FaX-RAG: Provenance-Centered Fairness and Explainability for Retrieval-Augmented Generation

Emilia Lenzi
Politecnico di Milano
Milan, Italy
TU Wien
Vienna, Austria
emilia.lenzi@polimi.it

Kostas Stefanidis
Data Science Research Centre
Tampere University
Tampere, Finland
konstantinos.stefanidis@tuni.fi

ABSTRACT

Retrieval-Augmented Generation (RAG) is becoming a dominant paradigm for grounding large language models in external knowledge. However, current RAG systems provide answers with citations rather than auditable accounts of how retrieved evidence shapes generated outputs. This affects both explainability and fairness: without visibility into evidence use, it is difficult to assess whether answers are faithfully grounded or whether certain sources, groups, or perspectives are over-represented, ignored, or disproportionately influential. In this vision paper we present FaX-RAG, a provenance-centered framework for fair and explainable RAG. FaX-RAG treats attribution traces as queryable provenance records linking generated claims to the evidence units that shaped them, produced during retrieval and generation, validated post-hoc, and reused to audit disparities between evidence exposure and evidence influence. We outline a modular architecture combining model-integrated explainability, post-hoc explainability, and post-hoc fairness, and we show why autoregressive, stochastic generation makes this provenance estimated rather than derived, a shift that opens data management challenges around trace validation, indexing, querying, and maintenance. FaX-RAG moves RAG from answers that cite to systems that account.

KEYWORDS

Retrieval-Augmented Generation, Data Provenance, Fairness Auditing

1 INTRODUCTION

Context & Motivation. Large Language Models (LLMs) are increasingly used in knowledge-intensive and high-stakes domains, where trustworthiness, transparency, and accountability are critical. Retrieval-Augmented Generation (RAG) has emerged as a dominant paradigm for grounding model outputs in external knowledge. From a data management perspective, RAG can be viewed as a retrieval–generation pipeline: a retriever selects evidence from an external corpus, and a generator uses this evidence to produce the final answer [18–20, 23]. While this design improves factuality and scalability, it also introduces new challenges. Evidence may be retrieved but not used, used but not cited, or amplified in ways that are difficult to observe. As a result, current RAG systems often provide answers with citations, but not necessarily auditable accounts of how evidence shaped the output. This lack of visibility affects both explainability and fairness. If we cannot trace which sources, passages, table cells, image regions, or perspectives influenced a generated claim, it becomes difficult to assess whether the system is faithfully grounded or whether some sources and groups are systematically over-represented, under-represented, or ignored. Conversely, fairness concerns give direction to explainability by identifying which evidence flows should be inspected and monitored. We therefore argue that explainability and fairness in RAG should not be treated as separate post-hoc properties, but as coupled dimensions of an auditable data pipeline.

Data Management Challenge. RAG systems raise a set of challenges that are naturally aligned with the data management community. First, generated outputs may depend on latent, fine-grained, and stochastic relations between user queries, retrieved evidence, prompt construction, model memory, and previous generated tokens. These relations are not captured by traditional document-level citations or by classical provenance models alone [7]. Second, retrieval and generation may introduce or amplify disparities in evidence exposure and influence, raising fairness risks that cannot be fully addressed by standard ranking or output-level evaluation alone [37]. Third, evaluating accountable RAG requires joint reasoning over retrieved data, attribution signals, explanations, fairness-relevant metadata, and generated outputs, making traditional evaluation pipelines insufficient [10]. These challenges call for new abstractions and system support for representing, storing, indexing, querying, and validating provenance and attribution traces in RAG. In particular, RAG systems should be designed as auditable data pipelines, where evidence selection, evidence use, explanation, and fairness auditing are explicitly connected.

FaX-RAG Vision. Existing work has largely studied explainability and fairness in RAG as separate concerns: explanation methods reveal which retrieved evidence supports an answer, while fairness methods balance retrieval exposure or mitigate output disparities. We argue the problems are coupled, namely, fairness cannot be assessed without understanding how evidence influences generation, and explanation becomes actionable when used to identify unfair evidential pathways [43], and we introduce FaX-RAG, a vision for fair and explainable RAG grounded in a common provenance-centered abstraction. Its core idea is to treat attribution as a first-class signal produced during retrieval and generation, validated post hoc, and reused for fairness auditing. Three components realize this and are described in Section 3: Model-Integrated Explainability, Post-Hoc Explainability, and Post-Hoc Fairness.

Contributions. We introduce a provenance-centered view of RAG in which generated claims, evidence units, attribution signals, validation results, and fairness-relevant metadata are queryable objects (Section 3); we distinguish evidence exposure from evidence influence and argue this distinction is necessary for fairness auditing, since balanced retrieval does not imply balanced influence (Sections 2.3, 3); we identify the properties of

EDBT '27, 6-9 April 2027, Lille (France)
© 2026 Copyright held by the owner/author(s). Published on OpenProceedings.org under ISBN 978-3-89318-106-3, series ISSN 2367-2005. Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.

language model generation that make this provenance estimated rather than derived (Section 3.5); and we outline the resulting data management agenda (Section 4).

