A Systematic Taxonomy of Failure Modes in Retrieval-Augmented Generation Systems

Anupama Garani
Independent Researcher
Austin, TX, USA
anupamagarani95@gmail.com

Abstract

Retrieval-Augmented Generation (RAG) systems fail in diverse, poorly characterized ways that single-stage evaluation metrics cannot detect. We present a systematic taxonomy of 33 failure modes across 7 pipeline stages — ingestion, representation, retrieval, generation, evaluation, deployment, and agentic orchestration — constructed through structured literature review of 48 sources spanning peer-reviewed publications and high-impact preprints. For each mode, we provide a formal definition, observable manifestation, and three-level evidence grading (Strong/Moderate/Limited). Our analysis reveals critical asymmetry in research attention: retrieval and generation failures are comparatively well-studied, while representation, evaluation, and agentic orchestration failures remain under-investigated despite frequent occurrence in production. We identify 12 failure modes with no dedicated peer-reviewed empirical evidence — all 8 agentic modes among them — constituting an evidence desert in the fastest-growing RAG deployment paradigm. Compared to prior work enumerating 7 failure points (Barnett et al., 2024) or 16 error types within partial pipeline runs (Cresswell et al., 2025), our taxonomy uniquely spans the full pipeline including agentic orchestration with explicit evidence-level grading.

Keywords: RAG, retrieval-augmented generation, RAG evaluation, failure analysis, LLM reliability, trustworthy NLP, agentic RAG, production AI systems, failure modes, failure taxonomy

1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as a predominant approach for integrating external knowledge into large language model (LLM) outputs (Lewis et al., 2020; Izacard and Grave, 2021; Gao et al., 2023), enabling enterprise question answering and domain-specific decision support in healthcare, finance, and law (Fan et al., 2024; Gupta et al., 2024). By coupling retrieval with generation, RAG addresses hallucination, knowledge staleness, and domain specificity limitations (Huang and Chang, 2023). However, rapid adoption has outpaced systematic understanding of how these systems fail.

Existing work has identified individual failure points (Barnett et al., 2024; Akkiraju et al., 2024) or narrow categories such as hallucination (Zhang et al., 2025) or retrieval errors (Cresswell et al., 2025), but no unified framework captures failures across the complete pipeline. This gap is consequential: RAG failures are rarely isolated—chunking errors propagate to retrieval failures, which cascade into generation errors, resisting single-point diagnosis. Consequences extend beyond user dissatisfaction: outdated retrieval surfaces superseded clinical guidelines (Zhang et al., 2026), retrieval inconsistency yields contradictory regulatory guidance (Liu et al., 2024), and poor attribution prevents verification of generated citations (S-RAG Team, 2025).

Despite the rapid adoption of RAG systems, the landscape of failure mechanisms remains fragmented across studies, making it difficult for practitioners and researchers to systematically diagnose reliability issues. This fragmentation motivates the need for a structured taxonomy of failure modes that captures the full lifecycle of RAG systems.

Contributions. This paper makes the following contributions:

1. **Comprehensive Taxonomy.** We present a structured taxonomy of 33 failure modes in Retrieval-Augmented Generation (RAG) systems across seven pipeline stages, extending beyond the 7 failure points of Barnett et al. (2024), the 15 control points of Akkiraju et al. (2024), and the 16 error types of Cresswell et al. (2025).

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 413–424

2. **Full-Pipeline Coverage Including Agentic RAG.** Our taxonomy spans ingestion through agentic orchestration, capturing 8 agentic failure modes representing the fastest-growing but least-studied deployment paradigm (Singh et al., 2025).

3. **Evidence-Level Grading.** We classify each failure mode as *Strong*, *Moderate*, or *Limited*, revealing that 12 modes (36%) lack peer-reviewed empirical investigation, including all 8 agentic modes.

4. **Research Road map.** We synthesize evidence asymmetry across pipeline stages to highlight under-explored failure classes requiring empirical validation.

2 Related Work

Prior work on RAG failures falls into three streams. Surveys and architectures (Lewis et al., 2020; Gao et al., 2023; Fan et al., 2024; Gupta et al., 2024; Manathunga and Illangasekara, 2025; Singh et al., 2025) have extensively studied RAG improvement but not systematic failure analysis. Evaluation frameworks including RAGAS (Es et al., 2024), RAGChecker (Ru et al., 2024), and ARES (Saad-Falcon et al., 2023) measure performance but do not characterize failure mechanisms. Closest to our work, Barnett et al. (2024) enumerate 7 failure points from engineering experience, Akkiraju et al. (2024) extend this to 15 control points, and Cresswell et al. (2025) present 16 empirically-validated error types spanning chunking through generation. As shown in Table 1, no prior work spans the full pipeline including agentic orchestration with explicit evidence-level grading.

3 Background

A standard RAG system operates across three stages: an indexing pipeline that ingests, chunks, embeds, and stores documents; a retrieval pipeline that encodes queries and returns top-k chunks; and a generation pipeline that produces grounded responses. Following Gao et al. (2023), we distinguish Naive, Advanced, Modular, and Agentic RAG paradigms (Singh et al., 2025), each with distinct failure profiles. Critically, RAG failures rarely occur in isolation—chunking errors propagate to retrieval failures, which make hallucination structurally inevitable (Cresswell et al., 2025; Ru et al., 2024). Figure 1 illustrates these pipeline stages with associated failure modes.

4 Methodology

This taxonomy was constructed through a structured literature review and systematic analysis of reported RAG system failures.

