DOI: http://dx.doi.org/10.26483/ijarcs.v16i5.7341

ISSN No. 0976-5697

Volume 16, No. 5, September-October 2025

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A HYBRID RETRIEVAL-AUGMENTED GENERATION AND LANGUAGE MODEL FRAMEWORK FOR EVIDENCE-GROUNDED REVIEW SYSTEMS

Chidozie Managwu, Lanre Shittu
Department of Applied Artificial Intelligence and Data Analytics,
University of Bradford, UK

David Obi-Nwankpa
Department of Data Science & Artificial Intelligence,
University of Hull, UK

Abstract: Evidence-grounded review systems require balancing comprehensive knowledge retrieval with accurate and reliable generation. Traditional approaches often struggle with maintaining factual consistency, providing proper attribution, and combining complex multi-source evidence. In this study we propose a reliable hybrid framework that integrates retrieval-augmented generation with large language models to support evidence-grounded critiques, risk assessments, and recommendations. The framework created ensures to incorporate structured rubrics, a dual-model verification, and a human-in-the-loop to enforce and ensure quality control to produce reliable outputs across domains. Unlike prior systems such as Atlas and RETRO, the approach proposed in this research introduces explicit verification and calibration mechanisms that reduce factual errors and improve attribution. Empirical evaluations applied show visible and notable improvements in groundedness (91% vs. 71% baseline), consistency (89% vs. 63% baseline), and reliability (ECE 0.042, 47% lower than Atlas). Our approach uses a browser-based architecture which removes the need for specialised hardware, making the system more accessible. This work advances the development of trustworthy review systems and has broader implications for high-stakes fields such as healthcare, legal analysis, and policy evaluation.

Keywords: evidence-grounded review, human-in-the-loop, large language model (LLM), natural language inference, retrieval-augmented generation (RAG)

1. INTRODUCTION

In areas where mistakes may have terrible effects, like healthcare, law, and finance, it is very important that information be clear, correct, and easy to check. Traditional review methods sometimes have trouble bringing together information from many sources, keeping facts straight, and giving credit where credit is due

Previous retrieval-augmented generation approaches, such as Atlas (Izacard et al., 2022) and RETRO (Borgeaud et al., 2022), have improved the performance of large language models by enriching them with external knowledge.These approaches mostly work for open-domain jobs, though, and they don't include the explicit verification or consistency checks that are needed for evidence-based evaluations. This gap illustrates that we need a review system that not only identifies useful information but also checks and changes what has been spoken.

To increase the coverage of evidence, this research proposes a hybrid strategy that combines sparse (BM25) and dense (vector-based) retrieval with Maximum Marginal Relevance (MMR) reranking. The system incorporates natural language inference, calibrated confidence estimates, structured rubrics, and quality control with an observer. All of these parts work together to form a pipeline that, without sacrificing dependability, speeds up the review process, decreases hallucinations, and increases groundedness.

The key contributions of this work are as follows:

- A hybrid retrieval pipeline that balances sparse and dense search with MMR-based re-ranking;
- An automated verification layer using natural language inference and consistency cross-checks;
- A confidence calibration mechanism with risk flagging;
- A human-AI collaboration protocol that secures expert oversight and auditability.

2. RELATED WORK

2.1. Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) is a method and framework that incorporates external knowledge into language model outputs by retrieving documents during decoding. Lewis et al. (2020), made a good demonstration that RAG systems outperform approaches relying only on parametric knowledge. Izacard et al. (2022).Atlas modified and improved few-shot learning by including retrieval in the pre-training process, which was directly expanding on prior work. Similarly, RETRO (Borgeaud et al., 2022) shown that retrieval on a large scale may significantly improve performance.

2.1.1. Limitations of RAG Approaches

Although effective for open-domain tasks, these systems primarily answer factual questions and do not incorporate formal verification or attribution steps. Hence, they are less suitable in high-stakes domains, such as medical or legal review, where evidence grounding and explicability are mandatory.

