Explanations, Evidence, and Overreliance in Retrieval-Augmented Generation: A Critical Literature Review

Adam Silva

University of the Potomac, Washington, DC, USA
adam.silva@student.potomac.edu
ORCID: 0009-0009-4906-2986

Preprint—not peer reviewed
DOI: https://doi.org/10.5281/zenodo.23328348

Abstract

The presence of evidence in output generated by retrieval-augmented generation (RAG) does not guarantee that the information is correct or that the users will evaluate it properly. This research examines the ethical relationship between explanations, source attribution, and overreliance in artificial intelligence systems. It synthesizes scholarly literature of retrieval and attribution, explanation faithfulness, and human-artificial intelligence interaction. It highlights four properties: factual correctness, evidential support, source credibility, and explanatory faithfulness. The study indicates evidence that explanations can increase acceptance of incorrect answers, however useful sources and lower verification costs can support appropriate reliance. The findings observed in this research do not establish a universal standard of citations or explanations in deployed RAG systems. The review argues that users should maintain the practical ability to question recommendations in responsibly designed systems. A conceptual framework is proposed by connecting evidence presentation to verification effort, opportunities for contestation, and responsibility. Key issues include identifying unsupported claims and source disagreement, distinguishing post-hoc attribution from causal explanation, and evaluating risk controls based on decision outcomes as well as reported trust. The contribution is an interpretation and research agenda, further experiments and empirical research needs to be done to extend this research.

Keywords: retrieval-augmented generation; explainable artificial intelligence; source attribution; overreliance; epistemic autonomy; AI ethics

1 Introduction

Retrieval-augmented generation (RAG) is the use of a large language model (LLM) that has access to external information. An influential work combining both parametric and retrieved knowledge for knowledge-intensive tasks was developed by Lewis et al. [1]. The opportunity for checking information that back up an answer is available through retrieval. There needs to be a relationship between an answer and a document, however, and that relationship needs to be assessed as well. In a 2023 generative search engine audit, it was found that some statements lacked citation support and some citations did not support the statements they accompanied [2]. The results it presents reflect the systems evaluated in 2023, and not the performance of all the current applications of RAG.

Another distinction is provided by explanations. A rationale can explain an answer without faithfully reflecting how the model produced it. Jacovi and Goldberg [3] differentiate between explanation faithfulness and plausibility, which constitutes a foundation for distinguishing between process fidelity and persuasiveness. This complicating factor of human interaction is illustrated in the work of Kim et al. [4] who discovered that explanations and sources were not equally effective in controlled question answering tasks. Overreliance means in this context taking “bad” AI advice. This behavioral measure is not the same as the user's reported trust and is not itself conclusive evidence of a user's unreasonable behavior. Technical advances in attribution won't necessarily lead to similar advances in users' judgements, then.

The ethical implications of these differences are not insignificant. Floridi's approach to AI ethics places explainability alongside issues such as autonomy, responsibility and trust [5]. In this context, this review raises the following questions: When do explanations and evidence provide a basis for a justified reliance on RAG and when does their provision make insufficiently supported answers more difficult to challenge? Here, "justified reliance" is treated as a normative criterion: reliance must correspond to the evidence, the uncertainty that may be relevant, and

the risks of being wrong. The paper relates technical evaluation to this standard and formulates research questions which can be started by conceptual and documentary analysis.

2 Review approach

In this targeted critical literature review, the sources were identified by web searches and by checking the records of scholarly publishers and proceedings, such as the ACL Anthology, ACM publications, NeurIPS, PMLR, Oxford University Press. Search included keywords related to: RAG, citations, explanation faithfulness, overreliance, and decision making. Other focused searches explored the work of Sarah Wiegreffe, Soheil Feizi and Heng Huang. The search was updated on 4 October 2026.

Relevance to the review question, identifiable publication status, and explicit findings or conceptual differences were favored for selection. There are 14 works cited: 13 published articles in conferences or journals and 1 published book. The synthesis was based on data from the publisher records and available full texts; claims were limited to those that could be verified. The book provides ethical context, but not any evidence of RAG performance.

