arab academic libraries; Source: Elsayed and Abusharhah (2025)

Table 4: Respondents' knowledge and optimism about artificial intelligence; Source: Haris et al. (2025).

Respondents' knowledge and optimism about AI		
Topic	Mean	SD
1. Conceptual knowledge of AI	3.44	1.45
2. Knowledge of trends in AI	3.38	1.45
3. An optimistic perspective on the potential of AI to enhance library services	3.99	1.56

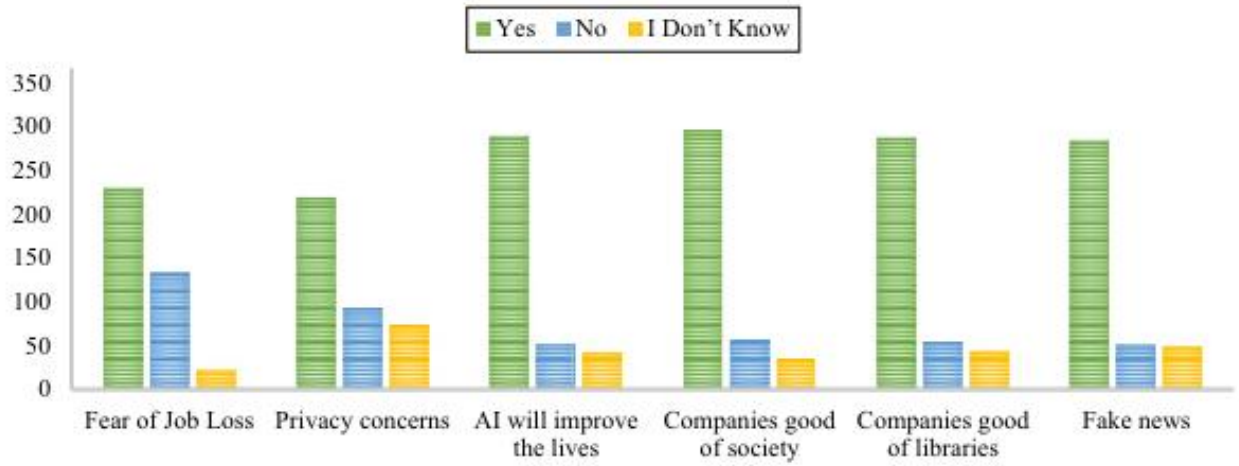

Fig. (3): Respondents' concern regarding artificial intelligence; Source: Haris et al. (2025).

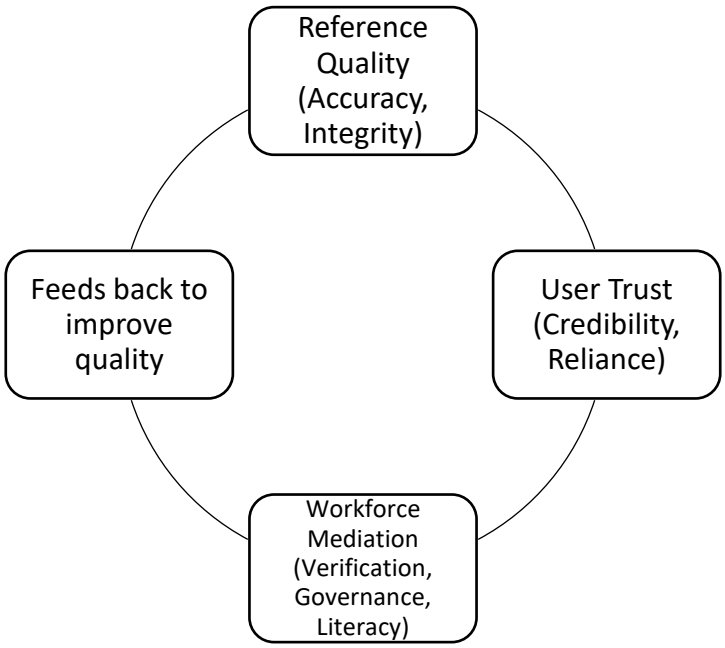

Fig. (4). Socio-technical integration model linking reference quality, workforce mediation, and user trust in AI-enabled academic libraries.