2 RESEARCH GAPS AND DESIGN QUESTIONS

Most RAG systems optimize relevance, answer quality, or citation quality, but expose little about how retrieved evidence is selected, transformed, and actually used during generation [23, 28]. Fairness-aware methods, in turn, address retrieval exposure or output-level disparities without tracing how specific sources, groups, or perspectives influence the final answer [12, 30, 38]. From a data management perspective, the gap is that RAG pipelines process evidence but are rarely designed to be auditable [4], in contrast to data and machine-learning pipelines, where lineage-based inspection and database repair have been used to debug fairness-relevant behaviour [14, 22, 35].

2.1 Beyond Chunk-Level Provenance

A first question concerns the granularity of provenance in RAG. In many current systems, provenance is approximated by returning the retrieved document or text chunk used to condition the answer [33, 40]. Consider a public-health analyst preparing an evidence brief on barriers to maternal healthcare access in low-resource settings, a running example we use throughout the paper and develop in Section 5. The corpus is heterogeneous and multimodal, including policy reports, NGO documents, journal articles, tables of regional indicators, and figures summarizing intervention outcomes, and each evidence unit carries metadata, such as region, country income group, institution type, publication year, venue, and language. A retrieved country report may cover access barriers, financing, and intervention outcomes, while only one paragraph supports the generated recommendation on transport distance. Chunk-level provenance is useful for readability and citation, but here it attributes the recommendation to three topics when one paragraph carried it.

The underlying issue is that retrieved evidence is not necessarily used evidence [29]. A chunk may sit in the prompt and have little or no effect on the output, and a highly ranked chunk may be reproduced from the model's parametric memory rather than actually consulted, as in row 3 of Table 1. Chunk-level provenance records what was exposed to the generator; it cannot record which evidence units influenced a specific claim. Fine-grained provenance is therefore needed to separate the two, a separation we formalize as exposure and influence in Section 3 and whose fairness consequences we take up in Section 2.3.

FaX-RAG therefore distinguishes between coarse provenance, such as document or chunk identifiers, and fine-grained attribution, such as sentence spans, table cells, figure panels, image regions, or OCR-linked visual elements. The goal is not to replace readable retrieved chunks, but to complement them with evidence-use signals that make the relation between generated claims and supporting evidence auditable.

2.2 Multimodal Provenance Granularity

A second question concerns what provenance means when the retrieved evidence is multimodal, as it is in the example above. For text, the relevant evidence unit may be a passage, sentence, or span. For tables, it may correspond to a row, column, cell, or group of cells. For images and figures, provenance is less straightforward: the relevant unit may be an entire image, a figure panel, a detected object, a labeled region, a visual trend, or an OCR-linked element [4]. The appropriate granularity is therefore task-dependent. Storing only an image identifier may be sufficient when the system cites the image as a whole. However, it is not sufficient when the generated answer depends on a specific subregion, object, label, panel, or visual comparison. For example, if a RAG system answers a question about a trend shown in one panel of a multi-panel figure, image-level provenance cannot indicate which panel or visual pattern supports the claim.

This motivates a hierarchical view of multimodal provenance, from coarse document- or image-level identifiers down to the modality-specific evidence units already introduced above, letting RAG systems adapt granularity to the task, the modality, and the reliability of available extraction methods.

2.3 From Evidence Exposure to Evidence Influence

A third question is how provenance supports fairness auditing. In the example above, a system may retrieve reports from several regions while producing a brief shaped almost entirely by English-language documents from large international organizations. For this reason, FaX-RAG distinguishes between two forms of disparity. *Exposure disparities* concern which sources are retrieved, ranked, or inserted into the prompt. *Influence disparities* concern how much those sources actually affect the generated output [17, 33, 38, 42]. This distinction enables a more precise form of fairness analysis: it becomes possible to ask not only whether different groups, regions, institutions, or perspectives were retrieved, but also whether they contributed meaningfully to the answer. The distinction also makes the coupling argued in Section 1 operational: explainability supplies the signals needed to measure influence, while fairness identifies which evidential pathways to inspect.

2.4 Architecture Dependence and Generality

A fourth question concerns the extent to which accountable RAG can be model-independent. Different RAG systems expose different internal signals: some provide access to retrieval scores, ranks, prompt positions, attention maps, gradients, or decoding states, while others are available only through black-box APIs[31, 39]. Therefore, a practical framework cannot assume a single retriever, generator, or decoding strategy.

