engineering proceedings

MDPI

Proceeding Paper

Comparative Evaluation of AI-Assisted Systematic Literature Reviews: NotebookLM, ChatGPT-4o, and Gemini 1.5 Pro in Wearable Technology, Smart Textiles, and Sustainable Fashion Research †

Selma Bulut [1,*], Beyza Buzol Mülayim [2] and Bozhana Stoycheva [3]

[1] Department of Computer Technologies, Technical Sciences Vocational School, Kirklareli University, 39100 Kirklareli, Turkey
[2] Department of Fashion Design, Technical Sciences Vocational School, Kirklareli University, 39100 Kirklareli, Turkey; beyza.mulayim@klu.edu.tr
[3] Department of Management and Social Activities, University of Ruse, 7017 Ruse, Bulgaria; bstoycheva@uni-ruse.bg
* Correspondence: selma.bulut@klu.edu.tr
† Presented at the International Conference on Electronics, Engineering Physics and Earth Science (EEPES2026), Bandirma, Turkey, 24–27 June 2026.

Abstract

This study compares three AI platforms, Google NotebookLM (Google Labs, May 2026 release/access), ChatGPT-4o, and Gemini 1.5 Pro, for systematic literature review (SLR) tasks in sustainable textile and fashion research. A corpus of 40 peer-reviewed articles (2019–2025) was queried using 12 standardized prompts across four dimensions: source traceability, thematic synthesis, research gap identification, and output utility. NotebookLM achieved the highest weighted score (8.63/10) due to its document-grounded, structurally constrained architecture, which produced no out-of-corpus citations under the conditions of this study. ChatGPT-4o (8.50/10) excelled in narrative synthesis; Gemini 1.5 Pro (7.78/10) produced superior structured tables. Five corpus-grounded and two inferred gaps, including XAI, Federated Learning, and Edge AI, were identified. A four-step hybrid AI workflow is proposed.

Keywords: NotebookLM; artificial intelligence; e-textile; wearable sensor; generative artificial intelligence

Check for updates

Academic Editor: Teodor Iliev, Ivaylo Stoyanov, Grigor Mihaylov and Selahattin Kosunalp

Published: 3 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The textile and fashion industry generates over 1.2 billion tons of CO_2 equivalent annually, more than the combined emissions of all international flights and maritime shipping, and its dyeing and treatment processes account for 20% of industrial water pollution globally [1], placing considerable pressure on manufacturers and designers to develop sustainable production models. Artificial intelligence (AI) technologies have emerged as transformative enablers across this sector, with demonstrated contributions in predictive materials design, automated defect identification, smart supply chain management, and wearable sensor integration [2–4].

The methodology by which the results of research are synthesized through the use of academic research synthesis is also being fundamentally changed. While it has been shown that generative AI systems can quickly and effectively process and analyze large

volumes of documents, there is still a significant lack of evidence regarding their ability to be reliable, reproducible, and provide accurate citations in domain-specific areas of research [5,6]. Google's release in 2023 of (and subsequent updates to) its NotebookLM system represents a new paradigm with regard to the way AI generates output. In contrast to traditional generative models that produce outputs as a result of generalizing across broad training data sets, NotebookLM produces outputs by referencing and limiting them to uploaded source documents provided by users [7]. The fact that NotebookLM is designed as a document-based architecture that limits output generation to user-provided source documents makes it very valuable for systematic literature reviews (SLRs), where accurate citation of sources is a critical component of the methodology. As a result of conducting a comprehensive review of the existing literature, it was found that to date, there are no empirical studies comparing the performance of NotebookLM, ChatGPT-4o, and Gemini 1.5 Pro in terms of generating abstracts using a domain-specific SLR workflow applied to sustainable textiles and fashion research. Studies conducted in the past have evaluated one algorithm in a specific application or have evaluated a single AI tool in a narrow domain, without providing domain-specific, multiple tool comparison frameworks [8–10]. Therefore, the purpose of this study is to create an empirical, repeatable evaluation framework directly applicable to engineering research workflows associated with the textile manufacturing industry.

1.1. Artificial Intelligence Applications in the Textile Sector

Artificial intelligence (AI) in textile engineering has been applied across two primary domains. In quality control and defect detection, CNN architectures (ResNet, InceptionV3) have achieved a defect detection accuracy above 95% under industrial conditions [11,12]; hybrid CNN-LSTM architectures for wearable e-textiles have improved real-time biometric signal classification by combining spatial feature extraction with temporal sequence modeling [13]. Biometric measurement in controlled stressor environments has further validated the utility of wearable sensor systems for stress and physiological monitoring in applied human-centered contexts [14]. In sustainable material development, AI has contributed on two complementary fronts: (1) high-accuracy signal processing enabling the use of biodegradable materials and environmentally friendly sensor substrates that reduce electronic waste, and (2) algorithmic optimization of production processes to minimize resource consumption [15]. Both areas align with the frameworks of Industry 4.0 and Industry 5.0, which emphasize human-centered and environmentally resilient production.

