Innovative Journal of Computing, Security and Digital Technologies
Volume 1, Issue 3, Sep – Dec 2026
Page no : 09-24

Seeing Where the Answer Comes From: Evidence Provenance, Calibrated Reliance, and Decision Quality in AI-Assisted Construction Project Risk Assessment

Bilal Khan
Department of Management Sciences, Abasyn University, Peshawar, Pakistan.

Abstract

Generative artificial intelligence is increasingly being explored for construction project risk assessment, including the preparation of risk registers, contract summaries, progress reports, and project documentation. However, retrieval-augmented generation does not guarantee that generated claims are fully supported by retrieved evidence. Unsupported claims can consequently complicate professional judgment, particularly when project managers must evaluate multiple sources of information under time pressure. Although technical approaches to factuality improvement have received growing attention, the influence of evidence provenance visibility on human reliance and decision quality remains an important area for empirical investigation. This study proposes a between-subjects scenario experiment examining whether provenance labels improve calibrated reliance and decision quality in AI-assisted construction project risk assessment. Drawing on trust-in-automation theory, research on retrieval-augmented generation, and emerging approaches to evidence-bound factual repair, the study investigates the effects of provenance visibility on unsupported-claim detection, calibrated reliance, and risk decision quality. It also examines the potential moderating roles of AI literacy and project managers' emotional intelligence. Construction and infrastructure professionals in Khyber Pakhtunkhwa, Pakistan, will evaluate AI-generated risk assessments under conditions with or without provenance labels. The planned analysis uses between-group comparisons and claim-level mixed-effects models, with mediation and moderation analyses treated as secondary and conditional on adequate statistical power. No empirical findings are claimed because data have not yet been collected. The study provides a framework for investigating whether making the evidential status of AI-generated claims visible can support more appropriate reliance on generative AI in construction project management.

Keywords: Evidence provenance, calibrated reliance, generative artificial intelligence, retrieval-augmented generation, construction project management, risk assessment, decision quality, AI literacy, emotional intelligence.

Publication Date : 21.09.2026

How to Cite

Seeing Where the Answer Comes From: Evidence Provenance, Calibrated Reliance, and Decision Quality in AI-Assisted Construction Project Risk Assessment. (2026). *Innovative Journal of Computing, Security and Digital Technologies*, *1*(3), 09-24. https://ijcdst.in/index.php/about/article/view/12

INTRODUCTION

Construction projects operate within complex environments characterized by technical uncertainty, interdependent activities, contractual obligations, resource constraints, and changing external conditions. Project managers must interpret information from contracts, schedules, inspection reports, progress records, engineering documents, and stakeholder communications to identify risks and determine appropriate responses. The quality of these decisions depends not only on the availability of information but also on the ability to distinguish reliable evidence from incomplete, inconsistent, or unsupported claims.

Generative artificial intelligence offers opportunities to assist these activities by summarizing project documents, organizing risk information, identifying potential conflicts, and preparing preliminary risk assessments. Retrieval-augmented generation (RAG) is particularly relevant because it combines information retrieval with language generation, allowing a system to produce responses using external documents rather than relying exclusively on information encoded during model training (Lewis et al., 2020). In construction project management, such an approach could help professionals locate relevant contractual provisions, summarize reported delays, and connect identified risks with documentary evidence.

Nevertheless, access to retrieved documents does not ensure that every generated statement accurately represents those documents. A language model may produce a plausible statement that is only partially supported by the retrieved material or that introduces details absent from the evidence. The resulting problem is not simply whether an AI system can generate a convincing answer. It is whether the professional receiving that answer can determine which claims deserve reliance and which require further verification.

Research on factual evaluation demonstrates the importance of assessing generated text at a granular level. Min et al. (2023), through FActScore, emphasize the evaluation of factual precision using atomic claims. Niu et al. (2024) introduce RAGTruth as a corpus for investigating hallucinations in retrieval-augmented language models. These contributions provide foundations for examining factual support at the claim level, although they do not by themselves establish how construction professionals respond to visible evidence labels.

Rajendran and Singaravelu (2026) propose an evidence-bound factual repair approach that distinguishes semantic localization from deterministic realization and describes provenance-related signals for generated claims. This approach offers a useful conceptual starting point for investigating whether users benefit when the evidential status of AI-generated statements is made visible. However, the effectiveness of such signals in professional project-risk decisions requires separate empirical evaluation.

