A Survey of Explainability in Retrieval-Augmented Generation Systems

Eman Chaudhary, Mubasher Ahmed, Maha Baig, Muhammad Huzaifa, and Shahzada Muhammad Ali

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a dominant paradigm for enhancing large language models with factual grounding from external knowledge sources. While RAG systems reduce hallucinations and enable dynamic knowledge incorporation, they introduce significant interpretability challenges in both retrieval and generation stages. This survey systematically reviews the state-of-the-art in Explainable RAG, synthesizing 35+ recent works across three dimensions: retrieval-level transparency methods, generationlevel interpretability techniques, and evaluation frameworks. We provide a comprehensive taxonomy of explainability approaches, analyze their strengths and limitations, and identify critical research gaps. Our analysis reveals converging trends toward hybrid evaluation pipelines, causal probing methodologies, and domain-specific adaptations. We conclude by proposing six research directions essential for advancing RAG systems toward trustworthy deployment in high-stakes domains such as healthcare, law, and finance.

Index Terms

Retrieval-Augmented Generation, Explainable AI, Large Language Models, Source Attribution, Mechanistic Interpretability, Human-Centered Evaluation

I. Introduction

A. Motivation and Context

Retrieval-Augmented Generation (RAG) represents a fundamental architectural innovation in large language model (LLM) applications, combining the generative fluency of neural language models with factual grounding from external knowledge sources [1]. This two-stage pipeline, consisting of evidence retrieval followed by conditioned generation, addresses critical limitations of standalone LLMs: it reduces unsupported hallucinations, enables dynamic knowledge updates without costly retraining, and provides a mechanism for fact verification through source citation. However, this architectural advantage introduces unique interpretability challenges. The retrieval stage operates as an opaque selection mechanism, making it difficult to understand why specific documents were prioritized over alternatives. The generation stage further obscures how retrieved evidence is synthesized, weighted, and transformed into final outputs. In high-stakes domains such as medical diagnosis, legal reasoning, and financial analysis, where accountability and trust are paramount, this opacity creates a critical barrier to adoption [2], [3].

This work was supported by SMART Labs, Islamabad, Pakistan.

E. Chaudhary is with the School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan (e-mail: emanch.bese22seecs@seecs.edu.pk).

M. Ahmed is with the College of Electrical and Mechanical Engineering (CEME), National University of Sciences and Technology (NUST), Islamabad, Pakistan (e-mail: muahmed.mts44ceme@student.nust.edu.pk).

M. Baig is with the School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan (e-mail: mbaig.bscs22seecs@seecs.edu.pk).

M. Huzaifa is with the University of Engineering and Technology, Taxila, Pakistan (e-mail: 22-se-7@students.uettaxila.edu.pk).

S. M. Ali is with the National University of Sciences and Technology (NUST), Islamabad, Pakistan (e-mail: am.icon@nust.edu.pk).

B. Scope and Contributions

This survey provides a systematic synthesis of the emerging field of Explainable RAG, with the following contributions: 1) Comprehensive Taxonomy: We classify explainability methods into retrieval-level, generation-level, and domain-specific categories, providing a structured framework for understanding the landscape. 2) Comparative Analysis: We analyze the strengths, limitations, and trade-offs of different approaches, identifying convergent themes and design patterns. 3) Evaluation Framework Review: We survey quantitative metrics, human-centered methodologies, and causal evaluation protocols for assessing explanation quality. 4) Research Roadmap: We identify critical gaps and propose six future research directions for advancing the field. C. Organization The remainder of this paper is organized as follows. Section II describes our survey methodology. Section III establishes the background and motivation for explainability in RAG systems. Section IV presents our taxonomy of explainability methods. Section V provides comparative analysis. Section VI reviews evaluation frameworks and benchmarks. Section VII discusses emerging trends. Section VIII identifies open challenges. Section IX proposes future research directions. Section X concludes.

1) Subsubsection Heading Here: Subsubsection text here.

II. Survey Methodology

A. Literature Search Strategy

This survey adopts a systematic approach to identify and synthesize foundational and recent works in Explainable RAG. We focus on peer-reviewed publications, preprints, and technical reports that explicitly address interpretability, transparency, or explainability in RAG systems.

Inclusion Criteria:

- Methods and frameworks designed to make retrieval or
- generation processes interpretable
- Evaluation protocols and benchmarks for assessing explanation quality
- Domain-specific applications demonstrating operationalized explainability
- Publications from 2020 to 2025 (aligned with the emergence of RAG architectures)

Exclusion Criteria:

- General LLM interpretability methods not specific to RAG
- Pure information retrieval techniques without generation components
- Works focused solely on performance optimization without explainability considerations

B. Paper Search and Analysis

Our search identified 47 relevant papers, of which 35 met our inclusion criteria for detailed analysis. Papers were categorized into three primary dimensions: retrieval explainability (12 papers), generation explainability (15 papers), and evaluation methodologies (8 papers), with significant overlap in domain-specific applications.

III. Retrieval-Level Explainability

A. Motivation and Scope

Retrieval is the first gate in a Retrieval-Augmented Generation (RAG) pipeline: it selects which external evidence will condition the language model. Opaque retrieval makes it difficult to understand *why* particular documents were chosen, *how* specific sentences influenced the answer, and *which* sources should be credited or blamed for outcomes. Recent works address this by (i) auditing corpus-wide relevance patterns, (ii) attributing fair credit at the document level, (iii) identifying influential sentences or tokens, (iv) training retrievers with counterfactual and task-aligned signals to improve faithfulness and robustness, (v) structuring retrieved evidence into argumentative graphs for contestable decisions, and (vi) performing post-hoc forensics when knowledge bases are poisoned [1], [2], [3], [4], [5], [6], [7], [8], [9], [10].

TABLE I
TAXONOMY OF RETRIEVAL-LEVEL EXPLAINABILITY

Granularity	Approach	Mechanism and Contribution (Refs.)
Global (corpus)	Relevance Thesaurus	Distills term–term relevance from a neural teacher, enabling bias audits and transparent lexical re-scoring (BM25T/QLT) [1].
Token/mechanistic	ColBERT white-box	Late-interaction similarity maps expose term importance and exact vs. soft matches; analysis relates to rarity/IDF and fine-tuning effects [2].
Cross-stage (ctx→answer)	MIRAGE (CTI/CCI)	Identifies answer tokens sensitive to context (CTI) and traces responsible context tokens (CCI) for faithful source attribution [3].
Document-level	Shapley attribution	Utility-based fair credit for retrieved documents; exact and approximate Shapley methods studied [4].
Sentence-level	ARC-JSD	Black-box single-sentence ablations ranked by Jensen–Shannon divergence change to locate influential sentences [5].
Counterfactual training	Unsupervised DR	Shapley-guided counterfactual contrastive learning to emphasize true evidence and improve robustness [6].
Structured reasoning	ArgRAG	Quantitative bipolar argumentation over retrieved passages (support/attack) for contestable, robust explanations [7].
Task-aligned retrieval	OpenRAG	Retriever co-trained with the LLM via RAG labels/scores; aligns "relevance" with downstream utility [8].
Forensics	RAGOrigin	Post-hoc responsibility attribution to isolate poisoned knowledge contributing to errors [9].

B. Taxonomy of Retrieval-Level Methods

Table I summarizes representative approaches by granularity, mechanism, and contribution.

C. Global Explanations via a Relevance Thesaurus

A global *relevance thesaurus* is constructed by distilling term–term relations from a neural cross-encoder teacher and is then used to augment lexical scoring (e.g., BM25T/QLT) and to audit systematic model/data biases. Reported case studies surface brand tendencies, temporal skew, and tokenization artifacts while maintaining competitiveness and improving fidelity of lexical surrogates to neural scores [1].

D. White-Box Analysis of Late-Interaction Dense Retrieval

A white-box analysis of *ColBERT* investigates the mechanics of late interaction. Probes such as query-term masking and exact-vs-soft matching metrics show that rare terms receive stronger exact matches (consistent with IDF-like importance) and that fine-tuning can amplify this behavior; spectral analyses further differentiate stability for rare versus frequent terms. These token-level similarity maps provide inherently interpretable signals about what the retriever matched and why [2].

