セクション階層を活用したRAGによるIoTデバイスのセキュリティ適合性評価

池上 裕香[1,a)]　長谷川 健人[2]　披田野 清良[2]　福島 和英[2]　橋本 和夫[1]　戸川 望[1]

概要：近年，IoT デバイスの急速な普及に伴い，それらに対するセキュリティ強化が国際的に求められている．日本では，情報処理推進機構 (IPA) により，IoT デバイスのセキュリティ適合性を評価・ラベリングする制度 “JC-STAR (Japan Cyber-Security Technical Assessment Requirements)” が開始された．しかし，IoT デバイスは用途や構造が多岐にわたる上，参照すべき製品ドキュメントも膨大かつ多様であるため，手動による評価は高度な専門知識と多大な工数が求められる．よって，評価の自動化が期待される．我々は，RAG (Retrieval-Augmented Generation) を活用し，IoT デバイスのドキュメントから関連情報を抽出し，LLM によって自動的に適合性を評価する手法を提案している．しかし，ドキュメントのページ数が膨大な場合，従来のドキュメント全体を対象としたフラットな検索では関連情報を十分に網羅できず，LLM の回答精度に限界があった．そこで，本稿では，セクション階層構造を活用した RAG ベースのセキュリティ適合性評価手法を提案する．提案手法では，従来の検索とドキュメントの階層構造を活用する検索を組み合わせた新たな検索手法を導入することで，関連情報の検索漏れを低減する．評価実験の結果，提案手法は従来手法と比較して，より関連度の高いチャンクの取得に成功し，LLM による適合性評価の精度が向上したことを確認した．

キーワード：IoT デバイス，セキュリティ，大規模言語モデル，RAG (Retrieval Augmented Generation)，階層検索

Security Compliance Evaluation of IoT Devices Using RAG with Section Hierarchy

Yuka Ikegami[1,a)]　Kento Hasegawa[2]　Seira Hidano[2]　Kazuhide Fukushima[2]
Kazuo Hashimoto[1]　Nozomu Togawa[1]

Abstract: In this paper, we propose a RAG-based security compliance evaluation method that leverages the section hierarchy of documents. Experimental results demonstrate that, compared to conventional methods, the proposed method successfully retrieves more relevant chunks and improves the accuracy of compliance evaluation by LLMs.

Keywords: IoT device, security, large language model, retrieval augmented generation, hierarchical retrieval

1. はじめに

近年，IoT (Internet of Things) デバイスの急速な普及に伴い，それらを標的としたセキュリティインシデントが数多く報告されている．例えば，2016 年には Mirai マルウェアが脆弱な IoT デバイスを大規模に悪用し，多数のサイトを長時間停止させる DDoS 攻撃を引き起こしたほか，2021 年に発見された “NAME:WRECK” 脆弱性は TCP/IP スタックに存在する DNS 処理の不備を悪用し，米国だけでも 18 万台以上のデバイスに被害を及ぼした [1]．

このような背景のもと，世界各国では IoT セキュリティに関する標準化が進められている．欧州における ETSI EN 303 645 [2] や ISO/IEC 27400 [3] をはじめとする国際規格は，IoT 製品に対する共通のセキュリティ基準を提供し，製品の安全性を担保するための基盤となっている．また，それらの基準に準拠した製品にラベルを付与する制度も国際的に整備されつつあり，フィンランドの Finnish

[1] 早稲田大学基幹理工学研究科情報理工・情報通信専攻
Dept. Computer Science and Communications Engineering, Waseda University
[2] 株式会社 KDDI 総合研究所
KDDI Research, Inc.
[a)] yuka.ikegami@togawa.cs.waseda.ac.jp

Cybersecurity Label [4] やシンガポールの Cybersecurity Labelling Scheme (CLS) [5] が例として挙げられる．これらの制度は，国際規格 (ETSI EN 303 645) に基づき製品のセキュリティ水準を評価し，基準を満たした製品にのみラベルを付与することで，消費者が安全な製品を選択できる環境を整えている．

日本でも同様の取り組みが進められており，2025 年 3 月より，情報処理推進機構 (IPA) が “セキュリティ要件適合評価及びラベリング制度 (JC-STAR)” [6] の運用を開始した．JC-STAR は，ETSI EN 303 645 や NISTIR 8425 などの国内外規格と整合性を保ちつつ，独自のセキュリティ要件に基づいて IoT 製品の適合性を評価し，その結果に応じてラベルを付与するものである．レベル 1 (☆ 1) の評価は，汎用的な IoT 製品に求められる最低限のセキュリティ要件に基づき，ベンダが所定の評価基準にしたがって自己評価を行う形式となっている．しかし，IoT 製品の種類や構造は多岐にわたり，評価に必要な技術文書も製品ごとに形式や記述内容が大きく異なる．これにより，手動による評価は評価者の知識や経験に左右されやすいという課題が想定される．

上記の課題を受け，膨大な情報を迅速かつ高精度で処理する能力を持つ大規模言語モデル (LLM) の活用が注目されている．LLM によって IoT デバイスの多様かつ膨大なドキュメントを効率的に解析することで，適合性評価の負担軽減が期待できる．

