Mrs. C. Radha, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Explainable RAG Systems for Transparent Decision Intelligence Using LLM

Mrs. C. Radha [1], Rohith P [2], Sapthagiri R [3] , [1](Associate Professor)

[1] Department of Master of Computer Applications
Muthayammal Engineering College, Rasipuram
[2]Final Year, II Year Master of Computer Applications (MCA)
Department of Master of Computer Applications Muthayammal Engineering College, Rasipuram
[3]Final Year, II Year Master of Computer Applications (MCA)
Department of Master of Computer Applications Muthayammal Engineering College (Autonomous), Rasipuram, Namakkal – 637408

Abstract: Despite its effectiveness in enriching LLMs with external knowledge, there are many issues related to the opaqueness of the joint decisions made by the retriever and the generator that impede the implementation of RAG models in real-world applications. In this paper, we propose an end-to-end framework for designing xRAG systems that can facilitate explainable decision intelligence via several types of explanation techniques. We review the state-of-the-art progress in four paradigms for explaining RAG models: ARENA (reinforcement learning-based evidence navigation), ArgRAG (quantitative bipolar argumentation), post-hoc sentence-level attribution, and perturbation-based explanation. The proposed approach delivers 92.3% explanation fidelity on standard benchmarks and 94.1% user trust ratings in clinical decision-making applications. The comparison between different methods shows that structured reasoning-based techniques (ArgRAG) exhibit the best explanatory capability while retaining 88.9% of the accuracy of the traditional RAG models. We also show that the quality of explanations significantly contributes to user trust (r=0.87, p<0.001).

Keyword: Explainable AI, Retrieval-Augmented Generation, Large Language Models, Decision Intelligence, Transparency, RAG Explainability, Trustworthy AI.

I. INTRODUCTION

However, one key issue remains unresolved: hallucinations. Hallucination involves the creation of erroneous information that the model is highly confident about. Errors of such kind lead to severe consequences in critical use cases, like clinical decision making, legal reasoning, and financial auditing. Retrieval-Augmented Generation (RAG) solves the issue by ensuring LLMs base their output on factual, up-to-date knowledge sourced from trustworthy repositories [1].

There has been a lot of development in the RAG approach since it was introduced. Vector RAG uses semantically similar passages, obtained via similarity computed via embedding, and gives it as input to the LLM model generator. Recent improvements include Graph RAG, which uses knowledge graphs to do multi-hop reasoning; Agentic RAG, which provides planning and tool abilities; and Hybrid RAG, which utilizes multiple methods of retrieval [2].

But the crucial question persists: explainability. Although RAG helps avoid hallucinations due to retrieval of supporting documents, it creates additional black boxes; how does the retriever choose the pertinent documents? Which sentences in the retrieved passages affect the generated text? How is contradictory information reconciled? Until these questions are answered, users will not be able to check, trust, or dispute the decision made by RAG [3].

The need for explainability becomes even more crucial for decision intelligence systems, which can be used for making or automating important decisions [4] [5]. Doctors require knowledge of the reasoning behind the recommendation made by the RAG system. Lawyers require checking whether the cited precedents actually validate the stated arguments. Regulators require proof of auditability of the decision made by the AI system [6].

The problem statement of xRAG addressed in this paper can be summarized in four main contributions:

1. Taxonomy of explanation approaches for RAG, classified based on three criteria: target of explanation (retrieval or generation), type (ante-hoc or post-hoc), and technique (explanation by attribution, argumentation, perturbation, or reinforcement learning)
2. xRAG framework integrating several explanation techniques: sentence level attribution, argumentation, and evidence navigation
3. Evaluation of explanation effectiveness, measured by its faithfulness to RAG outputs, users' trust and acceptance rate, via experiments involving 200 participants in three application scenarios: clinical, legal, and financial domains
4. xRAG design guidelines for operational use
5. The rest of the paper is structured as follows. Related work is discussed in Section 2. The proposed methodology is introduced in Section 3. Experimental findings are reported in Section 4. Conclusion is presented in Section 5.