1.2. AI Tools in Systematic Literature Reviews

A systematic literature review is considered the most authoritative methodology to conduct an evidence synthesis within engineering and applied sciences; nonetheless, the systematic process requires a substantial amount of mental effort since researchers have to scan thousands of articles, classify them, and then synthesize the results, usually while working under pressure to complete these activities in a timely manner [6]. The generative nature of AI platforms offers possible assistance in reducing the workload of the researcher; however, there are also some serious limitations regarding the use of these systems for generating citations in academic research, such as the fact that generative AI platforms tend to produce believable statements that are not true [10].

NotebookLM's document-constrained architecture limits the possibility of AI generating false claims when responding to questions about the documents in the loaded set of documents [7], thus creating a verifiable, closed system which is appropriate for the workflow associated with SLRs. On the other hand, both ChatGPT-4o and Gemini 1.5 Pro utilize open generative architectures that allow AI to include information from its

training data in each of its respective corpus-specific analyses, which increases the richness of synthesis as well as the chance of an AI generating false claims.

2. Method

2.1. Research Design

The study follows a three-phase study plan according to the guidelines for conducting systematic reviews and meta-analyses, as per PRISMA 2020 [15]; i.e., (1) the generation of a source text set; (2) the application of an assessment tool; and (3) comparative evaluation based on four dimensions.

This design emphasizes methodological comparability, reproducibility, and transparency so that other researchers may replicate this study.

2.2. Source Corpus Construction

A literature search was conducted using Boolean keyword combinations from Web of Science: ("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning") AND ("Smart Textiles" OR "Wearable Sensors" OR "Electronic Textiles" OR "Sustainable Fashion"). The search was limited to the years 2019–2025, initially yielding 162 records. Following a systematic title/abstract search and full-text review, 40 peer-reviewed articles were selected for the final compilation (See Supplementary Materials, Table S1). Table 1 summarizes the keyword set composition and the distribution of articles across thematic areas.

Table 1. Source corpus composition by keyword cluster (total $N = 40$).

Search Keyword Cluster	Initial Hits	Representative Subtopics	Final Articles Included
AI + Smart Textiles/E-Textiles	61	Deep learning for defect detection, CNN-LSTM wearable classification	18
AI + Wearable Sensors/Biometrics	44	Biometric signal processing, MXene sensors, biodegradable substrates	12
AI + Sustainable Fashion	35	Circular economy, zero-waste design, supply chain optimization	7
Deep Learning + Textile Quality Control	22	ResNet/InceptionV3 defect detection, real-time classification	3
TOTAL	162	After PRISMA 2020 screening and full-text review	40

Exclusion criteria were applied in two sequential stages, title/abstract screening and full-text review, consistent with PRISMA 2020 [16] (See Supplementary Materials, Table S2). Table 2 presents the six exclusion codes, corresponding exclusion reasons, and the number of records removed at each stage.

Table 2. PRISMA 2020 exclusion criteria and records excluded at each stage.

Code	Exclusion Criterion	Records Excluded	Stage Applied
E1	Language other than English or Turkish	1	Title/Abstract
E2	Publication date outside 2019–2025	11	Title/Abstract
E3	No full-text access available	67	Full-Text
E4	No AI/ML method applied	3	Full-Text

Table 2. *Cont.*

Code	Exclusion Criterion	Records Excluded	Stage Applied
E5	Insufficient textile/wearable sensor content	2	Full-Text
E6	Gray literature/non-peer-reviewed	38	Screening
Total Excluded		122	

Figure 1 presents the corresponding PRISMA 2020 flow diagram, summarizing the identification, screening, and exclusion stages described above in a format consistent with standard engineering systematic-review reporting.

