

サイバーセキュリティ関連情報のキュレーションに向けた 大規模言語モデルによる要約生成の人手評価方式の提案

松田 美慧^{1,2,a)} 宮本 耕平¹ 藤田 彬¹ 古本 啓祐¹ 井野 毅也¹ piyokango¹ 井上 大介¹

概要：サイバーセキュリティインシデントの発生時、迅速かつ正確なキュレーション記事の発信は類似事例の予防の観点から重要である。迅速かつ正確なキュレーションの実現には、情報の収集、突合、要約などを実行する自然言語処理の技術が欠かせない。なかでも大規模言語モデルは要約などにおいて革新的な成果を挙げており、複数文書要約による記事の自動生成への活用が期待される。しかし、大規模言語モデルの出力には、情報の誤変換や必要な情報の欠落などが頻発するため、入力と出力の内容の齟齬の有無を確認する必要がある。文書要約の評価には、一般的に複数文書間の類似性を判断する定量評価の指標が用いられるが、情報の誤変換や欠落の判断は最終的に人間による評価を要し、その評価手法は確立されていない。本研究では、大規模言語モデルを活用した複数文書要約を実装し、サイバーセキュリティ分野のキュレーション記事に対する評価のスキームを提案する。このスキームでは、攻撃の内容や被害規模などインシデントに関連する事象ごとに、収集された情報から記事に載せるべき重要な情報を抽出した評価シートをあらかじめ定め、評価シートに記載された内容と大規模言語モデルの出力との意味的な包含関係を人間が評価する。本研究の成果は、半自動的なキュレーションの効率化に貢献する。

キーワード：大規模言語モデル, キュレーション, OSINT

A Proposal for a Manual Evaluation Method for Curated Articles on Cybersecurity Information Generated by a Large Language Model

MISATO MATSUDA^{1,2,a)} KOHEI MIYAMOTO¹ AKIRA FUJITA¹ KEISUKE FURUMOTO¹ TAKEYA INO¹
PIYOKANGO¹ DAISUKE INOUE¹

Abstract: The curation of cybersecurity-related information plays an important role in incident prevention and post-incident response, and it must be fast and accurate. Natural language processing techniques for collecting, matching, sorting, and summarizing information are essential for rapid curation. Among these, large-scale language models have achieved innovative results in summarization and are expected to be used for summarization. However, the output of large-scale language models leaves a problem in the accuracy of information. Although existing metrics for quantitative evaluation have been proposed as general criteria for document summarization, human evaluation is ultimately required. In this study, we implement multiple document summarization using a large-scale language model and propose a scheme of human assessment. In this scheme, an evaluation sheet that extracts essential information that should be included in the article from the collected information is defined in advance, and the evaluation is performed by judging the implication with the contents described in the evaluation sheet. The results of this research will contribute to the realization of semi-automatic curation.

Keywords: Large Language Model, curation, OSINT

¹ 国立研究開発法人情報通信研究機構 National Institute of Information and Communications Technology

² 横浜国立大学大学院環境情報学府

Graduate School of Environment and Information Sciences, Yokohama National University

^{a)} m.matsuda@nict.go.jp

1. はじめに

サイバーセキュリティ分野において、巧妙化する攻撃の手法や複雑化する防御策に関する情報をまとめ、エンドユーザーや開発者などの対策の実施者に向けて発信する活動は、非常に重要である。特にセキュリティインシデントが発生した際に、インシデントによる被害の規模や原因、再発防止策といった情報を集約したキュレーション記事は、需要が高い。キュレーション記事を作成する工程には、セキュリティインシデントに関する情報の検索、被害を受けた組織からの公開情報や新聞記事などの複数の情報の確認、情報の取捨選択、記事の構成の決定、記事の作成などがある。これまでは人間の作業者がこれらの作業を実施してきたが、活動を継続するには膨大な時間的コストを要する。そのため、自然言語処理技術による自動化が必要である。

著者らは、記事の読者が必要とする情報を収集し、読者が求める情報が十分に含まれた記事として要約する自動キュレーション技術を開発している。キュレーションを自動化することで記事の作成にかかる時間の短縮が期待されるが、同時に要約生成される記事が、収集した情報(情報源)に対して正確であることが求められる。この正確性を確保するには、要約生成された記事と収集した情報との間で内容を突合し、情報源に対して相違がある箇所及び不足する情報を認識し、修正もしくは追記する必要がある。

