セキュリティオペレーションセンターにおける分析レポート評価のための大規模言語モデル活用手法の提案

岡田 裕幸[1,a)]　大庭 達海[1]　矢内 直人[1]

概要：セキュリティオペレーションセンター（SOC）では，インシデントに関する有用な分析レポートの作成支援のため，大規模言語モデル（LLM）の活用が期待される．しかし，LLM によるレポート作成支援の中でも分析官が作成したレポートの評価は，熟練分析官による評価と必ずしも一致しない点に課題が残る．本研究では，まず分析レポートに関するユーザ調査を通じて，LLM ベースの評価手法の課題を明らかにする．ユーザ調査で得た知見は二つである．第一に，従来の分析レポートに関するツールには限界があり，レポート評価のような高度な作業は代替できない．第二に，レポート評価には SOC における判断に必要な文脈や専門知識が不可欠である．次に, 上述した結果を踏まえ, 分析レポートの評価精度を向上させる LLM 活用手法を提案する．本手法の有効性を実験で確認したところ，従来評価手法を上回ることも示す．

Analysis Report Evaluation Framework Using Large Language Models for Security Operation Centers

Hiroyuki Okada[1,a)]　Tatsumi Oba[1]　Naoto Yanai[1]

Abstract: Security Operation Centers (SOCs) aim to leverage Large Language Models (LLMs) to support the creation of valuable analytical reports on security incidents. However, evaluations performed by LLMs do not always align with those conducted by experienced analysts, raising concerns regarding the reliability and validity of LLM-based assessments. In this study, we first conduct a user study on analytical report evaluation to identify the limitations of current LLM-based approaches. We then found two insights. First, current automation tools are insufficient for evaluating analysis reports. Second, the evaluation of analysis reports should be based on sufficient context and expert knowledge. We then propose a novel conceptual framework utilizing LLMs and develop a new method to enhance the accuracy of report evaluation. Comprehensive experiments demonstrate that our approach outperforms existing methods.

1. はじめに

サイバー攻撃の脅威が高まる昨今，セキュリティアラートを分析し，適切な対応を行うセキュリティオペレーションセンター（Security Operation Center，SOC）の設置は，各組織の喫緊かつ重要な課題となっている．SOC における重要な任務は，セキュリティアラートの分析結果の提示はもちろん，その脅威の背景にある文脈を明確かつ実践的に示す分析レポートの作成である．分析レポートとは，検知されたアラートやインシデントに関する事実・原因・影響・対応策などを整理し，関係者間で共有・意思決定に活用するための文書であり，SOC 内部や顧客との対応調整や報告に用いられる [1]．高品質な分析レポートは，非専門家を含む読者にアラート内容を理解させ，自身の業務と結び付けて捉えることを可能にする．これにより状況を自分事として認識しやすくなり，行動意欲が高まり，組織全体のセキュリティ意識向上につながる [2]．

一般に，高品質な分析レポートを継続的に提供するには，レポート自体を適切かつ自動的に評価する仕組みが重要となる [3]．この際，記述の不足を指摘するだけでなく，「レポートから導かれる対応が実行可能か」など専門家の観点も含めた評価が求められる [4]．しかし，SOC における分

[1] パナソニック ホールディングス株式会社
Panasonic Holdings Corporation
[a)] okada.hiroyuki001@jp.panasonic.com

析レポートの評価には専門的な知見が必要であり，評価の質は分析官の能力や経験に大きく依存する [5].

一方、サイバーセキュリティ分野において作業負荷を軽減する観点から，サイバーセキュリティ分野において大規模言語モデル（Large Language Model，LLM）の活用が注目されている [6]．しかしながら，SOC におけるセキュリティアラートの分析レポート評価には依然として課題が存在する．前述の通り，分析レポートの評価には専門家からの観点が必要だが，高品質な出力を生成するシステムの内容を適切に評価することが難しい [7]．たとえ標準的な基準を用いる LLM ベースの評価方法であっても，専門家の立場からレポートを評価することはしばしば困難である [8].

本研究では，SOC におけるセキュリティアラートの分析レポートを LLM によって評価することで，レポートの品質向上を目指す．具体的には，以下の研究課題に取り組む：

RQ1 SOC での分析レポート自動化の現状と期待は何か？

RQ2 SOC でのレポート品質の観点と評価基準とは何か？

RQ3 LLM は SOC の分析官と同様の観点からレポートを評価可能か？

これらの問いに答えるために，まず本研究は SOC の分析官を対象としたアンケートおよびインタビューによるユーザ調査を実施し，現在行われている分析レポートの評価方法や評価観点を調査した．その結果，SOC の運用においてフィードバックの不十分さや評価基準の曖昧さなど，複数の課題が明らかになった.

ユーザ調査の知見をもとに，多視点に基づくセキュリティ分析評価支援手法も提案する．提案手法は，LLM によるレポート評価に専門領域の知見を多視点的に組み込むことで，熟練分析官の認知に近い判断を可能にする．提案手法を用いた実験を通じて，既存手法 [7,9] と比べて本手法がより熟練分析官に近い評価性能を有することを確認した.