FaX-RAG is designed to be general at the level of abstraction, while allowing different implementation depths depending on system access. Its core abstraction is a provenance and attribution layer that links generated outputs to retrieved evidence units, metadata, and uncertainty signals.

The framework is therefore not tied to a model family or attribution method: it defines design principles, namely, expose evidence selection, estimate influence, validate attribution, and reuse it for fairness auditing, whose implementation depth varies with the access a system provides.

Implications for data management. These questions point to a broader data management agenda: if provenance and attribution become first-class objects, systems must represent, store, index, query, validate, and maintain them as corpora, retrieval indexes, and models evolve. We return to this agenda, and to the audit query patterns it must support, in Section 4.

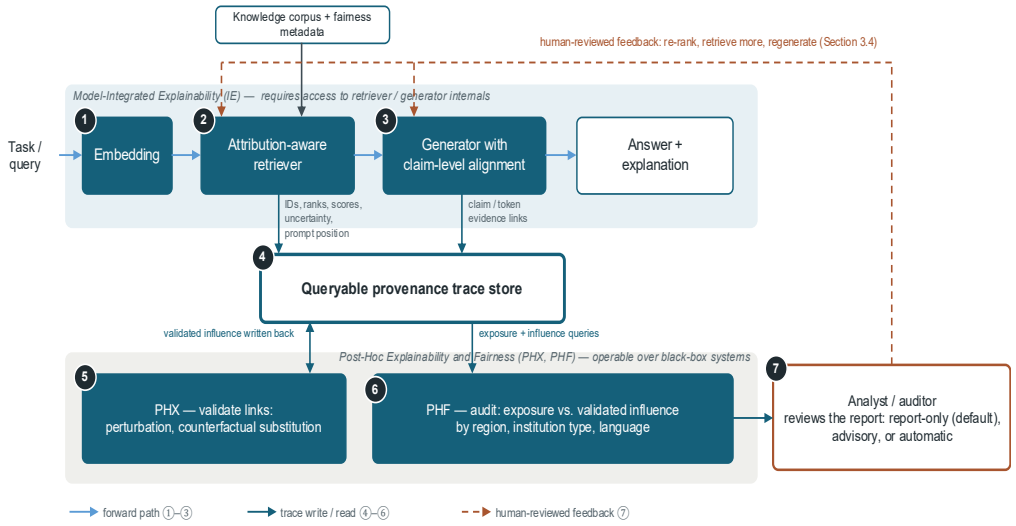


Figure 1: FaX-RAG. Forward path: (1) the task is embedded; (2) the attribution-aware retriever returns evidence units with identifiers, ranks, scores, uncertainty estimates, and prompt positions; (3) the generator produces the answer with claim-level alignment to evidence units; (4) IE writes both signal sets into the queryable provenance trace store. Audit path: (5) PHX validates highly attributed links through selective perturbation and counterfactual evidence substitution, writing validated influence $\hat{\alpha}$ back to the store; (6) PHF queries exposure against validated influence across metadata dimensions such as region, institution type, and language, and reports disparities; (7) the analyst reviews the diagnostic and, where warranted, triggers fairness-aware re-ranking, additional retrieval, or regeneration, producing a before-and-after comparison. Steps 1-4 populate traces and 5-7 consume them; only (2) and (3) require model-integrated access.

3 FAX-RAG: A PROVENANCE-CENTERED ARCHITECTURE

At the core of FaX-RAG lies the idea that explainability and fairness are properties of the retrieval–generation process itself, not external layers added after inference. As shown in Figure 1, the three modules introduced in Section 1 form a single auditable pipeline, built around the principle that attribution should not be produced once and discarded, but should become a reusable data object supporting explanation, validation, fairness auditing, and system debugging.

The shared object across the pipeline is the provenance trace, defined next. The rest of the architecture can be understood as a pipeline for producing, validating, aggregating, querying, and acting on these traces.

Provenance traces. Let q be a user query, $G_q = \langle g_1, \ldots, g_n \rangle$ the generated answer decomposed into claim-level units, and $E_q = \{e_1, \ldots, e_k\}$ the evidence units retrieved for q. An evidence unit is a modality-specific addressable region of a source: a character span, a table cell or cell group, a figure panel, an image region, or an OCR-linked label. A provenance trace is a finite relation T_q with schema: Trace(gen_unit, ev_unit, $meta$, $retrieval$, loc, α, u), where $meta$ holds source and fairness-relevant attributes such as region, institution type, language, venue, and source category; $retrieval$ holds rank, score, candidate-set membership, and prompt position; loc holds the modality-specific location; $\alpha(g, e) \in [0, 1]$ is an estimated influence weight of e on g; and $u(g, e)$ quantifies uncertainty in that estimate. IE produces traces, PHX annotates them with validated influence $\hat{\alpha}$, and PHF queries them.