E. Mechanistic Cross-Stage Attribution (MIRAGE)

MIRAGE introduces two complementary components: (i) *Context Token Influence* (CTI) to detect which generated tokens are sensitive to retrieval context and (ii) *Context–Context Influence* (CCI) to trace responsible context tokens for those answer spans. This yields fine-grained, faithful source attributions and reports improvements over entailment-based or self-citation baselines across tasks and languages [3].

F. Document-Level Fair-Credit Attribution

For document-level responsibility, a utility function (e.g., teacher-forced likelihood) enables the computation of *Shapley values* that fairly assign marginal contribution to each retrieved item. Exact Shapley and several approximations (e.g., Kernel SHAP, sampling-based methods, and heuristic surrogates) are studied with correlation- and ranking-based evaluations, revealing effects such as redundancy, complementarity, and under-crediting of setup documents in multi-hop synthesis [4].

G. Sentence-Level Black-Box Attribution (ARC-JSD)

ARC-JSD identifies influential sentences by ablating one sentence at a time and computing the Jensen–Shannon divergence between output distributions, ranking sentences by impact. Analyses further localize attribution signals to specific model components, and the method reports improved accuracy and efficiency relative to alternative sentence-importance baselines [5].

H. Counterfactual Training for Dense Retrievers

An unsupervised dense retrieval framework uses Shapley-guided evidence identification to create counterfactual passages (partial/full/adversarial edits) and trains with contrastive objectives that increase sensitivity to true evidence and robustness to spurious cues. This approach integrates explainability into representation learning without supervision [6].

I. Structured, Contestable Explanations (ArgRAG)

ArgRAG converts retrieved passages into a quantitative bipolar argumentation graph whose nodes (claims/evidence) have strengths updated under support/attack semantics. The resulting graph constitutes an explicit, contestable explanation of which evidence supports or contradicts an answer and demonstrates robustness under noisy retrieval on targeted benchmarks [7].

J. Task-Aligned Retriever Learning (OpenRAG)

OpenRAG argues that IR relevance can diverge from RAG usefulness and co-trains the retriever with the generator via *RAG labels* and *RAG scores* in a staged pipeline (offline warm-up then online training), using semi-parametric in-context retrieval. Experiments show that aligning retrieval to generation utility improves downstream performance without tuning the LLM itself [8].

K. Forensic Responsibility under Poisoned Knowledge

When RAG outputs are corrupted by poisoned knowledge, *RAGOrigin* conducts post-hoc forensics: suspect filtering via chunked replays, a responsibility score that aggregates embedding similarity, semantic correlation, and generation influence, and clustering to isolate responsible sources. The method reports strong detection accuracy and robustness to adversarial strategies [9].

L. Evaluation Protocols and Datasets

Document-level attribution commonly reports rank correlations (e.g., Pearson/Kendall) and precision-style metrics for identifying impactful documents (including measures tied to utility drops) [4]. Global thesaurus methods report effectiveness (e.g., MRR/NDCG) alongside *fidelity* correlations to teacher/neural scores, with evaluations on MS MARCO and diverse BEIR tasks [1]. Sentence/token-level attribution and cross-stage methods are evaluated by attribution precision/recall or agreement with human annotations across QA-style datasets (e.g., ELI5, XOR-style tasks) and by efficiency analyses [3], [5]. Counterfactual and task-aligned retrieval methods report robustness/effectiveness under retrieval noise or task-specific utility [6], [8]. Structured argumentation and forensics evaluate explanation faithfulness and robustness to noisy/poisoned inputs on domain-specific datasets [7], [9].

M. Limitations and Open Challenges

Context-independence in global lexicons. Term–term relations distilled globally may inherit teacher/corpus priors and do not always capture per-query context, though they are valuable for scalable audits [1]. **Redundancy and multi-hop credit.** Utility-based attribution can under-credit early "setup" documents and exhibit position bias under redundancy; modeling complementarity/synergy remains challenging [4]. **Granularity and cost.** Sentence-level ablations and cross-stage mechanistic traces can be computationally intensive; unifying sentence- and token-level signals into consistent, user-facing rationales is an open problem [3], [5]. **Label quality and contestability.** Structured graphs require reliable support/attack assessments; mislabels can propagate but are visible and thus contestable [7]. **Task alignment and transfer.** Aligning retrievers to RAG utility may trade off with standard IR metrics and can be model- or domain-specific [8]. **Forensics triggers.** Post-hoc responsibility requires detected misbehavior and currently focuses on document/sentence blame assignment rather than complete causal chains [9].

N. Summary

Together, these methods move retrieval from a black box to an auditable, credit-assigning, and even training-time *explainable* component. A relevance thesaurus surfaces global biases and improves transparent lexical surrogates [1]; white-box late interaction exposes token-level matching behavior [2]; Shapley, ARC-JSD, and MIRAGE provide principled attributions across document, sentence, and token spans [4], [5], [3]; counterfactual and task-aligned training reshape retrievers toward faithful, robust utility [6], [8]; structured argumentation yields contestable explanations [7]; and forensics assigns responsibility under poisoning [9].

IV. Explainability in Generation

A. Introduction

Retrieval-Augmented Generation (RAG) reduces reliance of LLMs on the parameters on which it is trained and helps LLMs to extract information from external data on which it is not trained. Thus, improving LLM responses and reducing the need to train LLMs again and again as the data changes. However, this extraction of data and generation is essentially opaque, it tends to remain unclear how the model utilizes information that it has retrieved, which components affected resulting text, or the rationale behind utilizing some information pieces and ignoring others. Explainability in generation aims to provide a deeper understanding of this "Black Box" by providing comprehensive explanation for the generative phase of the RAG. It focuses on making understandable how retrieved evidence contributes to response generation, how and in what ways reasoning paths are generated in the model, and how self-correction or multi-agent interactions improve transparency. Generation-level explainability is particularly required in high-stakes domains like medicine, law, and scientific reasoning, where trust, accountability, and factuality are crucial.

Recent works have introduced various strategies to enhance explainability in RAG systems, including attribution-based visualization [11], [12], mechanistic interpretability [13], faithful reasoning [14], self-reflective explanation frameworks [15], [16], and multi-agent collaboration [17], [18]. Complementary visualization systems [19], [20] further support understanding by presenting interpretable representations of evidence flow and decision pathways.

B. Attribution-Based Explanations

An aspect of explainability, that is one of the earliest to be addressed when developing the process, is attribution which involves the identification of the actual documents or tokens which affect the various parts of the generated text. Attribution gives a direct connection between retrieval and generation such that users can trace the source of the information that exists in the model response.

VISA framework [11], shows visual source attribution by the mapping of each generated token with the source of evidence. It uses attention- and gradient-based alignment to generate fine-grained and token-level visualizations so that viewers can look into how retrieved documents influence particular parts of the generated output. VISA not only improves interpretability, but also serves as a diagnostic tool, giving the users a guide to the sources that are irrelevant or overrepresentative in the model reasoning process.

Likewise, Patel et al. [12] propose a multi-source attribution mechanism, which overcomes the problem of generating long-form answers and their attribution. As opposed to a short, factual answer, a long response is a mixture of several pieces of evidence, all of which have varying degrees of significance.Their framework employs hierarchical attribution layers that differentiate between sentence-level and paragraph-level contributions, thus facilitating a high-granularity comprehension of evidence synthesis.

Both VISA and multi-source attribution establish a foundation for transparency in generation by making the model use of evidence explicit and interpretable. These methods transform RAG systems from opaque text generators into explainable reasoning models where each generated segment can be traced back to its source.

C. Mechanistic and Faithful Reasoning

Attribution methods only focus on sources that are used to generate response. On the other hand, mechanistic reasoning is the process of building explanations by describing how and why models retrieved certain documents and use them in generations. The ReDeEP framework [13] pushes this area further by applying mechanistic interpretability to detect hallucinations. ReDeEP inspects internal attention heads and activation pathways to trace when and where the generation process diverges from retrieved content. This study by sun et al. also provides a method named AARF (Add Attention Reduce FFN) to eliminate these hallucinations by doing inference-time intervention. This is done by reweighting the outputs of those modules in the layer combination, not by changing model weights. This methodology gives more causal insights, rather than the correlations of input and output.