著者らは、この記事の要約生成に大規模言語モデル(以下、LLM)を用いる。昨今のLLMは、自然言語で記述した命令(プロンプト)を入力すると、命令の内容に対して一定程度の正確性をもちながら流暢な表現で出力となるテキスト等を返す。一方で、LLMは、情報源に対して誤った情報を出力したり、使用者が必要とする情報が取り落とされたテキストを出力することがある。そのため、特に情報の正確性が求められるサイバーセキュリティ分野のキュレーション記事作成においては、LLMの性能の十分な検証と、情報の正確性を保証する要約評価手法が必要である。

しかし、既存の要約評価手法は、文書間の類似度を測ることを基本とする指標が多く、キュレーション記事に求められる要件を評価する指標にはなっていない。その上、自動評価の結果は人間の直感的な評価には反することも多く、最終的な評価は人間によって実施されることが多い。

本研究では、LLMを用いた要約を実施し、事前に事実確認がされた情報と生成文の内容を比較し、必要な情報の欠落や情報の誤変換の有無を評価する手法を提案した。提案手法では、人手評価シートを定義し、セキュリティインシデントが発生してから収束に至るまでに起こりうる事象ごとに、人間の評価者が評価を実施した。評価結果の分析は、評価者間での評価結果の比較、項目ごとの比較、生成方法ごとの比較という3つの観点で実施した。評価者間の比較では、評価者間でカッパ係数を算出し、評価結果につ

いて一部で高い一致度を確認した。項目ならびに生成方法ごとの比較では、評価結果の数量的比較と有意差検定を行い、複数の群間で有意な差が確認された。項目ごとの比較では、特にセキュリティインシデントの原因調査や対策に関する情報が欠落しやすい可能性が示された。本研究の成果は、サイバーセキュリティ分野のキュレーション記事について、LLMによる自動生成を実施し、適切な評価手法と出力方法についての知見を得た点にある。

2. 既存の評価手法

2.1 自動評価の指標

生成された要約文の評価には、いくつかの定量的な評価指標が提案されている。多くの評価指標は、評価対象となる要約文(候補文と呼称)と、要約の正解例とみなされる文(参照文と呼称)を比較し、両者の類似の程度を評価している。ここでは、代表的な評価指標のROUGE[1]とBERTScore[2]について説明する。

ROUGE(Recall-Oriented Understudy for Gisting Evaluation)は、候補文と参照文の比較に基づく、伝統的な評価指標の一つである。ROUGEには、ROUGE-LやROUGE-Sといったいくつかの種類があり、なかでも最も基本的な指標はROUGE-Nである。ROUGE-Nでは、Nを自然数として、文中の連続したNトークン(N-グラム)を用いて類似度を測る。ROUGE-NをはじめとするROUGEの算出では、基本的に比較する文書間の一致度をトークン単位の一貫に基づいて求め、これを指標とする。しかし、自然言語で記述された文章には、同じ意味であっても表層表現が異なる場合や言い回し表現などが複数存在する。そのため、トークン単位での一致によって算出される類似度は、文書間の意味の類似を正確に反映できないという欠点がある。

BERTScoreは、BERT[3]によって得られる各トークンのベクトル表現の類似度に基づいた文章生成の評価指標である。BERTはLLMと同様のTransformer[4]アーキテクチャを基礎としたベクトル表現の学習手法である。BERTScoreでは、候補文と参照文のそれぞれをBERTの入力として、それらに含まれるトークンのベクトル表現を得る。そして、候補文中のトークンと参照文中のトークンの組について、それらのベクトル表現の内積によって定義される類似度を用いて評価指標を構成する。BERTScoreはトークン自体の一貫ではなくBERTによるベクトル表現の類似度に基づいて評価するため、ROUGEによる評価では反映されにくい意味の類似も評価に反映できる。一方で、評価に用いるベクトル表現自体が機械学習によって獲得されたものであることから、その性能に結果の妥当性が左右される点が欠点である。この他の評価手法には、BLEU[5]やQAEval[6]などがある。

宮本ら[7]は、ROUGEなどの定量的かつ自動での評価が可能な評価指標を用いて、LLMが生成した要約文を評

価した。しかし、自動評価指標による評価結果は、人間が行う評価の良し悪しとは必ずしも一致しないとし、人間による評価手法の必要性を主張している。

2.2 人間による評価の指標の必要性

自動評価指標は、評価対象が大量にある場合や、評価基準の一貫性が重要視される場合に特に有用である。しかし、ROUGE が言い換え表現を適切に評価できないことをはじめとし、自動評価指標はしばしば人間の直感とは矛盾した結果を示す。そのため、人間が実際に評価対象を読んで評価を決定する評価方法は、自然言語処理分野において依然重要なものである。人間による評価には様々な観点があり、多くの評価方法、評価基準が提案されているが、統一的な評価方法は著者らの知る限りは提案されていない。本研究では、サイバーセキュリティインシデントに関する情報のキュレーションの観点に基いた評価方法を提案する。