The human dimension is especially important because AI-assisted decisions involve more than technical accuracy. Professionals must decide whether to accept, reject, verify, or disregard AI-generated recommendations. Lee and See (2004) argue that trust in automation should support appropriate reliance rather than indiscriminate acceptance or unnecessary rejection. Parasuraman and Riley (1997) similarly distinguish among

different patterns of automation use, including misuse and disuse. These perspectives suggest that the value of evidence provenance should be assessed through its relationship with actual reliance behavior and decision performance.

The problem becomes more complex when professionals differ in their understanding of AI systems and their capacity to interpret uncertain information. AI literacy may influence how effectively users recognize system limitations and evaluate generated content. Emotional intelligence may also be relevant to project management because interpersonal awareness, emotional regulation, and responses to uncertainty can influence professional judgment and coordination. Zhang et al. (2022) provide a meta-analytic examination of the relationship between project managers' emotional intelligence and project performance, offering a foundation for considering this characteristic in project-related decision research.

BACKGROUND OF THE STUDY

2.1 Generative AI and construction project risk assessment

Construction risk management involves identifying uncertain events, evaluating their likelihood and consequences, and selecting appropriate responses. These activities depend on documentary evidence and professional interpretation. An overlooked contractual condition, an incorrectly summarized delay, or an unsupported assumption about material availability can affect the assessment of project exposure.

Generative AI may assist project professionals by converting unstructured information into organized summaries and preliminary recommendations. RAG provides a relevant architecture because the system retrieves external information and conditions generation on the retrieved material (Lewis et al., 2020). This arrangement can make generated responses more contextually relevant and provide a potential basis for checking their factual support.

However, retrieval and generation are distinct operations. A document may be retrieved correctly while the generated statement misrepresents it. A response may also combine supported details with an unsupported conclusion, making the entire statement difficult to evaluate without examining its individual claims. The availability of source documents therefore does not eliminate the need for human verification.

Asai et al. (2024) investigate Self-RAG, an approach that incorporates retrieval, generation, and self-reflection. Their work illustrates efforts to improve the control and evaluation of retrieval-augmented responses. Zhang and Zhang (2025) review hallucination mitigation approaches for RAG systems, while Lv et al. (2024) investigate factuality improvement through contrastive decoding involving factual and hallucination prompts. These studies represent technical approaches to improving generated content. The present study addresses a related but different question: whether making the evidential status of claims visible changes the behavior of professionals using generated information.

This distinction matters in construction because an AI-generated risk report may influence decisions concerning project schedules, contractual responsibilities, resource allocation, and mitigation priorities. Even when a system produces a generally useful

summary, an unsupported claim may affect the interpretation of an individual risk. Evaluating the value of provenance visibility therefore requires attention to both the information displayed and the decisions made in response.

2.2 Evidence provenance and factual support

Evidence provenance concerns the relationship between a statement and the evidence from which it is derived. In an AI-generated report, provenance information can help users identify whether a claim is supported by the supplied documents, contradicted by them, or not verified from them.

Rajendran and Singaravelu (2026) describe an evidence-bound factual repair approach that separates semantic localization from deterministic realization. Their proposed mechanism provides a relevant conceptual basis for representing the evidential status of generated text. For the purposes of this study, provenance visibility refers to the presentation of labels indicating the support status of individual claims in an AI-generated risk assessment.

The study distinguishes provenance visibility from factual accuracy. Factual accuracy concerns whether a statement is correct, while provenance visibility concerns whether information about its evidential support is available to the user. A provenance label does not automatically make an unsupported claim accurate. Its potential value lies in helping a professional identify where verification is necessary.

This distinction also separates the proposed intervention from technical hallucination mitigation. A generation system may attempt to reduce unsupported content through retrieval strategies, self-reflection, or decoding methods. Provenance labels instead provide information that may help users evaluate the output that remains. These approaches may be complementary, but the effectiveness of the labels cannot be inferred solely from improvements reported for technical methods.

The study will operationalize provenance labels as predefined categories that indicate whether a claim is supported by the supplied project evidence, contradicted by the supplied project evidence, or not verified by the supplied project evidence. The exact labels and their definitions must be consistent across experimental materials. Their interpretation should also be tested during pilot work to ensure that participants understand the distinction among categories.