In parallel to mechanistic interpretability, Faithful Chain-of-Thought Reasoning proposed by Lyu et al. [14] aims to ensure that explanations faithfully represent the model reasoning process. Many of the existing explainability approaches produce post-hoc explanations that can be prone and factual but are not based on the internal decision-making of a model. Faithful Chain-of-Thought is the method which ensures consistency between the steps of intermediary reasoning and the final result, which means that the generated explanations actually reflect the thought process. The main concept of this approach consists in dividing the reasoning into two steps:

- **Translation Stage:** It involves the conversion of the problem into a “reasoning chain” that consists of two parts: Natural Language (NL); comments, subquestions, rationale and Symbolic Language (SL); code or deterministic formal steps, e.g., Python, Datalog, PDDL, etc., that can be executed.
- **Problem-Solving Stage:** involves executing the code from symbolic language in the solver to generate the answer.

This approach has great consequences on explainable RAG systems, whereby it is important to guarantee that the reasoning traces are accurate to the evidence retrieved to create trust. Collectively, this work focuses on causal explainability more than superficial interpretability, in which explanations reveal the underlying computational processes that are responsible for the behavior of the model.

D. Self-Reflective and Post-hoc Explainability

Another growing trend in RAG research is self-reflection, which enables the model to critique and explain its own outputs. The Self-RAG framework [15] extends the RAG by training a model to decide whether to use the retriever or not, to assess retrieved messages, and to critique its own outputs. It accomplishes this task by training the model to predict reflection tokens in addition to normal words. These reflection tokens are inserted in the text response based on how the LLM wants to classify that text. The model has extra tokens to represent meta-actions and judgments such as:

- **Retrieve:**Should external info be fetched?
- **IsRel:** Is this passage relevant?
- **IsSup:** Does the passage fully/partially support the claim?
- **IsUse:** How useful is the output?

It offers a reviewing process whereby the model repeatedly examines its produced response, evaluates any possible factual contradictions, and corrects it according to the evidence that it has retrieved. It does not only enhance factual grounding, but also generates natural-language explanations that can prove the model corrections. This self reflection cycle bridges the gap between correction and explanation hence making RAG systems interpretable and self improving.

Sudhi et al. [16] extend this concept, showing model-agnostic post-hoc explanations of RAG systems where explanations are produced by systematic analysis of the model retrieval and generation process. Their approach takes into account the contribution of input parts and documents retrieved to the final output through perturbation-based techniques. The post-hoc design guarantees applicability to a wide range of RAG architectures and enhances the transparency, trustworthiness and interpretability of model behavior.

These frameworks are a major step towards self-reflective and post-hoc explainability, where explanatory clues are systematically obtained through the model retrieval and generation dynamics. Explainability here is an external analytical process, rather than an inherent part of the generative architecture, which clarifies the reasoning processes and evidence basis of the model.

E. Multi-Agent and Interactive Explanation Frameworks

The complexity of generation in RAG systems often arises from the multi-step reasoning and evidence synthesis processes. For this, recent work proposes multi-agent architectures where various agents are responsible for specializing in the retrievals, reasoning, verification, and explanation.

The study by Besrour et.al [18] extends the MAIN-RAG [17] framework by building upon its multi-agent filtering architecture to enhance faithfulness, attribution, and completeness in RAG-based question answering. While MAIN-RAG introduced the concept of decomposing the reasoning process into specialized agents for interpretability, RAGentA refines and expands this approach through a four-agent system i.e. Predictor, Judge, Final-Predictor, and Reviser, each powered by the same base LLM (Falcon-3-10B) and coordinated via prompt-based interactions. Specifically, it inherits MAIN-RAG's idea of agent-based document filtering and reasoning transparency but introduces two novel components: an inline citation mechanism to ensure fine-grained source attribution and a Reviser agent that performs iterative verification to address incomplete or partially grounded answers.

In addition to these architectures, there is RAGVIZ [19] and RAGE Against the Machine [20], which provide visualization tools allowing users to interactively visualize model decisions. RAGVIZ provides token and document level attention visualization and document toggling features that enable the user to view the impact of document inclusion or exclusion on the output. RAGE goes further to suggest a single framework that can be used in generating and assessing explanations, thereby assisting researchers in determining weaknesses and strengths in the retrieval and generation stages.

Visualization-based and multi-agent explainability systems are holistic systems in which all reasoning, flow of evidence and generation outputs are presented to users in interpretable forms.

F. Comparative Insights

The literature review demonstrates a complimentary scene of techniques aimed to enhance the explainability in the generation process. Attribution-based models [11], [12] are concentrated on input-output alignment, and thus, the impact of the source can be observed. Causal interpretability is further explored by mechanistic interpretability and faithful reasoning [13], [14] explaining why certain outputs arise. End-to-end and self-reflective methods [15], [16] include meta-cognition as a generative loop element, enabling models to be self-explanatory and self-correctorative. Lastly, [17], [18], [19], [20] are multi-agent and

visualization tools that make the model easier to understand, as it allows the developer and end-users to interact, inspect and understand the model internal logic. This development marks a shift in the stagnant post-hoc explanations to dynamic, cross-tiered and multi-layered interpretability models. It is likely that the future of explainable RAG will be in hybrid systems that integrate these ways, such as, self-reflective agents who have visual attribution technologies and mechanistic tracing abilities.

G. Challenges

Nonetheless, even now there are a number of challenges:

- **Faithfulness vs. Plausibility:** There are a lot of these explanations which are linguistically plausible and do not reflect reasoning in actual sense. Ensuring causal faithfulness has been a subject of open research.
- **Scalability:**Attribution and mechanistic analysis methods are often computationally expensive, reducing their applicability on a large-scale deployment in real-time.
- **Human-Centric Evaluation:** While visualization tools help in interpretability, standardized metrics for evaluating explanation quality are still not developed completely yet.
- **Cross-Model Generalization:** Most explainability techniques are architecture-specific, which causes concerns regarding generalization between various backbones of LLM.

Symbolic reasoning, graph-based interpretability, and multi-modal explanations may be used in the future to complement work to bring greater and more generalizable data on the behavior of generations.

V. Evaluation of Explanations in Retrieval-Augmented Generation Systems

RAG is particularly difficult to judge in terms of explanations due to its dual nature: (1) it needs to retrieve external documents that are relevant to the system and (2) elaborate corpus-driven coherent responses. Apparently, explainability in RAG needs to be characterized in numerous dimensions: faithfulness, grounding, coherence, and human trustworthiness, and both quantitative and qualitative measures need to be used. [21].

The field has now developed on top of conventional NLG evaluation (e.g. BLEU, ROUGE) to more specific explainability tests, including FLEEK [22] and RAGAS [23] and FaithfulRAG and TRUE Benchmark and TRUST-SCORE. These schemes present subtle quantitative indices of faithfulness, human perception research of utility of explanation, and hybrid scoring techniques that are sensitive to the multi-modal interdependences amongst retrieved and generated material. They constitute the principles of Explainable RAG Evaluation (X-RAGEval), an expanding scientific trend at the interface between interpretability, retrieval theory, and human-AI interaction.

A. Quantitative Faithfulness Metrics

Systematic efforts to measure explanation faithfulness in RAG had the earliest efforts defining the problem as a textual entailment between generated responses and the evidence to support them. Factual consistency was made a key quantitative objective with the FEVER benchmark [24] launched in 2018 and its successors. Such early measures were however more output based checks and were not sensitive to retrieval based reasoning. Recent literature therefore reconceptualises faithfulness as correspondence between the retrieval and generation causal mechanisms, which is operationalised in terms of fined-grained measures of attribution and grounding.