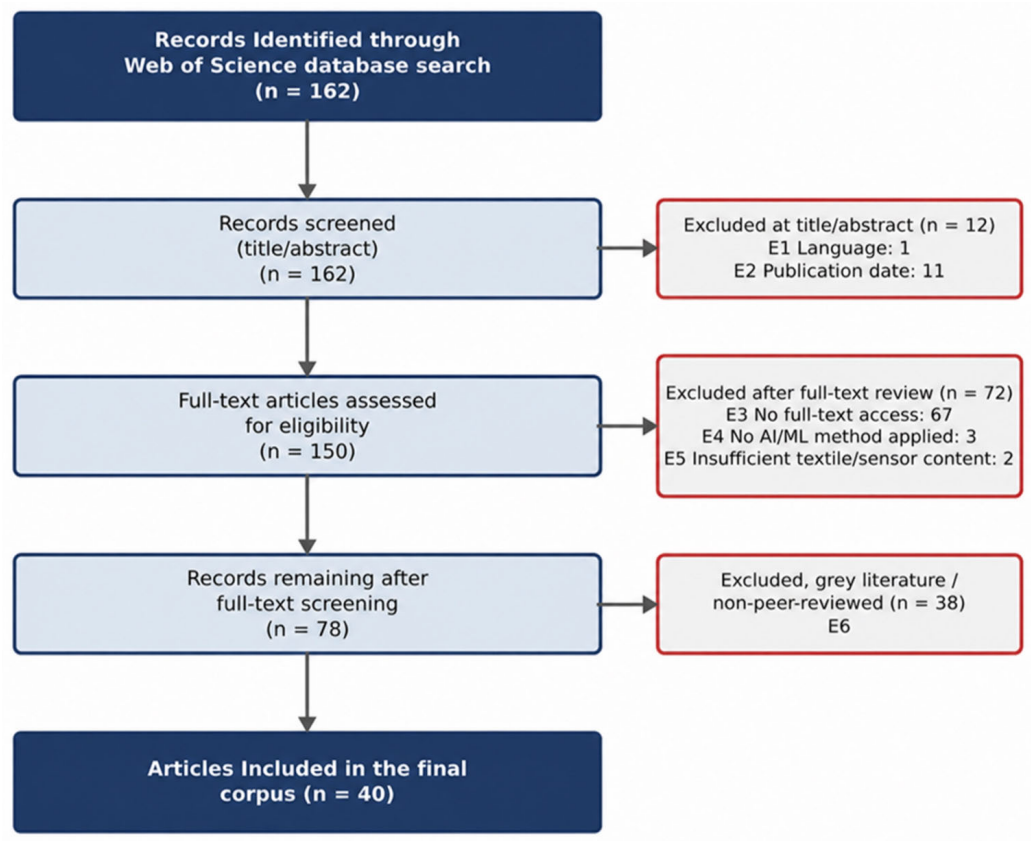


Figure 1. PRISMA 2020 flow diagram of record identification, screening, and exclusion (codes E1–E6).

2.3. Tool Application Protocol

The 40-article corpus was duplicated into three identical PDFs for use with the three different interfaces. One dedicated Notebook was created in NotebookLM; an environment in which to upload and access the full corpus was created using the Custom GPT option in ChatGPT-4o, since this is limited to about ten PDFs at one time through the standard user interface. With respect to Gemini 1.5 Pro, a Google AI Studio environment (version as of March 2025) was utilized, allowing for ingestion of all 40 articles as part of a single session due to its million-token context window.

A total of 12 standard prompts were used on each interface across four consecutive groups: (A) Thematic Mapping, (B) Research Gap Analysis, (C) Source Verification, and (D) Synthesis and Output Generation. See Table 3 for additional details regarding the prompt protocols(See Supplementary Materials, Table S3).

Table 3. Standardized 12-prompt query protocol applied across all three platform.

Block	Q#	Prompt (Abbreviated)	Dimension Assessed
A	Q1	List all AI/ML methods identified in the corpus with article counts	D2—Thematic Synthesis
A	Q2	Map thematic clusters: group articles by research domain	D2—Thematic Synthesis
A	Q3	Generate a structured table: AI method vs. sustainability metric	D2/D4
B	Q4	Identify underrepresented methods or topics in the corpus	D3—Research Gap
B	Q5	What future research directions are explicitly recommended?	D3—Research Gap
B	Q6	Which AI approaches are absent from this corpus?	D3—Research Gap
C	Q7	Provide exact quotations from specific article [11]	D1—Source Accuracy
C	Q8	List all articles in corpus that use CNN-based architectures	D1—Source Accuracy
C	Q9	Verify: Is ResNet used for defect detection in this corpus?	D1—Source Accuracy
D	Q10	Synthesize the five most important findings as bullet summary	D4—Output Format
D	Q11	Generate a structured literature review outline for a conference paper	D4—Output Format
D *	Q12	Generate Audio Overview (NotebookLM only): focus on AI in sustainable textiles	D4—Multimodal

* Q12 (Audio Overview) is exclusive to NotebookLM and was evaluated as a supplementary multimodal output capability within dimension D4. ChatGPT-4o and Gemini 1.5 Pro lack a comparable native audio generation feature. This introduces an asymmetry in D4 scoring, which is acknowledged as a limitation of the current evaluation design.

2.4. Four-Dimensional Evaluation Framework

Tool responses were scored using a weighted four-dimensional rubric. Dimension weights were assigned to reflect the relative importance of each criterion to the validity of systematic literature reviews in engineering research (See Supplementary Materials, Table S4). D1 (Source Accuracy, 30%) carries the highest weight because citation reliability is the foundational requirement of any SLR methodology; an SLR built on unverifiable or hallucinated citations is structurally compromised regardless of its analytical quality. D2 (Thematic Synthesis, 25%) and D3 (Research Gap Identification, 25%) share equal weight, reflecting the comparable importance of breadth of coverage and originality of contribution in engineering reviews. D4 (Output Format Utility, 20%) is weighted lower because formatting, while practically valuable, does not affect the epistemic validity of the review. For dimension D2 (Thematic Synthesis Quality), scoring was performed by two independent evaluators, a textile engineer with eight years of research experience and a computer engineer specializing in AI, following a structured calibration session conducted prior to independent scoring. During calibration, both evaluators independently scored three sample responses from outside the main corpus, then discussed discrepancies until a consensus was reached on rubric interpretation, particularly regarding the distinction between methodological enumeration and genuine thematic synthesis. This calibration protocol was designed to minimize evaluator-specific interpretation drift. Inter-rater agreement on the main corpus was substantial (Krippendorff's $\alpha = 0.82$), and the mean of the two scores was used for all D2 evaluations. Table 4 presents the complete evaluation framework with scoring methods for each dimension.