Two derived views make the exposure-influence distinction precise. For a metadata value v and a rank discount $disc(\cdot)$, $\mathrm{Exp}(v) = \frac{\sum_{e \in E_q : v \in meta(e)} disc(rank(e))}{\sum_{e \in E_q} disc(rank(e))}$, $\mathrm{Inf}(v) = \frac{\sum_{g \in G_q} \sum_{e : v \in meta(e)} \hat{\alpha}(g,e)}{\sum_{g \in G_q} \sum_{e \in E_q} \hat{\alpha}(g,e)}$, with exposure–influence gap $\Delta(v) = \mathrm{Inf}(v) - \mathrm{Exp}(v)$. Here $\mathrm{Inf}(v)$ aggregates validated attribution mass; it does not assume that influence decomposes additively over evidence units. Because retrieved evidence can be redundant or complementary (Section 3.5), the $\hat{\alpha}$ must come from a set-level allocation, for instance, a Shapley-style division of the joint effect, which distributes the total among units while accounting for interactions [26]. Where only leave-one-out estimates are available, $\mathrm{Inf}(v)$ should be read as an approximation whose error grows with redundancy in the retrieved set. Exposure is computed over the retrieved and ranked set and is observable in any RAG system; influence is computed over validated attribution weights and must be estimated. Fairness auditing in FaX-RAG is, at its core, the comparison of these two distributions across metadata dimensions, together with the diagnostics that establish whether $\hat{\alpha}$ can be trusted. Table 1 shows three trace instances for the running example. In row 3, the evidence unit is highly ranked and receives high raw attribution, yet removing it leaves the answer semantically unchanged and a closed-book control reproduces the claim, so its validated influence is near zero: the claim is consistent with the evidence but not caused by it. Distinguishing rows 1–2 from row 3 is not possible from citations or chunk-level provenance, and, as Section 3.5 explains, the effect is systematic rather than incidental.

Relation to existing provenance models. The trace is not a new provenance algebra but a two-layer construct. Its coarse layer is workflow provenance: retrieval, prompt construction, generation, and validation as activities, and queries, evidence units, prompts, answers, and explanation reports as entities, and is intended to be serializable in standard workflow-provenance vocabularies [27]. Its fine layer plays the role of where- and why-provenance over evidence units [6, 7]. Three departures from classical data provenance are essential. First, annotations are

Table 1: Illustrative provenance trace instances for Section 2.1. Rows 1–2 record evidence necessary for the generated claim. Row 3 records an evidence unit with high exposure and high raw attribution but near-zero validated influence.

Generated unit g	Evidence unit e	*loc*	*meta*	rank / pos.	α	$\hat{\alpha}$	u	PHX verdict
"transport distance is a primary barrier"	paragraph, national ministry report	span 412–587	region A; ministry; local language	4 / 6	0.41	0.38	low	evidence-necessary
"coverage rose after the voucher programme"	cell, regional indicator table	row 7, col 3	region A; multilateral; English	1 / 1	0.62	0.57	low	evidence-necessary
"facility-based delivery is widely recommended"	passage, international NGO report	span 88–210	region unspecified; NGO; English	2 / 2	0.55	**0.04**	low	evidence-consistent, not necessary

Table 2: Research agenda and open challenges for provenance-centered RAG.

Challenge	Risk	Design direction	Maturity
Provenance granularity	Chunk-level evidence may over-attribute generated claims to evidence that was retrieved but not actually used.	Use hierarchical provenance, from documents and chunks to spans, table cells, figure panels, image regions, and OCR-linked elements.	Partial: span-level attribution exists for text; claim-to-span linking is emerging.
Multimodal evidence	Evidence units differ across text, tables, figures, and images, making uniform provenance difficult.	Define task-dependent and modality-aware evidence units, with different levels of granularity depending on available extraction methods.	Early: grounding and OCR make regions addressable; no accepted evidence-unit taxonomy.
Attribution faithfulness	Internal attribution signals may reflect model heuristics rather than genuine evidential influence.	Validate explanations through selective perturbations, counterfactual retrieval, attribution shifts, citation changes, and semantic consistency checks.	Emerging and contested: perturbation protocols exist; agreement with internal signals unresolved.
Scalability	Fine-grained token-level traces may be expensive to store, index, query, and update.	Develop compressed, sampled, aggregated, or indexed representations of provenance and attribution traces.	Open: no trace-specific compression, indexing, or query optimization.
Architecture dependence	Different RAG systems expose different internal signals, and many deployed systems are black-box or proprietary.	Use a modular provenance abstraction, with model-integrated implementations for open systems and post-hoc approximations for black-box systems.	Partial: post-hoc approximation available for black-box systems; model-integrated signals only in open stacks.
Fairness operationalization	Fairness-relevant metadata may be incomplete, unavailable, or task-dependent; content-level biases may not be captured by source metadata alone.	Combine source metadata with content-level signals such as topical framing, stance, entity variety, language, and narrative emphasis.	Split: exposure-side measures mature from fair ranking; influence-side essentially unaddressed.
Utility trade-offs	Fairness interventions may reduce relevance, factuality, latency, or explanation faithfulness.	Treat fairness, accuracy, explainability, and cost as multi-objective trade-offs rather than isolated optimization targets.	Open: no benchmark jointly covering fairness, faithfulness, and cost.

