Boosting-Inspired Validation of Retrieval-Augmented Generation in Structured Scientific Knowledge Bases

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Language Models (LLMs) enhanced with Retrieval-Augmented Generation (RAG) achieve remarkable results, yet they often hallucinate or provide incomplete answers. This poses critical challenges in scientific knowledge domains where factuality and precision are essential. In this paper, we propose a boosting-inspired evaluation framework for RAG that combines iterative error reduction with forward-looking retrieval mechanisms from FLARE. Unlike existing work that primarily optimizes retrieval or ranking, our focus is on the validation loop itself. We validate the framework in a controlled scenario using Citavi, a structured literature management system, serving as a reproducible environment for testing. Results indicate that strict substring matching underestimates semantic correctness, while boosting-inspired metrics highlight when expansion is necessary. This proof-of-concept demonstrates technical feasibility and motivates iterative, semantic validation for future scientific assistants.

Human Foreword

This paper is a meta-experiment. It was created in a continuous collaboration where a human researcher acted as supervisor and generative AI systems (ChatGPT-4 and GPT-5) took on the role of creator of the scientific proof-of-concept. The entire workflow - from exploring the research question, through drafting code and structuring the validation scenario, to producing the manuscript - was conducted within a single ChatGPT chat.

To ensure transparency, the GitHub repository referenced in the acknowledgements provides open access to the complete chat history (in German and English translation), the validation project, and the Citavi test project.

1 Introduction

Large Language Models (LLMs) have rapidly advanced natural language processing and are increasingly applied in scientific and industrial domains. Despite their remarkable capabilities, a persistent challenge remains: LLMs tend to hallucinate, producing factually incorrect or unverifiable content [1]. This shortcoming is particularly problematic in scientific knowledge bases, where accuracy, reproducibility, and transparency are essential.

Retrieval-Augmented Generation (RAG) [2] mitigates this issue by combining parametric memory stored in model weights with non-parametric memory retrieved from external corpora. While RAG improves factual grounding, it lacks systematic validation loops to ensure that retrieved evidence is sufficient and that generated answers remain reliable. In practice, validation is often reduced to ranking metrics, leaving gaps in coverage and robustness unaddressed.

Boosting methods, such as Gradient Boosting [3] and XGBoost [4], demonstrate the effectiveness of iteratively reducing residual errors. Similarly, FLARE [5] introduced forward-looking retrieval, in which intermediate predictions guide expansion towards missing evidence. Both approaches highlight the importance of iterative refinement, a principle not yet fully leveraged in RAG validation.

This paper is motivated by the need for reliable validation mechanisms in knowledge-intensive environments. We therefore introduce a methodology that integrates boosting-inspired residual tracking with FLARE-style expansion, enabling a dedicated evaluator for RAG. To examine its feasibility, we conduct a pilot validation in a structured environment (Citavi), which provides citations, abstracts, and hierarchical knowledge suitable for controlled testing. The results reveal the limitations of strict string-matching metrics and highlight the necessity of semantic evaluation for future iterations. Together, these contributions demonstrate the potential of boosting-inspired validation as a new direction for improving the robustness of retrieval-augmented generation in scientific knowledge bases.

2 Related Work

Our work builds on three main strands of research: ensemble learning and boosting, retrieval-augmented models, and evaluation of hallucinations and factuality. Each area contributes important foundations, yet none addresses the specific problem of designing validation loops for Retrieval-Augmented Generation (RAG).

2.1 Boosting and Ensembles

Ensemble learning, and boosting in particular, has proven to be a powerful method for iterative error reduction. Friedman introduced Gradient Boosting [3], which was later extended in practical implementations such as XGBoost [4]. Further theoretical contributions, such as the comprehensive review by Bühlmann and Hothorn [6], and the classic textbook by Hastie, Tibshirani, and Friedman [7], emphasize the principle of repeatedly fitting residuals to improve predictive performance. This principle inspires our evaluator design, which aims to detect and act upon coverage gaps in retrieved evidence.