我々は，LLM を活用し，JC-STAR における☆ 1 の適合性評価項目に対するドキュメント評価の自動化手法 [7] を提案している．[7] では，RAG (Retrieval-Augmented Generation) [8] を用いて，IoT デバイスのユーザマニュアルなどから関連情報を検索し，その内容を LLM に渡すことで適合性を評価するアプローチをとっており，複数のデバイスを用いた実験を通じてその有効性が示されている [7], [9]．しかし，ドキュメントのページ数が膨大な場合，従来のドキュメント全体から類似度の高いチャンクを検索するようなフラットな検索では，関連情報を十分に網羅できず，LLM の判断に必要なコンテキストが欠落することがあった．

本稿では，セキュリティ適合性評価自動化手法 [7] を拡張し，従来の全体スコープ検索とドキュメントの階層構造を活用した新たなセクションスコープ検索を組み合わせて関連チャンクを検索することで，LLM へ入力するコンテキストの充実を図り，LLM の判断精度の向上を目指す．

本稿の貢献を以下に示す．

(1) JC-STAR における☆ 1 のセキュリティ適合性評価自動化手法 [7] を拡張し，ドキュメントの階層構造を活用した新たな検索手法を導入する．
(2) 従来の全体スコープ検索に加え，ドキュメントの階層構造を活用するセクションスコープ検索を組み合わせて関連チャンクを取得する．全体スコープ検索は網羅性を，セクションスコープ検索はまとまりを重視して情報を抽出することで情報の過不足を解消し，LLM の判定精度の向上を図る．
(3) 3 種類のデバイスに対する評価実験の結果，全てのデバイスにおいて，提案手法は比較手法と比べて全ての評価指標で同等または上回る性能を示した．特にユーザマニュアルが 4200 ページと最大規模であったデバイスに対し，提案手法は TPR と F-measure の両方で大幅に向上しており，長大なドキュメントにおいても効果的に情報を抽出できることが確認された．

2. 関連研究

RAG は，LLM の知識やコンテキスト長の制約を補完するために，外部のドキュメントから関連情報を検索し，その情報をもとに応答を生成する技術である．ユーザのクエリに対して適切な関連チャンクを検索できるかどうかが LLM 応答の品質に大きく影響するため，検索精度の向上が極めて重要である．特に大規模なドキュメントを扱う場合，ドキュメント全体を対象とした従来の検索手法では，関連する情報の一部が取得できたとしても，それだけでは判断材料として不十分なことが多く，LLM が正確に応答するためにはその前後や上位セクションなど，周辺の情報も不可欠となる．そのため，ドキュメントの階層構造を活用して，文脈的に一貫性のある情報をまとまりとして取得する検索アプローチが重要である．

ドキュメントの階層構造を活用した RAG 手法として [10], [11] が提案されている．[10] では，長文のドキュメントをセクション，サブセクション，段落といったツリー構造に変換し，その階層構造を活用して関連情報を検索する手法である．具体的には，Leaf-to-Root および Root-to-Leaf の両方向から各ノード (セクション，サブセクション，段落) の関連度を評価し，ローカル (ノード単位) およびグローバル (ツリー全体) な観点で関連スコアを算出する．そして，関連スコアが高い親ノード配下のチャンクをプロンプトに挿入する．

[11] も同様に，長文のドキュメントを LLM を用いてツリー構造へと変換し，Leaf-to-Root および Root-to-Leaf の双方向検索を行う．関連度の高いノードの親ノード配下にある全てのチャンクを取得し，リランキングによって最終的に使用する情報を選定する．

上記 2 つの手法はいずれも，ドキュメントの階層性を活用することで，関連性の高い情報をまとまりとして抽出し，RAG の精度を向上させた．

本稿では，セクション階層に基づく関連チャンクの検索と統合を行う．特に本稿では，対象とするドキュメント (IoT デバイスのユーザマニュアルや脆弱性開示ポリシー) が PDF や HTML 形式であり，明示的なセクション階層を有しているため，既存の Python ライブラリを用いることで高精度なセクション階層を取得できる．

提案手法では，従来の全体スコープ検索と新たに導入したセクションスコープ検索の 2 つのモードに分けて関連チャンクを検索する．全体スコープ検索は，ドキュメント全体を対象にセクション構造を考慮せず，クエリと直接的に関連する情報を網羅的に取得することを目的とする．一方，セクションスコープ検索では，事前に構築された階層ツリーに基づき，検索クエリに関連するセクションを特定し，その配下の情報をまとまりとして取得することで，LLM が正しく判断するために必要な周辺情報もあわせて取得する．2 つの検索モードを組み合わせることで，断片