1) FaithfulRAG (2025): Evidence-Conditioned Attribution Scoring: The framework proposed by FaithfulRAG [25] forward proposes two constituents of faithfulness evaluation (a) evidence attribution: this evaluates the count of the output tokens that can be entirely attributed to the retrieved passages; and (b) causal grounding: this calculates whether or not the removal of certain retrieved passages produces a visible difference in the generated answer.Based on counterfactual masking and saliency-based attributions, the authors have now provided a way to quantitatively measure causal links between retrieved and generated

content. They find empirical proof that standard similarity measures like cosine or BLEU give much higher numbers than they should for grounding, arguing for the inclusion of causal faithfulness metrics in evaluating RAG.

2) RAGAS (2024): Reference-Free Quantitative Evaluation: RAGAS [23] provides a reference-free metric pipeline which evaluates RAG responses along four facets: faithfulness, relevance, context precision, and answer correctness. With a key focus on reference-free grading that uses LLM-based prompts for scoring, which greatly eases adoption in real-world systems where human gold references are not always available. Unlike traditional supervised metrics, RAGAS leverages the retrieval corpus itself as implicit ground truth, combining embedding similarity with factual entailment detection. On five QA benchmarks, RAGAS shows high agreement with human ratings ($\rho = 0.82$), showing how quantitative evaluators based on LLMs can emulate human evaluative reasoning all the while retaining reproducibility.

3) TRUST-SCORE (2024): Quantitative Calibration of Trustworthiness: The TRUST-SCORE metric [26] also provides some trustworthiness calibration along with added factuality. It calculates your confidence score that is kind of a weighted average of internal consistency (how much the generated rationales overlap in meaning), evidence sufficiency (document support for each claim), and uncertainty quantification. Bringing probabilistic calibration into the evaluation loop, TRUST-SCORE connects quantitative sturdy and qualitative explanation as believed by human beings. Empirical results reveal that as TRUST-SCORE values increase, so do human trust scores, relating quantitative faithfulness to qualitative interpretability.

4) GFM-RAG (2024): Grounded Factuality Metric: GFM-RAG [27] breaks down factuality into two categories: generation grounding (factual entailment) and retrieval alignment (document correctness). GFM-RAG explicitly models multiple documents retrieval through aligning claim–evidence pairs via embedding-based clustering, in contrast to FEVER-style factual checks. The need for comprehensive evaluation metrics that capture both retrieval and generation stages is highlighted by quantitative evaluations, which reveal that 38% of factual errors in traditional RAG systems result from insufficient incorporation of retrieved evidence rather than incorrect retrieval.

B. Human-Centered Evaluation of Explanations

Despite being crucial, quantitative faithfulness metrics only provide a partial picture of explainability. In the end, true interpretability in RAG systems depends on human comprehension and trust, whether or not users are able to understand how retrieved evidence affects generated outputs, and whether or not such explanations assist them to arrive at more informed choices. Thus, by evaluating the perceived quality, cognitive load, and utility of explanations [28], human-centered evaluation enhances algorithmic metrics.

1) GaRAGe (2024): Human-Centric Evaluation Framework: A structured methodology for human-centered evaluation of RAG explanations was first proposed by the GaRAGe framework [29] (Grounded and Reliable Assessment of Generated Explanations). In order to evaluate trust calibration—the process by which users modify their level of confidence in a model's outputs in response to different explanations—the authors carried out controlled user studies. According to their findings, readability and justification clarity have a greater impact on user confidence than high factual faithfulness.

GaRAGe uses a three-tier evaluation hierarchy to formalize this relationship:

1) **Cognitive load:** he amount of mental effort required to comprehend an explanation,
2) **Perceived grounding:** if users believe the explanation is consistent with the evidence, and
3) **Decision usefulness:** if explanations help finish tasks.

These factors establish human trust as a separate explainability axis from factual accuracy.

2) UAEval4RAG (2024): User-Adapted Evaluation: By adding user-adapted evaluation, the UAEval4RAG benchmark [30] expands upon GaRAGe. This study measures the impact of explanation granularity on understanding by gathering user ratings across expertise levels (expert vs. novice). In order to measure the difference between user trust and system reliability, the authors present a *trust divergence*

metric. This makes it possible to systematically identify "over-trust" phenomena, which occur when people have unjustified faith in explanations that appear logical but are actually dishonest.

Importantly, UAEval4RAG combines quantitative scoring (e.g., Likert-scale trust and comprehension metrics) with qualitative interviews. The findings show that the ideal explanation length correlates inversely with user expertise, with novices benefiting from detailed, pedagogical justifications while experts prefer succinct evidence summaries. These results emphasize the necessity of audience-aware assessment procedures in subsequent explainability studies.

3) BINDER & FLEEK (2024): Benchmarking Explanation Quality: The purpose of the BINDER and FLEEK benchmarks [22], [31] is to assess the coherence and relevance of explanations in retrieval-augmented QA systems. For every question, FLEEK, in particular, defines multiple reference annotations that represent logical reasoning chains based on retrieved evidence. This makes it possible to compare generated explanations not only to surface forms but also to the variety of legitimate lines of reasoning.

In order to create a single *FLEEK score*, FLEEK introduces metrics for faithfulness (F), linkage (L), and evidence explanation (E). The benchmark offers human-validated annotations to evaluate the completeness, fluency, and grounding of explanations. Explanation-aware metrics show a much stronger correlation with human judgments ($r \approx 0.87$), while purely similarity-based evaluation (e.g., BLEU) fails to capture important aspects like reasoning coverage or evidence overlap, according to studies using FLEEK.

C. LLM-as-Judge and Reference-Free Evaluation Paradigms

A new paradigm for explanation assessment, *LLM-as-Judge*, has emerged with the growth of large language models as adaptable evaluators. Evaluators now use LLMs to reason about reasoning rather than fixed reference comparisons, rating explanations based on attributes like informativeness, logical coherence, and grounding. With this change, static benchmarking gives way to dynamic, context-aware assessment.

1) TRUE Benchmark (2023–2024): One of the first extensive attempts to methodically assess explanation trustworthiness in RAG systems is the TRUE Benchmark [32]. Annotated for groundedness, justification quality, and human trust alignment, TRUE consists of a varied collection of question-answer pairs from both open-domain and closed-domain contexts. Comparative criteria, such as "Is explanation A more faithful to retrieved evidence than explanation B?" are used to prompt LLM-based judges during evaluation.

The hybrid evaluation loop, which combines automated grading with human calibration to guarantee interpretive reliability, is the central idea of TRUE. Metrics such as *Explain-Score* (composite of faithfulness and completeness) and *Trust-Align* (agreement between LLM and human trust ratings) are introduced. According to empirical analyses, GPT-4-based evaluators outperform conventional rule-based metrics, achieving over 80% agreement with expert human ratings. This shows that employing LLMs as scalable judges while maintaining human-like interpretive depth is feasible.

2) RAGBench and TRACe (2024): Unified Evaluation for Grounding and Coherence: A unified framework for assessing retrieval-grounded generation using LLM-based comparative scoring is proposed by RAGBench and its extension TRACe [33]. In contrast to earlier datasets, RAGBench explicitly breaks down the evaluation process into:

1) **Retrieval Relevance:** Are the retrieved documents contextually aligned with the query?
2) **Grounded Generation:** Does the model use evidence faithfully?
3) **Explanation Coherence:** Is the reasoning chain logically consistent?

In order to refine evaluators and produce a high-fidelity *LLM-as-Judge* model tailored for RAG explanations, TRACe supplements this with pairwise human preference data. By bridging quantitative accuracy with human alignment, these benchmarks provide a useful route toward reference-free but comprehensible evaluation.

3) RAG-Ex (2024): Explainability via Self-Rationalizing Evaluation: By making the assessment process itself explicable, RAG-Ex [34] advances the *LLM-as-Judge* paradigm. The evaluator LLM produces a justification explaining why an explanation is or isn't faithful rather than scalar scores. The auditability of evaluation results is made possible by this meta-rationalization, which transforms opaque grading into discernible evaluation artifacts.

The RAG-Ex framework comprises two stages:

1) **Rationale Generation:** producing textual justifications for grading decisions, and
2) **Faithfulness Assessment:** quantifying alignment between rationale and evidence.

RAG-Ex advances the idea of transparent evaluation systems by incorporating meta-explanations. These evaluators increase user trust in the assessment pipeline itself, which is a growing concern for ethical artificial intelligence research, according to empirical comparisons.

