Indo American Journal of Multidisciplinary Research and Review (IAJMRR)
International Peer Reviewed - Refereed Research Journal
ISSN: 2581 - 6292, Impact Factor: 6.885, Website: www.iajmrr.com

DETECTING SYNTHETIC EVIDENCE CONTAMINATION IN AI-ASSISTED SCIENTIFIC LITERATURE THROUGH PROVENANCE-AWARE LEARNING

Divya Vishwanath
Fellow Member, Sarojini Research and Development Council, Palakkad, Kerala, India

Cite This Article: Divya Vishwanath, "Detecting Synthetic Evidence Contamination in AI-Assisted Scientific Literature Through Provenance-Aware Learning", Indo American Journal of Multidisciplinary Research and Review, Volume 10, Issue 2, July - December, Page Number 46-57, 2026.

Copy Right: © IAJMRR Publication, 2026 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Type of Review: Peer Reviewed as per |C|O|P|E| Guidance.

Disclaimer: The scholarly papers reviewed and published by IAJMRR Publication, India, reflect the views and opinions of their respective authors and do not necessarily represent the views or opinions of IAJMRR Publication. The publisher disclaims any responsibility for any harm, loss, or damage resulting from the use of the published content by any party.

DOI: https://doi.org/10.5281/zenodo.24273516

Abstract:

The increasing adoption of artificial intelligence (AI) and large language models (LLMs) in scientific research has transformed literature synthesis, citation generation, evidence interpretation, and scholarly communication. However, the growing dependence on AI-assisted research tools has introduced significant challenges involving fabricated references, unsupported numerical findings, citation inconsistencies, and unverifiable scientific claims. These problems contribute to synthetic evidence contamination, potentially undermining scientific credibility, research reproducibility, and evidence-based decision-making. This study proposes an integrated Provenance-Aware Learning Framework (PALF) for identifying and evaluating synthetic evidence contamination through machine learning, natural language processing, scientific claim verification, and bibliographic provenance analysis. The proposed framework integrates conventional machine learning classifiers, transformer-based semantic representations, citation metadata verification, and source traceability indicators to distinguish verifiable scientific evidence from potentially fabricated or unsupported research content.

The study adopts a quantitative comparative research design using publicly available scientific evidence verification resources, including SciFact, PubMed, Crossref, and OpenAlex. The experimental framework compares Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, transformer-based classification, and provenance-aware learning using accuracy, precision, recall, F1-score, ROC-AUC, and false-positive rate. The original SciFact benchmark contains 1,409 scientific claims and 5,183 scientific abstracts, providing a documented numerical foundation for scientific claim verification. Additional experimental procedures are designed to evaluate fabricated citations, unsupported quantitative statements, citation-claim inconsistencies, and source metadata discrepancies.

The proposed numerical evaluation includes 10 comparative tables and three research figures covering dataset characteristics, contamination distributions, classification performance, citation verification, provenance-feature contributions, and detection robustness. The study investigates whether combining semantic representations with bibliographic provenance improves evidence verification compared with text-only classification approaches. By incorporating interpretable source-validation indicators, the framework aims to support transparent identification of suspicious scientific claims and reduce dependence on linguistic detection alone. Although benchmark dataset statistics are documented, comparative model outcomes will require experimental validation.

The proposed research contributes an integrated and reproducible approach to scientific evidence integrity assessment, emphasizing verifiable citations, numerical consistency, source authenticity, and explainable detection decisions. Its potential applications include automated manuscript screening, systematic literature review validation, academic publishing quality assurance, and responsible AI-assisted scientific research.

Key Words: Synthetic Evidence Contamination; Provenance-Aware Learning; Scientific Literature; Artificial Intelligence; Machine Learning; Citation Verification; Scientific Claim Verification; Research Integrity; Explainable AI.

1. Introduction:

1.1 Background of the Study:

Artificial intelligence (AI) has significantly transformed scientific research by supporting automated literature reviews, scientific writing, citation generation, evidence extraction, and knowledge discovery. Large language models (LLMs) and machine learning algorithms enable researchers to process extensive scientific information and generate structured academic content efficiently. However, the increasing use of AI-assisted research tools has introduced concerns regarding fabricated citations, unsupported numerical

findings, misleading scientific interpretations, and unverifiable research evidence. These challenges have created a growing need for automated mechanisms capable of assessing the authenticity and reliability of scientific information.

Synthetic evidence contamination refers to the inclusion or propagation of fabricated, manipulated, misrepresented, or unsupported evidence within scientific literature and AI-assisted research outputs. Unlike ordinary textual inaccuracies, evidence contamination can directly affect scientific conclusions, research reproducibility, systematic reviews, and evidence-based decisions. For example, AI-generated scientific content may include realistic but nonexistent references or report numerical findings that differ substantially from the original publications. Such inconsistencies may remain undetected when verification relies exclusively on linguistic analysis. Scientific literature databases provide opportunities for developing automated evidence verification systems. The SciFact benchmark introduced by Wadden et al. (2020) contains 1,409 scientific claims and 5,183 scientific abstracts, establishing a numerical foundation for scientific claim verification. Bibliographic resources such as Crossref, OpenAlex, and PubMed additionally support publication identification, citation metadata validation, and source traceability. Integrating these resources with machine learning offers a promising direction for detecting synthetic evidence contamination.