2.3 Trust in automation and calibrated reliance

Trust influences whether individuals rely on automated systems, seek additional information, or reject automated recommendations. Lee and See (2004) conceptualize trust in automation around the goal of appropriate reliance. From this perspective, the objective is not to maximize trust without qualification but to support reliance that reflects the capabilities and limitations of the system.

Parasuraman and Riley (1997) describe misuse, disuse, and abuse as important concerns in human interaction with automation. Misuse can occur when users rely on automation inappropriately, whereas disuse occurs when useful automation is rejected or underused. Both patterns are relevant to AI-assisted construction decisions. A professional who accepts an unsupported contractual claim may make an avoidable error, while a

professional who rejects every AI-generated statement may lose the benefits of useful assistance.

Hoff and Bashir (2015) integrate empirical evidence on factors that influence trust in automation, emphasizing the importance of contextual and individual influences. Glikson and Woolley (2020) review empirical research on human trust in artificial intelligence and identify trust as a multidimensional issue influenced by characteristics of the technology and the interaction context. Together, these studies suggest that trust should be examined in relation to the conditions under which AI is used and the behavior that follows from it.

In the present study, calibrated reliance is defined operationally as the degree to which acceptance and reliance correspond to the actual evidential status of individual AI-generated claims. A participant demonstrates better calibration when supported claims are appropriately accepted and unsupported claims are appropriately rejected or subjected to verification. General confidence in AI, by contrast, may remain high even when a participant accepts an incorrect claim.

This operational distinction is central to the proposed experiment. Provenance labels may encourage appropriate reliance on supported claims while reducing acceptance of unsupported claims. Alternatively, users may misunderstand the labels or become overly cautious. The direction and magnitude of these effects therefore require empirical testing rather than assumption.

2.4 Decision quality in project risk assessment

Project risk decision quality concerns the extent to which a professional's judgments and responses are consistent with the available evidence and the requirements of the task. In a controlled experiment, decision quality can be assessed using a predefined rubric that considers claim verification, risk prioritization, and the appropriateness of proposed mitigation actions.

A decision should not be classified as high quality simply because it agrees with an AI-generated recommendation. The appropriate benchmark is the scenario evidence and the decision criteria established before data collection. This approach avoids treating the AI output as its own measure of correctness.

Evidence provenance may improve decision quality by directing attention toward claims that require additional checking. If participants use labels effectively, they may be less likely to accept unsupported information and more likely to prioritize verification where it is needed. However, labels may also create distraction or encourage overreliance on the labeling system itself. The relationship between provenance visibility and decision quality must therefore be examined directly.

Endsley (2017) discusses lessons from human–automation research in the development of autonomous systems. This work reinforces the importance of understanding human capabilities, limitations, and situation awareness when designing automated decision support. In construction risk assessment, visible evidence information may support situation assessment by helping professionals distinguish documented facts from statements requiring further scrutiny.

2.5 AI literacy and emotional intelligence

AI literacy is relevant because professionals may differ in their understanding of how generative systems operate, how retrieval influences responses, and why apparently coherent outputs may still be unsupported. Lintner (2024) systematically reviewed AI literacy scales and found variation in their measurement quality and suitability across target populations. The review can guide instrument selection, but it does not establish that any particular scale is valid for construction professionals in Pakistan. The selected instrument should therefore undergo expert review and, if translated, appropriate translation and cultural adaptation.

Professionals with greater AI literacy may be more likely to understand the limitations of generated content and interpret provenance labels as signals for further evaluation. However, higher literacy does not necessarily guarantee better decisions. Experienced users may still rely on automation under time pressure, while less experienced users may benefit from clear labels. The moderating role of AI literacy therefore remains an empirical question.

Emotional intelligence represents a different individual characteristic. Zhang et al. (2022) examine its relationship with project performance through meta-analysis. In project environments, emotional intelligence may be relevant to how professionals manage uncertainty, respond to conflicting information, and communicate concerns about risk. Nevertheless, its influence on the interpretation of provenance labels has not been established by the cited meta-analysis. The present study therefore treats emotional intelligence as a proposed moderator rather than an established determinant of calibrated reliance.

LITERATURE REVIEW

3.1 Technical approaches to factuality and provenance

Research on RAG has established an important foundation for connecting language generation with external evidence. Lewis et al. (2020) introduce retrieval-augmented generation for knowledge-intensive natural language processing tasks, demonstrating the relevance of combining parametric language models with retrieved information. This architecture provides a useful starting point for applications in which responses must be grounded in external documents.