real-valued and uncertainty-qualified rather than boolean set memberships, placing the semantics closer to approximate and probabilistic lineage than to exact why-provenance [1, 5]. Second, the generator is not a declarative operator, so provenance cannot be derived by static analysis of a query expression: it must be estimated from model signals and validated by intervention, which makes provenance an inference and measurement problem rather than a rewriting problem. Third, influence is not additively decomposable over evidence units (Section 3.5), so the compositional assumptions underlying semiring-based how-provenance do not hold [15]. What follows from provenance that is estimated rather than derived is a systems agenda: validation, compression, indexing, staleness under corpus and model updates, and approximate auditing with error bounds (Section 4).

3.1 Model-Integrated Explainability

Goal. The goal of Model-Integrated Explainability is to make evidence use observable during inference. Instead of treating retrieval scores, citations, or attention patterns as isolated proxies, IE aims to capture attribution signals that connect generated outputs to the evidence units retrieved and processed by the system.

Design principle. In open or modifiable RAG architectures, IE can be implemented by instrumenting both retrieval and generation. On the retrieval side, the system can expose document identifiers, ranks, scores, confidence or uncertainty estimates, prompt positions, and candidate-set information. On the generation side, the system can log token-to-evidence interactions, cross-attention patterns, or other attribution signals that indicate which evidence units contributed to specific generated claims. Possible implementations include attribution-aware retrievers, differentiable retrieval approximations, attention- or gradient-based attribution, and decoder-side alignment mechanisms [3, 8, 16, 18, 19, 41].

Open challenges. Model-integrated provenance is not deterministic causality but a probabilistic, multi-source influence signal (Section 3.5). A promising abstraction is to represent provenance as a weighted influence graph, where generated claims are connected to evidence units, prompt components, and contextual dependencies with uncertain attribution weights. Key challenges include non-differentiable retrieval, long-context dilution of attribution signals, storage overhead, and the varying availability of internal signals across RAG architectures.

3.2 Post-Hoc Explainability

Goal. The goal of Post-Hoc Explainability is to assess whether attribution signals are faithful and to transform dense provenance traces into explanations that humans can inspect. Model-integrated signals provide visibility, but they do not automatically guarantee that the attributed evidence was genuinely necessary or influential.

Design principle. PHX complements IE through validation and summarization. Validation can be performed through selective perturbations, guarded ablations, counterfactual retrieval, citation changes, or output comparisons. Rather than perturbing all retrieved evidence, FaX-RAG treats post-hoc validation as a selective auditing process: highly attributed spans, cells, regions, or documents are prioritized for testing. Summarization then aggregates fine-grained links into claim-level, source-level, cluster-level, or theme-level explanations, making attribution reports more interpretable and easier to query.

Open challenges. Perturbation-based validation can be computationally expensive because it may require additional retrieval and generation calls. This motivates caching, batching, approximate local validation, and selective auditing strategies. Another challenge is deciding whether a perturbation has changed the answer: lexical differences alone are insufficient, so PHX should consider semantic similarity, attribution shifts, citation changes, and task-aware consistency metrics. In black-box or proprietary

settings, PHX may not access internal model states, but it can still approximate explanation validation through input-level perturbations, output comparisons, saliency-based analyses, and metadata-aware auditing [24, 32–34, 36].

3.3 Post-Hoc Fairness

Goal. The goal of Post-Hoc Fairness is to use provenance and attribution traces to audit and mitigate unfairness in evidence exposure and evidence influence. This goes beyond evaluating whether the answer appears fair: it asks which sources were retrieved, which sources were used, and which perspectives shaped the output.

Design principle. PHF links attribution traces to fairness-relevant information. This includes available and appropriate source metadata, such as publication region, institution type, language, venue, or declared source category. In some tasks, sensitive attributes may be relevant for auditing, but FaX-RAG should not require inferring unavailable sensitive attributes; when proxies are used, they must be explicitly documented, uncertainty-aware, and subject to governance constraints. On top of these signals, PHF computes the Exp and Inf views of Section 3 and compares them. Comparing these two distributions allows FaX-RAG to detect cases where evidence is retrieved but weakly used, or where a narrow set of sources disproportionately shapes the answer. These diagnostics can guide interventions such as fairness-aware re-ranking, source diversification, exposure balancing, or fairness-aware decoding (under the review modes of Section 3.4) [12, 13, 21, 42].