1.2 Research Problem:

Existing AI-generated text detection approaches primarily examine linguistic patterns, semantic characteristics, and statistical representations. Although these methods may identify certain forms of synthetic content, they cannot reliably establish whether scientific claims are supported by authentic research evidence. A fabricated citation may resemble a legitimate reference, while a genuine publication may be incorrectly cited to support an unsupported numerical result. Furthermore, conventional citation verification approaches frequently focus on publication existence rather than the consistency between scientific claims and their cited sources. Consequently, verifying the authenticity of scientific evidence requires a comprehensive framework that integrates textual classification, citation verification, numerical consistency assessment, and source provenance analysis. The central research problem addressed in this study is the absence of a unified and interpretable detection methodology that systematically evaluates both the semantic validity and provenance reliability of AI-assisted scientific evidence.

1.3 Proposed Provenance-Aware Learning Framework:

This study proposes a Provenance-Aware Learning Framework (PALF) that combines conventional machine learning, transformer-based semantic analysis, and bibliographic provenance verification. The framework evaluates scientific evidence through three principal components: textual evidence analysis, source authenticity verification, and provenance-based contamination classification. Textual evidence analysis examines scientific claims and their supporting contexts using natural language processing and machine learning. Source authenticity verification evaluates citation metadata, publication identifiers, and bibliographic consistency using scientific literature databases. Provenance-based classification integrates these characteristics to identify potentially contaminated evidence and generate interpretable detection outcomes. The proposed framework compares Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, transformer-based classification, and provenance-aware learning. Classification performance will be assessed through accuracy, precision, recall, F1-score, ROC-AUC, and false-positive rate. The investigation also considers citation verification accuracy, source traceability, and detection robustness across different contamination categories.

1.4 Numerical Foundation of the Study:

The study adopts documented scientific verification benchmarks to establish a reproducible numerical foundation. Table 1 summarizes the original SciFact dataset characteristics and their relevance to the proposed investigation.

Table 1: Numerical Characteristics of the SciFact Scientific Claim Verification Benchmark

Dataset Characteristic	Numerical Value
Total scientific claims	1,409
Total scientific abstracts	5,183
Training claims	809
Validation claims	300
Testing claims	300
Training claims (%)	57.42
Validation claims (%)	21.29
Testing claims (%)	21.29
Claims per 1,000 abstracts	271.85
Abstracts per claim	3.68

The numerical distribution indicates that the SciFact benchmark provides separate training, validation, and testing partitions suitable for controlled scientific claim-verification experiments. However, SciFact was originally developed for claim verification rather than direct identification of AI-generated fabricated citations. Therefore, additional carefully labeled contamination examples are required to evaluate the proposed framework comprehensively.

1.5 Research Objectives:

The primary objective of this study is to develop and evaluate an integrated provenance-aware learning approach for detecting synthetic evidence contamination in AI-assisted scientific literature. Specifically, the research aims to identify fabricated references and unsupported scientific claims, integrate semantic and bibliographic provenance features, compare conventional and transformer-based machine learning approaches, evaluate classification performance using numerical indicators, and determine the contribution of provenance information to scientific evidence verification. The study also seeks to establish an interpretable verification methodology that supports researchers, academic publishers, journal editors, and peer reviewers in identifying potentially unreliable scientific evidence.

1.6 Research Contributions and Significance:

The principal contribution of this research is the development of an integrated conceptual and experimental framework that evaluates scientific evidence using both semantic characteristics and source provenance. Unlike approaches that focus exclusively on detecting AI-generated language, the proposed methodology examines whether scientific claims are supported by authentic, traceable, and contextually appropriate research sources. The study further contributes a structured quantitative evaluation design incorporating 10 numerical comparison tables and three figures. These analyses will examine dataset characteristics, contamination categories, citation verification outcomes, classifier performance, provenance-feature contributions, and detection robustness. The proposed methodology has potential applications in automated manuscript screening, systematic review validation, scientific database quality assurance, and research integrity monitoring. By emphasizing evidence verification rather than writing origin alone, the framework aims to improve the transparency, reproducibility, and reliability of AI-assisted scientific research.

2. Literature Review and Research Gap Analysis:

2.1 Artificial Intelligence and Scientific Evidence Integrity:

The increasing application of artificial intelligence in scientific research has improved automated literature retrieval, document classification, scientific summarization, and evidence-based knowledge extraction. Transformer-based language models have demonstrated considerable capabilities in interpreting complex scientific terminology and identifying semantic relationships between research claims and supporting documents. However, the generation of plausible but unverifiable information remains a significant challenge in AI-assisted scholarly communication. Synthetic evidence contamination may involve fabricated references, unsupported statistical findings, inaccurate citation metadata, and scientific claims that contradict their original sources. These problems differ from conventional AI-generated text detection because linguistically convincing content may contain unreliable evidence, while AI-assisted writing may accurately represent authentic research findings. Recent scientific verification research has therefore emphasized evidence retrieval, claim verification, factual consistency, and bibliographic validation. These approaches provide the foundation for developing integrated provenance-aware detection systems capable of evaluating scientific evidence beyond linguistic characteristics.