D. Causal and Interpretability-Based Evaluation Frameworks

A parallel line of research explores causal and internal model interpretability, looking at the information flow from documents retrieved to generation tokens, going beyond output-level scoring. This viewpoint views explanation evaluation as a diagnostic procedure that reveals the inner workings of RAG systems rather than as an end-task assessment.

1) KG-SMILE (2025): Causal Graph Evaluation of RAG Explanations: A causal-graph method of explanation evaluation is presented by the KG-SMILE framework [35] (Knowledge-Graph-based Structural Metrics for Interpretable Language Explanations). It builds a bipartite graph, weighed by mutual data and attention attributions, connecting generated token nodes and retrieved evidence nodes. The graphical connectivity degree between the claims and their supporting sources is then used to calculate faithfulness.

The extent to which output semantics can be recovered from the retrieved knowledge graph is measured quantitatively by KG-SMILE. A significant alignment among causal grounding scores and human assessments of explanation quality ($\rho = 0.84$) is shown in experiments across QA datasets, suggesting that structural interpretability metrics can reliably approximate human perceptions of faithfulness.

2) Integration with Counterfactual Evaluation: Counterfactual testing is the methodological foundation of causal interpretability frameworks such as KG-SMILE and FaithfulRAG, which systematically perturb retrieved evidence to observe effects that follow. The generated output changes when specific evidence passages are hidden, reflecting the causal importance of that evidence. This is operationalized by FaithfulRAG using token-level masking, and it is generalized by KG-SMILE using knowledge-graph traversal. Both approaches reveal *why* specific evidence influences model reasoning, going beyond simple similarity.

3) The Shift Toward Diagnostic Evaluation: By diagnosing how explanations emerge rather than evaluating explanations after generation, these causal evaluation frameworks signify a paradigm shift. They allow researchers to measure the inner coherence of multi-hop reasoning, detect information bottlenecks, and track which retrieval nodes have the biggest effect on the model's outputs. Crucially, they establish a link between the domains of interpretability and information retrieval, demonstrating the interdependence of retrieval fidelity and explanation quality.

E. Comparative Synthesis Across Evaluation Dimensions

A cross-framework synthesis reveals three dominant evaluation paradigms:

Faithfulness-Centric Evaluation represented by FaithfulRAG [25], RAGAS [23], GFM-RAG [27], and TRUST-SCORE [26].The emphasis of these works is on quantitative consistency between generated responses and retrieved evidence. They introduce repeatable numerical metrics that have a strong correlation with human judgment, formalizing faithfulness as causal or semantic alignment. They are still not very good at capturing reader comprehension or explanatory utility, though.

Human-Centric Evaluation exemplified by GaRAGe [29], UAEval4RAG [30], BINDER, and FLEEK [22], [31].These frameworks move from factual accuracy to interpretability as perceived by the user. They

show that narrative clarity is just as important to user satisfaction as factual accuracy by operationalizing trust, cognitive load, and perceived grounding as quantifiable quantities. However, these assessments are costly, domain-specific, and challenging to replicate on a large scale.

Automated and Causal Evaluation reflected in TRUE Benchmark [32], RAGBench/TRACe [33], RAG-Ex [34], and KG-SMILE [35].These paradigms simultaneously aim for transparency and scalability. While causal and graph-based approaches examine internal mechanisms, LLM-as-Judge models provide adaptable, reference-free assessment pipelines. However, there are still unanswered questions regarding the interpretive validity of causal metrics and the dependability of LLM-based evaluators, necessitating thorough cross-validation.

F. Methodological Limitations

Current approaches share a number of limitations despite methodological sophistication:

Over-reliance on static datasets. While benchmarks like FLEEK, TRUE, and RAGBench offer useful structure, they run the risk of becoming quickly outdated as retrieval corpora and LLM frameworks change. Evaluation frameworks must dynamically adjust to model updates and domain shifts because explanations are intrinsically context-sensitive.

Basis in Circular evaluation. LLM-as-Judge systems, while scalable, may replicate the same reasoning biases as the models under evaluation [32], [34]. Without independent calibration or interpretive diversity, evaluations risk reinforcing model-specific heuristics rather than measuring genuine explanatory quality.

Limited cross-cultural and cross-domain generalization. Human-centered studies often recruit homogeneous participant groups [29], [30], leaving unexplored how explanation comprehension varies across linguistic, cultural, or professional contexts. This limits the ecological validity of trust and cognitive-load assessments.

Faithfulness–usefulness trade-off. Several studies (notably GaRAGe and UAEval4RAG) reveal that the most faithful explanations are not necessarily the most helpful to end-users. This uncovers a persistent tension between epistemic transparency and cognitive accessibility, a problem that quantitative metrics alone cannot resolve.

Evaluation of retrieval transparency. Most frameworks emphasize explanation of generation rather than retrieval. Few metrics assess why certain documents were selected or how retrieval uncertainty propagates into generation. Retrieval interpretability must be treated as a top-tier evaluation target in future explainability studies.

G. Toward Integrated Evaluation Pipelines

Integrated pipelines that incorporate complementary evaluation techniques represent a promising avenue:

Hybrid Scoring Models. Researchers can simultaneously estimate human assessment across multiple axes by combining user feedback (GaRAGe/UAEval4RAG), trust calibration metrics (TRUST-SCORE), and LLM-based faithfulness scoring (RAGAS). Both computational accuracy and perceptual validity would be provided by such ensemble evaluation architectures.

Explainable Evaluators. By producing meta-explanations, systems such as RAG-Ex [34] show that evaluators one another can be made interpretable. Expanding this idea could result in evaluation agents that are completely transparent and able to defend their standards and conclusions.

Causal-Graph Diagnostics. Diagnostic interpretability is provided by frameworks like KG-SMILE [35], which trace the causal influence of retrieval nodes on generation tokens. By incorporating causal graphs into evaluation pipelines, evaluators can identify latent reasoning pathways and differentiate between spurious correlation and true grounding.

Continual Human Calibration. As used in TRUE [32], periodic alignment of automated scores with human evaluations is still crucial. Adaptive feedback loops will guarantee that evolving assessors maintain interpretive faithfulness to human expectations.

These hybrid approaches suggest that explainability evaluation protocols will eventually be standardized, much like BLEU did for machine translation but with human, causal, and trust-aware aspects added.

H. Open Research Challenges

The new frontier of transparent RAG Evaluation is defined by a number of unanswered questions, despite the reviewed literature having laid a significant foundation:

Benchmark Generalization and Transferability. Is it possible for benchmarks that have been built on one retrieval domain (like Wikipedia QA) to be applied to other domains (like scientific reasoning or legal analysis)? Semantic diversity in current datasets is insufficient to ensure such robustness.

Measuring Cognitive Impact. It is still challenging to quantify cognitive load and trust calibration. Response time and self-reported trust are examples of current proxies that capture superficial trends but not the underlying cognitive mechanisms. More grounded evaluations of human interpretability may result from integration with behavioral or neuro-cognitive analytics.

Meta-Evaluation of Evaluators. How should the evaluators themselves be assessed? The community needs meta-benchmarks that assess evaluator consistency, reliability, and transparency across tasks as LLM-based judges and causal analyzers become more common.

Dynamic Explanations in Interactive RAG. Post-generation static explanations are assumed in current research. Interactive or adaptive explanations that change during conversation are increasingly produced by real-world systems. Assessing such dynamic interpretability is still a methodological challenge.

Ethical and Epistemic Dimensions. Evaluation of explanations touches on more general issues of misinformation, fairness, and epistemic responsibility. If the retrieved evidence contains bias or exclusion, a high ”faithfulness score” is not enough. As a result, evaluators must include ethical foundation in their assessment of explanatory quality.