Subsequent research has explored ways to improve the reliability of retrieval-augmented responses. Asai et al. (2024) present Self-RAG, which incorporates retrieval, generation, and self-reflection. Niu et al. (2024) introduce RAGTruth to support the development and evaluation of trustworthy retrieval-augmented models. Min et al. (2023) develop FActScore to evaluate factual precision through fine-grained atomic claims. These contributions indicate that factual reliability can be examined at a more detailed level than the overall fluency or apparent coherence of a generated response.

Zhang and Zhang (2025) review hallucination mitigation techniques for retrieval-augmented language models, while Lv et al. (2024) investigate contrastive decoding using factual and hallucination prompts. Rajendran and Singaravelu (2026) propose evidence-bound factual repair that separates semantic localization from deterministic

realization. Collectively, these works provide a technical basis for examining the support status of generated claims.

However, the existence of factuality evaluation and mitigation methods does not establish whether human users interpret or act upon evidence signals effectively. Technical accuracy and human decision quality are related but distinct outcomes. A system may improve the presentation of evidence without necessarily improving user decisions, particularly if labels are misunderstood, ignored, or treated as guarantees.

The present study therefore focuses on the user-facing implications of provenance visibility. It does not attempt to establish that provenance labels eliminate hallucinations or outperform all technical mitigation strategies. Instead, it investigates whether the visibility of evidential status influences claim verification, calibrated reliance, and risk decisions.

3.2 Trust and reliance in AI-assisted decisions

Trust in automation research emphasizes the importance of aligning reliance with system capabilities. Lee and See (2004) argue that automation should support appropriate reliance, while Parasuraman and Riley (1997) identify the risks of both misuse and disuse. Hoff and Bashir (2015) further demonstrate that trust is influenced by multiple factors rather than a single, universal characteristic of an automated system.

Glikson and Woolley (2020) review research on human trust in AI and show the importance of examining trust in relation to system properties and interaction contexts. This perspective is particularly relevant to generative AI because language outputs may be persuasive even when their evidential support is incomplete. The fluency of an answer may therefore create a mismatch between perceived reliability and actual correctness.

Provenance labels may help reduce this mismatch by providing users with information that is not immediately apparent from the generated prose. A label indicating that a claim is unsupported could prompt verification, whereas a label indicating direct evidential support could help users identify information that is more suitable for reliance. However, these benefits depend on the accuracy and clarity of the labels and on users' ability to interpret them.

The study consequently proposes that provenance visibility will improve calibrated reliance, but it does not assume that greater trust in AI is inherently desirable. The relevant outcome is whether participants distinguish supported from unsupported claims more effectively and make decisions that correspond to the available evidence.

3.3 Individual differences in professional judgment

AI-assisted decision-making occurs in organizational environments where users possess different levels of knowledge and experience. AI literacy may affect the ability to understand the limitations of generative systems, evaluate evidence, and recognize uncertainty. Lintner's (2024) systematic review provides a basis for selecting an AI literacy instrument, although the selected scale must be appropriate for construction professionals and the study context.

Emotional intelligence may influence how professionals respond to uncertain situations and manage interpersonal aspects of project work. Zhang et al. (2022) examine the relationship between project managers' emotional intelligence and project performance, suggesting that the construct warrants consideration in project management research. However, project performance and calibrated reliance are not equivalent outcomes. Any proposed connection between emotional intelligence and provenance interpretation must be tested independently.

These individual differences may influence the effectiveness of provenance labels. Users with greater AI literacy might be better equipped to interpret evidence categories, while emotional intelligence could be associated with how participants manage uncertainty or competing demands. Alternatively, provenance labels may provide similar benefits across different levels of these characteristics. The proposed moderation tests are designed to investigate these possibilities.

3.4 Research gap and conceptual framework

The reviewed literature provides three relevant foundations. RAG and factuality research explains how generated claims can be grounded in external evidence and evaluated for support. Human–automation research explains why trust should be aligned with actual system reliability. Project management research establishes the relevance of individual characteristics such as emotional intelligence to professional outcomes.

What remains uncertain is whether exposing claim-level evidence provenance changes how construction professionals evaluate AI-generated risk assessments. The proposed study addresses this question by examining the relationship between provenance visibility, calibrated reliance, and decision quality.