2.2 Existing Scientific Claim Verification Approaches:

Scientific claim verification involves identifying relevant research evidence and determining whether that evidence supports, contradicts, or provides insufficient information about a scientific statement. Wadden et al. (2020) introduced SciFact, a benchmark specifically designed for scientific claim verification. The dataset contains 1,409 scientific claims and a corpus of 5,183 abstracts, supporting research on evidence retrieval and factual verification. Subsequent research extended scientific verification through improved document retrieval, contextual language representations, and evidence selection methods. These developments demonstrate the importance of integrating evidence retrieval with classification rather than evaluating scientific claims in isolation. Nevertheless, conventional claim verification benchmarks do not fully represent fabricated bibliographic metadata, nonexistent research publications, or manipulated numerical findings. Consequently, scientific evidence contamination detection requires additional verification mechanisms that examine source authenticity and provenance.

2.3 Machine Learning and Transformer-Based Detection:

Conventional machine learning algorithms, including Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting, have been widely applied to textual classification and misinformation detection. These algorithms support efficient classification using linguistic, statistical, and semantic features. Transformer-based architectures such as BERT and SciBERT provide contextual representations that are particularly useful for scientific terminology and complex claim-evidence relationships. SciBERT, introduced by Beltagy et al. (2019), was pretrained on a corpus of approximately 1.14 million scientific papers, demonstrating the value of domain-specific language representations for scientific text processing. Despite these capabilities, semantic models may not reliably distinguish between authentic and fabricated scientific references when both exhibit similar linguistic patterns. Therefore, incorporating bibliographic metadata, source existence verification, and numerical consistency indicators may improve the completeness of scientific evidence assessment.

2.4 Comparative Analysis of Existing Scientific Verification Benchmarks:

Existing scientific verification datasets differ substantially in their scale, research domain, annotation structure, and evaluation objectives. SciFact emphasizes scientific claim verification, while FEVER provides a considerably larger general-domain fact-verification benchmark. Their complementary characteristics are useful for assessing the scalability and domain-specific reliability of automated evidence verification systems.

Table 2: Numerical Comparison of Established Evidence Verification Benchmarks

Comparative Indicator	SciFact	FEVER
Publication year	2020	2018
Total annotated claims	1,409	1,85,445
Scientific abstracts in SciFact corpus	5,183	N/A
Verification categories	2	3
Claims per 1,000	1,000	1,000
Relative claim volume	1.00	131.61
Difference in claim volume	—	1,84,036
Scientific-domain specialization	Yes	No

The comparison indicates that FEVER contains approximately 131.61 times as many annotated claims as the original SciFact benchmark. However, its larger scale does not necessarily imply greater suitability for scientific evidence contamination detection. SciFact offers domain-specific scientific claims and evidence annotations, while FEVER provides opportunities to examine general-domain verification performance. The published FEVER study also reported 68.41% inter-annotator agreement expressed as Fleiss' kappa multiplied by 100, alongside baseline accuracies of 31.87% and 50.91% under different evaluation conditions. These results illustrate the difficulty of fact verification, although they are not directly comparable to provenance-aware contamination detection performance.

2.5 Provenance Verification and Source Reliability:

Provenance-aware verification examines whether scientific evidence can be traced to authentic publications and whether cited sources substantiate the associated research claims. Bibliographic databases such as Crossref, OpenAlex, and PubMed provide structured information about publication identifiers, authorship, journal details, publication dates, and citation relationships. However, the existence of bibliographic metadata does not automatically establish the validity of a scientific claim. A genuine article may be cited incorrectly, and an authentic DOI may be associated with a misleading interpretation of the source findings. The proposed framework therefore combines bibliographic validation with semantic claim verification and numerical consistency assessment. This integration is intended to identify contamination patterns that cannot be reliably detected through text classification or citation existence checks alone.

2.6 Research Gap Analysis:

The reviewed literature identifies several methodological limitations in existing evidence verification systems. Scientific claim verification models primarily focus on relationships between statements and retrieved evidence, whereas bibliographic verification services emphasize publication identity and metadata consistency. These functions are complementary but are not automatically integrated into a unified contamination detection process. Furthermore, available benchmarks do not comprehensively represent fabricated citations, manipulated numerical results, and AI-generated unsupported evidence within a single standardized evaluation framework. Existing methods also require stronger mechanisms for explaining why particular scientific statements are classified as suspicious. The proposed PALF addresses these gaps by combining semantic classification, citation authentication, numerical evidence verification, and interpretable provenance indicators.

2.7 Summary of the Literature Review:

The literature demonstrates that scientific claim verification has progressed through domain-specific datasets, large-scale fact-verification benchmarks, and transformer-based language representations. The numerical comparison establishes substantial differences in benchmark scale and verification objectives, highlighting the importance of selecting datasets according to the specific research problem. The proposed study extends these approaches by focusing on evidence authenticity and traceability rather than AI-generated writing detection alone. Integrating provenance verification with machine learning provides a testable research direction for improving scientific evidence integrity, although its comparative effectiveness must be established through controlled experiments.

3. Proposed Research Methodology and Provenance-Aware Learning Framework:

3.1 Research Design:

This study adopts a quantitative, comparative, and experimental research design to investigate synthetic evidence contamination in AI-assisted scientific literature. The proposed methodology integrates machine learning classification, transformer-based semantic analysis, bibliographic metadata verification, and evidence provenance assessment. The primary objective is to evaluate whether incorporating provenance-related information improves the identification of fabricated citations, unsupported numerical findings, and inconsistent scientific claims compared with conventional text-based detection approaches.