Table 4. Four-dimensional weighted evaluation framework.

Dimension	Weight	Description	Evaluation Method
D1: Source Accuracy	30%	Citation reliability and corpus traceability. Each cited reference is verified against the uploaded PDF corpus.	Correct citations/total cited references × 10
D2: Thematic Synthesis	25%	Thematic depth, sub-category richness, and ability to distinguish review papers from primary studies.	1–5 point rubric scored by two independent evaluators (a textile engineer with 8 years of research experience and a computer engineer specializing in AI); calibration session conducted prior to scoring using three sample responses external to the main corpus, with evaluators reconciling score discrepancies to align rubric interpretation; inter-rater agreement: Krippendorff's $\alpha = 0.82$ (substantial); mean score used for all D2 evaluations
D3: Research Gap Identification	25%	Identification of original, corpus-supported gaps not explicitly stated in any individual source document.	Number of corpus-grounded original gaps identified
D4: Output Format Utility	20%	Usability of structured outputs (tables, outlines, audio), including multimodal capabilities.	Score per functional output type (0–2 scale per sub-item)

3. Findings

3.1. D1: Source Accuracy and Citation Reliability

Source accuracy was assessed by verifying each citation generated in response to Q7, Q8, and Q9 against the uploaded corpus. NotebookLM demonstrated perfect citation fidelity: all 11 references generated were verified to correspond to actual corpus documents, with traceability provided at the file level (e.g., 17.pdf [11]). No out-of-corpus citations were produced, yielding a D1 score of 9.0/10.

ChatGPT-4o generated 14 citations across the three questions, of which 11 (79%) were correctly matched to corpus documents. However, three references were identified as out-of-corpus hallucinations, papers not present in the uploaded set but presented with plausible metadata. Gemini 1.5 Pro referenced 10 sources, of which 8 (80%) were correctly corpus-matched. Critically, Gemini introduced two non-corpus methodological examples (Box–Behnken and Taguchi experimental designs), likely sourced from its training distribution rather than the uploaded documents. Table 5 provides a detailed comparative breakdown.

Table 5. D1 source accuracy: detailed citation verification results.

Metric	NotebookLM	ChatGPT-4o	Gemini 1.5 Pro	Ideal Benchmark
Total citations generated (Q7–Q9)	11	14	10	—
Correctly matched to corpus	11 (100%)	11 (79%)	8 (80%)	100%
Out-of-corpus hallucinated citations	0	3	2 (method-level)	0
File-name/page-level traceability	Yes (e.g., 17.pdf)	No	No	Yes
D1 score	9.0 *	8.5	7.5	10.0

* = highest D1 score among the three platforms.

https://doi.org/10.3390/engproc2026154029

3.2. D2: Thematic Synthesis Quality

The authors judged the quality of thematic synthesis by evaluating the breadth, depth, and level of detail (methodological granularity) at which each tool recognized the categories of AI methods for Questions 1–3. The NotebookLM tool identified an overall larger number of AI methodologies than either the ChatGPT-4o or Gemini 1.5 Pro tools; this included low-visibility architectures (e.g., DBNs, RBMs, Variational Gaussian Processes) that neither of these two tools surfaced from their respective corpora. As a result, the NotebookLM received a D2 score of 8.0/10.

The ChatGPT-4o tool received the highest D2 score (8.5/10) due to its qualitatively different contribution: it spontaneously introduced a review-versus-primary paper taxonomy that provided analysis above simply enumerating methods. This taxonomy could be applied directly to literature review writing. Lastly, the Gemini 1.5 Pro tool produced the greatest amount of publication-ready table output on Question 3, creating a table summarizing AI methods relative to sustainability measures in a format that could be transferred into manuscript tables. A summary of methodology identification among all three tools can be found in Table 6.

Table 6. AI method recognition by tool across the 40-article corpus.

AI Method Category	# Articles in Corpus	NotebookLM Recognized	ChatGPT-4o Recognized	Gemini 1.5 Pro Recognized
CNN-based defect detection (ResNet, InceptionV3)	14	Yes	Yes	Yes
CNN-LSTM hybrid (wearable biometrics)	9	Yes	Yes	Yes
Graph Neural Networks (GNNs)	4	Yes	Yes	Partial
Generative AI (GAN, diffusion models)	5	Yes	Yes	Yes
Transformer/BERT-based NLP	3	Yes	Yes	Yes
DBN/RBM (low-visibility methods)	2	Yes *	No	No
Variational Gaussian Processes	1	Yes *	No	No
Reinforcement Learning	0 (gap; see Table 7)	Gap noted	Gap noted	Not noted

* = low-visibility method identified exclusively by NotebookLM; not surfaced by ChatGPT-4o or Gemini 1.5 Pro.

