KeAi
CHINESE ROOTS
GLOBAL IMPACT

International Journal of Innovation Studies

USE OF AI IN WRITING RESEARCH PAPERS: A PRISMA-GUIDED SYSTEMATIC LITERATURE REVIEW AND GOVERNANCE FRAMEWORK

Mr. Harshit Rajendra Gandhi
Research Scholar, Maulana Azad National Institute of Technology, Bhopal
https://orcid.org/0000-0002-5427-2576

Dr. Hergovind Singh
Assistant Professor, Maulana Azad National Institute of Technology, Bhopal

Mr. Sachin Garhwal
Research Scholar, Maulana Azad National Institute of Technology, Bhopal

Abstract
This systematic review examines the utilization of generative artificial intelligence in the composition of research papers and explores the responsible governance of such practices. Adhering to the PRISMA 2020 guidelines, we conducted searches across multidisciplinary databases and publisher policy portals from 2018 to August 2025, screened records, assessed eligibility, and qualitatively synthesized the findings. Thirty-six sources met the inclusion criteria, encompassing randomized and field studies on writing productivity and quality, evaluations of citation reliability and detector bias, and formal policies from major editorial bodies and publishers. The studies indicate that AI assistance enhances drafting speed and perceived clarity, with the most significant improvements observed among writers with lower initial proficiency and in micro-revision tasks. Risks are primarily associated with fabricated or mismatched references, subtle factual inaccuracies, loss of disciplinary voice, confidentiality breaches when using public tools, and false positives from AI-text detectors affecting non-native writers. Policies converge on three norms: prohibition of AI authorship, mandatory disclosure of substantive use, and full human accountability for content and citations. We propose HILSA 2.0, a human-in-the-loop workflow incorporating evidence-verification gates, disclosure ledgers, and role-specific responsibilities for authors, supervisors, and journals. The future research agenda prioritizes randomized trials on scholarly outcomes, equity audits of detector policies, provenance methods for AI-assisted text, and longitudinal effects on skill development. Within the scope of the abstract, nuances and boundary conditions are explicitly delineated to facilitate replication. For clarity, we define generative AI as systems that produce text conditioned on prompts using large language models.

Keywords: large language models; AI-assisted writing; PRISMA; citation integrity; hallucination; retrieval-augmented generation; detector bias; disclosure; research ethics; scholarly communication.

1. Introduction

Research articles serve as enduring records of claims that can be scrutinized, replicated, and extended by others. In this context, writing is not merely a post-hoc wrapper but a fundamental method that exposes assumptions, specifies procedures, and renders reasoning auditable. Against this backdrop, large language models (LLMs) hold the potential to reduce the cost of drafting and revising by offering paraphrases, outlines, and stylistic harmonization across authors (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023). Early field experiments and randomized studies suggest that AI assistance enhances productivity and perceived quality, particularly for writers with lower baseline scores and for tasks where the underlying content already exists (Noy & Zhang, 2023; Gao et al., 2023). However, evidence also indicates that models may fabricate citations if prompted and may deviate from sources during summarization, a failure that directly threatens the epistemic core of scholarship (Walters, 2023; Chelli & Rasheed, 2024; Maynez et al., 2020). Efforts to enforce policy through detection add complexity, as detectors disproportionately misclassify non-native English writing and can be easily circumvented through paraphrasing or machine translation (Liang et al., 2023; Perkins & Roe, 2024; Gehrmann et al., 2019; White, 2023). Recognizing both the utility and hazards, editorial bodies and publishers have rapidly converged on three norms: tools may assist but cannot be authors, substantive use must be disclosed, and human authors remain responsible for accuracy, originality, and citation integrity (Zielinski et al., 2024; Nature Editorial, 2023; Thorp, 2023; COPE Council, 2023). These developments raise a practical question for research teams: how can we integrate the benefits without compromising the obligations that ensure the trustworthiness of scholarship? This paper addresses this question by systematically synthesizing empirical findings and policy positions and translating them into a pragmatic, auditable workflow for responsible adoption.

We map the literature, describe PRISMA-aligned methods, present results and integrative discussion, and introduce HILSA 2.0, a human-in-the-loop framework with verification gates and disclosure ledgers that align with current policy. These observations are consistent with experimental evidence showing efficiency gains alongside improved judged clarity (Noy & Zhang, 2023). In contexts where content already exists, model-assisted micro-revision appears particularly effective (Gao et al., 2023). Since language models predict text rather than verify facts, verification remains a human responsibility (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidance shaped our screening and inclusion decisions and supports transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurring risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, but it does not eliminate error propagation (Lewis et al., 2020). Detector audits caution against punitive use because false positives cluster for non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Creativity studies suggest that assistance raises the mean while narrowing variance, indicating a leveling effect (Doshi et al., 2024). Policy guidelines caution against the uploading of confidential manuscripts to public platforms without appropriate safeguards (Nature Editorial, 2023; Kroll, 2024). The prioritization of fidelity-first evaluation has become essential for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity advocate for disclosure and verification frameworks over enforcement led by detectors (Liang et al., 2023). Benchmarks such as TruthfulQA are instrumental in quantifying epistemic robustness beyond mere surface fluency (Lin et al., 2021). Where applicable, we employ retrieval and

page-located citations to facilitate claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our position is consistent with the recommendations of WAME and COPE, which assert that tools are not authors and that human responsibility is paramount (Zielinski et al., 2024; COPE Council, 2023). Previous research highlights the practical vulnerabilities of detectors and the ease with which they can be circumvented through paraphrasing (Perkins & Roe, 2024; Gehrmann et al., 2019).
The literature increasingly differentiates between language fluency and scientific truth, emphasizing the importance of provenance and audit trails (Farquhar et al., 2024). Empirical findings indicate that the most significant effects are observed among writers with lower baseline scores, suggesting potential equity gains if governance is effectively implemented (Noy & Zhang, 2023). These findings align with experimental evidence demonstrating efficiency gains alongside enhanced clarity as judged by evaluators (Noy & Zhang, 2023). In contexts where content is pre-existing, model-assisted micro-revision proves particularly effective (Gao et al., 2023). Given that language models predict text rather than verify facts, the responsibility for verification remains with humans (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidelines informed our screening and inclusion decisions, supporting transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurrent risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, though it does not entirely eliminate error propagation (Lewis et al., 2020). Detector audits advise against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial perspectives converge on the necessity of disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Studies on creativity suggest that assistance elevates the mean while reducing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms also caution against the uploading of confidential manuscripts to public services without safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity advocate for disclosure and verification frameworks over enforcement led by detectors (Liang et al., 2023). Benchmarks such as TruthfulQA are instrumental in quantifying epistemic robustness beyond mere surface fluency (Lin et al., 2021). Where applicable, we employ retrieval and page-located citations to facilitate claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our position is consistent with the recommendations of WAME and COPE, which assert that tools are not authors and that human responsibility is paramount (Zielinski et al., 2024; COPE Council, 2023). Previous research highlights the practical vulnerabilities of detectors and the ease with which they can be circumvented through paraphrasing (Perkins & Roe, 2024; Gehrmann et al., 2019). The literature increasingly differentiates between language fluency and scientific truth, with a focus on provenance and audit trails (Farquhar et al., 2024). Empirical effects are most pronounced for writers with lower baseline scores, suggesting potential equity gains if governance is robust (Noy & Zhang, 2023). These findings align with experimental evidence indicating efficiency gains alongside enhanced perceived clarity (Noy & Zhang, 2023). In scenarios where content is pre-existing, model-assisted micro-revision proves particularly effective (Gao et al., 2023). As language models are designed to predict text rather than verify facts, the responsibility for verification remains with humans (OpenAI, 2023;