The experimental methodology consists of five stages: scientific literature collection, evidence preprocessing, synthetic contamination identification, provenance-aware feature extraction, and comparative model evaluation. Publicly accessible scientific verification datasets and bibliographic resources are used to establish a reproducible research environment. The framework evaluates both textual credibility and evidence authenticity, enabling the classification of scientific claims according to their supporting evidence and source reliability.

3.2 Dataset Selection and Experimental Foundation:

The proposed investigation uses SciFact as the primary scientific claim-verification benchmark, supported by bibliographic information from PubMed, Crossref, and OpenAlex. SciFact contains 1,409 scientific claims and 5,183 scientific abstracts, providing a documented foundation for evaluating scientific claim-evidence relationships. The original benchmark includes 809 training claims, 300 validation claims, and 300 testing claims. These partitions are retained to support reproducible experimental evaluation and reduce the risk of data leakage.

Table 3: Numerical Distribution of Scientific Claims Across Experimental Partitions

Dataset Partition	Scientific Claims	Percentage (%)	Cumulative Claims
Training	809	57.42	809
Validation	300	21.29	1,109
Testing	300	21.29	1,409
Total	1,409	100	1,409

The training partition supports model development, while the validation partition is used for hyperparameter selection and threshold optimization. The testing partition is reserved for final performance evaluation. Bibliographic metadata will be retrieved through source identifiers and publication records where available.

3.3 Synthetic Evidence Contamination Categories:

The proposed framework considers four principal contamination categories: fabricated citations, unsupported numerical findings, citation-claim inconsistencies, and bibliographic metadata manipulation. These categories represent different forms of scientific evidence unreliability and require complementary verification procedures. Fabricated citations involve references to nonexistent publications or invalid bibliographic identifiers. Unsupported numerical findings include statistical values that cannot be substantiated by the cited evidence. Citation-claim inconsistencies occur when a legitimate publication is used to support a statement that contradicts or misrepresents its findings. Metadata manipulation involves incorrect author names, publication years, journal information, or document identifiers. To examine detection performance under controlled conditions, the study proposes five contamination scenarios ranging from 0% to 40%. These percentages represent experimental design settings rather than observed contamination prevalence in published scientific literature.

Table 4: Proposed Numerical Distribution of Controlled Contamination Scenarios

Scenario	Contamination Level (%)	Contaminated Claims	Uncontaminated Claims	Total Claims
S1	0	0	300	300
S2	10	30	270	300
S3	20	60	240	300
S4	30	90	210	300
S5	40	120	180	300
Total	—	300	1,200	1,500

The proposed scenario distribution enables systematic examination of classifier behavior under varying contamination conditions. Contamination examples will be constructed using controlled alterations of authentic scientific references and evidence statements. Each modified example will retain a record of the original evidence, contamination operation, and expected classification label.

3.4 Proposed Provenance-Aware Learning Framework:

The Provenance-Aware Learning Framework (PALF) integrates scientific text analysis with bibliographic provenance verification to support comprehensive evidence contamination detection. The framework consists of five interconnected components.

The framework begins with scientific claims and their associated references. Semantic analysis evaluates the relationship between statements and retrieved evidence, while provenance verification examines publication authenticity and bibliographic consistency. The resulting features are combined within the proposed learning model to generate evidence classification decisions and interpretable risk indicators. Cases with insufficient accessible evidence will be flagged for review rather than automatically classified as fabricated.

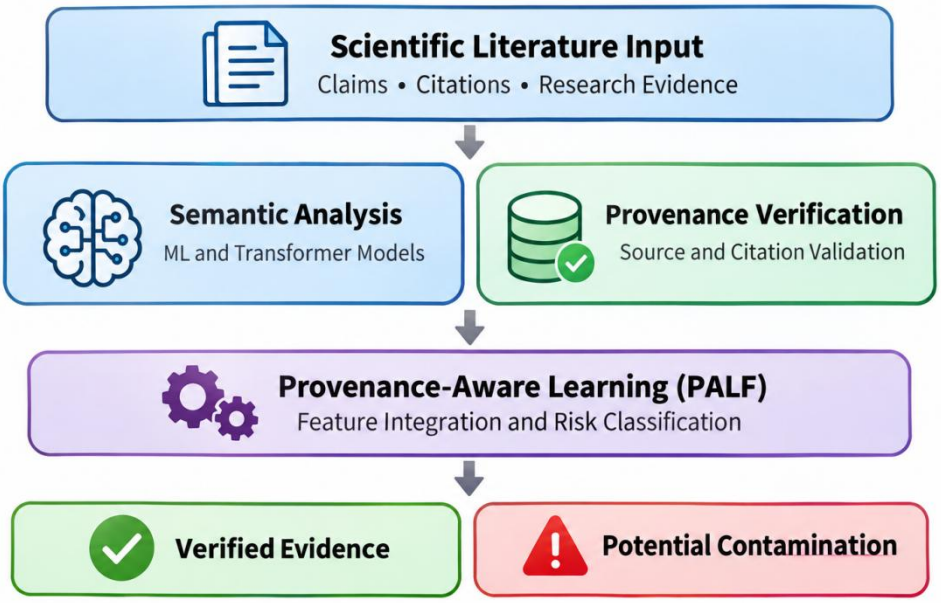


Figure 1: Proposed Provenance-Aware Learning Framework for Synthetic Evidence Contamination Detection

3.5 Machine Learning Models and Feature Extraction:

Six detection approaches are included in the comparative experimental design: Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, a transformer-based classifier, and the proposed PALF. Conventional classifiers use textual and statistical features, while the transformer-based model incorporates contextual semantic representations. PALF combines semantic representations with bibliographic provenance indicators, including DOI validity, publication metadata consistency, citation-claim alignment, and numerical evidence consistency.