3.3. D3: Research Gap Identification

Research gap identification capacity was evaluated based on the number of corpus-grounded, original gaps identified, that is gaps supported by direct textual evidence in the uploaded sources rather than inferred from the tool's training data (See Supplementary Materials, Table S5). NotebookLM identified five original gaps, each accompanied by direct quotations from specific corpus documents, achieving a D3 score of 9.0/10. ChatGPT-4o identified a sixth gap, physics-informed neural networks (PINNs) for textile material simulation, derived from cross-domain reasoning rather than direct corpus evidence, which is valuable for D2 but less credible for strict D3 criteria (8.5/10). Gemini 1.5 Pro did not identify sustainability computation costs as a gap and failed to surface several corpus-supported gaps (7.5/10). Table 7 presents all identified research gaps with domain descriptions and suggested future directions.

Table 7. Research gaps identified across AI tools with corpus support and future directions.

Gap Domain	Detailed Description	Corpus-Supported	Tool Identified	Suggested Future Direction
Explainable AI (XAI)	Black-box deep learning models lack interpretability for safety-critical wearable applications	Yes (NotebookLM)	NLM, GPT	Integrate SHAP/LIME into CNN-LSTM wearable pipelines
Federated Learning (FL)	Privacy-preserving distributed training absent in current textile AI literature	Yes (NotebookLM)	NLM, GPT	Apply FL to multi-site textile defect data without central data sharing
Green AI/Sustainable Compute	Energy costs of model training not reported; carbon footprint of AI ignored	Yes (NotebookLM)	NLM	Report FLOPs and CO_2 equivalents alongside model accuracy
Edge AI/TinyML	Low-power embedded deployment rarely addressed; most models designed for cloud	Yes (NotebookLM)	NLM, GPT	Develop TinyML-quantized models for on-textile microcontrollers (e.g., Arduino Nano BLE)
Dataset Diversity	Publicly available textile defects and biometric datasets are limited and non-diverse	Yes (NotebookLM)	NLM, GPT	Establish open benchmark datasets across textile types and demographic groups
Physics-Informed NNs (PINNs)	Material property modeling could benefit from physics-constrained learning	Inferred (GPT)	GPT only	Apply PINNs for fiber behavior simulation and smart material optimization
Reinforcement Learning for Process Optimization	RL for textile manufacturing process control (e.g., dyeing, cutting, assembly line scheduling) is entirely absent from the corpus	Not present (identified as absence)	NLM, GPT	Apply RL for adaptive control of textile production parameters (dyeing temperature, cutting speed, assembly scheduling)

3.4. D4: Output Format Utility

Output format utility assessed the usability and multimodal range of tool outputs. NotebookLM and ChatGPT-4o achieved equal scores (8.5/10), with NotebookLM producing compact structured outlines and ChatGPT-4o generating publication-ready narrative summaries. Gemini 1.5 Pro's tabular output quality is the highest among the three tools, with the Q3 AI-method/sustainability table directly usable in a manuscript without reformatting (8.0/10).

NotebookLM's Audio Overview feature, exclusive among the three platforms, generates a downloadable MP3 in which two AI speakers discuss the corpus content in podcast format (5–10 min). This feature was assessed as a unique multimodal knowledge dissemination mechanism with direct utility for conference presentations and educational outreach in engineering contexts.

3.5. Aggregate Scoring

The final weighted totals for each of the four factors are shown in detail in Table 8.

Table 8. Four-dimensional evaluation results—weighted scores (out of 10).

Dimension	Weight	NotebookLM	ChatGPT-4o	Gemini 1.5 Pro	Max
D1: Source Accuracy	30%	9.0 *	8.5	7.5	10.0
D2: Thematic Synthesis	25%	8.0	8.5 *	8.0	10.0
D3: Research Gap Identification	25%	9.0 *	8.5	7.5	10.0
D4: Output Format Utility	20%	8.5 *	8.5 *	8.0	10.0
WEIGHTED TOTAL	—	8.63 *	8.50	7.78	10.0

* = highest score in dimension.

The highest weighted total achieved was 8.63/10 from NotebookLM, with 8.50/10 from ChatGPT-4o and 7.78/10 from Gemini 1.5 Pro. No single tool achieved dominant performance across all dimensions, which provides the empirical basis for the hybrid workflow proposed in Section 4.3.

To support direct visual comparison of tool performance alongside the quantitative results in Table 8, Figure 2 plots the four weighted dimension scores and overall totals for the three platforms.