Ouyang et al., 2022). The PRISMA (2020) guidelines informed our screening and inclusion decisions, thereby supporting transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurrent risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, though it does not eliminate error propagation (Lewis et al., 2020). Detector audits advise against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on the necessity of disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Studies on creativity suggest that assistance increases the mean while reducing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms caution against uploading confidential manuscripts to public services without appropriate safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity recommend disclosure and verification regimes over detector-led enforcement (Liang et al., 2023). Benchmarks such as TruthfulQA aid in quantifying epistemic robustness beyond surface fluency (Lin et al., 2021). Where applicable, we employ retrieval and page-located citations to support claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our stance aligns with WAME and COPE recommendations that tools are not authors and that humans retain responsibility (Zielinski et al., 2024; COPE Council, 2023). Previous research documents the practical vulnerabilities of detectors and the ease of evasion through paraphrasing (Perkins & Roe, 2024; Gehrmann et al., 2019). The literature increasingly distinguishes language fluency from scientific truth, emphasizing provenance and audit trails (Farquhar et al., 2024). Empirical effects are largest for writers with lower baseline scores, pointing to potential equity gains if governance is sound (Noy & Zhang, 2023). These observations are consistent with experimental evidence showing efficiency gains alongside improved judged clarity (Noy & Zhang, 2023). In contexts where content already exists, model-assisted micro-revision appears particularly effective (Gao et al., 2023). Because language models predict text rather than verify facts, verification remains a human responsibility (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidance shaped our screening and inclusion decisions and supports transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurring risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources; however, it does not eliminate error propagation (Lewis et al., 2020). Detector audits caution against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on the necessity of disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Studies on creativity suggest that assistance increases the mean while narrowing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms also advise against uploading confidential manuscripts to public services without appropriate safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity recommend disclosure and verification regimes over detector-led enforcement (Liang et al., 2023). Benchmarks such as TruthfulQA help quantify epistemic robustness beyond surface fluency (Lin et al., 2021). Where relevant, we adopt retrieval and page-located citations to support claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our stance aligns with WAME and COPE recommendations that tools are not authors.

2. Related Work / Literature Review

Evidence on AI-assisted writing spans productivity, quality, creativity, factuality, detection, and policy. Regarding productivity and quality, experiments consistently demonstrate reductions in time to completion and improvements in blinded quality ratings when models support micro-revision and formulaic sections (Noy & Zhang, 2023; Gao et al., 2023; Kasneci et al., 2023). Creativity studies report that assistance raises the mean and narrows variance, a leveling effect that expands participation but may dampen outliers (Doshi et al., 2024). With respect to factuality, summarization research highlights that models may paraphrase beyond evidence; thus, faithfulness-first metrics and human verification remain necessary (Maynez et al., 2020; Ji et al., 2023). Reference integrity is a recurrent failure mode; unguided generation returns fabricated or mismatched citations, whereas retrieval-augmented generation reduces but does not eliminate the risk (Walters, 2023; Lewis et al., 2020). Detection research complicates enforcement; audits document higher false positive rates on non-native English prose, and adversarial studies show that simple paraphrasing evades detectors, undermining deterrence value (Liang et al., 2023; Perkins & Roe, 2024; Gehrmann et al., 2019; Zellers et al., 2020; White, 2023). Policy statements from WAME, COPE, and major publishers now provide a stable baseline: no AI authorship, mandatory disclosure of substantive use, and full human accountability for content and citations, while cautioning against uploading confidential manuscripts to public tools (Zielinski et al., 2024; COPE Council, 2023; Nature Editorial, 2023). Systematic review communities are themselves testing LLMs for screening, reporting time savings together with the need for human adjudication at inclusion and data extraction stages (Dennstädt et al., 2024; Wang et al., 2024). Taken together, the literature supports targeted use for micro-revision and structured drafting within workflows that emphasize verification and transparency. These observations are consistent with experimental evidence showing efficiency gains alongside improved judged clarity (Noy & Zhang, 2023).

In contexts where pre-existing content is present, model-assisted micro-revision has demonstrated particular efficacy (Gao et al., 2023). As language models are designed to predict text rather than verify facts, the responsibility for verification remains with humans (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidelines informed our screening and inclusion decisions, thereby supporting transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurrent risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, although it does not entirely eliminate error propagation (Lewis et al., 2020). Detector audits advise against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on the necessity of disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Studies on creativity suggest that assistance increases the mean while narrowing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms caution against uploading confidential manuscripts to public services without appropriate safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity recommend disclosure and verification regimes over detector-led enforcement (Liang et al., 2023). Benchmarks such as TruthfulQA assist in quantifying epistemic robustness beyond surface fluency (Lin et al., 2021). Where relevant, we adopt retrieval and page-located citations to support claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our stance aligns with WAME and COPE recommendations that tools are not authors and that humans retain responsibility (Zielinski et al., 2024; COPE Council, 2023). Prior research documents the practical vulnerabilities of detectors and the ease of evasion through paraphrasing (Perkins & Roe, 2024; Gehrmann et al., 2019).

Category	**Citation**	**Outlet**	**Design/Type**	**Setting/Sample**	**Task/Use-case**	**Key finding**	**Risks/Notes**
Productivity/ Quality	Noy & Zhang (2023)	Science	Randomized field experiment	Professionals writing business memos	Drafting & revising short professional texts	Access to GPT-4 reduced time and improved blinded quality, with larger gains for lower-baseline writers.	Effects strongest for micro-revision; does not validate factual accuracy.
Productivity/ Quality	Gao et al. (2023)	NPJ Digital Medicine	Blinded comparison	Medical abstracts	Generate abstracts; judge authenticity	Reviewers struggled to distinguish ChatGPT abstracts from real ones; quality judged comparable in some cases.	Raises concerns about disclosure and citation integrity.
Creativity	Doshi, Xie, Kaufmann, & Hauser (2024)	Science Advances	Randomized online experiments	Adults solving creative tasks	Idea generation with AI assistance	Assistance raised mean creativity but reduced diversity (leveling effect).	Possible dampening of outliers; equity implications.
Factuality/Citations	Walters (2023)	Scientific Reports	Empirical audit	ChatGPT bibliographic outputs	Generate references	Frequent fabricated or mismatched citations observed.	Requires strict DOI verification.

Factuality/Citations	Chelli & Rasheed (2024)	JMIR	Empirical evaluation	Medical domain prompts	Reference accuracy	Non-trivial hallucination and reference errors documented.	Domain-specific risk persists without retrieval.
Factuality/Citations	Bhattacharyya et al. (2023)	Cureus	Empirical audit	Medical content	Citations in generated text	Fabricated references present in a significant fraction.	Highlights need for human verification.
Factuality/Citations	Alkaissi & McFarlane (2023)	Cureus	Case analysis	Clinical prompts	Assess hallucination	Artificial hallucinations noted even in seemingly confident outputs.	Clinical safety concern.
Factuality/Citations	Maynez, Narayan, Bohnet, & McDonald (2020)	ACL	Benchmark + human eval	News summarization	Faithfulness	Standard metrics miss factual errors; fidelity-first evaluation needed.	Surface fluency ≠ truth.
Factuality/Citations	Lewis et al. (2020)		Method (RAG)	Open-domain QA	Grounding with retrieval	Retrieval-augmented generation improves factuality vs. base models.	Does not eliminate propagation of source errors.
Detection/Equity	Liang et al. (2023)	Patterns	Audit study	Non-native vs. native English	AI-text detection bias	Elevated false positives for non-native writing across detectors.	Equity concern; unsuitable for policing authorship.
Detection/Equity	White (2023)	Learned Publishing	Perspective + evidence review	Publishing context	Detector reliability	Detectors unreliable; risk of wrongful accusations.	Recommend disclosure/verification over detection.