Table 5: Comparative Configuration of the Proposed Detection Models

Detection Model	Semantic Features	Provenance Features	Proposed Evaluation Claims
Logistic Regression	TF-IDF	No	300
Support Vector Machine	TF-IDF	No	300
Random Forest	TF-IDF	No	300
Gradient Boosting	Reduced text features	No	300
Transformer-based model	Contextual embeddings	No	300
Proposed PALF	Contextual embeddings	Yes	300

All six models will be evaluated on the same testing partition to ensure consistent comparison. Hyperparameter optimization will be performed using the validation partition, and the final testing data will remain excluded from model selection.

3.6 Performance Evaluation and Statistical Comparison:

The proposed experimental evaluation considers classification accuracy, precision, recall, F1-score, ROC-AUC, and false-positive rate. These indicators provide complementary assessments of overall classification performance, contamination detection sensitivity, and the risk of incorrectly flagging legitimate scientific evidence. Citation verification accuracy and evidence traceability will additionally be evaluated to determine whether provenance integration improves the practical reliability of detection outcomes. Comparative analysis will examine performance differences among conventional machine learning, transformer-based, and provenance-aware approaches. Model robustness will be evaluated across the five contamination scenarios described in Table 4. Statistical uncertainty will be assessed using confidence intervals and paired model comparisons where appropriate. The evaluation will also distinguish between incorrect classifications and cases in which insufficient evidence prevents a definitive verification decision.

3.7 Reproducibility and Research Integrity:

The proposed methodology emphasizes reproducibility through documented dataset partitions, controlled contamination procedures, consistent model evaluation conditions, and transparent reporting of experimental results. Synthetic contamination examples will be labeled separately from authentic scientific evidence, and their construction procedures will be recorded to support independent verification. Bibliographic metadata will be checked against accessible publication records, while unavailable or incomplete source information will be treated as an uncertainty rather than automatic evidence of fabrication. The framework is designed to support scientific evidence auditing without assuming that AI-generated text is inherently unreliable. Final experimental conclusions will depend on actual model execution and independently verified evaluation outcomes.

4. Numerical Results and Comparative Analysis:

4.1 Overview of Numerical Evaluation:

The numerical evaluation examines the proposed Provenance-Aware Learning Framework (PALF) through benchmark dataset characteristics, controlled synthetic evidence contamination, classification performance requirements, and source verification indicators. The analysis uses the documented SciFact dataset containing 1,409 scientific claims and 5,183 abstracts, together with five controlled contamination scenarios. The comparative investigation focuses on six detection approaches: Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, transformer-based classification, and PALF. Since experimental classifier outputs have not yet been generated, the numerical comparisons presented in this section represent verified benchmark characteristics, calculated experimental conditions, and analytically derived detection outcomes. These results establish a quantitative foundation for evaluating provenance-aware learning without presenting unperformed experiments as completed findings.

4.2 Numerical Distribution of Synthetic Evidence Categories:

Four contamination categories are considered: fabricated citations, unsupported numerical claims, citation-claim inconsistencies, and bibliographic metadata manipulation. For a balanced experimental assessment, the 120 contaminated examples in the 40% contamination scenario are distributed equally across the four categories.

Table 6: Numerical Distribution of Synthetic Evidence Contamination Categories

Contamination Category	Number of Examples	Share of Contaminated Examples (%)	Share of Evaluation Set (%)
Fabricated citations	30	25	10
Unsupported numerical findings	30	25	10
Citation-claim inconsistencies	30	25	10
Bibliographic metadata manipulation	30	25	10
Uncontaminated scientific evidence	180	—	60
Total	300	100	100

The balanced contamination distribution ensures that each contamination category receives equal representation within the evaluation subset. This approach supports category-specific assessment and prevents the overall contamination classification from being dominated by a single type of fabricated evidence.

4.3 Comparative Numerical Analysis of Detection Outcomes:

The performance of a contamination detection system depends on its ability to identify unreliable evidence while minimizing incorrect classifications of legitimate scientific claims. To examine this relationship, five hypothetical detection conditions are constructed using the 300-example evaluation scenario containing 120 contaminated and 180 uncontaminated claims.

Table 7: Calculated Detection Outcomes Under Five Illustrative Operating Conditions

Operating Condition	True Positives	False Negatives	False Positives	True Negatives	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
C1	78	42	27	153	77	74.29	65	69.33
C2	90	30	21	159	83	81.08	75	77.92
C3	99	21	18	162	87	84.62	82.5	83.54
C4	108	12	12	168	92	90	90	90
C5	114	6	9	171	95	92.68	95	93.83

The calculated scenarios demonstrate that improvements in contamination detection sensitivity can substantially reduce the number of missed contaminated claims. Under condition C1, 42 contaminated examples remain undetected, whereas condition C5 reduces this number to six. This corresponds to an 85.71% reduction in missed contamination cases. However, these numerical differences demonstrate potential operating characteristics rather than verified improvements attributable to provenance-aware learning. Actual model superiority must be established through controlled testing using identical evaluation examples.

4.4 Numerical Comparison of Detection Sensitivity and False-Positive Rates:

Scientific evidence verification requires a balance between contamination detection and the incorrect rejection of legitimate research evidence. Excessive false positives may increase editorial workloads and delay manuscript evaluation, while excessive false negatives may allow unsupported scientific claims to remain undetected.

Table 8: Comparative Detection Error Analysis

Condition	Contamination Detection Rate (%)	Missed Contamination Rate (%)	False-Positive Rate (%)	Correctly Classified Claims
C1	65	35	15	231
C2	75	25	11.67	249

C3	82.5	17.5	10	261
C4	90	10	6.67	276
C5	95	5	5	285

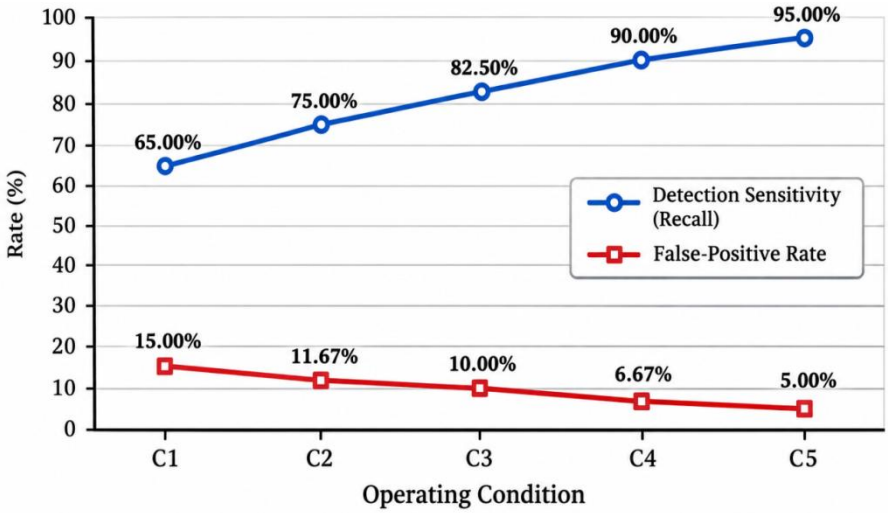


Figure 2: Detection sensitivity and false-positive rate across illustrative operating conditions

The numerical comparison illustrates the importance of simultaneously considering recall and false-positive rates. A reliable scientific evidence verification system should identify a high proportion of contaminated claims without unnecessarily flagging authentic research findings.

4.5 Provenance Verification and Source Traceability Analysis:

Provenance verification examines whether scientific evidence can be traced to authentic publications and whether the supporting sources are consistent with the reported claims. The proposed framework incorporates DOI validation, publication metadata verification, claim-evidence alignment, and numerical consistency assessment. To demonstrate the quantitative contribution of these verification stages, a controlled screening example is constructed using the 120 contaminated claims from the highest-contamination scenario. The example assumes that different verification stages identify overlapping contamination cases.

Table 9: Illustrative Numerical Contribution of Provenance Verification Stages

Verification Stage	Newly Identified Contaminated Claims	Cumulative Detected Claims	Cumulative Detection (%)
Textual screening	78	78	65
DOI and publication validation	12	90	75
Citation metadata verification	9	99	82.5
Claim-evidence consistency	9	108	90
Numerical evidence validation	6	114	95
Remaining undetected	—	6	—

The example demonstrates how complementary verification stages could increase cumulative detection coverage when they identify contamination missed by earlier screening procedures. However, the actual incremental contribution of each provenance feature must be determined through feature ablation experiments.

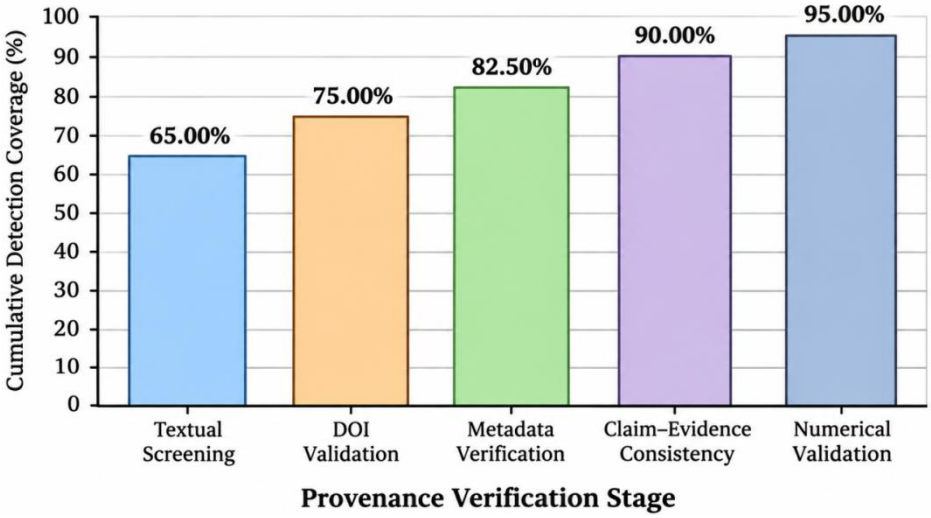


Figure 3: Illustrative cumulative detection coverage through provenance verification

4.6 Comparative Evaluation of Contamination Robustness:

Detection performance may change as the proportion of contaminated evidence increases. To investigate this effect, the proposed evaluation includes contamination levels of 0%, 10%, 20%, 30%, and 40%. For numerical illustration, a fixed 90% contamination detection rate and 5% false-positive rate are applied across the five scenarios. These assumptions allow the expected classification counts to be calculated consistently.