3. 評価スキーマの提案

3.1 データ

サイバーセキュリティ分野のキュレーション記事を発信する日本語ブログの代表例には、piyolog[8]がある。piyologは、警察による検挙を報じた報道記事や、被害を被った組織からの被害報告などの公開情報を元に、サイバーセキュリティインシデントに関する記事を発信する個人ブログで、サイバーセキュリティアワードを受賞するなど高く評価されている。各記事は原則として、記事の本文と、ブログの著者が記事の作成にあたって参照した参考文献の一覧を合わせた構成を取る。

本研究では、ブログ記事の本文と、参考文献欄に記載されたニュース記事等の URL からテキストデータを収集し、piyolog の記事ごとに両者を収容したデータセットを作成した。収集した記事の範囲は、2022 年 1 月から 2023 年 12 月 12 日までに公開された piyolog の全記事 74 件である。piyolog の記事本文は、参考文献に含まれる情報をブログの著者が要約した結果と解釈できる。よって、作成したデータセットは参考文献とその要約のペアで構成される要約タスクのデータセットとみなせる。piyolog 記事からの参考文献の URL の収集にあたっては、作業量を削減するため、収集する URL は最大 5 件とし、それ以上の収集は実施しない。収集した参考文献の URL からのテキストデータの収集では、情報の公開からの時間経過などの何らかの理由により、テキストデータが取得できないニュース記事については、データの収集を行わない。上記の場合に該当し、データを取得しなかった参考文献が存在する場合、piyolog の記事本文と参考文献のペアデータをデータセットから除外した。なお、要約文の生成以降の作業は、全ての参考文献が取得された piyolog 記事 17 件についてのみ実施した。

表 1 使用した言語モデルの一覧

本稿での表記	使用したモデル名
model 1	elyza/ELYZA-japanese-Llama-2-7b-fast-instruct
model 2	tokyotech-llm/Swallow-7b-instruct-hf
model 3	rinna/your-7b-instruction
model 4	moneyforward/houou-instruction-7b-v2

表 2 プロンプトにおける指示の一覧

prompt	指示
prompt 1	入力文章を要約してください。
prompt 2	入力文章を要約せよ。
prompt 3	入力文章を要約せよ。要約は箇条書きで記述せよ。
prompt 4	入力文章を要約せよ。要約は箇条書きで記述せよ。補足的な情報も項目として記述せよ。
prompt 5	入力文章を要約してください。要約文章以外は出力しないでください。

3.2 要約文の生成

要約文の生成では、4 種類のモデルと 5 種類のプロンプトを組み合わせた 20 通りの出力を得た。使用した言語モデルの一覧を表 1 に示す。使用した言語モデルは、いずれも 70 億パラメータ規模の言語モデルで、日本語能力の強化を意図した追加学習および、指示形式の追加学習を実施したモデルであり、HuggingFace^{*1}上で公開されている。

言語モデルへ入力したプロンプトは、指示文の後に、参考文献から収集されたテキストデータを記述する形式をとる。指示文と収集されたテキストデータとの境界には、「入力文章:」という文字列を記載し、言語モデルへの指示を明確に分けるように工夫した。指示文は、入力文を要約するように指示する日本語文で、文言が少しずつ異なる 5 種類を用意した。言語モデルへの指示の実施時は、いずれかの指示文を 1 種類ずつを用いて、指示を出す。以降では、各指示を組み合わせたプロンプトをそれぞれ Prompt1~5 と呼称する。各指示の一覧を表 2 に示す。prompt 1 と prompt 2 は同一の内容を口調を変えた指示文である。prompt 3 では出力を箇条書きで行うように指示を追加し、prompt 4 では箇条書きに加えてより詳細な内容を出力に含むよう促している。prompt 5 では余分な出力を抑制するための指示を加えている。想定している余分な出力は、「入力文章を要約すると次のようになります」というような定型文である。評価結果の分析では、モデルとプロンプトごとに出力結果に対する評価に差異が生じるかを検証する。

3.3 人手評価シートの作成

LLM の出力に対して、情報の誤変換と必要な情報の欠落の有無を評価するため、人手評価シートを作成し、評価に導入した。人手評価シートには、piyolog の記事に記載された内容を元に、LLM の出力に含まれることが望ましい内容が、項目ごとにあらかじめ定められている。項目は、サイバーセキュリティインシデントの発生から収束に至る