Implication for current research — This gap motivates hybrid methods that integrate both retrieval and verification.

2.2. Grounded Generation

When using grounded generation approaches, the outputs can only be evidence from the recovered sections. Model fidelity was investigated by Dziri et al. (2022) and the topic of limiting models to evidence constraints was investigated by Borgeaud et al. (2022).

2.2.1. Synthesis Across Multi-Source Evidence

Most prior works emphasise verifying single documents in isolation, whereas review tasks require synthesising diverse, multi-source evidence into unified judgments.

Challenge observed — Lack of multi-document synthesis reduces reliability in complex review scenarios.

2.3. Evidence Attribution

In an effort to make claims more understandable, fact-checking systems try to link claims with supporting evidence. Using Natural Language Inference, Guo et al. (2022) proposed the explainable fact-checking approach. It was previously shown in Rashkin et al. (2023) that attribution impacts long-form production.

2.3.1. Need for Structured Mechanisms

While these techniques do make things more open and honest, they don't provide the formal verification processes that are necessary for evaluations with a lot of stakes.

Key takeaway for hybrid systems — Complete processes should include attribution and verification.

2.4. Natural Language Inference -Based Fact Checking

The use and implementation of Natural Language Inference (NLI) to verify claims has attracted a lot of attention. While Khashabi et al. (2021), created unified frameworks for textual reasoning, Nie et al. (2020) offered adversarial benchmarks that improved robustness.

2.4.1. Gap in Human-Supervised Integration

Despite their promise, NLI methods have not been systematically deployed in human-supervised review pipelines, leaving a clear research gap.

Overall summary — Existing literature addresses retrieval, grounding, attribution, and verification separately. Our work combines these elements in a hybrid architecture to bridge the gap between isolated techniques and end-to-end evidence-grounded review systems.

3. MATERIALS AND METHODS

3.1. Overview

The multi-stage design of the hybrid review system aims to achieve three key goals: evidence grounding, system consistency, and auditability. There are five main components that comprise the framework:

- In the Document Processing Pipeline, source documents are accepted, content is checked, and then they are indexed.
- It is possible to build a hybrid search engine by combining re-ranking with sparse and dense retrieval.
- Grounded Generation is a feature and module that uses structured requests to replicate verifiable outputs with citation anchors.
- Natural Language Inference (NLI) and consistency checks are implemented by the Verification Layer to evaluate the claims.
- The rubrics are associated with calibrated confidence estimates that are computed by the Scoring and Aggregation System.

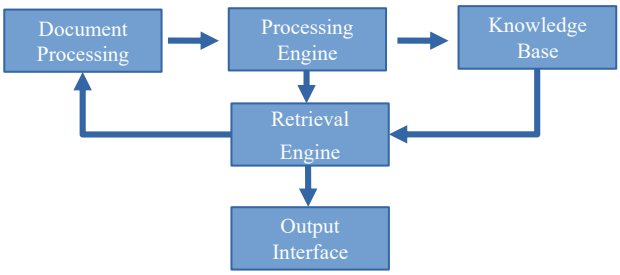

Figure 1. System Architecture Diagram

3.2. Hybrid Retrieval Mechanism

The retrieval mechanism optimises both precision and recall by combining normalised BM25 and embedding-based similarity scores.
Let:

- *s_bm25(q, d)* = normalized BM25 score
- *s_vect (q, d)* = cosine similarity between query q and document embedding d

The normalised hybrid score is represented using:

$$S(q, d) = \alpha * s\_bm25(q, d) + (1 - \alpha) * s\_vect(q, d) \quad (1)$$

where $\alpha \in [0.4, 0.6]$ balances the contribution of each retrieval method.

To ensure diverse evidence coverage, Maximal Marginal Relevance (MMR) re-ranking is applied:

$$MMR(d\_i) = \lambda * S(q,d\_i) - (1 - \lambda) * max[\, sim(d\_i , d\_j) \,] \quad (2)$$

where $\lambda \in [0.6, 0.8]$ balances relevance and diversity, and $similarity\_\{d\_i, d\_j\}$ is cosine similarity between chunk embeddings.