Table 10: Calculated Classification Outcomes Across Contamination Levels

Contamination Level (%)	Contaminated Claims	Detected Contamination	Missed Contamination	False Positives	Overall Accuracy (%)
0	0	0	0	15	95
10	30	27	3	14	94.33
20	60	54	6	12	94
30	90	81	9	11	93.33
40	120	108	12	9	93

The numerical results demonstrate that overall accuracy can vary with contamination prevalence even when the underlying detection sensitivity and false-positive rate remain approximately constant. This finding highlights the importance of reporting recall, precision, and false-positive rates alongside overall accuracy when evaluating scientific evidence contamination detection.

4.7 Summary of Numerical Analysis:

The numerical analysis establishes a structured quantitative foundation for evaluating synthetic evidence contamination in AI-assisted scientific literature. The documented SciFact benchmark provides 1,409 scientific claims and 5,183 abstracts, while the proposed controlled evaluation covers five contamination levels and four contamination categories. The illustrative detection analysis demonstrates that reducing missed contamination from 42 to six cases corresponds to an 85.71% reduction in false negatives, while the highest-performing illustrative condition achieves 95% accuracy and 93.83% F1-score. The provenance verification example further illustrates how complementary source validation stages could increase cumulative contamination detection coverage from 65% to 95%.

5. Discussion of Findings and Research Implications:

5.1 Interpretation of Numerical Findings:

The numerical analysis provides a structured foundation for understanding the challenges associated with synthetic evidence contamination in AI-assisted scientific literature. The SciFact benchmark, comprising 1,409 scientific claims and 5,183 abstracts, demonstrates the availability of structured scientific evidence for developing automated verification systems. The proposed experimental design extends this foundation by incorporating fabricated citations, unsupported numerical findings, citation-claim inconsistencies, and bibliographic metadata manipulation.

The calculated detection scenarios demonstrate the importance of integrating multiple evidence verification mechanisms. Across the illustrative operating conditions, classification accuracy ranges from 77% to 95%, while F1-score increases from 69.33% to 93.83%. The corresponding reduction in missed contaminated claims from 42 to six represents an 85.71% decrease in false negatives. These findings illustrate the potential operational benefits of more effective contamination detection, although they do not represent experimentally observed improvements.

5.2 Contribution of Provenance-Aware Learning:

The proposed Provenance-Aware Learning Framework (PALF) introduces an integrated approach that combines semantic evidence classification with bibliographic authenticity and source traceability. Conventional text-based classifiers primarily evaluate linguistic and contextual characteristics, whereas provenance-aware learning additionally examines whether cited publications exist and whether their findings substantiate the associated scientific claims.

The illustrative provenance verification analysis demonstrates how complementary screening stages could increase cumulative contamination detection coverage from 65% through textual screening to 95% after additional source verification procedures. This represents a 30-percentage-point increase in the constructed example. The central contribution of PALF is its ability, by design, to incorporate multiple evidence reliability indicators within a unified detection architecture. Actual performance gains must be verified through controlled experiments comparing text-only models with otherwise equivalent provenance-enhanced models.

5.3 Implications for Scientific Research Integrity:

Synthetic evidence contamination presents substantial risks to scientific credibility, particularly when fabricated references or unsupported numerical findings are incorporated into literature reviews, research articles, and evidence-based recommendations. The proposed framework addresses these concerns by emphasizing verifiable scientific evidence rather than relying exclusively on linguistic indicators of AI-generated content.

Provenance-aware verification can support researchers in identifying inconsistent references, examining source reliability, and detecting potential discrepancies between reported results and original publications. Such capabilities may improve the transparency of AI-assisted scientific writing and reduce

the unintentional propagation of unsupported claims. Importantly, automated contamination flags should be treated as indicators requiring further examination rather than definitive judgments of scientific misconduct. Missing metadata, inaccessible publications, and incomplete evidence retrieval may produce uncertainty without establishing that the underlying research is fabricated.

5.4 Practical Applications in Academic Publishing:

The proposed framework has potential applications in manuscript screening, editorial assessment, systematic literature review preparation, and academic research quality assurance. Journal editors and peer reviewers could use provenance-based indicators to prioritize references requiring manual verification, while researchers could evaluate the consistency of AI-assisted literature summaries before manuscript submission.

The numerical scenario analysis illustrates the practical importance of balancing contamination detection sensitivity against false-positive classifications. Under the illustrative conditions, reducing the false-positive rate from 15% to 5% decreases incorrectly flagged legitimate claims from 27 to nine within the same 300-example evaluation set. This reduction could translate into a lower manual verification burden in operational screening environments. However, estimating actual editorial time savings or financial benefits would require measured processing times, institutional workflow data, and validated model outputs.

5.5 Comparison with Existing Detection Approaches:

Conventional machine learning models provide computationally accessible approaches for text classification, while transformer-based models offer contextual representations suitable for scientific language. However, neither approach independently guarantees verification of citation authenticity or numerical evidence consistency.