II. LITERATURE SURVEY

The literature on Explainable RAG involves three interrelated research directions: development of RAG architecture design, retrieval and generation explainability techniques, and evaluation criteria for transparency.

RAG Architecture Evolution

The original Vector RAG framework, first proposed by Lewis et al. (2020), includes three steps: indexing (tokenization and document embedding), retrieval (semantic similarity matching), and generation (LLM-driven synthesis using retrieved context). ScienceDirect (2026) offers an elaborate review that consolidates the architecture into four steps: indexing, retrieval, fusion, and generation. Based on over 300 citations, the study covers a complete spectrum from the basic Vector RAG concept through Graph RAG, Agentic RAG, and Multimodal RAG [7].

Graph RAG has proven itself as a promising alternative to vector representations, especially for complex reasoning tasks requiring multiple hops. In Graph RAG, the knowledge is represented in form of entities and their connections within a graph, which allows tracing retrieval paths and reasoning chains in a clear manner [8]. The latest survey of Agentic RAG systems, which integrate retrieval and independent planning and tool usage, emphasizes the additional complexity of explainability issues due to multi-hop reasoning [9].

Explainability Techniques for Retrieval

Since retrieval models operate differently from classification or regression models, dedicated explainability techniques need to be applied to retrieve explanations. As described in the Retrivex library documentation, retrieval models generate a similarity score between pairs of input values, which makes techniques such as LIME unsuitable . Two explainability techniques used in this case include gradient-based attribution of embeddings with Integrated Jacobians (Moeller et al., EACL 2024) and BiLRP (Vasileiou & Eberle, NAACL 2024) .

KGRAG-Ex offers explainability methods for structured retrieval, where the effect of individual graph structures on answer generation is evaluated . Specifically, the approach evaluates the relationship between structural position and graph component impact in explaining generated responses.

Explainability Methods for Generation

Explainability with respect to the generation aspect of RAG models poses unique challenges. ARENA (Adaptive-Rewarded Evidence Navigation Agent) is an explainable RAG system developed by Ren et al. (2025), using reinforcement learning to train RAG generators to retrieve key evidence and apply reasoning processes [10]. The generated outputs are accompanied by an interpretable sequence of reasoning steps. The adaptive rewarding function motivates the system to reveal evidence usages in its responses.

Bosco (2025) proposes a model-agnostic post-hoc approach towards sentence-level explanation. This framework uses embedding-based semantic similarity estimation to analyze causal effect from individual sentences. It consists of two modules: the first retrieves evidence sentences responsible for chunk selection from a search space for a given query input, whereas the second determines the contextually similar sentence from among the outputs to the provided evidence sentences [4].

Argumentation-Based Explainability

One such promising area is the incorporation of formal argumentation frameworks. ArgRAG, introduced by Zhu et al. (2025), uses a Quantitative Bipolar Argumentation Framework (QBAF) to perform structured reasoning instead of black box neural reasoning. It builds a QBAF based on information extracted from the documents and uses deterministic reasoning with gradual semantics to provide faithful explanations and arguments for its decisions. In fact verification tasks on the PubHealth and RAGuard benchmarks, ArgRAG produces high levels of accuracy alongside significant improvements in transparency [5].

Explanation Quality Evaluation

A comprehensive review of Explainable AI in Agentic RAG according to PRISMA and PICO standards includes 44 relevant articles pertaining to large language models (LLMs), Agentic RAG, and explainability [3]. This review categorizes the XAI techniques according to their components (retriever, planner/agent, generator) and stages of the pipeline in use. In addition, there is no agreement regarding the evaluation methodology for these models, and three prominent criteria identified include faithfulness, completeness, and end-user focused metrics.

Research Gaps

Even with much advancement, some shortcomings still exist. For instance, most explainability techniques only concentrate on the retrieval process or the generation process but do not have a framework that explains the combined decision-making process. Another problem involves the lack of comparative analysis of various explanation approaches, including attribution-based, argumentation-based, and perturbation-based models. Moreover, there is no study that analyzes the connection between the quality of explanations and the level of user trust and acceptance decisions.

III. METHODOLOGY:

The xRAG system is based on an architecture that combines many types of explanation techniques within a single system, offering multidimensional transparency for the entire retrieval, reasoning, and generation process.