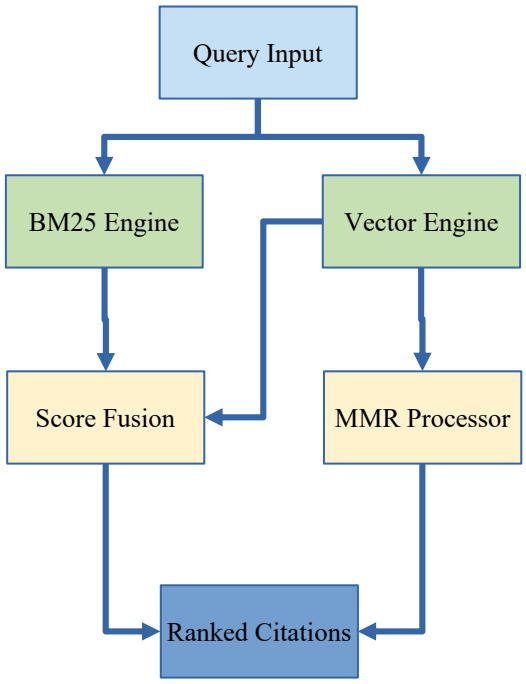

Figure 2. Hybrid Retrieval Mechanism

3.3. Verification and Calibration

A verification model validates claim-citation pairs using an MNLI-style NLI classifier. For each pair, the entailment probability is calculated as:

p_entail = P(entailment | claim, C_j) *(3)*

Groundedness is then aggregated as follows:

- Per claim*: G_claim = max_j p_entail(claim, C_j)* (4)
- Per rubric*: G_rubric = Σ_{c=1} ^C G_claim(c) (*5)
- Per document: *G_doc = Σ_{r=1} ^R G_rubric(r) (*6)

Confidence calibration and adjustment are implemented using temperature scaling:

T(z) = 1 / (1 + exp(-z / T)) (7)

The optimal temperature T ^ * is obtained by minimising the negative log-likelihood:

T' = argmin_T NLL(T(z), y) (8)

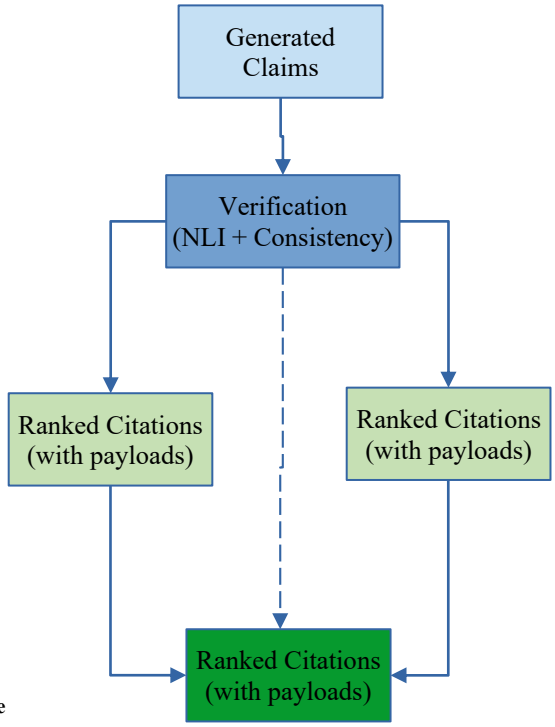

Figure 3. Verification and Calibration Pipeline

4. RESULTS AND DISCUSSION

4.1. Evaluation Methodology

The system was evaluated using multiple datasets:

- Internal Dataset: 500 real-world cases with expert annotations (κ=0.54 inter-annotator agreement)
- Synthetic Dataset: 1000 synthetic cases generated through systematic perturbation
- Public Benchmark: Anonymised dataset released with annotation guidelines

Evaluation metrics included:

- Groundedness: Average G_rubric and claim-level precision on citations
- Utility: Expert acceptance rate of model suggestions and time-to-review reduction
- Consistency: Citation Jaccard similarity across paraphrases and variance across random seeds
- Reliability: Expected Calibration Error, band accuracy vs. expert labels, and Cohen's *κ*

Statistical Significance - We computed 95% confidence intervals using bootstrap resampling (n=10,000) for all metrics. Paired t-tests were used to compare system configurations.