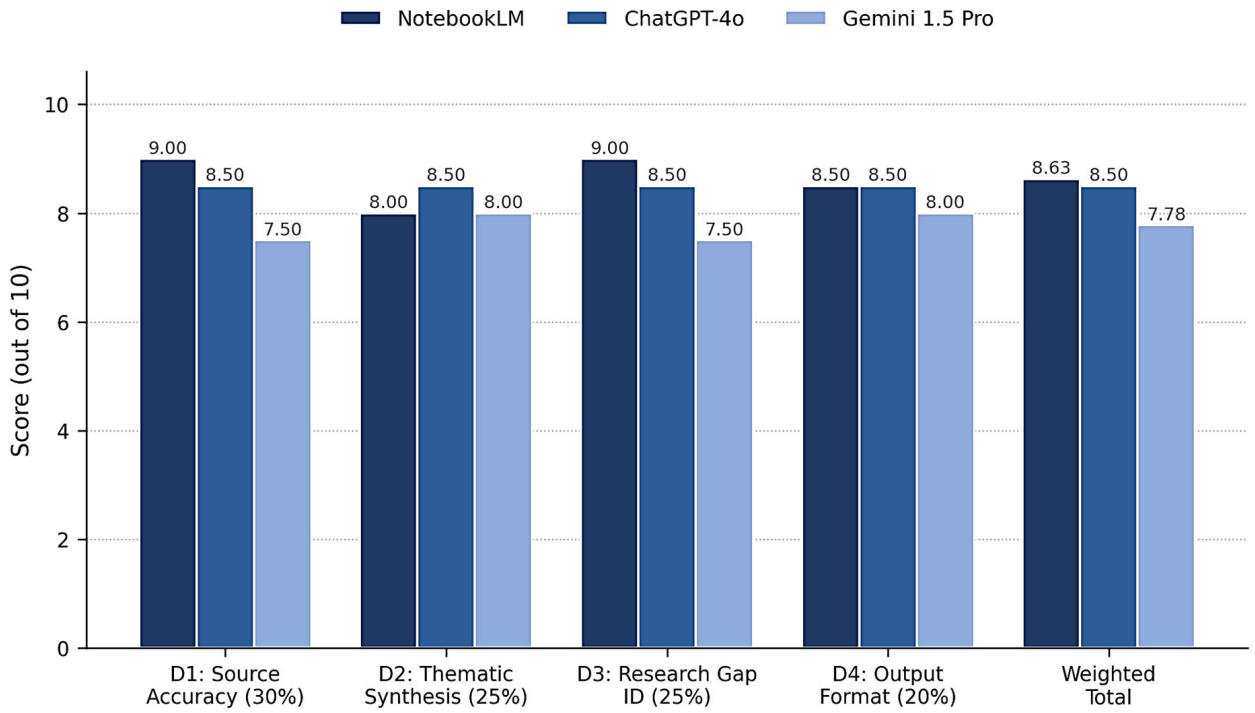


Figure 2. Four-dimensional weighted evaluation scores (D1–D4) and overall weighted totals for NotebookLM, ChatGPT-4o, and Gemini 1.5 Pro.

4. Discussion

4.1. Architectural Differences and Research Implications

The findings indicate that the choice of tools to use does not constitute an innocuous methodological decision in this regard; each of these architectures produces different results, which are both qualitatively and quantitatively unique based on their respective capabilities

and shortcomings. Although Gemini 1.5 Pro is capable of producing a wide variety of high-quality outputs, NotebookLM was specifically designed to be a document-based tool. Its architecture structurally constrains output generation to the uploaded corpus, substantially reducing, though not categorically eliminating, the risk of out-of-corpus citation under conditions where prompts remain within the scope of the loaded documents. Therefore, NotebookLM provides citation-level traceability that is critical in establishing valid systematic reviews. Thus, while Gemini 1.5 Pro can create many types of high-quality content, NotebookLM is the preferred tool when academic integrity is most important and when validating a large body of information is required (e.g., corpus verification).

The majority of current SLR tools still rely on relatively outdated classification methods (e.g., SVM, Naive Bayes) and text representation techniques such as bag-of-words, while the integration of LLMs into these tools remains in its early stages; although LLM-based systems show promising potential for literature search and academic writing support, the need for verification mechanisms persists due to fundamental challenges such as hallucination and inconsistent outputs [5]. In the present study, Gemini 1.5 Pro's million-token context window facilitated large-scale corpus ingestion; however, the model did not consistently prioritize uploaded corpus content over parametric knowledge, limiting its reliability in citation-sensitive tasks.

4.2. Textile Research-Specific Findings

The gap analysis identified five corpus-supported and one inferred research gap within the textile AI literature. The absence of explainable AI (XAI) techniques in wearable sensor systems is particularly significant given the safety-critical nature of health monitoring applications; black-box models that cannot provide output explanations are unsuitable for regulatory and clinical deployment contexts. Similarly, the absence of Federated Learning approaches represents a missed opportunity for privacy-preserving multi-institutional textile data collaboration. The lack of Edge AI and TinyML architectures indicates that most current models are designed for cloud or server-side inference, incompatible with resource-constrained on-garment microcontrollers.

4.3. Proposed Hybrid AI Research Framework

The empirical findings indicate that no single tool achieves optimal performance across all evaluation dimensions. This motivates a four-step hybrid workflow that allocates tasks to the platform best suited for each function. Table 9 presents the hybrid framework with estimated efficiency gains relative to unassisted manual methods.