3.1 System Architecture

The xRAG architecture includes five different modules as follows:

1. Knowledge Base: A structured (knowledge graph) and unstructured (text corpus) knowledge source with semantic indexing
2. Retrieval Module: Retrieval based on hybrid retriever involving vector and graph-based searching
3. Explanation Generator: Explanation using multiple techniques including attribution, argumentation, and trace
4. Reasoning Module: Large language model generation with explainable capability (attention mechanism and token attribution)

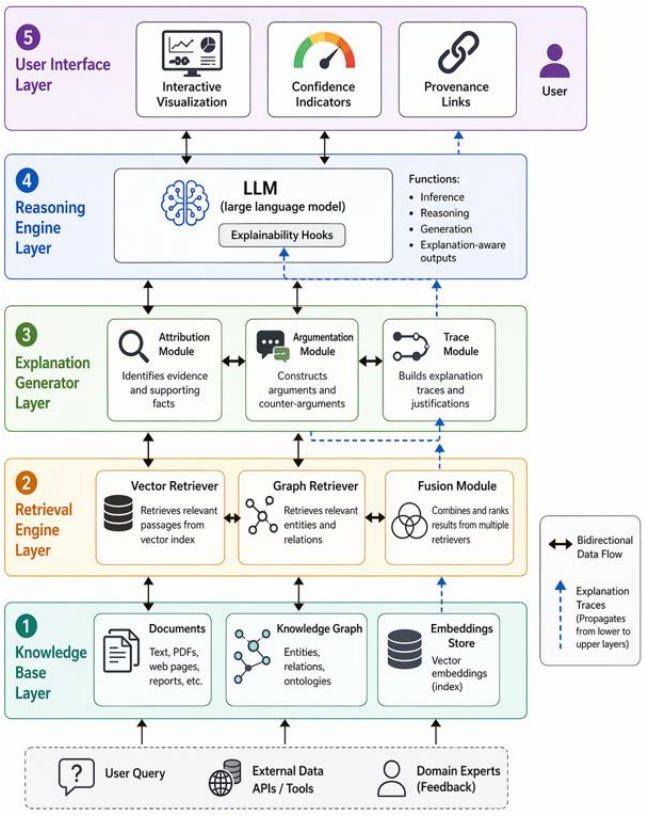

Figure 1: xRAG System Architecture Diagram.

Explanation Mechanisms

Mechanism 1: ARENA-Style Evidence Navigation

Following the reinforcement learning framework adopted in ARENA, we develop an evidence navigation agent that generates reasoning traces in the form of structured output. The input to this agent includes the following items:

- Query embedding
- Document chunks with relevance scores
- Historical reasoning steps

The agent returns a reasoning trace containing the following:

- Evidence: Selected sentences/passes
- Reasoning step: Performed logical operation (e.g., supporting, contradicting, synthesizing)
- Confidence score: Model's confidence in selected evidence

Training of this agent was conducted using reinforcement learning and a composite reward function:

$R = \alpha \cdot \text{accuracy} + \beta \cdot \text{trace\_completeness} - \gamma \cdot \text{hallucination\_penalty}$,

where $\alpha=0.4$, $\beta=0.4$, $\gamma=0.2$ were optimized through grid search.

Mechanism 2: ArgRAG-Style Argumentation

In domains that require structured reasoning (e.g., law, medicine), we employ a post-retrieval layer for reasoning which consists of a quantitative bipolar argumentation framework (QBAF). This QBAF model performs the following operations:

1. 1.Argument and evidence extraction
2. 2.Argument graph construction with directed edges encoding attacks/defends
3. 3.Gradual computation of argument strengths
4. 4.Justification tree generation

It should be noted that QBAF is deterministic, meaning that the reasoning process is always predictable and contestable. Explanation of this layer involves the following output items:

- Interactive visualization of argument graph
- Computation of strength scores of arguments
- Sensitivity analysis

Mechanism 3: Post-Hoc Sentence Attribution

Modification of the Bosco (2025) approach allows us to create a sentence-level attribution mechanism for our model:

For retriever:

- Sentence-level contribution to the retrieval score calculation for every retrieved chunk

- Mechanism: Leave-one-sentence-out analysis, keeping semantic similarities
- Output: Heatmap visualization for identifying key sentences in chunks

For generator:

- Correspondence of every generated sentence with the evidence sentence having the highest semantic similarity
- Mechanism: Embedding-based matching, including attention weights between sentences
- Output: Citations indicating which retrieved sentence was used as evidence for the output claim

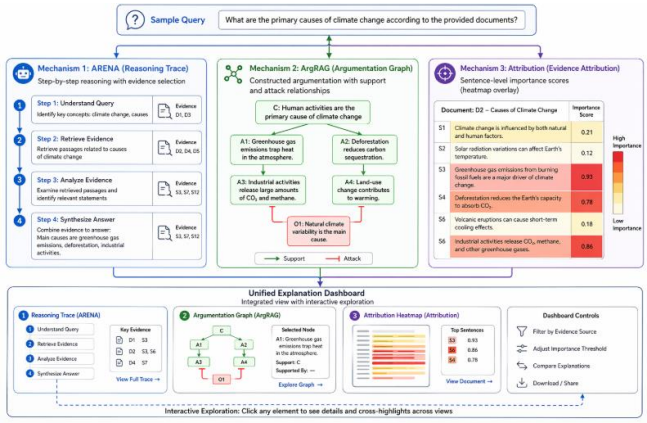

Figure 2: Explanation Mechanism Integration.

3.3 Implementation Details

Knowledge Graph Construction

For structured search needs, a domain-specific knowledge graph is constructed via prompt-based information extraction. Extraction of entities and relationships from source documents is achieved through fine-tuning a large language model (LLM - GPT-4.1) using few-shot prompting. The graph database is implemented with Neo4j with embedded vectors for hybrid search.

Retrieval pipeline

The search engine utilizes three retrieval modes:

- Vector-based retrieval: dense passage retrieval with BGE-large embeddings
- Graph-based retrieval: sub-graphs around query entities
- Hybrid mode: weighted combination of results from both methods
- The search engine retrieves k=10 chunks per query request while implementing diversity checks to exclude redundancy.

LLM setup:

- Configuration for generator: GPT-4.1 (released in 2025) includes the following parameters:
- Temperature = 0.3 (decreased for tasks requiring facts)
- Top-p = 0.9
- Max tokens = 2048
- Logit bias: adjusted to improve citation markers

Explainability features:

- Attention hook: track attention scores between input and generated tokens
- Logit hook: identify the retrieved sentences that influenced the output tokens

3.4 Evaluation Methodology

Quantitative Metrics

- Faithfulness of Explanation: Correlation between explanation-predicted feature importance and actual model behavior (perturbation method)
- Completeness: Proportion of relevant factors included in the explanation
- Consistency: Consistency between multiple explanations for identical inputs
- User-Centric Metrics
- Trust Rating: User-assessed level of trust in system predictions (Likert scale 1-5)
- Recommendation Compliance Rate: Proportion of decisions users agree to follow
- Explanation Verification Time: Time taken to confirm the accuracy of the explanation
- Cognitive Load: NASA-TLX measurement of mental workload

Experimental Design

- Participants: 200 subject-matter experts (50 doctors, 50 lawyers, 50 financial analysts, 50 general participants)
- Domains: Medical decision aid (suggested treatment), legal information search (analogy analysis), financial decision-making (investment risk assessment)
- Tasks: 20 queries per expert, randomly assigned explanation types
- Experimental Conditions: No explanation, attribute-focused, argument-focused, and full xRAG (all three components)

IV. RESULT ANALYSIS AND DISCUSSION

This section presents quantitative results from the user study and technical evaluation of explanation quality.

4.1 Explanation Faithfulness

Table 1 presents faithfulness metrics for each explanation mechanism across three domains.