The interpretive comparison focuses on what each study evaluates, the evidence supporting its conclusions, and the inference needed to apply those conclusions to RAG users. Technical benchmarks, behavioral experiments and normative arguments are distinctively dealt with. No database screening, independent coding or formal quality score or pooled effect estimate was conducted. The resulting structure is not intended to provide a quantitative estimate of the prevalence of overreliance for deployments, but rather a review of selected evidence and suggests unaddressed assumptions.

3 What explanations and evidence establish

3.1 Correctness and evidential support

Factual correctness involves the accuracy of an assertion in the context in question. Evidential support deals with making it clear whether the cited material supports the claim as it is stated. The properties can vary: A right answer can include an irrelevant citation, and an inaccurate source can support an incorrect answer. The number of times each occurs is not an estimate, it is a conceptual distinction. Gao et al. [6] implement separate evaluation through the Automatic LLMs' Citation Evaluation (ALCE) benchmark, which assesses fluency, correctness and citation quality. Similarly, Liu et al. [2] make the distinction between whether statements have adequate citation support and whether individual citations support statements.

One example of post-hoc attribution is provided below. Balasubramanian et al. [7] propose DecompTune that decomposes a previously generated answer into units and correlates them with supporting context. The work presents better results in terms of attribution in various sets of questions answering scenarios that were evaluated. This can lead to more explicit connections between claims and passages. However, it is not enough to add supporting sentences after generation to show that these sentences were a cause of the answer or that the content is correct. These are different assessment questions, rather than defects established by the attribution results.

The evaluation of the source credibility is therefore subject to one's own judgment. When assessing the weight of a document, its authorship, date, scope and connection to other evidence are all relevant. Think of a hypothetical system that correctly refers to an outdated institutional policy. The statement displayed might be backed up by the citation but may not provide justification for recommendations for current needs. The ethical issue is that the system provides a more limited level of assurance than what is requested by users. It is important that a design presents that mismatch, instead of having one verification label to conceal it.

3.2 Faithfulness and plausibility

An explanation can be clear and convincing without including factors that affected an output. Balasubramanian, Basu, and Feizi [8] describe problems in explicitly describing implicit biasing cues such as image-based cues in vision-language models. This is to caution against assuming that generated reasoning is a full description of the

behavior of the model. It does not imply that all rationales are unfaithful, nor does it indicate the amount of risk for any specific RAG application.

Palta et al. [9] study another characteristic: changes to perceived plausibility when generated arguments are present. A study by their ACL 2026 finds that human plausibility judgments for commonsense questions change based on arguments for or against the answers. The outcome is a judgment of plausibility, rather than an observed consequential action. Therefore, its relevance here concerns changes in perceived plausibility, rather than direct evidence of overreliance on RAG. When read together with research on explanation faithfulness, it invites reflection on whether users are given reasons for assessing the text independently, or persuasive text that has an unclear connection to the production process.

The four differences should be apparent during an evaluation. Evaluating a source's credibility cannot be replaced by determining if a quoted sentence substantiates a statement. The assessment of readers' beliefs to a text is not a valid substitute for faithfulness assessment. No single property is the determining factor of whether a particular decision should rely on it. The suggested guidelines do not assume that these can be reduced to one score for use in trusts, but rather each is a distinct variable that can inform judgment.

Figure 1 summarizes the four assessment properties and the practical conditions relevant to justified reliance.