The conceptual framework proposes that provenance visibility improves discrimination between supported and unsupported claims and thereby improves calibrated reliance. Better calibrated reliance is expected to be associated with higher decision quality. AI literacy and emotional intelligence are considered potential moderators of the effect of provenance visibility on calibrated reliance. These propositions are hypotheses, not established findings. The model distinguishes the experimental treatment (provenance visibility), behavioral outcomes (claim-level acceptance and rejection), the calibration construct (alignment between reliance and evidence status), and the decision outcome (risk-assessment quality).

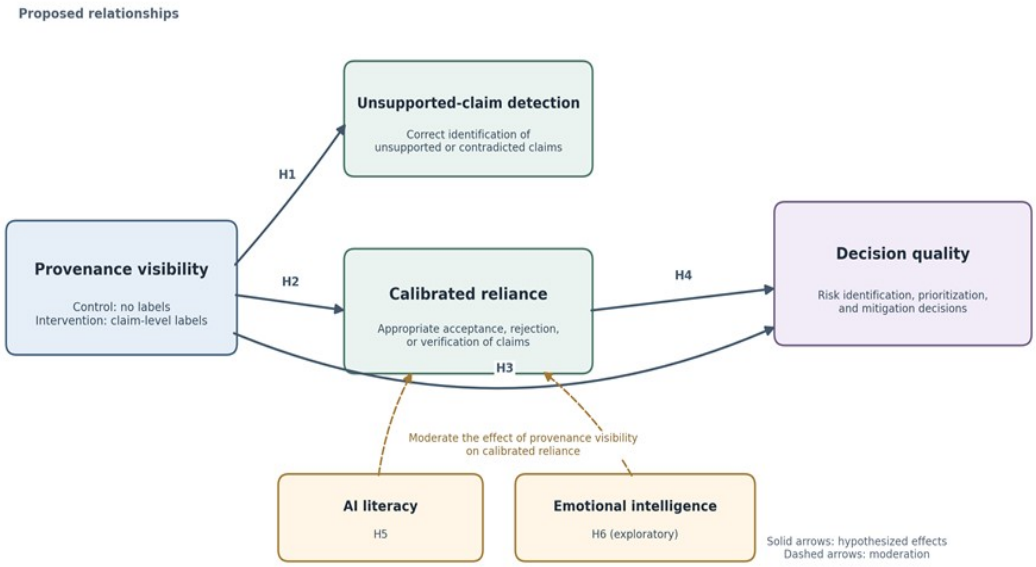


Figure 1. Conceptual framework linking provenance visibility, calibrated reliance, and decision quality.

The following hypotheses will guide the confirmatory analyses. H1 proposes that participants who receive claim-level provenance labels will demonstrate higher accuracy in identifying unsupported or contradicted claims than participants who receive no labels. H2 proposes that provenance labels will improve calibrated reliance by increasing appropriate acceptance of supported claims and appropriate rejection or verification of unsupported claims. H3 proposes that participants in the provenance-label condition will achieve higher decision-quality scores than participants in the control condition. H4 proposes that calibrated reliance will be positively associated with decision quality. H5 proposes that AI literacy will moderate the effect of provenance visibility on calibrated reliance. H6 proposes that emotional intelligence will moderate this relationship; this final hypothesis will be treated as exploratory because the cited project-performance literature does not directly establish a relationship with calibrated reliance.

METHODOLOGY

4.1 Research design

The study will employ a randomized, between-subjects scenario experiment. Participants will be assigned to either a control condition, in which AI-generated construction risk assessments are presented without claim-level provenance labels, or an intervention condition, in which the same assessments include provenance labels. Random assignment will be computer-generated and concealed until the participant begins the task. If recruitment or delivery constraints prevent individual randomization, the study must report the alternative allocation procedure and address the resulting risk of selection bias.

The experimental design is intended to isolate the effect of provenance visibility. The content, length, factual claims, formatting, and task instructions should be held as consistent as possible across conditions. Any systematic differences between the two versions beyond the presence of labels could otherwise confound the results.

The proposed setting is construction and infrastructure project management in Khyber Pakhtunkhwa, Pakistan. This setting allows the scenarios to reflect recognizable project documentation and risk-management tasks. Recruitment should cover more than one organization and professional role where feasible. Because the study uses a regional sample, findings should be generalized only to populations and settings supported by the sampling frame.

4.2 Population and sample

The target population will consist of construction and infrastructure professionals involved in project planning, implementation, monitoring, or risk management. Eligible participants may include project managers, construction engineers, risk managers, contract professionals, and other practitioners who regularly interpret project information.