本研究の貢献は以下の通りである：

- SOC の分析レポートに関するユーザ調査を実施し，既存ツールの限界と LLM 評価の課題を明らかにした.
- レポート評価には，SOC における判断に必要な文脈と専門知識、及び、明確なフィードバックの出力が必要なことを確認した.
- 新たな SOC 分析レポート評価手法を提案し，既存手法よりも熟練分析官に近い評価が可能なことを示した.

以下は本稿の構成である． 2 節では技術的背景と問題設定を示す． 3-4 節ではそれぞれユーザ調査の設定と結果を示す． 5 節では提案手法の詳細を， 6 節ではその評価結果を示し， 7 節で本研究の結論を述べる.

2. 背景と関連研究

本節では，LLM と SOC の概要について，それらの関連研究を含めて述べる．また，本研究の問題設定として，SOC における分析レポートの評価に関する課題を整理する.

2.1 大規模言語モデル

自然言語処理とデータ生成を組み合わせた機械学習モデルである LLM は各分野で利用され，その精度は熟練のデータ分析官に匹敵し得る [10]．この性能を応用し，LLM によるデータ評価の取り組みも進んでおり，GPTScore や G-Eval [7,9] 等の手法が注目されている．LLM はサイバーセキュリティ研究でも多様な応用が進んでいる．具体的には，データ処理，攻撃検知，レポート生成等がある [6]．本研究ではこれらの知見を踏まえ，LLM を用いたサイバーセキュリティ分析レポートの評価手法に焦点を当てる.

2.2 セキュリティオペレーションセンター

セキュリティオペレーションセンター（SOC）はサイバー攻撃に関わる様々な活動を担う．主な役割は，サイバー攻撃の検知，セキュリティアラートのトリアージと調査，アラートに対する最終的な判断である．このとき，SOC の業務は 1）分析官による評価，2）管理者による判断の二つに分類できる．この二つをつなぎ相互に支援する意味で，分析レポートの作成は SOC では重要である [5].

近年の SOC は業務負荷の軽減のために自動化ツールの導入が進む一方，依然として分析官の負荷は高い [11]．上述した背景から，LLM を含む AI 技術による SOC 業務の支援が検討されているが，現状では LLM による分析レポートの作成や分析に掛かる助言は限界がある [12].

2.3 SOC における分析レポートの評価と問題設定

本研究では，SOC における分析レポートの評価を主要な問題と設定する．実際，現在利用されているレポート生成ツールにはいくつかの限界がある．例えば，LLM は必ずしも高品質な分析レポートを生成できるわけではなく [12]，また，レポートやそれに紐づくセキュリティアラートに関するデータの内容自体にも制約がある [5]．本研究では，SOC におけるセキュリティアラートに対する分析レポートの生成に向けた初期検討として，LLM がこれらのレポートを効果的に評価できるか明らかにする．分析レポートの評価は依然として困難である．なぜなら，これらのレポートはドメイン固有の文脈を含むことが多く，汎用的な LLM では人間の判断と整合しない場合があるからである [13].

本研究の問題設定は，分析レポートを入力とし，評価結果を出力とするシステムの設計である．出力には，定量的な評価結果とともに，熟練分析官の視点に基づく定性的な観察結果（例えば，実行可能性のあるフィードバック）が含まれることを想定する．なお，分析レポートの読者は管理者や最終提出先の顧客，両方を想定し，レポートの生成自体は本研究の対象外である.

3. ユーザ調査方法

本節では，実際の SOC に所属する分析官を対象とした

表 1　参加者情報.

ID	専門性と経験	職務	対象環境
1	高：7 年以上	SOC 管理者，アーキテクト	工場，ビル
2	高：3〜5 年	エンジニア，上級分析官	ビル，家電
3	高：7 年以上	SOC 管理者，アーキテクト	IT，工場
4	高：5〜7 年	上級分析官	IT
5	中：3〜5 年	エンジニア，上級分析官	工場，家電

ユーザ調査の実施に関する調査目的や各種設定を説明する．

3.1　調査目的と設定

本調査の目的は，RQ1 および RQ2 への答えを得ることにある．具体的には，分析レポートの品質評価に関する評価基準を含め，レポート作成ツールの現状を明らかにする．そのために，SOC において分析官が分析レポートを作成する際に直面する課題について調査を行う．

本調査の全体像としては，SOC の現場における実際の声を取り入れるために量的調査と質的調査を織り交ぜた混合研究法を用いる [14]．具体的には，事前ヒアリングによる調査を実施し，その内容に基づいたアンケートによる調査を行い，その内容について必要に応じてフォローアップインタビューを実施した．本調査の各工程は，SOC の業務に関するユーザ調査 [2, 5, 12, 15] を参考に実施している．以降では各調査設定について説明する．

3.2　参加者

情報系の SOC に所属する 5 名の分析官に本調査への参加を依頼した．いずれも著者の所属機関と関係のあるグループ企業，またはセキュリティ関連のパートナー企業から協力を得ており，男女の配分を含む多様な経歴の参加者で構成される（詳細は表 1 参照）．表において, 専門性の欄では経験年数が 3 年以上であり, かつ, CISSP 等の国際資格または同等の資格を保有する参加者を「高」と分類した．これらの参加者は，分析レポートの評価に関する判断が可能な熟練分析官を中心に選出された．参加者には報酬は支払っていない一方, 研究目的，データ処理内容と匿名化等に関する情報を提供した上で参加への同意を得た．以降で説明する調査方法では，同一の参加者に対していずれの調査も実施した．これは，結果の解釈における一貫性を確保することで，それぞれの結果を相互に補足するためである．

3.3　調査方法

本調査では，参加者の実務経験や背景に根差した具体的な課題を的確に抽出するため，まず事前ヒアリングを実施し，その結果からアンケート項目を設計した．選択式と自由記述式質問を交えたアンケートを実施した後，その結果を解釈するため，30 分程度のフォローアップインタビューを実施し，各参加者の結果の背景にある理由や具体事例を深掘りした．この調査方法は，幅広い傾向の把握と詳細な背景理解の両立を可能にする方法として知られる [14].