This framework is designed to be platform-agnostic in the sense that individual tools can be replaced with successor models as they are released, while the functional logic of the four-step sequence remains stable. Note: The estimated time savings reported in Table 9 are based on the authors' direct experience conducting this study and should be interpreted as indicative approximations rather than empirically validated benchmarks; controlled pilot studies comparing assisted and unassisted SLR workflows are recommended for future quantification.

Figure 3 renders the four-step hybrid workflow of Table 9 as a process diagram, making the sequential handoff between tools, and the corresponding engineering outputs and efficiency gains, immediately legible for practitioners designing their own SLR pipelines.

Table 9. Proposed four-step hybrid AI research framework for systematic literature reviews.

Step	Tool	Function	Output	Estimated Time Saving
1	NotebookLM	Corpus upload and source-grounded citation verification	Verified citation list with file-level traceability	~60–70% vs. manual screening
2	ChatGPT-4o	Narrative synthesis and interdisciplinary inference	Discussion/Conclusion draft sections	~50% vs. unassisted writing
3	Gemini 1.5 Pro	Structured data extraction and table generation	Publication-ready tables (AI method × sustainability metric)	~40% vs. manual table construction
4	NotebookLM	Audio Overview for conference dissemination	MP3 podcast-style summary (5–10 min, two AI speakers)	Novel multimodal capability

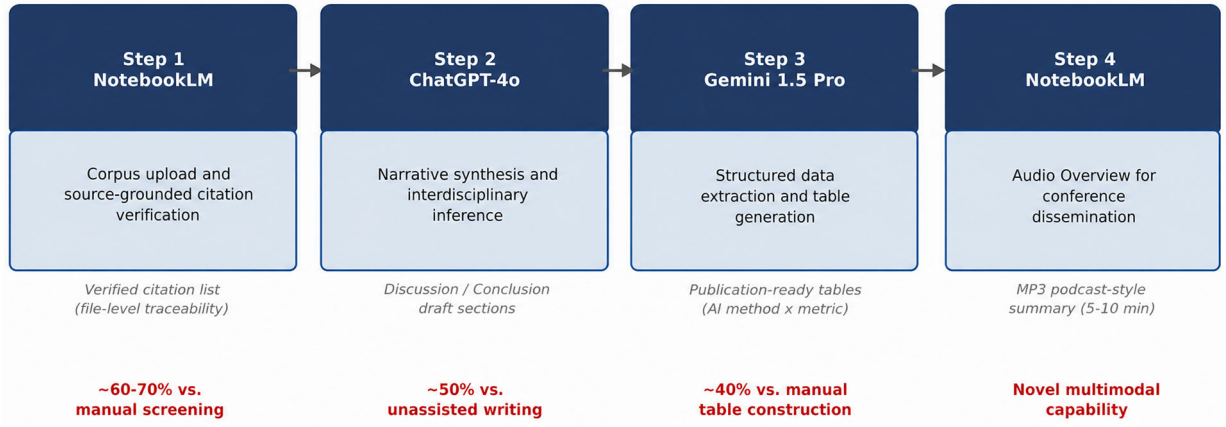


Figure 3. Proposed four-step hybrid AI research workflow for systematic literature reviews.

5. Conclusions

This study presents the first systematic, multi-dimensional comparison of NotebookLM, ChatGPT-4o, and Gemini 1.5 Pro applied to a domain-specific systematic literature review in sustainable textile and fashion research. Using a weighted four-dimensional rubric applied to a PRISMA 2020-selected 40-article corpus, the study demonstrates that: (1) NotebookLM achieves superior source traceability with structurally reduced hallucination risk (zero out-of-corpus citations were observed in this study under the applied prompt conditions, though this result should not be generalized as an absolute property of the platform); (2) ChatGPT-4o excels in narrative synthesis and cross-domain analytical inference; and (3) Gemini 1.5 Pro produces the highest-quality structured tabular outputs. Five empirically grounded research gaps, XAI in wearables, Federated Learning, Green AI/sustainable compute, Edge AI/TinyML, and dataset diversity, are formally identified from corpus evidence.

The proposed four-step hybrid AI framework provides a replicable, efficiency-enhancing workflow for SLR practitioners in engineering and materials science domains. Limitations include the restriction to a single database (Web of Science) and a corpus of 40 articles. Furthermore, given the rapid development cycle of generative AI platforms, the performance characteristics reported here reflect the specific model versions available in early 2025; results may differ as underlying models are updated or replaced. Future work should extend the corpus across multiple databases, incorporate inter-rater reliability coefficients for rubric scoring, and conduct longitudinal re-evaluations as AI tool capabilities evolve.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/engproc2026154029/s1, Table S1: Full 40-Article Source Corpus (PRISMA 2020 Selected); Table S2: PRISMA 2020 Screening Flow: Records at Each Stage; Table S3: Full 12-Prompt Query Protocol with Verbatim Prompt Texts; Table S4: Four-Dimensional Evaluation Rubric: Full Scoring Detail; Table S5: Research Gaps: Full Descriptions, Corpus Support, and Future Directions.