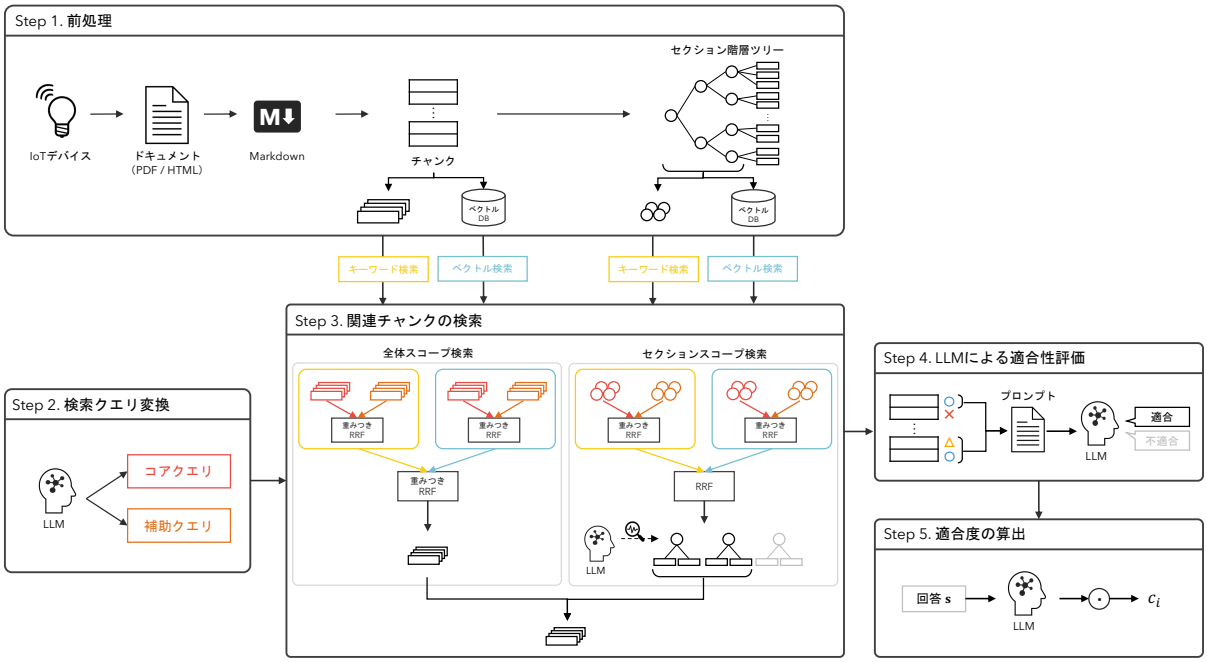

図 1: 提案手法の全体フロー.

的な情報に基づく誤判断が軽減され，LLM による応答の正確性が向上することが期待される．

3. 提案手法

本章では，LLM を用いた JC-STAR の適合性評価自動化手法を提案する．提案手法のフローを図 1 に示す．検査対象デバイスのドキュメントから抽出したデバイス情報を LLM に渡すことで，そのデバイスが各評価項目を満たすか否かを判定する．提案手法の流れを以下に示す．

3.1 Step 1. 前処理

まず，IoT デバイスのドキュメントを用いてベクトルデータベースとセクション階層ツリーを構築する．ベクトルデータベースおよびセクション階層ツリーはそれぞれ全体スコープ検索とセクションスコープ検索で使用される．

Step 1 は以下の 3 つの手順で構成される．

Step 1-1. ソースデータの Markdown 変換: IoT デバイスのドキュメントをソースデータとして RAG にロードする．このとき，PDF や HTML 形式のドキュメントを Markdown に変換する．IoT デバイスのユーザマニュアルなどのドキュメントでは “セットアップ” や “ファームウェアアップデート” などのセクションが階層的に整理されている場合が多いため，Markdown に変換することで，それらの構造的な情報 (セクション，段落) を保持できる．PDF の Markdown 変換には pymupdf4llm [12] を，HTML の Markdown 変換には markdownify [13] を使用した．

Step 1-2. ソースデータの分割: Step 1-1 でロードしたドキュメントのテキストをチャンクと呼ばれる小さな単位で分割する．チャンク分割には LangChain の MarkdownHeaderTextSplitter [14] と RecursiveCharacterTextSplitter [15] を使用する．まず，MarkdownHeaderTextSplitter により Markdown 中の見出しを基準にしてセクション単位で分割し，文章量が多いセクションは RecursiveCharacterTextSplitter によって段落ごとに分割する．各チャンクはメタデータとして，そのチャンクが位置するセクション名と，そのセクションの階層 (セクションレベル) を表す H1〜H6 の文字列情報を保持する．

Step 1-3. ベクトルデータベース・階層ツリーの構築: ベクトルデータベースはエンベディングによりベクトル表現に変換されたチャンクを用いて構築される．このデータベースはベクトル空間内での各チャンクの位置を表現するため，効率的な検索と類似度の計算を可能にする．セクション階層ツリーは各チャンクのメタデータに付与されたセクションレベル (H1〜H6) に基づいて構築される．各セクションレベル (H1〜H6) はツリー構造における深さに対応しており，上位のセクションが下位のセクションを包含する構造となる．例えば，IoT デバイスのマニュアルにおいて，“H1: はじめに，H2: デバイスの概要，H3: 同梱物” という階層情報を持つチャンクと，“H1: はじめに，H2: デバイスの概要，H3: 製品仕様” という階層情報を持つチャンクが存在する場合，構築される階層ツリーは，“はじめに” を最上位ノードとし，その下に “デバイスの概要”，さらにその下に “同梱物” と “製品仕様” が並列に配置された形になる．そして，最下位のセクションノード下には各チャンクのテキストが保持される．

3.2 Step 2. LLM による検索クエリ変換

Step 2 では，Rewrite-retrieve-read [16] に従い，セキュリティ項目の内容を基に，Step 3 で関連チャンクを検索する時に使用する 2 種類の検索クエリを LLM によって生成する．ここで，i 番目に評価するセキュリティ項目 (評価

項目と呼ぶ) を $list_i$ とする．例えば，評価項目 $list_i$ は “適切な認証機構の使用を要件とする項目” などが該当する．2 種類の検索クエリのうち，一つは，ドキュメント内で最も関連性の高いセクションを特定するために不可欠なキーワードで構成されるコアクエリ集合である．もう一つは，関連語，派生語，類義語，上位概念といったより広い情報を捉える補助クエリ集合である．これにより，コアクエリが精度を担保しつつ，補助クエリが情報の網羅性を補完する構成が実現される．Step 2 で使用するプロンプトのテンプレートを以下に示す．

Step 2. プロンプト

You are an expert in searching through technical documents, such as PDF user manuals. We are evaluating whether an IoT device satisfies the following assessment item based on its user manual. If you were to search through a PDF document to find relevant information about this item, what search queries (or keywords) would you use?

