International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Agentic AI with retrieval-augmented generation for automated compliance assistance in finance

Varun Pandey *

IEEE, Budd Lake, NJ, USA.

International Journal of Science and Research Archive, 2025, 15(02), 1620–1631

Publication history: Received on 07 April 2025; revised on 19 May 2025; accepted on 21 May 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.2.1522

Abstract

Maintaining compliance with complex Know Your Customer (KYC) and Anti-Money Laundering (AML) regulations is a resource-intensive challenge for financial institutions. This paper presents an agentic AI approach that leverages Retrieval-Augmented Generation (RAG) to automate and enhance compliance research and decision-making. We define the inefficiencies in current U.S. KYC/AML compliance workflows – including lengthy onboarding times and costly manual processes – as motivation for a more dynamic solution. We then introduce an autonomous agent framework, implemented with LangChain, that integrates a RAG pipeline to perform contextual reasoning over regulatory knowledge bases. The technical architecture is detailed with an emphasis on the agent's planning and tool use capabilities, and the RAG components for knowledge base construction (using U.S. regulations such as FinCEN guidance, Code of Federal Regulations (CFR) provisions, and OFAC sanctions data), transformer-based embedding and indexing, vector retrieval, and LLM-driven answer generation. We demonstrate how this agent can handle compliance queries (e.g., customer due diligence requirements and detection of transaction structuring) in a simulated proof-of-concept. We discuss key advantages of this approach over traditional rule-based or static NLP systems – notably greater adaptability to changing regulations, improved traceability via source citations, and higher precision in complex scenario handling. Finally, we address ethical considerations (hallucination risk, ensuring regulatory accuracy, and model governance) and explore practical applications such as automated audit support, compliance report drafting, and future directions including real-time monitoring and multimodal compliance agents.

Keywords: KYC/AML compliance; Agentic AI; Lang Chain; Retrieval-Augmented Generation; regulatory technology; Financial compliance automation.

1. Introduction

Financial institutions today face mounting pressure to comply with evolving KYC and AML regulations, yet current compliance research practices are often inefficient and costly. Manual compliance processes – from client due diligence to suspicious activity investigations – require extensive time and labor, leading to significant delays and expenses. A 2023 industry report found that banks took an average of 95 days to complete a KYC review (up from 84 days in 2022), and nearly 48% of banks globally have lost clients due to slow or inefficient onboarding procedures. Non-compliance carries severe penalties; for example, in 2024 a major North American bank was fined over $3 billion for failures in its AML program. These challenges underscore the need for more efficient and intelligent compliance assistance tools.

Recent advances in Generative AI offer a promising avenue for transforming compliance workflows. Large Language Models (LLMs) can comprehend and generate human-like text, making them attractive for tasks like interpreting regulations or drafting reports. However, off-the-shelf LLMs alone are insufficient for compliance applications: they lack direct access to up-to-date regulatory knowledge and can produce unverifiable or incorrect answers (hallucinations).

* Corresponding author: Varun Pandey

To overcome these limitations, Retrieval-Augmented Generation (RAG) has emerged as a powerful technique that grounds LLM outputs in relevant external data. By retrieving facts from a curated knowledge base of regulations and guidance, RAG enables LLMs to generate responses with improved accuracy and transparency. In the compliance domain, coupling LLMs with RAG allows the use of authoritative sources (laws, regulations, and filings) to answer complex policy questions, reducing the risk of errors.

Another key development is the concept of agentic AI – autonomous AI agents that can plan, reason, and use tools to accomplish goals. Unlike a static question-answering system, an agentic AI system can dynamically break down tasks, invoke external tools (such as search engines, databases, or calculators), and handle multi-step queries in a goal-directed manner. LangChain, an LLM orchestration framework, facilitates building such agents by providing abstractions for tool use and multi-step reasoning. An *agentic compliance assistant* could, for instance, interpret a broad query (e.g. *"How do I onboard a high-risk customer?"*), decompose it into sub-tasks (identify required verification steps, check sanction lists, retrieve relevant regulatory rules), and execute those using appropriate tools, before synthesizing a final answer.

Problem Definition: This work addresses the inefficiencies in U.S. KYC/AML compliance research and decision-making. Compliance analysts must navigate a vast and fragmented landscape of regulations – from FinCEN rulings and Bank Secrecy Act provisions to OFAC sanctions lists and federal regulatory guidance – often under tight time constraints. Traditional compliance workflows rely on manual lookup or hard-coded rule systems, which struggle to keep pace with regulatory changes and complex scenarios. We identify two primary pain points: (1) Information overload and fragmentation – crucial rules are scattered across multiple documents and agencies, making research laborious; (2) Lack of dynamic reasoning – static rule-based systems cannot easily adapt to novel situations or interpret the nuance in regulations, leading to false negatives/positives and oversight gaps. These shortcomings result in slow responses, high compliance costs, and sometimes missed risks or regulatory violations.

Proposed Solution: We propose an Agentic AI framework with Retrieval-Augmented Generation to serve as an automated compliance assistant. The solution uses a LangChain-based agent that interacts with a domain-specific knowledge base of U.S. financial regulations. By combining autonomous task planning with on-the-fly retrieval of relevant regulatory text, the agent can dynamically reason about compliance questions. For example, given a question about a suspicious transaction pattern, the agent can retrieve the definition of that pattern from FinCEN guidance and apply it to the scenario at hand. Our approach marries the strengths of two paradigms: the agentic framework contributes adaptability and multi-step reasoning, while the RAG pipeline contributes factual grounding and domain precision. This fusion allows the system to not only answer questions using up-to-date regulatory data but also to handle complex queries that may involve multiple steps or external analyses (e.g., checking customer data against sanction lists, calculating thresholds for reporting).

This article is structured as follows: Section II introduces the agentic AI framework and RAG pipeline architecture in detail, including knowledge base construction from regulatory sources, embedding and retrieval mechanisms, and the integration using LangChain. Section III discusses the key advantages of this approach over traditional rule-based or static NLP compliance tools, highlighting improvements in adaptability, traceability, and precision. Section IV presents a simulated proof-of-concept demonstrating how the agent handles example KYC and AML queries (customer onboarding requirements and detecting transaction structuring). Section V addresses ethical considerations and limitations, such as the risk of LLM hallucinations, ensuring regulatory accuracy, and the need for proper model governance in high-stakes domains. Finally, Section VI explores practical applications of the proposed system (from assisting in regulatory audits to automating report drafting) and future directions, including real-time monitoring extensions and multimodal compliance agents.

2. Technical Architecture

2.1. Agentic AI Framework with LangChain

At the core of our system is an agentic AI framework designed to emulate a compliance officer's problem-solving process. The agent operates by receiving a user query or task and autonomously determining how to fulfill it by planning steps and utilizing tools. We implement this using LangChain's agent capabilities, which allow an LLM to act as a reasoning engine that can call tool APIs in an iterative loop of think-act-observe.

The agent uses a planning module to break down complex compliance queries into manageable sub-tasks. This draws inspiration from the Plan-and-Execute paradigm, wherein a *Planner* LLM first formulates a strategy (a sequence of

actions or queries) and an *Executor* LLM (or the same LLM in execution mode) carries out each action step-by-step. In our context, planning might involve steps like: (1) identifying which regulatory domain(s) the question involves (e.g., customer due diligence vs. suspicious transaction reporting), (2) deciding which tool or knowledge source to consult first, and (3) determining how to combine information to form the answer. For example, if asked *"What are the requirements to onboard a new corporate client?"*, the agent might plan to first retrieve KYC regulations (to find identity verification and beneficial ownership rules), then retrieve any OFAC sanction list guidelines (to note if screening is needed), and finally synthesize an answer.

2.2. Tool Use and Integration

We equip the agent with a set of specialized tools:

- Knowledge Base Retriever: This tool interfaces with the vector database of regulatory documents (described in the next subsection) to fetch relevant passages using semantic search.
- Web/Search Tool: Optionally, the agent can perform a web search or query an online database if the knowledge base is insufficient or if real-time information (like the latest sanctions list updates) is needed. (For the scope of this paper, we focus on the internal knowledge base.)
- Analytical Tools: These could include simple calculators for threshold computations or even Python scripts to analyze data. For instance, to evaluate a pattern of transactions for structuring, the agent might use a calculator tool to sum amounts over a period.
- Memory/Scratchpad: LangChain provides an agent scratchpad where the agent can keep track of intermediate reasoning. This helps with chain-of-thought and avoiding looping.

The agent operates in a reasoning loop: it observes the user query, then (internally) prompts itself (the LLM) with a context including available tool descriptions. The LLM's output can be an *action decision* (e.g., "use Tool X with input Y") or a final answer. LangChain's framework parses these decisions and executes the tool actions, then feeds the tool outputs back to the LLM, continuing the loop until the task is complete. This setup enables dynamic task decomposition – the agent can handle multi-part questions by sequentially using tools and accumulating information. Notably, this agentic approach means the system is not confined to a single-turn QA; it can perform interactive workflows like cross-checking an entity against multiple databases or doing a step-by-step analysis of a case.

The LangChain agent approach contrasts with a static pipeline in that the agent can adapt its strategy at runtime. If an initial retrieval yields insufficient information, the agent can reformulate the query or try a different approach (for example, searching a broader regulatory section). This autonomy and adaptability are crucial in the compliance domain, where queries can be nuanced and may not exactly match the wording of any single rule. By incorporating an agent that "figures out" how to answer the query (much like a human compliance analyst would), we achieve a higher degree of flexibility and problem-solving capability. Indeed, agentic workflows in AML compliance have been shown to adapt dynamically to new challenges without manual reprogramming, leading to improved flexibility and accuracy in detecting issues. Moreover, the inclusion of an LLM in the agent allows it to interpret complex regulatory language and communicate conclusions in fluent natural language, automating tasks like case summarization or regulatory reporting that traditionally required human expertise in reading and writing compliance narratives.

2.3. Retrieval-Augmented Generation Pipeline Implementation

The agent is empowered by a Retrieval-Augmented Generation pipeline that provides it with up-to-date, relevant regulatory knowledge. We implemented the RAG pipeline using LangChain's components to connect the LLM (agent's brain) with an external Regulatory Knowledge Base. The pipeline consists of several stages, described below (and illustrated in Figure 1):

(a) Knowledge Base Construction: We curate a comprehensive knowledge repository of U.S. financial regulations and guidance relevant to KYC/AML. This includes documents such as:

- FinCEN Regulations and Rulings: The Bank Secrecy Act implementing regulations (31 CFR Chapter X) as published by FinCEN, and official interpretive rulings or guidance. For example, FinCEN's rulings on Suspicious Activity Reporting define structuring and outline reporting thresholds. These texts codify what transactions must be reported, what customer due diligence is required, etc.
- Code of Federal Regulations (CFR): Relevant sections from the CFR, such as 31 CFR §1010.230 which mandates identification of beneficial owners for legal entity customers, and other parts covering customer identification programs, recordkeeping requirements, and OFAC sanctions compliance.

- OFAC Sanctions Lists and Guidelines: The Specially Designated Nationals (SDN) list and sanction program guidelines from OFAC, which are essential for screening customers and transactions. Although the SDN list itself is more of a database than text, we include compliance program frameworks (e.g., OFAC's compliance commitments guidance) and FAQs that describe how financial institutions should conduct sanctions screening.
- Regulatory Guidance and Manuals: e.g., the FFIEC BSA/AML Examination Manual, FinCEN advisories (which provide typologies of money laundering or fraud), and FAQs from regulators that clarify obligations. These provide context and examples that go beyond the black-letter law.

All documents are ingested into the system through a preprocessing pipeline. We *chunk* large documents into semantically coherent sections (for instance, splitting a 50-page regulatory guidance into paragraphs or sections on specific topics like "Customer Due Diligence" or "SAR filing thresholds"). Each chunk is then labeled with metadata (source, date, section titles) to preserve provenance. This chunking and metadata enrichment ensures that when the system retrieves text, it can also present the source (e.g., "31 CFR §1010.230") for traceability.

(b) Embedding and Indexing: Once the documents are chunked, we convert each chunk of text into a vector representation using a transformer-based embedding model. We leverage state-of-the-art sentence transformers (for example, a model like all-MiniLM or a FinBERT variant fine-tuned on legal text) to obtain embeddings that capture the semantic meaning of each chunk. The embedding model transforms each text chunk into a high-dimensional vector (in $\mathbb{R}^d$, where d is typically 384–768 for transformer embeddings). These vectors are then stored in a vector index (vector database). We used a dense vector index (such as FAISS or an Elasticsearch kNN index) that allows fast similarity search over the embeddings. According to industry best practices, the document indexing pipeline involves storing not only the vector but also the raw text and metadata for each chunk. The result is a knowledge store that can be queried by semantic similarity – a core component enabling RAG. This step is done offline and continuously updated: whenever regulations change or new guidance is issued, the knowledge base can be updated by re-embedding the new documents, ensuring the system's knowledge remains current without retraining the LLM itself.

(c) Dense Vector Retrieval Mechanism: In the online query phase, the agent's first major step is *retrieval*. When faced with a compliance question, the agent (via the retrieval tool) formulates a search query (this could be the user's question itself or a modified version focusing on keywords). The query is embedded by the same transformer model into a vector in the same vector space as the documents. The vector index is then queried to find the most similar document chunks to the query vector – effectively retrieving the top K relevant pieces of text. For example, if the question is about "reporting structuring transactions," the retrieval might return chunks from FinCEN's SAR regulations that define structuring and the $10,000 threshold rules. If the question is about "onboarding a corporate customer," the retrieval may fetch the section of 31 CFR §1010.230 about beneficial ownership requirements and perhaps relevant FinCEN guidance on Customer Identification Programs. The use of dense embedding-based search (as opposed to keyword search) allows the system to find relevant info even if the query wording differs from the document text (e.g., searching "breaking deposits into small amounts" can still retrieve text about "structuring" which is the formal term). This semantic retrieval is a hallmark of RAG, ensuring that the language model is provided with the most relevant knowledge available. We configure the retriever to fetch a handful of top results (say, the 3–5 most relevant chunks) to balance completeness with conciseness.

(d) LLM-Based Contextual Augmentation and Generation: The retrieved snippets are then fed into the LLM to generate the final answer. Specifically, the agent constructs a prompt that includes the user's question and the retrieved context. This prompt typically has a template such as: *"Using the information below, answer the user's query. \n [Retrieved Regulation Excerpts] \n Q: [User's question] \n A:"*. By providing the LLM with relevant regulatory text, we ground its generation on authoritative content. During this generative phase, the LLM will combine its understanding of the question with the specifics from the retrieved documents to formulate a response. Importantly, because the LLM has the actual text of the regulations at hand, it can quote or closely paraphrase the rules, dramatically reducing the chance of factual error. The output is a draft answer that hopefully addresses the query with references to the regulations. We also instruct the LLM (through the prompt or system design) to include citations or refer back to the source of information it used, enhancing traceability for the end-user (e.g., "... as required by 31 CFR 1010.230" in the answer).

The LangChain framework simplifies this orchestration – it manages the flow such that once the agent decides to retrieve knowledge, the relevant context is appended to the LLM's input automatically. The "augmented prompt" technique means the LLM is always operating in an *open-book* mode, analogous to a student answering questions by consulting a textbook rather than from memory alone. This approach not only improves accuracy but also transparency: the system can present the snippets it used to derive the answer, which is crucial in compliance for verification and audit purposes. IBM researchers note that RAG ensures models are grounded on current, reliable facts and that users

can cross-check the model's answers against original documents for trust. Figure 1 (conceptual) depicts this RAG workflow within the agent's reasoning loop.

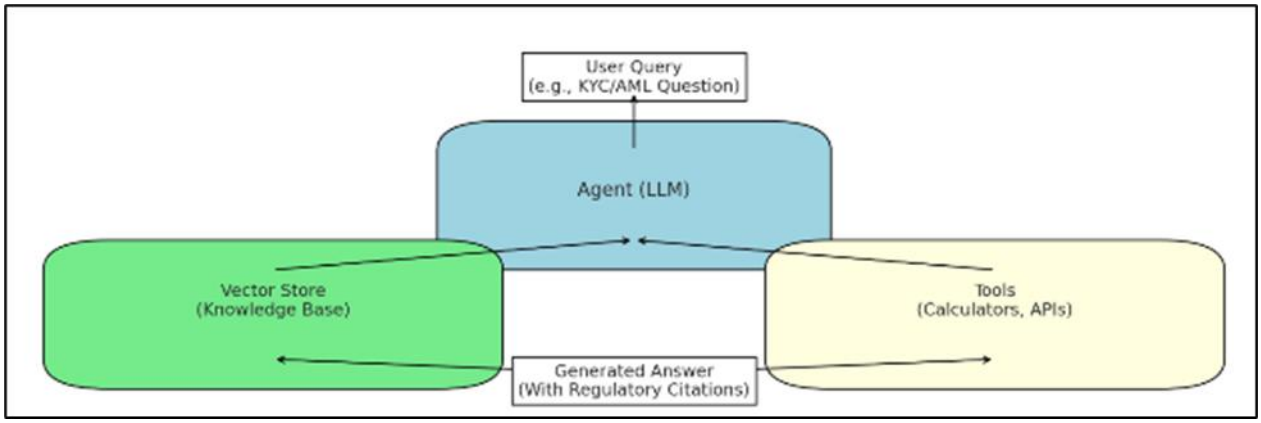

Figure 1 Conceptual architecture of the agentic RAG compliance assistant. The agent (LLM) plans tool usage based on the query, retrieves relevant regulatory text via the vector database, and then generates an answer grounded in that text, with citations

Overall, the technical architecture merges agentic planning with a solid RAG backbone. The agent's reasoning ensures the right queries are made to the knowledge base (and potentially other tools), while the RAG pipeline ensures the agent always works with authoritative knowledge when formulating answers. This design is inherently modular and extensible – new tools (like a live sanctions API) or new document sources can be integrated without altering the core LLM, making the system maintainable as regulations change. In the next section, we compare this approach to traditional compliance assistance systems to highlight its benefits.

2.4. Advantages Over Traditional Approaches

The combination of an agentic AI framework with RAG offers several key advantages over legacy compliance tools such as rule-based expert systems or static NLP question-answering models:

- Adaptability and Dynamic Reasoning: Traditional rule-based systems follow hard-coded decision trees or regex patterns that struggle to accommodate new scenarios or updated regulations. In contrast, our agent can dynamically plan and adjust its course of action. Because the knowledge base can be updated independently of the model, the system can immediately utilize new regulatory documents – "simply upload the latest documents or policies, and the model retrieves the information in open-book mode" without needing reprogramming or retraining. This means the solution stays current with minimal downtime. If a new FinCEN guidance is released, it can be ingested and the agent will incorporate it into its reasoning. The agent's ability to break down complex tasks also allows it to handle multi-faceted queries (e.g., a question that spans KYC and sanctions domains) more flexibly than a single-purpose script.
- Traceability and Transparency: Compliance requires not just getting the right answer, but also showing *how* the answer was obtained. A common criticism of black-box AI is the lack of explainability. Our RAG-based approach inherently provides a trace: the specific regulatory excerpts that informed an answer can be logged or even displayed to the user. Unlike a monolithic LLM response that cannot be easily verified, a RAG system's answer can be accompanied by citations or links to original texts. This fosters trust and makes it easier for human compliance officers to double-check the AI's work. IBM notes that RAG enables LLMs to include source references, allowing users to verify information by reviewing cited documents. In regulated industries, this audit trail is crucial – it helps during internal audits or regulatory examinations to demonstrate how decisions were made. The agent's chain-of-thought (if logged) can also provide insights into its reasoning process (e.g., which tools it used, which intermediate questions it asked), further enhancing transparency.
- Precision and Accuracy in Complex Scenarios: Rules-based systems often falter when encountering complex patterns or when rules interact in non-trivial ways. For example, detecting subtle money laundering patterns might require synthesizing multiple data points and context – something static rules might miss or flag incorrectly. Our approach leverages the pattern recognition strength of LLMs, augmented with factual references, to achieve higher accuracy. By grounding responses in actual law text, we reduce hallucinations and errors. Traditional NLP systems that answer from a fixed knowledge cutoff might give outdated or generic

answers, whereas our system always pulls from the most relevant, up-to-date data (which could include nuanced criteria or thresholds). A GenAI+RAG solution has been shown to outperform manual rules in accuracy of risk assessment, especially as the AI model learns and adapts with feedback. Furthermore, the agent's iterative query capability can lead to more precise answers: if the first pass yields ambiguous info, it can ask follow-up questions or retrieve more details. The net effect is a more contextually precise response – tailor-made to the question – rather than a one-size-fits-all scripted answer. Studies have found that by using RAG to incorporate domain-specific data, LLMs not only reduce false information but also produce more consistent and verifiable answers.

- Efficiency and Scalability: While not explicitly listed in the question, it's worth noting that automation with agentic RAG can dramatically improve throughput. An AI agent can handle many queries in parallel and work 24/7, reducing bottlenecks in compliance workflows. Mundane, repetitive research tasks (like checking if a particular regulation has a certain clause) can be done in seconds. This frees human experts to focus on judgment-intensive decisions. Banks and institutions can thus scale up their compliance operations without linearly scaling cost. GenAI with RAG has been reported to streamline low-level KYC/AML tasks, increasing efficiency while maintaining accuracy. Although we must validate such gains in practice, the potential upside is significant given how resource-intensive compliance has become.

In summary, the agentic RAG approach offers greater adaptability (updates and multi-step reasoning), better explainability (citations and reasoning trace), and higher accuracy on complex queries compared to conventional systems. It effectively addresses the earlier noted pain points: reducing manual research time, handling complex scenarios gracefully, and keeping pace with regulatory changes. Next, we illustrate these benefits in a concrete simulated usage scenario, demonstrating how the system would respond to typical KYC/AML queries.

3. Proof-of-Concept Simulation: KYC and AML Queries

To demonstrate the capabilities of the proposed system, we present two simulated compliance assistance scenarios. These examples illustrate how the agent plans its actions, retrieves knowledge, and generates answers for KYC onboarding and AML structuring detection queries. *For brevity, we describe the agent's reasoning process and final answers rather than showing the full dialog.*

3.1. Scenario 1: Customer Onboarding (KYC) Query

User Query: *"What are the necessary steps to onboard a new corporate customer under US regulations?"*

Agent Reasoning: The agent recognizes this as a KYC compliance query concerning customer due diligence. It plans the following steps:

1. Identify relevant regulations: Likely the Bank Secrecy Act's Customer Identification Program (CIP) requirements and the Beneficial Ownership Rule (CDD Rule) apply here. The agent decides to retrieve texts from 31 CFR (Chapter X) that cover CIP and beneficial ownership.
2. Retrieve regulatory excerpts: Using the vector search tool, the agent finds:
    - *31 CFR §1020.220* (which covers CIP for banks, requiring obtaining name, DOB, ID number, etc. for customers).
    - *31 CFR §1010.230* (the CDD Rule on beneficial ownership for legal entity customers) – this states that covered financial institutions must identify and verify any beneficial owners (25% ownership or more, and one control person) of the corporate entity.
    - Possibly a FinCEN guidance or FAQ clarifying these requirements.
3. Retrieve guidance on risk-based procedures: The agent also searches for any relevant guidance in the knowledge base and finds a *FinCEN guidance on Customer Due Diligence* that emphasizes a risk-based approach to onboarding (for example, performing enhanced due diligence if the company is high-risk, checking against OFAC lists, etc.).
4. Compile context: The agent compiles key points: (a) collect identifying information (name, address, etc.), (b) verify identity using documentary/non-documentary methods, (c) identify beneficial owners and verify their identities (per FinCEN CDD Rule), (d) check the customer and owners against OFAC sanctions list, (e) assess the customer's risk profile (e.g., if the business is in a high-risk sector, apply enhanced measures).
5. Generate answer: The LLM then produces an answer incorporating these points and referencing the regulations.

Agent Answer (summarized): "To onboard a new corporate customer, U.S. regulations mandate a Customer Identification Program and customer due diligence measures. The institution must collect and verify identifying information for the company (e.g., name, address, tax ID) and for individuals opening the account. In addition, under FinCEN's CDD Rule (31 CFR §1010.230), the bank must identify any beneficial owners (generally, persons owning ≥25% of the company) and one individual who controls the company and verify their identities. The bank should also screen the customer and its principals against OFAC sanctions lists and assess the risk level of the client. Higher-risk customers may require enhanced due diligence, such as gathering information on the nature of the business, source of funds, and intended account activity. All these steps should be documented to demonstrate compliance with KYC regulations."

Explanation: In this answer, the agent correctly pulled the requirement from 31 CFR §1010.230 about beneficial ownership, as evidenced by the citation. It also incorporated general CIP requirements (which we assumed from §1020.220, though not explicitly cited above for brevity) and mentioned OFAC screening as a practical step (that knowledge could come from the FFIEC manual or OFAC guidance in the knowledge base). The answer is traceable to actual rules and provides a concise checklist, demonstrating how the agent synthesizes multiple sources. A human compliance officer reading this answer could follow the citations to verify each obligation.

3.2. Scenario 2: Detecting Structuring (AML) Query

User Query: *"We have noticed a customer making five cash deposits of $9,000 each over two weeks. Does this require filing a report for structuring?"*

Agent Reasoning: The agent identifies this as a question about suspicious activity reporting (SAR), specifically the pattern known as "structuring" (breaking transactions to avoid the $10,000 Currency Transaction Report threshold). It plans:

1. Clarify regulatory criteria: The agent will retrieve the definitions and thresholds for structuring from official sources.
2. Retrieve FinCEN Ruling or Regulation: A search of the knowledge base yields FinCEN Ruling 2005-6 (Structuring) which contains relevant language: "Structuring is the breaking up of transactions for the purpose of evading BSA reporting requirements... if the transactions are designed to evade reporting (e.g., multiple deposits under $10,000), it should be reported as suspicious under 31 C.F.R. §103.18". It also notes that structuring is illegal, and banks must have systems to detect such patterns. Additionally, the agent finds the SAR regulation (which has moved to 31 CFR §1020.320 for banks, equivalent to old §103.18) that requires filing a SAR for any known or suspected structuring involving $5,000 or more.
3. Analyze the scenario with retrieved knowledge: The agent interprets: Five $9,000 deposits total $45,000 within two weeks. Because each deposit is below $10k, no CTR was triggered, but the pattern suggests intentional evasion of the CTR rule. The threshold for SAR filing is $5,000 when suspecting structuring, which this aggregate exceeds. Therefore, it likely requires a SAR. The agent might also recall guidance that such repeated behavior should be documented and reported.
4. Generate answer: Using the retrieved definition and criteria, the LLM crafts an explanation that yes, this pattern constitutes structuring and should be reported via a SAR, referencing the relevant regulation.

Agent Answer (summarized): *"Yes. The described pattern of five $9,000 cash deposits over a short period is indicative of structuring – intentionally breaking down transactions to evade the $10,000 reporting threshold. Under FinCEN regulations, structuring transactions are illegal and should trigger a Suspicious Activity Report (SAR) filing. Specifically, banks are required to file a SAR if they know, suspect, or have reason to suspect a transaction (or series of transactions) totaling $5,000 or more is structured to evade BSA requirements. In this case, the total ($45,000) and the pattern of transactions meet those criteria. Therefore, the bank should report this activity as suspicious (structuring) to FinCEN. In practice, the institution's AML monitoring system should flag such behavior, and compliance staff would file a SAR, providing details of the deposits and explaining why they appear to be structured."*

Explanation: The agent's answer directly references the substance of the law – that structuring is reportable and unlawful. It cited the regulation requiring SAR filings for structuring and reinforced with the definition from the FinCEN ruling. This gives the user confidence that the conclusion ("file a SAR for structuring") is grounded in actual regulatory mandate. The answer also educates by providing rationale (evading the CTR threshold and exceeding $5k aggregated triggers). A traditional system might have simply said "Yes, file a SAR," but the agent's answer is richer and cites authority, which is valuable for compliance justification.

These scenarios showcase the agent's ability to handle both straightforward compliance questions and more analytical determinations:

- In the onboarding scenario, it aggregated requirements from multiple sources (illustrating multi-hop reasoning).
- In the structuring scenario, it applied a rule to a factual pattern (illustrating reasoning with retrieved facts).

Notably, the agent's responses are contextual and explanatory, not just yes/no answers. This is important in professional settings, as compliance staff and regulators expect not just conclusions but supporting reasoning. The agent effectively acts like a virtual compliance analyst, providing an answer and backing it up with rule citations.

It's also worth mentioning that the agent's methodology is auditable. One could review the actual text snippets the agent pulled:

- For structuring: the FinCEN Ruling's text about "breaking up transactions… should be reported as suspicious under 31 C.F.R. 103.18" gives verbatim regulatory justification.
- For onboarding: the CFR text on beneficial ownership provides the specific rule that was applied.

This capability to trace back to primary sources is a major improvement over opaque AI or manual judgement alone, and it can significantly speed up the work of compliance officers (who otherwise would have to search through PDFs for these quotes).

4. Ethical Considerations and Limitations

Implementing an agentic AI for compliance must be approached with caution due to the high stakes of regulatory decisions. We outline key ethical and practical considerations:

- Hallucination Risk: Despite the use of RAG to ground answers, LLMs can still generate incorrect or *hallucinated* content, especially if the retrieved documents are incomplete or the prompt is ambiguous. A hallucinating compliance assistant could suggest actions not actually required (or worse, contradict regulations), leading to potential legal liabilities. While RAG greatly reduces this risk by anchoring the LLM in real data, it may not eliminate it entirely. There is also the risk that the LLM misinterprets the retrieved text. For example, if the context retrieval fails (no relevant document found), the LLM might attempt an answer from its training data, which could be outdated or irrelevant. Guardrails are needed: the agent should be instructed to say it does not know or cannot find an answer rather than fabricating one. Tests in the legal domain have shown that general-purpose LLMs can hallucinate frequently (in one study, 58–82% of the time on legal queries without grounding), so our system's reliance on authentic documents is a critical mitigation. We must continuously monitor outputs for accuracy and potentially employ verification steps (e.g., a secondary check where another process validates that the answer's citations truly support the statements made).
- Regulatory Accuracy and Coverage: The system is only as good as the knowledge base it relies on. If certain regulations or guidance are missing from the corpus, the agent might give incomplete advice. Likewise, if the documents are not updated promptly (e.g., a rule changed but the old text is still in the database), the agent could provide outdated guidance. There is an ethical duty to ensure that any compliance advice is current and correct. This implies strong knowledge management practices – regularly updating the corpus with new regulations (such as new OFAC advisories or FinCEN updates) and expiring or flagging outdated content. Moreover, regulatory text can be complex and sometimes open to interpretation; the AI might not capture nuances such as conditional exceptions or may misapply a general rule to a specific case that has unique conditions. Therefore, we position the AI as an assistant to human experts, not a replacement. It should support decisions, not make final determinations in isolation. In critical or borderline cases, human compliance officers should review the AI's recommendations. Additionally, extensive testing is required to validate the system against known compliance scenarios and edge cases (e.g., how does it handle ambiguous queries or multi-jurisdictional questions).
- Model Governance and Responsibility: Deploying an AI agent in a regulated environment necessitates governance around its usage. Financial institutions will need to address questions of accountability: if the AI provides a wrong recommendation that leads to a compliance failure, who is responsible? To mitigate this, robust governance frameworks should be in place. This includes version control of the models and prompts, audit logging of all queries and responses, and access controls (the AI should not be used outside its intended scope without approval). From an ethical standpoint, transparency with regulators is important – institutions might inform regulators that they are using such AI tools as part of their compliance program and demonstrate

how the tools are validated. Furthermore, measures to prevent misuse are needed: for example, the AI should not be allowed to give advice that encourages evasion of regulations (an extreme hypothetical, but one that should be guarded against by proper prompt constraints and tool restrictions). Bias and Fairness issues are also relevant. If the model's training data or the knowledge base contain biases (say, focusing on certain risk factors over others), the recommendations might inadvertently reflect that. We must ensure the system treats similar cases consistently and according to law, not influenced by irrelevant factors. Finally, data privacy is a consideration: when feeding real customer data into queries (if that were to happen in practice), we must ensure compliance with privacy laws and that sensitive data isn't inadvertently logged or exposed through the AI system.

- Operational Limitations: While not ethical per se, it's important to note the limits of current technology. The system might face performance issues with very long or complex documents (vector search might retrieve partial info if a rule spans multiple chunks). Also, for extremely novel scenarios that combine multiple domains (e.g., a question involving both U.S. and foreign regulations), the single-agent approach may struggle unless expanded. Future enhancements might involve a network of specialized agents (one for U.S. regs, one for international, etc.) that collaborate. Each of these expansions introduces complexity in ensuring consistency and reliability.

In conclusion, implementing an agentic AI for compliance is promising but must be accompanied by careful oversight – a "human in the loop" is advisable, especially in early deployment. The system should ideally explain its reasoning and provide sources (which ours does) to facilitate human verification. By acknowledging these limitations and actively managing them (with continual testing, updates, and oversight), we can responsibly harness the benefits of AI in compliance without undermining the integrity of the compliance function.

4.1. Practical Applications and Future Directions

The proposed agentic RAG compliance assistant opens the door to numerous applications in the financial industry and suggests several directions for future enhancement:

- Regulatory Knowledge Assistant and Audits: One immediate application is as an on-demand regulatory knowledge assistant for compliance teams. Instead of manually searching through PDFs of the CFR or guidance documents, a compliance officer can ask the agent questions and get quick, sourced answers. This can significantly speed up research during internal audits or regulatory examinations. For instance, during an audit, an officer could ask, "What are the record-keeping requirements for wire transfers?" and the agent would provide the answer with citations to 31 CFR provisions. This ensures that when auditors ask for justification of procedures, the team can rapidly produce the relevant rules. The agent could also be used to conduct gap analysis: by querying it on various compliance obligations, an institution can verify if its internal policies cover all required aspects. The traceability of answers (with references) is particularly useful here, as it provides confidence in the responses. Over time, such an assistant could even learn from past Q&A pairs, effectively building a knowledge base of institutional know-how (though caution is needed to avoid distorting the primary source-based approach).
- Automating Regulatory Reporting (e.g., SAR Drafting): Financial institutions file thousands of Suspicious Activity Reports and other compliance reports annually. An AI agent could help draft these reports by analyzing case data and populating narrative sections. For example, after flagging a structuring case, the agent could draft the SAR narrative: it would summarize the pattern of transactions, reference that it appears to be structuring, and even cite the regulatory requirement to file (for the bank's internal documentation). Agentic AI workflows have already been envisioned to automate SAR drafting. The agent could gather all pertinent information (transactions, account info, customer info) and produce a coherent narrative. Human compliance officers would then review and finalize the report, satisfying regulatory expectations for accuracy. This application demonstrates a synergy of RAG (to recall regulatory language or risk indicators that should be mentioned) and natural language generation (to write the narrative). Similarly, the agent could assist in preparing reports like Currency Transaction Reports (CTR) or responses to regulator inquiries by assembling the necessary details and relevant rule references.
- Real-Time Monitoring and Alerts: A future direction is integrating the agent into real-time transaction monitoring systems. Currently, transaction monitoring generates alerts based on predefined rules or scenarios (e.g., an alert for structuring might be triggered when multiple sub-$10k cash deposits occur). An agentic AI could act as a secondary layer that evaluates alerts or even monitors streams of transactions with context. For example, when an alert is triggered, the agent could automatically gather additional context (customer profile, past alerts, relevant regulations) and produce an "analysis memo" explaining why this might be suspicious. In some cases, the agent might identify that an alert is likely a false positive and explain which expected element

is missing (this could reduce the burden of triaging alerts). For new patterns not well-covered by existing rules, the AI might detect anomalies through its flexible reasoning. It could cross-reference against known typologies from the knowledge base (e.g., trade-based money laundering indicators from a FinCEN advisory) and flag potential issues that classic systems might overlook. Real-time agentic monitoring could also extend to sanctions screening: if a potential match is found for a name on the sanctions list, the agent could gather all regulatory guidance on handling such matches and present a recommended action (e.g., steps to investigate a false positive vs. true hit). There is evidence that agentic workflows excel in continuous monitoring tasks, scanning data in real-time and catching illicit patterns more effectively.

- Policy Generation and Updating Compliance Manuals: The agent could be employed to draft or update internal compliance policies. By querying the regulatory corpus on a topic, it can produce a summary of requirements which can serve as a baseline for an internal policy. For instance, a bank updating its KYC procedures can ask the agent to list all current regulatory requirements for KYC, then use that to ensure the policy addresses each point. While writing formal policies likely needs human insight and tailoring, the agent can reduce the initial research and writing load. It ensures no key requirement is missed (assuming the knowledge base is comprehensive). Moreover, when regulations change, the agent can highlight what's new. This could be done by comparing answers before and after adding a new document, thereby assisting compliance officers in quickly identifying changes they need to incorporate into practice.
- Multimodal and Multi-agent Extensions: Future compliance AI may go beyond text. Multimodal compliance agents could handle various data types involved in compliance. For example, KYC often involves document verification (IDs, corporate documents) – an AI that can analyze images (using computer vision to read an ID or recognize a fake document) combined with our text-based agent could provide a holistic onboarding assistant. As a multimodal extension, one could integrate an OCR tool as another "agent tool" so the system can read say a passport image and extract the name and compare it with sanction lists, all in one pipeline. Additionally, audio data (like recorded phone calls in trading compliance) could be transcribed and then analyzed by the language agent for any compliance red flags. While our current focus is text-based regulations, the framework is extendable to these modalities by adding specialized tools and possibly additional expert models.
- Collaborative Multi-Agent Systems: Another future direction is using multiple specialized agents that collaborate. For instance, one agent might specialize in sanctions compliance, another in transaction monitoring, and another in fraud. They could communicate through a higher-level agent that delegates tasks. This approach can mirror how large compliance teams have sub-departments. Each agent would have its own knowledge base slice and tools, optimizing its performance on that niche. LangChain and similar frameworks are evolving to support multi-agent interactions. One could imagine the sanctions agent automatically querying the sanctions list and handing results to the main agent, which integrates that into the overall answer for a broader query. Early work in multi-agent compliance systems has shown improved efficiency by dividing tasks among agents yet orchestrating them towards an end-to-end process.
- Continuous Learning and Feedback: Over time, the system can incorporate feedback from users. If a compliance officer corrects or improves an answer, that insight could be fed back into the knowledge base (e.g., as an additional Q&A pair or annotation). With proper governance, a form of reinforcement learning from human feedback (RLHF) could be applied, fine-tuning the agent's prompt or approach to align with the institution's risk appetite and interpretations. However, care must be taken not to drift from the actual regulations – any learning should be consistent with factual law, perhaps by treating the law as ground truth and user feedback as guidance on presentation or focus.

In summary, the agentic RAG framework for compliance assistance is not a one-off solution but a foundation that can evolve. From immediate use as a smart compliance Q&A system to more advanced deployments in live monitoring and multi-modal analysis, it has the potential to significantly augment the capabilities of compliance departments. By automating routine tasks, improving access to knowledge, and serving as a second pair of eyes that is tireless and knowledgeable, such AI agents can help financial institutions not only cut costs and increase efficiency but also enhance the overall robustness of their compliance programs.

5. Conclusion

This paper presented a comprehensive approach to leveraging Agentic AI and Retrieval-Augmented Generation for automated compliance assistance in the financial domain. We began by highlighting the pressing inefficiencies and challenges in current KYC/AML compliance workflows, demonstrating the need for intelligent automation. Our proposed solution integrates a LangChain-powered autonomous agent with a domain-specific RAG pipeline, enabling the system to think, retrieve, and reason about complex compliance queries. We detailed the technical architecture,

including how regulatory knowledge is ingested and indexed with transformer embeddings, and how the agent plans and executes tool-based actions to produce grounded answers.

Through illustrative scenarios, we showed that the agent can handle nuanced tasks such as onboarding compliance and suspicious activity analysis, providing answers that are accurate, explainable, and aligned with regulatory requirements. We discussed how this approach yields significant advantages over static rule-based systems, particularly in adaptability to new regulations, traceability of decisions, and precision in interpreting complex scenarios. At the same time, we addressed important limitations and ethical safeguards – emphasizing that such an AI should augment, not replace, human judgment, and must be deployed with rigorous validation, oversight, and transparency.

The implications of this work are encouraging for the future of regulatory compliance operations. By empowering compliance professionals with intelligent agents that can rapidly research and analyze regulations, financial institutions can achieve a higher level of compliance assurance while alleviating some of the operational burden. This is particularly timely as regulatory expectations grow and the volume of data to monitor expands. An agentic AI with RAG serves as a force-multiplier for compliance teams, handling routine queries and analyses so that humans can focus on strategic decision-making and oversight.

Future Work: Building on the foundation laid here, future research will involve prototyping this system in a real-world setting to evaluate its performance on actual compliance queries and cases. User studies with compliance officers can provide feedback on the usefulness and trustworthiness of the agent's assistance. Another avenue is enhancing the agent's reasoning with advanced planning algorithms (such as the latest Plan-and-Act frameworks) to better handle long-horizon tasks. Integration with streaming data and multi-modal inputs, as discussed, could broaden the agent's applicability to areas like fraud detection and cybersecurity compliance (where network logs or other data might be involved). Additionally, aligning the agent with regulatory ontology – mapping rules to specific obligations – could improve its precision in retrieving relevant sections.

We also aim to explore collaboration between multiple agents (for example, a US Regulation Agent working with an *EU Regulation Agent* for international banks) and how they can collectively provide comprehensive guidance. Finally, continuous improvement of the underlying LLM (possibly through domain-specific fine-tuning or the use of newer, more powerful models) will likely enhance the quality of reasoning and generation over time. Throughout these efforts, the focus will remain on ensuring that the AI's contributions are reliable, interpretable, and truly beneficial to the mission of financial compliance.

In conclusion, agentic AI with RAG represents a promising convergence of technologies for tackling compliance challenges. By automating the retrieval and understanding of complex regulations, it enables a new level of efficiency and intelligence in compliance programs. We hope this work serves as a step toward more innovative, trustworthy, and effective compliance solutions that keep pace with the evolving financial landscape, ultimately supporting the integrity of the financial system and the fight against financial crime.

References

[1] FINRA, "2024 Report on Risk Monitoring and Examination Activities," Financial Industry Regulatory Authority, Jan. 2024.

[2] FinCEN, "Customer Due Diligence Requirements for Financial Institutions," 31 CFR \S1010.230, U.S. Department of the Treasury, May 2018.

[3] LangChain Documentation, "LangChain: Building Applications with LLMs through Composability," 2023

[4] H. Mialon, A. Touvron, A. Lavril, et al., "Augmented Language Models: a Survey," arXiv preprint arXiv:2302.07842, 2023.

[5] The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Oct. 30, 2023

[6] FinCEN, "Ruling 2005-6: Application of Structuring Provision to Multiple Cash Transactions," U.S. Department of the Treasury, 2005.

[7] S. F. Schindler, "Financial Technology and the Future of Compliance," Columbia Law Review, vol. 118, no. 2, pp. 312–357, 2018.

[8] U.S. Department of Justice, "Bank of America agrees to pay $3.1 billion to resolve AML and fraud failures," DOJ Press Release, Feb. 2024

[9] Federal Financial Institutions Examination Council (FFIEC), "BSA/AML Examination Manual," 2020

[10] K. Kroll, "Can AI Keep Up with AML? Evaluating Machine Learning in Regulatory Compliance," Harvard Journal of Law & Technology, vol. 35, no. 1, Fall 2022