Author Contributions: Conceptualization, S.B. and B.B.M.; methodology, S.B.; software, S.B.; validation, S.B., B.B.M. and B.S.; formal analysis, B.B.M. and B.S.; investigation, S.B., B.B.M. and B.S.; resources, B.B.M.; data curation, S.B.; writing—original draft preparation, S.B. and B.B.M.; writing—review and editing, S.B. and B.B.M.; visualization, S.B.; supervision, B.S.; project administration, S.B. and B.S.; funding acquisition, B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partially financed by the European Union—Next GenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No. BG-RRP-2.013-0001-C01.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank the anonymous reviewers for their constructive feedback.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SLR	Systematic Literature Review
AI	Artificial Intelligence
CNN	Convolutional Neural Network
GNN	Graph Neural Network
GAN	Generative Adversarial Network
DBN	Deep Belief Network
RBM	Restricted Boltzmann Machine
XAI	Explainable AI

References

1. Ellen MacArthur Foundation. A New Textiles Economy: Redesigning Fashion's Future. EMF Report, 2017. Available online: https://www.ellenmacarthurfoundation.org/assets/downloads/publications/A-New-Textiles-Economy_Full-Report.pdf (accessed on 1 April 2026).
2. Ramos, L.; Rivas-Echeverría, F.; Pérez, A.G.; Casas, E. Artificial intelligence and sustainability in the fashion industry: A review from 2010 to 2022. *SN Appl. Sci.* **2023**, *5*, 387. [CrossRef]
3. Xu, Y.; Sun, R.; Zhi, C.; Liu, Z.; Chen, J.; Ke, Z.; Yu, L. Zero-defect manufacturing in the textile industry: A review of current advances and challenges. *Text. Res. J.* **2025**, *95*, 194–219. [CrossRef]
4. Nisa, H.; Van Amber, R.; English, J.; Islam, S.; McCorkill, G.; Alavi, A. A Systematic Review of Reimagining Fashion and Textiles Sustainability with AI: A Circular Economy Approach. *Appl. Sci.* **2025**, *15*, 5691. [CrossRef]
5. Bolaños, F.; Salatino, A.; Osborne, F.; Motta, E. Artificial intelligence for literature reviews: Opportunities and challenges. *Artif. Intell. Rev.* **2024**, *57*, 259. [CrossRef]
6. Chelli, M.; Descamps, J.; Lavoué, V.; Trojani, C.; Azar, M.; Deckert, M.; Raynier, J.L.; Clowez, G.; Boileau, P.; Ruetsch-Chelli, C. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *J. Med. Internet Res.* **2024**, *26*, e53164. [CrossRef] [PubMed]
7. Google. NotebookLM: AI-Powered Research Assistant. Available online: https://notebooklm.google.com (accessed on 10 April 2026).

8. Li, M.; Sun, J.; Tan, X. Evaluating the effectiveness of large language models in abstract screening: A comparative analysis. *Syst. Rev.* **2024**, *13*, 219. [CrossRef] [PubMed]
9. Scherbakov, D.; Hubig, N.; Jansari, V.; Bakumenko, A.; Lenert, L.A. The emergence of large language models as tools in literature reviews: A large language model-assisted systematic review. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 1071–1086. [CrossRef] [PubMed]
10. Alshami, A.; Elsayed, M.; Ali, E.; Eltoukhy, A.E.E.; Zayed, T. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems* **2023**, *11*, 351. [CrossRef]
11. Alruwais, N.; Alabdulkreem, E.; Mahmood, K.; Marzouk, R.; Assiri, M.; Abdelmageed, A.A.; Abdelbagi, S.; Drar, S. Hybrid mutation moth flame optimization with deep learning-based smart fabric defect detection. *Comput. Electr. Eng.* **2023**, *108*, 108706. [CrossRef]
12. Malathi, S.; Aiswarya, P.U. Comparative Analysis of Suitability of Deep Learning Models in Quality Assurance of Fabrics. *Int. Res. J. Multidiscip. Technovation* **2025**, *7*, 41–57. [CrossRef]
13. Xu, L.; Chen, W.; Li, Q.; Gupta, R.; Bagchi, B.; Lovat, L.B.; Tiwari, M.K. Printing Nacre-Mimetic MXene-Based E-Textile Devices for Sensing and Breathing-Pattern Recognition Using Machine Learning. *Adv. Funct. Mater.* **2025**, *35*, e08370. [CrossRef]
14. Antonova, D.; Beloeva, S. Controlled Anxiety for Stress Management and HR Innovation in Creative Industries (Measurement by Biometrics in a Controlled Stressor Environment). In Proceedings of the International Conference on Business Excellence, Bucharest, Romania, 1 July 2025; pp. 3649–3655. [CrossRef]
15. Zhao, P.; Song, Y.; Hu, Z.; Zhong, Z.; Li, Y.; Zhou, K.; Qin, T.; Yan, Y.; Hsiao Hsuan, H.; Han, S.-T.; et al. Artificial intelligence enabled biodegradable all-textile sensor for smart monitoring and recognition. *Nano Energy* **2024**, *124*, 110118. [CrossRef]
16. Page, M.J.; Moher, D.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ* **2021**, *372*, n160. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.