2.2 Retrieval-Augmented Models

In parallel, retrieval-augmented models have become central to modern language technologies. Lewis et al. presented RAG [2], combining parametric knowledge embedded in model weights with non-parametric retrieval. Guu et al. extended this line with REALM [8], where retrieval is interleaved during pretraining, and Karpukhin et al. introduced Dense Passage Retrieval (DPR) [9]. Izacard and Grave proposed Fusion-in-Decoder (FiD) [10], while Borgeaud et al. scaled retrieval to trillions of tokens in RETRO [11]. More recently, Izacard et al. introduced FLARE [5], which uses forward-looking predictions to actively expand retrieval. Dialogue-focused systems [12] and domain-specific adaptations such as scientific question answering [13] illustrate the breadth of RAG applications. These advances strengthen factual grounding, but none of them explicitly incorporates validation mechanisms that monitor adequacy and completeness.

2.3 Corrective and Adaptive Retrieval

Corrective and adaptive retrieval approaches show growing awareness of this gap. Corrective Retrieval Augmentation (CRA) [14] integrates error signals into retrieval, and Self-RAG [15] combines retrieval, generation, and reflection in a unified loop. Adaptive retrieval methods [16] explore query reformulation and contextual retrieval to minimize drift. All share conceptual ground with boosting in that they iteratively improve results. However, their focus remains on generation rather than on dedicated validation.

2.4 Learning to Rank

Learning-to-rank methods contribute another relevant dimension. LambdaMART [17] and listwise approaches [18, 19] provide effective techniques for ranking retrieval candidates, while large-scale challenges such as the Yahoo! Learning to Rank dataset [20] established benchmarks for progress.

These methods optimize retrieval quality but do not address the broader question of whether retrieved evidence is sufficient to validate generated answers.

2.5 Evaluation and Hallucinations

Finally, evaluation of hallucinations and factuality in natural language generation has gained increasing attention. Ji et al. surveyed hallucination phenomena [1], while Maynez et al. [21] and Zhao et al. [22] analyzed factuality in summarization and question answering. Classical metrics such as precision, recall, and nDCG [23] remain standard, yet they rely on strict string matching and often underestimate semantic adequacy. Surveys of retrieval-augmented methods [24–26] and benchmarks like BEIR [27] provide useful overviews, but none establish explicit validation loops. Recent initiatives such as FEVER [28] and Izacard et al.'s active retrieval paradigm [29] further underline the need for iterative, validation-oriented approaches.

Taken together, the literature reveals three key insights. Boosting highlights the power of iterative error reduction, retrieval-augmented models enhance factual grounding, and evaluation research exposes the limitations of current metrics. What is still missing is an integrated framework that connects these strands by validating retrieval adequacy through iterative mechanisms. Closing this gap is the objective of the methodology described in the following section.

3 Methodology

The goal of this work is to develop a validation framework for Retrieval-Augmented Generation (RAG) that integrates principles from boosting and FLARE. Unlike prior research that primarily optimizes retrieval or generation, our focus is on the evaluation loop itself: determining whether retrieved evidence is sufficient, identifying residual gaps, and deciding when expansion is necessary. The methodology is designed to be dataset-agnostic and can be applied to any structured knowledge base. In this section we describe the design principles, system architecture, graph representation, evaluator logic, and performance indicators before introducing the validation scenario in Section 4.

3.1 Design Principles

As discussed in Section 2, three strands of research motivate our design: boosting demonstrates the power of iterative error reduction [3, 4, 6], retrieval-augmented models such as RAG, REALM, and FLARE improve factual grounding [2, 8, 5], and evaluation studies expose the limitations of current metrics [1, 21]. From boosting we adopt the idea of residual tracking: in each step, what remains uncovered is treated as error to be addressed. From FLARE we adopt forward-looking expansion: when residuals exceed a threshold, additional retrieval is triggered. Together, these principles transform validation into an iterative process rather than a static one-time assessment.