Please output the search queries in JSON format with two categories: 1. “core_keywords”: Select only the **most critical and directly relevant** terms or phrases that are **essential for locating the most relevant sections** of the document. These should be limited in number (typically 3-5).
2. “supplementary_keywords”: Include additional related terms, variations, synonyms, broader concepts, or indirect indicators that could also help, but are not as central as the core keywords.

Follow these instructions carefully:
1. Output only **one JSON object**.
2. DO NOT include any explanation, comment, or additional text before or after the JSON.
3. Ensure the JSON is syntactically valid and properly formatted.

Assessment item:
{ 評価項目 $list_i$ の記述 }

Output format (JSON):
{“core_keywords”: [“exact phrase 1”, “exact phrase 2”, ...], “supplementary_keywords”: [“related term 1”, “related term 2”, ...]}

3.3 Step 3. 関連チャンクの検索

Step 3 では，全体スコープ検索とセクションスコープ検索の 2 つのモードに分けて関連チャンクを検索する．

3.3.1 全体スコープ検索

全体スコープ検索では，ドキュメント全体に対し，[7] で提案した 2 段階の重みつき RRF (Reciprocal Rank Fusion) による検索を実施する．第 1 段階として，まず Step 2 で取得したコアクエリ集合と補助クエリ集合を用いてそれぞれキーワード検索を実施し，2 つの検索結果をコア集合を重視した重み ($w_{core} : w_{sup} = 0.8 : 0.2$) で RRF により統合する．$w_{core}$ と w_{sup} はコアクエリ集合と補助クエリ集合を用いたそれぞれの検索結果に対する RRF の重みを示す．同様にして，コアクエリ集合と補助クエリ集合を用いてそれぞれベクトル検索を実施し，2 つの検索結果をコア集合を重視した重み ($w_{core} : w_{sup} = 0.8 : 0.2$) で RRF により統合する．第 2 段階として，上記の統合処理をキーワード検索とベクトル検索の結果の双方に対して実施し，キーワード検索を重視した重み ($w_{keyword} : w_{vector} = 0.8 : 0.2$) で再度 RRF により統合する．$w_{keyword}$ と w_{vector} はキーワード検索とベクトル検索の検索結果に対する RRF の重みを示す．

IoT デバイスのユーザマニュアルなどのドキュメントは，定型的な用語で記述されていることが多く，語彙一致に強いキーワード検索が有効に機能する傾向にある．一方，ベクトル検索による意味的類似性の検索は，主に補助クエリ集合が包括する語彙や概念の多様性によって補完されるため，補助的な位置付けとなっている．これにより，検索クエリと直接的に関連するチャンクを網羅的に取得できる．

3.3.2 セクションスコープ検索

セクションスコープ検索では，まずキーワード検索およびベクトル検索により関連するセクション候補を絞り込み，その中で特に関連度の高いセクションを LLM に選定させる．そして，選定されたセクション配下のチャンクを取得することで，文脈的にまとまりのある情報を収集する．

まずセクション階層ツリーのうちすべての最下位セクションを抽出し，そのセクションパス集合を取得する．例えば，“H1: はじめに，H2: デバイスの概要，H3: 同梱物” という階層が存在する場合，“はじめに > デバイスの概要 > 同梱物” というセクションパスが得られる．次に，セクションパスの集合に対し，全体スコープ検索と同様の 2 段階の重みつき RRF を実行し，候補となるセクションを検索する．ただし，全体スコープ検索の 2 段階目の RRF はキーワード検索を重視した重みを用いて統合したが，セクションスコープ検索の 2 段階目の RRF はベクトル検索とキーワード検索の結果を等しい重みで統合する．これは，セクションスコープ検索においては，検索対象がドキュメント全体ではなくセクションパスのみに限定されているため，キーワード一致の精度が相対的に低下しやすく，意味的な類似性を捉えるベクトル検索の重要性が高まるためである．

その後，上記で検索されたセクションパスの中から特に関連度の高いものを LLM によって選定する．その時のプロンプトを以下に示す．

Step 3. プロンプト (セクションスコープ検索)

You are an expert in IoT devices. Your task is to determine which sections are especially relevant for evaluating whether an IoT device satisfies the following assessment item.

Respond **only** with a JSON array of the exact section names (verbatim, including any special characters like ‘*’, ‘_’, etc.) from the provided list that are highly relevant for judging compliance. If none are relevant, return an empty array (‘[]’). Do **not** include any explanations or extra text. Do **not** modify, remove, or alter the formatting of the provided section names in any way.
Select at most 3 section names. Return no more than 3 items in the array.

