Pharmacophore

ISSN-2229-5402

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

RETRIEVAL-AUGMENTED LANGUAGE MODELS FOR CHECKING SAFETY CONSISTENCY IN PHARMACEUTICAL PRODUCT LABELS

Claire Dupont[1]*, Julien Martin[1]

1. *Department of Computational Pharmacy and Therapeutics, Faculty of Pharmacy, University of Bordeaux, Bordeaux, France.*

ARTICLE INFO

Received:
01 February 2025
Received in revised form:
28 April 2025
Accepted:
29 April 2025
Available online:
28 June 2025

Keywords: Retrieval-augmented generation, Pharmaceutical product label, Regulatory science, Pharmacovigilance, Safety consistency, Contradiction detection

ABSTRACT

Pharmaceutical labels serve as the authoritative source of approved drug safety information, guiding prescribing, dispensing, monitoring, and patient counseling; however, safety statements are often scattered across warnings, adverse reactions, contraindications, and interaction sections, creating potential internal inconsistencies. Manual cross-checking of all safety-relevant sections is slow, repetitive, and subject to reviewer variability, while keyword searches alone cannot reliably detect semantic equivalence, missing safety concepts, or contradictions expressed with different clinical terminology. To address these challenges, this article proposes a conceptual retrieval-augmented language model system that ingests the full product label, indexes safety-relevant statements, and answers consistency queries by retrieving and comparing authoritative passages. The system integrates a structured label parser, section-aware chunking, semantic indexing, a vector database, retrieval-augmented answer generation, contradiction detection, and a human review interface, with each generated consistency judgment linked to the supporting label passages. Such an approach could accelerate label review, enhance traceability, and help reviewers identify discrepancies that might otherwise remain hidden in lengthy regulatory documents, provided it is carefully grounded, expert-validated, deployed in a privacy-preserving manner, and governed appropriately. By combining semantic retrieval with human adjudication, retrieval-augmented label consistency checking has the potential to become a practical tool for regulatory affairs and pharmacovigilance teams, supporting continuous surveillance of safety information integrity.

To Cite This Article: Dupont C, Martin J. Retrieval-Augmented Language Models for Checking Safety Consistency in Pharmaceutical Product Labels. Pharmacophore. 2025;16(3):1-11. https://doi.org/10.51847/FZeUaxELuK

Introduction

Pharmaceutical product labels communicate the approved evidence-based safety profile of a medicinal product and therefore occupy a central role in prescribing decisions, risk minimisation, and pharmacovigilance governance. Computational work on label annotation has shown that safety content in approved labeling can be structured and reused for downstream surveillance, as illustrated by systems that annotate adverse events in drug product labeling with MedDRA terminology [1]. More recent labeling-focused language models such as RxBERT have further emphasised the importance of section-sensitive analysis of regulatory text, because the same clinical concept may carry different regulatory meaning depending on whether it appears in warnings, adverse reactions, or interactions [2]. A consistency-checking system must therefore treat the label not as free text alone, but as a structured safety document whose internal coherence directly affects patient-facing and prescriber-facing communication.

Current label authoring and review processes still depend heavily on manual cross-referencing across sections, especially when post-marketing evidence prompts updates to warnings, adverse reactions, contraindications, or interaction language. FDA-oriented tools such as LabelComp demonstrate that artificial intelligence can assist reviewers by identifying adverse-event changes in labeling, which indicates a broader opportunity for automated support during safety text revision [3]. Earlier pharmacovigilance extraction pipelines also showed that label text can be transformed into structured safety concepts, but these systems were primarily designed to detect or classify safety mentions rather than to reason over whether two sections are mutually consistent [4, 5]. The proposed EAI focus is therefore not automated labeling replacement, but an expert-assistive layer that highlights where human review should concentrate.

Retrieval-augmented language models are well matched to this problem because they can condition their answers on passages retrieved from a specific authoritative document rather than relying only on parametric model knowledge. Biomedical RAG

Corresponding Author: Claire Dupont. Department of Computational Pharmacy and Therapeutics, Faculty of Pharmacy, University of Bordeaux, Bordeaux, France. E-mail: claire.dupont@gmail.com

systems such as BiomedRAG show how retrieval can connect large language models to domain corpora for biomedicine, while broader reviews describe RAG as a practical strategy for improving biomedical LLM applications through grounding and traceability [6, 7]. Benchmarking work on retrieval-augmented biomedical language models also underscores the need to examine robustness and self-awareness, which are particularly important when a system must state that evidence is insufficient rather than invent a regulatory conclusion [8]. These properties align closely with label review, where every consistency claim must be anchored in approved source text.

This article proposes a specialised retrieval-augmented language model system that embeds safety statements from a drug label into a retrievable knowledge base and uses structured prompts to compare related statements across sections. The concept builds on drug-label-specific models such as PharmBERT, which demonstrate the value of domain adaptation for label language, and on general biomedical models such as BioBERT and PubMedBERT, which support semantic representation of clinical and biomedical terminology [9-11]. It also draws on scientific fact-verification and rationale-generation work, where claim checking depends on retrieved evidence and interpretable justifications rather than ungrounded assertions [12, 13]. The central thesis is that a RAG system could help regulatory professionals detect, review, and document potential safety inconsistencies while preserving human authority over final labeling decisions.

Background

Pharmaceutical Product Labels as Structured Safety Documents

Pharmaceutical product labels are structured regulatory documents that organise safety information into predictable sections such as Boxed Warning, Contraindications, Warnings and Precautions, Adverse Reactions, and Drug Interactions. The development of annotated Structured Product Label resources has shown that these sections can be treated as machine-readable sources for adverse drug reaction identification and safety knowledge extraction [14]. Section classification studies using regulatory labeling documents further demonstrate that automated systems can identify predefined label sections, a prerequisite for comparing a warning statement with an adverse reaction listing or an interaction statement [15]. A RAG-based label consistency system would use this structure to retrieve evidence from the appropriate section rather than collapsing all label language into an undifferentiated text stream.

Safety Information Inconsistencies and Their Consequences

Safety inconsistencies in labeling may arise when new post-marketing information is added to one section but not propagated to related sections, or when similar risks are expressed with incompatible levels of specificity. Label comparison tools for adverse event changes illustrate the practical need to track how safety language evolves across versions and sections, because missing or mismatched updates can affect review quality and downstream safety communication [3]. Work on automated product label annotation also suggests that label content is sufficiently complex that terminology normalisation and expert oversight are needed to avoid misinterpretation of adverse-event statements [1]. A system that identifies potential discrepancies would be expected to reduce review burden, but it should present findings as candidates for regulatory judgement rather than definitive determinations.

NLP and Information Extraction from Drug Labels

NLP methods for drug labels have progressed from rule-based recognition and dictionary mapping toward machine-learning systems that extract adverse reactions, relations, and safety concepts from Structured Product Labels. ADE Eval provided a comparative framework for evaluating adverse-event extraction from drug labels, showing the importance of systematic assessment when label text is used for pharmacovigilance [4]. Other work has applied event-based extraction to Structured Product Labels, while relation extraction studies have combined expert, crowd, and machine input to construct structured knowledge from DailyMed labels [5, 16]. Knowledge-base construction efforts such as OnSIDES extend this trajectory by extracting adverse drug events from labels, supporting the idea that label content can be computationally represented for consistency analysis [17].

Retrieval-Augmented Generation for Regulatory Text

Retrieval-augmented generation is especially relevant for regulatory text because it can restrict language-model reasoning to retrieved passages from an approved corpus. Biomedical RAG research has proposed architectures in which retrieval and generation are coupled to answer domain questions with supporting evidence, and systematic reviews have described RAG as a way to improve factual grounding in biomedical LLM applications [6, 7]. Clinical and biomedical summarisation studies have also shown that large language models can produce useful text but require careful evaluation for fidelity to source evidence [18, 19]. In label consistency checking, the RAG layer would therefore serve as a guardrail that keeps the answer tied to the current approved label rather than to general biomedical knowledge. A further design consideration is that label consistency checking should not depend on generation quality alone, but on the quality of retrieval, evidence selection, and verification logic. FDA labeling has been described as a rich regulatory-science resource containing structured information on safety, efficacy, interactions, and precision-medicine use cases, which supports treating labels as an authoritative evidence corpus rather than ordinary narrative text [20]. General RAG methods show that combining parametric language models with retrieved non-parametric evidence can improve knowledge-intensive generation, while dense passage retrieval and retrieval-augmented pretraining demonstrate why semantic retrieval is central when relevant evidence may be expressed through

different wording [21–23]. For safety consistency review, this retrieval layer should be paired with claim-verification principles, because fact-checking benchmarks require systems to distinguish supported, refuted, and insufficiently evidenced claims rather than forcing a definitive answer [24]. Recent FDA-labeling information-extraction work also supports the practical value of NLP pipelines that integrate multiple public labeling resources and structured paragraph classification, reinforcing the need for section-aware retrieval before any generated regulatory judgement is produced [25].

Fact-Checking and Contradiction Detection in Biomedicine

Fact-checking and contradiction detection require models to compare a claim against evidence, determine whether the evidence supports, refutes, or fails to address the claim, and provide a rationale for that judgement. Scientific claim verification work such as SciFact provides a conceptual foundation for evidence-grounded verification, while public-health claim checking demonstrates the importance of explainable outputs in safety-relevant contexts [12, 26]. Rationale benchmarks such as ERASER reinforce that explanations should point to evidence spans, not merely produce a plausible answer [13]. In pharmaceutical labeling, these ideas translate into a system that compares safety statements across sections and explicitly identifies whether the retrieved text supports consistency, suggests discrepancy, or remains inconclusive.

System Architecture Overview

High-Level Design

The proposed system operates in two complementary modes: a batch consistency audit that systematically checks predefined safety assertion pairs across label sections, and an interactive query mode that allows a regulatory professional to ask a natural-language question about a specific risk. Label-focused AI systems such as AskFDALabel show how users can query FDA labeling documents through language-model interfaces, suggesting that interactive access to regulatory text is feasible when the model is anchored in labeling content [27]. In batch mode, the same retrieval and comparison logic would be applied repeatedly to risk topics such as hepatotoxicity, QT prolongation, hypersensitivity, renal impairment, or embryo-fetal toxicity. In interactive mode, the system would return a concise answer with cited passages and a recommendation for human adjudication when the evidence is conflicting or incomplete.

Core Data Flow

The core data flow begins with label XML or PDF ingestion, followed by structured parsing, section segmentation, chunking, embedding, and storage in a retrievable vector index. Section classification and regulatory text mining methods support the need to preserve section identity during preprocessing, because a retrieved adverse-reaction table entry has a different interpretive role from a warning paragraph [2, 15]. At query time, the system embeds the user's question, retrieves relevant chunks, optionally re-ranks them, and conditions the language model on those chunks to generate a source-grounded response. The output would contain the judgement category, a short explanation, and links to the exact label passages used in the comparison.

Figure 1 illustrates the proposed retrieval-augmented safety consistency checking architecture for pharmaceutical product labels, showing how label sections are parsed, indexed, retrieved, compared, cited, and routed to human regulatory review.

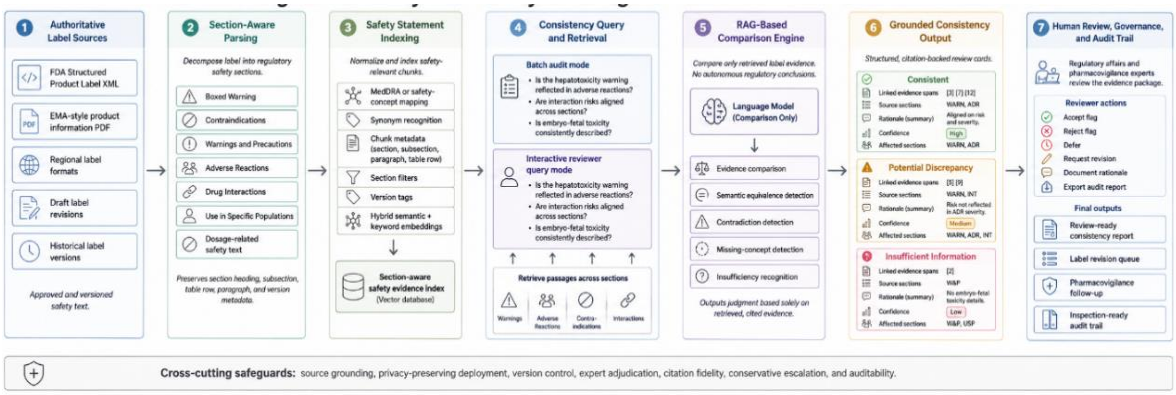

Figure 1. Retrieval-Augmented Safety Consistency Checking Architecture for Pharmaceutical Product Labels

Design Principles

The design principles are source grounding, semantic comparison, human-in-the-loop adjudication, and privacy-respecting deployment for sensitive or pre-market labeling materials. Privacy-preserving medical information retrieval research demonstrates that LLM-enabled retrieval systems can be designed with security and confidentiality in mind, which is essential when drafts or proprietary labels are involved [28]. Implementation guidance for healthcare LLMs similarly emphasises balancing control, collaboration, cost, and security, all of which are relevant to regulated pharmaceutical workflows [29]. The system should therefore avoid autonomous regulatory conclusions and instead provide review-ready evidence packages that human experts can accept, reject, or refine.

Label Document Ingestion and Indexing

Parsing Structured and Unstructured Labels

The ingestion layer would extract text from FDA Structured Product Label XML, EMA-style product information PDFs, and other regional formats while preserving section headings, subsections, tables, lists, and cross-references. Annotated SPL datasets demonstrate that product labels can be converted into structured corpora suitable for adverse reaction identification, but they also highlight the importance of maintaining the link between extracted text and its original label context [14]. Studies that classify regulatory text into predefined sections further support mapping diverse label layouts into a common ontology so that safety statements can be compared across equivalent sections [15]. This common representation is necessary because a consistency query should retrieve the Warnings and Precautions statement, the Adverse Reactions entry, and the Drug Interactions language as separate but related evidence objects.

Chunking and Semantic Annotation

After parsing, the system would split the label into paragraph-level, table-row-level, or entity-focused chunks while attaching metadata such as section type, drug name, indication, population, safety topic, and version. Drug-label-specific models such as PharmBERT indicate that language from product labels has distinctive terminology and structure, supporting the need for label-aware semantic annotation rather than generic document chunking [9]. MedDRA-based annotation work also shows that safety concepts in labels often require terminology normalisation, because the same clinical event may be expressed through synonyms or hierarchical variants [1]. Metadata-rich chunks would allow filtered retrieval, such as limiting a hepatotoxicity query to warnings, adverse reactions, and interactions rather than retrieving unrelated pharmacology text.

Embedding and Vector Store Construction

Each chunk would be embedded using a biomedical or pharmaceutical language model and stored in a vector database that supports semantic retrieval, keyword constraints, and section filters. BioBERT provides a foundation for biomedical text mining, PubMedBERT demonstrates the value of domain-specific pretraining from biomedical literature, and RxBERT extends this principle toward drug-labeling analysis [2, 10, 11]. The vector store would be expected to capture similarity between expressions such as “liver injury,” “hepatitis,” and “transaminase elevations,” while metadata filters would preserve regulatory context. Hybrid search would be preferable because exact regulatory phrases, controlled terminology, and semantic paraphrases are all important in label review.

Continuous Update with Label Versions

The indexing layer should preserve version history so that outdated chunks can be distinguished from the current approved label and from pending draft revisions. Label comparison work such as LabelComp shows the value of identifying changes in adverse-event language across labeling versions, which is directly relevant to avoiding stale evidence in consistency review [3]. Systems such as OnSIDES also depend on extracting structured safety information from labels, making currency and provenance central to the reliability of downstream knowledge bases [17]. A version-aware RAG system would therefore retrieve from the current approved text by default while allowing reviewers to compare proposed updates against prior language when needed.

Table 1 presents the section-aware retrieval architecture required to convert pharmaceutical labels into traceable evidence objects for safety consistency checking.

Table 1. Section-Aware Retrieval Architecture for Safety Consistency Checking Across Pharmaceutical Product Labels

System Layer	Regulatory Function	Required Label Representation	Consistency-Checking Contribution	Failure Risk if Weakly Designed	Reviewer-Facing Output
Label ingestion layer	Converts approved, draft, and historical labels into analyzable text	Preserved document source, region, product, version, approval status, and file format	Establishes the authoritative corpus from which all evidence must be retrieved	Stale or incorrect label versions may be used as evidence	Current-label corpus with explicit version provenance
Section parser	Separates safety text into regulatory sections	Boxed Warning, Contraindications, Warnings and Precautions, Adverse Reactions, Drug Interactions, Use in Specific Populations, and dosage-related safety sections	Enables comparison of related safety claims across their correct regulatory locations	Safety statements may be interpreted without section context	Section-tagged passages ready for review
Chunking and metadata layer	Converts long label sections into retrievable evidence units	Paragraphs, table rows, list items, subsection headings, cross-references, and sentence-level spans	Allows precise retrieval of the safety passage supporting or	Overly broad chunks may hide contradictions; overly small chunks	Passage-level evidence cards with section and location metadata

			contradicting a claim	may lose clinical context	
Safety concept normalization	Aligns related safety terminology	Preferred terms, synonyms, clinical variants, laboratory abnormalities, mechanisms, and population qualifiers	Helps detect semantic overlap between differently worded risks	Lexically different but clinically related risks may be missed	Normalized safety-topic clusters, such as hepatotoxicity or QT prolongation
Hybrid search index	Stores retrievable safety evidence	Dense embeddings, keyword constraints, section filters, and version filters	Combines semantic similarity with exact regulatory phrase retrieval	Pure keyword search may miss paraphrases; pure semantic search may overretrieve irrelevant text	Ranked candidate passages for each query
Retrieval and reranking layer	Selects evidence passages for comparison	Query-specific safety topic, target sections, retrieved chunks, and ranking scores	Prioritizes the most relevant passages before generation	The language model may compare incomplete or irrelevant evidence	Evidence bundle grouped by label section
RAG comparison layer	Generates consistency judgement from retrieved passages	Retrieved text only, structured prompt, judgement schema, and citation requirements	Determines whether sections are aligned, discrepant, or insufficiently informative	Ungrounded generation may create false regulatory conclusions	Structured judgement with rationale and cited evidence
Human adjudication interface	Places AI-generated flags under expert control	Retrieved passages, model rationale, confidence category, affected sections, and reviewer action log	Converts model output into reviewable regulatory work product	Users may overtrust or dismiss AI output without structured review	Accept, reject, defer, request revision, or document rationale
Audit and governance layer	Records system and reviewer decisions	Query, retrieved passages, generated judgement, reviewer decision, timestamp, user role, and label action	Supports reproducibility, inspection readiness, and governance oversight	Inability to reconstruct why a flag was raised or dismissed	Inspection-ready audit trail and consistency review report

Retrieval-Augmented Consistency Checking Engine

Query Interpretation and Decomposition

The consistency engine would first interpret a reviewer's query, identify the safety topic, and decompose the request into retrieval tasks targeting the relevant label sections. Prompt-based extraction and machine reading comprehension methods in clinical NLP suggest that natural-language questions can be transformed into structured information needs, which is useful when a reviewer asks whether a warning is reflected elsewhere in the label [30]. Clinical named-entity recognition with large language models also supports the idea that a query can be normalised to key biomedical concepts before retrieval [31]. For example, a hepatotoxicity query would trigger retrieval from Warnings and Precautions, Adverse Reactions, Contraindications, and Drug Interactions, followed by section-aware comparison.

Retrieval and Ranking

Retrieval would combine dense semantic search with re-ranking so that passages using clinically related but lexically different terminology are prioritised for comparison. Biomedical RAG benchmarking stresses that retrieval quality, robustness, and model self-awareness are central to reliable domain-specific question answering [8]. BiomedRAG likewise frames retrieval as the mechanism that connects model outputs to relevant biomedical evidence rather than leaving answers to unsupported generation [7]. In the label setting, this means that "drug-induced liver injury," "hepatitis," and "elevated aminotransferases" could be retrieved as related safety evidence while still being presented with their original section provenance.

LLM-Based Contradiction Detection

After retrieval, the language model would compare the evidence passages and generate a structured judgement such as "consistent," "potential discrepancy," or "insufficient information." Medical LLM research shows that large language models can encode clinical knowledge, but clinical decision-making evaluations also caution that limitations must be mitigated before use in safety-critical settings [32, 33]. For that reason, the model should be prompted to justify each judgement using retrieved source sentences and to avoid resolving regulatory ambiguity without human review. The contradiction module would not claim that the label is wrong; it would identify where the retrieved evidence appears mismatched, absent, or semantically divergent.

Handling Semantic Variability

Safety consistency checking must handle semantic variability because labels may describe the same risk through symptoms, diagnoses, laboratory abnormalities, mechanisms, or population-specific language. Public-health fact-checking and scientific claim-verification work show that evidence comparison must account for paraphrase and implication, not only exact word overlap [12, 26]. In biomedical QA resources such as PubMedQA, models are similarly challenged to infer whether evidence supports, contradicts, or does not answer a biomedical claim [34]. A label-focused system would apply this reasoning to regulatory sections, recognising that "increased transaminases" may be relevant to a hepatotoxicity warning while still requiring expert judgement about whether the adverse reaction language is sufficiently aligned.

Answer Grounding, Citation, and Conflict Resolution

In-Line Citation and Source Traceability

Every assertion in the system output should be linked to the exact label section, paragraph, table row, or sentence from which it was retrieved, because traceability is essential when generated language is used in regulated safety review. Evidence-grounded summarisation studies show that useful biomedical language generation must remain aligned with the source passages rather than merely sounding clinically plausible [18]. Rationale-oriented NLP benchmarks similarly support the principle that a system should expose the evidence span behind a conclusion, especially when a reviewer must decide whether a generated explanation is valid [13]. In practice, the answer interface would allow a reviewer to move directly from a consistency judgement to the source statements in Warnings and Precautions, Adverse Reactions, Drug Interactions, or Contraindications.

Handling Conflicting or Ambiguous Evidence

When retrieved passages appear to disagree, the system should flag the conflict explicitly and avoid converting ambiguity into an artificial conclusion. Scientific claim-verification methods distinguish between supported, refuted, and insufficiently evidenced claims, a distinction that maps naturally to label consistency checking when one section contains a risk statement and another section is silent or less specific [12]. Public-health fact-checking work also highlights the need for explainability when automated systems assess safety-related claims, because unsupported certainty could mislead users [26]. A regulatory RAG system should therefore state that the evidence is conflicting, incomplete, or inconclusive and route the issue to human review.

Confidence Scoring and Flagging for Human Review

The system could assign confidence categories to consistency judgements based on retrieval coverage, semantic agreement between passages, and the clarity of the generated rationale, but these categories should be treated as prioritisation aids rather than regulatory determinations. Biomedical RAG benchmarking emphasises that models should be evaluated for robustness and self-awareness, including their ability to recognise when retrieved evidence is insufficient [8]. Clinical LLM evaluation studies further caution that model outputs in safety-critical contexts require mitigation strategies and expert oversight before they can support real-world decisions [33]. Low-confidence or high-impact flags would therefore be escalated to regulatory affairs or pharmacovigilance experts for adjudication.

Human-in-the-Loop Validation and Label Governance

Collaborative Review Dashboard

A collaborative dashboard would present each flagged inconsistency with the retrieved source passages, the model's structured rationale, the affected label sections, and reviewer actions such as accept, reject, defer, or request revision. Human-in-the-loop design is consistent with clinical information extraction research showing that LLMs can support expert workflows while still requiring verification of extracted concepts and contextual meaning [31]. Work using large language models to identify social determinants of health in electronic health records also illustrates the broader principle that domain experts remain essential when model outputs concern nuanced clinical interpretation [35]. For label governance, the dashboard should make review decisions easy to document without obscuring the distinction between AI-generated suggestions and expert regulatory judgement.

Feedback Loop for Model Improvement

Accepted and rejected flags would provide domain-specific correction data that could be used to refine prompts, retrieval filters, section mappings, and model behaviour over time. Drug-label relation extraction research has shown that combining expert input with machine methods can improve structured knowledge creation from DailyMed labels, suggesting a practical foundation for iterative expert correction [16]. Prompt-based clinical relation extraction also demonstrates that information needs can be encoded and refined through task-specific instructions rather than relying only on general-purpose model behaviour [30]. In this system, feedback should be curated and governed so that future updates improve consistency checking without weakening source grounding or regulatory conservatism.

Audit Trail for Regulatory Submissions

All queries, retrieved passages, generated judgements, reviewer decisions, and subsequent label actions should be logged to create an audit trail suitable for internal governance, inspection readiness, and regulatory submission support. Healthcare LLM

implementation guidance stresses the importance of control, collaboration, cost, and security, which implies that model-assisted decisions must be reproducible and reviewable rather than opaque [29]. Privacy-preserving information retrieval work further supports secure handling of structured medical content, a necessary condition when draft labels, confidential submissions, or pre-approval safety information are processed [28]. The audit trail should therefore record both the evidence used by the system and the human decision that determined whether a label change was warranted.

Integration into Regulatory and Pharmacovigilance Workflows

Pre-Submission Label Review

In pre-submission workflows, the system would be run as a standard consistency audit before final label approval, producing a review report that complements manual cross-checking. Label comparison methods such as LabelComp show that AI can help identify adverse-event language changes in FDA labeling, supporting the feasibility of automated review assistance during label revision [3]. Section-classification approaches for regulatory documents also indicate that AI can organise label text into reviewable components, which would help teams confirm that proposed safety language is reflected consistently across sections [15]. The expected role is to accelerate reviewer focus and documentation, not to replace medical, legal, or regulatory sign-off.

Post-Marketing Safety Label Surveillance

In post-marketing pharmacovigilance, the system could support safety signal follow-up by checking whether proposed label updates are propagated coherently across warnings, adverse reactions, contraindications, and interaction language. Adverse-event extraction systems for labeling, including ADE Eval and OnSIDES, show how structured safety concepts can be derived from product labels and reused for pharmacovigilance-oriented analysis [4, 17]. AskFDALabel also illustrates how question-answering over FDA labeling documents can make regulatory text more accessible for safety review tasks [27]. When a new risk signal prompts label revision, the RAG system could identify which sections may require aligned updates and provide citations for expert review.

Evaluation Strategy

Detection Accuracy

The system should be evaluated conceptually through expert-created test cases containing known or seeded label inconsistencies, with reviewers assessing whether the system identifies the relevant discrepancy category. Prior label NLP evaluations demonstrate the need for structured assessment when extracting adverse events or mapping labeling language to standard terminology [1, 4]. Because this EAI proposal does not report experiments or performance results, detection accuracy should be framed as a future evaluation dimension rather than as a measured outcome. The key evaluation question is whether the system could reliably distinguish consistent safety statements from potential discrepancies and insufficient evidence under expert-defined criteria.

Retrieval and Answer Quality

Retrieval and answer quality should be evaluated by examining whether the system retrieves the correct label sections, uses the retrieved passages faithfully, and cites the evidence that actually supports the generated judgement. Biomedical evidence summarisation research shows that answer quality depends not only on fluent generation but also on faithfulness to the source material [18]. Retrieval-augmented biomedical LLM studies similarly indicate that retrieval relevance and answer grounding should be assessed together, because a plausible answer can still be unsafe if it is not supported by the retrieved evidence [6, 7]. Reviewers should therefore evaluate source selection, citation fidelity, and whether the generated rationale accurately reflects the label passages. **Figure 2** shows the proposed evaluation logic for assessing whether the retrieval-augmented label-review system identifies seeded inconsistencies, retrieves the correct label evidence, and generates citation-grounded consistency judgements.

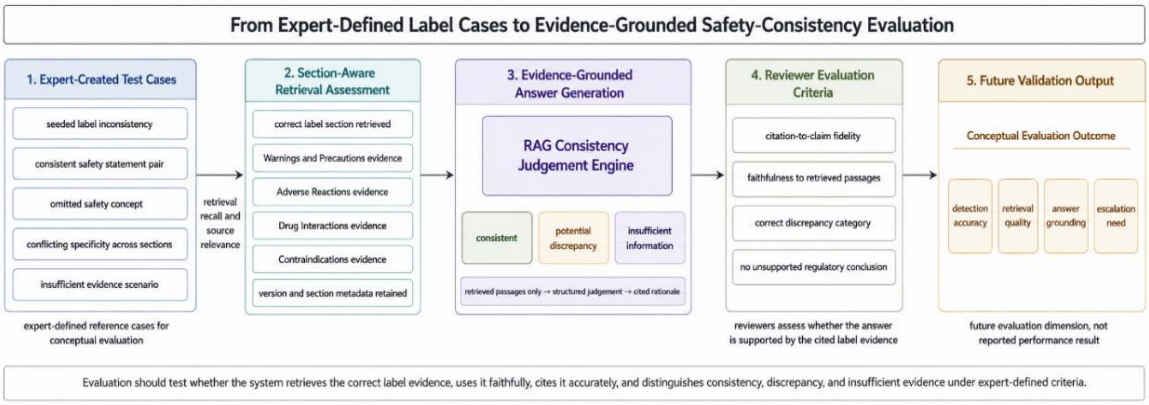

Figure 2. Evaluation Logic for Retrieval Accuracy, Answer Grounding, and Safety-Consistency Detection

Workflow Efficiency and User Satisfaction

Workflow evaluation should examine whether the system helps regulatory professionals complete consistency review more efficiently, understand flagged issues more clearly, and document decisions more consistently. Studies of clinical LLM applications show that model usefulness depends on fit with professional workflow and expert trust, not only technical capability [19, 32]. Healthcare implementation guidance also emphasises that collaboration and operational control are central to responsible deployment of LLM systems [29]. Evaluation should therefore include user-centred review of dashboard usability, alert burden, adjudication clarity, and the extent to which AI-generated flags contribute to meaningful label governance.

Table 2 defines an evaluation and governance framework for determining whether retrieval-augmented label consistency checking is accurate, traceable, and suitable for expert-regulated safety review.

Table 2. Evaluation and Governance Framework for Retrieval-Augmented Label Consistency Checking

Evaluation Domain	Core Question	Practical Metric or Review Criterion	Evidence Required	Governance Threshold	Regulatory or Pharmacovigilance Value
Section retrieval accuracy	Did the system retrieve the correct label sections for the safety topic?	Proportion of expert-identified relevant sections retrieved per query	Gold-standard reviewer annotations for seeded label cases	High recall required for warnings, contraindications, adverse reactions, and interactions	Reduces missed cross-section safety discrepancies
Citation fidelity	Do cited passages actually support the generated judgement?	Reviewer rating of citation-to-claim alignment	Generated answer, cited passages, source section labels	Any unsupported safety claim requires rejection or revision	Prevents plausible but ungrounded regulatory statements
Semantic consistency detection	Can the system recognize equivalent safety concepts expressed differently?	Accuracy for synonym, hierarchy, symptom, diagnosis, laboratory, and mechanism variants	Expert-created cases with paraphrased safety risks	Must distinguish clinically related wording from unrelated similarity	Improves detection beyond keyword matching
Contradiction and omission detection	Can the system identify mismatched, missing, or less specific safety statements?	Classification into consistent, potential discrepancy, or insufficient information	Seeded inconsistency scenarios across label sections	High-impact discrepancies require human escalation regardless of confidence	Prioritizes reviewer attention during label revision
Insufficiency recognition	Does the system know when evidence is inadequate?	Frequency of appropriate “insufficient information” outputs	Queries where label evidence is absent, ambiguous, or incomplete	Unsupported conclusions should be blocked	Reduces hallucinated regulatory conclusions
Version awareness	Does the system distinguish current approved text from draft or historical language?	Correct retrieval from intended label version	Current label, proposed revision, and historical comparison set	Current approved label must be default evidence source	Prevents outdated safety language from shaping review
Reviewer usability	Does the system help experts complete consistency review efficiently?	Time to adjudication, perceived clarity, alert burden, and reviewer satisfaction	User testing with regulatory affairs and pharmacovigilance reviewers	System should reduce review burden without increasing low-value alerts	Supports workflow adoption and expert trust
Human oversight quality	Are final decisions clearly separated from AI suggestions?	Presence of reviewer action, rationale, and final disposition	Audit log and adjudication record	No AI flag should become a label action without expert decision	Maintains human authority over regulated safety communication
Privacy and access control	Are confidential draft labels and	Access logs, role permissions, deployment	Security review and system governance documentation	Proprietary or pre-approval materials	Enables use in sensitive regulatory and

	safety materials protected?	boundary, and data retention controls		require controlled deployment	pharmacovigilance environments
Auditability	Can the full reasoning path be reconstructed later?	Completeness of query, evidence, output, reviewer decision, and timestamp records	System logs and exported review reports	All high-impact flags must be reproducible	Supports inspection readiness and internal quality governance

Limitations

Semantic Complexity and Regulatory Nuance

A major limitation is that regulatory labels intentionally repeat or vary safety language across sections for emphasis, legal precision, population specificity, or clinical context, and a model may initially misclassify these differences as discrepancies. Clinical decision-making evaluations warn that large language models can produce incomplete or misleading interpretations when task nuance exceeds their grounding or reasoning capacity [33]. Medical LLM research also indicates that encoded clinical knowledge does not by itself guarantee safe regulatory reasoning, especially when a conclusion depends on subtle distinctions between adverse-event listing, warning language, and contraindication wording [32]. Consequently, the system should be configured to flag potential issues conservatively and require human experts to determine whether a difference is meaningful.

Label Format Heterogeneity

A second limitation is format heterogeneity, because labels may appear as structured XML, semi-structured PDF, regional product information, historical labeling, or legacy documents with inconsistent tables and section headings. SPL annotation and section-classification studies show that structured parsing is feasible, but they also imply that reliable downstream analysis depends on preserving document organisation during ingestion [14, 15]. Relation extraction from DailyMed labels and adverse-event extraction from Structured Product Labels further demonstrate that information quality depends on the consistency of source formatting and annotation conventions [5, 16]. For older or regionally divergent labels, preprocessing limitations could reduce retrieval coverage and make some consistency judgements incomplete.

Conclusion

Retrieval-augmented language models offer a promising conceptual foundation for checking safety consistency across pharmaceutical product labels. By ingesting the full label, indexing safety-relevant statements, and comparing retrieved passages across sections, such a system could help reviewers identify potential contradictions, omissions, and semantic mismatches.

The proposed architecture is strongest where regulatory review requires traceability, evidence grounding, and expert oversight. Its outputs would be citation-backed suggestions rather than autonomous regulatory decisions, allowing human reviewers to verify each finding against the authoritative label text.

Important challenges remain, including semantic nuance, regional label heterogeneity, privacy requirements, and the need to establish reviewer trust in safety-critical settings. These challenges argue for careful governance, conservative deployment, and structured evaluation before routine operational use.

Pilot implementation within regulatory affairs and pharmacovigilance departments would help define practical requirements for AI-assisted label review. Collaboration among pharmaceutical companies, regulators, informaticians, and safety scientists will be needed to develop standards for retrieval quality, citation fidelity, auditability, and human adjudication.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Ly T, Pamer C, Dang O, Brajovic S, Haider S, Botsis T, et al. Evaluation of natural language processing (NLP) systems to annotate drug product labeling with MedDRA terminology. J Biomed Inform. 2018;83:73-86.
2. Wu L, Gray M, Dang O, Xu J, Fang H, Tong W. RxBERT: enhancing drug labeling text mining and analysis with AI language modeling. Exp Biol Med (Maywood). 2023;248(21):1937-43.

3. Neyarapally GA, Wu L, Xu J, Zhou EH, Dang O, Lee J, et al. Description and validation of a novel AI tool, LabelComp, for the identification of adverse event changes in FDA labeling. Drug Saf. 2024;47(12):1265-74.
4. Bayer S, Clark C, Dang O, Aberdeen J, Brajovic S, Swank K, et al. ADE eval: an evaluation of text processing systems for adverse event extraction from drug labels for pharmacovigilance. Drug Saf. 2021;44(1):83-94.
5. Pandey A, Kreimeyer K, Foster M, Dang O, Ly T, Wang W, et al. Adverse event extraction from structured product labels using the event-based text-mining of health electronic records (ETHER) system. Health Inform J. 2019;25(4):1232-43.
6. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025;32(4):605-15.
7. Li M, Kilicoglu H, Xu H, Zhang R. BiomedRAG: a retrieval augmented large language model for biomedicine. J Biomed Inform. 2025;162:104769.
8. Li M, Zhan Z, Yang H, Xiao Y, Zhou H, Huang J, et al. Benchmarking retrieval-augmented large language models in biomedical NLP: application, robustness, and self-awareness. Sci Adv. 2025;11(47):eadr1443.
9. ValizadehAslani T, Shi Y, Ren P, Wang J, Zhang Y, Hu M, et al. PharmBERT: a domain-specific BERT model for drug labels. Brief Bioinform. 2023;24(4):bbad226.
10. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40.
11. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2021;3(1):1-23.
12. Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or fiction: verifying scientific claims. In: Proc Conf Empir Methods Nat Lang Process; 2020. p. 7534-50.
13. DeYoung J, Jain S, Rajani NF, Lehman E, Xiong C, Socher R, et al. ERASER: a benchmark to evaluate rationalized NLP models. In: Proc Annu Meet Assoc Comput Linguist; 2020. p. 4443-58.
14. Demner-Fushman D, Shooshan SE, Rodriguez L, Aronson AR, Lang F, Rogers W, et al. A dataset of 200 structured product labels annotated for adverse drug reactions. Sci Data. 2018;5(1):180001.
15. Gray M, Xu J, Tong W, Wu L. Classifying free texts into predefined sections using AI in regulatory documents: a case study with drug labeling documents. Chem Res Toxicol. 2023;36(8):1290-9.
16. Shingjergji K, Celebi R, Scholtes J, Dumontier M. Relation extraction from DailyMed structured product labels by optimally combining crowd, experts and machines. J Biomed Inform. 2021;122:103902.
17. Tanaka Y, Chen HY, Belloni P, Gisladottir U, Kefeli J, Patterson J, et al. OnSIDES database: extracting adverse drug events from drug labels using natural language processing models. Med (N Y). 2025;6(7).
18. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. npj Digit Med. 2023;6(1):158.
19. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30(4):1134-42.
20. Fang H, Harris SC, Liu Z, Zhou G, Zhang G, Xu J, et al. FDA drug labeling: rich resources to facilitate precision medicine, drug safety, and regulatory science. Drug Discov Today. 2016;21(10):1566-70.
21. 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.
22. Karpukhin V, Oğuz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense passage retrieval for open-domain question answering. In: Proc Conf Empir Methods Nat Lang Process; 2020. p. 6769-81.
23. Guu K, Lee K, Tung Z, Pasupat P, Chang MW. REALM: retrieval-augmented language model pre-training. In: Proc Int Conf Mach Learn; 2020. p. 3929-38.
24. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a large-scale dataset for fact extraction and verification. In: Proc Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol; 2018. p. 809-19.
25. Shi Y, Zhang X, Mu H, Chen M, Liu Z, Ning B, et al. Information extraction from FDA drug labeling to enhance product-specific guidance assessment. Front Res Metr Anal. 2021;6:670006.
26. Kotonya N, Toni F. Explainable automated fact-checking for public health claims. In: Proc Conf Empir Methods Nat Lang Process; 2020. p. 7740-54.
27. Wu L, Fang H, Qu Y, Xu J, Tong W. Leveraging FDA labeling documents and large language model to enhance annotation, profiling, and classification of drug adverse events with AskFDALabel. Drug Saf. 2025;48(6):655-65.
28. Wiest IC, Ferber D, Zhu J, van Treeck M, Meyer SK, Juglan R, et al. Privacy-preserving large language models for structured medical information retrieval. npj Digit Med. 2024;7(1):257.
29. Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. npj Digit Med. 2025;8(1):143.
30. Peng C, Yang X, Yu Z, Bian J, Hogan WR, Wu Y. Clinical concept and relation extraction using prompt-based machine reading comprehension. J Am Med Inform Assoc. 2023;30(9):1486-93.
31. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc. 2024;31(9):1812-20.

32. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
33. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613-22.
34. Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: a dataset for biomedical research question answering. In: Proc Conf Empir Methods Nat Lang Process Int Jt Conf Nat Lang Process; 2019. p. 2567-77.
35. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, et al. Large language models to identify social determinants of health in electronic health records. npj Digit Med. 2024;7(1):6.