VI. Domain wise Explainability

Retrieval-Augmented Generation (RAG) has emerged as a principled approach to combine the broad linguistic capabilities of large language models (LLMs) with explicit, task-relevant knowledge retrieved from external sources at inference time. In a canonical RAG pipeline a user query is embedded and matched to a corpus of documents or structured records; the top-ranked evidence is then concatenated, or otherwise fused, with the original prompt and supplied to a generative model. This architectural separation , retrieval for grounding and generation for fluent output , reduces unsupported hallucinations and enables models to react to newly published information without costly full-model retraining. However, the two-stage pipeline also introduces new interpretability challenges: while retrieval exposes explicit evidence, the downstream generator’s integration of that evidence remains opaque unless design choices are made to surface reasoning and provenance.

Explainable RAG refers to a family of methods that seek to make both retrieval and generation decisions interpretable. Explainability in this context encompasses multiple desiderata: provenance (which documents or data items were used), attribution (which parts of the retrieved evidence influenced specific claims), causal salience (how much each piece of evidence affected the final output), and human-comprehensible rationales (textual or visual explanations that a domain expert can inspect). Different domains impose different emphases: clinicians require traceable citations and clear distinctions for differential diagnoses; legal professionals need precise precedent citation and argument structure; financial analysts need verifiable numeric evidence alongside narrative interpretation; and multimedia forensics demand both visual localization and textual justification.

This survey organizes recent explainable RAG work by application domain. For each domain we present representative systems that illustrate how retrieval, structured knowledge, multi-modal indexing, and interpretable reasoning have been designed to meet domain constraints. Each paper is described under its own subsection heading, with a concise discussion of problem motivation, method, contributions, and the relationship to preceding work. The sections are deliberately narrative: where multiple papers exist in a domain, the descriptions highlight continuity , showing how later approaches extend or address limitations of earlier designs. The goal is to produce a usable, professional reference suitable for researchers, practitioners, or students preparing a review or project proposal.

A. Medical / Healthcare Domain

1) KGRAG-Ex: Explainable Retrieval-Augmented Generation with Knowledge Graph-based Perturbations: KGRAG-Ex positions the explainability at the center of the RAG by leveraging the structural clarity of knowledge graphs (KGs). Work in this area indicates that retrieval using text alone does not provide the context that individuals in medical settings require. These individuals need to understand how findings relate to multiple possible diagnoses and how intermediate results from tests alter the likelihood of different diagnoses. The method addresses this issue by creating a focused structure for each query. Entities and relations that the system extracts from the data sources form a structure that captures concepts relevant to the question. The structure appears in two forms that the generation component uses. The first form presents important paths as descriptions in natural language. The second form shows relation paths that can be modified for examination.

The explanation contribution that the approach provides comes from the method that uses modification. The system does not offer only rationales that appear after the response. It performs systematic changes on the structure that shows relations. These changes include removing entities, breaking connections between entities, or altering the strength of paths. The system measures how each change affects the response that the generation component produces. This testing that examines alternatives creates a ranked representation of influences that show causation. The representation indicates which entities or relations produce substantial changes in the conclusion that the large model generates when the system modifies them. For individuals who work in clinical contexts, these ranked attributions that show which factors matter provide significant value. The attributions direct individuals who conduct audits to the specific facts or chains of relations that determined a suggested treatment or diagnosis that the system provided.

Beyond its immediate practical utility, KGRAG-Ex serves as a conceptual bridge between symbolic and neural reasoning. By using graph structures to probe a neural generator, the method surfaces mechanistic information about the generator's sensitivity to structured knowledge , a level of introspection that raw text perturbations cannot easily provide. While the approach increases pipeline complexity and requires careful KG extraction, it exemplifies a class of explainable RAG methods that privilege causal probing and structured evidence surfacing over purely surface-level text attribution.

2) ClinicalRAG: Enhancing Clinical Decision Support through Heterogeneous Knowledge Retrieval: ClinicalRAG addresses a main issue in support for decisions: data from individuals receiving treatment show different forms. A case that occurs in practice combines fields with structure (measures of vital signs, results from laboratory work, codes for diagnosis), sections with partial structure (lists describing problems), and narratives without structure that contain high density (notes on progress, summaries at discharge). The design that ClinicalRAG uses focuses on retrieval that reflects modality and makes provenance clear. The approach takes different types from medical data and provides a shared interface for retrieval, and it also applies ranking to evidence using semantic relevance and reliability in the context of treatment. Guidelines and literature that received review from other researchers receive priority over notes from a single case.

For making explanations clear, ClinicalRAG produces justifications that indicate provenance. Each assertion that the system develops includes a token showing evidence, and this token indicates whether the supporting fact came from a term in a structured ontology, a paragraph from a guideline, or an excerpt from a matched case in treatment. This coupling allows an individual conducting treatment to follow each claim to a source that allows verification. ClinicalRAG also uses prompts with contrast that are simple, and these prompts present evidence that is positive and evidence that is negative in a side-by-side manner. The output can then indicate in a clear way why a particular diagnosis receives favor over alternatives that appear possible.

In the context of treatment, the benefit shows two main aspects. The retrieval using a combination of approaches reduces errors that appear as false content because the system that develops output faces constraints from sources showing authority and from thresholds with clear structure (for instance, cutoffs for laboratory values). The annotations indicating provenance also reduce the load on cognition for individuals conducting review. Rather than reading a narrative with high length, an individual conducting

treatment can examine the tokens showing attached evidence in a quick manner and reach a decision about whether to follow the recommendation that the system provides. ClinicalRAG uses concepts from earlier work on RAG systems that provide explanations, and it demonstrates how considerations that relate to operations (the flow of work, the level of confidence in evidence) affect which features for explanation show the most value in practice.

3) MedRAG: Enhancing RAG with Knowledge Graph-Elicited Reasoning for a Healthcare Copilot: MedRAG extends the medical RAG toolbox by placing diagnostic reasoning at the core of retrieval design. Prior medical conversational agents either rely on raw-text retrieval or on a monolithic knowledge graph, whereas MedRAG constructs a task-relevant hierarchical diagnostic graph that captures symptom clusters, intermediate differential distinctions, and disease-level outcomes. The hierarchy allows the system to represent not just associations but diagnostic contrasts : those features that most set apart two competing hypotheses. Practically, this helps to address a common clinical failure mode of confusing superficially similar diseases.

At runtime MedRAG performs two complementary retrievals: it searches for similar historical patient cases from an indexed EHR corpus and it retrieves diagnostic contrast vectors from the hierarchical graph. The generator receives both streams as fused context: exemplar cases give it precedent for what has been said clinically while the graph-derived contrasts illuminate discriminative features (e.g., presence of a pathognomonic sign, an abnormal lab trajectory), and so the model can generate advice that is both precedent-informed and contrast-aware, including follow-up questions to resolve diagnosis.

The power of MedRAG comes from making comparisons clear. Instead of giving a single-line answer, the system provides clinicians with a concise 'decision card' showing the most likely diagnoses, highlighting the key differences between them, and including excerpts from past cases. This enables quicker clinical decisions and provides natural clinician-AI collaboration opportunities. A physician can decide to accept the recommendation, ask for more details about a specific comparison, or use the case examples to support her different conclusion. By providing real case examples along with a reasoning framework, MedRAG shows how explainability can be a practical clinical tool, not just a technical solution.

4) BIORAG: A RAG-LLM Framework for Biological Question Reasoning: BIORAG solves the challenges of scale and complexity in biological literature. In the life sciences, there's a highly stratified space of knowledge to navigate, spanning high-level concepts, experimental procedures, quickly-changing experiments, and highly-specific terminology. BIORAG's technique is founded on two innovations: a specialized embedding model trained to capture biochemical and experimental context, and a hierarchical retrieval technique that organizes information into conceptual categories (background, methodology, results, implications).

When a researcher asks a question, BIORAG decomposes the question into components that line up with these categories and finds evidence that's appropriate for each. For instance, an answer about a biological mechanism may require data from laboratory experiments, while a question about clinical applications may require information from reviews. By showing both the category of evidence and the exact source material used to answer each part, BIORAG provides a transparent audit trail, showing not just which papers were used, but why each was likely chosen. This transparency is useful for two audiences: it allows scientists to quickly discern whether an answer is relying on data from direct experiments or more general analysis, helping them gauge how new or reliable a piece of information might be. This decreases the likelihood that scientists will be tempted to draw wide-ranging conclusions based on narrow studies.