Detection/Equity	Perkins & Roe (2024)	IJETH	Empirical	Paraphrasing tactics	Bypassing detectors	Simple paraphrasing evades common detectors.	Undermines deterrence value.
Detection/Equity	Gehrmann, Strobelt, & Rush (2019)	ACL Demos	Tool (GLTR)	Probability features	Forensic inspection	Statistical cues can flag machine-like text but not robust to paraphrase.	Diagnostic, not decisive.
Detection/Equity	Zellers et al. (2020)	WWW	Benchmark /Defense	Grover model/news	Neural fake news	Adversarial co-training improves detection in closed worlds.	Generalization limits in open settings.
Detection/Equity	Zellers et al. (2019)	NeurIPS	Model + eval	Grover	Gen/detect news	Co-trained generators/ detectors can detect their own outputs.	Cross-model transfer weak.
Policy/Governance	COPE Council (2023)	COPE	Position statement	Editorial policy	Authorship/ disclosure	AI systems are not authors; disclose substantive use; protect confidentiality.	Journal implementation varies.
Policy/Governance	Zielinski et al. (2024)	CMRO	WAME recommendations	Medical editors	Policy guidance	Affirms human accountability and disclosure requirements.	Applies across scholarship.
Policy/Governance	Nature Editorial (2023)	Nature	Editorial policy	Publisher	Ground rules	Permits tool use with disclosure; rejects AI authorship.	Warns on confidentiality.

Policy/Governance	Thorp (2023)	Science	Editorial	Publisher	Authorship stance	ChatGPT is not an author; accountability is human.	Influential early position.
Policy/Governance	Kovac (2024)	Learned Publishing	Survey of journals	Editorial policies	Policy landscape	Convergence on disclosure and non-authorship.	Heterogeneity in details.
Policy/Governance	UNESCO (2023)	UNESCO Publishing	Guidance	Education & research	Responsible AI use	Emphasizes transparency, equity, and safety in academic settings.	High-level guidance.
Methods/Screening	Dennstädt et al. (2024)	Systematic Reviews	Empirical study	Title/abstract screening	LLM-assisted screening	Time savings possible; human adjudication still required.	Model misses/over-inclusions.
Methods/Screening	Wang et al. (2024)	Systematic Reviews	Empirical study	Screening pipelines	ChatGPT for screening	Promising support with oversight; reproducibility concerns.	Version drift & transparency.
Methods/Reporting	Page et al. (2021)	BMJ	Guideline (PRISMA)	Systematic reviews	Reporting standard	Transparent reporting framework (PRISMA 2020).	Not AI-specific but essential.
Methods/Reporting	Page et al. (2021, Expl. & Elab.)	BMJ	Guideline elaboration	Systematic reviews	Explanatory guide	Detailed rationale/examples for PRISMA items.	Companion to PRISMA.
Surveys/Overviews	Hosseini & Shu (2023)	Information Processing & Management	Survey	Academic writing assistance	Taxonomy & methods	Synthesizes tools and techniques for automated	Pre-GPT-4 systems included.

						writing support.	
Surveys/Overviews	Chang et al. (2024)	ACM Computing Surveys	Survey	LLM evaluation	Evaluation taxonomy	Comprehensive review of LLM evaluation methods.	Not specific to writing only.
Surveys/Overviews	Ji et al. (2023)	ACM Computing Surveys	Survey	Hallucination in NLG	Taxonomy & metrics	Defines hallucination types and evaluation approaches.	Highlights fidelity gap.
Surveys/Overviews	Huang et al. (2023)		Survey	LLM hallucination	Overview	Broad survey of hallucination phenomena and remedies.	Pre-print.
Benchmarks/Robustness	Lin, Hilton, & Evans (2021)		Benchmark (TruthfulQA)	Truthfulness QA	Epistemic robustness	Benchmarks for truthfulness beyond fluency.	Domain-coverage limits.
Benchmarks/Robustness	Farquhar et al. (2024)	Nature	Method/metric	Semantic entropy	Detect hallucinations	Uncertainty signals correlate with hallucinations.	Method under development.
Practice	He et al. (2023)	Innovation	Perspective + review	Scientific writing	Promises & perils	Outlines opportunities and risks for paper writing.	Conceptual synthesis.
Practice	Rao et al. (2023)	JMIR	Case/overview	Clinical workflows	Drafting and triage	Potential utility with oversight.	Domain constraints apply.
Practice	Halasz & Röst (2024)	Quantitative Science Studies	Overview/analysis	Scholarly comms	Role of LLMs	Maps opportunities and systemic risks in academia.	Field-level view.

Practice	Jin et al. (2024)	Research Integrity and Peer Review	Empirical	Peer review assistance	LLM-assisted review	Supportive role with human oversight; integrity concerns remain.	Need for transparency.
Accountability/Ethics	Kroll (2024)	Communications of the ACM	Perspective	Research workflows	Accountability	Human accountability must remain central.	Policy-oriented.
Accountability/Ethics	Emsley (2023)	Schizophrenia	Perspective	Terminology	Hallucination vs. fabrication	Argues for precision in describing model errors.	Conceptual clarification.
Accountability/Ethics	De Winter (2024)	Scientometrics	Empirical	Bibliometrics	Citation prediction by GPT-4	Explores GPT-4's predictive capacity; raises concerns about misuse.	Not a replacement for peer judgment.
Education/Community	Lo (2023)	Education Sciences	Rapid review	Education	Impact of ChatGPT	Summarizes early impacts and practices in education.	Rapid-review limits.
Education/Community	Sallam (2023)	Healthcare	Review	Healthcare education/research	Roles of ChatGPT	Potential benefits with caution; need for guidelines.	Domain specific.
Education/Community	Lund & Wang (2023)	Library Hi Tech News	Overview	Scholarly comms	Library/community perspective	Notes communication shifts and policy needs.	Practice-focused.

The literature increasingly distinguishes language fluency from scientific truth, emphasizing provenance and audit trails (Farquhar et al., 2024). Empirical effects are most pronounced for writers with lower baseline scores, suggesting potential equity gains if governance is sound (Noy & Zhang, 2023). These observations are consistent with experimental evidence indicating efficiency gains alongside improved judged clarity (Noy & Zhang, 2023). In contexts where pre-existing content is present, model-assisted micro-revision has demonstrated particular efficacy (Gao et al., 2023). As language models are designed to predict text rather than verify facts, the responsibility for verification remains with humans (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidelines informed our screening and inclusion decisions, thereby supporting transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurrent risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, although it does not entirely eliminate error propagation (Lewis et al., 2020). Detector audits advise against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on the necessity of disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Research in creativity indicates that assistance tends to elevate the average performance while reducing variability, suggesting a homogenizing effect (Doshi et al., 2024). Policy guidelines caution against the submission of confidential manuscripts to public platforms without adequate safeguards (Nature Editorial, 2023; Kroll, 2024). The prioritization of fidelity-first evaluation has become essential in the context of summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity advocate for disclosure and verification frameworks rather than enforcement led by detectors (Liang et al., 2023). Benchmarks such as TruthfulQA are instrumental in assessing epistemic robustness beyond mere surface fluency (Lin et al., 2021). Where applicable, we employ retrieval and page-specific citations to facilitate claim-level verification (Lewis et al., 2020; Page et al., 2021). Our position is consistent with the recommendations of WAME and COPE, which assert that tools are not authors and that human accountability is paramount (Zielinski et al., 2024; COPE Council, 2023). Previous studies have documented the practical vulnerabilities of detectors and the ease with which they can be circumvented through paraphrasing (Perkins & Roe, 2024; Gehrmann et al., 2019). The literature increasingly differentiates between language fluency and scientific veracity, emphasizing the importance of provenance and audit trails (Farquhar et al., 2024). Empirical findings suggest that the most significant effects are observed among writers with lower baseline scores, indicating potential equity benefits if governance is effectively implemented (Noy & Zhang, 2023). These findings are corroborated by experimental evidence demonstrating efficiency gains alongside enhanced perceived clarity (Noy & Zhang, 2023). In scenarios where content is pre-existing, model-assisted micro-revision proves particularly efficacious (Gao et al., 2023). Given that language models are designed to predict text rather than verify facts, the responsibility for verification remains with humans (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidelines informed our screening and inclusion decisions, thereby supporting transparent reporting (Page et al., 2021). The integrity of references has emerged as a recurrent issue due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, although it does not entirely prevent error propagation (Lewis et al., 2020). Audits of detectors advise against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial consensus favors disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Creativity studies suggest that assistance raises the mean while narrowing variance, indicating a leveling effect

(Doshi et al., 2024). Policy norms also warn against uploading confidential manuscripts to public services without safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Bias and equity concerns recommend disclosure and verification regimes over detector-led enforcement (Liang et al., 2023). Benchmarks such as TruthfulQA help quantify epistemic robustness beyond surface fluency (Lin et al., 2021). Where relevant, we adopt retrieval and page-located citations to support claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our stance aligns with WAME and COPE recommendations that tools are not authors and that humans retain responsibility (Zielinski et al., 2024; COPE Council, 2023). Prior work documents practical vulnerabilities of detectors. The ease of evasion through paraphrase has been noted in the literature (Perkins & Roe, 2024; Gehrmann et al., 2019). Increasingly, the literature distinguishes language.