3.3.1　事前ヒアリング

事前ヒアリングでは，後述のアンケート設計に向けて, 半構造化インタビューによる探索的定性調査を実施した．アンケートの質問項目は著者を含む SOC に関連する専門家 3 名による試験調査を通じて策定し，「日常的な分析レポート作成プロセス」「作成時の課題」「評価・改善」に関する質問を含む想定である．これらの質問は，アラートの分析レポートに関する記述を含む先行研究 [2, 5, 12, 15] も参考にしているが，先行研究はレポート作成・評価に関する SOC 現場の課題把握には不十分であった．このため, 事前ヒアリングではアンケート設計の精度を高めることを目的として，先行研究を補完し現場の実態を直接確認する．

ヒアリングでは，SOC 内外のやりとりも含め具体的事例を収集した．ヒアリングは質問の偏りを軽減するため，著者含む専門家 2 名により実施をした．ヒアリングは，各 SOC 環境での詳細な意見収集と，相互影響による回答バイアスの抑制を目的として，3 名以内の 2 グループで実施した．グループごとに 1 回あたり 1 時間，計 2 回ヒアリングを実施した．記録には参加者の同意を得た上で，Microsoft Teams を用い，転記作業の負担軽減のため録音および自動書き起こしと手動の修正を行った．

事前ヒアリングのデータに対しては，テンプレート分析（TA） [16] の枠組みを参考に，事前に想定されたカテゴリとデータ駆動型の発見を組み合わせた柔軟性を持つテーマ別分析を実施した．SOC に関する既存研究では，質的データ分析に TA を適用した事例が報告されており [15], 本研究の対象である SOC 分析レポートの自動化のような収集できるサンプル数に限りがある場合や，研究蓄積があまり多くない場合に有用とされる [15]．まず，先述の先行研究 [2, 5, 12, 15] について，分析レポート作成・評価に関する主要テーマを整理した．これらのテーマ（例：レポート作成プロセス，作成時の主な課題，評価および改善の取り組み等）を初期テンプレートとして設定した．次に，事前ヒアリングの記録を精読し，初期テンプレートを用いて著者が第 1 回コーディングを実施した．この際，LLM を補助的に用いてコード候補や類似表現の整理を行い，網羅性と一貫性の確保に努めた．得られた結果は，SOC 業務の知見を有する 2 名の専門家による確認を受けた．それらの反応を加味し，第 2 回コーディングを行い，冗長性や表現の曖昧さを排除した最終的なコードブックを確定した．最後にそれらの結果をユーザ調査に経験のある専門家により確認してもらい，結果の妥当性を補強した．

3.3.2　アンケート

本アンケートの作成は，先行研究 [17] に従い，参加者の負担を軽減するため，30 分〜1 時間以内で完了できる設計とした．事前ヒアリングで得られた全テーマのうち，研究目的と直接関連し，かつ既存研究で十分に検証されてい

ないテーマを中心に選定し，不要な重複や低関連の項目は除外した．選定したテーマは関連する先行研究で指摘された課題と照合し，先行例では解決できていない点であることを整理し，再度カテゴリ修正を行った上で質問項目を作成した．作成した内容は，著者を含む SOC 業務の知見を有する専門家 3 名による確認を経て確定した．アンケート項目には，3 段階のリッカート尺度による選択式設問と，自由記述欄による意見記入の両方を含めた．アンケート内容に関するやり取り（参加者からの内容確認，提出等）は Microsoft Teams を通して行った．

3.3.3 フォローアップインタビュー

アンケート結果の背後にある理由や具体事例，または曖昧な記述内容の真意を明らかにするため，各参加者に対し 30 分間のフォローアップインタビューを実施した．実施は事前ヒアリングと同様の方法で音声記録および逐語化を行った．分析は，事前ヒアリングで作成したコードブックを初期テンプレートとして，アンケート自由記述およびフォローアップインタビューの記録に適用した．既存のカテゴリ構造に基づき一貫性を保ちつつ，新たに浮上した課題や視点は必要に応じてカテゴリを修正・追加した．本分析は事前ヒアリングと同様の体制で実施し，著者による 2 回のコーディングと SOC 業務経験を有する専門家 2 名の結果に対する反応を踏まえて結果を確定した．