High levels of audit data represented a huge relief against the fear of social bias. As an example, Wang et al. (2025) analysed more than 2,000 reference interactions per model in six large language models (LLMs) and found no statistically significant racial or ethnic bias and slight traces of gender-based stereotypes in one of the models. The apparent linguistic variation was associated with the institutional functions, and professionals explained the variation in an institutional manner as natural accommodation instead of discrimination. Although these results preserve some of the concerns on fairness, they motor out the unresolved issues on the matters of accuracy and verifications. In turn, the data define trust as an essential moderating factor that defines the manner in which users will engage with AI despite significant constraints.

4.5 Integrative Synthesis: Interdependence of Quality, Workforce, and Trust

Generative AI, in turn, can be considered a socio-technical institution where failures observed in the reference during the experimentation process will discredit the user, leading to the need to mediate and indirect the workforce to the necessary responses. Losses in reference accuracy undermine trust, which leads to greater librarian inquiry and more effective mediation stabilises trust and is allowed to adopt selectively. Altogether, the application of generative AI is both realistic and impractical as a source of academic information when used in a non-monitored mode. Librarians thus remain as validators, educators as well as moral gatekeepers. Transparency, accuracy and institutional framing is a source of user trust, thus requiring that sustainable integration harmonises AI application with professional values, governance models, and evidence-based standards of service (see Figure 4).

5. DISCUSSION

This systematic review provides a synthesised and critical insight into the transformation of academic library services, professional functions, and user confidence brought about by generative AI, specifically large language models like ChatGPT. Lai (2023), Chen (2025), and Kim (2025) show that the effect of generative AI is neither consistently positive nor inherently disruptive. It depends on the task's complexity, the organisation, and the role of humans, which reiterates broader socio-technical views in information science. As stated by Lai (2023), Saeidnia et al. (2024), and Mugaanyi et al. (2024), generative AI is reliable with low-complexity reference queries; however, it cannot meet the scholarly requirements of complex, research-focused assignments, especially those that require contextual sensitivity, institutional expertise, or reference search. These align with Bawden and Robinson (2022), who identified that automated tools are better at performing surface-level retrieval but struggle with the interpretive and evaluative aspects of information work. The citation DOI errors are substantial in the literature review. Ji et al. (2023) findings are also aligned with this review. They identified that hallucination is a natural property of probabilistic language models and is not introduced by implementation failure. These findings contradict techno-optimism storeys according to which scale is the solution to quality issues. Instead, they support the argument raised by Floridi et al. (2018), to the effect that epistemic reliability in information systems relies on human-in-the-loop validation, particularly where authority, accountability, and scholarly credibility are demanded. The implementation of AI in academic libraries, where reference services often serve as epistemically gatekeeping systems, risks undermining the fundamental professional and institutional values.