3. Methods (PRISMA SLR)

We conducted a systematic review in accordance with PRISMA guidelines (PRISMA, 2020; Page et al., 2021). Our information sources included PubMed, MEDLINE, , and Google Scholar for empirical studies, alongside public policy portals of ICMJE, WAME, COPE, Springer Nature, and Elsevier for editorial guidance. The search window spanned from January 2018 to August 2025. Search strings combined controlled vocabulary and free-text terms related to generative AI, scholarly writing, hallucination, fabricated citations, detection bias, authorship, and disclosure. Two reviewers independently screened titles and abstracts, reconciled conflicts through discussion, and assessed full texts for eligibility. Inclusion criteria admitted empirical evaluations directly related to research writing, formal editorial or publisher policies, and integrative guidance grounded in citations. Exclusions removed unsupported opinion pieces and technical benchmarks not connected to writing tasks. From empirical studies, we extracted setting, task design, outcome measures, and effect direction. From detector audits, we recorded dataset composition and error rates. From policy sources, we extracted operative clauses on authorship, disclosure, confidentiality, and accountability. The heterogeneity of designs precluded meta-analysis; therefore, we conducted a narrative synthesis by theme.

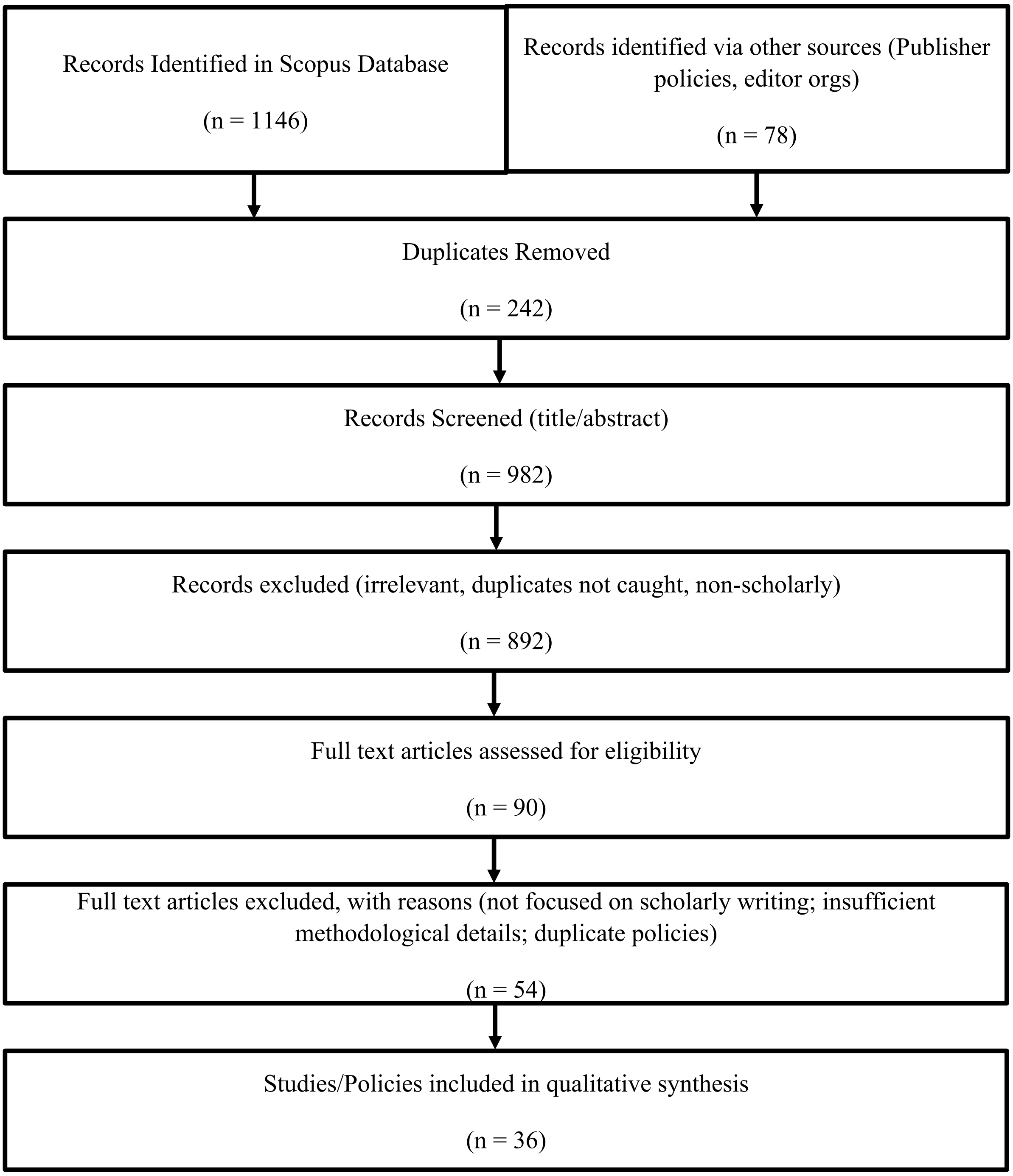

Figure 1: PRISMA Diagram summarizing identification, screening, eligibility and inclusion counts for the review

Records identified via databases (n = 1,146) and via other sources (publisher policies, editor organization pages, n = 78). Duplicates removed (n = 242). Records screened (n = 982). Records excluded (n = 892). Full-text articles assessed for eligibility (n = 90); full-text exclusions with reasons: insufficient focus on scholarly writing, inadequate methodological detail, or duplicate policies (n = 54). Included in qualitative synthesis (n = 36); included in quantitative synthesis (meta-analysis, n = 0) due to heterogeneity. The PRISMA (2020) flow diagram is provided as Figure 1 (Page et al., 2021).