Open challenges. Fairness in RAG is task-dependent and often difficult to operationalize, especially when sensitive attributes are incomplete, implicit, unavailable, or inappropriate to infer. In such settings, FaX-RAG can combine available metadata with content-based proxies, but these proxies must be used carefully and transparently. Another challenge is balancing fairness with relevance, factuality, and explanation faithfulness. Fairness interventions should therefore be treated as soft constraints or multi-objective trade-offs rather than as absolute rules. Since PHF relies on attribution traces, its reliability also depends on the faithfulness of IE and PHX, making fairness auditing inseparable from explanation validation.

3.4 Audit Loops and Human Review

In the forward path, the user query is embedded, evidence is retrieved, the generator produces an answer, and IE records provenance traces; PHX validates and summarizes them into explanation reports; PHF queries the same validated traces to compare exposure against influence across sources, groups, or perspectives. These results are then surfaced rather than automatically applied. Because PHF consumes estimates produced by IE and PHX, closing the loop automatically means optimizing the pipeline against a noisy and contestable measurement, which invites two failures: metric gaming, in which the reported influence distribution is balanced without evidence use changing, and rationalization, in which an intervention improves the audit report rather than the answer [2, 9]. We therefore distinguish three modes. In report-only mode, the default, diagnostics are presented to an analyst or auditor and the pipeline is unchanged. In advisory mode, the system proposes candidate interventions: fairness-aware re-ranking, additional retrieval from under-represented sources, citation revision, balanced regeneration, or further perturbation tests for uncertain high-impact claims, each with a before-and-after preview over the answer, citations, influence distribution, and source balance, and a human decides. In automatic mode, only pre-approved, reversible, and logged policies may be applied without review, and the applied policy is itself recorded in the trace so that the intervention remains auditable.

This makes the loop an accountability mechanism rather than an optimizer, and clarifies FaX-RAG's role as a vision architecture: systems may instantiate the modules at different levels of access and sophistication, but the provenance-centered logic is common: expose evidence use, validate attribution, audit fairness, and present the results for adjudication.

3.5 Why Language Model Generation Breaks Generic Provenance

The abstractions above could in principle be applied to any black-box evidence-processing system. They are not sufficient for RAG: the generator is autoregressive, stochastic, and independently knowledgeable, and each property breaks an auditing assumption in a direction that matters for fairness.

Estimation. Generation is sequential, so evidence influences a late claim both directly and through earlier generated text; a scalar attribution conflates the two and under-credits evidence that sets an answer's framing without supporting any single claim. Generation is also stochastic, so any single-run attribution is a sample statistic, and perturbation-based validation cannot separate a change caused by removing evidence from one caused by resampling. Influence must therefore be an expectation over replicates under controlled seeds, and traces must be generation-ordered.

Directional bias. Three effects skew auditing systematically rather than noisily. Removing an evidence unit may not change the answer because the model already knows the fact (Table 1, row 3), so leave-one-out under-attributes well-represented topics and requires a closed-book control and a necessity typology rather than a single score. Redundant and complementary evidence defeat leave-one-out symmetrically, and since redundancy concentrates on widely reproduced sources, set-level attribution becomes a requirement, and is what makes the influence shares of Section 3 well defined. Attention also varies with prompt position, so without position debiasing a mechanical property of the generator is reported as a fairness finding about groups whose sources land mid-context [25].

Uncertainty. Answer uncertainty, whether a claim is correct, is the target of work on calibration, self-consistency, and semantic entropy [11]; attribution uncertainty, the field u of the trace, concerns whether an evidence unit actually influenced a claim and has no counterpart there. High answer confidence with low attribution confidence is the case most dangerous for accountability, and FaX-RAG uses the pair to triage validation effort.

4 RESEARCH AGENDA AND OPEN CHALLENGES

Table 2 summarizes the main risks and corresponding design directions that should guide future work. A key systems challenge is that naive token-level provenance can be prohibitively large. FaX-RAG therefore calls for storage-aware trace representations. Possible directions include claim-level aggregation, sparse top-k attribution links, provenance sketches, hierarchical indexes over

evidence units, and materialized audit views for recurring fairness dimensions. This opens database research questions around trace compression, incremental maintenance under corpus and model updates, approximate auditing with error bounds, and provenance query optimization. Because exposure depends on the corpus and retriever while influence depends on the generator, a model version change invalidates influence annotations while leaving exposure intact, so incremental maintenance should support selective re-validation of high-impact links rather than full recomputation.

Evaluating provenance-centered RAG requires a multidimensional methodology that treats explanations, fairness diagnostics, and system overhead as part of the data pipeline. Benchmarking accountable RAG therefore requires not only datasets and generated answers, but also metadata-rich corpora, claim-level evidence annotations, perturbation protocols, fairness-aware audit tasks, and provenance query workloads. Table 2 also indicates where the community stands. Black-box PHX and PHF over metadata-rich corpora are feasible today; a persistent trace store with claim-level aggregation and an audit query interface is a realistic near-term target; and model-integrated attribution as a first-class retriever and generator output, fairness-aware decoding, and incremental trace maintenance under model updates remain longer-term goals.