In addition, the evidence implies that, the professional labour has to be reorganised, and the role of a librarian will be queued as the agent of AI, validator and moral guardian instead of being replaced. This is consistent with the theory of skilled jurisdiction as developed by Abbott (1988) that is the response of professions to the change in technologies to renegotiate tasks but not to delegate power. The transition of transactional reference work to oversight, AI literacy education, and governance is in line with the trends outlined by Cox et al. (2019). These authors also observe that there is a need to study other aspects of the digital transformation like the development of discovery systems and algorithmic recommender systems.However, Gmiterek and Kotuła (2025) and Elsayed and Abusharhah (2025) indicate a chronic capacity and governance gap. Despite the high interest and perceived benefits, most institutions lack formal AI policies, and staff skills are not balanced. This aligns with other criticisms of the effects of policy lag in educational technologies, as noted by Rafiq-uz-Zaman (2025), who argues that a new resource is adopted without the creation of regulatory and ethical protections. The lack of well-developed institutional frameworks places an excessive burden on individual librarians for risk mitigation, accompanied by related issues of sustainability, accountability, and professional burnout. Trust appears to be the most prominent moderating factor of AI adoption in other user-friendly studies. The paradoxical nature of using both high and low at the same time, and having high and low trust, is defined by Wen (2024) as conditional trust; the users are pragmatic about systems but not confident in their efficacy in terms of epistemics. This finding is consistent with Fubenger et al. (2022), who indicate that educational and healthcare stakeholders adopt algorithmic tools to optimise efficiency but are not yet willing to do away with cognitive tasks entirely. The concept of algorithmic bias is also relevant to the review despite the allegations of issues of accuracy and verification that have been there long enough. Weidinger et al. (2021) find a small demographic bias on contemporary large language models in the setting of academic references. These results suggest that fairness does not give birth to trustful relations; an effective system of credibility in academic libraries requires transparency and traceability of the sources and compliance to academic standards.

THEORETICAL CONTRIBUTION: FROM TOOL EVALUATION TO SOCIO-TECHNICAL GOVERNANCE

This review supports the academic discourse by providing a socio-technical account of generative AI in libraries by overcoming tool-focused or adoption-oriented discourse. The view on generative AI in academic libraries as an innovation that allows saving time and resources or a technology, which needs to be ethically regulated, largely depends on current

reviews, including works by Cox (2023) and Bittle and El Geaya (2025). Although important, such accounts often treat the quality of references, workforce change, and user trust as parallel issues rather than interdependent dynamics. Combining experimental, survey, and qualitative evidence, this review shows that the problem of failures in reference quality, specifically citation hallucination and contextual misinterpretation, is the central event that triggers erosion of trust among two-thirds of the workforce, which in turn prompts the need to have more mediation in the workforce and institutional governance. According to Campbell et al. (2025), such holistic solution results in an AI failure interpretation based on governance, the inaccuracy should be redefined as an organisational risk, which requires policy adjustments, workflow reshaping, and professional controls. Notably, the review clarifies that librarians are reconsidered as epistemic arbitrators who are required to justify knowledge assertions, haggle with AI outcomes, and protect academic values. This stance is in tandem with socio-technical explanations of information infrastructure, based on the explanation by Orlikowski (2007) that human actors stabilise trust in complex systems by providing supervision and norm-setting. By doing so, the review establishes a theoretical one-way link between evidence of AI performance and professional governance, providing a framework through which previously fragmented reviews could not be empirically examined.

STRENGTHS AND LIMITATIONS

The primary strength of the reviewed literature is the variety of methodology used, as there are studies that apply experimental design (Lai, 2023, Mugaanyi et al., 2024, Wang et al., 2025), large-scale quantitative surveys (Choudhury & Shamszare, 2023, Haris et al., 2025, Elsayed & Abusharhah, 2025), qualitative interviews (Chen, 2025, Kim, 2025). This triple-bottom-line makes the evidence base more robust, as it combines objective performance indicators with the perceptions of professionals and users and makes the interpretation of the effect of the generative AI subtle. External validity to libraries of the Western perspective is achieved by the broad geographic coverage of the study, which comprises North America, Europe, Asia, the Middle East, and the Global South (Chen, 2025, Hussain, 2024, Elsayed and Abusharhah, 2025).