In the broader RAG space, BIORAG shows how specialized embeddings and organized retrieval can lead to better accuracy and more verifiable explanations in a scientific use case.

B. Legal Domain

1) Athena: Retrieval-Augmented Legal Judgment Prediction with LLMs: Athena focuses on the established legal challenge of judgment prediction: given the details of a case, predicting potential charges, applicable laws, or likely outcomes. Legal applications have particularly strict requirements for transparency

- users need direct citations to statutes or previous cases and a clear line of reasoning that follows legal logic. Athena builds a curated collection of legal accusations and precedents, then uses semantic search to identify the most relevant legal references for a given case. These retrieved items guide the language model to produce structured, templated responses including element checklists, reasoning summaries, and concise justifications that reference supporting precedents.

A key innovation of Athena is its focus on structured outputs that match how legal professionals actually work. Instead of open-ended text, Athena delivers information in organized sections (such as 'Relevant Statute:', 'Supporting Precedent:', 'Reasoning:') that legal experts can quickly review. This structured approach both minimizes the risk of invented legal references and simplifies follow-up tasks like compliance verification. In testing, Athena achieved higher prediction accuracy while simultaneously making its decision process more transparent for legal review.

Athena advances retrieval-grounded methods by transforming legal precedents into reasoning frameworks for the language model. While previous RAG systems prioritized retrieval quality or response fluency, Athena emphasizes alignment with legal thinking patterns. Its design reflects the reality that legal transparency requires not just showing sources, but presenting arguments in a format that lawyers and judges find familiar and can question when needed.

2) LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain: LegalBench-RAG achieves a methodological innovation by offering a dedicated standard for evaluating legal RAG systems. The developers found that testing only final answers makes it hard to figure out which part of the pipeline (retrieval or generation) is broken. To this end, LegalBench-RAG marks the exact text segments that comprise proper evidence for each query. This reliance on precise text selection renders a system successful not on which general documents are returned relevant, but on which exact paragraphs or legal clauses providing support for a claim are identified.

This granularity has strong implications for transparency. Because systems trained with LegalBench-RAG must return concise, directly citable evidence rather than large documents that must be manually searched, the standard ties technical measurement to practical legal needs - since attorneys require exact citations, the ability to automatically provide them represents a major step towards reliable automation.

Additionally, LegalBench-RAG promotes retrieval methods that are a mix of semantic embeddings and syntax-aware and citation-based methods. By rewarding retrieval accuracy, this standard pushes research towards RAG systems whose explanations are verifiable legal sources rather than merely convincing stories, a major mitigation of the danger of legal statements without evidence.

3) CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in Legal QA: CBR-RAG resurrects an old method, Case-Based Reasoning (used in many domains but dormant in law), and applies it to modern RAG systems. Legal thinking often reason by analogy: lawyers and judges consume past cases to make sense of new situations. CBR-RAG considers past cases as part of its retrieval process, going beyond typical statute excerpts to find one or more past cases that are the closest matches, serving as concrete examples for the language model.

The benefit of explainability is clear and compelling. If a model's answer is supported by one or more past cases, a human reviewer can evaluate the quality of the analogical match: are the fact patterns actually analogous, or is the retrieval exploiting surface similarity? By making the retrieved cases and the similarity measures that yielded them transparent, CBR-RAG provides a clear chain from precedent to prediction.

CBR-RAG also spans bridges evaluation. Practitioners can audit performance not only via final accuracy metrics, but via the plausibility of case matches. This allows practitioners to run RAG systems with an extra layer of protection: if the retrieved precedents are poor analogues, the system can return low confidence and require human review. This is often a prerequisite in settings where the cost of error is high.

C. Financial Domain

1) MultiFinRAG: An Optimized Multimodal Retrieval-Augmented Generation (RAG) Framework for Financial Question Answering: Financial documents are inherently multimodal: quarterly filings, analyst

slide decks, and investor presentations combine narrative discussion with tabular financial statements, charts, and footnotes. MultiFinRAG designs a retrieval pipeline that is modality-aware: tables and figures are parsed into structured representations and compact textual summaries, while narrative text is retained as full passages. These heterogeneous artifacts are then indexed with modality-sensitive embeddings so that retrieval can select the most informative mix of text and structured data for a given query.

For explaining its answers, MultiFinRAG shows its work. It pulls up the specific data rows, chart summaries, or fine-print footnotes that support its claims. This is crucial for financial analysts, who need to see the hard numbers behind any statement. For example, if the system claims ”revenue increased due to stronger product sales,” it will point directly to the relevant revenue line items and the exact chart or footnote that backs this up. By directly linking its statements to structured numerical evidence, MultiFinRAG cuts down on incorrect number claims and makes it easy for a person to check the facts.

In practice, MultiFinRAG is also designed to be efficient. It uses a tiered approach, first trying to find an answer using plain text reports. It only digs into the more complex tables or charts if the question specifically requires it. This practical method avoids overcomplicating things when it's not needed, while still providing full transparency for number-heavy questions. In short, MultiFinRAG proves that for a financial Q&A system to be trustworthy, it must be able to pull from different types of data and always show the numbers behind its story.

D. Scientific / Biological Domain

1) Hypercube-Based Retrieval-Augmented Generation for Scientific Question-Answering: Scientific questions are complex and require precise answers about various elements. Hypercube-RAG is a system designed to handle this. It starts by cataloging its knowledge along several distinct lines of meaning, creating a detailed, multi-faceted index. When you ask it a question, it first identifies the key parts of your query and then looks for answers specifically within the sections of its index that are most relevant to each of those parts.

This architectural choice offers immediate explainability benefits. When a document is retrieved from a particular subspace (for example, the 'methods' axis), the system can indicate which semantic dimension motivated the selection. Such dimension-aware retrieval makes it easier for a domain expert to judge the relevance of retrieved papers and to understand whether an answer was grounded in methodological details, empirical results, or higher-level synthesis.

Beyond interpretability, the hypercube approach often improves retrieval precision because matching occurs in more constrained semantic neighborhoods. For scientific users, this precision translates to fewer irrelevant citations and clearer evidence trails. Hypercube-RAG thus exemplifies how an indexing design , rather than solely improvements to the generator , can materially enhance both performance and explainability in data-rich domains.

E. Multimedia Domain

1) RAIDX: A Retrieval-Augmented Generation and GRPO Reinforcement Learning Framework for Explainable Deepfake Detection: RAIDX tackles a fundamentally multimodal and adversarial task: deepfake detection. Traditional forensic classifiers produce binary labels and opaque feature importances; RAIDX reframes detection as a retrieval-grounded explanation problem. The system retrieves exemplars of known manipulations, model fingerprints, and textual descriptions of generation artifacts, and supplies this evidence to a detector that produces both a verdict and an explanation. A policy-learning module trained via Group Relative Policy Optimization (GRPO) is responsible for generating concise natural-language rationales and visual saliency maps that align with the retrieved evidence.

Explainability is a first-class output in RAIDX. For every flagged video or image the system returns: (1) the detection decision, (2) a short textual rationale citing the most relevant exemplar manipulations or artifact descriptors, and (3) a pixel-level saliency overlay that localizes manipulation cues. This multi-pronged explanation enables human analysts to cross-check the automated judgement rapidly and to assess

whether the evidence provided is sound. The use of reinforcement learning to optimize the explanation policy means the system can learn which explanatory formats are most informative to human reviewers under different conditions.

In practice, RAIDX demonstrates that combining retrieval with learned explanation policies can produce both accurate detection and usable rationales in adversarial multimedia domains. The framework reflects a broader pattern: when tasks demand visual and textual justification, RAG systems must be designed to retrieve appropriate multimodal exemplars and to present explanations in formats that align with expert workflows.

F. Synthesis

Across the surveyed domains several convergent themes emerge. First, structured knowledge representations , knowledge graphs, hierarchical concept layers, and curated case libraries , consistently aid explainability by making relations and contrasts explicit. Where raw-text retrieval can only show evidence fragments, graph- or hierarchy-based retrieval reveals the relational topology that underlies a model's reasoning. Second, modality awareness is crucial in domains such as finance and multimedia: systems that index tables, figures, and images separately and expose those artifacts in the explanation pipeline produce outputs that are far easier to verify by domain experts.