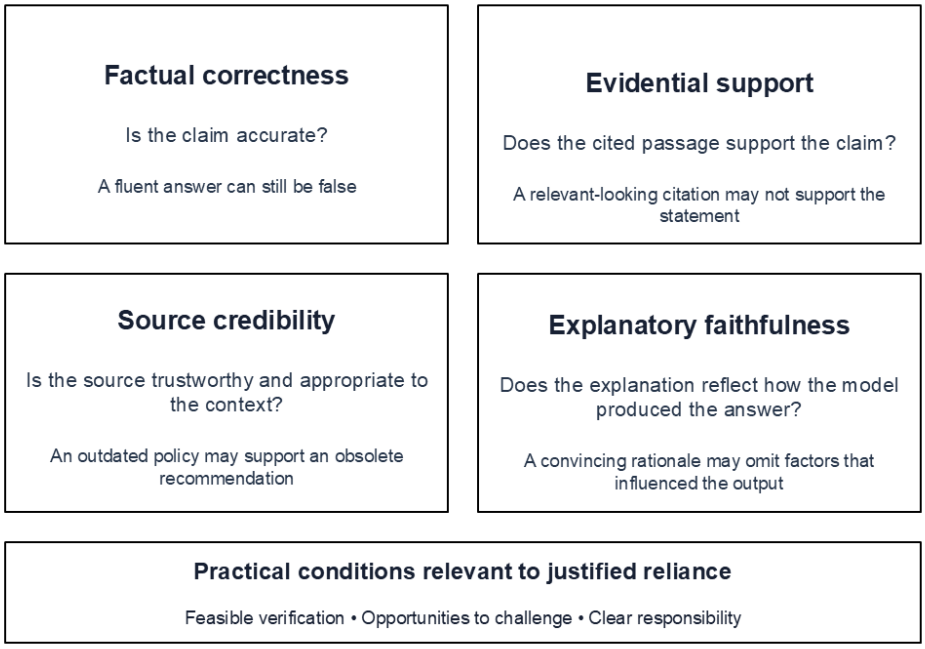


Fig. 1 Four assessment properties and practical conditions relevant to justified reliance in retrieval-augmented generation. The diagram presents a conceptual synthesis

4 Overreliance and its ethical implications

4.1 Explanations can help or hinder judgment

A claim that explanations are protective or harmful faces resistance by behavioral literature. In the evaluated tasks, Bansal et al. [10] found that explanations boosted acceptance of AI suggestions, even when the suggestions were false, but did not significantly improve team performance compared with displaying the AI's recommendation and confidence without an explanation. More relevant LLM evidence is provided by Kim et al. [4]. A controlled experiment revealed that explanations led to higher reliance on correct as well as on incorrect answers while giving sources led to lower reliance on incorrect answers (N = 308). Kim et al.'s experiment used prepared answers to objective questions, which limits generalization to open-ended, ongoing interactions with deployed RAG systems.

Spatharioti et al. [11] also look at decisions based on LLM-powered search. Participants were able to carry out product-comparison tasks faster, but the wrong information generated may lead to wrong decisions. A highlighting intervention improved decision accuracy on the challenging task in the second experiment. This study focuses on the relationship between information presentation and behavior, but does not explore other domains, error patterns, or users for which the relationship between the two might be as similar. It shows, too, why speed and satisfaction aren't enough to show that a decision is of good quality.

Similarly, evidence of beneficent explanations is pertinent. Vasconcelos et al. [12] demonstrate through simulated AI and maze tasks that costs and benefits of checking an explanation impact overreliance. Buçinca et al. [13] found that cognitive forcing reduced overreliance in a nutrition-related decision task, but did not eliminate it and involved trade-offs in participants' subjective ratings. These studies indicate that the amount of work needed to confirm recommendations is important. They are not specific to a particular RAG interface. The synthesis helps support a conditional proposition: to be useful from an ethics point of view, explanations should be ones that allow for discriminating assessment that users can reasonably carry out.

4.2 Epistemic autonomy and verification burdens

There is no concern with a user taking help from a machine. Delegation of information may be reasonable. The concern is whether an interface makes it easier to accept it without being aware of the limitations that are relevant to the decision. Epistemic autonomy, in the sense employed here, refers to one's ability to evaluate, interrogate and rework the basis for one's judgment in practice. More than just a hyperlink is required.

Before a citation is usable evidence, a person has to find the right place, decode technical vocabulary, find the right qualification or compare the accounts from different sources. This review therefore suggests evaluating verification through tasks that reflect the practical demands of checking evidence. The findings in the above references [12, 13] are cost sensitive and drive this proposal, but do not determine the distribution of those demands between the RAG users. How language proficiency, expertise, accessibility needs and time constraints affect verification should be investigated rather than assumed.