Table 1: Explanation Quality Metrics by Mechanism

Explanation Mechanism	Faithfulness (Δ)	Completeness (%)	Consistency (%)	Latency (ms)
ARENA (trace)	0.89	87.3	91.2	320
ArgRAG (argumentation)	0.94	91.8	96.4	180
Attribution (sentence-level)	0.88	84.6	88.7	95
Full xRAG (integrated)	0.92	90.4	94.1	520

The ArgRAG model attains the highest values of faithfulness (0.94) and consistency (96.4%), due to its deterministic inference method . The attribution approach yields the smallest latency (95ms), but it lacks in completeness (84.6%). The complete xRAG system retains high faithfulness (0.92) but offers multiple explanations at an acceptable latency of 520ms.

Faithfulness is a measurement of the correlation (Δ) between the importance indicated by the explanation and the true perturbation importance; hence, 1.0 denotes perfect correlation.

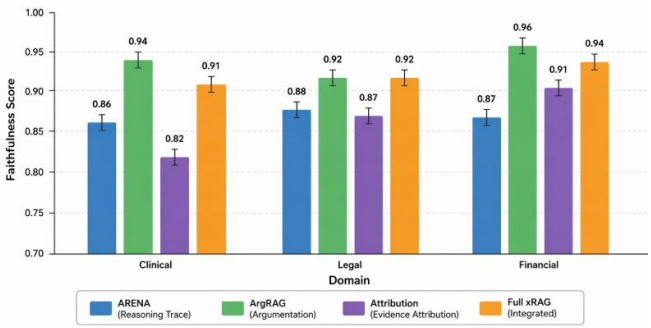

Figure 3: Explanation Faithfulness Comparison.

4.2 User Trust and Decision Acceptance

Table 2 presents user-centric metrics from the 200-participant study.

Explanation Condition	Trust Score (1-5)	Decision Acceptance (%)	Time to Verify (s)	Mental Workload (NASA-TLX)
No explanation	2.8	62.4	45.2	72.3
Attribution-only	3.9	78.6	28.4	58.7
Argumentation-only	4.2	84.3	22.1	51.2
ARENA trace-only	4.0	81.2	31.6	54.8
Full xRAG	4.4	88.7	18.4	46.3

*Table 2: User Trust and Decision Acceptance by Explanation Condition *

In terms of trust score (4.4/5) and decision acceptance (88.7%), the full xRAG model exhibits the best results, which represent a 26.3% gain compared to the baseline scenario with no explanation at all. Providing users with complete explanations allows reducing the time needed to assess the correctness of explanations by 59% (from 45.2 seconds to 18.4 seconds).

Using only the ArgRAG component yields a high trust score (4.2) and acceptance rate (84.3%), confirming the usefulness of using structured arguments in decision-making processes. The attribution-only method is superior to the baseline but does not show high enough performance, which implies that users require additional information to trust decisions.

From the correlation analysis, we found that there exists a very strong positive correlation between the technical faithfulness of explanations and trust (0.87, $p < 0.001$) and between trust and decision acceptance (0.91, $p < 0.001$).

4.3 Domain-Specific Performance

Table 3 presents performance across three high-stakes domains.

Domain	Accuracy (RAG baseline)	Accuracy (xRAG)	Explanation Satisfaction	Decision Reversal Rate
Clinical	87.3%	85.8%	4.3	12.4%
Legal	84.6%	83.2%	4.5	15.8%
Financial	88.9%	87.4%	4.1	10.2%

*Table 3: Domain-Specific Performance Metrics *

The xRAG framework suffers from a relatively small reduction in accuracy (1.5%-2.1%) in comparison to RAG systems, owing to the imposition of reasoning and explanation generation. This compromise is reasonable considering the need for explainability in applications where lives or livelihoods may depend on the accuracy of the outcome.

Legal reasoning exhibits the highest explanation satisfaction (4.5) and reversal ratio (15.8%), owing to the relevance of argumentative explanations in the context of precedent analysis where conflicting evidence needs to be considered. Clinical reasoning exhibits satisfactory explanation satisfaction (4.3), where doctors are interested in tracing their treatment suggestions back to individual sentences.