The sample size will be determined before recruitment through an a priori power analysis for the primary between-group comparison. The calculation will specify the smallest effect considered practically meaningful, the selected significance level, desired statistical power, allocation ratio, and allowance for incomplete responses. The proposed range of 150–250 participants may be used only as a provisional recruitment target; it is not a substitute for the power calculation. Moderation analyses require additional power and should be treated as secondary unless the final power analysis supports them.

Recruitment procedures should specify eligibility criteria, professional experience requirements, and the approach used to invite participants. Participants should be informed about the study's purpose, voluntary nature, confidentiality arrangements, and the handling of their responses.

4.3 Experimental materials and procedure

The experimental materials will consist of realistic construction project risk scenarios prepared from a controlled set of project documents. These materials may include contractual extracts, progress reports, schedule updates, and records describing potential project risks. AI-generated assessments will be constructed so that some claims are supported by the supplied documents and others contain identifiable unsupported statements.

The support status of every claim must be established before the experiment using a claim-evidence matrix. Each claim will be classified as (a) supported by an identifiable passage in the supplied documents, (b) contradicted by the supplied documents, or (c) not verifiable from the supplied documents. These categories must not be conflated: absence of evidence is not automatically evidence of contradiction. At least two independent reviewers with construction or project-management expertise should classify the claims, with disagreements resolved through documented adjudication. Inter-rater agreement should be reported. The final evidence map and scoring key must be frozen before data collection.

Participants in the control condition will receive the risk assessment without provenance labels. Participants in the intervention condition will receive the same text with labels attached to individual claims. The labels will be defined in the instructions as

"supported" (an identifiable source passage substantiates the claim), "contradicted" (the supplied documents contain evidence inconsistent with the claim), and "not verified" (the supplied documents do not establish whether the claim is true). The interface, wording, source documents, task order, and time limits will otherwise be identical. Labels must be generated from the prevalidated evidence map rather than assigned by the language model during the experiment.

Participants will review the materials and indicate for each claim whether they would rely on it, reject it, or seek further verification. They will then identify and prioritize the project risks and select or propose mitigation actions. The study will record claim-level decisions, response times where feasible, confidence ratings, and the final risk assessment. The primary outcome and scoring rules will be specified before examining the data.

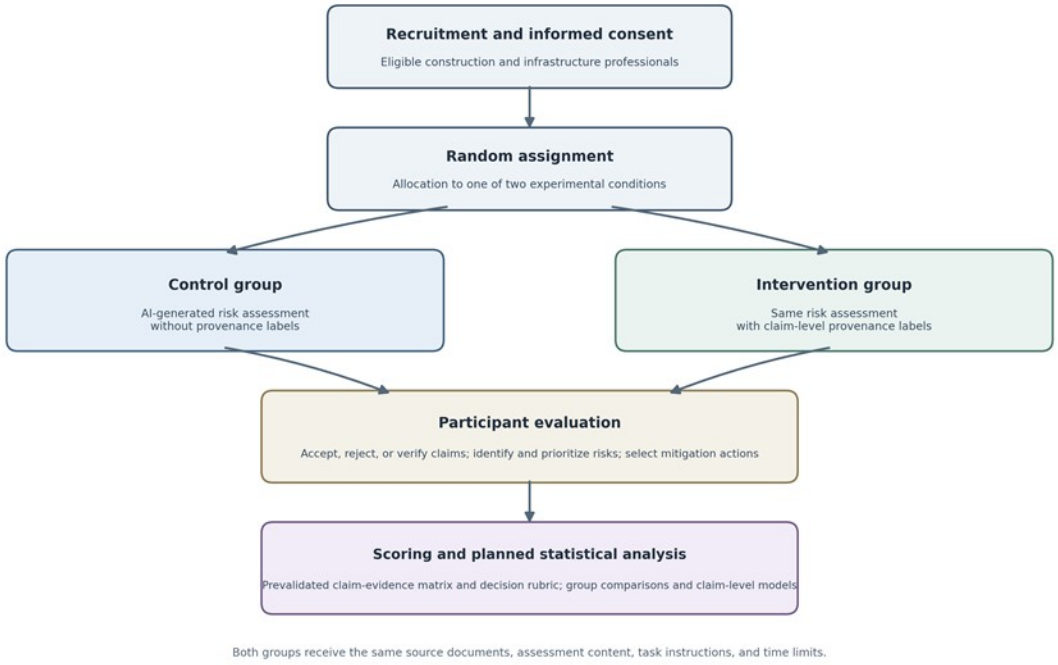


Figure 2. Experimental procedure for evaluating provenance visibility in construction project risk assessment.