4. ユーザ調査結果

本節では各調査の結果を示したのち，それらの全分析結果を統合し，RQ1 および RQ2 に対する回答を導く．

4.1 事前ヒアリング結果

事前に設定した主要質問を基盤としてインタビュー結果を分析したところ，7 つのテーマが抽出された．その中で今回の RQ に関連する SOC アナリストの業務実態と課題を説明する以下の 5 つのテーマを選定した．以下にテーマと参加者コメント例を記載する．1）ツール活用の限界：**「定型的処理向けの自動化ツールは運用しているが，より分析業務を効率化できるツールが望ましい．」**2）レポート作成への AI 活用の課題・リスク：**「*AI* による文章生成への期待はあるが，誤情報混入への懸念が強く，現状は導入に慎重．」**3）分析レポート作成の粒度・深度：**「作成者やインシデント内容等によってレポート内容にばらつきがある．」**4）品質保証・多段レビューの負荷：**「品質担保のための多段階レビューは作業負荷が高い．」**5）レポート作成の運用面での整備：**「フォーマットや記載手順はあるが，詳細は人の能力依存．」**

これらのテーマを，先行研究で指摘された SOC レポート作成・評価の課題と照合し，次の 3 カテゴリに整理した（対応関係は括弧内に示す）．カテゴリ 1「分析レポート作成に用いられているツールの現状と課題」（テーマ 1,2)）：既存ツールの機能的限界や AI 活用時のリスクを指摘した先行研究 [2] と対応するが，先行研究では具体的な運用実態や改善要求までは十分に把握されていないため，その点を調査するためのカテゴリとなる．カテゴリ 2「分析レポートの構成および SOC における運用実態」（テーマ 3,5)）：明確かつ構造化されたレポート定義の必要性を示す先行研究 [2,12] に関連するが，アラート分析レポートを想定した現場レベルの運用課題等は十分に検証されていないため，その点を調査するためのカテゴリとなる．カテゴリ 3「分析レポートの評価基準とその課題」（テーマ 4,5)）：評価基準の一貫性欠如やレビュー負荷を指摘する先行研究 [2,5,15] と対応するが，評価実務での具体的影響や改善策は調査されず未解明である．この整理を基にアンケートを設計した．

4.2 アンケート結果

選択形式の設問結果を表 2 に示す．各設問は「はい」「中立」「いいえ」の 3 段階のリッカート尺度で構成される．表によれば，参加者の大多数は，現行の分析レポート作成プロセスに何らかの課題があることを認識している．

まず，多くの SOC 分析官は自動化ツールを利用しているが，深い分析や判断を要する業務の効率化には限界がある（1–2 行目）．また，LLM 活用への期待はあるものの，精度面の課題から実務利用は困難との意見が示された（3 行目）．さらに，レポート構成やテンプレートは整備されている一方で，フィードバック体制や品質評価基準は不十分である（4–6 行目）．これらから，レポート作成・評価は依然として分析官の経験や知識に依存しており，品質のばらつきが生じていることが分かる．自由記述欄の内容は，以降のフォローアップインタビューの節で述べることとする．

4.3 フォローアップインタビュー結果

自由記述欄とそのフォローアップインタビューに基づくテーマ別分析から，SOC における分析レポート作成に関して以下の 3 つの主要な困難が明らかになった：1）既存ツールによる手作業の代替に限界があること，2）レポート評価基準の曖昧さと共通認識が欠如していること，3）分析レポートの品質が特定個人に依存していること．以下に、参加者のコメントを交え、それぞれの詳細を述べる．

既存ツールによる手作業の代替の限界：表 2 の 1-2 行目に示されている通り，多くの SOC 分析官は分析レポート作成に自動化ツールを利用している一方で，それらに対する不満や限界も存在している．とくに，深い分析や判断を伴う作業において，既存の自動化ツールは人的労力を大きく削減するには不十分である．いくつかの参加者は，月間の通信量や認証失敗回数のような定量的な出力を伴うレポート作成に関しては自動化が進んでいると述べている [5]．**「アラート件数や *KPI* などは，チケッティングシステムを通じて自動生成される．」**—**ユーザ** *ID3*

表 2 レポート作成状況および課題に関する調査結果.

#	質問項目	選択肢	回答数
1	レポート作成を効率化するために使用しているツールや自動化システムはありますか？	はい／中立／いいえ	5 ／ 0 ／ 0
2	レポート作成時に使用しているツールに不満や課題はありますか？	はい／中立／いいえ	5 ／ 0 ／ 0
3	レポート作成に LLM を活用することに高い期待を持っていますか？	はい／中立／いいえ	2 ／ 3 ／ 0
4	共通のレポート形式やテンプレートや作成手順を整備していますか？	はい／中立／いいえ	5 ／ 0 ／ 0
5	レポート内容について（顧客や熟練分析官などから）十分なフィードバックがありますか？	はい／中立／いいえ	2 ／ 2 ／ 1
6	レポートの品質を測定するための評価基準を確立していますか？	はい／中立／いいえ	0 ／ 1 ／ 4

また，いくつかの SOC では，LLM を活用した要約処理も試験的に導入されている．**「*IDS* から出力されるアラートを要約する目的で *LLM* を利用中.」—ユーザ** *ID4*

しかし，参加者はこれらのツールが有効なのは統計情報作成や要約までであり，「調査の方向性の示唆」など，いわゆる解釈的な作業には依然として多大な人的負荷がかかっていると指摘する．例えば，LLM を用いた分析に対して期待を寄せたが，精度の不足から実務利用は難しいという声もあった（表 2 の 3 行目参照）．**「*LLM* を分析目的で評価したが，正確性が低く本格運用には早い.」—ユーザ** *ID2*