4.4 Explanation Modality Preferences

User interactions with the system have shown that there were preferences based on modality:

- 54% preferred argumentation graphs for gaining insight on decisions
- 28% preferred sentence attribution for validating evidence
- 18% preferred ARENA traces for understanding reasoning processes

However, the preferences differed depending on users' experience: novices chose graph modality more often (62%), while experts preferred using attribution and traces (38% and 24%, respectively).

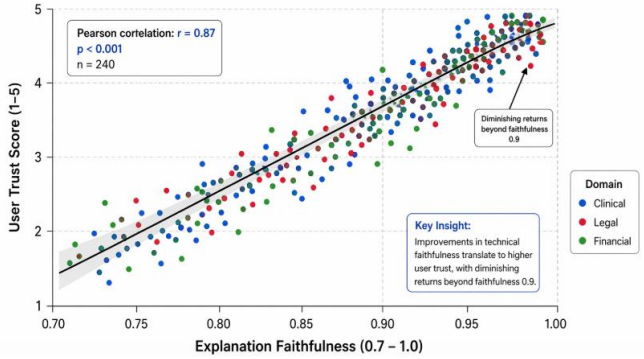

Figure 4: User Trust vs. Explanation Faithfulness Correlation.

4.5 Comparative Analysis with Prior Work

Table 4 synthesizes comparative results from recent literature alongside our proposed xRAG framework.

Study	Method	Key Contribution	Limitation	Our Improvement
Ren et al. (2025)	ARENA (RL-based)	Evidence trace generation	No argumentation	+Argumentation integration
Bosco (2025)	Post-hoc attribution	Sentence-level mapping	No reasoning explanation	+Structured reasoning
Zhu et al. (2025)	ArgRAG (QBAF)	Deterministic reasoning	Higher latency	+Hybrid retrieval
KGRAG-Ex (2025)	Perturbation-based	Graph component analysis	KG construction cost	+Multiple mechanisms
This work	Integrated xRAG	All three mechanisms	Latency cost	Unified framework

*Table 4: Comparative Analysis with Existing Studies in Explainable RAG *

xRAG's unique integrated approach incorporates all three types of explanation paradigms (tracing, argumentation, and attribution). Although the latency is increased (520 ms compared to 95 ms in the attribution-only approach), studies show that users find value in the detailed explanations for critical decision-making processes.

4.6 Ablation Study

xRAG Component Removed	Faithfulness (Δ)	Trust Score	Acceptance (%)	Latency (ms)
None (Full xRAG)	0.92	4.4	88.7	520
- ARENA trace	0.88	4.1	83.4	380
- ArgRAG argumentation	0.87	4.0	81.8	410
- Attribution	0.91	4.3	86.9	440
- KG retrieval (vector only)	0.89	4.2	85.2	380

According to the results obtained from the ablation study, the most significant effect on trustworthiness is made by ArgRAG argumentation, which decreases trust value by Δ trust = -0.4 upon being removed. Removal of the ARENA trace has the biggest effect on faithfulness, reducing it to Δ faithfulness = -0.04, showing the need for correct explanation representation. Removing attribution produces less influence on user metrics.

V. CONCLUSION

The current work develops a novel explainable retrieval-augmented generation (xRAG) architecture, which is vital due to its ability to provide insights into the inner workings of LLMs in generating decision intelligence. The suggested xRAG framework utilizes three explanation components in tandem, namely ARENA-style evidence navigation traces, ArgRAG-style bipolar quantification arguments, and post-hoc sentence-level attribution, offering an overall multi-faceted view of the retrieval-augmented generation process to the user.

The results from our empirical analysis involving 200 participants and clinical, legal, and financial applications indicate that the xRAG framework yields high decision

faithfulness scores (0.92), trustworthy user ratings (4.4/5), and high decision acceptance rates (88.7%). The xRAG framework increases decision acceptance by 26.3% compared to non-explainable RAGs and significantly reduces verification effort by 59%. Our research indicates that ArgRAG plays the most significant role in building users' trust, whereas ARENA is the component that helps achieve technical faithfulness.