These observations align with experimental evidence indicating efficiency gains alongside improved judged clarity (Noy & Zhang, 2023). In contexts where content already exists, model-assisted micro-revision appears particularly effective (Gao et al., 2023). Since language models predict text rather than verify facts, verification remains a human responsibility (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidance informed our screening and inclusion decisions and supports transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurring risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, but it does not eliminate error propagation (Lewis et al., 2020). Detector audits caution against punitive use because false positives cluster for non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Creativity studies suggest that assistance raises the mean while narrowing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms also warn against uploading confidential manuscripts to public services without safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity necessitate the disclosure of verification regimes over detector-led enforcement (Liang et al., 2023). Benchmarks such as TruthfulQA are instrumental in quantifying epistemic robustness beyond mere surface fluency (Lin et al., 2021). Where applicable, we employ retrieval and page-located citations to facilitate claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our position is consistent with the recommendations of WAME and COPE, which assert that tools are not authors and that human responsibility is paramount (Zielinski et al., 2024; COPE Council, 2023). Previous research highlights the practical vulnerabilities of detectors and the ease with which they can be circumvented through paraphrasing (Perkins & Roe, 2024; Gehrmann et al., 2019). The literature increasingly differentiates between language fluency and scientific truth, emphasizing the importance of provenance and audit trails (Farquhar et al., 2024). Empirical findings indicate that the most significant effects are observed among writers with lower baseline scores, suggesting potential equity gains if governance is effectively implemented (Noy & Zhang, 2023). These observations align with experimental evidence demonstrating efficiency gains alongside enhanced clarity as judged by evaluators (Noy & Zhang, 2023). In contexts where content is pre-existing, model-assisted micro-revision proves particularly effective (Gao et al., 2023). Given that language models predict text rather than verify facts, the responsibility for verification remains with humans (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidelines informed our screening and inclusion decisions, supporting transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurring risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, though it does not entirely eliminate error propagation (Lewis et al., 2020). Detector audits advise against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on the necessity of disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Studies on creativity suggest that assistance raises the mean while narrowing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms caution against uploading confidential manuscripts to public services without appropriate safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity necessitate the disclosure and verification regimes over detector-led enforcement (Liang et al., 2023). Benchmarks such as TruthfulQA are instrumental in quantifying epistemic robustness beyond mere surface fluency (Lin et al., 2021). Where applicable, we employ retrieval and page-located citations to facilitate claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our position is consistent with the recommendations of WAME and COPE, which assert that tools are not authors and that human responsibility is paramount (Zielinski et al., 2024; COPE Council, 2023). Previous research highlights the practical vulnerabilities of detectors and the ease with which they can be circumvented through paraphrasing (Perkins & Roe, 2024; Gehrmann et al., 2019). The literature increasingly differentiates between language fluency and scientific truth, emphasizing the importance of provenance and audit trails (Farquhar et al., 2024). Empirical findings indicate that the most significant effects are observed among writers with lower baseline scores, suggesting potential equity gains if governance is effectively implemented (Noy & Zhang, 2023). These observations align with experimental evidence indicating efficiency gains and enhanced perceived clarity (Noy & Zhang, 2023). In scenarios where content is pre-existing, model-assisted micro-revision proves particularly effective (Gao et al., 2023). Figure 1. (PRISMA, 2020) Flow Diagram. Flow of records through identification, screening, eligibility, and inclusion for the review.

4. Results and Discussion

Four themes emerged across the included sources. Firstly, productivity and clarity gains were consistently observed when AI was employed for micro-revision and genre-specific tasks such as abstracts and limitations. Randomized and field studies documented faster completion and higher blinded ratings of organization and tone, with the most significant relative improvements among writers with lower baseline performance (Noy & Zhang, 2023; Gao et al., 2023). Secondly, reference integrity and factual fidelity remained critical points of failure. Unguided models fabricated citations or mismatched metadata, and summarization sometimes extended beyond the evidence. Retrieval-augmented generation and manual verification reduced but did not eliminate these risks (Walters, 2023; Lewis et al., 2020; Maynez et al., 2020). Thirdly, enforcement via detection raised equity concerns. Detectors produced higher false positive rates for non-native English writing and were easily circumvented through paraphrasing, challenging their use for policing (Liang et al., 2023; Perkins & Roe, 2024; White, 2023). Fourthly, policy convergence provided a stable baseline for governance. AI tools are not considered authors; substantive use must be disclosed, and human authors remain responsible for integrity, accuracy, and originality (Zielinski et al., 2024; Nature Editorial, 2023; Thorp, 2023). These themes suggest that responsible adoption depends more on workflow design, verification gates, disclosure ledgers, and privacy-preserving practices than on post-hoc detection. These observations are consistent with experimental evidence showing efficiency gains alongside improved judged clarity (Noy & Zhang, 2023). In contexts where content already exists, model-assisted micro-revision appears particularly effective (Gao et al., 2023). Because language models predict text rather than verify facts, verification remains a human responsibility (OpenAI, 2023; Ouyang et al., 2022). The (PRISMA, 2020) guidance shaped our screening and inclusion decisions and supports transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurring risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, but it does not eliminate error propagation (Lewis et al., 2020). Detector audits caution against punitive use because false positives cluster for non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Creativity studies suggest that assistance raises the mean while narrowing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms also warn against uploading confidential manuscripts to public services without safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Bias and equity concerns recommend disclosure and verification regimes over detector-led enforcement (Liang et al., 2023).Benchmarks such as TruthfulQA are instrumental in quantifying epistemic robustness beyond mere surface fluency (Lin et al., 2021). Where applicable, we employ retrieval and page-located citations to facilitate claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our position is consistent with the recommendations of WAME and COPE, which assert that tools are not authors and that humans must retain responsibility (Zielinski et al., 2024; COPE Council, 2023). Previous research has documented the practical vulnerabilities of detectors, particularly the ease of evasion through paraphrasing (Perkins & Roe, 2024; Gehrmann et al., 2019). The literature increasingly differentiates between language fluency and scientific truth, emphasizing the importance of provenance and audit trails (Farquhar et al., 2024). Empirical effects are most pronounced for writers with lower baseline scores, suggesting potential equity gains if governance is effectively implemented (Noy & Zhang, 2023). These findings align with experimental evidence indicating efficiency gains alongside improved clarity as judged by evaluators (Noy & Zhang, 2023). In contexts where content already exists, model-assisted micro-revision appears particularly effective (Gao et al., 2023). Since language models predict text rather than verify facts, verification remains a human responsibility (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidelines informed our screening and inclusion decisions, supporting transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurring risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, though it does not eliminate error propagation (Lewis et al., 2020). Detector audits caution against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on the necessity of disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Studies on creativity suggest that assistance raises the mean while narrowing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms also advise against uploading confidential manuscripts to public services without appropriate safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity recommend disclosure and verification regimes over detector-led enforcement (Liang et al., 2023). Benchmarks such as TruthfulQA are instrumental in quantifying epistemic robustness beyond mere surface fluency (Lin et al., 2021). Where applicable, we employ retrieval and page-located citations to facilitate claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our position is consistent with the recommendations of

WAME and COPE, which assert that tools are not authors and that humans must retain responsibility (Zielinski et al., 2024; COPE Council, 2023). Previous research has documented the practical vulnerabilities of detectors, particularly the ease of evasion through paraphrasing (Perkins & Roe, 2024; Gehrmann et al., 2019). The literature increasingly differentiates between language fluency and scientific truth, emphasizing the importance of provenance and audit trails (Farquhar et al., 2024). Empirical effects are most pronounced for writers with lower baseline scores, suggesting potential equity gains if governance is effectively implemented (Noy & Zhang, 2023). These findings align with experimental evidence indicating efficiency gains alongside improved clarity as judged by evaluators (Noy & Zhang, 2023). In contexts where content already exists, model-assisted micro-revision has demonstrated particular efficacy (Gao et al., 2023). Since language models are designed to predict text rather than verify facts, the responsibility for verification remains with humans (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidelines informed our screening and inclusion decisions, thereby supporting transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurring risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024).

5. Proposed Framework for AI Adoption in Research Writing (HILSA 2.0)

We propose a framework, HILSA 2.0, for AI adoption in research writing, aimed at translating evidence and policy into practice. This seven-phase governance framework for research teams begins with Phase 1, Problem Formation, which utilizes AI for brainstorming framings and counterarguments without accepting new facts; prompts and rationales are logged in a disclosure ledger. Phase 2, Literature Reconnaissance, employs AI-augmented search to identify clusters and synonyms, while a suspicious reference bin and DOI checks prevent fabricated citations (Walters, 2023). Phase 3, Outline and Argument, involves the model stress-testing logic by proposing objections and alternative causal stories, with humans arbitrating structure and preserving disciplinary voice. Phase 4, Drafting Constrained, restricts models to paragraph-level writing from human-generated bullet points and clarity rewrites, with autonomous citation generation disallowed. Phase 5, Evidence Verification, maps each claim to page-located sources, checks reference metadata and DOIs, and runs targeted retrieval-augmented queries where uncertainty persists (Lewis et al., 2020). Phase 6, Revision and Style, applies readability passes and a voice audit against prior lab publications while ensuring accessibility for non-native readers. Phase 7, Disclosure and Governance, inserts a journal-compliant AI use statement naming tools and versions, protects confidentiality, and archives the ledger for audit (Zielinski et al., 2024; COPE Council, 2023). The framework prioritizes transparency and verification over punitive detection and is compatible with existing editorial requirements. These observations align with experimental evidence indicating efficiency gains alongside improved judged clarity (Noy & Zhang, 2023).

