Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

RETRIEVAL-AUGMENTED GENERATION FOR PHARMACEUTICAL REGULATORY INTELLIGENCE: A NARRATIVE REVIEW

Luca Bianchi[1]*, Marco Rossi[1], Giulia Romano[2]

1. ***Department of Pharmaceutical Informatics and AI, Faculty of Pharmacy, University of Milan, Milan, Italy.***
2. ***Department of Intelligent Drug Systems, Faculty of Engineering, Polytechnic University of Turin, Turin, Italy.***

ARTICLE INFO

Received:
21 February 2026
Received in revised form:
14 May 2026
Accepted:
15 May 2026
Available online:
28 June 2026

Keywords: Retrieval-augmented generation, Regulatory intelligence, Pharmaceutical regulation, Grounding, Hallucination control, Traceability

ABSTRACT

Regulatory intelligence depends on the accurate retrieval, interpretation, and synthesis of complex pharmaceutical guidance, product documentation, and evolving compliance expectations. The growing volume and specificity of regulatory text have made manual synthesis increasingly difficult for development, quality, and submission teams. Retrieval-augmented generation emerged as a response to the factual limitations of large language models by linking generated answers to external knowledge sources. This shift was especially relevant to pharmaceutical regulation, where unsupported synthesis can mislead decision-making. Modern RAG systems combine dense retrieval, passage ranking, prompt control, and verification layers to produce answers that are both fluent and traceable. They are now being explored for guideline interpretation, submission preparation, compliance checking, and other regulatory intelligence workflows. Despite these advances, RAG systems remain vulnerable to retrieval errors, ambiguous source language, incomplete document corpora, and residual hallucination. Trust depends not only on answer quality but also on source transparency, auditability, and expert review. Future regulatory RAG systems are likely to combine structured knowledge graphs, real-time guidance monitoring, validated evaluation frameworks, and privacy-preserving deployment. These capabilities may support increasingly autonomous but still accountable regulatory intelligence tools. This narrative review traces the evolution of RAG from a technical solution for hallucination mitigation to a credible framework for pharmaceutical regulatory intelligence. It emphasizes grounding, traceability, evaluation, and governance as the foundations of trustworthy regulatory AI.

To Cite This Article: Bianchi L, Rossi M, Romano G. Retrieval-Augmented Generation for Pharmaceutical Regulatory Intelligence: A Narrative Review. Pharmacophore. 2026;17(3):44-52. https://doi.org/10.51847/SMA6GzNcaq

Introduction

Pharmaceutical regulatory intelligence has always depended on the ability to interpret large, heterogeneous, and frequently revised bodies of text. Development teams must synthesize international guidelines, agency expectations, product labels, clinical protocols, quality documentation, and safety communications into practical decisions about evidence generation and submission strategy. The growing scale of regulatory documentation resembles the broader information-management challenges seen in biomedical informatics, where domain-specific retrieval and expert interpretation are essential for safe decision support [1]. In this setting, regulatory intelligence is not simply search; it is the disciplined transformation of distributed textual evidence into accountable recommendations.

The arrival of large language models intensified interest in automating this synthesis, but it also exposed a fundamental limitation: fluent language is not the same as factual reliability. Clinical and biomedical studies of large language models showed that such systems could encode useful domain knowledge while still requiring careful safeguards before use in high-stakes environments. Research on factuality in summarization similarly demonstrated that generated text may appear coherent while deviating from source evidence [2, 3]. For pharmaceutical regulation, where a misread requirement can affect dossier quality or compliance posture, this distinction became central rather than incidental.

Retrieval-augmented generation introduced a practical paradigm shift by coupling language generation with external document retrieval. Instead of relying only on parametric memory, RAG systems retrieve relevant passages and condition generation on

Corresponding Author: Luca Bianchi; Department of Pharmaceutical Informatics and AI, Faculty of Pharmacy, University of Milan, Milan, Italy. E-mail: luca.bianchi@gmail.com.

those passages, an approach formalized in early architectures for knowledge-intensive natural language processing [4]. Dense passage retrieval, retrieval-augmented pretraining, and passage-conditioned generation supplied the technical foundation for answers that could be linked back to evidence [5, 6]. This made RAG attractive for regulatory intelligence because it aligned a computational architecture with a professional norm: claims should be supported by identifiable source text.