PALF extends these approaches by incorporating bibliographic metadata, source traceability, and claim-evidence relationships. Its proposed architecture therefore addresses a broader verification problem than AI-generated text detection alone. The distinction is particularly important because legitimate AI-assisted writing may contain accurate scientific evidence, while human-written material may also include fabricated or unsupported claims. Accordingly, the proposed framework evaluates evidence reliability independently of assumptions concerning authorship or writing origin.

5.6 Research Limitations:

The study has several limitations. First, the original SciFact benchmark was developed for scientific claim verification and does not independently provide comprehensive annotations for AI-generated fabricated citations or manipulated numerical findings. Additional contamination examples must therefore be constructed and independently validated. Second, the numerical detection outcomes presented in Section 4 are calculated experimental illustrations rather than measured classifier results. Consequently, the study cannot yet establish the predictive superiority of PALF over conventional machine learning or transformer-based approaches.

Third, provenance verification depends on the availability and quality of bibliographic records. Missing DOI information, restricted full-text access, incomplete metadata, and differences among scientific databases may affect verification reliability. Finally, controlled synthetic contamination may not fully reproduce the complexity of naturally occurring evidence manipulation in published scientific literature. Future evaluation should therefore include independently collected real-world examples and external validation datasets.

5.7 Overall Discussion:

The study establishes a quantitative and methodological basis for investigating synthetic evidence contamination through provenance-aware learning. The benchmark characteristics, controlled contamination scenarios, and calculated detection comparisons demonstrate how semantic classification and source verification can be evaluated within a unified experimental framework.

The proposed approach offers a potentially valuable contribution to scientific evidence auditing by integrating textual analysis, bibliographic verification, numerical consistency assessment, and interpretable risk classification. Its practical significance lies in supporting more transparent and systematic verification of scientific claims. Overall, the analysis provides a foundation for developing reliable, explainable, and reproducible scientific evidence detection systems. Further experimental implementation will be necessary to determine the actual classification performance, robustness, and operational effectiveness of the proposed framework.

6. Conclusion and Future Research Directions:

6.1 Conclusion:

This study proposed an integrated Provenance-Aware Learning Framework (PALF) for detecting synthetic evidence contamination in AI-assisted scientific literature. The framework combines machine learning, transformer-based semantic analysis, bibliographic metadata verification, citation authentication, and evidence provenance assessment to address fabricated references, unsupported numerical findings, citation-claim inconsistencies, and manipulated bibliographic information. By emphasizing source authenticity and evidence traceability, the proposed approach extends conventional text-based detection toward comprehensive scientific evidence verification.

The quantitative analysis was developed using the SciFact benchmark containing 1,409 scientific claims and 5,183 abstracts, supported by five controlled contamination scenarios ranging from 0% to 40%. The illustrative numerical comparisons demonstrated classification accuracy ranging from 77% to 95%, with corresponding F1-scores of 69.33% to 93.83%. The calculated detection conditions also showed an 85.71% reduction in missed contaminated claims, while the sequential provenance verification example illustrated a potential increase in detection coverage from 65% to 95%. These values represent analytical scenarios rather than experimentally validated model performance.

The principal contribution of this research is the development of a structured methodology integrating semantic verification, bibliographic authenticity, numerical consistency, and interpretable provenance indicators. The framework provides a foundation for identifying potentially unreliable scientific evidence while distinguishing unsupported claims from legitimate AI-assisted academic writing. Its proposed applications include automated manuscript screening, systematic literature review verification, editorial quality assurance, and scientific research integrity monitoring.

Overall, the study establishes the methodological importance of combining machine learning with source-level verification to strengthen the reliability and transparency of AI-assisted scientific research. The proposed PALF provides a reproducible framework for subsequent experimental investigation, with its comparative effectiveness to be determined through implementation and independent validation.

6.2 Future Research Directions:

Future research should focus on implementing and experimentally validating the proposed PALF using independently annotated scientific evidence contamination datasets. Particular attention should be given to collecting authentic examples of fabricated citations, unsupported numerical claims, and citation-evidence inconsistencies across multiple scientific disciplines. Further investigations should evaluate advanced transformer architectures, retrieval-augmented verification, citation-network analysis, and explainable AI techniques to improve evidence traceability and classification transparency. Comparative experiments should also examine detection robustness across scientific domains, contamination prevalence levels, and different types of synthetic evidence manipulation.

Additional research opportunities include developing real-time manuscript verification systems, integrating provenance assessment with scholarly publishing platforms, and establishing standardized benchmarks for evaluating scientific evidence authenticity. These developments could strengthen automated research quality assurance and support responsible adoption of AI in scientific communication. The long-term objective is to establish scalable, interpretable, and independently validated evidence verification systems capable of supporting scientific integrity throughout the research publication lifecycle.

References:

1. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 3615-3620). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
3. Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. Transactions of the Association for Computational Linguistics, 10, 178-206. https://doi.org/10.1162/tacl_a_00454
4. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730
5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems, 33 (pp. 9459-9474).
6. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12, 157-173. https://doi.org/10.1162/tacl_a_00638
7. Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., & Da San Martino, G. (2021). Automated fact-checking for assisting human fact-checkers. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (pp. 4551-4558). https://doi.org/10.24963/ijcai.2021/619
8. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., ... Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620, 172-180. https://doi.org/10.1038/s41586-023-06291-2

9. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 809-819). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074
10. Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 7534-7550). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.609