3.2 System Architecture

The framework is organized into four stages. First, the *ingest stage* prepares structured input and artifacts. Second, the *graph construction stage* initializes a knowledge graph that captures elements and relations in a compact form. Third, the *retriever stage* combines sparse retrieval (BM25) with dense embeddings for semantic similarity, similar to approaches in open-domain QA [9]. Finally, the *evaluator stage* applies the boosting- and FLARE-inspired logic that distinguishes our approach from existing retrieval systems.

3.3 Graph Representation

Knowledge is represented as a graph to enable transparency and incremental updates. Nodes correspond to citations, documents, or categories, while edges capture references, group membership, or hierarchical relations, as is common in knowledge graph construction [**? ?**]. The initial graph is deliberately small, containing only citations and linked documents. Expansion introduces categories or additional documents as new nodes, increasing search space and recall. By tracking which nodes have been covered, the graph directly supports boosting-style residual measurement.

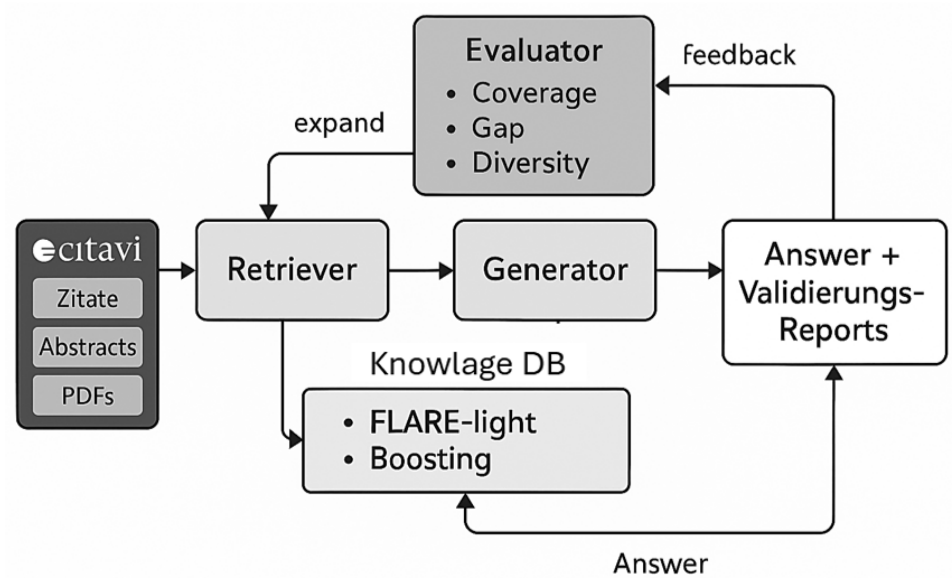

Figure 1: Workflow of the validation framework with Citavi input, RAG retrieval, and FLARE/boosting-inspired evaluation. The dataset-specific component (Citavi) is applied only in the validation scenario described in Section 4.

3.4 Evaluator Logic

The evaluator is the methodological core. Inspired by boosting, it calculates residuals by measuring the gap between retrieved evidence and gold references. Inspired by FLARE [5], it then decides whether to expand or stop: if the gap is large, the graph is expanded and retrieval is repeated; if coverage is sufficient, the loop halts. Related approaches such as Corrective Retrieval Augmentation (CRA) [14] and Self-RAG [15] share elements of this idea, but they focus on improving generation rather than providing a dedicated validation loop. Our evaluator reframes these principles as an explicit mechanism for adequacy checking.

3.5 Key Performance Indicators

Several key performance indicators operationalize validation. Coverage@k measures whether gold evidence appears among the top-k retrieved items. Because strict matching is often too rigid, we extend this to *Semantic Coverage*, which uses cosine similarity in embedding space. Normalized Discounted Cumulative Gain (nDCG) [23] captures ranking quality and is widely applied in retrieval evaluation [27]. Two additional metrics extend FLARE principles: *Gap-FLARE* quantifies the proportion of uncovered evidence that should trigger expansion, and *Diversity-FLARE* measures the variance among retrieved results to avoid redundancy. Together, these indicators provide a multidimensional perspective on validation that goes beyond classical IR metrics.