However, there remain several limitations. Most sources focus on ChatGPT or similar large language models Lai, (2023); Chen, (2025); Mugaanyi et al., (2024), thereby overlooking alternative or locally developed AI-based solutions. Survey-based research generally relies on self-reported notions, and the approach cannot be seen as precise regarding the assessment of actual behaviour or the level of competency (Deschênes and McMahon, 2024; Haris et al., 2025). Typically, experimental studies are cross-sectional or screenshot studies, which restrict our understanding of the longitudinal process of adaptation and learning (Lai, 2023; Wang et al., 2025). Moreover, although governance and skills gaps are often reported Gmiterek & Kotuawa (2025); Elsayed & Abusharhah (2025), empirical research has yet to test the effectiveness of individual policy or training interventions. Last but not least, although new fairness audits are emerging Wang et al. (2025), the intersectional and discipline-specific relations in trusts are under-researched, providing evidence for future research directions.

CONCLUSION

This systematic review explored how generative AI has impacted on academic library reference services, workforce roles, and user trust, based on recent empirical evidence across geographic and institutional settings. The results indicate that generative AI is associated with evident benefits in applications to tasks characterised by low complexity that require reference. In spite of its contribution, the field is limited to complex, research-based investigations since the issues of inaccuracy, hallucination, and lack of proper contextualisation persist. Instead of replacing librarians, generative AI is transforming professional practice by intensifying the need to mediate, verify, ethically sanction, and be AI literate, which only strengthens the continued centrality of human expertise in the provision of academic information services. The acceptance by users is pervasive though conditional, and trust turns out to be rather a weak but conclusive factor influenced by transparency, the credibility of the source, and the framing by institutions. The boundary of the future research should be the focus on longitudinal studies because they would help evaluate the long-term effect of the AI integration on the reference services quality, professional identity, and the nature of user-AI interactions. It is also necessary to compare the assessments of different generative AI services, such as ChatGPT, to identify which of the limitations are systemic or model-specific. In addition, practical advice regarding the responsible and sustainable adoption of AI by libraries would be provided through empirical research concerning the governance models, training programmes, and AI literacy training programmes.

ETHICAL CONSIDERATIONS

This study used secondary data from published sources and followed established ethical principles, with no requirement for institutional or ethical approval.

FUNDING

None

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

Abbott, A. (1988). The System of Professions: An Essay on the Division of Expert Labour, University of Chicago Press, Chicago. Abbott, The System of Professions: An Essay on the Division of Expert Labour, 1988. https://press.uchicago.edu/ucp/books/book/chicago/S/bo5965590.html

Bawden, D., & Robinson, L. (2022). Introduction to information science. https://openaccess.city.ac.uk/id/eprint/3224/4/Into%20to%20Info%20Sci%20Chap%201.pdf

Chen, S. C. (2025). Transforming Reference Services through ChatGPT: Insights from University Libraries in Taiwan. New Review of Academic Librarianship, 1-21. https://doi.org/10.1080/13614533.2025.2586271

Chigwada, J., & Pasipamire, N. (2024). Perception and use of large language models by library and information science students. International Journal of Librarianship, 9(3), 75-89. https://doi.org/10.23974/ijol.2024.vol9.3.385

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. Journal of Medical Internet Research, 25, e47184. https://doi.org/10.2196/47184

Cox, A. (2025). AI and the UK Library Profession: Survey Report 2025. https://orda.shef.ac.uk/articles/report/AI_and_the_UK_Library_Profession_Survey_Report_2025/30069412/1/files/57724372.pdf

Cox, A. M., Pinfield, S., & Rutter, S. (2019). The intelligent library: Thought leaders' views on the likely impact of artificial intelligence on academic libraries. Library Hi Tech, 37(3), 418-435. https://doi.org/10.1108/LHT-08-2018-0105

Del Castillo, M. S., & Kelly, H. Y. (2024). Can AI become an ally in information literacy? A survey of library instructor perspectives on ChatGPT. https://crl.acrl.org/index.php/crl/article/view/26938/34834

Deschênes, A., & McMahon, M. (2024). A survey on student use of generative AI chatbots for academic research. Evidence-based library and information practice, 19(2), 2-22. https://doi.org/10.18438/eblip30512