Assessment item:
{ 評価項目 $list_i$ の記述 }

Available section names:
{ 検索されたセクションパスの集合 }

Respond with a JSON array. The format should be:
- If relevant sections exist: [“Section Name A”, “Section Name B”, ...] (maximum of 3)
- If no relevant sections: []

LLM によって選定されたセクションの配下にある全てのチャンクを取得し，全体スコープ検索で取得したチャンクと統合する．意味的なまとまりを保持するため，同一セクション内のチャンクは 3 個ずつ統合し，新たな一つのチャンクとした．

3.4 Step 4. LLM による適合性評価

Step 4 では, Chain-of-Thought (CoT) [17] を用いて LLM によって対象デバイスが評価項目 $list_i$ を満たすかどうかを段階的に判定する．Step 4 は以下の 2 つの手順で構成される．

Step 4–1. 関連チャンクのフィルタリング: Step 4–1 では，Step 3 で取得した関連チャンク一つずつに対し，評価項目 $list_i$ との関連度を LLM が判定し，“High”，“Medium”，“Low” の 3 段階でラベル付けを行う．これにより不要な情報やノイズを早期に除去でき，最終判断におけるハルシネーションの低減を図る．Step 4–1 で使用するプロンプトのテンプレートを以下に示す．

Step 4–1. プロンプト

You are an expert in IoT devices. Your task is to evaluate how essential the given document is in determining whether the IoT device satisfies the following assessment item.

Respond with only a JSON object with a single key ”relevance”, whose value is one of the following:
- “High”: The information is critical for judging whether the IoT device meets the assessment item. Without this information, the assessment cannot be accurately made.
- “Medium”: The information is somewhat helpful but not essential. It may support the assessment but is not sufficient on its own.
- “Low”: The information is unrelated or only marginally related to the assessment and does not contribute meaningfully to determining compliance.

Do not include any explanations, reasoning, or code.

Assessment item:
{ 評価項目 $list_i$ の記述 }

Document:
“““
{ 関連チャンク }
”””

Example output:
- {“relevance”: “High”}
- {“relevance”: “Medium”}
- {“relevance”: “Low”}

Step 4–2. 適合性の評価: Step 4–2 では，Step 4–1 で関連度が “High” または “Medium” と判定されたチャンクのみを関連情報として与え，評価項目 $list_i$ を満たしているか否かを LLM が最終的に判定する．使用するプロンプトのテンプレートを以下に示す．

Step 4–2. プロンプト

You are an expert in IoT device evaluation. Your task is to determine whether the IoT device satisfies the assessment item below, based on the provided information.

Example answers:
- Yes, the IoT device satisfies the assessment item because ...
- No, the IoT device does not satisfy the assessment item because ...

Follow these instructions carefully:
1. If the information explicitly confirms that the assessment item is satisfied, answer “Yes”.
2. If the information does not confirm that the assessment item is satisfied, or if there is insufficient information to make a judgment, answer “No”.
3. Avoid making assumptions or inferences beyond what is explicitly stated in the information.
4. Provide a single sentence explaining your answer, using only the information given.

Assessment item:
{ 評価項目 $list_i$ の記述 }

Extracted relevant information:
{Step 4–1 で関連度が “High” または “Medium” と判断されたチャンク }

3.5 Step 5. 適合度の算出

Step 5 では，Step 4 で取得した LLM の回答 $\mathbf{s}$ をテキスト分類 LLM によって定量的に分析し，評価項目 $list_i$ に対する適合度 c_i を算出する．テキスト分類 LLM とは, MNLI データセット [18] によってファインチューニングされた LLM であり，入力された 2 つのテキスト $\mathbf{s}, \mathbf{s}'$ の関係性が “含意”，“中立”，“矛盾” である確率を出力する．含意の確率が高ければ $\mathbf{s}$ が真なら，$\mathbf{s}'$ も必ず真，中立の確率が高ければ $\mathbf{s}$ が真でも，$\mathbf{s}'$ が真か偽かは不明，矛盾の確率が高ければ $\mathbf{s}$ が真なら，$\mathbf{s}'$ は必ず偽と判断できる．

回答に “Yes” が含まれていれば適合度を 1，“No” が含まれていれば適合度を 0 とするのが最も簡単な定量化手法である．しかしながら，“The IoT device satisfies the assessment item.” のように回答内に “Yes” または “No” が明記されていない場合がある．そこで，文脈を理解できるテキスト分類 LLM を活用することで，より柔軟な適合度の算出が可能となる．

評価項目 $list_i$ に対する適合度 c_i は以下の手順で求まる．

(1) 回答 $\mathbf{s}$ と “Yes, the IoT device satisfies the assessment item.” という固定文のペアをテキスト分類 LLM に入力し，回答が固定文に対して “含意”，“中立”，“矛盾” である確率ベクトル $(p_i^e,\ p_i^n,\ p_i^c)$ を求める．

(2) 評価項目 $list_i$ に対する適合度 c_i は，式 (1) のように上記で得られた確率ベクトルと重みベクトル $(1, 0.5, 0)$

表 1: 実装に使用したライブラリやモデル.

実装項目	使用したライブラリやモデル
統合フレームワーク	LangChain [19]
ベクトルデータベース	FAISS [20]
Embedding モデル	sentence-transformers/stsb-xlm-r-multilingual [21]
テキスト生成 LLM	Qwen/Qwen3-8B [22]
テキスト分類 LLM	microsoft/deberta-v2-xlarge-mnli [23]

表 2: 検査対象デバイスとドキュメント情報.