There is an ethical dilemma that ensues. Making evidence simple can help to verify it, but stripping away context can obscure uncertainty or disagreement. On the other hand, providing all the passages that have been recovered might place an undue burden on checking. Responsible presentation should be used to facilitate focused inspection whilst still allowing access to context. This is normative design goal based on analysis but not an optimal interface. Whether or not users can recognize the implications of its limitations and respond to them – such as selecting an alternative source or rejecting a suggestion – is what determines its adequacy.

4.3 Responsibility and the limits of transparency

Another concern is how the responsibility is distributed. Providing citations should not be considered enough reason to push all verification responsibility to the end user. The argument is normative, that is that developers and deploying organizations make choices regarding retrieval, selection of evidence, presentation of uncertainty, and opportunities for review. These decisions dictate the circumstances in which users must make a choice. A user should thus be able to challenge a recommendation, while an organization should have the responsibility and practical ability to investigate and correct failures.

Consequential uses should preserve sufficient detail to recreate what was presented, where the information was obtained, and any limitations that were shared, allowing for appropriate privacy protections. These records would help hold responsible people and organizations accountable for the interaction without claiming to provide insight into the model's full internal logic. An inspectable record of evidence does not necessarily reveal what caused the model's output. Providing more explanatory text does not, by itself, establish that responsibilities have been fulfilled.

5 Implications for design and future research

5.1 Make the limits of support inspectable

Technical and behavioral observations suggest a design hypothesis that users may better judge an answer when they are given access to the evidence supporting each consequential claim. Candidate features include links to specific passages, visible qualifications, flagging unsupported claims, and access to conflicting evidence. Attribution research provides techniques to link answer units to context [7] and behavioral research provides reasons to consider actual human use of this link [4]. This hypothesis needs to be tested as a combined RAG intervention.

There should also be a distinction between various reasons for uncertainty in risk communication. Each of these scenarios requires a different response when evidence is missing, contradictory evidence or when the quality of sources is unclear. Those differences might not be apparent through a generic confidence indicator. The advice is to make the reason for caution clearly understandable, and offer a viable action to take next, such as, checking another source or seeking review. This is not a calibrated probability of error nor a guarantee of safety.

5.2 Connect resource allocation to consequences

Verification is complicated by resource limitations. Xue et al. [14] present R2-Router, which chooses a language model and output-length budget to optimize the answer quality and cost. This is relevant to the feasibility of allocating computation, but the study does not evaluate human overreliance or harm control in RAG. This is a research proposal, not an outcome of the previous paper on resource-aware routing, for extending the concept of resource-aware routing to the area of evidence verification.

An extension like this would have to be clarified regarding what events would cause further checks to be carried out and who would benefit from the allocation. A quick response might be acceptable if the information needed is something that can be easily reversed, but not if a decision has serious consequences. Additional computation would also have to demonstrate benefits; a larger model or a longer explanation does not, by itself, establish adequate protection. Evaluators should not conflate cost of operations with the ethical issue of whether there is a tolerance for uncertainty.

5.3 A qualitative research agenda

The research agenda explores three questions. First, what is meant by evidence, grounding and explanation in RAG papers and do the assessment tools align with these? A structured documentary comparison could highlight cases in which the concepts of support, truth and faithfulness are incorrectly treated as synonymous. Second, what are the responsibilities entrusted to users, developers and institutions by suggested interfaces? Qualitative analysis of published descriptions of design can be conducted to see if there are opportunities to meet verification responsibilities.

Third, what is the balance between effort to verify and access to context in published interventions? It might be possible to see from existing studies, if there are any, in which areas proposed safeguards rely upon expertise, time or exceptionally clear evidence. Later behavioral validation may be used to further clarify this framework through these analyses. If such validation is possible, incorrect advice acceptance and correct advice rejection should be included as outcomes, along with the effort expended on the verification and consequences of the decision. Trust reporting would not resolve the central question that may be raised: do users have sufficient reason to rely on the actions that they undertake?