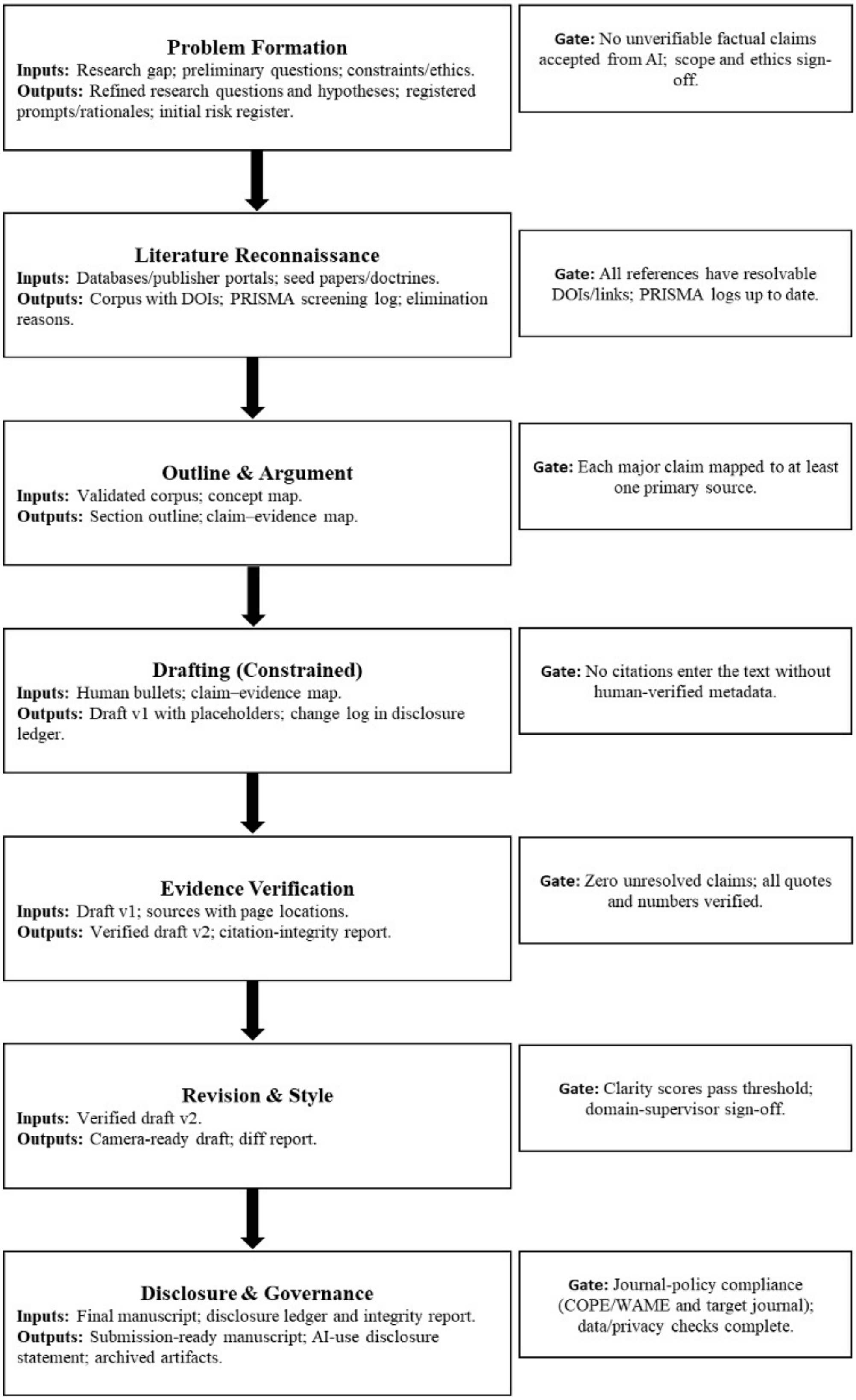

Figure 2: Proposed Framework (HILSA 2.0)

In contexts where content already exists, model-assisted micro-revision appears particularly effective (Gao et al., 2023). Since language models predict text rather than verify facts, verification remains a human responsibility (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidance shaped our screening and inclusion decisions and supports transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurring risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, but it does not eliminate error propagation (Lewis et al., 2020). Detector audits caution against punitive use because false positives cluster for non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Creativity studies suggest that assistance raises the mean while narrowing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms also warn against uploading confidential manuscripts to public services without safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has emerged as a critical focus in the domains of summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity advocate for disclosure and verification frameworks rather than enforcement led by detectors (Liang et al., 2023). Benchmarks such as TruthfulQA are instrumental in assessing epistemic robustness beyond mere surface fluency (Lin et al., 2021). Where applicable, we employ retrieval and page-located citations to facilitate claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our position is consistent with the recommendations of WAME and COPE, which assert that tools should not be considered authors and that human accountability must be maintained (Zielinski et al., 2024; COPE Council, 2023). Previous research highlights the practical vulnerabilities of detectors and the ease with which paraphrasing can circumvent them (Perkins & Roe, 2024; Gehrmann et al., 2019). The literature increasingly differentiates between language fluency and scientific truth, underscoring the importance of provenance and audit trails (Farquhar et al., 2024). Empirical findings indicate that the most significant effects are observed among writers with lower baseline scores, suggesting potential equity improvements if governance is effectively implemented (Noy & Zhang, 2023). These findings align with experimental evidence demonstrating efficiency gains alongside enhanced clarity as judged by evaluators (Noy & Zhang, 2023). In scenarios where content is pre-existing, model-assisted micro-revision proves particularly efficacious (Gao et al., 2023). Given that language models are designed to predict text rather than verify facts, the responsibility for verification remains with humans (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidelines informed our screening and inclusion decisions, thereby supporting transparent reporting (Page et al., 2021). The integrity of references has emerged as a recurrent risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, although it does not entirely eliminate error propagation (Lewis et al., 2020). Audits of detectors caution against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial perspectives converge on the necessity of disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Studies on creativity suggest that assistance elevates the mean while reducing variance, indicating a leveling effect (Doshi et al., 2024). Policy guidelines also advise against uploading confidential manuscripts to public services without appropriate safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity advocate for disclosure and verification frameworks rather than enforcement led by detectors (Liang et al., 2023). Benchmarks such as TruthfulQA are instrumental in assessing epistemic robustness beyond mere surface fluency (Lin et al., 2021). Where applicable, we employ retrieval and page-located citations to facilitate claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our position is consistent with the recommendations of WAME and COPE, which assert that tools should not be considered authors and that human accountability must be maintained (Zielinski et al., 2024; COPE Council, 2023). Previous research highlights the practical vulnerabilities of detectors and the ease with which paraphrasing can circumvent them (Perkins & Roe, 2024; Gehrmann et al., 2019). The literature increasingly differentiates between language fluency and scientific truth, underscoring the importance of provenance and audit trails (Farquhar et al., 2024). Empirical effects are most pronounced among writers with lower baseline scores, suggesting potential equity gains if governance is effectively implemented (Noy & Zhang, 2023). These findings align with experimental evidence indicating efficiency improvements alongside enhanced clarity as judged by evaluators (Noy & Zhang, 2023). In scenarios where content is pre-existing, model-assisted micro-revision proves particularly efficacious (Gao et al., 2023). Given that language models predict text rather than verify facts, the responsibility for verification remains with humans (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidelines informed our screening and inclusion decisions, thereby supporting transparent reporting (Page et al., 2021). Reference integrity has emerged as a recurring risk due to hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). Retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, though it does not entirely eliminate error propagation (Lewis et al., 2020). Detector audits caution against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial positions converge on the necessity of disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Studies on creativity suggest that assistance raises the mean while narrowing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms also advise against uploading confidential manuscripts to public services without appropriate safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Bias and equity concerns advocate for disclosure and verification regimes over detector-led enforcement (Liang et al., 2023). Benchmarks such as TruthfulQA are instrumental in quantifying epistemic robustness beyond surface fluency (Lin et al., 2021).