さらに，フォレンジックレポートのように高い正確性が求められる領域では，LLM は熟練分析官の代替にはなり得ないという点も指摘されている [12]．このように，自動化は構造化された情報の処理には有効だが，分析的思考や文脈理解といった高度な判断には，依然として分析官の専門性が不可欠である．参加者は，文脈に即した判断を支援するようなツールの実現を期待しているが，現時点ではそのような支援は不十分である．

レポート評価基準の曖昧さと共通認識の欠如: 表 2 の 6 行目の内容の通り，SOC 内部では分析レポートの評価基準が統一されていない．その結果，SOC 管理者ごとに判断が異なり，分析官から見た際の評価に一貫性がなく，公平性の欠如につながっている．このような状況では，分析レポートの継続的な品質向上も難しい [5]．したがって，明確な評価基準の整備と，それに基づいた評価プロセスの確立が必要である．SOC 管理者が状況を理解し判断するためには，レポート内に文脈が明確に記載されている必要があるという指摘もあった．**「レポートには，問題点，状況，必要なアクションが明記されているべきで，顧客が何をすべきかをすぐ理解できる形が望ましい.」—ユーザ** *ID1*

また，SOC 管理者の意思決定には，リスクや負荷など実務的観点の提示が有用であるとの意見も得られた．**「影響，リスク緩和，予算，作業量といった実行可能なインサイトがあると意思決定しやすい.」—ユーザ** *ID3*

一方で，レポートに十分な文脈が含まれていないと，品質が著しく低下する [2,5]．しかし，文脈情報の記載は作成者依存であり品質が安定しない．**「文脈は判断に不可欠だが，作成のための情報収集・分析等のスキルに依存されるだけでなく，環境・発生事象等にも影響される．必要な情報が無いレポートは『だから何？』となる.」—ユーザ** *ID3*

同様に，文脈の有無や適切さを客観的に評価すること自体が難しく，判断はレビューアに大きく依存する傾向がある [11]．**「必要な情報の記述漏れのチェックは重要だが，レビューアの専門知識に依存する.」—ユーザ** *ID1*

したがって，分析官の経験と知識にレポート評価を依存しすぎる状況を是正するためには，評価基準とその視点を明示し，体系的な評価枠組みを導入する必要がある．特に，過去事例や専門知識に基づいて評価基準を言語化するための LLM の活用は有望である．

分析レポートの品質が特定個人に依存: レポート作成・評価は，各分析官の経験や知識に大きく依存しており，レポートの品質にもばらつきがある．実際，レポートの構成やテンプレートは整備されているものの，レビューアの業務負荷などにより十分なフィードバックが得られないケースもある（表 2 の 4,5 行目参照）．以下のコメントは，インシデントの内容によって分析の深度や対応内容が変わることを示している．**「分析結果の記載の詳細度はインシデントによって異なる．例えば，ランサムウェアと単純なワーム感染では，分析の深さや対策も変わる.」—ユーザ** *ID3*

また，レポートの形式面は整備されているが具体的な内容の記載は困難である．**「記載手順やフォーマットが整備されても，何を書けばいいかわからないことがある．過去のレポート等から検討することも必要.」—ユーザ** *ID4*

このように，形式の整備だけではレポート品質を十分に担保できず，内容そのものを支援・評価できる仕組みが求められている．**「過度な形式化は異なるインシデントを同一視する等のリスクがある．*LLM* 等でインシデントに応じて動的に記述の補助をして欲しい.」—ユーザ** *ID2*

また，熟練分析官による評価やフィードバック体制は存在するものの，日々の業務負荷により効果的に機能していないとの指摘もあった．**「レビューは重要だが，日常業務が忙しくて適切に対応できないことがある.」—ユーザ** *ID1*

このような観点は，SOC 管理者と分析官でレポート品質の認識が異なることを示唆する先行研究とも関連する [5]．また，サイバーセキュリティ人材育成においても [18]，SOC 管理者による組織的観点に加え，分析官個人の成長支援の観点での適切なフィードバック体制が重要である．

まとめ: 以上の結果から，レポート自動評価は現状では

困難であるが，この実現に向けて取り組むための以下の 3 つの主要な課題が明らかとなった．1）明確な評価基準の欠如，2）分析官の知識・経験への依存とフィードバックの不足，3）熟練分析官と現行ツールとの間に存在するギャップ．つまり，単なるツール導入だけではレポートの品質向上には不十分であり，レポート内容を適切に評価するための仕組みの確立が必要不可欠である．

4.4 RQ1,2 への回答

RQ1：現行の自動化ツールは一部タスクには有効だが，分析レポート作成や評価といった高度な作業は依然として手作業が必要である．ユーザは LLM による完全自動化の可否に関心を示す一方，レポート作成支援やフィードバック活用にも強い期待を寄せていることが分かった．

RQ2: 高品質な分析レポートには十分な文脈と実行可能な示唆が必要だが，明確かつ共通の評価基準がなく，文脈が曖昧になりやすい．品質や評価は専門家個人の能力に依存しており，明確な基準も存在しないため，現行の LLM 等のツールでは適切に評価するのが困難だと考えられる．