4.1 Literature Search and Selection

We conducted a structured literature review (January 2025–February 2026) searching ACL Anthology, IEEE Xplore, Semantic Scholar, and arXiv using terms including "RAG failure," "RAG evaluation," and "RAG robustness." The search covered major NLP and machine learning venues including ACL, EMNLP, NeurIPS, KDD, and SIGKDD, as well as high-impact preprints and domain-specific journals in healthcare and finance. Papers were included if they described, evaluated, or diagnosed Retrieval-Augmented Generation failures, or proposed benchmarks that revealed RAG system vulnerabilities. Papers were excluded if they focused solely on generic large language model hallucination without involving retrieval components. From 48 included papers, we extracted failure instances and organized them thematically into candidate modes grouped by pipeline stage. These papers included both peer-reviewed conference publications and high-impact preprints widely cited within the RAG research community. Throughout Section 5 we explicitly note when supporting evidence for a failure mode rests on preprints or non-peer-reviewed sources, particularly for modes assigned Limited evidence. A *failure mode* is a distinct causal mechanism; sub-variants sharing the same mechanism were consolidated.

4.2 Evidence Classification

Each failure mode was assigned an evidence level based on the type of supporting research available in the literature.

Strong indicates a dedicated peer-reviewed empirical study that isolates the failure mechanism quantitatively.

Moderate indicates supporting evidence from broader studies, benchmark analyses, or practitioner reports.

Limited indicates architectural analysis or insights from adjacent fields where no dedicated RAG benchmark currently exists.

Evidence levels were assigned by the lead author

Work	Failures	Pipeline Coverage	Evidence Type	Agentic
Barnett et al. (2024)	7	Retrieval + Generation	Case studies	No
Akkiraju et al. (2024)	15 points	Full pipeline	Case studies	No
Cresswell et al. (2025)	16 types	Chunking → Generation	Empirical + tool	No
RAGChecker (Ru et al., 2024)	Metrics	Retrieval + Generation	Benchmark	No
De Lima et al. (2024)	Dataset taxonomy	Retrieval-focused	Dataset analysis	No
This work	**33**	**Full pipeline + Agentic**	**48 sources, graded**	**Yes**

Table 1: Comparison with existing RAG failure analyses.

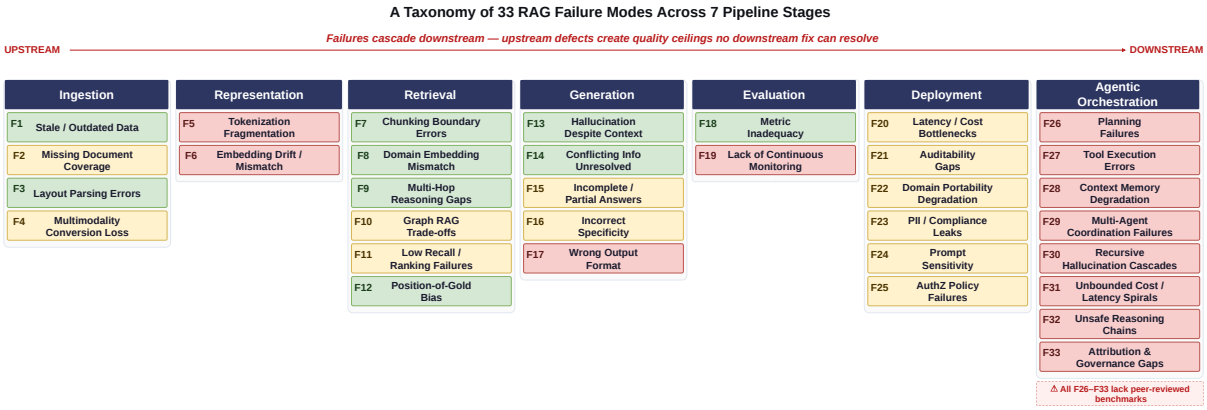

Figure 1: A taxonomy of 33 RAG failure modes across 7 pipeline stages.
Card shading: Strong, Moderate, Limited evidence.

through structured re-reading of each cited source against the three-level rubric. To mitigate single-rater bias, we applied conservative downgrading: a mode was assigned *Strong* only when at least one peer-reviewed study isolated the failure mechanism quantitatively; ambiguous cases defaulted to *Moderate*; and modes resting only on architectural reasoning, practitioner reports, or evidence from adjacent fields were assigned *Limited* regardless of the volume of supporting material. This conservative protocol biases the taxonomy toward under-claiming evidence strength rather than over-claiming. This evidence grading approach is inspired by evidence classification practices commonly used in systematic literature reviews in software engineering and applied AI research.

5 A Taxonomy of RAG Failure Modes

We organize failure modes according to major stages of the RAG pipeline, including representation, retrieval, generation, evaluation, and agentic orchestration components. We present 33 failure modes organized into 7 categories. Table 2 summarizes the complete taxonomy of 33 failure modes and their corresponding evidence levels. The following subsections detail each category, including definitions, observable manifestations, and evidence-levels.

5.1 Ingestion and Data Failures (F1–F4)

Ingestion failures corrupt or degrade the knowledge base before any retrieval or generation takes place, creating a performance ceiling that downstream components cannot overcome.