3.6 Abstraction and Outlook

A crucial property of the methodology is that it remains independent of the specific dataset. It can be applied to enterprise document collections, scientific repositories, or any other structured corpus. In this paper, we use Citavi only as a controlled testbed to examine feasibility, not as part of the method itself. The overall workflow is illustrated in Figure 1, which also anticipates the next section where Citavi is introduced as the validation scenario.

4 Validation Scenario

The proposed methodology was validated in a controlled pilot study. The goal of this validation was not to achieve competitive performance, but to demonstrate the feasibility of boosting-inspired evaluation in a structured environment. This section describes the scope and constants of the experiment, the setup of the validation run, the role of Citavi as a structured testbed, the obtained results, and their interpretation.

4.1 Scope and Constants of the Validation

The validation was deliberately constrained in order to focus on the core question of feasibility. Several restrictions were imposed: iterative graph updates were disabled, the number of queries was limited to five, and user feedback was excluded. These choices reduced complexity and ensured that the experiment could be reproduced reliably.

Certain aspects of the setup were treated as constants. The graph was limited to citations and documents, excluding higher-level categories. Citation types in Citavi served as proxy labels, which avoided manual annotation but introduced rigidity. Retrieval was fixed to a combination of BM25 and embedding similarity. Together, these constants provided a stable environment, even though they also introduced biases.

Within this controlled setting, the element under validation was the evaluator. The experiment was designed to test whether boosting-inspired residual tracking and FLARE-style expand/stop logic could be operationalized in practice. The consistent decisions made by the evaluator serve as evidence of feasibility, even if the metrics themselves reveal limitations.

4.2 Experimental Setup

The validation run was implemented as a snapshot experiment. At initialization, a small graph was constructed containing citations and their associated documents. Retrieval was carried out using BM25 and dense embeddings, with the two lists merged before evaluation. The evaluator then computed Coverage@5, Semantic Coverage, nDCG@5, Gap-FLARE, and Diversity-FLARE. Logs, result CSV files, and summary JSON files were generated to provide full transparency of the run. In total, five queries were executed, each paired with a gold citation to serve as reference evidence.

4.3 Citavi as Structured Testbed

Citavi was chosen as the validation environment because of its structured organization of knowledge. Citations, abstracts, and full-text PDFs are stored in a unified project file, with categories and groups providing additional hierarchical structure. These features map naturally onto graph representations: citations and documents become nodes, while references and categories form edges. Furthermore, citation types (direct quote, summary, paraphrase) function as proxy labels for relevance, allowing evaluation without manual labeling. This makes Citavi an effective testbed for feasibility studies, even though it is not part of the methodology itself.

4.4 Results

The outcomes of the validation run are summarized in Table 1. Exact string matching yielded no correct hits, while semantic inspection revealed partial correctness in one case. In all cases, the evaluator returned the decision to expand.

4.5 Interpretation

The validation run shows that the evaluator operated consistently and as designed. All five queries resulted in expand decisions, reflecting the detection of residual gaps. Coverage@5 remained at zero under strict string matching, while manual semantic inspection indicated partial adequacy in at least one case. The gap between exact and semantic coverage demonstrates a limitation of the applied metrics. These findings establish the technical feasibility of the evaluation loop and provide the empirical basis for the broader discussion in Section 5.

Table 1: Validation setup, key performance indicators, and results. Gold labels are citations from the Citavi project. Coverage is reported for exact and semantic matching.

Query	Gold Label	Exact Cov.@5	Sem. Cov.@5	nDCG@5
What is FLARE?	FLARE iteratively uses a prediction	0	1	0.0
How does RAG combine memory?	RAG combines parametric memory	0	0	0.0
What is Gradient Boosting?	Gradient boosting is a generalization	0	0	0.0
What is LambdaMART?	LambdaMART combines gradient boosting	0	0	0.0
What does REALM interleave?	REALM interleaves knowledge retrieval	0	0	0.0

5 Discussion and Future Work

The validation presented in Section 4 provides a narrow but informative demonstration of the framework. In this section, we move beyond the specific scenario and discuss what the results imply for the methodology introduced in Section 3 and for the broader research gap identified in Section 2.

