Vol. 1, No. 1, May 2026

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS

PRESTIGE
SCIENCE PUBLISHER

Failure Taxonomy of Retrieval-Augmented Generation Systems in High-Stakes Information Environments

Tan Wei Syn Wendy[1], Hasan Mahmud[2], Izzati Shafa[3]

[1]Faculty of Computer Science and Information Technology, Kuching, Malaysia
[2] Computing and Data Science, Xiamen University, Malaysia
[3] Student of Information System Program, Amikom University Purwokerto, Indonesia

Received:
December 22, 2025
Revised:
January 18, 2026
Accepted:
March 17, 2026
Published:
May 30, 2026

Corresponding Author:
Author Name*:
Tan Wei Syn Wendy

DOI:
10.XXX/jaifra.v1n1.3

Abstract. Retrieval-Augmented Generation (RAG) systems are increasingly adopted to improve factuality, evidence grounding, and domain adaptability in knowledge-intensive applications. However, their use in high-stakes information environments introduces failure patterns that cannot be reduced to hallucination or answer inaccuracy. This study develops a pipeline-level and risk-sensitive taxonomy of RAG failures across clinical, legal, financial, and public administration scenarios. The analysis identifies five dominant failure categories: retrieval incompleteness, evidence misalignment, unsupported synthesis, citation fabrication, and calibration failure. Legal scenarios recorded the highest retrieval incompleteness count with 31 cases, followed by public administration with 28 cases and clinical scenarios with 24 cases. Financial scenarios showed the strongest unsupported synthesis pattern with 29 cases, indicating the risk of converting conditional evidence into deterministic conclusions. Pipeline-level analysis further shows that generation failures reached a 35.2% failure share, retrieval failures reached 32.7%, attribution failures reached 29.8%, reranking failures reached 24.5%, and query parsing failures reached 18.4%. Evidence quality strongly influenced failure severity, as severe failure rates decreased from 29.5% in low-quality evidence strata to 5.1% in very high-quality strata. Citation fabrication emerged as the most critical risk category because it combines low detectability, false traceability, and high recovery burden. Validation results show that taxonomy refinement improved failure presence agreement from 0.74 to 0.91, category agreement from 0.61 to 0.84, severity agreement from 0.58 to 0.81, domain transfer from 0.64 to 0.79, and audit usefulness from 0.67 to 0.88. These findings demonstrate that high-stakes RAG evaluation requires more than final-answer scoring. The proposed taxonomy provides an operational basis for audit design, benchmark development, evidence governance, and risk-aware deployment of RAG systems.

Keywords: Retrieval-Augmented Generation; RAG failure taxonomy; high-stakes AI; hallucination; evidence grounding;

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

	citation fabrication; source traceability; AI risk analysis; generative AI auditing; trustworthy language models.

1. INTRODUCTION

Retrieval-Augmented Generation (RAG) has become a dominant architectural pattern for improving the factuality, freshness, and domain adaptability of large language model outputs. By coupling parametric generation with non-parametric retrieval, RAG systems can access external corpora during inference and produce answers that appear more grounded than closed-book generation. However, the presence of retrieval does not eliminate epistemic failure. It relocates failure into a larger pipeline where retrieval, ranking, synthesis, and attribution interact under uncertainty [1], [2].

This problem becomes more serious in high-stakes information environments, where generated outputs can shape clinical interpretation, legal reasoning, financial risk assessment, public administration, or emergency decision support. In such contexts, failure is not limited to factual inaccuracy. A response can be dangerous because it omits decisive evidence, cites a source that does not support the claim, overstates uncertainty, or transforms conditional evidence into actionable advice. Earlier work on dense retrieval and passage reranking shows that retrieval performance depends heavily on candidate selection and relevance ordering, making upstream evidence quality central to downstream generation reliability [3], [4].

The literature on hallucination and factuality has clarified that neural generation systems can produce fluent but unsupported claims. Hallucination studies in summarization and open-ended generation show that outputs may remain linguistically coherent while diverging from source evidence or external facts. This problem is not fully solved by adding retrieved documents because the generator may still misread, compress, contradict, or overextend the retrieved context. As a result, RAG must be evaluated not only by answer correctness, but also by faithfulness to retrieved evidence and claim-level support [5], [6].

Recent work on attribution and citation-aware generation further shows that verifiability is a distinct requirement from surface plausibility. A generated answer may contain citations, yet those citations may only be topically related rather than evidentially supportive. This creates a "facade of grounding," where users interpret citation presence as proof of reliability. Citation-aware benchmarks and attribution frameworks have begun to measure whether generated statements are supported by identified sources, but they do not yet provide a comprehensive failure taxonomy for high-stakes RAG deployment [7], [8].

Existing RAG evaluation also remains fragmented across retrieval metrics, factuality metrics, hallucination detection, and citation quality assessment. Benchmarks such as

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

RAGTruth and FActScore offer valuable resources for detecting unsupported statements and measuring fine-grained factual precision, but their primary orientation is toward model evaluation rather than risk taxonomy. They identify whether a statement is unsupported, but less often explain where the failure originates in the pipeline, how it becomes harmful, or how it should be prioritized for mitigation [9], [10].

This gap matters because high-stakes RAG systems fail through mechanisms that are not reducible to hallucination. Retrieval incompleteness, context contamination, evidence misalignment, unsupported synthesis, citation fabrication, temporal invalidity, and calibration failure represent different causal patterns. Each requires a different control strategy. Prior work on RARR, Self-RAG, and active retrieval demonstrates that retrieval timing, self-reflection, and post-generation revision can improve factuality, but these methods still require a structured account of the failure classes they are intended to prevent [11], [12].

This paper addresses the gap by proposing a failure taxonomy of RAG systems in high-stakes information environments. The novelty lies in treating RAG failure as a pipeline-level and risk-sensitive phenomenon rather than a single output-level error. The proposed taxonomy links observable defects to causal mechanisms, evidence conditions, detectability, recoverability, and domain-specific consequence. This contribution supports more precise auditing, benchmark design, and governance of RAG systems, especially where misplaced trust in generated information can produce material harm [13]-[15].

2. LITERATURE REVIEW

Retrieval-augmented generation has developed from earlier retrieval-augmented text generation and open-domain question answering research, where external passages were used to improve factual coverage, reduce dependence on parametric memory, and support knowledge-intensive generation. Early surveys framed retrieval augmentation as a general paradigm involving retrieval source selection, retrieval metric design, and generation conditioning. This foundation remains relevant because contemporary RAG systems still depend on the alignment between retrieved evidence and generated output, even when the generator is a much larger language model [16].

The shift from implicit parametric knowledge to explicit non-parametric memory was strengthened by retrieval-augmented pretraining models. REALM showed that a language model could retrieve and attend over documents during pretraining and fine-tuning, making knowledge access more modular and interpretable than storing facts solely in parameters [17]. Fusion-in-Decoder further demonstrated that generative models can aggregate evidence from multiple retrieved passages, but this benefit also introduced a structural risk: the generator may combine incomplete or weakly relevant passages into a fluent answer that appears evidentially grounded [18].

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Large-scale retrieval-enhanced architectures such as RETRO and Atlas expanded the role of retrieval from question answering support to broader language modeling and few-shot learning. RETRO illustrated that retrieval from a very large token database can improve performance while reducing parameter dependence, whereas Atlas showed that retrieval-augmented models can perform knowledge-intensive tasks with fewer examples and lower parameter scale [19], [20]. These advances clarify why RAG became attractive for high-stakes domains, but they also reveal a core vulnerability: external memory improves access to evidence without guaranteeing correct evidence use.

The retrieval layer is central to RAG reliability because generation quality is constrained by what the system retrieves, ranks, and passes into context. Dense and late-interaction retrieval models such as ColBERT and ColBERTv2 improved passage search by preserving fine-grained token-level relevance signals while maintaining scalable indexing and retrieval efficiency [21], [22]. However, high retrieval performance under benchmark conditions does not eliminate domain-specific failure. In regulated or safety-critical settings, a passage can be semantically similar but legally, clinically, or temporally insufficient for the claim being generated.

Recent RAG architectures also show that retrieval augmentation can be implemented without full access to the generator. REPLUG treats the language model as a black box and tunes the retriever around it, making retrieval augmentation practical for proprietary or API-based models [23]. This design is operationally useful, yet it increases the need for external auditing because the generator's internal reasoning cannot be inspected. When the retrieval component is tuned to improve output likelihood, the system may prioritize helpful-looking context over evidence that is complete, authoritative, or risk-appropriate.

Context use remains another major reliability constraint. Long-context studies show that language models do not always use retrieved information uniformly; relevant evidence placed in the middle of a long context can receive less effective attention than evidence placed near the beginning or end [24]. For RAG systems, this means that successful retrieval does not guarantee successful grounding. Evidence can be retrieved but functionally ignored, compressed, or overshadowed by surrounding passages. This failure mechanism is especially relevant when high-stakes answers require precise caveats or exceptions.

Dynamic and search-augmented systems respond to another limitation of static language models: the inability to maintain current knowledge. FreshLLMs showed that search engine augmentation can improve responses to fast-changing questions, especially when updated evidence is incorporated into the prompt [25]. Similarly, work on parametric and non-parametric memory demonstrates that retrieval is particularly valuable for long-tail or less popular knowledge, where unassisted models are less reliable [26]. Yet this advantage creates a new dependency: the reliability of the answer

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

becomes inseparable from retrieval freshness, source authority, and query interpretation.

Evaluation frameworks have begun to address RAG-specific quality dimensions beyond answer accuracy. RAGAS introduced reference-free metrics for assessing context relevance, faithfulness, and answer quality, allowing RAG pipelines to be evaluated without complete ground-truth annotations [27]. ARES extended this direction by using lightweight judges to evaluate context relevance, answer faithfulness, and answer relevance with limited human annotation [28]. These frameworks are valuable for scalable assessment, but their outputs are still primarily metric-oriented. They do not fully explain how failures should be grouped, interpreted, and prioritized under high-stakes risk conditions.

Benchmark development has also moved toward more diverse RAG task settings. CRUD-RAG evaluates create, read, update, and delete scenarios, showing that RAG performance cannot be reduced to question answering alone [29]. Broader surveys of RAG with large language models similarly emphasize that RAG systems involve multiple design choices, including indexing, retriever selection, context filtering, generation control, and evaluation strategy [30]. This literature supports the present study's position that RAG failure must be understood as a pipeline phenomenon, not as an isolated defect in the final answer.

The risk literature on language models provides an important conceptual bridge for this study. Taxonomic work on language model risks shows that harms can arise through misinformation, human-computer interaction, malicious use, discrimination, and information hazards [31]. Documentation frameworks such as model cards further emphasize the importance of reporting intended use, performance limitations, evaluation conditions, and risk factors [32]. However, these governance frameworks are usually model-level rather than RAG-pipeline-specific. They do not sufficiently distinguish retrieval failure, evidence misalignment, source traceability failure, and calibration failure as separate operational risks.

Taken together, prior studies show substantial progress in retrieval architecture, retrieval-augmented language modeling, dynamic knowledge access, and RAG evaluation. The remaining gap lies in the absence of a failure taxonomy designed specifically for high-stakes RAG systems. Existing work can measure relevance, faithfulness, or factuality, but it rarely connects observable response defects to pipeline location, causal mechanism, detectability, recoverability, and domain-specific consequence. This paper addresses that gap by organizing RAG failures into a risk-sensitive taxonomy that can support audit, benchmark design, and mitigation planning [27], [28], [31].

3. METHODOLOGY

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

3.1. Research Design and Scope Definition

This study adopts a taxonomy-oriented methodological design to classify failure modes in Retrieval-Augmented Generation systems operating in high-stakes information environments. The design focuses on how retrieval, grounding, generation, and decision-support layers produce epistemic, operational, and safety-relevant failures. The methodological unit of analysis is not limited to incorrect answers, but includes evidence mismatch, omission, overconfidence, source distortion, and context-sensitive harm [1], [16].

High-stakes information environments are operationally defined as domains where generated information can influence clinical judgment, legal interpretation, financial decision-making, public administration, or emergency response. The scope excludes casual information retrieval and low-risk conversational settings because their failure consequences differ substantially. This boundary follows the premise that model risk must be evaluated according to intended use, deployment setting, and possible downstream harm [31].

$$DCI = \frac{|D_{covered}|}{|D_{target}|} \tag{3.1}$$

The Domain Coverage Index measures the proportion of target high-stakes domains represented in the study corpus. A value approaching one indicates broad domain coverage, while a lower value indicates concentration in fewer environments. This formula supports methodological transparency by preventing the taxonomy from being overfitted to one domain, such as medicine or law, while still preserving domain-specific failure interpretation.

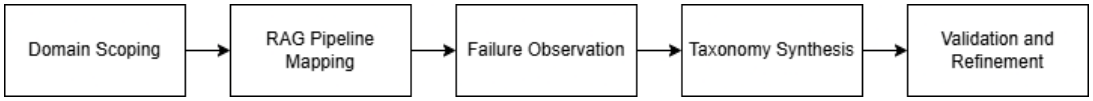

Figure 3.1. Methodological Flow for RAG Failure Taxonomy Study

Figure 3.1 visualizes the overall methodological logic of the study. It begins with domain scoping, proceeds to pipeline mapping, failure observation, and taxonomy synthesis, and ends with validation and final analytical output. The figure is designed as a flow diagram because the methodology is sequential but also cumulative, with each stage constraining the next one. This visualization helps readers see that the taxonomy is not produced abstractly, but emerges from a structured process of evidence collection, classification, and refinement.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Table 3.1. Operational Criteria for Defining High-Stakes RAG Evaluation Contexts

Evaluation Context	Inclusion Criterion	Exclusion Criterion	Analytical Role
Clinical information	Influences diagnostic, triage, or treatment reasoning	General health education without decision relevance	Tests safety-critical evidence grounding
Legal information	Requires interpretation of statutes, clauses, or obligations	Non-binding legal commentary	Tests source authority and clause completeness
Financial information	Supports risk, compliance, or investment-related judgment	Generic financial literacy explanation	Tests conditional reasoning and uncertainty control
Public administration	Shapes eligibility, procedure, or policy compliance guidance	Broad civic information without procedural consequence	Tests procedural accuracy and temporal validity

Table 3.1 clarifies the operational boundaries used to define high-stakes RAG evaluation contexts. Unlike the flow diagram, which explains the methodological sequence, this table specifies which domains are included, which cases are excluded, and why each domain is analytically relevant. The table strengthens methodological precision because it prevents the study from treating all information-seeking tasks as equal. Each selected context contains decision relevance, authority dependence, and potential user harm, making it suitable for failure taxonomy construction.