F1. Outdated/Stale Data. The knowledge base contains documents that are no longer current, causing responses based on obsolete information. Ouyang et al. (2025) developed HoH, the first benchmark specifically measuring how outdated information degrades RAG performance, demonstrating that stale data both reduces accuracy and can mislead models into generating harmful outputs even when current information is available. Multiple surveys recognize dynamic knowledge management as a fundamental RAG challenge (Gao et al., 2023; Fan et al., 2024; Gao et al., 2024). *Evidence: Strong.*

F2. Missing/Incomplete Documents. The knowledge base lacks comprehensive domain coverage, creating gaps where relevant information does not exist in the retrieval corpus. Barnett et al. (2024) identify "Missing Content" as a primary failure point. The RARE benchmark (Zeng et al.,

ID	Category	Failure Mode	Evidence
F1	Ingestion	Outdated/Stale Data	Strong
F2	Ingestion	Missing/Incomplete Documents	Moderate
F3	Ingestion	Layout Parsing Errors	Strong
F4	Ingestion	Multimodality Conversion Loss	Moderate
F5	Representation	Tokenization Fragmentation	Limited
F6	Representation	Embedding Drift / Model Mismatch	Limited
F7	Retrieval	Chunking Boundary Errors	Strong
F8	Retrieval	Domain Embedding Mismatch	Strong
F9	Retrieval	Multi-Hop Reasoning Gaps	Strong
F10	Retrieval	Graph RAG Trade-offs	Moderate
F11	Retrieval	Low Recall / Ranking Failures	Moderate
F12	Retrieval	Position-of-Gold Bias	Strong
F13	Generation	Hallucination Despite Context	Strong
F14	Generation	Conflicting Info Unresolved	Strong
F15	Generation	Incomplete / Partial Answers	Moderate
F16	Generation	Incorrect Specificity	Moderate
F17	Generation	Wrong Output Format	Limited
F18	Evaluation	Metric Inadequacy	Strong
F19	Evaluation	Lack of Continuous Monitoring	Limited
F20	Deployment	Latency / Cost Bottlenecks	Moderate
F21	Deployment	Auditability Gaps	Moderate
F22	Deployment	Domain Portability Degradation	Moderate
F23	Deployment	PII / Compliance Leaks	Moderate
F24	Deployment	Prompt Sensitivity	Moderate
F25	Deployment	Authorization / Policy Failures	Moderate
F26	Agentic	Planning Failures	Limited
F27	Agentic	Tool Selection/ Execution Errors	Limited
F28	Agentic	Context Memory Degradation	Limited
F29	Agentic	Multi-Agent Coordination Failures	Limited
F30	Agentic	Recursive Hallucination Cascades	Limited
F31	Agentic	Unbounded Cost / Latency Spirals	Limited
F32	Agentic	Unsafe Reasoning Chains	Limited
F33	Agentic	Attribution & Governance Gaps	Limited

Table 2: Complete taxonomy overview: 33 failure modes across 7 categories. Evidence levels indicate empirical support strength (Strong = dedicated empirical study; Moderate = supporting evidence; Limited = architectural or practitioner-based evidence).

2025) demonstrates that RAG systems are unexpectedly sensitive to scenarios where relevant content is systematically absent. Note: this mode concerns corpus-level missingness and is conceptually distinct from F22 (Domain Portability Degradation), which concerns performance decline when documents exist but were optimized for a different domain. *Evidence: Moderate.*

F3. Layout Parsing Errors. Documents with heterogeneous layouts (tables, figures, multi-column text, equations) resist uniform processing. Ma et al. (2025) benchmark multimodal retrieval for long documents averaging 65.1 pages, showing layout-unaware methods fail to retrieve correct elements. Zhou et al. (2025) demonstrate that explicit structural awareness significantly improves evidence acquisition over treating passages as isolated chunks. *Evidence: Strong.*

F4. Multimodality Conversion Loss. Information conveyed through non-text modalities (images, diagrams, charts) is lost when documents are converted to text-only format. Park et al. (2025) show performance varies significantly with document complexity. Zhang et al. (2024) demonstrate that OCR errors cascade through RAG pipelines, compounding retrieval and generation failures. *Evidence: Moderate.*

5.2 Representation Failures (F5–F6)

Representation failures occur when encoding of documents or queries into vector representations degrades semantic fidelity, causing systematic retrieval degradation. These failures are uniquely insidious because they degrade retrieval quality while remaining invisible to generation-level metrics, making them hard to detect and easy to misattribute to later stages.

F5. Tokenization Fragmentation. Tokenizers split semantically meaningful domain-specific units (medical terms, abbreviations, compound

words) into subword fragments, degrading embedding coherence. Pavlick et al. (2025) identify tokenizer fragmentation as a dominant bottleneck for non-English text, reporting that replacing fragmented sequences with semantic slots improves recall-at-1 from 0.104 to 0.430. *Evidence: Limited — no RAG-specific studies directly tie tokenizer fragmentation to retrieval failure; evidence comes from adjacent cross-lingual work.*

F6. Embedding Drift / Model Mismatch. Vector embeddings become inconsistent over time as models are updated, or different tokenizers between retrieval encoder and generation model cause misalignment at the retrieval-generation interface. In practice, such representation shifts can silently erode retrieval quality, leading to cascading failures that standard generation-focused evaluations fail to surface. *Evidence: Limited — discussed in engineering guides (Akkiraju et al., 2024) but no dedicated peer-reviewed analysis.*

5.3 Retrieval Failures (F7–F12)

Retrieval failures represent the most extensively studied category, occurring when the retrieval component fails to surface relevant documents or surfaces irrelevant ones.

F7. Chunking Boundary Errors. Documents are segmented into chunks that break semantic coherence or split critical information across boundaries. Stankovic (2026) demonstrates that poor chunking caps quality and degrades faithfulness. Comparative evaluation in clinical decision support (Gomez-Cabello et al., 2025) shows adaptive chunking achieves retrieval precision of 0.50 vs. 0.17 for fixed-length baselines. Sarthi et al. (2025) show inappropriate granularity introduces noise that particularly degrades weaker LLMs. Cresswell et al. (2025) identify underchunking and overchunking as distinct error types with different mitigation strategies. *Evidence: Strong.*

F8. Domain Embedding Mismatch. General-purpose embedding models fail to capture domain-specific semantic relationships. Gupta et al. (2023) demonstrate through TACL that domain adaptation significantly improves retrieval-augmented QA. Liu et al. (2024) introduce FinMTEB, showing no correlation between general-benchmark and financial-benchmark performance. *Evidence: Strong.*

F9. Multi-Hop Reasoning Gaps. Queries requiring connections across multiple documents are not served by single-step retrieval. Liu et al. (2025) show HopRAG achieves up to 76.78% higher answer accuracy on multi-hop benchmarks, with EM gains of 15–30 points over non-planning baselines on 2Wiki and HotpotQA. *Evidence: Strong.*

F10. Graph RAG Trade-offs. Knowledge graph structures for retrieval introduce complexity-coverage trade-offs. Six et al. (2025) identify a structure-content trade-off where structural coherence reduces content coverage. Wang et al. (2025b) note that graph-based RAG methods sacrifice recall for higher precision. *Evidence: Moderate.*

F11. Low Recall / Ranking Failures. Relevant documents exist but rank below the top-k cutoff or are not retrieved at all. Cresswell et al. (2025) identify missed retrieval as a distinct error type, finding it accounts for a significant proportion of errors in their empirical analysis. *Evidence: Moderate.*

F12. Position-of-Gold Bias. LLMs disproportionately attend to retrieved documents based on context position rather than relevance, exhibiting U-shaped attention that ignores middle-placed content. Byerly and Khashabi (2025) demonstrate up to 65% fewer LLM queries needed with position-aware reordering and 34% accuracy gains on 400-fact contexts. Multiple recent papers (Wang et al., 2025a; Yu et al., 2025) confirm and mitigate this bias. *Evidence: Strong.*

5.4 Generation Failures (F13–F17)

F13. Hallucination Despite Retrieved Context. The generator produces claims not supported by retrieved documents despite having relevant context available. This is the most extensively studied RAG failure, addressed by faithfulness evaluation in RAGAS (Es et al., 2024), RAGChecker (Ru et al., 2024), and numerous mitigation works (Zhang et al., 2025). Cresswell et al. (2025) further decompose this into content fabrication, parameter override, and context misalignment subtypes. Note: F13 characterizes single-turn hallucination and is mechanistically distinct from F30 (Recursive Hallucination Cascades), which describes how agentic multi-step workflows amplify hallucinations through fabricated intermediate outputs. *Evidence: Strong.*

F14. Conflicting Information Unresolved. Retrieved documents contain contradictory informa-

tion and the generator fails to resolve or acknowledge the conflict. Lee et al. (2025) demonstrate that generators struggle to resolve contradictions, while the RGB benchmark (Chen et al., 2023) evaluates robustness under conflicting information conditions. BordaRAG (BordaRAG Team, 2025) proposes resolution through voting mechanisms. *Evidence: Strong.*

F15. Incomplete / Partial Answers. Responses address only part of the query, missing relevant information available in retrieved context. Barnett et al. (2024) identify this as one of their seven failure points. Cresswell et al. (2025) identify incomplete answers as a distinct generation error type. *Evidence: Moderate.*

F16. Incorrect Specificity. Responses are too general or too specific relative to query intent. Barnett et al. (2024) identify this as a failure point; however, it has not been isolated with granularity-specific metrics. *Evidence: Moderate.*

F17. Wrong Output Format. The response does not conform to required structural constraints (e.g., JSON schema, tabular format, citation template), even when the retrieved content is correct. While semantic accuracy may be preserved, structural non-compliance can break downstream parsing or tool-invocation components. Prompt-sensitivity studies (Zhu et al., 2024; Arabzadeh and Clarke, 2025) demonstrate brittleness under format constraints, providing indirect empirical support. *Evidence: Limited.*

5.5 Evaluation Failures (F18–F19)

F18. Metric Inadequacy. Standard metrics (BLEU, ROUGE, exact match) fail to capture generation quality, faithfulness, or utility. RAGChecker (Ru et al., 2024) demonstrates significantly better correlation with human judgments than prior metrics. MIRAGE (MIRAGE Team, 2025) proposes comprehensive multi-metric evaluation. Yu et al. (2024) catalog limitations of existing metrics. *Evidence: Strong.*

F19. Lack of Continuous Monitoring. Systems are evaluated at development time but lack production monitoring, causing silent quality degradation as data and models evolve. *Evidence: Limited — emerging concern in RAG operations literature.*

5.6 Deployment and Operations Failures (F20–F25)

Deployment failures differ from upstream pipeline failures in that they emerge only when RAG systems operate under real-world constraints of scale, governance, and multi-user access.

F20. Latency / Cost Bottlenecks. End-to-end RAG latency exceeds acceptable thresholds, or improving quality requires disproportionate cost increases. Es et al. (2024) analyze system trade-offs. DIRC-RAG (DIRC-RAG Team, 2025) addresses edge deployment latency. Akkiraju et al. (2024) report empirical accuracy-latency trade-offs between large and small LLMs. Note: F20 concerns steady-state latency and cost under expected load and is distinct from F31 (Unbounded Cost/Latency Spirals), which concerns unbounded cost growth from recursive agentic execution. *Evidence: Moderate.*

F21. Auditability Gaps. RAG systems lack sufficient logging and tracing to verify which documents influenced specific outputs. S-RAG Team (2025) achieve 94.1% accuracy in detecting unauthorized data use in RAG systems. Note: F21 concerns system-level logging infrastructure and is distinct from F33 (Attribution and Governance Gaps), which concerns semantic attribution across multi-step reasoning chains. *Evidence: Moderate.*

F22. Domain Portability Degradation. RAG systems optimized for one domain show significant degradation in other domains. Domain adaptation is widely studied (Gupta et al., 2023) but portability failures are not isolated as a distinct RAG failure mode. *Evidence: Moderate.*

F23. PII / Compliance Leaks. Retrieved documents or generated responses expose personally identifiable or regulated data. Discussed in trustworthiness frameworks (TrustworthyRAG Team, 2024) but limited RAG-specific empirical analysis exists. Note: F23 concerns content-level exposure, whereas F25 concerns policy-layer failures where access rules are not enforced even when content is not inherently sensitive. *Evidence: Moderate.*

F24. Prompt Sensitivity. Small prompt variations cause significant performance changes. Zhu et al. (2024) introduce PromptSensiScore showing discrepant responses across models. Arabzadeh and Clarke (2025) demonstrate prompt variations significantly impact LLM relevance judgments. *Evidence: Moderate.*

F25. Authorization & Policy Enforcement Failures. Retrieved or generated outputs violate role-based access controls despite technically correct retrieval and generation. Unlike F23 (content-level exposure), authorization failures arise from insufficient enforcement of user-specific permissions or document-level access rules. *Evidence: Moderate* — Ammann et al. (2025) identify authorization enforcement as a primary retrieval-layer risk in RAG-specific security frameworks; Berini et al. (2026) corroborate authorization gaps as a recurring compliance failure pattern in LLM security surveys.

5.7 Agentic Orchestration Failures (F26–F33)

Agentic RAG extends standard RAG with autonomous planning, tool use, and multi-step reasoning (Singh et al., 2025; Masterman et al., 2024). These capabilities introduce entirely new failure categories. Empirical evidence for agentic failures is nascent; the modes below are identified primarily from architectural analysis and emerging practitioner reports. This evidence gap itself is a key finding of our taxonomy.

F26. Planning Failures. Agents generate plans that are overly complex, overly simplistic, or circular. Note: F26 concerns single-agent planning quality and is distinct from F29 (Multi-Agent Coordination Failures), which concerns inter-agent orchestration; a well-planned individual agent can still fail F29 in a multi-agent context. *Evidence: Limited.*

F27. Tool Selection and Execution Errors. Agents choose inappropriate tools, provide malformed parameters, or fail to handle tool timeouts and non-deterministic outputs. Masterman et al. (2024) discuss tool use failures in agentic AI architectures. Note: F27 concerns correctness of individual tool calls (selection, parameters, error handling) and is distinct from F31 (Unbounded Cost/Latency Spirals), which concerns the boundedness of tool-call sequences. *Evidence: Limited.*

F28. Context Memory Degradation. Agents lose prior context (forgetting), over-rely on stale cache, or allow context leakage across queries. *Evidence: Limited.*

F29. Multi-Agent Coordination Failures. Orchestrators enter deadlocks, agents produce contradictory outputs, or coordination overhead exceeds parallelism benefit (Singh et al., 2025). *Evidence: Limited.*

F30. Recursive Hallucination Cascades. Hallucinated intermediate results trigger subsequent queries based on fabrications, creating cascading chains of increasingly fabricated information. Zhang et al. (2025) identify recursive hallucination as a distinct failure mode in agentic RAG loops. *Evidence: Limited.*

F31. Unbounded Cost / Latency Spirals. Agentic workflows lack guardrails on execution depth, allowing cascading costs through recursive tool calls and retrieval operations. *Evidence: Limited.*

F32. Unsafe Reasoning Chains. Multi-step reasoning produces harmful outputs through individually innocuous steps, where the composite action violates safety constraints not visible at any single step. *Evidence: Limited.*

F33. Attribution and Governance Gaps. Agents fail to trace generated claims back to source documents across multi-step workflows, and collectively breach compliance regulations despite individually compliant actions. *Evidence: Limited.*

Grounding the Agentic Modes. Each mode is grounded in classical multi-agent systems and AI safety literature. Measurable proxies include: F26 — plan length vs. query complexity; F27 — tool error rate; F28 — context staleness score; F29 — output agreement rate; F30 — hallucination rate per reasoning depth; F31 — cost-per-query variance; F32 — safety constraint satisfaction at composite vs. step level; F33 — attribution coverage rate.

6 Analysis and Discussion

6.1 Evidence Distribution

Table 3 summarizes the distribution of evidence levels across our taxonomy. Of 33 identified failure modes, 9 (27%) have strong empirical evidence, 12 (36%) have moderate evidence, and 12 (36%) have limited or no dedicated peer-reviewed investigation. The well-studied failures concentrate in retrieval (4 of 6 modes strong) and generation (2 of 5 strong). The under-studied failures concentrate in representation (both limited), evaluation (1 of 2 limited), and especially agentic orchestration (all 8 modes limited).

Category	Strong	Moderate	Limited	Total
Ingestion	2	2	0	4
Representation	0	0	2	2
Retrieval	4	2	0	6
Generation	2	2	1	5
Evaluation	1	0	1	2
Deployment	0	6	0	6
Agentic	0	0	8	8
Total	**9 (27%)**	**12 (36%)**	**12 (36%)**	**33**

Table 3: Evidence Level Distribution by Category.

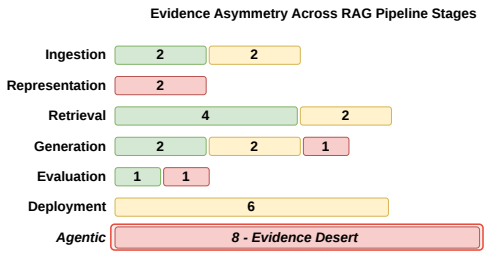

Figure 2: Evidence asymmetry across RAG pipeline categories. Bar width proportional to number of failure modes.

6.2 Key Insights

Research attention asymmetry. The distribution reveals a stark imbalance. Retrieval and generation failures—the stages most amenable to benchmarking—have attracted disproportionate research attention. Representation failures, despite creating systematic performance ceilings, have no dedicated RAG-specific studies. Deployment failures, despite being the primary concern of practitioners building production systems, have only moderate evidence levels drawn from broader engineering reports.

The agentic evidence gap. Agentic RAG represents the fastest-growing deployment paradigm (Singh et al., 2025), yet all 8 identified agentic failure modes lack qualifying peer-reviewed empirical analysis. This creates a dangerous situation where the most complex and failure-prone architectures receive the least scientific scrutiny. This finding aligns with Singh et al. (2025)'s observation that agentic RAG evaluation frameworks significantly lag behind system development.

Cascade blindness. A fundamental limitation of current RAG evaluation is its stage-local nature: benchmarks measure retrieval quality, faithfulness, or answer accuracy in isolation, masking multi-stage failure propagation. When chunking errors (F7) degrade embedding quality, retrieval recall (F11) falls silently; when retrieval fails, hallucination (F13) becomes structurally inevitable yet generation-level metrics misattribute the root cause. We define *cascade blindness* as the systematic underestimation of end-to-end failure rates that occurs when evaluation focuses on single-stage metrics rather than their composition across pipeline stages. Developing cascade-aware evaluation benchmarks with multi-stage labeled traces is a critical open problem identified by this taxonomy. Our taxonomy is complementary to Cresswell et al. (2025)-their work provides empirical depth on specific benchmarks while ours provides breadth across all pipeline stages including agentic orchestration.

6.3 Under-Studied Failure Modes

Of the 12 *limited*-evidence modes (F5, F6, F17, F19, F26–F33), none have dedicated peer-reviewed benchmarks isolating causal impact. Representation failures (F5–F6) systematically degrade retrieval yet lack RAG-specific ablations. Most critically, all 8 agentic failure modes (F26–F33) lack controlled empirical validation despite representing the fastest-growing RAG paradigm (Singh et al., 2025). Future work should prioritize: (1) synthetic benchmarks isolating representation-layer distortions; (2) cascade-aware evaluation frameworks; (3) agentic loop benchmarks for recursive hallucination and planning instability.

6.4 Implications for Practitioners

Our taxonomy serves as a diagnostic checklist for RAG system development:

Pre-deployment audit. Systematically evaluate susceptibility to each relevant failure mode based on domain, scale, and architecture choices. The evidence-level grading indicates which failures have known mitigations (strong evidence) versus which require defensive engineering (limited evidence).

Root cause analysis. When end-to-end performance degrades, trace failures upstream through the taxonomy to identify actual root causes rather than surface symptoms. Generation-stage symptoms often have ingestion or retrieval root causes.

Architecture selection. Choose between naive, advanced, modular, and agentic RAG based on the failure profiles most relevant to the deployment domain, informed by the evidence levels indicating how well each failure class is understood.

Diagnostic procedure. Practitioners should identify the observed symptom, localise it to a pipeline stage, trace upstream candidate failure modes using Table 2, select a mitigation tier, and validate using stage-specific metrics.

6.5 Worked Example: Diagnosing a Cascading Failure

To illustrate how practitioners can use this taxonomy diagnostically, we present an illustrative scenario synthesizing patterns commonly observed in production RAG deployments. The scenario is hypothetical and does not reproduce any specific deployment.

Observed symptom. A clinical decision-support RAG system serving care teams begins producing responses with hallucinated citations to clinical guidelines. Users report that the system "gives the textbook answer, not the answer for this patient."

Stage-local diagnosis. Standard generation-stage evaluation indicates faithfulness scores in the acceptable range on a curated test set. Retrieval recall on the same test set is high. Stage-local metrics suggest no failure.

Trace upstream through Table 2. The hallucinated citations point initially to F13 (Hallucination Despite Context). However, F13 with high faithfulness scores on test data suggests the test set itself does not represent production queries—a signal of upstream propagation. Moving up the cascade: F8 (Domain Embedding Mismatch) is plausible because clinical terminology is poorly represented in general-purpose embeddings; F7 (Chunking Boundary Errors) is plausible because clinical guidelines contain dense tabular dosing information; F3 (Layout Parsing Errors) is plausible because guidelines are PDF-native with multi-column layouts and structured tables.

Identified root cause. Inspection of parsed documents reveals that table boundaries in dosing tables are broken during ingestion (F3), producing chunks where dosage values are separated from the conditions they apply to (F7). At retrieval time, queries about specific clinical scenarios match these structurally broken chunks at high cosine similarity (because the surrounding text is topically relevant), but the chunks lack the conditional context required for faithful generation. The symptom presents as F13; the root cause is F3. The test set, drawn from manually curated examples, did not contain the affected document types and was therefore unable to surface the cascade.

Mitigation. Fixing F3 through layout-aware parsing, validated by F19 (Lack of Continuous Monitoring) on a production-representative evaluation set, resolves the visible F13 symptom without prompt-level intervention. This pattern—symptoms presenting at one stage while root causes reside upstream—illustrates the cascade blindness phenomenon discussed in Section 6.2 and motivates the value of full-pipeline taxonomies for diagnostic routing.

6.6 Limitations

Our review, while covering 48 papers, may miss work in non-English or domain-specific venues. RAG is rapidly evolving and some failure modes may be resolved by architectural advances. Agentic modes (F26–F33) rest on architectural reasoning rather than empirical studies. Unlike Cresswell et al. (2025), we prioritize breadth over empirical depth on specific stages.

The evidence grading protocol relied on a single rater (the lead author). While the conservative-downgrading rules described in Section 4.2 bias the taxonomy toward under-claiming evidence strength, formal inter-rater agreement was not established. Future replication with independent raters would strengthen the construct validity of evidence-level assignments.

Claims regarding the relative frequency, severity, or production prevalence of specific failure modes rest on secondary interpretation of cited studies and practitioner reports rather than direct measurement. The taxonomy is therefore most usefully read as a structured map of the failure space, not as a quantitative ranking of failure prevalence. Future work should develop targeted benchmarks and evaluation datasets to empirically validate the failure modes identified in this taxonomy.

7 Broader Impact

This taxonomy is intended primarily as defensive infrastructure: a diagnostic resource for practitioners building, auditing, and governing production RAG systems in high-stakes domains including healthcare, finance, and law. By making failure modes explicit and mapping evidence asymmetry, the taxonomy supports more rigorous pre-deployment review, more targeted reliability re-

search, and more transparent communication between technical teams and non-technical stakeholders.

We acknowledge a dual-use consideration: explicit enumeration of failure modes—particularly under-studied modes in agentic orchestration—could in principle inform adversarial probing of deployed systems. We judge this risk to be small relative to the defensive value for two reasons. First, the failure modes documented here are already known to motivated attackers through deployment observation; the taxonomy primarily benefits defenders who lack the structured view that attackers naturally develop through targeted testing. Second, the under-studied modes are flagged precisely because they lack benchmarks; the taxonomy points to research gaps rather than exploitation procedures.

We encourage practitioners using this taxonomy in high-stakes deployments to combine it with domain-specific risk assessment, human oversight at decision points where the taxonomy identifies Limited evidence, and continuous monitoring for cascade effects that single-stage evaluation cannot detect.

8 Conclusion

We present a systematic taxonomy of 33 RAG failure modes across 7 pipeline stages, with explicit evidence-level grading revealing that 12 modes—all 8 agentic modes among them—lack peer-reviewed empirical investigation. This taxonomy serves as both a diagnostic framework for practitioners and a research roadmap. Future work should prioritize empirical validation of under-studied modes, cascade-aware benchmarks, and integration with auto-evaluation tools such as RAGChecker (Ru et al., 2024). As RAG systems evolve toward agentic architectures, failure analysis must evolve from component-level benchmarking to end-to-end reliability science. We hope this taxonomy serves as a foundation for future benchmarking efforts and reliability-focused evaluation of retrieval-augmented generation systems.

References

Rama Akkiraju, Anbang Xu, Deepak Bora, Tian Yu, and 1 others. 2024. FACTS about building retrieval-augmented generation-based chatbots.

L. Ammann, S. Ott, C. R. Landolt, and M. P. Lehmann. 2025. Securing RAG: A risk assessment and mitigation framework.

Negar Arabzadeh and Charles L. A. Clarke. 2025. A human-AI comparative analysis of prompt sensitivity in LLM-based relevance judgment.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval-augmented generation system. In *Proceedings of the IEEE/ACM International Conference on AI Engineering (CAIN)*.

A. D. E. Berini, N. Jamil, A. Benrazek, A. Lakas, L. Ismail, M. A. Ferrag, and K. Lam. 2026. Security and privacy in LLMs: A comprehensive survey of threats and mitigation strategies. *Computers & Security*.

BordaRAG Team. 2025. BordaRAG: Resolving knowledge conflict in retrieval-augmented generation. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Adam Byerly and Daniel Khashabi. 2025. Turning positional bias into signal for multi-document LLM reasoning.

Jiawei Chen, Hongyu Deng, Jiansen Bian, Ruichu Qiu, Bin Wu, Tao Shi, Ruining Ma, Yaliang Lv, and Yong Ma. 2023. Benchmarking large language models in retrieval-augmented generation (RGB).

Joseph Cresswell and 1 others. 2025. Classifying and addressing the diversity of errors in retrieval-augmented generation systems.

Renan Tiago De Lima, Shubham Gupta, Carlos Berrospi, and 1 others. 2024. Know your RAG: Dataset taxonomy and generation strategies for evaluating RAG systems.

DIRC-RAG Team. 2025. DIRC-RAG: Accelerating edge RAG with robust high-density digital in-ReRAM computation.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated evaluation of retrieval-augmented generation. In *Proceedings of the EACL System Demonstrations*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey.