#	ベンダ	デバイス	ドキュメント種類	ドキュメント形式	ページ数
1	ベンダ A	ワイヤレスルータ	ユーザマニュアル	PDF	150
			脆弱性開示ポリシー	PDF	5
2	ベンダ B	業務用ネットワーク機器	ユーザマニュアル	PDF	964
			脆弱性開示ポリシー	HTML	–
3	ベンダ C	業務用ネットワーク機器	ユーザマニュアル	PDF	4200
			脆弱性開示ポリシー	HTML	–

との内積で求める.

$$s_i = \begin{bmatrix} p_i^e \\ p_i^n \\ p_i^c \end{bmatrix}^{\mathsf{T}} \cdot \begin{bmatrix} 1 \\ 0.5 \\ 0 \end{bmatrix} = p_i^e + 0.5 \times p_i^n \qquad (1)$$

JC-STAR の☆ 1 のドキュメント評価に該当するすべての項目に対して Step 2 から Step 5 が繰り返される.

4. 評価実験

本章では，4.1 節で実験環境，4.2 節で実験結果を説明する.

4.1 実験環境

本稿では，JC-STAR の☆ 1 (レベル 1) チェックリストとして公開されている “チェックリスト (2025.05.05) 版 [24]” を使用し，チェックリストの評価ガイドに記載されている評価項目の記述を使用した.

評価実験では，Step 5 で得られた適合度と正解値を比較し評価した．各評価項目の正解値は，検査対象デバイスのドキュメントをもとに，適合する項目には “1”，適合しない項目またはドキュメントからは判断できない項目には “0”，評価適用外の項目には “NA (Not Applicable)” と手動で設定した．NA の項目は，該当する機能や前提条件自体が検査対象デバイスに存在しないケースであり，本稿では評価対象から除外する．これは，本稿の目的が評価項目の適合性を判断することであり，個々の評価項目が検査対象に適用可能かどうか (適用性) の判断はスコープ外であるためである.

比較実験を行うため，従来の全体スコープ検索のみを採用した手法 [7] も同じ条件下で実験した．提案手法と比較手法で共通して使用したライブラリやモデルを表 1 に示す．検査対象デバイスとしては異なるベンダの 3 種類のデバイスを使用し，デバイスやそのドキュメント情報を表 2 に示す.

実験で使用した評価指標について説明する．適合性評価は 2 値分類であるため, 不適合である項目を不適合と正しく判断した数を TN (True Negative)，不適合である項目を適合と誤って判断した数を FP (False Positive)，適合である項目を不適合と誤って判断した数を FN (False Negative)，適合である項目を適合と正しく判定した数を TP (True Positive) とする.

上記の指標を用いて, TP/(TP+FN) で定義される TPR, TN/(TN+FP) で定義される TNR, (TP+TN)/(TP+FN+FP + TN) で定義される Accuracy，TP/(TP + FP) で定義される Precision，2(TPR × Precision)/(TPR + Precision) で定義される F-measure を評価指標として用いる.

4.2 実験結果

提案手法と比較手法で得られた実験結果を表 3 に示す．表 3 より，全てのデバイスにおいて，提案手法は比較手法と比べて全ての評価指標で同等または上回る性能を示した．特に F-measure において一貫した改善が見られ，これは提案手法が関連情報をより適切に取得できていることを示唆する.

ベンダ B の業務用ネットワーク機器に対する実験では，提案手法によって FP を 0 に抑えることができ，その結果として TNR および Precision がともに 1.00 となった．これは，LLM に与える情報が適切に制限され，不必要な情報に起因するハルシネーションが抑制されたためである.

さらに，ユーザマニュアルのページ数が 4200 ページと最もドキュメントの規模が大きいベンダ C のデバイスにおいても，提案手法は比較手法に対して TPR を 0.391 から 0.522，F-measure を 0.563 から 0.686 へと大きく向上させた．この結果は，特にドキュメントが長大である場合に，提案手法がドキュメント階層を活用して必要な情報を効果的に抽出できることを示す.

以下では各評価項目に対する結果を詳細に述べる．ベンダ B の業務用ネットワーク機器について，各評価項目の正解値と提案手法を用いて算出した適合度を表 4 に示す．例えば，項目 No.1 ②–B について，比較手法により算出した適合度は 0.002 である一方，提案手法により算出した適合度は 0.993 となり，正解値 1 に近い値をとった．評価項目の内容は “複数の認証要素を利用した多要素認証機能が実装されているか否か” を問うものである．まず，Step 2 で評価項目の内容から検索クエリが以下のように生成された.

Step 2. LLM の回答 (No.1 ②–B)

```
{
“core_keywords”: [
    “multi-factor authentication”,
    “access control method”,
    “authentication implementation”,
    “security features”
],
“supplementary_keywords”: [
    “two-factor authentication”,
    “user authentication”,
    “security protocols”,
    “device authentication”,
    “login security”
]
}
```

Step 3 の全体スコープ検索では，多要素認証で使用される Time-based One-Time Password (TOTP) に関するチャンクは検索されなかったが，セクションスコープ検索では，LLM によって “Understanding TOTP” や “Configuring Admin/User Realm to Associate a TOTP Authentication

表 3: 全デバイスに対する提案手法と比較手法の実験結果.