This narrative review traces how RAG moved from general natural language processing into pharmaceutical regulatory intelligence. It follows the technical milestones of retrieval, ranking, grounding, and verification while connecting them to use cases such as guideline interpretation, submission drafting, compliance checking, and label review. Recent pharmaceutical and healthcare RAG studies show that the method is now being adapted to regulated biomedical workflows rather than remaining a generic question-answering technique [7-9]. The review therefore emphasizes not only what RAG can generate, but how its outputs can be governed, verified, and trusted in regulatory practice.

Figure 1 illustrates the chronological evolution of retrieval-augmented generation from early hallucination-control architectures into a governed framework for pharmaceutical regulatory intelligence.

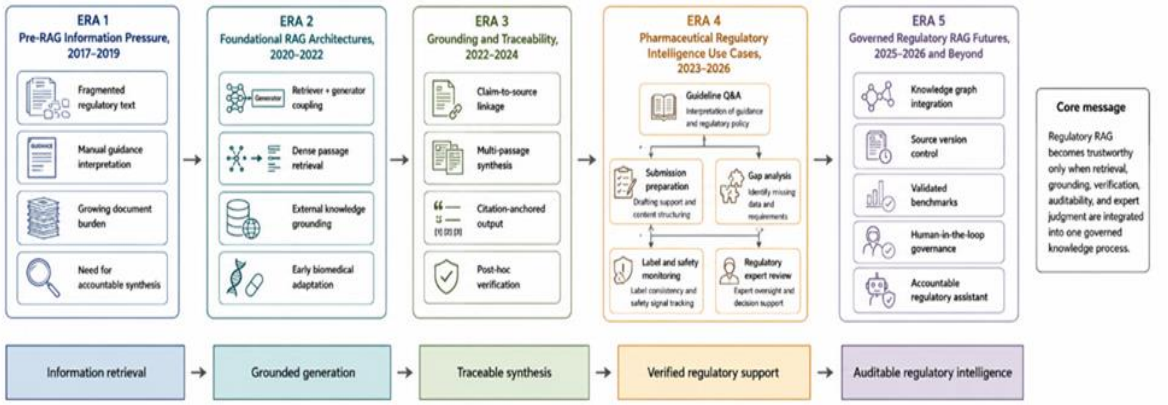

Figure 1. Chronological evolution of retrieval-augmented generation toward pharmaceutical regulatory intelligence

The Rise of Retrieval-Augmented Generation (C. 2020–2022)

The Hallucination Problem and the Retrieval Solution

RAG emerged from the recognition that language models alone could not reliably answer knowledge-intensive questions when factual precision mattered. The original RAG formulation combined a neural retriever with a sequence generator so that answers were shaped by retrieved evidence rather than only by internal model parameters [4]. REALM extended this logic by integrating retrieval into language model pretraining, making retrieval part of representation learning rather than a post-hoc add-on [5]. Together, these developments reframed hallucination control as an architectural problem: the model should be supplied with the right evidence before it writes.

Table 1 distinguishes successive maturity layers of regulatory RAG, showing how technical capability must be matched with regulatory function, trust contribution, failure control, and governance discipline.

Table 1. Conceptual Maturity Matrix for Retrieval-Augmented Generation in Pharmaceutical Regulatory Intelligence

Maturity layer	Core technical capability	Regulatory intelligence function	Trust contribution	Primary failure risk	Governance implication
Manual regulatory search	Keyword search, document browsing, expert reading	Locating guidance, precedent, labeling language, and compliance expectations	Relies on human expertise and professional judgment	Slow synthesis, missed documents, inconsistent interpretation across teams	Requires documentation standards and reviewer accountability
Retrieval-supported question answering	Semantic retrieval, passage matching, document ranking	Finds relevant passages from regulatory or pharmaceutical corpora	Improves access to large document libraries	Retrieval may return incomplete, outdated, or superficially relevant passages	Requires curated corpora, metadata tagging, and jurisdiction/version control
Basic RAG answer generation	Retrieved context conditions LLM output	Produces readable answers to guideline, protocol, labeling, or compliance questions	Reduces dependence on model memory alone	Generated answer may overstate, misread, or combine evidence incorrectly	Requires source-restricted prompting and refusal rules
Citation-anchored RAG	Claim-to-source linkage, passage citation, answer attribution	Supports inspectable regulatory interpretation and assisted drafting	Makes outputs reviewable by regulatory experts	Citation may not fully support the generated claim	Requires claim-level citation checking and reviewer inspection

Verified RAG	Factuality checks, contradiction detection, claim decomposition	Identifies unsupported statements before use in controlled work	Strengthens reliability beyond fluent generation	Automated verification may miss nuanced regulatory interpretation	Requires expert validation and documented correction workflows
Workflow-integrated regulatory RAG	Integration with submission, labeling, compliance, and surveillance workflows	Supports drafting, gap analysis, label comparison, and guidance monitoring	Connects RAG output to practical regulatory decisions	User over-reliance may occur if AI output appears authoritative	Requires human-in-the-loop review, escalation pathways, and decision logs
Governed regulatory intelligence assistant	Private deployment, audit trail, knowledge graph, continuous corpus refresh	Monitors regulatory change, routes affected documents, drafts recommendations, and supports accountable review	Converts RAG into a managed regulatory knowledge process	Autonomy without validation could create compliance or submission risk	Requires benchmark evaluation, access control, model/version logging, and GxP-aligned auditability

Early Architectures: Dense Retrieval + Generation

The early RAG stack depended on dense retrieval systems capable of finding semantically relevant passages beyond exact keyword overlap. Dense Passage Retrieval showed that dual-encoder representations could retrieve answer-bearing passages for open-domain question answering, creating a foundation for later regulatory document search [6]. ColBERT refined this retrieval landscape through contextualized late interaction, improving the balance between efficiency and fine-grained textual matching [10]. At infrastructure level, billion-scale similarity search and hierarchical navigable small-world graphs helped make vector retrieval feasible for large corpora, including the kind of document libraries maintained by pharmaceutical organizations [11, 12].

First Adaptations to Biomedical and Regulatory Domains

Biomedical and regulatory adoption began when researchers recognized that guideline-heavy domains reward retrieval-conditioned answers more than unconstrained generation. Medical RAG benchmarking showed that healthcare questions require not only relevant retrieval but also faithful synthesis and transparent evaluation [1]. Early clinical and laboratory regulatory applications demonstrated that RAG could interpret complex procedural requirements by grounding answers in source regulations rather than generic medical knowledge [13]. Pharmaceutical examples then extended this idea toward immunogenicity data, drug information, and clinical trial protocols, indicating that RAG could support both scientific and regulatory interpretation [7, 8].

Grounding and Traceability: Making RAG Trustworthy (2022–2024)

Improving Retrieval Quality

Trustworthy RAG begins with retrieval quality because even a well-behaved generator can produce poor answers from irrelevant or incomplete passages. Fusion-in-Decoder showed how generative models could leverage multiple retrieved passages, strengthening the link between retrieval breadth and answer synthesis [14]. Retrieval from very large corpora further demonstrated that model performance could improve when generation was supported by massive external text stores, although such scale also intensified the need for ranking and filtering [15]. In regulatory intelligence, these lessons translate into hybrid search, re-ranking, and domain-adapted embeddings that can distinguish binding guidance from background commentary.

Citation and Attribution Mechanisms

Citation and attribution became central because professional users need to know not only what an AI system says but where each claim came from. In healthcare RAG, citation-anchored outputs are increasingly treated as a usability and safety feature, especially when answers may influence clinical or operational decisions [9]. Pharmaceutical regulatory compliance systems have similarly emphasized traceable links between generated statements and the underlying drug information, protocol language, or regulatory requirement [8]. The practical goal is not decorative referencing but claim-level accountability, so that a reviewer can inspect the exact passage supporting an interpretation.

Post-Hoc Fact-Verification

Post-hoc fact-verification developed as a second layer of defense after retrieval and generation. FEVER established a broader benchmark tradition for verifying claims against evidence, while later factuality work in summarization highlighted that generated statements must be checked against source documents rather than judged only for fluency [2, 16]. Consistency-detection methods such as FactCC and SummaC pushed this field toward automated identification of unsupported or contradictory claims [3, 17]. For regulatory RAG, these methods suggest workflows in which an answer is generated only after retrieval, then decomposed into claims and verified before being shown to a user.

Hallucination-Control Strategies for High-Stakes Text

Prompt-Engineering Approaches

Prompt engineering became one of the first practical mechanisms for constraining RAG outputs in high-stakes settings. A regulatory assistant can be instructed to answer only from retrieved passages, state when evidence is insufficient, and separate direct source interpretation from inference. Healthcare RAG reviews describe this refusal behavior as essential because retrieval alone does not guarantee that the model will faithfully use retrieved context [18]. In pharmaceutical regulatory work, such prompts are most valuable when they transform uncertainty into an explicit response rather than allowing the model to improvise a confident answer.

Chain-of-Thought and Self-Consistency

Reasoning strategies added another layer to hallucination control by encouraging models to work through complex queries more deliberately. Chain-of-thought prompting showed that language models could improve reasoning by generating intermediate steps, although such steps require grounding when used in regulated domains [19]. Self-consistency then demonstrated that sampling multiple reasoning paths and selecting convergent answers can reduce some reasoning errors [20]. In regulatory intelligence, the useful adaptation is not to expose speculative reasoning as authority, but to combine retrieved evidence, structured reasoning, and consistency checks before drafting a final answer.

Table 2 shows a structured overview of key reasoning strategies used to mitigate hallucination in language models and their adapted roles in regulatory intelligence workflows.

Table 2. Reasoning Strategies for Reducing Hallucination and Improving Reliability in Regulatory Intelligence Systems

Strategy	Core mechanism	Benefit	Key limitation	Adaptation for regulatory intelligence
Chain-of-thought prompting	Generates intermediate reasoning steps before final output	Improves multi-step reasoning and problem decomposition	Intermediate steps may be ungrounded or speculative	Use only as internal scaffolding combined with retrieved evidence, not as standalone justification
Self-consistency	Samples multiple reasoning paths and selects the most consistent answer	Reduces random reasoning errors and improves robustness	Computationally expensive; may still converge on biased answers	Apply consistency checks across multiple evidence-grounded reasoning traces
Retrieval-grounded reasoning	Integrates external evidence into the generation process	Reduces hallucination by anchoring outputs in real data	Depends on quality and completeness of retrieved sources	Prioritize validated regulatory, clinical, or manufacturing datasets as anchors
Structured reasoning pipelines	Separates retrieval, reasoning, and generation stages	Improves transparency and auditability	More complex system design	Align with regulatory audit requirements and decision traceability
Consistency verification	Cross-checks outputs against multiple evidence sources or rules	Detects contradictions and unsupported claims	May miss subtle domain-specific inconsistencies	Apply rule-based and evidence-based validation before final response generation

Human-in-the-Loop and Guardrails

Human-in-the-loop review remains indispensable because regulatory interpretation often depends on context, product history, jurisdiction, and risk tolerance. Medical RAG evaluations and clinical trial screening studies show that expert assessment is needed to judge whether retrieved evidence and generated recommendations are appropriate for the workflow [1, 21]. Research on revising model outputs through retrieval illustrates how systems can generate, check, and correct claims, but the final decision in high-stakes environments still requires accountable professional judgment [22]. For pharmaceutical organizations, guardrails therefore include source restrictions, escalation rules, review checklists, and audit trails rather than model behavior alone.

RAG In Pharmaceutical Regulatory Intelligence

Guideline Interpretation and Q&A

Guideline interpretation is one of the most intuitive applications of RAG because regulatory professionals routinely ask targeted questions about complex documents. A RAG system can retrieve relevant ICH, FDA, EMA, or pharmacopeial passages and synthesize an answer that preserves the relationship between the question, the source text, and the regulatory interpretation. Pharmaceutical studies on immunogenicity data and drug information show how RAG can transform specialized document collections into queryable knowledge resources [7, 8]. The value lies not in replacing regulatory expertise, but in accelerating navigation through dense guidance while preserving source visibility.

Supporting Submission Preparation

Submission preparation requires consistent use of regulatory language across modules, summaries, protocols, and responses to agency questions. RAG can support this work by retrieving relevant precedents, guidance excerpts, and internal source documents before drafting or revising dossier text. Pharmaceutical compliance research has begun to frame RAG as a tool for aligning drug information and clinical trial documentation with explicit regulatory expectations [8]. In practice, the safest role

for such systems is assisted drafting: generating source-anchored language that regulatory writers and subject-matter experts revise, approve, and document.

Compliance Checking and Gap Analysis

Compliance checking and gap analysis require comparing product-specific text against external requirements and internal standards. RAG systems can retrieve the relevant requirement, identify the corresponding section of a protocol, label, or quality document, and propose whether the evidence appears complete, inconsistent, or missing. The QA-RAG approach to pharmaceutical regulatory compliance reflects this shift from simple question answering toward structured quality assurance over regulatory processes [23]. Because compliance judgments can have inspection or submission consequences, these systems must show the retrieved requirement, the assessed document passage, and the reasoning bridge between them.

Label Consistency and Post-Marketing Surveillance

Label consistency and post-marketing surveillance extend RAG from pre-submission preparation into lifecycle regulatory intelligence. Safety updates, class labeling changes, agency communications, and regional labeling differences create a dynamic environment in which document retrieval and synthesis must remain current. Medical and healthcare RAG frameworks illustrate how retrieval-conditioned generation can support specialized decision support when evidence is fragmented across sources [9, 24]. In pharmaceutical surveillance, the same pattern can help identify whether emerging safety information or revised regulatory language should trigger review of product labeling, risk-management documents, or internal compliance records.

Evaluation of RAG Systems in Regulatory Contexts

Domain-Specific Benchmarks

Evaluation became a defining challenge as RAG moved from general question answering into medicine and regulation. Medical RAG benchmarks emphasized that systems must be judged on retrieval relevance, answer correctness, evidence use, and safety rather than on linguistic fluency alone [1]. Broader surveys of RAG highlighted the need for task-specific evaluation because performance depends heavily on corpus quality, query type, retrieval depth, and generation constraints [25, 26]. For pharmaceutical regulatory intelligence, this implies benchmarks built around real guidance interpretation, compliance checking, label comparison, and submission-readiness questions rather than generic biomedical trivia.

Table 3 provides a governance and evaluation framework for assessing whether pharmaceutical regulatory RAG systems are complete, current, grounded, factually consistent, useful, private, human-reviewed, and auditable.

Table 3. Governance and Evaluation Framework for Trustworthy Pharmaceutical Regulatory RAG

Evaluation domain	What must be assessed	Regulatory-specific question	Suitable evidence or metric	Risk if ignored	Recommended control
Corpus completeness	Whether the system contains the documents needed to answer the query	Does the corpus include the relevant current guidance, historical precedent, internal document, and jurisdiction-specific source?	Corpus inventory, source coverage audit, version history, missing-source logs	The system may confidently answer from incomplete evidence	Maintain controlled corpus registers and scheduled source-refresh procedures
Source version accuracy	Whether retrieved documents reflect the applicable effective version	Is the answer based on current, superseded, draft, regional, or internal guidance?	Effective-date metadata, document hierarchy tags, supersession tracking	Outdated or non-binding text may be treated as current requirement	Add version tags, jurisdiction filters, and source-status labels
Retrieval relevance	Whether retrieved passages actually answer the regulatory question	Did the retriever identify the passage that contains the operative requirement or interpretation?	Top-k relevance review, expert retrieval grading, recall at k, precision at k	A fluent answer may be generated from irrelevant or partial evidence	Use hybrid retrieval, re-ranking, query expansion, and expert test sets
Claim grounding	Whether every generated claim is supported by retrieved text	Can each recommendation, requirement, or interpretation be traced to a specific passage?	Claim-level citation audit, evidence-support classification, unsupported-claim rate	Citations may appear credible while failing to support the claim	Require claim-to-source mapping and citation verification
Factual consistency	Whether generated statements contradict source documents	Does the output preserve the source meaning without adding unsupported conditions or conclusions?	Factuality scoring, contradiction detection, expert fact-checking	Small distortions may create incorrect regulatory interpretation	Add post-generation verification and contradiction review
Uncertainty handling	Whether the system recognizes insufficient or ambiguous evidence	Does the answer state when evidence is missing, ambiguous, jurisdiction-dependent, or requires expert interpretation?	Refusal accuracy, uncertainty-label review, ambiguity-handling tests	The system may present uncertainty as certainty	Use refusal prompts, confidence qualifiers, and escalation rules

Regulatory usefulness	Whether the answer supports a real workflow decision	Does the output help with guideline interpretation, submission drafting, compliance checking, label review, or surveillance triage?	Expert usefulness rating, task-completion testing, workflow simulation	Technically correct answers may still be operationally weak	Evaluate against realistic regulatory tasks, not generic QA
Human oversight	Whether accountable experts review outputs before use	Who reviewed the answer, what was changed, and what decision followed?	Reviewer logs, approval status, revision history, escalation records	Users may over-rely on AI-generated regulatory conclusions	Require role-based review and documented approval workflows
Privacy and access control	Whether confidential regulatory and product information is protected	Are proprietary documents, embeddings, prompts, logs, and outputs kept within controlled infrastructure?	Access logs, deployment architecture review, data-flow assessment	Confidential strategy or product data may be exposed	Use private deployment, encryption, role-based access, and retention policies
Auditability	Whether the full AI-assisted decision path can be reconstructed	Can the organization reconstruct the query, retrieved passages, prompt, model version, answer, reviewer, and final decision?	Audit trail completeness, timestamp records, model/prompt version logs	AI-supported decisions may not withstand inspection or internal quality review	Preserve full traceability from query to approved regulatory action

Expert-Based Assessment

Expert-based assessment is especially important because many regulatory questions do not have a single mechanically obvious answer. Clinical and healthcare RAG studies show that expert reviewers are needed to judge whether an answer is faithful to retrieved evidence, clinically or operationally appropriate, and sufficiently cautious for decision support [21, 27]. Local deployment studies in radiology consultation further illustrate that safety evaluation must consider not only answer accuracy but also whether the system appropriately handles uncertainty and domain-specific constraints [28]. In regulatory settings, expert review supplies the interpretive layer that automated factuality metrics cannot yet replace.

Deployment and Governance Considerations

On-Premise and Privacy-Preserving Deployment

Pharmaceutical RAG deployment often involves sensitive regulatory correspondence, investigational product information, clinical protocols, quality records, and commercially confidential strategy documents. This creates strong incentives for on-premise or private-cloud architectures in which retrieval indexes, embeddings, prompts, logs, and generated outputs remain within controlled environments. Healthcare RAG literature increasingly recognizes deployment setting as part of safety, because model behavior cannot be separated from data governance, access controls, and corpus curation [9, 18]. In pharmaceutical organizations, privacy-preserving RAG is therefore less a technical preference than a prerequisite for using proprietary regulatory knowledge at scale.

Audit Trails, Model Updates, and GMP Alignment

Governance also requires audit trails that document which source passages were retrieved, which model version generated the answer, and which human reviewer approved or rejected the output. Compliance-oriented pharmaceutical RAG research suggests that regulatory systems must preserve traceability across the full workflow, from query formulation to final decision record [8, 23]. Methods for fact verification and claim revision further support this governance model by making it possible to identify unsupported statements before they enter controlled documentation [16, 22]. When aligned with good documentation and data-integrity expectations, RAG becomes not only a writing aid but a managed knowledge process.

Remaining Challenges and Limitations

Retrieval Failures and Out-of-Domain Queries

Retrieval failure remains the most basic limitation of RAG because the generator can only ground itself in what the system retrieves and permits it to use. Dense retrieval and late-interaction methods improved semantic matching, but they do not eliminate failures caused by ambiguous queries, outdated corpora, missing documents, or jurisdiction-specific terminology [6, 10]. Large-corpus retrieval methods show that scale can increase coverage, yet regulatory intelligence still depends on careful filtering between binding guidance, draft guidance, historical precedent, and internal interpretation [15]. Out-of-domain queries are therefore dangerous when the system retrieves superficially similar text and presents it as regulatory evidence.

Residual Hallucination and Over-Reliance

Residual hallucination persists even when retrieval is present because a model may misread, overgeneralize, or combine passages in unsupported ways. Research on factuality in summarization and atomic factual precision shows that fluent text can contain small but consequential distortions, especially when outputs are long or synthesize multiple sources [2, 29]. Contrastive evidence methods further indicate that systems must learn to distinguish supporting evidence from plausible but non-

supporting evidence [30]. In regulatory intelligence, the human risk is over-reliance: users may trust a cited answer without checking whether the citation actually supports the claim.

Standardization and Interoperability

Standardization remains underdeveloped across regulatory RAG systems. Different implementations use different chunking strategies, metadata schemas, retrieval pipelines, citation formats, and evaluation criteria, making results difficult to compare across organizations or vendors. RAG surveys describe this fragmentation as a general problem for the field, while healthcare applications show that domain-specific interoperability is necessary for safe workflow integration [25, 26]. Pharmaceutical regulatory intelligence will require shared conventions for source versioning, jurisdiction tagging, document hierarchy, claim attribution, and expert-review status.

Future Directions and Emerging Paradigms

Integration with Knowledge Graphs

The next stage of regulatory RAG is likely to move beyond flat passage retrieval toward hybrid systems that combine documents with structured knowledge graphs. Such systems could represent relationships among guidance requirements, product attributes, study designs, quality commitments, safety signals, and regional labeling obligations. Medical RAG benchmarks and healthcare frameworks already suggest that complex questions benefit from combining retrieved text with domain structure and expert validation [1, 9]. For pharmaceutical regulatory intelligence, knowledge graphs may help distinguish related but non-equivalent concepts, such as recommendation, requirement, commitment, deviation, and precedent.

Real-Time Guideline Updates and Continuous Learning

Regulatory knowledge changes continuously as agencies revise guidance, publish question-and-answer documents, issue safety communications, and update review expectations. RAG is well suited to this dynamic environment because updating the retrieval corpus can be safer and more transparent than retraining an entire language model. Surveys of retrieval-augmented language models emphasize this advantage: external memory allows systems to incorporate new information through controlled indexing and retrieval workflows [25, 26]. In pharmaceutical use, continuous learning should therefore mean governed corpus refresh, version tracking, and re-validation rather than uncontrolled model adaptation.

Toward Autonomous Regulatory Assistants

The long-term vision is an autonomous regulatory assistant that can monitor new guidance, identify affected products or documents, draft source-grounded recommendations, and route them for expert review. Early pharmaceutical RAG systems for immunogenicity querying, protocol assessment, and compliance evaluation show the first steps toward this broader assistant model [7, 8, 23]. Clinical and surgical RAG frameworks similarly demonstrate how retrieval-conditioned generation can become part of decision-support workflows when evidence, constraints, and review processes are explicit [21, 24]. The credible path to autonomy is therefore incremental: systems must first become reliable collaborators before they can manage submission-readiness tasks with limited supervision.

Strengths and Limitations of this Narrative

This narrative review offers a chronological synthesis of how RAG evolved from a general solution for knowledge-intensive natural language processing into an emerging architecture for pharmaceutical regulatory intelligence. Its strength is breadth: it connects foundational retrieval architectures [4, 5], vector search and re-ranking methods [10-12], factuality and verification research [3, 16, 17, 29, 30], and recent pharmaceutical or healthcare RAG applications [1, 7-9, 13, 18, 21, 23, 24, 27, 28]. Its limitation is that it is intentionally narrative rather than systematic, so it does not quantify performance, conduct meta-analysis, or claim exhaustive coverage of all regulatory AI systems. This framing is appropriate for a developing field whose central questions concern trust, traceability, and governance as much as benchmark accuracy.

Conclusion

RAG has traveled a notable path from a technical response to hallucination toward an emerging foundation for regulatory intelligence. Its central contribution is the simple but powerful idea that generated answers should be grounded in retrievable, inspectable evidence. In pharmaceutical regulation, that idea aligns closely with professional expectations for documentation, justification, and accountability. The result is a technology that feels unusually well matched to the needs of guidance interpretation and controlled decision support.

The key advances have been retrieval quality, grounding discipline, traceability, and verification. Dense retrieval made semantic access to large document collections practical, while citation mechanisms made answers easier to inspect. Fact-checking and revision layers added further safeguards against unsupported synthesis. Together, these developments moved RAG from a conversational interface toward a governed knowledge-management architecture.