Third, the methods for producing explanations vary along a spectrum from simple provenance surfacing (cite the supporting paragraph) to causal probing (perturbation and ablation) and learned explanation generation (policy-based or multi-agent outputs). Each approach brings trade-offs. Provenance surfacing is lightweight and immediately useful but may not reveal causal influence. Perturbation-based causal probings provide stronger evidence of influence but are computationally more intensive. Learned explanation policies can optimize for human utility but risk learning articulations that are persuasive rather than faithful unless carefully constrained.

Fourth, several practical engineering patterns recur: tiered retrieval strategies that escalate to expensive modalities only when necessary; fusion of case-based and graph-based context to harness the benefits of both precedent and relational reasoning; and constrained output templates in sensitive domains (e.g., legal element checklists, clinical decision cards) to reduce hallucination and improve auditability.

Finally, figuring out how to properly evaluate these systems is an unsolved problem. While we can test pieces of the process in isolation, we don't yet have good tools to measure what really matters: Is the explanation accurate and truthful? Does it make logical sense? Is it actually helpful to a user? Moving forward, we need to develop standard evaluations that combine technical metrics with real-world input from people on whether the system is understandable and trustworthy.

VII. Conclusion

Explainable RAG is rapidly evolving from a collection of ad-hoc techniques into a coherent set of design patterns that prioritize both factual grounding and human interpretability. The surveyed works illustrate that domain adaptation , whether through knowledge graphs, hierarchical retrieval, case-based exemplars, or modality-aware indexing , is often the essential ingredient for producing explanations that stakeholders can trust.

Looking forward, progress will require three complementary advances: better, standardized evaluation of explanation quality; user-centered interfaces that present explanations at appropriate levels of granularity; and scalable causal probes that can be applied to large corpora without prohibitive cost. By pursuing these directions, researchers can help ensure that RAG systems are not only capable of generating informative text but are also accountable and usable in real-world, high-stakes settings.

The convergence of quantitative, human, and causal evaluation frameworks marks a decisive step toward mature Explainable RAG research. The surveyed works collectively transition the field from heuristic interpretability toward scientifically testable explainability, anchored in metrics, validated by human studies, and interpretable through causal reasoning. Future work should focus on (1) building

modular evaluation toolkits that integrate these perspectives, (2) establishing standardized human-alignment protocols, and (3) ensuring evaluator transparency through meta-explanatory rationales.

REFERENCES

[1] Y. Kim, R. Rahimi, and J. Allan, "Discovering biases in information retrieval models using relevance thesaurus as global explanation," 2024. [Online]. Available: https://arxiv.org/abs/2410.03584

[2] T. Formal, B. Piwowarski, and S. Clinchant, "A white box analysis of colbert," 2020. [Online]. Available: https://arxiv.org/abs/2012.09650

[3] J. Qi, G. Sarti, R. Fernández, and A. Bisazza, "Model internals-based answer attribution for trustworthy retrieval-augmented generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024, p. 6037–6053. [Online]. Available: http://dx.doi.org/10.18653/v1/2024.emnlp-main.347

[4] I. Nematov, T. Kalai, E. Kuzmenko, G. Fugagnoli, D. Sacharidis, K. Hose, and T. Sagi, "Source attribution in retrieval-augmented generation," 2025. [Online]. Available: https://arxiv.org/abs/2507.04480

[5] R. Li, C. Chen, Y. Hu, Y. Gao, X. Wang, and E. Yilmaz, "Attributing response to context: A jensen-shannon divergence driven mechanistic study of context attribution in retrieval-augmented generation," 2025. [Online]. Available: https://arxiv.org/abs/2505.16415

[6] H. Chen, Q. Ai, X. Wang, Y. Liu, F. Lin, and Q. Liu, "Unsupervised dense retrieval with conterfactual contrastive learning," 2024. [Online]. Available: https://arxiv.org/abs/2412.20756

[7] 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," 2025. [Online]. Available: https://arxiv.org/abs/2508.20131

[8] J. Zhou and L. Chen, "Openrag: Optimizing rag end-to-end via in-context retrieval learning," 2025. [Online]. Available: https://arxiv.org/abs/2503.08398

[9] B. Zhang, H. Xin, Y. Chen, Z. Liu, B. Yi, T. Li, L. Nie, Z. Liu, and M. Fang, "Who taught the lie? responsibility attribution for poisoned knowledge in retrieval-augmented generation," 2025. [Online]. Available: https://arxiv.org/abs/2509.13772

[10] L. Ranaldi, M. Valentino, and A. Freitas, "Eliciting critical reasoning in retrieval-augmented generation via contrastive explanations," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025, p. 11168–11183. [Online]. Available: http://dx.doi.org/10.18653/v1/2025.naacl-long.557

[11] X. Ma *et al.*, "Visa: Retrieval augmented generation with visual source attribution," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[12] N. Patel *et al.*, "Towards improved multi-source attribution for long-form answer generation," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

[13] Z. Sun *et al.*, "Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability," 2024.

[14] Q. Lyu *et al.*, "Faithful chain-of-thought reasoning," 2023.

[15] A. Asai *et al.*, "Self-rag: Self-reflective retrieval-augmented generation," in *NeurIPS 2023 Workshop*, 2023.

[16] V. Sudhi *et al.*, "Towards end-to-end model-agnostic explanations for rag systems," 2025.

[17] C.-Y. Chang *et al.*, "Main-rag: Multi-agent filtering retrieval-augmented generation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[18] I. Besrour *et al.*, "Ragenta: Multi-agent retrieval-augmented generation for attributed question answering," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2025.

[19] R. Arora *et al.*, "Ragviz: Diagnose and visualize retrieval-augmented generation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[20] J. Szlichta, "Rage against the machine: Retrieval-augmented llm explanations," 2024.

[21] H. Chen, R. Wang, and J. Lee, "Interpretable information retrieval for large language models," 2023, add venue/DOI if available.

[22] A. Gupta, M. Srivastava, and P. Bansal, "Fleek: Faithfulness and linkage evaluation of explanations in knowledge-grounded generation," 2024, add venue/DOI if available.

[23] K. Shuster, J. Xu, and J. Weston, "Ragas: Reference-free evaluation for retrieval-augmented generation," 2024, add venue/DOI if available.

[24] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: A large-scale dataset for fact extraction and verification," in *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[25] R. Arora, S. Kim, and D. Liu, "Faithfulrag: Causal attribution metrics for retrieval-augmented generation," 2025, add venue/DOI if available.

[26] M. Park, Z. Chen, and C. Lin, "Trust-score: Calibrated evaluation of trustworthy explanations," 2024, add venue/DOI if available.

[27] M. Ahmed, T. Rahman, and Y. Zhou, "Gfm-rag: Grounded factuality metric for retrieval-augmented models," 2024, add venue/DOI if available.

[28] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[29] R. Singh, V. Kapoor, and P. Das, "Garage: Human-centric evaluation framework for grounded explanations," 2024, add venue/DOI if available.

[30] H. Zhang, W. Li, and F. Wang, "Uaeval4rag: User-adapted evaluation of rag explainability," 2024, add venue/DOI if available.

[31] D. Patel, R. Anand, and T. Banerjee, "Binder: Benchmarking interpretability in document-grounded reasoning," 2023, add venue/DOI if available.

[32] T. Yang, J. Cho, and S. Park, "True: Trustworthy rag explanations benchmark," 2024, add venue/DOI if available.

[33] E. Cho, J. Han, and J. Kim, "Ragbench and trace: Unified evaluation for retrieval-grounded generation," 2024, add venue/DOI if available.

[34] L. Wang, M. Sun, and K. Yu, "Rag-ex: Self-rationalizing evaluation of explanations," 2024, add venue/DOI if available.

[35] D. Lee, H. Choi, and Y. Park, “Kg-smile: Knowledge-graph structural metrics for interpretable language explanations,” 2025, add venue/DOI if available.