Elsayed, A. M., & Abusharhah, M. M. (2025). Artificial Intelligence adoption, perceptions, and ethical literacy among Arab academic librarians: A survey. The Journal of Academic Librarianship, 51(5), 103083. https://doi.org/10.1016/j.acalib.2025.103083

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. Information Systems Research, 33(2), 678-696. https://doi.org/10.1287/isre.2021.1079

Gmiterek, G., & Kotuła, S. D. (2025). Generative artificial intelligence in the activities of academic libraries of public universities in Poland. The Journal of Academic Librarianship, 51(3), 103043. https://doi.org/10.1016/j.acalib.2025.103043

Grams, M. K. (2024). Students' Perspective of the Advantages and Disadvantages of ChatGPT Compared to Reference Librarians. https://doi.org/10.18438/eblip30518

Haris, M., Ansari, A. J., Malik, B. A., Lund, B. D., & Ali, N. (2025). Artificial intelligence in academic libraries: a survey of users' perception and adoption. Global Knowledge, Memory, and Communication. https://doi.org/10.1108/GKMC-09-2024-0585

HEPI (2025). Student Generative AI Survey 2025. https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/? Accessed on (29 December 2025)

Lame, G. (2019, July). Systematic literature reviews: An introduction. In Proceedings of the Design Society: International Conference on Engineering Design (Vol. 1, No. 1, pp. 1633-1642). Cambridge University Press. https://doi.org/10.1017/dsi.2019.169

Holmes, W., & Miao, F. (2023). Guidance for generative AI in education and research. Unesco Publishing. https://books.google.com/books?hl=en&lr=&id=cKnYEAAAQBAJ&oi=fnd&pg=PA2&dq=UNESCO.+(2023).+Guidance+on+generative+AI+in+education+and+research.+UNESCO+Publishing.&ots=wmJgbeO9iP&sig=sJqnCVG5DjTvwOMZ_0VP3tqzJJA

Hussain, A. (2025). Examining Artificial Intelligence (AI) Literacy among University Library Professionals in Pakistan: The Case of Khyber Pakhtunkhwa. Available at SSRN 5278625. https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=5278625

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38. https://dl.acm.org/doi/pdf/10.1145/3571730

Kim, J. (2025). Academic library with generative AI: From passive information providers to proactive knowledge facilitators. Publications, 13(3), 37. https://doi.org/10.3390/publications13030037

Lai, K. (2023). How well does ChatGPT handle reference inquiries? An analysis based on question types and question complexities. College & Research Libraries, 84(6), 974. https://crl.acrl.org/index.php/crl/article/download/26102/34011

Mugaanyi, J., Cai, L., Cheng, S., Lu, C., & Huang, J. (2024). Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. Journal of Medical Internet Research, 26, e52935. https://doi.org/10.2196/52935

Rafiq-uz-Zaman, M. (2025). Between Adoption and Ambiguity: Navigating the AI Policy Vacuum in Pakistani Higher Education. Research Journal for Social Affairs, 3(6), 877-885. https://orcid.org/0009-0002-4853-045X

Saeidnia, H. R., Kozak, M., Lund, B. D., & Hassanzadeh, M. (2024). Evaluation of ChatGPT's responses to information needs and information seeking of dementia patients. Scientific Reports, 14(1), 10273. https://doi.org/10.1038/s41598-024-61068-5

Sayed, M. (2024). How AI will change the job of librarians: Galala University case study. Cybrarians Journal, (73), 234-243. https://doi.org/10.70000/cj.2024.73.626

Wang, H., Clark, J., Yan, Y., Bradley, S., Chen, R., Zhang, Y., ... & Tian, Z. (2025). Fairness Evaluation of Large Language Models in Academic Library Reference Services. arXiv preprint arXiv:2507.04224. https://arxiv.org/pdf/2507.04224?

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359. https://arxiv.org/pdf/2112.04359

Wen, J. (2024). On Epistemic Trust. https://figshare.mq.edu.au/ndownloader/files/47005000