5 ILLUSTRATIVE SCENARIO: FAIR EVIDENCE SUMMARIZATION

To illustrate the accountability FaX-RAG supports, we return to the public-health analyst of Section 2.1. The central question in preparing the brief is not which sources were retrieved, but which sources actually influenced the answer.

In the FaX-RAG workflow, the IE module records traces linking generated claims to evidence units. For example, a claim about transport barriers may be linked to a specific paragraph in a local ministry report, while a claim about intervention outcomes may be linked to a table cell or figure panel in a policy evaluation. The PHX module then validates and summarizes these links through selective perturbations and counterfactual evidence substitutions, checking whether highly attributed evidence is genuinely influential rather than merely present in the prompt.

The PHF module uses the same validated traces to compare exposure and influence across fairness-relevant dimensions such as region, institution type, and language. This may reveal that some sources are retrieved but have little influence on the generated brief, while others dominate the final framing. For example, an audit may show that highly ranked English-language reports from large international organizations receive substantial exposure and high raw attribution, yet much of that attribution does not survive validation, because the claims they appear to support are generic recommendations the model reproduces from parametric memory (Table 1, row 3). The region-specific claims that carry the brief are instead supported by lower-ranked local ministry reports (Table 1, row 1). The disparity runs in both directions: the sources that shape the visible framing are not the ones doing the evidential work, and the sources doing that work are the least exposed. This discrepancy is difficult to detect from citations alone. Then the analyst reviews a report showing how the answer, citations, evidence influence, and source balance would change under fairness-aware re-ranking or source diversification, and decides whether to apply it.

The same machinery applies to other fairness dimensions. In legal research, metadata may include jurisdiction, court level, language, and publisher type. An audit may reveal that lower-court and dissenting opinions receive substantial exposure, while the summary is shaped mainly by appellate majority opinions from one jurisdiction, a disparity hidden by a citation list.

6 CONCLUSION

RAG systems are increasingly used as interfaces to external knowledge, yet their current reliance on retrieved chunks and citations provides only limited accountability. In this paper, we presented FaX-RAG, a provenance-centered vision for fair and explainable RAG, in which attribution traces become first-class, queryable objects linking generated claims to the evidence units that shaped them. Its central claim, developed in Section 3.5, is that provenance in this setting cannot be derived from the structure of a declarative query but must be estimated and validated, with every fairness diagnostic inheriting that estimate's uncertainty. This is what makes accountable RAG a data management problem rather than only a modeling one, opening a concrete agenda: representations and indexes for attribution traces, approximate auditing with error bounds, incremental maintenance as corpora and models change, and benchmarks that treat explanation faithfulness, fairness diagnostics, and system cost as joint objectives.

ARTIFACTS

This paper is a vision paper and does not introduce an implemented system, experimental benchmark, dataset, or software artifact. The contribution is conceptual: it defines the FaX-RAG framework and outlines a research agenda for provenance-centered fair and explainable retrieval-augmented generation.

REFERENCES

[1] Parag Agrawal, Omar Benjelloun, Anish Das Sarma, Chris Hayworth, Shubha U. Nabar, Tomoe Sugihara, and Jennifer Widom. 2006. Trio: A System for Data, Uncertainty, and Lineage. In *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006.* ACM, 1151–1154.

[2] Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97).* PMLR, 161–170.

[3] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf. 2022. XAI for Transformers: Better Explanations through Conservative Propagation. In *International Conference on Machine Learning.* PMLR, 435–451.

[4] Sihem Amer-Yahia, Jasmina Bogojeska, Roberta Facchinetti, Valeria Franceschi, Aristides Gionis, Katja Hose, Georgia Koutrika, Roger Kouyos, Matteo Lissandrini, Silviu Maniu, Katsiaryna Mirylenka, Davide Mottin, Themis Palpanas, Mattia Rigotti, and Yannis Velegrakis. 2025. Towards Reliable Conversational Data Analytics. In *Proceedings of the 28th International Conference on Extending Database Technology (Advances in database technology EDBT Proceedings, 3).* 962–969.

[5] Omar Benjelloun, Anish Das Sarma, Alon Y. Halevy, and Jennifer Widom. 2006. ULDBs: Databases with Uncertainty and Lineage. In *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006.* ACM, 953–964.

[6] Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. 2001. Why and Where: A Characterization of Data Provenance. In *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*, Vol. 1973. 316–330.

[7] James Cheney, Laura Chiticariu, and Wang Chiew Tan. 2009. Provenance in Databases: Why, How, and Where. *Found. Trends Databases* 1, 4 (2009), 379–474.