6 Limitations

Coverage is restricted by purposive selection, and selection bias exists, including deliberate attention to the work of three researchers. The review does not establish that the cited findings represent all relevant evidence. Older, foundational studies focus on systems and tasks that are different from modern RAG, and the newer technical papers may not evaluate human actions. Publication and accessible-content checks are not a replacement for replication nor a formal appraisal of all studies. The review focuses on information seeking; extending its conclusions to autonomous agents requires separate analysis of their actions and consequences. The proposed framework and recommendations are conceptual, and not validated measures, interventions or guarantees of harm reduction.

7 Conclusion

Explanations and citations provide information, and the ethical value they have depends on what they allow people to judge. The reviewed literature supports distinguishing factual correctness, evidential support, source credibility, and how faithfully explanations reflect the model's reasoning process. Studies on behavior show both dangers of acceptance and chances to make better judgments depending on the information and tasks being used. Because of this responsible RAG should be judged based on the opportunities for users to check and challenge important claims along with how well it works technically. The important research goal right now is to link these opportunities to clear duties and clear limits of what can be proven. A careful detailed comparison can make these needs clearer before more experiments are done.

References

1. 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.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33, 9459–9474 (2020). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
2. Liu, N.F., Zhang, T., Liang, P.: Evaluating Verifiability in Generative Search Engines. Findings of the Association for Computational Linguistics: EMNLP 2023, 7001–7025 (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.467
3. Jacovi, A., Goldberg, Y.: Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? Proceedings of ACL 2020, 4198–4205 (2020). https://doi.org/10.18653/v1/2020.acl-main.386
4. Kim, S.S.Y., Vaughan, J.W., Liao, Q.V., Lombrozo, T., Russakovsky, O.: Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–26 (2025). https://doi.org/10.1145/3706598.3714020
5. Floridi, L.: The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford University Press, Oxford (2023). https://doi.org/10.1093/oso/9780198883098.001.0001
6. Gao, T., Yen, H., Yu, J., Chen, D.: Enabling Large Language Models to Generate Text with Citations. Proceedings of EMNLP 2023, 6465–6488 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.398
7. Balasubramanian, S., Basu, S., Goswami, K., Rossi, R.A., Manjunatha, V., Santhosh, R., Zhang, R., Feizi, S., Lipka, N.: Decomposition-Enhanced Training for Post-Hoc Attributions in Language Models. Proceedings of EACL 2026, Volume 1: Long Papers, 5070–5084 (2026). https://doi.org/10.18653/v1/2026.eacl-long.236
8. Balasubramanian, S., Basu, S., Feizi, S.: A Closer Look at Bias and Chain-of-Thought Faithfulness of Large (Vision) Language Models. Findings of the Association for Computational Linguistics: EMNLP 2025, 13406–13439 (2025). https://doi.org/10.18653/v1/2025.findings-emnlp.723
9. Palta, S., Rankel, P.A., Wiegreffe, S., Rudinger, R.: Arguments that Alter Minds: LLM Rationales Sway Human (and LLM) Notions of Plausibility. Proceedings of ACL 2026, Volume 1: Long Papers, 13132–13152 (2026). https://doi.org/10.18653/v1/2026.acl-long.599
10. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D.S.: Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–16 (2021). https://doi.org/10.1145/3411764.3445717
11. Spatharioti, S.E., Rothschild, D., Goldstein, D.G., Hofman, J.M.: Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–15 (2025). https://doi.org/10.1145/3706598.3714082
12. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M.S., Krishna, R.: Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proceedings of the ACM on Human-Computer Interaction 7(CSCW1), Article 129, 1–38 (2023). https://doi.org/10.1145/3579605
13. Buçinca, Z., Malaya, M.B., Gajos, K.Z.: To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction 5(CSCW1), Article 188, 1–21 (2021). https://doi.org/10.1145/3449287
14. Xue, J., Lou, Q., Xing, J., Huang, H.: R2-Router: A New Paradigm for LLM Routing with Reasoning. Proceedings of the 43rd International Conference on Machine Learning, PMLR 306, 143255–143268 (2026). https://proceedings.mlr.press/v306/xue26h.html