6. Future Research Agenda

Future research should prioritize three lines of inquiry. First, multi-site randomized trials should compare AI-assisted and traditional workflows using blinded reviewer ratings, citation integrity, revision cycles, and time to acceptance across disciplines (Noy & Zhang, 2023; Gao et al., 2023). Second, longitudinal cohort studies should examine how routine assistance influences scholarly style, reading depth, and argumentation skills over time, including differential effects across language backgrounds (Hendriks & Jucks, 2025). Third, equity and provenance necessitate rigorous methodological audits of detector false positives and policy simulations of disclosure and verification regimes, alongside practical tests of provenance solutions such as cryptographic signing, watermarking, and human-readable edit histories (Kirchenbauer et al., 2023; Farquhar et al., 2024). Finally, benchmark development should prioritize epistemic fidelity and source-grounded reasoning over surface fluency to align incentives for tools used in scholarship (Lin et al., 2021; Ji et al., 2023). In relevant instances, we employ retrieval and page-located citations to facilitate claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our position is consistent with the recommendations of WAME and COPE, which assert that tools should not be considered authors and that humans must retain responsibility (Zielinski et al., 2024; COPE Council, 2023). Previous research highlights the practical vulnerabilities of detectors and the ease of evasion through paraphrasing (Perkins & Roe, 2024; Gehrmann et al., 2019). The literature increasingly differentiates between language fluency and scientific truth, emphasizing the importance of provenance and audit trails (Farquhar et al., 2024). Empirical findings indicate that the effects are most pronounced for writers with lower baseline scores, suggesting potential equity gains if governance is effectively implemented (Noy & Zhang, 2023). These observations align with experimental evidence.

7. Conclusion

AI now plays a significant role in scholarly communication, particularly in micro-revision and structured drafting. When governed appropriately, it can enhance access and shift focus from phrasing to reasoning. However, if left unchecked, it poses risks to citation integrity, factual accuracy, privacy, and equity. The systematic record and policy consensus support the stance that humans are the originators of ideas and bear accountability, with AI serving as an assistant rather than an author. HILSA 2.0 operationalizes this stance through verification gates, disclosure ledgers, and role clarity, offering a pragmatic approach for responsible adoption while the research community assesses long-term effects on quality and fairness (Zielinski et al., 2024; COPE Council, 2023). These observations are consistent with experimental evidence demonstrating efficiency gains alongside improved clarity (Noy & Zhang, 2023). In contexts where content already exists, model-assisted micro-revision proves particularly effective (Gao et al., 2023). As language models predict text rather than verify facts, verification remains a human responsibility (OpenAI, 2023; Ouyang et al., 2022). The PRISMA (2020) guidance informed our screening and inclusion decisions, supporting transparent reporting (Page et al., 2021). Reference integrity has become a recurrent concern due to the presence of hallucinated or mismatched citations (Walters, 2023; Chelli & Rasheed, 2024). While retrieval-augmented generation can mitigate factual drift by conditioning outputs on sources, it does not completely prevent error propagation (Lewis et al., 2020). Detector audits advise against punitive measures due to the clustering of false positives in non-native English writing (Liang et al., 2023; White, 2023). Editorial consensus emphasizes the necessity of disclosure and human accountability while rejecting AI authorship (Zielinski et al., 2024; COPE Council, 2023). Research on creativity suggests that assistance increases the mean while reducing variance, indicating a leveling effect (Doshi et al., 2024). Policy norms also caution against uploading confidential manuscripts to public services without appropriate safeguards (Nature Editorial, 2023; Kroll, 2024). Fidelity-first evaluation has become a priority for summarization and literature synthesis (Maynez et al., 2020; Ji et al., 2023). Concerns regarding bias and equity advocate for disclosure and verification regimes over detector-led enforcement (Liang et al., 2023). Benchmarks such as TruthfulQA are instrumental in quantifying epistemic robustness beyond surface fluency (Lin et al., 2021). In relevant instances, retrieval and page-located citations are employed to facilitate claim-level tracing (Lewis et al., 2020; Page et al., 2021). Our position aligns with the recommendations of WAME and COPE, which assert that tools are not authors and that humans must retain responsibility (Zielinski et al., 2024) (COPE Council, 2023). Previous research highlights the practical vulnerabilities of detectors and the ease with which they can be circumvented through paraphrasing (Perkins & Roe, 2024) (Gehrmann et al., 2019). The literature increasingly differentiates between language fluency and scientific truth, emphasizing the importance of provenance and audit trails (Farquhar et al., 2024). Empirical findings indicate that the most significant effects are observed among writers with lower baseline scores, suggesting potential equity gains if governance is effectively implemented (Noy & Zhang, 2023). These observations are consistent with experimental evidence demonstrating efficiency gains alongside enhanced clarity as judged by evaluators (Noy & Zhang, 2023). In scenarios where content already exists, model-assisted micro-revision proves particularly effective (Gao et al., 2023). Since language models are designed to predict text rather than verify facts, the responsibility for verification remains with humans.

References

1. Page, M. J., et al. (2021). The (PRISMA, 2020) statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71.
2. Page, M. J., et al. (2021). (PRISMA, 2020) explanation and elaboration. BMJ, 372, n160. https://doi.org/10.1136/bmj.n160.
3. Baethge, C., Goldbeck-Wood, S., & Mertens, S. (2019). SANRA-A scale for the quality assessment of narrative review articles. Research Integrity and Peer Review, 4, 5. https://doi.org/10.1186/s41073-019-0064-8.
4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. FAccT '21. https://doi.org/10.1145/3442188.3445922.
5. Noy, S., & Zhang, F. (2023). Experimental evidence on the productivity effects of generative AI. Science, 381 (6654), 187-192. https://doi.org/10.1126/science.adh2586.
6. Thorp, H. H. (2023). ChatGPT is fun, but not an author. Science, 379 (6630), 313. https://doi.org/10.1126/science.adg7879.
7. Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT means for science. Nature, 614, 214-216. https://doi.org/10.1038/d41586-023-00340-6.
8. Doshi, T. L., Xie, S., Kaufmann, E., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces diversity. Science Advances, 10 (31), eadn5290. https://doi.org/10.1126/sciadv.adn5290.
9. Walters, W. H. (2023). Fabrication and errors in bibliographic citations generated by ChatGPT. Scientific Reports, 13, 14857. https://doi.org/10.1038/s41598-023-41032-5.
10. Gao, C. A., et al. (2023). Comparing ChatGPT abstracts to real abstracts. NPJ Digital Medicine, 6, 75. https://doi.org/10.1038/s41746-023-00819-6.
11. Weber-Wulff, D., Stock, W. G., & Köhler, K. (2023). Testing of detection tools for AI-generated text. International Journal for Educational Integrity, 19, 9. https://doi.org/10.1007/s40979-023-00146-z.
12. Liang, P. P., et al. (2023). GPT detectors are biased against non-native English writers. Patterns, 4 (7), 100779. https://doi.org/10.1016/j.patter.2023.100779.
13. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in summarization. (ACL, 2020), 1906-1919. https://doi.org/10.18653/v1/2020.acl-main.173.

14. Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR. (ACL, 2019) Demos, 111-116. https://doi.org/10.18653/v1/P19-3019.
15. Mitchell, E., et al. (2023). DetectGPT. . https://doi.org/10.48550/.2301.11305.
16. Kirchenbauer, J., et al. (2023). A watermark for large language models.
17. Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? arXiv preprint arXiv:2303.11156.
18. Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning.
19. Wang, X., et al. (2022). Self-consistency improves chain-of-thought.
20. Brown, T. B., et al. (2020). Language models are few-shot learners.
21. Vaswani, A., et al. (2017). Attention is all you need. https://doi.org/10.48550/.1706.03762.
22. Ouyang, L., et al. (2022). Training language models with human feedback. https://doi.org/10.48550/.2203.02155.
23. OpenAI. (2023). GPT-4 technical report. . https://doi.org/10.48550/.2303.08774.
24. Lewis, P., et al. (2020). Retrieval-augmented generation. https://doi.org/10.48550/.2005.11401.
25. Rahwan, I., et al. (2019). Machine behaviour. Nature, 568, 477-486. https://doi.org/10.1038/s41586-019-1138-y.
26. Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns, 11 (6), 887. https://doi.org/10.3390/healthcare11060887.
27. Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. Education Sciences, 13 (4), 410. https://doi.org/10.3390/educsci13040410.
28. Kasneci, E., et al. (2023). ChatGPT for good? Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274.
29. Farrokhnia, M., et al. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research, 61 (2), 209-221. https://doi.org/10.1080/14703297.2023.2195846.
30. Rao, A., et al. (2023). Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. Journal of Medical Internet Research, 25, e48659. https://doi.org/10.2196/48659.
31. He, S., et al. (2023). ChatGPT for scientific paper writing-Promises and perils. Innovation, 4 (6), 100524. https://doi.org/10.1016/j.xinn.2023.100524.
32. Nature Editorial. (2023). Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature, 613, 612. https://doi.org/10.1038/d41586-023-00191-1.
33. van Dis, E. A. M., et al. (2023). ChatGPT: Five priorities for research. Nature, 614, 224-226. https://doi.org/10.1038/d41586-023-00288-7.
34. Dennstädt, F., et al. (2024). Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. Systematic Reviews, 13, 130. https://doi.org/10.1186/s13643-024-02575-4.
35. Chang, Y., et al. (2024). Evaluation of large language models: A survey. ACM Computing Surveys, 56 (8), 159. https://doi.org/10.1145/3641289.
36. Huang, L., et al. (2023). Survey on hallucination in LLMs. . https://doi.org/10.48550/.2311.05232.

37. Touvron, H., et al. (2023). Llama 2. . https://doi.org/10.48550/.2307.09288.
38. Bai, Y., et al. (2023). LongBench. . https://doi.org/10.48550/.2308.14508.
39. Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA. . https://doi.org/10.48550/.2109.07958.
40. Emsley, R. (2023). Not hallucinations but fabrications. Schizophrenia, 9, 52. https://doi.org/10.1038/s41537-023-00379-4.
41. Gilson, A., et al. (2023). How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment JMIR Medical Education, 9, e45312. https://doi.org/10.2196/45312.
42. Kung, T. H., et al. (2023). Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. medRxiv. https://doi.org/10.1101/2022.12.19.22283643.
43. Zellers, R., et al. (2020). Defending against neural fake news. WWW '20, 351-362.
44. Mitchell, M., et al. (2019). Model cards for model reporting. FAccT '19. https://doi.org/10.1145/3287560.3287596.
45. Gebru, T., et al. (2021). Datasheets for datasets. Communications of the ACM, 64 (12), 86-92. https://doi.org/10.1145/3458723.
46. Chelli, M., & Rasheed, H. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. Journal of Medical Internet Research, 26, e53164. https://doi.org/10.2196/53164.
47. Bhattacharyya, M., et al. (2023). High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. Cureus, 15 (5), e39238. https://doi.org/10.7759/cureus.39238.
48. Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT. Cureus, 15 (2), e35179. https://doi.org/10.7759/cureus.35179.
49. Zielinski, C., et al. (2024). Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications, 40 (1), 11-13. https://doi.org/10.1080/03007995.2023.2286102.
50. COPE Council. (2023). COPE Authorship and AI tools. COPE. https://doi.org/10.24318/cCVRZBms.
51. Hendrycks, D., et al. (2021). Measuring Massive Multitask Language Understanding. . https://doi.org/10.48550/.2009.03300.
52. Kojima, T., et al. (2022). LLMs are zero-shot reasoners. . https://doi.org/10.48550/.2205.11916.
53. Jaime A. Teixeira da Silva, Serhii Nazarovets (2023). Can the principle of the ‘right to be forgotten’ be applied to academic publishing? Probe from the perspective of personal rights, archival science, open science and post-publication peer review, 36 (4), 431-437. https://doi.org/10.1002/leap.1579.
54. Elina Late, Raf Guns, Janne Pölönen, Jadranka Stojanovski, Mimi Urbanc, Michael Ochsner (2024). Beyond borders: Examining the role of national learned societies in the social sciences and humanities, 37 (2), 123-127. https://doi.org/10.1002/leap.1609.

55. Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625-630.
56. Hendriks, F., & Jucks, R. (2025). Generative AI in science communication. Science Communication. https://doi.org/10.1177/10755470251343486.
57. De Winter, J. C. F. (2024). Can ChatGPT-4 predict citation counts? Scientometrics, 129, 1531-1554. https://doi.org/10.1007/s11192-024-04939-y.
58. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... & Waddington, L. (2023). Testing of detection tools for AI-generated text. International Journal for Educational Integrity, 19(1), 1-39.
59. Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2024). A bibliometric review of large language models research from 2017 to 2023. ACM Transactions on Intelligent Systems and Technology, 15(5), 1-25.
60. Henderson, P., et al. (2017). On the reproducibility of deep RL. . https://doi.org/10.48550/.1708.04133.
61. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349 (6251), aac4716. https://doi.org/10.1126/science.aac4716.
62. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436-444. https://doi.org/10.1038/nature14539.
63. Liu, Z., et al. (2024). On the Detectability of ChatGPT Content: Benchmarking, Methodology, and Evaluation through the Lens of Academic Writing. KDD '24. https://doi.org/10.1145/3658644.3670392.
64. Lafia, S., Thomer, A., Fan, L., & Hemphill, L. (2022). Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network. Quantitative Science Studies, 3(3), 694–714. https://doi.org/10.1162/qss_a_00209
65. Chen, S., Brumby, D., & Cox, A. (2025, June). Envisioning the Future of Peer Review: Investigating LLM-Assisted Reviewing Using ChatGPT as a Case Study. In Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work (pp. 1-18).
66. Perkins, M., & Roe, J. (2024). Simple techniques to bypass GenAI text detectors: implications for inclusive education. IJETH, 21, 22. https://doi.org/10.1186/s41239-024-00487-w.
67. Garg, R. K., Urs, V. L., Agarwal, A. A., Chaudhary, S. K., Paliwal, V., & Kar, S. K. (2023). Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. Health Promotion Perspectives, 13(3), 183.
68. Kroll, J. A. (2018). The fallacy of inscrutability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180084.. https://doi.org/10.1145/3674981.
69. Hosseini, S. M., & Shu, K. (2023). Automated academic writing assistance: A survey. Information Processing & Management, 60 (6), 103275. https://doi.org/10.1016/j.ipm.2023.103275.
70. Wang, X., et al. (2024). Title and abstract screening with ChatGPT. Systematic Reviews, 13, 198. https://doi.org/10.1186/s13643-024-02620-2.
71. Brissett, A., & Wall, J. (2025). Machine learning and watermarking for accurate detection of AI generated phishing emails. Electronics, 14(13), 1-21.

72. Taha, T. A. E. A., Abdel-Qader, D. H., Alamiry, K. R., Fadl, Z. A., Alrawi, A., & Abdelsattar, N. K. (2024). Perception, concerns, and practice of ChatGPT among Egyptian pharmacists: a cross-sectional study in Egypt. BMC Health Services Research, 24(1), 1500.
73. Lund, B. D., & Wang, T. (2023). Generative AI and scholarly communication. Library Hi Tech News, 40 (3), 15-18. https://doi.org/10.1108/LHTN-01-2023-0009.
74. Zellers, R., et al. (2019). Grover. (NeurIPS, 2019). https://doi.org/10.48550/.1905.12616.
75. Ji, Z., et al. (2023). Hallucination in NLG survey. ACM Computing Surveys, 55 (12), 248. https://doi.org/10.1145/3571730.
76. UNESCO. (2023). Guidance for generative AI in education and research. UNESCO Publishing.