5.1 Implications for the Methodology (Section 3)

The validation confirmed that two central design elements of the methodology are operational: boosting-inspired residual tracking and FLARE-style expand/stop decisions. These findings support the feasibility of treating adequacy as a residual and of embedding expansion as a control mechanism in validation. At the same time, the scope of the experiment revealed which aspects of the methodology remain untested. Iterative updates, semantic coverage metrics, and richer graph representations were not exercised in the pilot run. Their absence does not invalidate the design, but highlights the areas where further empirical work is required. The validation therefore partially substantiates the methodology, while pointing to open components.

5.2 Connection to the Research Gap (Section 2)

The limitations observed in Section 4 resonate with prior critiques in the literature. Classical metrics such as Coverage and nDCG underestimated semantic adequacy, echoing findings from hallucination and factuality research [1, 21]. Benchmarks such as BEIR [27] have already called for richer evaluation, but they lack an explicit validation loop. Our framework contributes in this direction by treating validation as an iterative process, informed by residuals and expansion. While the Citavi pilot is minimal, it illustrates that the research gap identified in Section 2 can be addressed with a concrete operational design.

5.3 Limitations of the Present Study

The present study is constrained by deliberate design choices: a small number of queries, reliance on proxy labels, and the exclusion of iteration and user feedback. These restrictions were necessary to ensure reproducibility in a proof-of-concept, but they limit the generalizability of the results. The implication is not that the methodology is invalid, but that further studies are required to evaluate its robustness in larger and more diverse settings.

6 Conclusion

6.1 Summary

This paper proposed a validation framework for Retrieval-Augmented Generation (RAG) that integrates boosting-inspired residual tracking with FLARE-style expand/stop logic. The methodology shifts the focus from optimizing retrieval or generation to validating adequacy itself, treating uncovered evidence as residuals and using expansion as a control mechanism.

A pilot validation in a Citavi-based testbed confirmed technical feasibility. The evaluator consistently identified residual gaps and triggered expand decisions, demonstrating that the two guiding principles of the methodology can be implemented in practice. At the same time, the restricted scope—five queries, proxy labels, no iterative updates—revealed limitations: classical string-based metrics such as Coverage@k and nDCG underestimated semantic adequacy, and expand decisions could not influence retrieval outcomes. These findings establish a foundation for iterative, feedback-driven validation but stop short of a full performance benchmark.

6.2 Future Work

Future work will extend the framework along several directions. First, iterative cycles must be enabled so that residuals and expansion interact dynamically across multiple retrieval rounds. Second, semantic similarity measures will be integrated to capture adequacy beyond surface-level matching, ensuring that paraphrases and equivalent formulations are recognized. Third, richer graph structures should be employed, incorporating categories and cross-document relations to broaden coverage. Fourth, user feedback can be leveraged as an additional residual signal, bridging automated evaluation with practical relevance. Finally, the framework should be applied to larger and more diverse benchmarks such as BEIR as well as to industrial document collections, to assess robustness and scalability.

Taken together, these steps will move the approach from a controlled proof-of-concept toward a practical methodology for improving the reliability of retrieval-augmented generation in scientific and industrial contexts.

References

[1] Ziwei Ji, Nayeon Lee, Jason Fries, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 2023.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.

[3] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[5] Gautier Izacard, Fabio Petroni, Lucas Hosseini, et al. Active retrieval-augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[6] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[8] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020.

[9] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[10] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models. *arXiv preprint arXiv:2101.00294*, 2021.

[11] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katharine Millican, et al. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*, 2022.

[12] Kurt Shuster, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive dialogue. *arXiv preprint arXiv:2101.00168*, 2021.

[13] Michihiro Yasunaga et al. Retrieval-augmented scientific question answering. *arXiv preprint arXiv:2203.08115*, 2022.

[14] Omer Ram et al. Corrective retrieval augmentation. *arXiv preprint arXiv:2303.09858*, 2023.