デバイス	手法	TN	FP	FN	TP	TPR	TNR	Accuracy	Precision	F-measure
ベンダ A ワイヤレスルータ	提案手法	33	0	8	6	0.429	1.00	0.830	1.00	0.600
	比較手法	33	0	9	5	0.357	1.00	0.809	1.00	0.526
ベンダ B 業務用ネットワーク機器	提案手法	26	0	12	8	0.400	1.00	0.739	1.00	0.571
	比較手法	24	2	13	7	0.350	0.923	0.674	0.778	0.483
ベンダ C 業務用ネットワーク機器	提案手法	25	0	11	12	0.522	1.00	0.771	1.00	0.686
	比較手法	25	0	14	9	0.391	1.00	0.708	1.00	0.563

表 4: ベンダ B の業務用ネットワーク機器に対する実験結果.

#	項目 No.	小項目	正解値	適合度	#	項目 No.	小項目	正解値	適合度
1	1	①	1	0.992	25		②	0	0.001
2		②-B	1	0.993	26		③	0	0.002
3		②-C	1	0.994	27	11	①	1	0.001
4		②-D	1	0.001	28		②	1	0.001
5		②-E	1	0.995	29		③	1	0.001
6		②-F	0	0.001	30		④	0	0.001
7	2	①	0	0.001	31		⑤	0	0.002
8		②	0	0.001	32	12	①-A	1	0.001
9	3	①	0	0.002	33		①-B	0	0.001
10		②	0	0.002	34		②	0	0.002
11	5	①	1	0.992	35		③	1	0.001
12		②	1	0.990	36	13	①-A	0	0.001
13		③	1	0.991	37		①-B	NA	–
14		④	1	0.991	38		①-C	NA	–
15	7	①	0	0.002	39		②	0	0.002
16		②	0	0.001	40		③	0	0.001
17		③	1	0.002	41	15	①-A	1	0.001
18		④	1	0.001	42		①-B	1	0.001
19	8	①	0	0.001	43		①-C	0	0.001
20		②	0	0.001	44		②	0	0.001
21		③	0	0.001	45	16	①	1	0.002
22		④	0	0.001	46		②	0	0.002
23		⑤	1	0.001	47		③	0	0.001
24	9	①	0	0.002	48		⑤	0	0.002

Server as Secondary Authentication Server” といったセクションが選定されたことで，TOTP に関するチャンクが取得できた.

Step 4–1 では TOTP やリスクベース認証 (RBA) に関するチャンクに対し，関連度が “High” または “Medium” と判断されたことで，Step 4–2 で LLM は適合すると適切に判断できた.

Step 4–2. プロンプト (No.1 ②–B)

You are an expert in IoT device evaluation. Your task is to determine whether the IoT device satisfies the assessment item below, based on the provided information.

Example answers:
- Yes, the IoT device satisfies the assessment item because ...
- No, the IoT device does not satisfy the assessment item because ...

...
Assessment item:
The authentication or access control method used includes the following implementation: B) Implementation of multi-factor authentication functionality using multiple authentication factors.

Extracted relevant information:
<Relevance: High>
Document 1: #### **Enabling RSA Risk Based Authentication (RBA) Support with** [redacted]
RSA SecurID Risk-Based Authentication is a token less, multi-factor enterprise authentication solution.
...
Document 2: ##### **Configuring Admin/User Realm to Associate a TOTP Authentication Server as Secondary Authentication Server**
...
3. Check the **Enable additional authentication** server option
4. Under Additional Authentication Server, select any already created **TOTP** authentication-server from the Authentication #2 dropdown, as shown in the following figure.
...

<Relevance: Medium>
Document 3: #### **Understanding TOTP**
...
Customers can use TOTP authentication as part of their MFA policy, and strengthen their authentication mechanism for secure access scenarios.

Step 4–2. LLM の回答 (No.1 ②–B)

Yes, the IoT device satisfies the assessment item because the extracted information confirms the implementation of multi-factor authentication using TOTP as a secondary authentication factor.

Step 5 では，上記の Step 4 の回答について，テキスト分類 LLM により $(p^e,\ p^n,\ p^c) = (0.988, 0.012, 0.001)$ が得られ，式 (1) より，適合度は $c_i = 0.993$ となった.

4.3 今後の課題

実験結果より，FN (False Negative) が依然として発生しており，その主な原因を以下に示す.

まず，Step 4 における LLM による関連チャンクのフィルタリングおよび適合性判断の精度に課題が見られた．ドキュメントから適合性判断に必要な情報が適切に検索されていたにも関わらず，それらに対して関連度が “Low” と判定されたことでプロンプトに挿入されず，本来 “適合” とすべき項目を誤って “不適合” と判定するケースが確認された．また，プロンプトで関連度の高いチャンクが与えられていた場合でも，LLM が “適合” とすべき項目を誤って “不適合” と判定する事例も見受けられた．これらの原因として，プロンプトに挿入した関連チャンクの数が過剰であったことや，LLM の内部知識の不足が考えられる．今後は，関連チャンクの選別手法の最適化や判定結果を再評価するフェーズの導入が必要である.