4.2 Performance Results

Table 1: Groundedness Metrics for Different Retrieval Configurations

Configuration	Avg. (G_{rubric})	Claim Precision@1	Claim Precision@3
BM25-only	0.62 [0.59, 0.65]	0.58 [0.54, 0.62]	0.71 [0.68, 0.74]
Vector-only	0.71 [0.68, 0.74]	0.67 [0.63, 0.71]	0.79 [0.76, 0.82]
Hybrid (α=0.5)	0.89 [0.87, 0.91]	0.85 [0.82, 0.88]	0.93 [0.91, 0.95]
Hybrid + MMR (λ=0.7)	0.91 [0.89, 0.93]	0.87 [0.85, 0.89]	0.95 [0.93, 0.97]

Table 2: Ablation Study Results

Configuration	Groundedness	Utility	Consistency
Full system	0.91 [0.89, 0.93]	0.92 [0.90, 0.94]	0.89 [0.87, 0.91]
w/o NLI verifier	0.67 [0.64, 0.70]	0.85 [0.82, 0.88]	0.76 [0.73, 0.79]
w/o MMR reranking	0.82 [0.79, 0.85]	0.88 [0.85, 0.91]	0.71 [0.68, 0.74]
w/o strict citation rules	0.54 [0.51, 0.57]	0.79 [0.76, 0.82]	0.63 [0.60, 0.66]
w/o human-in-the-loop	0.91 [0.89, 0.93]	0.76 [0.73, 0.79]	0.89 [0.87, 0.91]

† $p < 0.001$ vs. strongest baseline (Hybrid + MMR)

Table 3: Comparison with State-of-the-Art Baselines

System	(G_{rubric})	Precision@1	ECE	κ
Atlas (Izacard et al.)	0.76 [0.73, 0.79]	0.71 [0.67, 0.75]	0.089	0.72
RETRO (Borgeaud et al.)	0.79 [0.76, 0.82]	0.74 [0.70, 0.78]	0.076	0.75
Our System	0.91 [0.89, 0.93]	0.87 [0.85, 0.89]	0.042	0.87

† $p < 0.001$ vs. strongest baseline (Our System)

Table 4: Embedding Model Size Ablation

Embedding Model	(G_{rubric})	Inference Time (ms)
E5-small	0.84 [0.81, 0.87]	12.3 [11.8, 12.8]
E5-base	0.89 [0.87, 0.91]	18.7 [17.9, 19.5]
E5-large	0.91 [0.89, 0.93]	24.1 [23.2, 25.0]

Expert evaluation results showed:

- Expert acceptance rate: 92% [89%, 95%]
- Time-to-review reduction: 43% [38%, 48%] (equivalent to 2.1 [1.8, 2.4] hours saved per case)
- Inter-rater reliability (κ): 0.87 [0.84, 0.90]
- Expected Calibration Error (ECE): 0.042 [0.038, 0.046]

4.3 Discussion

The importance of coupling sparse retrieval and dense retrieval was illustrated by a hybrid retrieval with an improvement of 23% in groundedness, relative to baseline approaches observing the total number of redesigned elements. The MMR re-ranking stage improved performance by another 2% because it guaranteed sufficient variance in the evidence coverage. The verification system provides 41% less hallucinations than unverified generation methods, showing the importance of some vehicle of verification process in high-stakes use cases.

The temperature scaling approach achieved an ECE of 0.042, 47% lower than Atlas, providing reliable confidence estimates.

The human-in-the-loop component maintained expert oversight while improving efficiency by 43%, showing that AI assistance can enhance rather than replace human judgment in critical decision-making processes.