[15] Weijia Shi et al. Self-rag: Learning to retrieve, generate, and reflect. *arXiv preprint arXiv:2310.11511*, 2023.

[16] Angeliki Lazaridou et al. Adaptive retrieval for question answering. *arXiv preprint arXiv:2204.11117*, 2022.

[17] Christopher JC Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2006.

[18] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 129–136, 2007.

[19] Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2009.

[20] Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 14, pages 1–24, 2011.

[21] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1917, 2020.

[22] Wayne Zhao et al. Calibrated measures of hallucination for open-domain qa. *arXiv preprint arXiv:2102.01521*, 2021.

[23] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[24] Luyu Gao et al. A survey on retrieval-augmented generation. *arXiv preprint arXiv:2307.03172*, 2023.

[25] Huayang Li, Luyu Gao, et al. Retrieval-augmented text generation: A survey. *arXiv preprint arXiv:2202.01110*, 2022.

[26] Richard Evans and Edward Grefenstette. Language models as knowledge bases? *arXiv preprint arXiv:2102.01096*, 2021.

[27] Nandan Thakur, Nils Reimers, Johannes Daxenberger, et al. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 43rd European Conference on IR Research (ECIR)*, pages 3–22, 2021.

[28] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: Fact extraction and verification. In *Proceedings of NAACL-HLT*, pages 809–819, 2018.

[29] Gautier Izacard et al. Towards active retrieval for language models. *arXiv preprint arXiv:2202.07227*, 2022.

A Technical Appendices and Supplementary Material

Technical appendices with additional results, figures, graphs and proofs may be submitted with the paper submission before the full submission deadline, or as a separate PDF in the ZIP file below before the supplementary material deadline. There is no page limit for the technical appendices.

Agents4Science AI Involvement Checklist

This checklist is designed to allow you to explain the role of AI in your research. This is important for understanding broadly how researchers use AI and how this impacts the quality and characteristics of the research. **Do not remove the checklist! Papers not including the checklist will be desk rejected.** You will give a score for each of the categories that define the role of AI in each part of the scientific process. The scores are as follows:

- **[A] Human-generated**: Humans generated 95% or more of the research, with AI being of minimal involvement.
- **[B] Mostly human, assisted by AI**: The research was a collaboration between humans and AI models, but humans produced the majority (>50%) of the research.
- **[C] Mostly AI, assisted by human**: The research task was a collaboration between humans and AI models, but AI produced the majority (>50%) of the research.
- **[D] AI-generated**: AI performed over 95% of the research. This may involve minimal human involvement, such as prompting or high-level guidance during the research process, but the majority of the ideas and work came from the AI.

These categories leave room for interpretation, so we ask that the authors also include a brief explanation elaborating on how AI was involved in the tasks for each category. Please keep your explanation to less than 150 words.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "Agents4Science AI Involvement Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[B]**

   Explanation: The central idea – combining boosting-inspired validation with RAG and using Citavi as the structured testbed – originated from the human researcher. Generative AI contributed by exploring alternative framings and drafting formulations, but it required significant clarification and supervision before aligning with the intended approach. Thus, hypothesis development was primarily human-driven, with AI providing supportive input.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[C]**

   Explanation: The experimental design and implementation were generated mostly by AI. Generative AI produced the Docker Compose setup, Python application code, and evaluator logic for the validation loop. The human researcher supervised, ensured executability, and applied minimal adjustments (e.g., correct handling of SQLite rows from Citavi and path alignment). Thus, while the technical foundation came from AI, the human role was critical for validation and final operability.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[D]**

   Explanation: The analysis of data and the interpretation of the results were carried out exclusively by the generative AI. The AI processed the outputs of the validation experiments, produced explanations, and articulated the interpretation of adequacy and gaps in coverage.