Yunfan Gao and 1 others. 2024. Retrieval-augmented generation for AI-generated content: A survey.

Carlos A. Gomez-Cabello, Sanjana Prabha, Syed A. Haider, Andrea Genovese, Bruno G. Collaco, Nathan G. Wood, Shreya Bagaria, and Antonio J. Forte. 2025. Comparative evaluation of advanced chunking for retrieval-augmented generation in large language models for clinical decision support. *Bioengineering*, 12(11):1194.

Porus Gupta, Vishwajeet Basu, Anchit Puri, and Rishabh Iyer. 2023. Improving the domain adaptation of retrieval-augmented generation (RAG) for open-domain question answering. *Transactions of the Association for Computational Linguistics*.

Shailja Gupta, Rasika Ranjan, and Shubhanshu Singh. 2024. A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions.

Shengding Huang and Ming Chang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.

Gautier Izacard and Edouard Grave. 2021. 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*.

Kaiyue Lee and 1 others. 2025. Retrieval-augmented generation with conflicting evidence.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Haoyan Liu, Zheng Wang, Xin Chen, Zhiyuan Li, Fei Xiong, Qing Yu, and Wei Zhang. 2025. HopRAG: Multi-hop reasoning for logic-aware retrieval-augmented generation.

Xiangyu Liu and 1 others. 2024. Do we need domain-specific embedding models? an empirical study on financial text.

Siming Ma, Tao Jiang, Zhuo Wang, Hao Wang, Zheng Liu, and Maosong Sun. 2025. MMDocIR: Benchmarking multimodal retrieval for long documents. In *Findings of the Association for Computational Linguistics: EMNLP*.

Shashanka Manathunga and Yasiru Illangasekara. 2025. Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers.

Tula Masterman, Sanjana Besen, Mason Sawtell, and Alex Chao. 2024. The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey.

MIRAGE Team. 2025. MIRAGE: A metric-intensive benchmark for retrieval-augmented generation evaluation. In *Findings of the Association for Computational Linguistics: NAACL*.

Jiaxin Ouyang, Tianyuan Pan, Meng Cheng, Rui Yan, Yanan Luo, Jing Lin, and Qian Liu. 2025. HoH: A dynamic benchmark for evaluating the impact of outdated information on RAG. In *Proceedings of the Association for Computational Linguistics*.

Jaeho Park, Marco Rossi, and Ankit Gupta. 2025. Benchmarking retrieval-augmented multimodal generation on complex PDF documents.

Ellie Pavlick, Percy Liang, and 1 others. 2025. Languages are modalities: Cross-lingual alignment via encoder bridge (LLINK).

Dongyu Ru, Linyi Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuohang Chang, and 1 others. 2024. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation.

S-RAG Team. 2025. S-RAG: A novel audit framework for detecting unauthorized use of data in RAG systems. In *Proceedings of the Association for Computational Linguistics*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. ARES: An automated evaluation framework for retrieval-augmented generation systems.

Yuvraj Sarthi and 1 others. 2025. Optimize the chunking granularity for retrieval-augmented generation systems. In *Proceedings of the International Conference on Computational Linguistics*.

Aditi Singh and 1 others. 2025. Agentic retrieval-augmented generation: A survey on agentic RAG.

Victor Six, Alexander Pichler, Don Tuggener, and Thomas Hofmann. 2025. The structure-content tradeoff in knowledge graph retrieval.

Marko Stankovic. 2026. Cross-document topic-aligned chunking for retrieval-augmented generation.

TrustworthyRAG Team. 2024. Trustworthiness in retrieval-augmented generation systems: A comprehensive framework.

Yue Wang, Fei Xiong, Yang Wang, Lei Li, Xingming Chu, and Daniel D. Zeng. 2025a. Position bias mitigates position bias: Mitigate position bias through inter-position knowledge distillation.

Yunqi Wang, Haoyang Li, Jianling Sun, and Shu Tang. 2025b. Graph retrieval-augmented generation: A survey. *ACM Computing Surveys*.

Hao Yu, Aoran Gan, Kai Zhang, and 1 others. 2024. Evaluation of retrieval-augmented generation: A survey.

Yiwei Yu, Hao Jiang, Xiao Luo, Qiang Wu, Chin-Yew Lin, Deng Li, Yi Yang, Yu Huang, and Lizi Qiu. 2025. Mitigate position bias in large language models via scaling a single dimension.

Yue Zeng, Tao Cao, Dong Wang, Xu Zhao, Zheng Qiu, Mojtaba Ziyadi, Tongshuang Wu, and Liunian Li. 2025. RARE: Retrieval-aware robustness evaluation for retrieval-augmented generation systems.

Junwen Zhang, Qian Zhang, Bing Wang, and 1 others. 2024. OCR hinders RAG: Evaluating the cascading impact of OCR on retrieval-augmented generation.

Min Zhang, Lu Li, Zhen Wang, and Jing Zhang. 2026. Graph-RAG-enabled local LLM for gestational diabetes education. *JMIR Diabetes*, 11:e76454.

Weiwei Zhang, Ying Hou, Dan Liu, and Heng Ji. 2025. Hallucination mitigation for retrieval-augmented generation: A review. *Mathematics*.

Yuan Zhou, Chen Zhang, Han Wu, and Jia Li. 2025. Equipping retrieval-augmented LLMs with structure-aware document routing (RDR2). In *Findings of the Association for Computational Linguistics: EMNLP*.

Jieyi Zhu and 1 others. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP*.