5. 提案手法

本節では前節で述べた課題に対する提案手法を述べる．

5.1 主な着想

レポート評価の自動化に向けた 3 つの主要な課題を解決するための主な着想を以下に示す．

第 1 の課題である明確な評価基準の欠如に対し，本研究は熟練分析官の視点を反映した評価用チェックリストを導入する．これは提案手法のプロンプトとしてだけでなく，SOC における評価タスクにも応用できる．

第 2 の課題である分析官依存とフィードバック不足については，従来型のチェックリストを利用した LLM 評価の限界が指摘されている [13]．本研究はこれに対応するため，分析官の知識やフィードバックを反映できる詳細化ガイドラインを提案する．これにより，LLM は熟練分析官の視点に沿って文脈を理解し，より深い評価が可能となる．

第 3 の課題である熟練分析官と既存ツールのギャップには，多視点評価 LLM で対応する．本提案手法は，LLM による表層的な評価とガイドラインに基づく詳細な評価とを統合し，多視点的な評価を実現する．この構造は人間の認知過程に対応し，表層情報で知識を活性化し，次に文脈を統合して理解する [19]．多視点評価 LLM はこの過程を模倣することで熟練分析官に近いレポート評価を実現する．

5.2 手法の詳細

以降では図 1 に示す提案手法の主な 3 つの技術を述べる．

5.2.1 評価用チェックリスト

評価用チェックリストは，熟練分析官の視点で分析レポートを評価する項目群である．4 節で述べた通り，十分な文脈を含むかの評価は重要かつ困難であり，本研究ではこれに対応するため新たに設計した．

チェックリストの作成は NIST SP800-61 [1] など 9 つのガイドラインと参加者の SOC 実務者 2 名の意見を基に，分析レポートに含まれるべき要素（例：「インシデントの状況把握・対応に必要な情報」等）の抽出，初期案作成，要否評価と理由付け，多数決による整理，最終レビューという多段階プロセスで実施した．最終的に，11 個のカテゴリにまたがる約 50 項目から成るチェックリストが得られた．誌面上，チェックリストの一例のみ以下に示す．「チェックリストの項目タイプ：基本情報」は，レポート内にアラートの内容，対象機器等の基本情報が記載されているか確認する．一方「タイプ：影響評価」は，発生事象による現場への影響の具体性，根拠の妥当性等を確認する．これら各チェックリストの項目は後述の詳細化ガイドラインと紐づけられる．このチェックリストは，LLM へのプロンプトはもちろん SOC 業務での人手評価にも応用できる．

5.2.2 詳細化ガイドライン

詳細化ガイドラインは評価用チェックリストと対応し，各観点の評価深度を確保するために設計した．LLM に「何を重視し，どういった手順で評価するか」などの指針をカテゴリ別に示し，チェック項目を詳細化することを目的とする．例えばカテゴリ「仮説の検証」では，仮説に対する証拠の適切性や文脈情報の十分性を評価させる．一方でカテゴリ「パターンおよび比較分析」では，インシデントが過去の事例や既知のパターンと比較されているかを評価させる．5 つのカテゴリが設定され，それぞれがチェックリストの各項目と紐づいている．

LLM は本ガイドラインをプロンプトとして各チェック項目を深く理解し，重要記述を特定・評価する．その観点は「項目の目的」「確認すべき情報の種類と箇所」「評価方法」等である．設計手順は，1）SOC の分析プロセスに関する先行研究 [11, 20, 21] や現場プロセスを参考に初期設計，2）10 件の SOC レポートを LLM（GPT-4o）で評価した結果と著者含む SOC 関係者 5 名の評価結果の差異を分析・修正，3）カテゴリやプロンプトを調整し著者含む SOC 関係者 3 名でレビュー，の 3 段階で行った．詳細化ガイドラインにより，LLM は単なる記述の有無確認にとどまらず，文脈の整合性や合理性まで評価可能となる．

5.2.3 LLM による多視点評価

多視点評価 LLM は，LLM による「表層的評価」と「詳細評価」を統合し，熟練分析官に近い評価を実現する手法である．以下に各評価の概要と統合方法を示す．

表層的評価 LLM: 表層的評価では，LLM が単純な情報の有無や文法の正確さなど，形式的な要素に基づいて分析レポートを評価する．この種の評価においては，LLM の出力と人間による評価との相関が高いことが知られているお

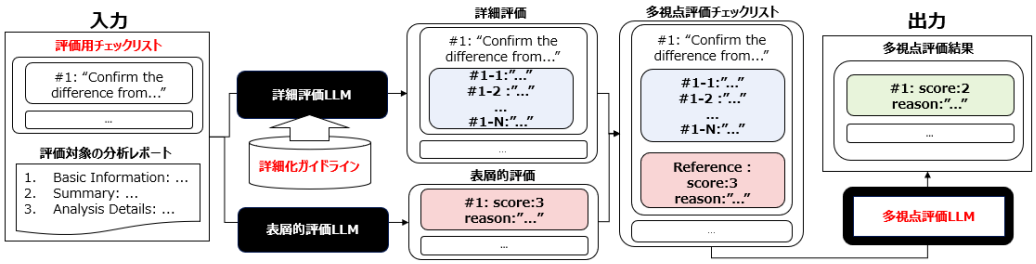

図 1 提案手法の概要.