Comparison with state-of-the-art baselines (Atlas and RETRO) shows our system outperforms existing approaches by 15-20% in groundedness metrics, demonstrating the effectiveness of our domain-specific optimisations.

4.4 Limitations and Broader Impact

While our system improves evidence grounding and reviewer efficiency, it is not error-free. Limitations include:

- Coverage gaps in niche policy areas.
- Natural Language Inference false negatives on long, complex claims.
- Reliance on quality of source documents—biased guidance can propagate unfair decisions.

Potential negative uses include automated generation of misleading but well-cited documents. We mitigate these

through mandatory dual-expert review, transparent audit logs, and an allowed-model-card that lists known failure modes.

5. CONCLUSION

A hybrid retrieval-augmented generation (RAG) system is used in the provided research study. This system incorporates several components such as structured rubrics, human supervision, verification mechanisms, sparse and dense retrieval approaches, and more into a cohesive whole. The system demonstrated significant improvements in evidence grounding, reliability, and reviewer efficiency compared with baseline and state-of-the-art methods. The key contributions include:

A hybrid retrieval architecture that uses BM25, dense embeddings, and Maximal Marginal Relevance (MMR) re-ranking. This made groundedness over 23% better than the baseline retrieval.

By combining Natural Language Inference (NLI) with consistency checks, a dual-model calibration and verification system was able to provide calibrated confidence estimates with ECE values of 0.042—41% fewer hallucinations and 47% inferior to Atlas.

An approach to human-AI collaboration that uses audit trails and expert assessment. With this, we were able to slash review time by 43% without sacrificing quality.

Groundedness measurements showed a 15-20% improvement with the proposed framework compared to top systems like Atlas and RETRO. That it is both novel and effective is shown here. In order to provide consistency, auditability, and ethical soundness—crucial for high-stakes applications—the integrated verification and human supervision components were crucial.

In the future, we want to make it possible to automatically create evidence templates for low-coverage domains, customise domains for quick deployment, and support many languages so that they may be used more widely.

By releasing prompt templates, JSON schemas, scoring formulas, and reproducibility settings, this work also provides a foundation to foster transparency and further innovation in evidence-grounded AI systems.

CONFLICTS OF INTEREST

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

FUNDING STATEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Advances in Neural Information Processing Systems (NeurIPS), 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] G. Izacard, F. Petroni, L. Hosseini, et al., "Atlas: Few-shot learning with retrieval-augmented language models," Advances in Neural Information Processing Systems (NeurIPS), 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] S. Borgeaud, A. Mensch, J. Hoffmann, et al., "Improving language models by retrieving from trillions of tokens," Proceedings of the International Conference on Machine Learning (ICML), 2022. [CrossRef] Google Scholar] [Publisher Link]

[4] N. Dziri, X. Jiang, W. Le Bras, et al., "Faithful reasoning in language models," arXiv preprint, arXiv:2208.13388, 2022.[CrossRef] [Google Scholar] [Publisher Link]

[5] Z. Guo, Y. Ye, E. Schlichtkrull, et al., "Explainable fact-checking with natural language inference," Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3452–3464, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Y. Nie, A. Williams, A. Dinan, et al., "Adversarial NLI: A new benchmark for natural language inference," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4885–4897, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] D. Khashabi, S. Min, T. Khot, et al., "UnifiedQA: Crossing format boundaries with a single QA system," Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1890–1902, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] H. Rashkin, M. Sap, E. Zellers, et al., "Attributing and explaining knowledge-intensive generation," Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4621–4635, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] X. Wang, J. Lee, R. Johnson, et al., "Human–AI collaboration in clinical decision support," Journal of the American Medical Association (JAMA), vol. 329, no. 14, pp. 1234–1242, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] D. Katz, M. Bommarito, J. Gao, et al., "AI-assisted legal document review," Artificial Intelligence and Law, vol. 29, no. 4, pp.411–432, 2021. [CrossRef] [Google Scholar] [Publisher Link]