また，“電子政府における調達のために参照すべき暗号のリスト (CRYPTREC 暗号リスト)” に関する知識を必要とする評価項目において，全デバイスに共通して FN が発生した．一部の評価項目では，CRYPTREC 暗号リストに関する知識が求められるが，使用した LLM がこれらの情

報を内部知識として保持しておらず，その結果として正確な判断が困難となるケースが見られた．また，たとえ現時点で内部知識として保持していたとしても，CRYPTREC 暗号リストは今後更新される可能性がある．したがって，今後は必要に応じてインターネット検索などにより最新の外部ソースを取得し，動的に情報を補完できる仕組みの導入が重要な課題となる．

5. おわりに

本稿では，セキュリティ適合性評価自動化手法 [7] を拡張し，従来の全体スコープ検索とドキュメントの階層構造を活用した新たなセクションスコープ検索を組み合わせて関連チャンクを検索することで，断片的な情報に基づく誤判断を軽減し，LLM 応答の精度向上を図った．3 種類のデバイスに対する評価実験の結果，全てのデバイスにおいて，提案手法は比較手法と比べて全ての評価指標で同等または上回る性能を示した．特にユーザマニュアルが 4200 ページと最大規模であったデバイスに対し，提案手法は TPR と F-measure の両方で大幅に向上しており，長大なドキュメントにおいても効果的に情報を抽出できることが確認された．今後の課題として，LLM による関連チャンクの選別および適合性判断精度の向上，ならびに CRYPTREC 暗号リストのような外部知識を補完できる仕組みの導入が挙げられる．

謝辞 本研究成果は，一部，国立研究開発法人情報通信研究機構 (NICT) の委託研究 (JPJ012368C08101) により得た．

参考文献

[1] K. Aucklah, A. Mungur, S. Armoogum, and S. Pudaruth, “The impact of internet of things on the domain name system,” in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 449–454, IEEE, 2021.

[2] ETSI, “CYBER; Cyber Security for Consumer Internet of Things: Baseline Requirements ETSI EN 303 645 V2.1.1.” `https://www.etsi.org/deliver/etsi_en/303600_303699/303645/02.01.01_60/en_303645v020101p.pdf`. Accessed on August 11, 2025.

[3] “Cybersecurity—IoT Security and Privacy—Guidelines ISO/IEC 27400:2022.” `https://www.iso.org/standard/44373.html`. Accessed on August 11, 2025.

[4] TRAFICOM, “Cybersecurity Label.” `https://tietoturvamerkki.fi/en/cybersecurity-label`. Accessed on August 11, 2025.

[5] Cyber Security Agency of Singapore (CSA), “Cybersecurity Labelling Scheme for IoT - CLS(IoT).” `https://www.csa.gov.sg/our-programmes/certification-and-labelling-schemes/cybersecurity-labelling-scheme/`. Accessed on August 11, 2025.

[6] 情報処理推進機構 (IPA), “セキュリティ要件適合評価及びラベリング制度 (JC-STAR).” `https://www.ipa.go.jp/pressrelease/2024/press20240930.html`, 2024.

[7] 池上裕香, 長谷川健人, 披田野清良, 福島和英, 橋本和夫, and 戸川望, “IoT デバイスのドキュメントに基づく LLM を用いたセキュリティ適合性スコアリング,” **電子情報通信学会研究技術報告**, 2025.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[9] 池上裕香, 長谷川健人, 披田野清良, 福島和英, 橋本和夫, and 戸川望, “多様な IoT デバイスを対象とした LLM によるセキュリティ適合性自動評価手法の検証,” **電子情報通信学会研究技術報告**, 2025.

[10] J. Jin, X. Li, G. Dong, Y. Zhang, Y. Zhu, Y. Wu, Z. Li, Q. Ye, and Z. Dou, “Hierarchical document refinement for long-context retrieval-augmented generation,” *arXiv preprint arXiv:2505.10413*, 2025.

[11] W. Tao, X. Xing, Y. Chen, L. Huang, and X. Xu, “Treerag: Unleashing the power of hierarchical storage for enhanced knowledge retrieval in long documents,” in *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 356–371, 2025.

[12] “pymupdf4llm.” `https://github.com/pymupdf/RAG/tree/main/pymupdf4llm`.

[13] “python-markdownify.” `https://github.com/matthewwithanm/python-markdownify`.

[14] LangChain, “MarkdownHeaderTextSplitter.” `https://python.langchain.com/api_reference/text_splitters/markdown/langchain_text_splitters.markdown.MarkdownHeaderTextSplitter.html`.

[15] LangChain, “RecursiveCharacterTextSplitter.” `https://python.langchain.com/api_reference/text_splitters/character/langchain_text_splitters.character.RecursiveCharacterTextSplitter.html`.

[16] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, “Query rewriting for retrieval-augmented large language models,” *arXiv preprint arXiv:2305.14283*, 2023.

[17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[18] A. Williams, N. Nangia, and S. R. Bowman, “A broad-coverage challenge corpus for sentence understanding through inference,” *arXiv preprint arXiv:1704.05426*, 2017.

[19] “LangChain.” `https://www.langchain.com/`.

[20] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, “The Faiss library,” 2024.

[21] “sentence-transformers/stsb-xlm-r-multilingual.” `https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual`.

[22] “Qwen/Qwen3-8B.” `https://huggingface.co/Qwen/Qwen3-8B`.

[23] “microsoft/deberta-v2-xlarge-mnli.” `https://huggingface.co/microsoft/deberta-v2-xlarge-mnli`.

[24] 情報処理推進機構 (IPA), “セキュリティ要件適合評価及びラベリング制度 (JC-STAR) ☆ 1 (レベル 1) チェックリスト (2025.05.05) 版.” `https://www.ipa.go.jp/security/jc-star/tekigou-kizyun-guide/label1/index.html`, 2025.