り，専門知識を適用することでかえって評価品質が低下する可能性がある [22]．ここで用いるプロンプトは先行例 [9] を参考に，1）対象レポート，2）評価用チェックリスト項目，3）評価タスク設定の 3 要素から構成される．評価には 5 段階のリッカート尺度を用い，人間に近い定性的分析と定量的スコアリングの信頼性を両立させる [13, 22]．

詳細評価 LLM: 詳細評価では，LLM にチェックリスト各項目の重要記述を理解させつつ評価させるプロンプトを設計し，レポート文脈を踏まえた実質的な内容を評価する．これにより，文脈情報を活かした厳密・妥当な評価と，評価過程の解釈可能性向上を実現できる [7, 8]．詳細評価で用いるプロンプトは以下の 3 要素で構成される：1）評価対象の分析レポート，2）評価用チェックリスト，3）詳細化ガイドライン．LLM はこれらの情報をもとに，レポートの内容に即した詳細化された評価結果を生成する．

多視点評価 LLM への統合: 多視点評価では，表層的評価と詳細評価を統合し、熟練分析官に近い立場から分析レポートを評価する．異なる特性を持つ両者の出力を活用することで，信頼性と解釈可能性を高め，相互参照により片方では見落としやすい情報や偏りを軽減できる．プロンプトは，1）評価対象の分析レポート，2）表層的・詳細評価結果，3）多視点評価方法の構成である．「多視点評価方法」では，表層的評価のスコアと理由に加え，詳細評価で得られた各項目のスコアと理由を統合して再検討する．そのうえで，各判断理由の重要度を考慮し，最終的な 5 段階リッカート尺度でスコアをその理由と共に出力する．

6. 実験

本節では提案手法の評価のための実験設定と結果を示す．

6.1 実験設定

以降で実験に使用したデータセット，評価指標，ベースライン手法について述べる．

データセット：本研究では，工場・ビル・IT インフラを監視する複数の実環境 SOC による分析レポートから成る非公開データセットを用いた．各 SOC は運用体制が大きく異なり，詳細は研究倫理上非公開である．本データセットは計 20 件の分析レポートで構成されている（工場：5，ビル：10，IT：5，期間：2022 年 4 月-2024 年 4 月）．レポート 1 件あたりの平均文字数は約 2,200 文字であり，誤検知に関する内容である．機密性と倫理に配慮し，IP アドレスや固有名詞等の機微情報を匿名化した．各レポートは次のような要素を含んでいる：日付等の基本情報，アラート概要，詳細分析結果，現場への影響情報等．手法評価用の正解データは，著者と 3 節の参加者を含む 5 名がチェックリスト項目を 5 段階で評価し，平均を最終スコアとした．項目数は 1 レポートあたり 10〜15 に限定し，文脈的に重要な内容に絞ることで，負担を抑えつつ精度を維持した．

評価指標：本研究では，5 段階リッカート尺度を用いた人手評価スコアとモデル出力スコアとの相関を主要な評価指標とする [7–9]．性能評価は全レポートのチェックリスト項目を統合し，約 260 項目のスコアから以下の指標を算出した：スピアマン相関（ρ），ケンドール相関（τ），ピアソン相関（r），RMSE，Jensen–Shannon ダイバージェンス（JSD）．また前述の人による評価スコアと各手法の評価スコアの分布をバイオリンプロットで示す．これにより人に近い形状の分布ほど性能が高いと直感的に比較できる．

ベースライン：本研究では，提案手法に加え，以下の 4 つのベースライン手法を実装した．すべての手法は，OpenAI の GPT-4o モデル を用いて実装しており，温度と top-p パラメータは 0 に設定している．各手法の結果は，10 回の実行で得られた最終スコアの平均値として集計している．

手法 1：表層的評価のみ．本手法は 5.2.3 節で示した表層的評価を行う．これは GPTScore [9] に準拠するプロンプト設計に基づいた，単純な LLM による評価となっている．

手法 2：詳細評価（w/o 詳細化ガイドライン）のみ．本手法は 5.2.3 節で提示した詳細評価を行うが，詳細化ガイドラインは用いない．代わりにタスク説明と評価基準に基づき，LLM が CoT（AutoCoT）で段階的推論を行う形式であり，G-Eval [7] に着想を得ている．

手法 3：詳細評価（w/詳細化ガイドライン）のみ．本手法は，5.2.3 節で提示した詳細評価を行う．手法 2 に比べて，より制御された評価が可能であるが，表層的評価は実施せず，多視点評価も行わない．

手法 4：表層的評価と詳細評価（w/詳細化ガイドライン）による多視点評価．本手法では，手法 1 および手法 2 を統合し，最終スコアを出力する．提案する多視点評価の構成の一部を模倣したものである．

表 3　各手法の比較.

Name	ρ	τ	r	RMSE	JSD
手法 1 [9]	0.68	0.57	0.67	0.94	0.26
手法 2 [7]	0.59	0.46	0.60	1.30	0.47
手法 3	0.66	0.52	0.67	0.93	0.26
手法 4	0.69	0.56	0.69	0.94	0.26
提案手法	**0.71**	**0.57**	**0.70**	**0.89**	**0.24**

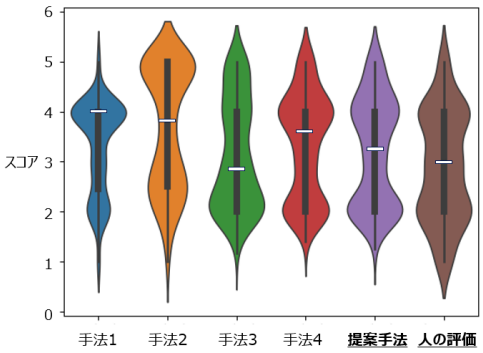

図 2　各手法と人による評価スコアのバイオリンプロット.