Important gaps remain. RAG systems can still retrieve the wrong passages, miss jurisdictional nuance, misinterpret ambiguous language, or produce answers that appear more authoritative than they are. Human oversight remains essential, particularly

when outputs influence submissions, labeling, compliance decisions, or interactions with regulators. The most responsible systems will be those that make uncertainty visible and preserve expert accountability.

The next priority for the field is to formalize evaluation standards for regulatory RAG. Shared benchmarks, source-versioning practices, audit expectations, and expert-review protocols will be needed before autonomous regulatory assistants can be trusted more broadly. Progress should therefore be ambitious but disciplined, advancing automation while preserving the evidentiary culture that pharmaceutical regulation requires. Responsible regulatory AI will not replace judgment; it will make knowledge work more traceable, current, and accountable.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. Findings of the Association for Computational Linguistics: ACL 2024. 2024:6233–51.
2. Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:1906–19.
3. Kryściński W, McCann B, Xiong C, Socher R. Evaluating the factual consistency of abstractive text summarization. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020:9332–46.
4. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inf Process Syst. 2020;33:9459-74.
5. Guu K, Lee K, Tung Z, Pasupat P, Chang M. Retrieval augmented language model pre-training. In: International conference on machine learning. PMLR; 2020:3929–38.
6. Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). 2020:6769–81.
7. Advani N, Bhat AG, Balu-Iyer S, Ramanathan M. Retrieval augmented generation (rag) for natural language querying of immunogenicity data for protein drugs. AAPS J. 2026;28(2):51.
8. Waikar S, Bhat AG, Ramanathan M. Retrieval Augmented Generation (RAG) for Evaluating Regulatory Compliance of Drug Information and Clinical Trial Protocols. CPT Pharmacometrics Syst Pharmacol. 2026;15(3):e70201.
9. Yang R, Ning Y, Keppo E, Liu M, Hong C, Bitterman DS, et al. Retrieval-augmented generation for generative artificial intelligence in health care. NPJ Health Syst. 2025;2(1):2.
10. Khattab O, Zaharia M. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020:39–48.
11. Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. IEEE Trans Big Data. 2019;7(3):535-47.
12. Malkov YA, Yashunin DA. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Trans Pattern Anal Mach Intell. 2018;42(4):824-36.
13. Nanua S, Steward R, Neely B, Datto M, Youens K. Retrieval-augmented generation for interpreting clinical laboratory regulations using large language models. J Pathol Inform. 2025;100520.
14. Izacard G, Grave E. 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. 2021:874–80.
15. Borgeaud S, Mensch A, Hoffmann J, Cai T, Rutherford E, Millican K, Silver D, et al. Improving language models by retrieving from trillions of tokens. Proceedings of the 39th International Conference on Machine Learning (ICML). 2022:2206–40.
16. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a large-scale dataset for fact extraction and verification. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 2018:809–19.
17. Laban P, Schnabel T, Bennett PN, Hearst MA. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. Trans Assoc Comput Linguist. 2022;10:163-77.
18. Neha F, Bhati D, Shukla DK. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. AI. 2025;6(9):226.
19. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824-37.

20. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171. 2022.
21. Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval-augmented generation–enabled GPT-4 for clinical trial screening. NEJM AI. 2024;1(7):AIoa2400181.
22. Gao L, Dai Z, Pasupat P, Chen A, Chaganty AT, Fan Y, 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). 2023:16477–508.
23. Kim J, Hur M, Min M. From rag to qa-rag: Integrating generative ai for pharmaceutical regulatory compliance process. In: Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing. 2025:1293–5.
24. Ong CS, Obey NT, Zheng Y, Cohan A, Schneider EB. SurgeryLLM: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. NPJ Digit Med. 2024;7(1):364.
25. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997. 2023;2(1):32.
26. Huang Y, Huang JX. A survey on retrieval-augmented text generation for large language models. ACM Comput Surv. 2024.
27. Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia AT, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. 2025;8(1):187.
28. Wada A, Tanaka Y, Nishizawa M, Yamamoto A, Akashi T, Hagiwara A, et al. Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. NPJ Digit Med. 2025;8(1):395.
29. Min S, Krishna K, Lyu X, Lewis M, Yih WT, Koh P, 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. 2023:12076–100.
30. Schuster T, Fisch A, Barzilay R. Get your vitamin C! robust fact verification with contrastive evidence. In: Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies. 2021:624–43.