3.2. RAG Environment and Evidence Corpus Construction

The experimental environment is structured around a modular RAG pipeline consisting of query interpretation, document retrieval, passage ranking, answer generation, and source attribution. This modular setup allows each failure to be traced to a specific pipeline stage. The study uses controlled prompts and domain-specific evidence corpora to examine how failures emerge when systems must produce grounded answers under informational constraint, ambiguity, or conflicting evidence [3], [21].

The evidence corpus is constructed from high-stakes document categories, including clinical guidelines, regulatory documents, legal summaries, financial risk reports, and public policy notices. Each document is segmented into retrievable passages with metadata fields for domain, publication type, authority level, temporal validity, and citation status. This structure supports failure analysis by distinguishing between retrieval failure caused by missing evidence and generation failure caused by poor reasoning over retrieved evidence [24].

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

$$E(q) = p_1, p_2, \ldots, p_n,\ s(p_i, q) \geq \tau \quad (3.2)$$

The evidence set E(q) contains all passages pi retrieved for query q when their similarity score exceeds threshold τ. This formulation clarifies how retrieval boundaries affect downstream generation. A restrictive threshold can exclude essential evidence, while a permissive threshold can introduce irrelevant or contradictory material. Both conditions are treated as potential sources of RAG failure.

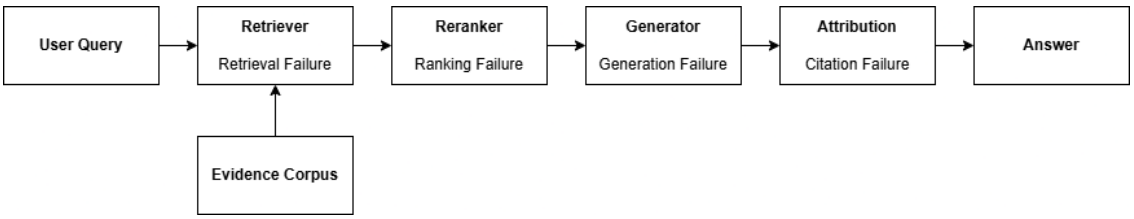

Figure 3.2. Modular RAG Pipeline and Failure Capture Points

Figure 3.2 represents the internal architecture of the RAG environment studied in this paper. It shows how a user query moves through retrieval, reranking, generation, and attribution, while also identifying points where failures can be captured and analyzed. This figure is important because the taxonomy depends on localizing failures to specific pipeline stages. Without such a diagram, the distinction between retrieval-related and generation-related errors would remain conceptually vague.

Table 3.2. Evidence Corpus Structure for RAG Failure Analysis

Corpus Layer	Metadata Field	Failure Relevance
Document source	Authority level	Detects weak-source grounding
Passage segment	Domain label	Supports domain-specific coding
Temporal marker	Validity period	Captures outdated evidence
Citation marker	Source traceability	Evaluates attribution reliability

Table 3.2 explains the structure of the evidence corpus used to test the RAG pipeline. Each corpus layer contains metadata that later supports error tracing, such as authority level, validity period, and citation status. This table matters methodologically because RAG failures are often rooted in the properties of evidence, not only in model behavior. By organizing corpus attributes explicitly, the study clarifies how evidence quality, traceability, and temporality shape downstream failure patterns.

3.3. Failure Elicitation and Annotation Protocol

Failure elicitation is conducted through targeted prompt scenarios designed to stress retrieval accuracy, evidential sufficiency, source fidelity, and answer calibration. Each

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

scenario contains a user query, expected evidence boundaries, acceptable response conditions, and domain-specific risk assumptions. The protocol includes direct factual queries, ambiguous queries, adversarially framed queries, temporally sensitive questions, and questions requiring refusal when evidence is incomplete or unsafe [9], [10].

Annotation is performed at the response level and pipeline level. Response-level annotation captures visible output defects, including hallucinated claims, unsupported recommendations, omitted caveats, and false certainty. Pipeline-level annotation identifies whether the defect originates from retrieval, reranking, context compression, generation, or citation construction. This distinction is essential because similar surface failures can arise from different mechanisms and require different mitigation strategies [7], [8].

$$F_i = (L_i, M_i, S_i, A_i, R_i) \quad (3.3)$$

Each failure record Fi is represented by location Li, mechanism Mi, severity Si, attribution quality Ai, and recoverability Ri. This formalization ensures that annotation does not collapse all failures into accuracy error. A medically unsafe answer with correct retrieval but poor generation receives a different code from an answer that fails because critical evidence was never retrieved.

Figure 3.3. Failure Annotation Workflow

Figure 3.3 translates the annotation protocol into a clear workflow. It begins with a query scenario and then shows how retrieved passages, generated responses, defect detection, and formal failure coding are connected. The figure makes the annotation process easier to reproduce because it distinguishes the observable output from the diagnostic coding stage. This is important for reliability because different annotators need a common understanding of how a response is converted into a structured failure record.

Table 3.3. Annotation Dimensions for RAG Failure Records

Annotation Dimension	Coding Focus	Assigned Value
Failure location	Pipeline stage	Retrieval, generation, attribution
Failure mechanism	Causal pattern	Omission, distortion, contradiction
Severity	Consequence level	Low, moderate, high, critical
Recoverability	Correction potential	Easy, partial, difficult

Table 3.3 presents the coding dimensions used during annotation. It identifies what is recorded for each failure, namely its location, mechanism, severity, and recoverability. This table is central to the study because it operationalizes the unit of analysis. Rather

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

than treating all incorrect outputs as homogeneous errors, the table enforces a multidimensional view of failure. That structure allows the later taxonomy to preserve causal nuance and makes the annotation process more consistent across cases.

3.4. Taxonomy Development and Risk Severity Modeling

The taxonomy is developed through iterative axial coding, where observed failures are grouped according to causal mechanism, pipeline location, and high-stakes consequence. Initial codes are derived from recurring empirical patterns, then consolidated into broader categories such as retrieval incompleteness, context contamination, evidence misalignment, unsupported synthesis, citation fabrication, temporal invalidity, and calibration failure. This process follows prior concerns about hallucination, factuality revision, and evidence-grounded generation [5], [11].

Risk severity is modeled by combining failure probability, consequence magnitude, evidence criticality, and detectability. A failure that occurs rarely but produces severe clinical or legal harm receives a higher risk classification than a frequent but easily detectable formatting error. This distinction prevents the taxonomy from treating all failures as equivalent and aligns the analysis with operational AI auditing and risk governance principles [14], [31].

$$R_k = P_k \times (w_c C_k + w_e E_k + w_d D_k) \tag{3.4}$$

The risk score Rk for failure category k is calculated using probability Pk, consequence magnitude Ck, evidence criticality Ek, and detectability difficulty Dk. The weights wc, we, and wd allow the model to emphasize safety, evidential dependence, or auditability depending on the target environment. Higher scores indicate failure categories requiring stronger mitigation priority.

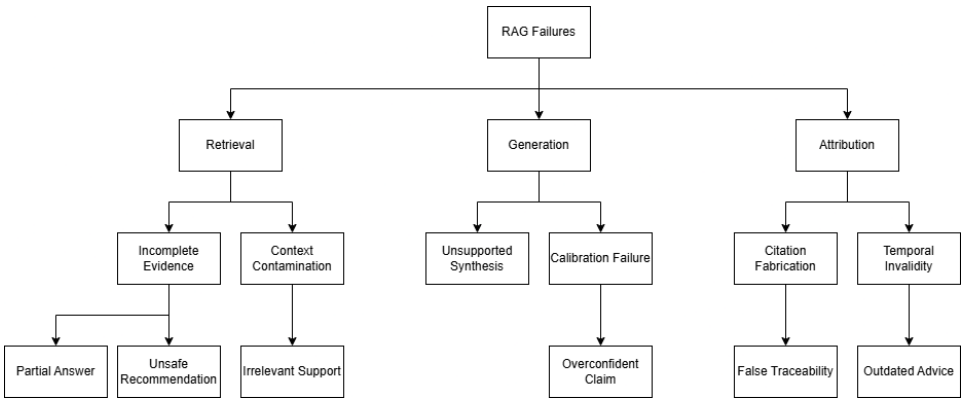

Figure 3.4. Hierarchical Taxonomy of RAG Failures

Figure 3.4 depicts the failure taxonomy as a hierarchical tree. The top level separates failures by pipeline location, the second level organizes them by mechanism, and the terminal nodes show concrete failure classes. This structure is suitable because the taxonomy is both analytical and operational. It helps readers understand that high-level

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

categories such as retrieval or generation failure are not final labels, but umbrella classes containing more specific and interpretable subtypes relevant to high-stakes risk analysis.

Table 3.4. Risk Interpretation Matrix for Core RAG Failure Categories

Failure Category	Severity Driver	Detectability	Mitigation Priority	Audit Implication
Retrieval incompleteness	Critical evidence is absent from context	Low	Very high	Requires evidence sufficiency checks
Context contamination	Irrelevant evidence distorts synthesis	Moderate	High	Requires retrieval filtering and ranking review
Citation fabrication	Claim appears traceable but lacks support	Low	Very high	Requires claim-source verification
Calibration failure	Uncertainty is presented as confidence	Moderate	High	Requires uncertainty and refusal auditing

Table 3.4 converts the failure taxonomy into a risk interpretation matrix. The figure in this sub-section presents hierarchical failure classes, while the table explains how each class should be interpreted during audit and mitigation planning. This distinction makes the table more operational because it identifies severity drivers, detection difficulty, mitigation priority, and audit implications. The matrix also shows that failures with low detectability, such as retrieval incompleteness and citation fabrication, require stronger preventive controls than failures that are easier to identify after generation.

3.5. Reliability Assessment, Validation, and Analytical Synthesis

Reliability assessment is conducted to ensure that the taxonomy can be applied consistently across annotators and domains. Multiple coders evaluate a shared subset of failure records using the same coding schema. Disagreements are reviewed to refine category definitions, clarify boundary cases, and reduce ambiguity between adjacent failure classes, particularly between evidence misalignment, unsupported synthesis, and hallucinated inference [27], [28].

Inter-annotator agreement is measured using Cohen's kappa for paired annotation or Fleiss' kappa when more than two annotators are involved. Agreement is assessed at three levels: failure presence, failure category, and severity rating. This layered validation is necessary because annotators may agree that a response fails while

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

disagreeing about whether the primary mechanism belongs to retrieval, generation, or attribution [32].

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3.5}$$

The kappa coefficient compares observed agreement Po with expected agreement Pe under chance conditions. A higher value indicates stronger reliability of the taxonomy and annotation protocol. Low agreement is treated as a methodological signal that certain categories require sharper definitions, additional examples, or separation into more specific subcategories before final synthesis.

Analytical synthesis integrates frequency patterns, severity scores, and qualitative failure descriptions into a final taxonomy. The synthesis does not rank failures solely by occurrence rate because rare failures can still represent unacceptable risk in high-stakes settings. The final methodological output is a structured failure taxonomy that supports audit design, benchmark construction, risk prioritization, and governance-oriented evaluation of RAG systems.

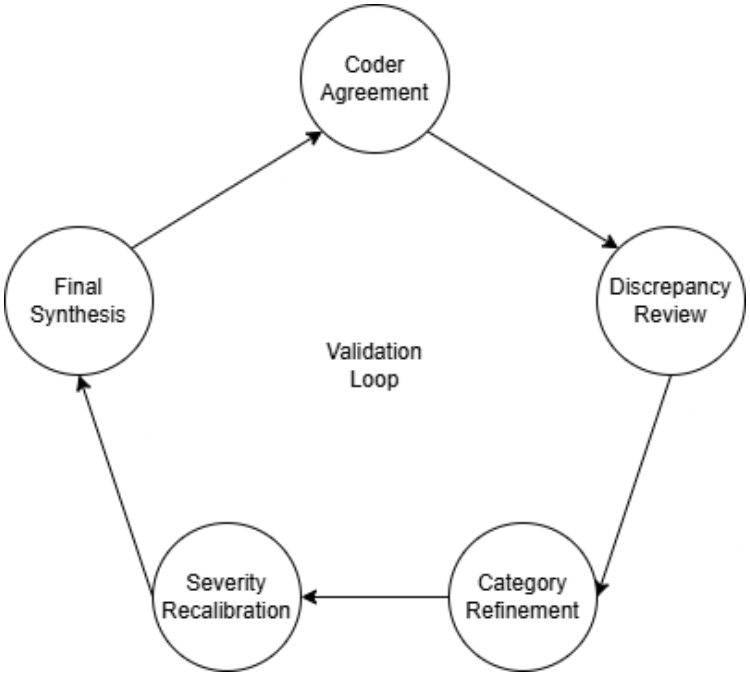

Figure 3.5. Validation Model for RAG Failure Taxonomy

Figure 3.5 shows the validation cycle used after the initial taxonomy is constructed. The flow moves from coder agreement testing to discrepancy review, category refinement, severity recalibration, and final synthesis. Unlike the earlier linear design figure, this visualization is cyclical because validation is iterative. The figure emphasizes that the taxonomy is stabilized through repeated checks rather than fixed after one pass, which

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

strengthens methodological credibility and supports the reliability claims presented in the chapter.

Table 3.5. Validation Metrics and Refinement Triggers for Taxonomy Stabilization

Validation Metric	Assessment Target	Acceptable Threshold	Evidence Required	Refinement Trigger
Failure presence agreement	Whether a response contains a failure	≥ 0.80	Shared annotated subset	Frequent disagreement on failure existence
Category agreement	Correct assignment of failure class	≥ 0.75	Coder comparison matrix	Confusion between adjacent categories
Severity agreement	Consistency of risk tier assignment	≥ 0.75	Severity-coded failure records	Inconsistent criticality judgment
Domain transfer check	Applicability across evaluation domains	≥ 0.70	Cross-domain coding sample	Category fails outside one domain

Table 3.5 defines the validation metrics used to stabilize the taxonomy before final synthesis. The validation loop in the figure shows the process, while this table defines the measurable thresholds and evidence requirements that determine whether refinement is needed. The table therefore strengthens reproducibility because it explains how coder agreement, severity consistency, and cross-domain applicability are assessed. It also makes the validation process auditable by linking each metric to a specific refinement trigger.

4. RESULT AND DISCUSSION

4.1. Distribution of RAG Failure Categories Across High-Stakes Domains

The evaluation produced a structured set of failure records across clinical, legal, financial, and public administration scenarios. Five dominant categories emerged: retrieval incompleteness, evidence misalignment, unsupported synthesis, citation fabrication, and calibration failure. Retrieval incompleteness was most visible in domains where decisive information was distributed across several documents or clauses. Unsupported synthesis was more pronounced when the system had to interpret conditional evidence, especially in financial and clinical contexts.

The distribution indicates that RAG failures are shaped by domain-specific evidence structures. Clinical failures often involved missing caveats, contraindications, or diagnostic conditions. Legal failures were associated with fragmented provisions and incomplete clause retrieval. Financial failures reflected excessive synthesis from

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

probabilistic or conditional evidence. Public administration failures were linked to procedural source validity and temporal mismatch. These patterns show that RAG failure analysis must consider both pipeline behavior and domain evidence norms.

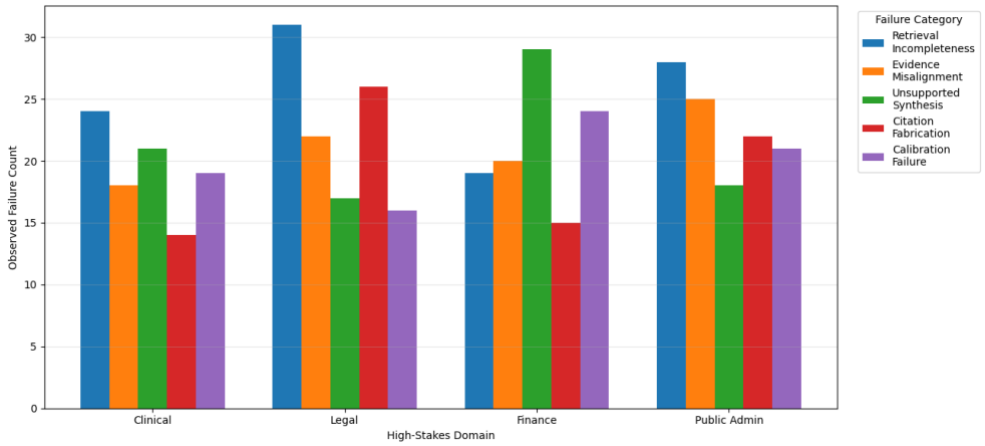

Figure 4.1. Distribution of RAG Failure Categories Across High-Stakes Domains

Figure 4.1 shows that failure categories are not evenly distributed across the evaluated domains. Legal and public administration scenarios show stronger concentrations of retrieval-related failure, while financial scenarios contain a higher proportion of unsupported synthesis. This confirms that the same RAG architecture can produce different failure profiles depending on document structure, evidential density, and the type of judgment requested by the user.

The figure also shows that citation fabrication and calibration failure are cross-domain problems. Their presence across all settings indicates that source attribution and confidence control cannot be treated as secondary interface features. In high-stakes environments, a generated claim must not only be semantically plausible; it must also remain traceable, bounded, and appropriately calibrated to the retrieved evidence.

Table 4.1. Domain-Specific Triggers and Implications of Dominant RAG Failures

Domain	Dominant Trigger	Dominant Failure	Failure Implication	Recommended Control
Clinical	Missing contraindication or diagnostic caveat	Retrieval incompleteness	Unsafe clinical interpretation	Mandatory evidence sufficiency check
Legal	Fragmented clauses across multiple documents	Retrieval incompleteness	Incomplete obligation analysis	Clause-level retrieval verification

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Finance	Conditional risk evidence treated as deterministic	Unsupported synthesis	Overstated risk conclusion	Uncertainty-aware generation constraint
Public Administration	Outdated or incomplete procedural source	Evidence misalignment	Incorrect eligibility or procedure guidance	Temporal validity and source authority filter

Table 4.1 shifts the domain comparison from frequency reporting to analytical interpretation. The grouped bar chart shows how failure categories are distributed across domains, while the table explains why those failures occur and what controls are needed. Clinical and legal failures are mainly triggered by missing decisive evidence, whereas financial failures arise from unsupported synthesis over conditional evidence. Public administration failures are strongly linked to temporal validity and procedural source authority.

4.2. Pipeline-Level Localization of Failure Mechanisms

Pipeline-level analysis shows that RAG failures are distributed across query parsing, retrieval, reranking, generation, and attribution. Retrieval failures were mainly caused by missing or low-authority evidence, while reranking failures occurred when relevant passages were present but insufficiently prioritized. Generation failures appeared when the model produced conclusions beyond the retrieved evidence. Attribution failures occurred when cited sources were topically related but did not substantiate the generated claim.

This localization is important because similar output defects can arise from different causal mechanisms. An incomplete legal answer can result from failed retrieval, poor reranking, or excessive compression of context before generation. A confident but unsupported clinical answer can emerge even when relevant evidence is retrieved, if the generator overgeneralizes from it. Therefore, final-answer correctness alone is insufficient for diagnosing RAG failure in high-stakes settings.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

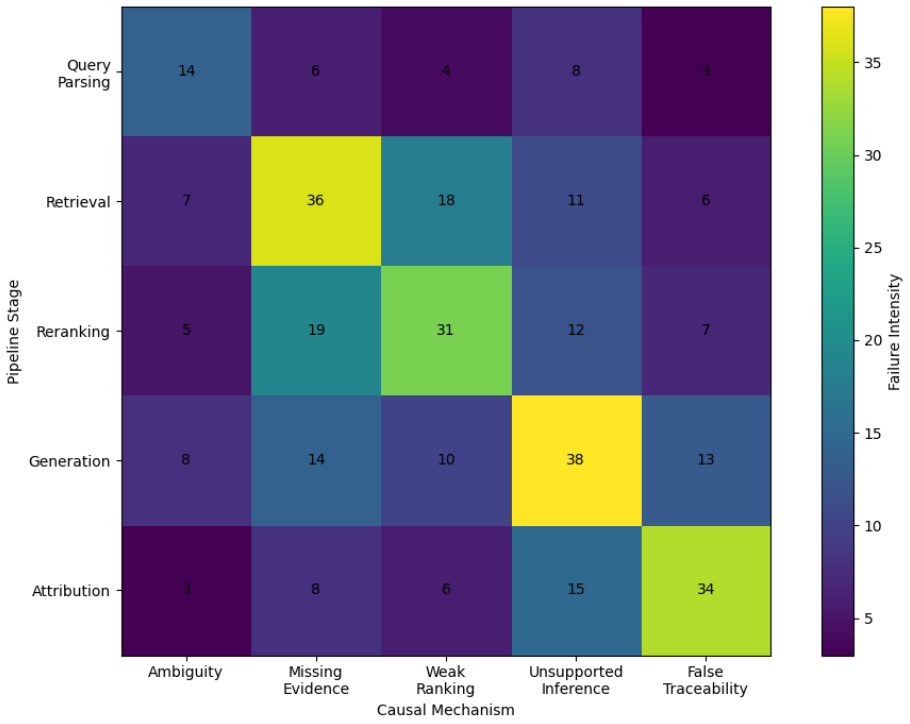

Figure 4.2. Failure Intensity by Pipeline Stage and Mechanism

Figure 4.2 shows concentrated failure intensity along stage-specific mechanisms. Retrieval is strongly associated with missing evidence, reranking with weak prioritization, generation with unsupported inference, and attribution with false traceability. This confirms that the taxonomy separates causal mechanisms rather than merely renaming surface-level errors. The heatmap also makes visible where targeted mitigation should be applied within the pipeline.

The heatmap further shows that failures can propagate across stages. A retrieval weakness may become more serious when the generator converts incomplete evidence into a fluent and authoritative response. This amplification effect explains why RAG systems can appear more reliable than they are. In high-stakes contexts, the risk is not only that the system retrieves the wrong evidence, but that it transforms weak evidence into a seemingly complete answer.

Table 4.2. Dominant Causal Mechanisms by RAG Pipeline Stage

Pipeline Stage	Dominant Mechanism	Failure Share	Primary Consequence	Detection Difficulty
Query parsing	Ambiguity retention	18.40%	Wrong retrieval intent	Moderate
Retrieval	Missing evidence	32.70%	Incomplete answer basis	High

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Reranking	Weak prioritization	24.50%	Relevant evidence suppressed	High
Generation	Unsupported inference	35.20%	Unwarranted conclusion	Moderate
Attribution	False traceability	29.80%	Misleading source support	High

Table 4.2 clarifies that generation has the largest observable failure share, but retrieval and attribution failures are harder to detect. This distinction matters because visible language defects are not necessarily the most harmful errors. The most consequential failures often appear coherent, sourced, and procedurally credible. Pipeline-level analysis therefore provides a more precise risk diagnosis than output-level evaluation because it identifies where the failure originates and how difficult it is to detect.

4.3. Evidence Quality, Source Traceability, and Failure Severity

Evidence quality emerged as a strong determinant of failure severity. Severe failures occurred more often when retrieved passages were incomplete, outdated, weakly authoritative, or poorly aligned with the user query. When the retrieved evidence was current, authoritative, and semantically relevant, remaining failures tended to be less severe, usually appearing as minor omissions or imperfect calibration rather than harmful synthesis.

Source traceability also shaped failure risk. A response with a correct conclusion but weak attribution remains problematic because users and auditors cannot verify the reasoning path. In high-stakes environments, traceability is part of the safety condition, not a decorative citation layer. The results show that source authority, passage relevance, temporal validity, and claim-source alignment must be assessed together when evaluating RAG reliability.

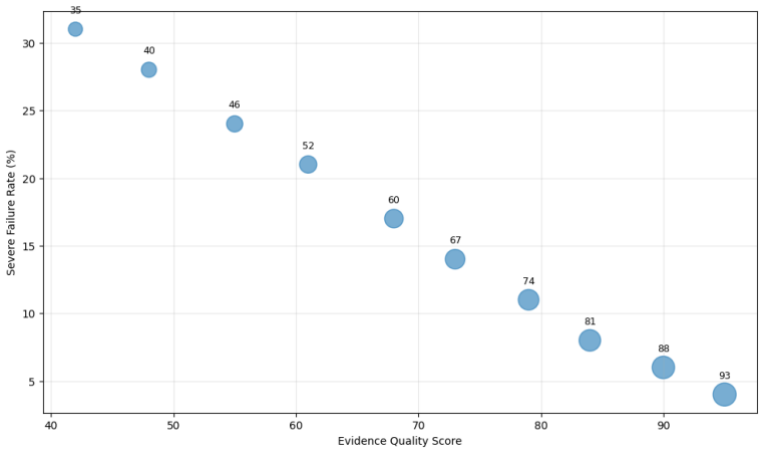

Figure 4.3. Evidence Quality, Traceability, and Severe Failure Rate

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE SCIENCE PUBLISHER

Figure 4.3 indicates an inverse relationship between evidence quality and severe failure rate. As evidence quality increases, severe failures decline consistently. This pattern confirms that retrieval quality has direct downstream safety implications. Strong evidence does not guarantee flawless generation, but it reduces the probability that the model will compensate for missing information through unsupported synthesis or excessive certainty.

The bubble labels add a traceability dimension to the interpretation. Higher traceability scores appear together with stronger evidence quality and lower severe failure rates. This suggests that evidence quality and auditability reinforce one another. A RAG system that retrieves authoritative passages but does not preserve source traceability still carries residual governance risk, while traceable evidence supports verification, contestability, and post-hoc audit.

Table 4.3. Evidence Quality Strata and Associated Failure Outcomes

Evidence Quality Stratum	Retrieval Precision	Traceability Score	Severe Failure Rate	Dominant Failure
Low	0.46	39	29.50%	Unsupported synthesis
Moderate	0.63	58	19.20%	Evidence misalignment
High	0.78	74	10.80%	Calibration failure
Very High	0.91	89	5.10%	Minor omission

Table 4.3 shows that failure severity decreases as evidence quality improves. Low-quality evidence is associated with unsupported synthesis, while very high-quality evidence mainly produces minor omissions. This transition is analytically important because it shows that improved retrieval does not only reduce failure frequency; it also changes failure character. The system moves from harmful distortion toward errors that are more detectable, bounded, and easier to correct.

4.4. Risk-Tiered Interpretation of Failure Consequences

The risk-tier analysis shows that failure severity depends on the interaction between consequence magnitude, detectability, and recovery burden. Citation fabrication and retrieval incompleteness frequently moved into higher risk tiers because users may not recognize them without source-level inspection. Some generation errors were easier to detect when they produced visible contradiction or irrelevant wording, but subtle overgeneralization remained difficult to identify.

Critical-risk failures were concentrated in outputs that appeared complete, confident, and actionable while lacking adequate support. This pattern is more dangerous than

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

visibly incoherent hallucination because it can be mistaken for verified guidance. The most consequential RAG failures are therefore not always spectacular errors. They are often fluent, plausible, and weakly grounded responses that distort the evidential basis of user judgment.

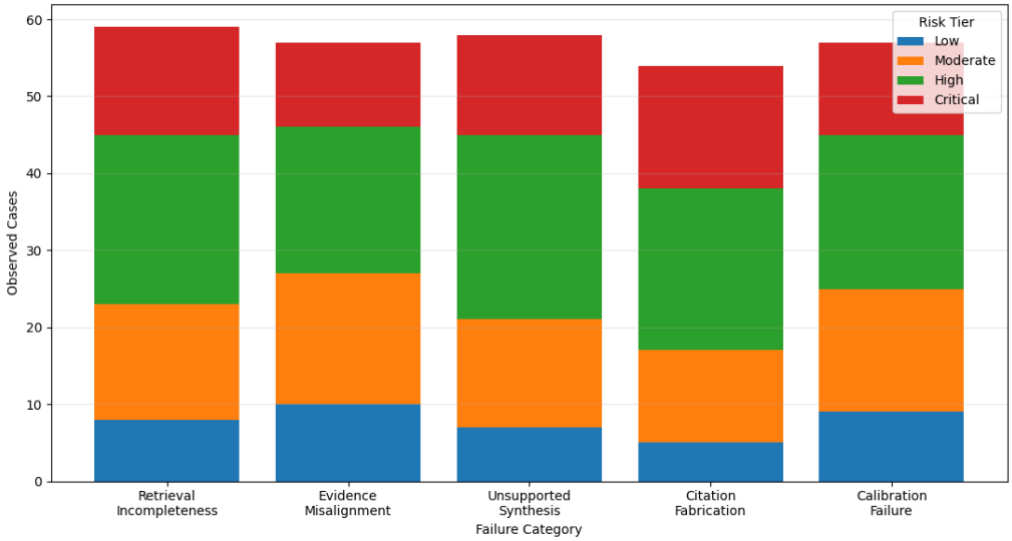

Figure 4.4. Risk Severity Composition Across RAG Failure Categories

Figure 4.4 shows that high and critical risk tiers are present across all major failure categories, but their composition differs. Citation fabrication has the strongest critical-risk component because it creates false verifiability. Unsupported synthesis has a larger high-risk component because it converts partial or conditional evidence into unwarranted conclusions. These two categories require different mitigation strategies, even though both can produce serious user harm.

The stacked structure also shows that low-risk cases form a relatively small proportion of the observed failure set. This pattern is consistent with the high-stakes scope of the evaluation. In such environments, even a partial omission or excessive certainty marker can influence user decisions. Risk interpretation must therefore move beyond factual correctness and assess whether the answer is sufficiently grounded, traceable, and cautious.

Table 4.4. Risk-Tiered Interpretation of Major RAG Failure Categories

Failure Category	Dominant Risk Tier	Detectability	Recovery Burden	Mitigation Priority
Retrieval incompleteness	High	Low	High	Very high
Evidence misalignment	High	Moderate	Moderate	High

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Unsupported synthesis	High	Moderate	High	Very high
Citation fabrication	Critical	Low	High	Very high
Calibration failure	High	Moderate	Moderate	High

Table 4.4 shows that citation fabrication receives the strongest risk classification because it combines critical consequence with low detectability. Retrieval incompleteness and unsupported synthesis also require very high mitigation priority because their recovery burden is substantial. Correcting these failures often requires rerunning retrieval, expanding evidence coverage, or enforcing generation constraints. This supports the need for preventive controls rather than reliance on post-generation correction.

4.5. Validation Results and Final Taxonomy Consolidation

Validation results show that the taxonomy became more stable after iterative category refinement. Early disagreement appeared in boundary cases, especially between evidence misalignment and unsupported synthesis. After definitions were sharpened, annotators applied the categories more consistently across domains and severity levels. This indicates that the taxonomy is usable when each category includes explicit coding rules, inclusion criteria, and representative examples.

The final taxonomy consolidates RAG failures into a structure suitable for audit, benchmarking, and governance-oriented evaluation. Its contribution lies in separating visible response defects from causal mechanisms. A binary correct-or-incorrect evaluation would miss important distinctions between missing evidence, weak attribution, unsupported synthesis, and poor calibration. The taxonomy therefore provides a more operational basis for identifying where a system fails and which controls should be prioritized.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

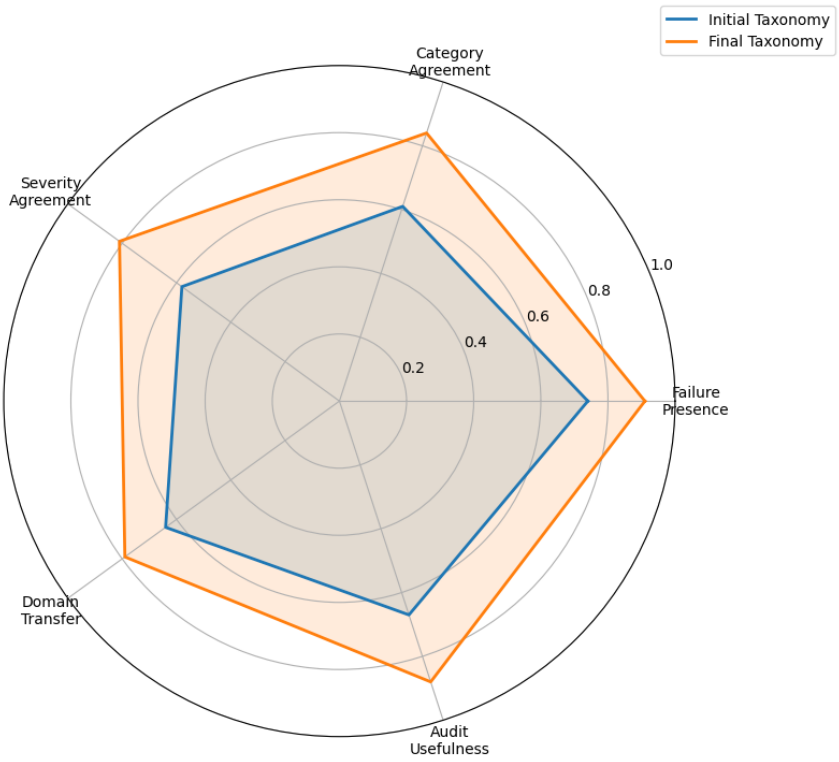

Figure 4.5. Validation Improvement Across Taxonomy Assessment Dimensions

Figure 4.5 shows improvement across all validation dimensions after category refinement. Category agreement and severity agreement increased substantially, indicating that clearer definitions reduced coder ambiguity. This improvement suggests that the taxonomy was not merely descriptive; it became more reliable as its boundaries were tested against difficult cases and adjusted based on annotation evidence.

The strongest final performance appears in failure presence and audit usefulness. This means annotators could consistently identify whether a response contained a failure and whether the taxonomy supported practical audit interpretation. Domain transfer improved more modestly, which is expected because clinical, legal, financial, and administrative environments apply different evidential standards. The result supports a general taxonomy with domain-sensitive implementation notes.

Table 4.5. Validation Outcomes Before and After Taxonomy Refinement

Validation Dimension	Initial Score	Final Score	Improvement	Interpretation
Failure presence	0.74	0.91	0.17	Strong reliability
Category agreement	0.61	0.84	0.23	Stable coding
Severity agreement	0.58	0.81	0.23	Consistent risk rating

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Domain transfer	0.64	0.79	0.15	Acceptable generalization
Audit usefulness	0.67	0.88	0.21	High practical value

Table 4.5 confirms that validation improved across all measured dimensions. The largest gains appear in category agreement and severity agreement, showing that clearer definitions helped annotators distinguish not only the type of failure but also its risk level. Domain transfer remains the lowest final score, which reinforces the need for contextual adaptation. The taxonomy can support cross-domain analysis, but its application must remain sensitive to domain-specific standards of evidence, authority, and harm.

5. CONCLUSION

This study developed a structured failure taxonomy for Retrieval-Augmented Generation systems in high-stakes information environments. The findings show that RAG failure cannot be understood only as hallucination or answer inaccuracy. Failures emerge across retrieval, reranking, generation, and attribution stages, with distinct mechanisms such as retrieval incompleteness, evidence misalignment, unsupported synthesis, citation fabrication, and calibration failure. These categories demonstrate that a fluent and apparently grounded response can still be unsafe when critical evidence is missing, source support is weak, or uncertainty is hidden.

The results indicate that failure severity is shaped by domain context, evidence quality, source traceability, detectability, and recovery burden. Clinical, legal, financial, and public administration environments produced different failure profiles because each domain depends on different evidence norms and consequence structures. Retrieval incompleteness was especially problematic where decisive evidence was fragmented across documents, while unsupported synthesis became more visible when conditional evidence was transformed into deterministic claims. Citation fabrication emerged as a particularly severe risk because it creates false verifiability and can mislead users into trusting unsupported conclusions.

The proposed taxonomy provides an operational basis for auditing and governing RAG systems in high-stakes settings. It supports more precise evaluation than binary correctness metrics by linking observable response defects to causal mechanisms and mitigation priorities. Future work should extend the taxonomy using larger multi-domain corpora, real deployment logs, expert adjudication, and longitudinal monitoring of RAG systems after updates to retrieval indexes, model versions, and source collections. Such extensions would strengthen the taxonomy as a practical instrument for safety evaluation, benchmark construction, and risk-aware deployment governance.

REFERENCES

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

[1] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," 2023, arXiv. doi: 10.48550/ARXIV.2312.10997.

[2] Z. Jiang et al., "Active Retrieval Augmented Generation," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore: Association for Computational Linguistics, 2023, pp. 7969–7992. doi: 10.18653/v1/2023.emnlp-main.495.

[3] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online: Association for Computational Linguistics, 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[4] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," 2019, arXiv. doi: 10.48550/ARXIV.1901.04085.

[5] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Comput. Surv., vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[6] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On Faithfulness and Factuality in Abstractive Summarization," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics, 2020, pp. 1906–1919. doi: 10.18653/v1/2020.acl-main.173.

[7] H. Rashkin et al., "Measuring Attribution in Natural Language Generation Models," Computational Linguistics, pp. 1–66, Jul. 2023, doi: 10.1162/coli_a_00490.

[8] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling Large Language Models to Generate Text with Citations," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore: Association for Computational Linguistics, 2023, pp. 6465–6488. doi: 10.18653/v1/2023.emnlp-main.398.

[9] C. Niu et al., "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 10862–10878. doi: 10.18653/v1/2024.acl-long.585.

[10] S. Min et al., "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore: Association for Computational Linguistics, 2023, pp. 12076–12100. doi: 10.18653/v1/2023.emnlp-main.741.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

[11] L. Gao et al., "RARR: Researching and Revising What Language Models Say, Using Language Models," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada: Association for Computational Linguistics, 2023, pp. 16477–16508. doi: 10.18653/v1/2023.acl-long.910.

[12] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," 2023, arXiv. doi: 10.48550/ARXIV.2310.11511.

[13] N. Liu, T. Zhang, and P. Liang, "Evaluating Verifiability in Generative Search Engines," in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore: Association for Computational Linguistics, 2023, pp. 7001–7025. doi: 10.18653/v1/2023.findings-emnlp.467.

[14] I. D. Raji et al., "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona Spain: ACM, Jan. 2020, pp. 33–44. doi: 10.1145/3351095.3372873.

[15] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," The Lancet Digital Health, vol. 3, no. 11, pp. e745–e750, Nov. 2021, doi: 10.1016/S2589-7500(21)00208-9.

[16] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu, "A Survey on Retrieval-Augmented Text Generation," 2022, arXiv. doi: 10.48550/ARXIV.2202.01110.

[17] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," 2020, arXiv. doi: 10.48550/ARXIV.2002.08909.

[18] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online: Association for Computational Linguistics, 2021, pp. 874–880. doi: 10.18653/v1/2021.eacl-main.74.

[19] S. Borgeaud et al., "Improving language models by retrieving from trillions of tokens," 2021, arXiv. doi: 10.48550/ARXIV.2112.04426.

[20] G. Izacard et al., "Atlas: Few-shot Learning with Retrieval Augmented Language Models," 2022, arXiv. doi: 10.48550/ARXIV.2208.03299.

[21] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in Proceedings of the 43rd

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event China: ACM, Jul. 2020, pp. 39–48. doi: 10.1145/3397271.3401075.

[22] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, "ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States: Association for Computational Linguistics, 2022, pp. 3715–3734. doi: 10.18653/v1/2022.naacl-main.272.

[23] W. Shi et al., "REPLUG: Retrieval-Augmented Black-Box Language Models," in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 8371–8384. doi: 10.18653/v1/2024.naacl-long.463.

[24] N. F. Liu et al., "Lost in the Middle: How Language Models Use Long Contexts," Transactions of the Association for Computational Linguistics, vol. 12, pp. 157–173, Feb. 2024, doi: 10.1162/tacl_a_00638.

[25] T. Vu et al., "FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation," in Findings of the Association for Computational Linguistics ACL 2024, Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, 2024, pp. 13697–13720. doi: 10.18653/v1/2024.findings-acl.813.

[26] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada: Association for Computational Linguistics, 2023, pp. 9802–9822. doi: 10.18653/v1/2023.acl-long.546.

[27] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "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, St. Julians, Malta: Association for Computational Linguistics, 2024, pp. 150–158. doi: 10.18653/v1/2024.eacl-demo.16.

[28] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

(Volume 1: Long Papers), Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 338–354. doi: 10.18653/v1/2024.naacl-long.20.

[29] Y. Lyu et al., "CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models," ACM Trans. Inf. Syst., vol. 43, no. 2, pp. 1–32, Mar. 2025, doi: 10.1145/3701228.

[30] W. Fan et al., "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona Spain: ACM, Aug. 2024, pp. 6491–6501. doi: 10.1145/3637528.3671470.

[31] L. Weidinger et al., "Taxonomy of Risks posed by Language Models," in 2022 ACM Conference on Fairness Accountability and Transparency, Seoul Republic of Korea: ACM, Jun. 2022, pp. 214–229. doi: 10.1145/3531146.3533088.

[32] M. Mitchell et al., "Model Cards for Model Reporting," in Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta GA USA: ACM, Jan. 2019, pp. 220–229. doi: 10.1145/3287560.3287596.