A pilot study should be conducted before the main data collection to assess the clarity of instructions, realism of scenarios, comprehensibility of labels, and approximate completion time. The pilot should also identify claims that are ambiguous or could reasonably be interpreted in multiple ways.

4.4 Measurement of variables

The independent variable will be provenance visibility, operationalized as the presence or absence of claim-level provenance labels.

Unsupported-claim detection will be measured by comparing participants' responses with the predetermined support status of each claim. The measure should distinguish correct rejection of unsupported claims from incorrect rejection of supported claims. This

is important because a participant who rejects every claim may achieve high unsupported-claim rejection while performing poorly overall.

Calibrated reliance will be operationalized as the extent to which participants' reliance decisions correspond to the independently established evidential status of claims. The scoring must penalize both false acceptance of unsupported or contradicted claims and false rejection of supported claims. A calibration score should be calculated from prespecified claim-level decision rules, while self-reported trust in AI will be treated as a separate supplementary construct. The manuscript should avoid using the terms trust and reliance interchangeably.

Decision quality will be assessed independently of the claim-detection score using a prespecified rubric. The rubric will assess whether participants identify the material risks supported by the scenario, prioritize them consistently with the stated likelihood and consequence information, and select mitigation actions proportionate to the documented risk. At least two qualified assessors blinded to experimental condition should score open-ended responses. Inter-rater reliability and the procedure for resolving scoring disagreements will be reported.

AI literacy will be measured using a validated scale selected for relevance to generative AI and information evaluation. The selection should be based on the available evidence concerning reliability, validity, and suitability for the participant population. Lintner (2024) provides a starting point for this process.

Emotional intelligence will be assessed using an established instrument suitable for professional research. The choice of instrument should reflect the study's theoretical definition and avoid unnecessary overlap with general personality, self-confidence, or emotional regulation measures.

Cognitive workload may be included as a supplementary variable if the questionnaire and experimental procedure support its measurement. Its inclusion should not be allowed to expand the model beyond what the sample size and research objectives can reasonably support.

4.5 Ethical considerations and study limitations

Before recruitment, the research protocol should receive approval from the relevant institutional research ethics committee. Participants should provide informed consent and be told that participation is voluntary, that they may withdraw without penalty, and how their data will be stored, protected, and reported. No identifiable project records, confidential contracts, or commercially sensitive documents should be used without authorization. Scenario materials should be fictionalized or appropriately de-identified. The study should also disclose that some AI-generated claims are intentionally unsupported for experimental purposes and provide a debriefing after participation.

The proposed experiment has limitations. A scenario-based task cannot reproduce the full pressures, accountability structures, and consequences of live construction projects. A regional sample from Khyber Pakhtunkhwa may limit generalizability. Participants' responses to visible labels may also differ from their behavior during routine work. These limitations should be acknowledged when interpreting the findings.

PLANNED DATA ANALYSIS

No empirical data have been supplied or collected for this manuscript. Accordingly, this section describes the planned analyses and does not report observed effects, sample characteristics, descriptive statistics, hypothesis tests, mediation estimates, or moderation estimates. The study must not present anticipated outcomes as findings.

The primary analysis will compare claim-level detection performance between the randomized conditions. Because each participant evaluates multiple claims, a mixed-effects logistic regression is appropriate for binary claim-level correctness, with participant and claim treated as random effects where the data support this structure. The model will include experimental condition as the principal fixed effect and prespecified covariates only where justified. Results should report effect estimates, confidence intervals, and exact sample counts.

The second analysis will examine calibrated reliance using the prespecified score that accounts for both false acceptance and false rejection. The analysis will compare conditions and report the separate error components so that a high score cannot be achieved merely by rejecting every claim.

The third analysis will compare decision-quality scores between conditions using a regression model appropriate to the score distribution. The scoring rubric will be applied by assessors blinded to condition. Any adjustment for baseline professional experience or other covariates will be prespecified and justified.

The fourth analysis will examine whether calibrated reliance statistically mediates the relationship between provenance visibility and decision quality. Because the mediator is not independently randomized, mediation results will be interpreted cautiously and will not be treated as definitive proof of a causal mechanism. Indirect effects and bootstrap confidence intervals will be reported if the sample and measurement quality support this analysis.