[8] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.* Association for Computational Linguistics, Florence, Italy, 276–286.

[9] Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. 2020. You Shouldn't Trust Me: Learning Models Which Conceal Unfairness from Multiple

Explanation Methods. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, Vol. 325. IOS Press, 2473–2480.

[10] S. Es, J. James, L. E. Anke, and S. Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 150–158.

[11] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nat.* 630, 8017 (2024), 625–630.

[12] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (Sept. 2024), 1097–1179.

[13] Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. 2022. FAIR: Fairness-aware information retrieval evaluation. *Journal of the Association for Information Science and Technology* 73, 10 (2022), 1461–1473.

[14] Stefan Grafberger, Shubha Guha, Julia Stoyanovich, and Sebastian Schelter. 2021. MLINSPECT: A Data Distribution Debugger for Machine Learning Pipelines. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*. 2736–2739.

[15] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. 2007. Provenance semirings. In *Proceedings of the Twenty-Sixth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 11-13, 2007, Beijing, China*. 31–40.

[16] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. arXiv:2004.11207 [cs.CL]

[17] Frank Hopfgartner and Jialie Shen. 2026. Trustworthy Personalized Retrieval in the LLM Era: Fairness, Bias Awareness, Privacy, and Multimodality. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Australia) *(SIGIR '26)*. Association for Computing Machinery, New York, NY, USA, 5328–5331. doi:10.1145/3805712.3808647

[18] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. arXiv:2007.01282 [cs.CL]

[19] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.* 24, 1, Article 251 (Jan. 2023), 43 pages.

[20] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, others, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

[21] To Eun Kim and Fernando Diaz. 2025. Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*. 33–43.

[22] Emilia Lenzi and Kostas Stefanidis. 2025. ADAPT: Fairness & diversity for sequential group recommendations. *Inf. Syst.* 133 (2025), 102572. doi:10.1016/J.IS.2025.102572

[23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL]

[24] Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Understanding Neural Networks through Representation Erasure. arXiv:1612.08220 [cs.CL]

[25] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguistics* 12 (2024), 157–173.

[26] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[27] Luc Moreau, Paolo Missier, Khalid Belhajjame, et al. 2013. *PROV-DM: The PROV Data Model*. W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

[28] Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, Ryan Rossi, Franck Dernoncourt, Md Mehrab Tanjim, Nesreen Ahmed, Xiaorui Liu, Wenqi Fan, Erik Blasch, Yu Wang, Meng Jiang, and Tyler Derr. 2025. Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2502.06872* (2025). arXiv:2502.06872

[29] Nilay Patel, Shivashankar Subramanian, Siddhant Garg, Pratyay Banerjee, and Amita Misra. 2024. Towards Improved Multi-Source Attribution for Long-Form Answer Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Mexico City, Mexico, 3906–3919.

[30] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2022. Fairness in rankings and recommendations: an overview. *VLDB J.* 31, 3 (2022), 431–458. doi:10.1007/S00778-021-00697-Y

[31] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. *preprint arXiv:1904.07531* (2019).

[32] Joel Rorseth, Parke Godfrey, Lukasz Golab, Mehdi Kargar, Divesh Srivastava, and Jaroslaw Szlichta. 2023. CREDENCE: Counterfactual Explanations for Document Ranking. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*.

[33] J. Rorseth, P. Godfrey, L. Golab, D. Srivastava, and J. Szlichta. 2024. RAGE against the machine: Retrieval-augmented LLM explanations. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 5469–5472.

[34] Alexis Ross, Ana Marasović, and Matthew E. Peters. 2021. Explaining NLP Models via Minimal Contrastive Editing (MiCE). arXiv:2012.13985 [cs.CL]

[35] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*. 793–810.

[36] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human Interpretation of Saliency-based Explanation Over Text. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. 611–636.

[37] Robik Shrestha, Yang Zou, Qiuyu Chen, Zhiheng Li, Yusheng Xie, and Siqi Deng. 2024. FairRAG: Fair Human Generation via Fair Retrieval Augmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 11996–12005.

[38] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings *(KDD '18)*. Association for Computing Machinery, 2219–2228.

[39] Jaspreet Singh, Megha Khosla, Zhenye Wang, and Avishek Anand. 2021. Extracting per query valid explanations for blackbox learning-to-rank models. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR)*. 203–210.

[40] Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. 2024. RAG-Ex: A Generic Framework for Explaining Retrieval Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 2776–2780.

[41] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 3319–3328.

[42] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part I: Score-Based Ranking. *ACM Comput. Surv.* 55, 6, Article 118 (Dec. 2022), 36 pages.

[43] Yuying Zhao, Yu Wang, and Tyler Derr. 2023. Fairness and explainability: bridging the gap towards fair model explanations. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. Article 1275.