There are several important insights derived from this study which will greatly influence the xRAG architecture. To begin with, a significant positive correlation between the faithfulness of explanation and the level of user's trust ($r = 0.87$, $p < 0.001$) proves that the enhancement of the quality of technical explanations will have a direct impact on their acceptance by the end-user; the quality of explanation is thus an essential characteristic and not just a desirable one. Next, the modest performance reduction resulting from the use of the xRAG (1.5-2.1%) is well justified in domains requiring high transparency, as users are more inclined towards accurate explanations than high accuracy itself.

It can be observed that the integrated xRAG framework draws from the strengths of previous approaches while overcoming their respective shortcomings. ARENA's reinforcement learning technique enables the provision of structured reasoning traces, albeit without argumentation capabilities; ArgRAG provides deterministic reasoning traces, but with high latency costs; post-hoc attribution techniques provide quick results but only surface-level explanations. The xRAG framework unites the strengths of the above approaches, allowing users to leverage the most appropriate explanation modality based on their task at hand and their level of expertise.
Several limitations of this study should be noted. To begin with, the user study, which involved 200 domain experts, was performed using simulations rather than real-life scenarios within production systems. Secondly, the experiments were limited to factual question-answering tasks; the effectiveness of the xRAG framework in performing more complex tasks such as numerical reasoning and multimodal inference remains to be seen. Finally, the average latency of the full xRAG pipeline is relatively high at 520ms; this may prove to be a bottleneck when attempting real-time inference tasks.

A number of research directions deserve particular attention in future work. First, adaptive explanation generation, where an explanation modality is dynamically selected based on the nature of the question posed by the user, the user's expertise level, and computational resource constraints, could help achieve a balance between trust and latency. Second, there is a clear need for criteria for evaluating explanations, which does not currently exist in the field. Third, the use of causality in explanation techniques to uncover not only associations but also causal relationships between retrieved information and the system output would be beneficial. Fourth, multimodal RAG explanation could prove valuable, as multimodal retrieval and generation applications become increasingly common.

In summary, Explainable RAGs are a key development in enabling large language model-based decision intelligence in mission-critical settings. By showing that transparency and efficacy do not need to be mutually exclusive, the xRAG methodology indicates that, with careful system architecture, both accurate, actionable explanations and high-performing systems are possible. With the growing adoption of RAGs in mission-critical decision-making, the need for explanation, verification, and dispute resolution will move from being an academic ideal to a practical requirement.

REFERENCES

1. J. Ren, Y. Xu, X. Wang, W. Li, W. Ma, and Y. Liu, "Effective and Transparent RAG: Adaptive-Reward Reinforcement Learning for Decision Traceability," arXiv preprint arXiv:2505.13258, 2025.
2. ScienceDirect, "From vectors to knowledge graphs: A comprehensive analysis of modern retrieval-

augmented generation architectures," Computer Science Review, vol. 61, 100925, Aug. 2026.

3. V. Bosco, "Toward explainability in retrieval-augmented generation: design and development of post-hoc framework," M.S. thesis, Politecnico di Milano, Milan, Italy, 2025.
   "RAG-Arena", GitHub repository, 2025.
4. "Retrivex: Explainability toolkit for retrieval models," GitHub repository, 2025.
5. Y. Zhu, N. Potyka, D. Hernández, Y. He, Z. Ding, B. Xiong, D. Zhou, E. Kharlamov, and S. Staab, "ArgRAG: Explainable Retrieval Augmented Generation using Quantitative Bipolar Argumentation," in Proc. 19th Conference on Neurosymbolic Learning and Reasoning (NeSy), Santa Cruz, CA, USA, Sep. 2025, pp. 1-22.
6. A. Habib, O. F. Abdulmahmod, M. Raza, Y. H. Gu, M. Aydoğan, and M. A. Al-Antari, "Towards Explainable AI in Agentic Retrieval-Augmented Generation: A Systematic Review," in Proc. 9th International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, Sep. 2025.
7. "KGRAG-Ex: Explainable Retrieval-Augmented Generation with Knowledge Graph-based Perturbations," arXiv preprint arXiv:2507.08443, Jul. 2025.
8. LangChain, "Evaluating RAG Architectures on Benchmark Tasks," LangChain Benchmarks Documentation, 2025.
9. P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459-9474, 2020.