The fifth and sixth analyses will test whether AI literacy and emotional intelligence moderate the effect of provenance visibility on calibrated reliance. Interaction terms will be evaluated using effect estimates and confidence intervals. These analyses will be treated as secondary or exploratory unless the power analysis demonstrates adequate power. Emotional intelligence will not be described as an established determinant of calibrated reliance; it is an exploratory moderator motivated by its broader relevance to project-management performance.

DISCUSSION

The proposed study addresses an important distinction between the reliability of AI-generated content and the ability of professionals to evaluate that content appropriately. Because data have not yet been collected, the discussion below identifies the implications that would follow under different possible result patterns rather than interpreting observed findings. Technical research has examined ways to improve factual precision, strengthen retrieval-grounded generation, and mitigate hallucinations. The present study extends this line of inquiry by focusing on how the presentation of evidence affects human evaluation and decision-making.

If the experiment demonstrates that provenance labels improve unsupported-claim detection, such a result would support the practical value of presenting evidential status at the point where users evaluate AI-generated claims. Such a finding would complement research on factual evaluation and mitigation by showing that information about support status may also influence the behavior of people using AI systems.

If provenance labels improve calibrated reliance, such a result would be consistent with the principle of appropriate reliance articulated by Lee and See (2004). The important implication would not be that professionals trust AI more in general, but that their acceptance and rejection decisions become more consistent with the evidence. Such an outcome would also be relevant to the concerns about misuse and disuse identified by Parasuraman and Riley (1997).

However, the proposed study cannot assume that provenance labels will produce these outcomes. Users may misunderstand the labels, overlook them under time pressure, or treat them as definitive guarantees. Labels may also create excessive caution, leading participants to reject supported information. The empirical analysis must therefore examine both the benefits and the potential unintended effects of provenance visibility.

CONCLUSION

This study proposes an investigation of evidence provenance, calibrated reliance, and decision quality in AI-assisted construction project risk assessment. Drawing on research concerning retrieval-augmented generation, factuality evaluation, and trust in automation, it examines whether making the evidential status of AI-generated claims visible can help professionals distinguish supported information from unsupported statements.

The proposed experiment compares construction and infrastructure professionals who evaluate AI-generated risk assessments with and without provenance labels. It also examines whether calibrated reliance is associated with decision quality and whether AI literacy and emotional intelligence influence the relationship between provenance visibility and calibrated reliance.

The central contribution is the shift from treating AI reliability solely as a property of the generation system toward examining how evidence information influences professional reliance. Provenance labels may provide a useful mechanism for supporting verification, but their effectiveness cannot be assumed. Their value depends on the accuracy of the labels, the clarity of their presentation, and the ability of users to incorporate them into sound decisions.

The study will contribute empirical evidence only after data have been collected and analyzed. Until then, the proposed relationships should be regarded as hypotheses rather than established conclusions. By maintaining this distinction, the research can provide a defensible foundation for evaluating evidence-aware AI workflows in construction project management.

REFERENCES

1. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024).*
2. Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors, 59*(1), 5–27. https://doi.org/10.1177/0018720816681350
3. Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057
4. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570
5. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.
7. Lintner, T. (2024). A systematic review of AI literacy scales. *npj Science of Learning, 9*, Article 50. https://doi.org/10.1038/s41539-024-00264-4
8. Lv, B., Feng, A., & Xie, C. (2024). Improving factuality by contrastive decoding with factual and hallucination prompts. *Sensors, 24*(21), Article 7097. https://doi.org/10.3390/s24217097
9. Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-T., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12076–12100.
10. Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K.-S., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing and benchmarking large language models in retrieval-augmented generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 10862–10878). Association for Computational Linguistics.
11. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886
12. Rajendran, S., & Singaravelu, D. (2026). *Evidence-bound factual repair in retrieval-augmented LLM answers: Separating semantic localization from deterministic realization* (Version 1) [Preprint]. Preprints.org. https://doi.org/10.20944/preprints202609.0490.v1

13. Zhang, Q., Hao, S., & Chung, K. S. K. (2022). The impact of project manager's emotional intelligence on project performance: A meta-analysis. *Chinese Management Studies, 17*(3), 461–487. https://doi.org/10.1108/CMS-05-2021-0195

14. Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics, 13*(5), Article 856. https://doi.org/10.3390/math13050856