The human researcher did not perform independent analysis but only supervised the process from a meta-level. Thus, data analysis and interpretation were exclusively AI-driven.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[D]**

   Explanation: The entire writing process—including generation of text, creation of figures and tables, and compilation of the reference list—was carried out exclusively by the generative AI. The human researcher did not contribute to the manuscript text itself but acted only in a supervisory role. Thus, the writing of the paper was exclusively AI-driven.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: The observed limitations of generative AI in this project were varied, but the documentation also allows the reader to grasp them for themselves. The link to Github with the entire chat history with ChatGPT, as well as the validation scenario and the paper, is published in the acknowledgements for this paper. Because this chat history already showed a strong trend toward a scientific feasibility review of a concept before the call was launched, it was logical to make it available to your project. The following points from the conversation were particularly noticeable: On the one hand, ChatGPT had difficulty consistently reusing information throughout the entire workflow. For example, towards the end, results from the beginning of the chat history were hardly considered during the paper creation process. ChatGPT frequently relied on its pre-trained background knowledge instead of using the provided project files. Only after explicit inquiries did ChatGPT indicate that it would only superficially review the file. Particularly with papers used for training and also considered in this project, it was impossible to deviate from existing knowledge. Taking the direct approach without critically questioning assumptions, even though critical doubt and methodological rigor are essential in scientific work, was the greatest difficulty in this project. Towards the end of the process, repeated inquiries from the human supervisor were necessary to ensure that the AI had partially considered the provided information and integrated it into the paper. Post-correction for the paper was deliberately omitted. The overall quality of the results was complete and assessable as an independent work with strong support for, for example, an academic paper, but was more reminiscent of a satisfactory bachelor's thesis grade 3.0, as it severely lacked depth, consistency, and critical reflection. This was also influenced by the special setup: the entire workflow was carried out in a single chat, from the brief idea of a term, its contextualization, deriving possible synergies, identifying the use case and research question, and finally creating the paper itself. In cases where individual steps are examined over several sessions, the process of creating a scientific paper was deliberately followed from start to finish in a continuous dialogue. This structure created a workflow similar to supervised student work: The human took on the role of supervisor, while the AI took on the role of the students and was guided to do scientific work. The AI was able to create depth for clearly defined sub-goals, but often lost the overall overview and rushed into creating final versions, which is why several loops were created. The ChatGPT fluctuated between superficial overviews and repeated refinements of simple to-dos with limited added value. These dynamics—including strengths and weaknesses—are documented in the chat transcript included in the repository formatted for readability.

Agents4Science Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **Papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers and area chairs. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given. In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "Agents4Science Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [No]

   Justification: The idea and solution approach are well structured, and the validation scenario was chosen with a very strong scope. However, the results did not provide evidence that this approach delivers clear added value.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The paper reflects openly on boundaries and challenges of the chosen proof-of-concept scope, but these are discussed mainly in a general scientific reflection rather than presented as method-specific limitations.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not contain formal theoretical results or proofs. It presents a proof-of-concept study with generative AI and human supervision rather than theorem-driven research.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: The validation presented in the paper is rudimentary and yields only weak results; reproducibility of the experiments is not ensured within the paper itself. While the supplementary chat log and Citavi test project, as well as the GitHub reference in the acknowledgements, provide useful material for further research on AI workflows, the scientific question is only conceptually outlined and not supported by reproducible, verifiable results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The GitHub repository provides open access to the validation project, the complete chat history with ChatGPT, and the Citavi test project, along with instructions. These resources allow other researchers to faithfully reproduce the experimental setup and verify the main claims.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the relevant experimental setting for the proof-of-concept, including the restricted scope (five queries, no user feedback, no iterative tuning) and the use of a Citavi test project. While continuous retraining of data was not implemented in the validation, the conditions under which the presented results were obtained are sufficiently described.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: This paper does not report error bars or statistical significance measures. However, the restricted scope of the validation (five queries, no feedback, no retraining) was explicitly defined and used consistently to evaluate the proposed method.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The validation scenario was lightweight and could be executed on standard computing resources, that why the paper does not provide explicit details on hardware type, memory, or runtime.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: This paper is conforms fully with the Agents4Science Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper is the direct outcome of a ChatGPT-assisted scientific writing process and does not include a dedicated discussion of potential positive or negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies.