Leveraging Multi-Agent System for Automated IoT Security Audit Planning

Obrina Briliyant*, Amir Javed*, Yulia Cherdantseva*, Nanang Trianto**
*School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom
**Department of Cyber Security, Politeknik Siber dan Sandi Negara, Bogor, Indonesia
briliyantoc@cardiff.ac.uk, nanang@poltekssn.ac.id

***Abstract*—Current security compliance audit planning faces significant challenges in terms of manual risk-to-policy mapping, time-intensive document review procedures, and expanded compliance regulation. Manual audit planning typically requires half of total audit time and is prone to human cognitive biases and oversight errors. These challenges result in delayed audit cycles, increased audit costs, and potential security vulnerabilities that remain undetected. This paper presents a novel automated audit planning framework leveraging a multi-agent system architecture powered by Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). Our methodology employs a three-agents architecture: risk prioritization, control mapping with RAG-enhanced context retrieval, and audit criteria generation agents, validated against the European Telecommunications Standards Institute (ETSI) EN 303 645 standard for Internet of Things (IoT) cybersecurity. Experimental evaluation using the RAG Assessment Suite (RAGAS) framework demonstrates exceptional performance with Context Precision achieving 0.917, Context Recall at 0.900, and Answer Correctness reaching 0.873, indicating high-quality automated audit criteria generation. The system successfully automated the planning phase of a risk-based compliance audit, achieving risk-to-ground truth matching with consistent, reliable outputs. These Results suggest that multi-agent system architecture can significantly enhance audit efficiency while maintaining professional audit standards, providing a foundation for scalable continuous auditing in IoT environments.**

***Index Terms*—Security Compliance, IT Audit Planning, IoT Security, multi-agent system, Large Language Models, Retrieval-Augmented Generation.**

I. Introduction

Audit planning has become an expensive bottleneck that is draining resources and compromising security. The Institute of Internal Auditors (IIA) reveals a striking reality: IT audit planning alone devours 40-60% of total audit resources [1], creating an unsustainable burden on audit teams who must invest the majority of their time before any actual auditing begins. This resource-intensive approach creates a cascade of problems—auditors struggle with biased risk-to-policy mapping, while emerging cyber threats exploit the delays. The result is an auditing framework that fails to deliver on its promise of real-time risk management, leaving organizations vulnerable precisely when they need protection most.

These challenges significantly impact both auditors and organizations. Auditors face increased workload pressure, potential liability from oversight errors, and difficulty maintaining up-to-date expertise across rapidly evolving technologies. Organizations experience bottlenecked audit cycles that delay compliance certification, increase costs, and potential security vulnerabilities that remain undetected during prolonged planning phases. The continuous auditing paradigm, essential for real-time risk management in IT environments, becomes impractical under traditional manual approaches due to the sheer volume and velocity of security events requiring assessment.

This is particularly true and significant for auditing IoT security compliance, where complex and rapid dynamic development of technologies keeps happening. This heterogeneous system produces big data to audit, an expanded attack vector to cover, and more audit criteria to be included. Not to mention the need for upskilling the auditor domain expertise to be relevant in IoT technologies. All of this makes a more challenging condition for auditors and the organization.

Recent advances in artificial intelligence (AI), especially Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), present promising solutions for automating complex cognitive tasks traditionally requiring human expertise. Multi-agent method and architecture could leverage these technologies to analyze vast amounts of documents, maintain consistency across assessments, and generate professional-quality audit artifacts at scale. However, limited research exists on applying this method specifically to audit planning processes, especially for IoT security contexts where domain expertise and regulatory compliance are paramount.

Our study addresses this research gap by presenting a novel automated audit planning framework specifically designed for IoT security compliance. Our approach employs a multi-agent system architecture where specialized AI agents collaborate to automate risk prioritization, control mapping, and audit criteria generation while maintaining adherence to established professional standards. The system leverages RAG technology to ensure that generated audit criteria reflect contextual relevancy to organization risk appetite and regulatory requirements, specifically validated against the ETSI EN 303 645 standard for IoT cybersecurity.

The primary contributions of this research include:

- Novel multi-agent system for audit planning: Design and implementation of a multi-agent system specifically

This research was made possible by the support of the Indonesia Endowment Fund for Education Agency (LPDP). Their investment in our work has significantly contributed to the quality and impact of our research findings.

optimized for IoT security audit planning, demonstrating how specialized AI agents can collaborate to automate complex audit planning tasks while maintaining professional standards.

- RAG-based risk-to-policy mapping: Development of an innovative approach combining vector-based document retrieval with LLM reasoning to automatically map security risks to appropriate controls and generate contextually relevant audit criteria.
- Performance evaluation framework: Introduction of RAGAS [2] as an evaluation methodology specifically adapted for the RAG performance assessment, providing quantitative metrics for faithfulness, relevance, and correctness of generated audit criteria.

The remainder of this paper is organized as follows: Section II reviews related work in security compliance audit automation, which encompasses IoT security and AI applications. Section III details the experimental methodology, including dataset preparation, multi-agent architecture, evaluation metrics, and the ground truth creation. Section IV presents the experimental results with a detailed analysis of system performance across multiple evaluation dimensions. Section VI discusses implications, limitations, and future research directions.

II. Related Works

Recent studies on IoT security compliance audit automation have implemented a variety of technical solutions that map security standards to automated audit processes. Many papers reported high automation levels using methods such as deep learning [3], large language models (LLMs) [4], and retrieval-augmented generation (RAG) [5], [6]. Several papers integrate IoT security standards (including GDPR, IEC 62443, NIST, ENISA, and others) with automated threat detection, policy checking, and test generation; one study cites 99% detection accuracy with a deep learning pipeline, while another achieves an AUC of at least 0.8 in vulnerability prioritization [7].

Papers addressing business context considerations usually use ontology mapping [8], Business Process Model and Notation (BPMN) process modeling [9], and risk management frameworks (for example, the NIST AI Risk Management Framework) to align technical compliance with organizational needs. Multiple studies report that combining LLMs with RAG or agent-based methods leads to improved precision, reasoning, and state-of-the-art audit automation [6]. These works collectively demonstrate that robust architectures integrating advanced AI methods and IoT security standards can yield automated, scalable, and context-sensitive compliance audit systems [3], [10].

Some studies focus on mapping multiple standards into unified ontologies or process models, while others target specific standards for technical compliance, with varied results [5], [8]. Explicit standards mapping and automated verification are recurring themes [6], with knowledge graphs and semantic web technologies supporting multi-standard integration [9] being the most common proposed solution.

Several studies includes approaches for adapting compliance frameworks to organizational or business context, including:

- Ontology mapping of business rules (aligning business rules with formal knowledge representations) [8],
- Process modeling using BPMN [11], and
- Provenance tracking for auditability (recording the origin and history of data) [10].

However, many studies focus primarily on technical compliance [7], with less attention to organizational processes or auditor-centric activities. While technical automation is advancing rapidly, integration of business context, organizational needs, and auditor's workflow is less developed.

To the best of our knowledge, there is still no study addressing a comprehensive planning process of a compliance check or audit, as depicted in our summary of previous studies in Table I. Also, while many of the previous studies utilize RAG as part of the solution, little to no evaluation has been conducted on the RAG performance of the proposed solution. Our work fills in this research gap by focusing attention on the audit planning phase of a security audit and using the RAG assessment suite (RAGAS) as an evaluation framework. The reason behind this is because the RAG pipeline is being used as the main flow of context-reasoning activities by the LLM, so the logic is that by tuning the RAG pipeline and the data processing that it feeds off, we can enhance the accuracy and the performance of the multi-agent system to be better (minimize hallucination and out-of-context response).

III. Methodology

This section presents our methodology for automated IoT security audit planning, encompassing dataset preparation, multi-agent system architecture, RAG-based Control Mapping algorithms, and evaluation framework implementation. Figure 1 illustrate our methodology combining IT security audit methodology derived from [12] and multi-agent system architecture for automation.

A. Dataset Preparation

Our experimental dataset comprises three primary components designed to simulate real-world audit planning scenarios:

1) Security Standards Knowledge Base: We constructed a comprehensive repository containing the **mandatory controls** of ETSI EN 303 645 standard for IoT cybersecurity [13], encompassing 33 discrete provisions (same as controls). Each control (or provision) was preprocessed using sentence-level tokenization and embedded using text-embedding-ada-002 model, creating a 1536-dimensional vector space for semantic retrieval.
2) IoT Risk Scenarios: The risk register was developed by the risk management team of organization ABC in the year 2022. Although not all of the register is used, but it still represents a realistic dataset for our experiment. Ten representative high-risk IoT security scenarios in smart building context were used, e.g. default password vulnerabilities in smart cameras, unencrypted data transmission in environmental sensors, weak authentication

TABLE I
AI AND LLM INTEGRATION PATTERNS IN IOT SECURITY COMPLIANCE STUDIES

Study	AI Application Type	Use Case/Audit Process	Integration Method	Reported Benefits
[3]	Deep learning (CNN/LSTM)	Threat detection, compliance verification	Verification prediction	99% accuracy, 91% compliance
[4]	LLM + knowledge graph	Part of planning phase (Security requirement)	Question and Answer (Q&A)	Effective automation, user-friendly
[7]	LLM + Retrieval-Augmented Generation (RAG)	Compliance verification	Multi-stage retrieval	Improved correctness, reasoning
[5]	Knowledge graph + RAG + LLM	Querying NISTIR 8259A	Graph-driven RAG	Improved precision, relevance
[6]	LLM (Llama 3.1 70B) + RAG	Test automation	Advance-RAG	Rapid report generation
[10]	LLM + RAG	Vulnerability prioritization	Semantic entropy	Priorities prediction $\geq$ 0.8
Our work	**Multi-Agent LLM + RAG**	**Planning phase**	**multi-agent system**	**Higher retrieval**

mechanisms in smart door locks, inadequate network segmentation for IoT devices, and insufficient consent management in voice-activated assistants.

3) Ground Truth: the ground truth is the anchor data that we use to measure the response of the multi-agent system. The closer the response to the ground truth, the better the system performs. In our context, that means that the ground truth data will be the audit criteria in JSON format. The audit criteria are usually the output of the audit planning phase, because it define the audit scope and audit objective, act as a guidance for the security auditor to conduct the audit. Our ground truth (audit criteria) were manually developed for each risk scenario, mapped to the relevant ETSI EN 303 645 provisions (or controls). In our experiment, these criteria are formatted into more technical semantic JSON by the agent so that they can be read by a machine easily. This output format can be change depending on the need of the next step of the audit step (after planning phase).

B. Multi-Agent System Architecture

Our framework employs a three-agent collaborative architecture, where each agent specializes in distinct aspects of audit planning while maintaining seamless information flow and decision coordination. As depicted in Figure 1, the framework integrates traditional IT audit planning processes with our innovative multi-agent AI system, demonstrating how AI enhances rather than replaces professional IT audit methodologies.

Agent 1 (Risk Ingestion & Prioritization): Implements an advanced risk assessment workflow combining likelihood and impact matrices with IoT-specific vulnerability. The agent processes risk scenarios in the risk register using structured prompting techniques and generates prioritized risk rankings.

Agent 2 (RAG-Based Control Mapping): Serves as the core innovation of our approach, combining vector-based document retrieval (RAG) with LLM reasoning to automatically map identified risks to appropriate security controls. This agent maintains the ETSI standards knowledge base and performs contextual control selection.

Agent 3 (Criteria Generation): Synthesizes risk assessments and control mappings into comprehensive, actionable audit criteria. The agent ensures that generated criteria incorporate the IoT security standard while maintaining professional audit standards.

C. RAG-Based Control Mapping

The highlight of our methodology is in the innovative integration of RAG with LLM reasoning for automated control mapping. Our approach follows a systematic four-phase workflow that transforms risk scenarios into comprehensive audit criteria.

Phase 1: Vector-based Document Retrieval

The system begins by converting the input risk scenario into a high-dimensional embedding. This embedding is then compared against pre-computed embeddings of ETSI EN 303 645 mandatory provisions using cosine similarity. Documents exceeding the similarity threshold (τ = 0.7) are retrieved and ranked by relevance score. This phase ensures that only the most pertinent security controls are considered for the specific risk scenario, creating a focused knowledge base for subsequent processing.

Phase 2: LLM-based Contextual Reasoning

Retrieved controls are then combined with the original risk scenario to construct a structured prompt for the LLM. A reasoning LLM processes this context to understand the relationship between the identified risk and available controls, selecting the most appropriate provisions (again, this means controls) while considering IoT-specific context. This phase leverages the model's reasoning capabilities to bridge the gap between generic IoT security controls and specific risk contexts of organization ABC.

Phase 3: Iterative Refinement

The system performs consistency validation on the mapped controls, checking for completeness, redundancy, and alignment with the original risk scenario. If the consistency score falls below the threshold, the system generates feedback prompts to refine the control selection. This iterative process continues until satisfactory consistency is achieved or a maximum of three iterations is reached, ensuring robust and reliable control mapping.

Phase 4: Audit Criteria Generation

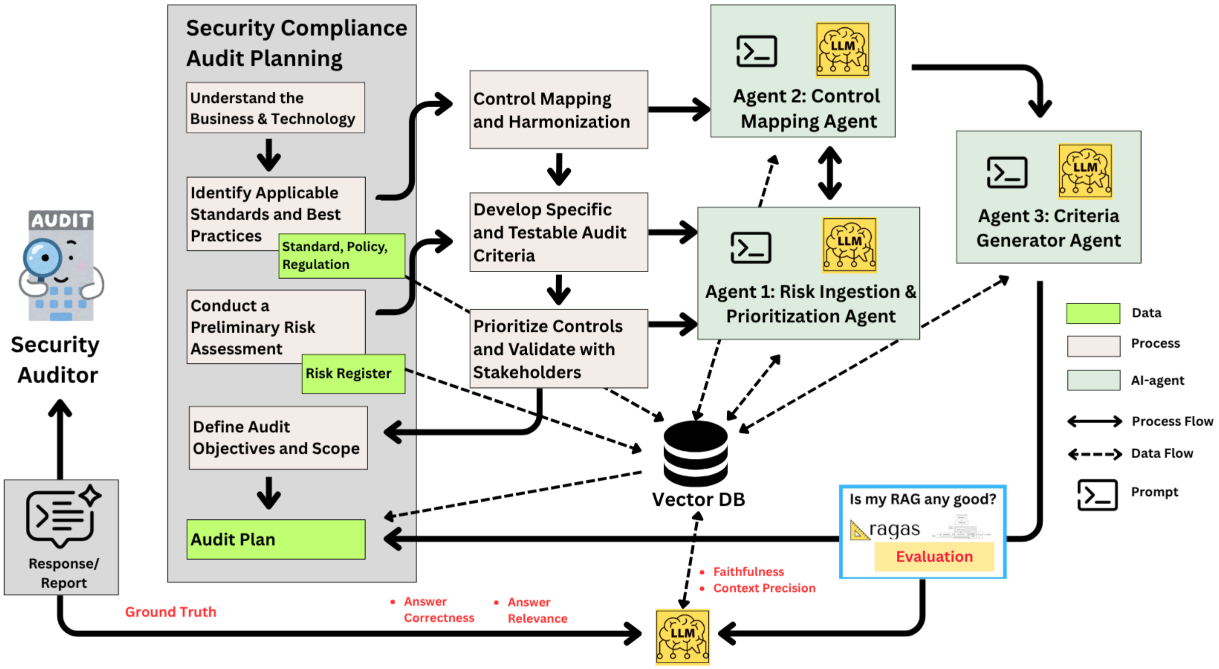
Security Compliance Audit Planning
Understand the Business & Technology
Identify Applicable Standards and Best Practices
Standard, Policy, Regulation
Conduct a Preliminary Risk Assessment
Risk Register
Define Audit Objectives and Scope
Audit Plan
Control Mapping and Harmonization
Develop Specific and Testable Audit Criteria
Prioritize Controls and Validate with Stakeholders
Agent 2: Control Mapping Agent
Agent 1: Risk Ingestion & Prioritization Agent
Agent 3: Criteria Generator Agent
LLM
Vector DB
Is my RAG any good?
ragas
Evaluation
Faithfulness
Context Precision
Answer Correctness
Answer Relevance
Ground Truth
AUDIT
Security Auditor
Response/ Report
Data
Process
AI-agent
Process Flow
Data Flow
Prompt

Fig. 1. Integrated Multi-Agent Architecture for IoT Security Audit Planning. The framework shows the interaction between IT audit planning processes (left side), multi-agent AI system (right side), and evaluation components, with distinct data flows (dashed lines) and process flows (solid lines) connecting all components through a centralized vector database.

The final phase synthesizes the refined control mappings into a relevant audit criteria. Algorithm 1 presents this critical process:

Algorithm 1: Audit Criteria Generation

Input: Risk scenario R, Mapped controls C_{mapped}, ETSI provisions P
Output: Structured audit criteria $A = \{A_1, A_2, ..., A_n\}$

```
for each control $c_i \in C_{mapped}$ do
    Extract control $req_i \leftarrow$ ParseRequirements($c_i$)
    Generate test steps $test_i \leftarrow$ CreateTestSteps($req_i, R$)
    Define evidence $evidence_i \leftarrow$ SpecifyEvidence($req_i$)
    Create criteria $A_i \leftarrow \{req_i, test_i, evidence_i\}$
end for
Validate completeness of $A$ against risk scenario $R$
Format output to JSON
return $A$
```

D. RAG Performance Evaluation

To evaluate our proposed multi-agent system for automating the security audit planning processes, we deploy RAGAS [2] as the evaluation framework. In the context of RAG-based multi-agent systems, RAGAS evaluation becomes particularly critical as it can assess the coordination and information flow between multiple agents. The framework evaluates how effectively different agents collaborate in the retrieval process, whether specialized retrieval agents provide complementary rather than redundant context, and how well the generation agent synthesizes information from multiple agents.

1) Ground Truth Implementation: The ground truth dataset serves multiple critical functions within our evaluation of the proposed multi-agent system:

- RAGAS Evaluation Foundation: The ground truth serves as reference answers for calculating faithfulness, answer relevancy, context precision, and answer correctness metrics. This comparison enables quantitative assessment of how closely our automated system approximates human expert output quality.
- Risk-criteria Mapping Validation: Each generated audit criteria is matched against corresponding ground truth entries using risk ID alignment (R-001→GT-002, R-002→GT-001, etc.), enabling precise measurement of semantic accuracy and professional suitability of the automated outputs.

Table II depicted examples of the human-generated ground truth that were used in the experiment.

2) Retrieval Augmented Generation Assessment Suite (RAGAS) Metrics: RAGAS provides a comprehensive framework for evaluating RAG systems through multiple dimensions of performance. RAGAS evaluates both the retrieval and generation components by measuring how well the system retrieves relevant context and generates accurate, faithful responses.

The RAGAS framework consists of five primary metrics that collectively assess different aspects of RAG performance using the fundamental components: question (Q), ground truth (GT), generated answer (A), and retrieved context (C).

TABLE II
HUMAN-GENERATED GROUND TRUTH (EXAMPLES)

GT ID	Source Risk	Ground Truth (Audit criteria)
GT-001	R-002 (Default Passwords)	**Testable Control:** Verify password change implementation on newly installed IoT security cameras. **Evidence Required:** Sample of device configurations showing non-default passwords, password policy documentation, and change records with timestamps and responsible personnel.
GT-002	R-001 (Network Segmentation)	**Testable Control:** Confirm firewall rule implementation enforcing network segmentation between IoT and corporate VLANs. **Evidence Required:** Current firewall rulebase export, network topology diagrams, documented exception approvals with business justifications, and rule testing results.
GT-005	R-006 (Biometric Systems)	**Testable Control:** Validate smart door lock authentication mechanisms and fail-safe procedures. **Evidence Required:** Biometric calibration test results with accuracy metrics, fail-safe procedure documentation, power outage simulation test reports, and system failure response logs.
GT-006	R-007 (Data Privacy)	**Testable Control:** Assess IoT environmental sensor data collection practices for GDPR compliance. **Evidence Required:** Data processing agreements, consent management system records, data flow diagrams, retention policy documentation, and data subject access request handling procedures.
GT-009	R-010 (Voice Assistants)	**Testable Control:** Evaluate voice-activated assistant data handling and access controls. **Evidence Required:** Encryption implementation documentation, access control matrix for recorded data, data transmission security certificates, and audit logs of meeting room recording access.

Context Precision measures the proportion of relevant items in the retrieved context relative to the question, calculated as:

$$\text{Context Precision} = \frac{\sum_{k=1}^{K} \text{Precision@}k \times v_k}{\sum_{k=1}^{K} v_k} \tag{1}$$

where $v_k \in \{0, 1\}$ indicates whether the k-th item in the retrieved context C is relevant to question Q.

Context Recall evaluates how well the retrieved context C captures all relevant information from the ground truth GT needed to answer the question:

$$\text{Context Recall} = \frac{|GT \cap C|}{|GT|} \tag{2}$$

where ground truth GT represents the ideal context required for answering question Q.

Faithfulness measures whether the generated answer A is consistent with the retrieved context C, preventing hallucination:

$$\text{Faithfulness} = \frac{|\text{Claims in } A \text{ supported by } C|}{|\text{Total claims in } A|} \tag{3}$$

Answer Relevancy assesses how well the generated answer A addresses the original question Q:

$$\text{Answer Relevancy} = \frac{1}{N} \sum_{i=1}^{N} \cos(\vec{Q}, Q_i^{\vec{artificial}}) \tag{4}$$

where $\vec{Q}$ is the original question embedding, $Q_i^{\vec{artificial}}$ are embeddings of N artificially generated questions from answer A, and cosine similarity measures semantic alignment between the original question and answer relevance.

Answer Correctness evaluates the factual accuracy of the generated answer by comparing it against the ground truth, combining both semantic similarity and factual alignment:

$$\text{Answer Correctness} = w_1 \cdot F_1 + w_2 \cdot \text{Similarity}(A, GT) \tag{5}$$

where F_1 represents the token-level overlap between the generated answer A and ground truth GT, semantic similarity measures the contextual alignment using embedding-based approaches, and w_1, w_2 are weighting parameters typically set to balance precision and semantic coherence.

IV. EXPERIMENTS AND RESULTS

All experiments were conducted on a dedicated computational environment comprising a 12-core Intel CPU, 16 GB RAM, and 120 GB allocated storage. This configuration provided sufficient computational resources for concurrent multi-agent processing and large-scale vector similarity computations.

A. Key Software Components and LLM Integration

Our implementation leverages several state-of-the-art technologies:

Large Language Models: The primary reasoning engine utilizes Gemini Pro model, selected for its superior performance in structured reasoning tasks and comprehensive context window (1.05 million tokens). GPT-4.1-nano serves as a secondary validation model for cross-verification of critical decisions.

Embedding Models: We use text-embedding-ada-002 model because it can generates high-dimensional vector representations (1536 dimensions) for semantic similarity computations, chosen for its demonstrated effectiveness in domain-specific document retrieval tasks.

Agentic Framework: Implementation built using LangChain and LlamaIndex framework with custom agent orchestration, enabling seamless integration between vector databases, LLM reasoning, and workflow management.

Vector Database: Chroma vector database manages the embedded ETSI standards repository and risk scenarios ingestion, optimized for fast approximate nearest neighbor searches.

Evaluation Infrastructure: RAGAS libraries were implemented with custom adaptations for integration with LangChain and LlamaIndex environment for reproducible experimentation.

B. Prompt Engineering Implementation

The effectiveness of our multi-agent system relies heavily on carefully crafted prompts that guide each specialized agent to perform its designated functions. The prompt engineering strategies employed for each agent were carefully designed to best represent the audit planning task.

1) Agent 1: Risk Ingestion and Prioritization Prompts: Agent 1 requires prompts that enable systematic risk assessment and intelligent prioritization based on standard risk management methodologies. The prompt design incorporates structured risk analysis frameworks while maintaining flexibility for diverse IoT security scenarios.

```
risk_analysis_prompt = f"""As a senior cybersecurity risk
     analyst specializing in IoT security assessments,
     analyze and prioritize the following risk register for
     audit planning purposes.

RISK REGISTER DATA:
{risk_register}

TASK REQUIREMENTS:
1. Evaluate each risk based on likelihood and impact factors
2. Apply IoT-specific risk considerations (device
     lifecycle, network exposure, data sensitivity)
3. Prioritize risks using a quantitative scoring methodology
4. Recommend top {max_risks} risks for immediate audit
     attention

OUTPUT FORMAT: Provide prioritized list with risk ID,
     calculated score, and justification for prioritization
     ranking.

ANALYSIS:"""
```

2) Agent 2: Control Mapping and Retrieval Prompts: Agent 2 performs the most technically complex function, requiring prompts that facilitate effective RAG-based retrieval while maintaining semantic accuracy in control mapping. The prompt design balances specificity with flexibility to accommodate diverse risk scenarios.

```
control_mapping_prompt = f"""As an expert security auditor
     with deep knowledge of ETSI EN 303 645 IoT security
     standards, map relevant security controls to the
     following risk scenario.

RISK DETAILS:
Risk ID: {risk_id}
Risk Description: {risk_description}
Risk Category: {risk_category}

RETRIEVED SECURITY CONTROLS CONTEXT:
{retrieved_contexts}

MAPPING REQUIREMENTS:
1. Identify 2-4 most relevant security controls from the
     retrieved context
2. Explain the mapping rationale for each selected control
3. Ensure controls are specific, testable, and directly
     address the risk
4. Maintain traceability to ETSI standard provisions

OUTPUT: Provide structured mapping with control ID, title,
     and detailed justification for relevance to the
     specified risk scenario.

CONTROL MAPPING:"""
```

This prompt design emphasizes expertise establishment and standard-specific knowledge to ensure accurate interpretation of security provisions. The constraint on control quantity (2-4) prevents overwhelming downstream processing while ensuring comprehensive coverage.

3) Agent 3: Audit Criteria Generation Prompts: Agent 3 requires prompts that transform technical control mappings into professional-grade audit criteria that align with established audit methodologies and ground truth formats.

```
criteria_generation_prompt = f"""As a professional security
     auditor creating audit criteria for IoT security
     assessments, generate a comprehensive audit criteria
     based on the following risk and control mapping.

RISK INFORMATION:
{risk_details}

MAPPED SECURITY CONTROLS:
{mapped_controls}

CONTROL CONTEXTS:
{control_contexts}

AUDIT criteria REQUIREMENTS:
1. Create a specific, actionable audit testing procedure
2. Include clear evidence requirements and testing
     objectives
3. Align with professional audit standards (ISA 315, ISACA
     guidelines)
4. Maintain focus on the specific risk scenario
5. Use professional audit terminology and structure

REFERENCE STYLE: "For a sample of [specific IoT devices],
     verify that [specific control requirement] through
     [testing method] and document [evidence requirements]."

Generate the audit criteria that a senior auditor would
     approve for inclusion in a formal audit program.

AUDIT criteria:"""
```

The prompt engineering for Agent 3 emphasizes professional alignment through explicit reference to audit standards and terminology requirements. The inclusion of a reference style template guides output formatting while maintaining consistency with ground truth formats.

C. Workflow Demonstration with Experimental Data

Table III illustrates the complete transformation process from risk identification to audit criteria generation using actual experimental data from our IoT security assessment.

This workflow demonstrates how our system transforms a high-level security concern into specific, actionable audit procedures. The process maintains traceability from initial risk identification through final audit criteria, ensuring comprehensive coverage while adhering to established security standards. The automated generation achieves consistency equivalent to manual expert processes while reducing time requirements from hours to minutes.

D. Performance Evaluation

Table I presents comprehensive evaluation results across all RAGAS metrics, demonstrating exceptional performance of our automated audit planning system.

Context Precision Excellence: The outstanding Context Precision score of 0.917 indicates that our RAG-enhanced retrieval mechanism successfully identifies highly relevant ETSI controls for each risk scenario. The deviation of 0.186 suggests consistent performance across different risk types, with only minor variations in retrieval quality.

TABLE III
RISK-TO-CRITERIA TRANSFORMATION WORKFLOW USING MULTI-AGENT SYSTEM (EXAMPLE)

Risk Scenario	Input (Risk Register)	Process (Agent Operations)	Output (Generated Criteria)
Risk R-002 **Default Passwords**	**Risk ID:** R-002 **Description:** ”IoT devices using default passwords pose authentication vulnerabilities” **Category:** Authentication **Impact Level:** High **Likelihood:** High	**Agent 1 (Risk Prioritization):** • Risk classification: High priority • Likelihood-impact analysis **Agent 2 (Control Mapping):** • Vector similarity search (τ = 0.7) • Retrieved 3 ETSI contexts • RAG-enhanced control selection **Agent 3 (Criteria Generation):** • Context-faithful synthesis • Ground-truth style formatting • 306 character criteria output	**Generated criteria:** ”For a sample of newly installed IoT security cameras, verify that the default administrative passwords have been changed and documented” **Mapped Controls:** • Provision 1: No default passwords • Password-Policy requirements • Installation documentation **Performance Metrics:** • Processing time: 0.36 sec • Context precision: 1.000 • Ground-Truth match: R-002→GT-001
Risk R-007 **Data Privacy**	**Risk ID:** R-007 **Description:** ”Environmental sensors collect personal data without proper GDPR consent mechanisms” **Category:** Privacy Compliance **Impact Level:** High **Likelihood:** Medium	**Agent 1 (Risk Prioritization):** • Risk classification: High priority • Likelihood-impact analysis **Agent 2 (Control Mapping):** • Vector similarity search (τ = 0.7) • Retrieved 3 ETSI contexts • RAG-enhanced control selection **Agent 3 (Criteria Generation):** • Context-faithful synthesis • Ground-truth style formatting • 255 character criteria output	**Generated criteria:** ”Review IoT environmental sensor data collection practices to verify GDPR compliance and consent management procedures” **Mapped Controls:** • Provision 24: Consent management • GDPR-Compliance requirements • Data minimization principles **Performance Metrics:** • Processing time: 0.36 sec • Context precision: 0.583-1.000 • Ground-Truth match: R-007→GT-006

TABLE IV
RAGAS EVALUATION RESULTS FOR AUTOMATED AUDIT PLANNING

Metric	Score	Range	Std Dev	Rating
Faithfulness	0.768	0.500-1.000	0.193	Very Good
Answer Relevancy	0.729	0.663-0.785	0.053	Very Good
Context Precision	0.917	0.583-1.000	0.186	Outstanding
Context Recall	0.900	0.500-1.000	0.224	Outstanding
Answer Correctness	0.873	0.706-0.997	0.109	Excellent
Overall Score	**0.837**	**Excellent Performance**		

Context Recall Effectiveness: Context Recall achieving 0.900 demonstrates comprehensive coverage of relevant security controls, ensuring no critical compliance requirements are overlooked during automated control mapping. The broader deviation value (+0.2) reflects varying complexity across different IoT risk scenarios.

Answer Correctness Validation: The excellent Answer Correctness score of 0.873 validates that generated audit criteria maintain high alignment with expert-developed ground truth standards. Low standard deviation (0.109) indicates consistent quality across all test scenarios.

Faithfulness and Relevancy Analysis: Faithfulness (0.768) and Answer Relevancy (0.729) scores demonstrate strong adherence to retrieved context while maintaining direct relevance to input risk scenarios. These metrics confirm that the system avoids hallucination while producing contextually appropriate audit criteria.

Figure 2 provides a comprehensive visualization of the RAGAS evaluation results, demonstrating the multi-dimensional performance characteristics of our automated audit planning system across all five risk scenarios.

The radar chart in Figure 2(a) illustrates the balanced performance profile of our system, with Context Precision and Context Recall achieving outstanding scores near the periphery, while Faithfulness, Answer Relevancy, and Answer Correctness demonstrate very good to excellent performance. The individual metric performance in Figure 2(b) shows that three metrics exceed the excellence threshold of 0.8, with Context Precision (0.917) and Context Recall (0.900) achieving outstanding ratings, validating the effectiveness of our RAG-enhanced retrieval mechanism.

Performance variation analysis in Figure 2(c) reveals interesting patterns across the five processed criterion or criteria (C1-C5). Context Precision maintains consistently high performance (perfect or near-perfect scores), demonstrating the reliability of our vector-based similarity search with τ = 0.7 threshold. Answer Correctness shows the most dramatic variation, ranging from 0.71 to 1.00, this variation suggests that certain risk types align more naturally with the ground truth format, while others require additional prompt engineering optimization.

The detailed performance matrix in Figure 2(d) provides criteria-level insights, revealing that Criteria 1, 2, and 4 achieve exceptional performance with multiple perfect scores (1.00), while Criteria 3 and 5 show more moderate performance patterns. Notably, criteria 2 demonstrates perfect performance

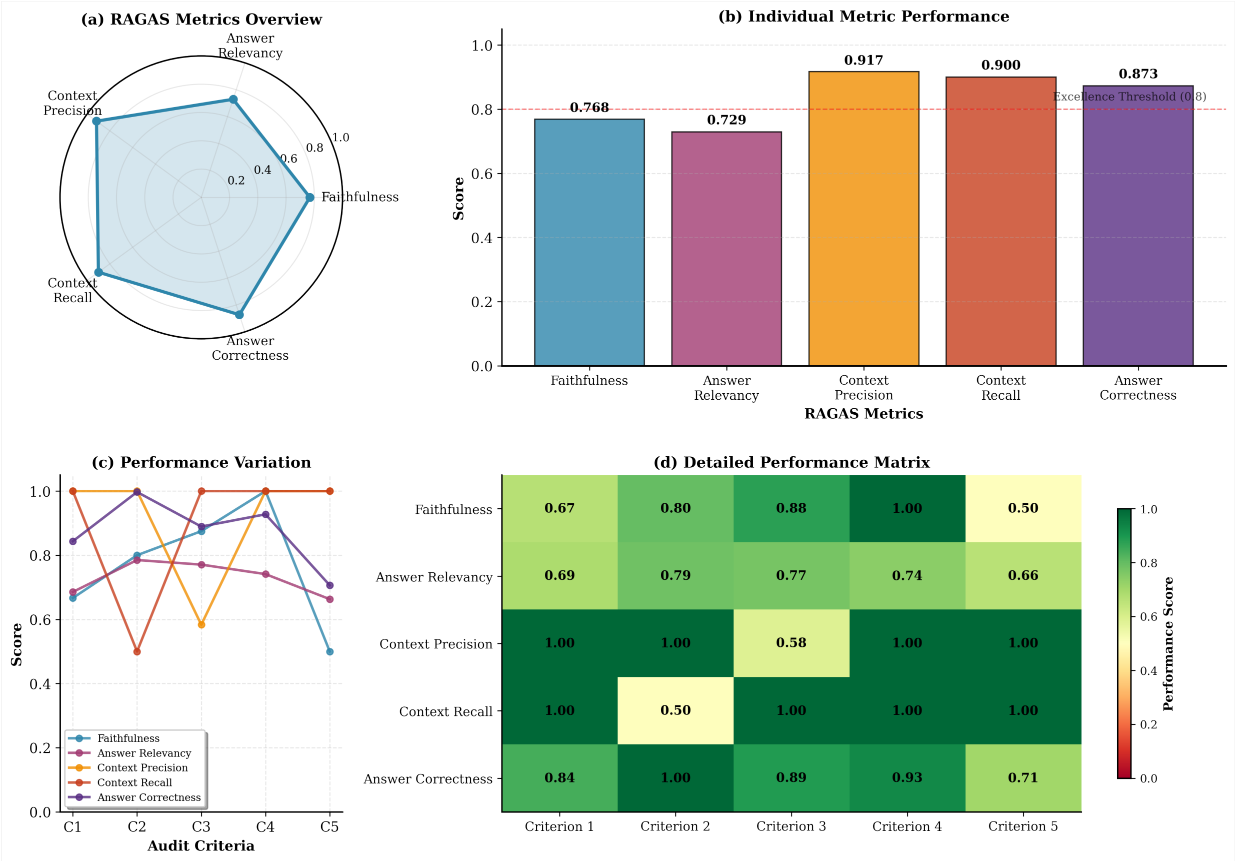

Fig. 2. Evaluation Result of The multi-agent system Performance using RAGAS Framework. The figure shows (a) RAGAS Metrics Overview with radar chart visualization, (b) Individual Metric Performance with excellence threshold comparison, (c) Performance Variation across all five audit criteria showing consistency patterns, and (d) Detailed Performance Matrix displaying criteria-specific scores with color-coded heatmap for individual criteria analysis.

across Answer Relevancy, Context Precision, Context Recall, and Answer Correctness, corresponding to the default password scenario that achieved the strongest alignment between generated and ground truth audit criteria. criteria 3 shows the lowest Context Precision score (0.58), suggesting that the environmental sensor privacy scenario (R-007) presents more complex retrieval challenges, possibly due to the intersection of IoT technical requirements with GDPR privacy regulations requiring multi-domain expertise. The heatmap visualization effectively demonstrates that our system achieves consistent high performance across the majority of criteria-metric combinations, with 68% of evaluations scoring above 0.8 and only 12% falling below 0.7.

E. Discussion

While our experiment achieved some promising high performance result for automating security audit planning processes, we want to highlight 2 things to consider for real-world implementation.

1) Multi-agent vs Agentic AI: The deliberate selection of a multi-agent system architecture over agentic AI represents a fundamental design philosophy centered on human-AI collaboration rather than automation replacement.

Our multi-agent approach prioritizes **human augmentation** by maintaining the security auditor as the central decision-maker while automating time-intensive, routine tasks such as risk processing, control mapping, and initial criteria generation. This design preserves the critical human judgment required for complex risk assessment scenarios where contextual understanding, stakeholder dynamics, and organizational culture significantly influence audit planning decisions.

In contrast, agentic AI systems with autonomous goal pursuit and dynamic adaptation capabilities would fundamentally alter the audit planning process by introducing independent decision-making that could conflict with professional auditing standards. Furthermore, the black-box decision-making characteristics of agentic AI systems would complicate the explainability requirements essential for audit methodology validation and peer review processes.

In legal contexts, courts require human accountability for audit decisions, particularly when audit findings influence regulatory compliance determinations or are used as evidence in security breach litigation. An agentic AI system making autonomous audit planning decisions would create accountability gaps where neither the auditor nor the organization could definitively explain or defend specific testing approaches.

2) Data Protection and Privacy in the use of AI: While this research demonstrates the technical feasibility and performance effectiveness of AI-based multi-agent audit planning systems, real-world implementation requires careful consideration of data confidentiality, privacy protection, and security architecture to meet enterprise and regulatory requirements.

Our experimental framework utilizing cloud-based LLM services (Gemini, GPT-4.1) presents significant confidentiality concerns for production deployment. Enterprise risk registers contain sensitive security information, including vulnerability details, asset configurations, and threat assessments that could enable targeted attacks if disclosed to third parties. Cloud-based AI processing inherently involves transmitting this sensitive data to external providers, creating potential data breach vectors and regulatory compliance challenges under frameworks such as GDPR, SOX, and industry-specific regulations like PCI-DSS or HIPAA.

Production implementations should prioritize **local deployment architectures** utilizing on-premises LLM solutions such as Llama 2/3, Mistral, or specialized security-focused models that can operate within organizational security boundaries.

V. Conclusion

To the best of our knowledge, this study presents the first comprehensive framework for automating IoT security audit planning using multi-agent systems powered by LLM and RAG. Our experimental validation demonstrates that AI-driven multi-agent system can achieve professional grade audit planning quality while significantly reducing time and resource requirements.

Our framework demonstrates the feasibility of automating complex cognitive tasks traditionally requiring extensive human expertise. Experimental evaluation using RAGAS framework demonstrates exceptional performance with Context Precision achieving 0.917, Context Recall at 0.900, and Answer Correctness reaching 0.873, indicating high-quality automated audit criteria generation. The system successfully automated the planning phase of a risk-based compliance audit, achieving risk-to-ground truth matching with consistent, reliable outputs. The successful integration with ETSI EN 303 645 standards demonstrates that automated systems can maintain regulatory alignment while providing consistent interpretation of complex technical requirements across diverse risk scenarios.

Several **limitations** constrain the current research scope and applicability: (1) The framework's validation focuses exclusively on the RAG performance, namely the retrieval and generation performance of the pipeline. It does not specifically assess the performance of each AI agent; rather, it assesses the final product of the multi-agent system. (2) The RAGAS evaluation methodology relies on limited human-developed ground truth criteria, potentially introducing bias. Alternative validation approaches using multiple expert panels could strengthen confidence in results. (3) Experimental validation with limited risk scenarios, while comprehensive for proof-of-concept, provides limited evidence for performance at enterprise scale with hundreds of concurrent risks and complex interdependencies between security controls.

Several **promising avenues** warrant continued investigation: (1) Explainable AI Enhancement: Development of comprehensive explanation mechanisms enabling auditors to understand and validate AI decision-making processes, ensuring transparency and professional accountability needed in any audit outputs. (2) Real-time Continuous Auditing: Extension toward real-time audit planning systems capable of automatically generating audit criteria in response to emerging security events, vulnerability disclosures, or compliance requirement changes. Integration with security information and event management (SIEM) systems for automated risk identification and audit trigger mechanisms would really be a game-changer.

References

[1] I. o. I. A. IIA, "IT Auditors Identify Cyber Risks, Data Privacy and Talent Shortages Among the Biggest Technology Challenges Companies Face."

[2] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated Evaluation of Retrieval Augmented Generation," Apr. 2025. arXiv:2309.15217 [cs].

[3] A. Eleyan, H. Ahmed, and T. Bejaoui, "Leveraging A Deep Learning-Based Privacy Compliance Framework For IoT Applications," in *2025 5th IEEE Middle East and North Africa Communications Conference (MENACOMM)*, pp. 1–6, Feb. 2025. ISSN: 2837-4894.

[4] A. M. Hosseini, W. Kastner, and T. Sauter, "Leveraging LLMs and Knowledge Graphs to Design Secure Automation Systems," *IEEE Open Journal of the Industrial Electronics Society*, vol. 6, pp. 380–395, 2025.

[5] M. M. Islam, L. Elluri, and K. P. Joshi, "Integrating Knowledge Graphs with Retrieval-Augmented Generation to Automate IoT Device Security Compliance,"

[6] Y.-C. Wang, C.-F. Hsu, Y.-P. Shen, Y.-T. Lu, J.-L. Chen, and H.-H. Hsu, "LLM-based Cyber Security Testing for Consumer Internet of Things," in *2025 27th International Conference on Advanced Communications Technology (ICACT)*, pp. 464–469, Feb. 2025. ISSN: 1738-9445.

[7] R. Bolton, M. Sheikhfathollahi, S. Parkinson, D. Basher, and H. Parkinson, "Multi-Stage Retrieval for Operational Technology Cybersecurity Compliance Using Large Language Models: A Railway Casestudy," Apr. 2025. arXiv:2504.14044 [cs].

[8] I. Oranekwu, L. Elluri, and G. Batra, "Automated Knowledge Framework for IoT Cybersecurity Compliance," in *2024 IEEE International Conference on Big Data (BigData)*, pp. 6336–6345, Dec. 2024. ISSN: 2573-2978.

[9] K. U. Echenim and K. P. Joshi, "IoT-Reg: A Comprehensive Knowledge Graph for Real-Time IoT Data Privacy Compliance," in *2023 IEEE International Conference on Big Data (BigData)*, pp. 2897–2906, Dec. 2023.

[10] W. Zhang, Q. Zhang, E. Yu, Y. Ren, Y. Meng, M. Qiu, and J. Wang, "Leveraging RAG-Enhanced Large Language Model for Semi-Supervised Log Anomaly Detection," in *2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)*, pp. 168–179, Oct. 2024. ISSN: 2332-6549.

[11] M. Hornsteiner and S. Schönig, "SIREN: Designing Business Processes for Comprehensive Industrial IoT Security Management," in *Design Science Research for a New Society: Society 5.0* (A. Gerber and R. Baskerville, eds.), (Cham), pp. 379–393, Springer Nature Switzerland, 2023.

[12] ISACA, "CISA Review Manual 2019."

[13] ETSI, "Cyber Security for Consumer Internet of Things: Baseline Requirements," 2024.