6.2　結果

実験結果を表 3 と図 2 に示す．表から明らかなように，提案手法は，すべての評価指標において他の手法を上回る性能を示している．さらに，図 2 では，各手法および人手評価に対するスコア分布をバイオリンプロットで可視化しており，提案手法のスコア分布が人手による評価と最も近い形状を示していることが確認できる．

6.3　RQ3 への回答

提案手法が全ベースラインを上回ったことは，以下の 2 つの要素の有効性を示唆している：1）評価用チェックリストと詳細化ガイドラインの活用．2）表層的評価と詳細評価を組み合わせた多視点評価．この結果より，提案手法は熟練分析官に近い出力を実現した．

制限：1）非公開データセットだけでなく公開データでの性能検証が必要，2）レポート毎のチェック項目自動選択機構が必要，3）長文・多様なレポートへの適用検証が必要，4）分析官に有益なフィードバックの妥当性確認が必要．

7.　結論

本研究では，まずユーザ調査を通じて現状の分析レポート評価の課題を明らかにし，得られた 2 つの知見を基に評価手法を提案した：1）現行の自動化ツールはレポート評価に不十分である，2）十分な文脈が評価に不可欠である．提案手法は「評価用チェックリスト」「詳細化ガイドライン」「多視点評価 LLM」の三要素から構成される．実験の結果，提案手法は既存手法を上回る性能を示し，熟練分析官と同等の性能を発揮し得る．今後は SOC において提案手法が効果的に機能するかを実証的に検証する必要がある．

参考文献

[1] K. A. Scarfone *et al.*, “Sp 800-61 rev. 1. computer security incident handling guide,” 2008.

[2] A. Jawad *et al.*, “” i’m getting information that i can act on now”: Exploring the level of actionable information in tool-generated threat reports,” in *Proc. of EuroUSEC 2024*.

[3] B. B. Klebanov *et al.*, “Automated evaluation of writing–50 years and counting,” in *Proc. of ACL 2020*, 2020.

[4] J. J. Ryan *et al.*, “Quantifying information security risks using expert judgment elicitation,” *Computers & Operations Research*, 2012.

[5] F. B. Kokulu *et al.*, “Matched and mismatched socs: A qualitative study on security operations center issues,” in *Proc. of CCS 2019*.

[6] Y. Chen *et al.*, “A survey of large language models for cyber threat detection,” *Computers & Security*, 2024.

[7] Y. Liu *et al.*, “G-eval: NLG evaluation using gpt-4 with better human alignment,” in *Proc. of EMNLP 2023*.

[8] Y. Lee *et al.*, “Checkeval: Robust evaluation framework using large language model via checklist,” *arXiv preprint arXiv:2403.18771*, 2024.

[9] J. Fu *et al.*, “Gptscore: Evaluate as you desire,” *arXiv preprint arXiv:2302.04166*, 2023.

[10] G. Sandoval *et al.*, “Lost at c: A user study on the security implications of large language model code assistants,” in *Proc. of USENIX Security 2023*.

[11] L. Kersten *et al.*, “A field study to uncover and a tool to support the alert investigation process of tier-1 analysts,” in *Proc. of USEC 2025*.

[12] G. Michelet *et al.*, “Chatgpt, llama, can you write my report? an experiment on assisted digital forensics reports written using (local) large language models,” *Forensic Science International: Digital Investigation*, 2024.

[13] C.-H. Chiang *et al.*, “A closer look into using large language models for automatic evaluation,” in *Proc. of EMNLP 2023*.

[14] M. D. Fetters *et al.*, “Achieving integration in mixed methods designs—principles and practices,” *Health services research*, 2013.

[15] B. A. Alahmadi *et al.*, “99% false positives: A qualitative study of SOC analysts’ perspectives on security alarms,” in *Proc. of USENIX Security 2022*.

[16] J. Brooks *et al.*, “Doing template analysis: evaluating an end of life care service,” *Sage research methods cases*, 2014.

[17] E. M. Redmiles *et al.*, “A summary of survey methodology best practices for security and privacy researchers,” 2017.

[18] S. Nepal *et al.*, “Burnout in cybersecurity incident responders: Exploring the factors that light the fire,” *Proc. of the ACM on Human-Computer Interaction*, 2024.

[19] W. Kintsch, “The role of knowledge in discourse comprehension: a construction-integration model.” *Psychological review*, 1988.

[20] A. Sundaram, “An introduction to intrusion detection,” *Crossroads*, 1996.

[21] L. Kersten *et al.*, “’give me structure’: Synthesis and evaluation of a (network) threat analysis process supporting tier 1 investigations in a security operation center,” in *Proc. of SOUPS 2023*.

[22] B. Murugadoss *et al.*, “Evaluating the evaluator: Measuring llms’ adherence to task evaluation instructions,” in *Proc. of AAAI 2025*.