*1 <https://huggingface.co/>

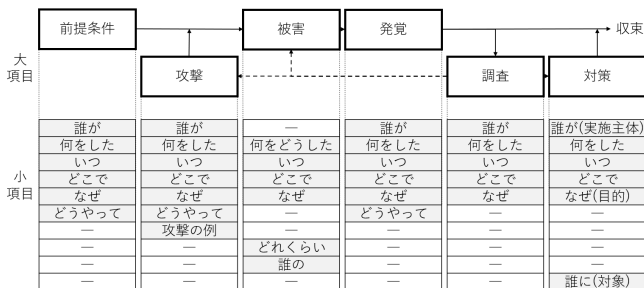


図 1 評価項目の体系

までに起こりうる事象に即した 6 種類の大項目 [前提条件 / 攻撃 / 発覚 / 被害 / 調査 / 対策] と、10 種類の小項目 [誰が / 誰の / 誰に / 何を / いつ / どこで / なぜ / どうやって / どれくらい / 攻撃の例] で構成される。小項目は必要に応じ、大項目に選択的に付随する。評価項目の体系を図 1 に示す。図 1 中において灰色でマークされた小項目は、各大項目に付随していることを表し、各大項目につき、最大 7 つ、最小 5 つの小項目が付随する。

評価シートの作成では、まず piyolog の記事本文を文節単位で区切り、各評価項目の該当箇所のいずれかに振り分けた。そして振り分けた記事の本文の内容をより簡潔に記述した文章（以下では、正例文と呼称）を評価項目ごとに用意した。要約文の生成には、piyolog の参考文献に記述された内容を用いるが、評価シートの作成には、piyolog の記事本文のみを用い、原則として参考文献欄にある URL から収集したテキストデータの内容は参照しない。piyolog に記述された情報は、複数の公開情報からニュースバリューがあると判断され、選別された情報である。このことから、piyolog の記事本文のみを参考に正例文を作成することで、冗長さを排した状態での評価を目指す。なお、piyolog に該当する記述のなかった大項目ならびに小項目は、正例文を作成できないため、本評価手法の対象から除外する。評価シートの生成は、原則として一人の人間の作業が実施し、もう一人の作業が評価項目ごとに記述内容に過不足がないことなどを確認した。作成した評価シートの例を図 2 に示す。図 2 中の左から 5 列目には piyolog の記事本文が記載されている。また評価作業では、左から 4 列目の太枠で強調した箇所に評価結果を記入する。

3.4 評価の実施

評価作業では、3 名の評価者が同一の指示のもと、作業を実施した。評価者へ指示された内容は、人手評価シートに記入した正例文と LLM の出力を比較し、正例文が記載された評価項目ごとに含意関係を判定するというものである。含意関係の判定とは、LLM の出力における情報の誤変換や必要な情報の欠落の有無を評価することを指す。LLM の出力には、不自然な文体やキュレーション記事として不適切な表現が含まれる場合があるが、これらの評価は行わ

大項目	小項目	正例文	評価	piyologの記事本文の該当箇所
攻撃	誰が	近畿大学病院より受付業務の委託を受けていたエヌジェシーの元従業員		不正行為を行ったのは、近畿大学病院より受付業務の委託を受けていたエヌジェシーの従業員(当時、既に退職しているため元従業員と記載)
	何を	患者の診療情報診療情報(氏名、生年月日、診療記録)や受診の状況を知人などに共有した		・元従業員は私物のスマートフォンを使って患者1名の電子カルテシステムの画面を動画で数十秒間撮影し、二人の共通の知人(元従業員と患者も知り合い)に対してSNSを通じてその動画を送信した。
	どうやって	私物のスマートフォンを使って患者1名の電子カルテシステムの画面を動画で数十秒間撮影した		・元従業員及び知人は家族や別の知人に対して動画を見せたり、口頭で患者の受診について共有を行っており、最終的に当該患者の診療情報(氏名、生年月日、診療記録)や受診について少なくとも6人へ流出した。
	いつ	2022年11月5日以降		<関連タイムライン>
	どこで	SNS、口頭		
	なぜ	友情、リテラシーの欠如		元従業員は不正行為を行った動機について、友人であったことから安易に迷ってしまったと話している。
発覚	誰が	患者1名		動画送信を受けた知人は患者に対して事実を伝え、動画を見せたところ患者は病院側へ抗議を行ったことから流出の事象が発覚した。
	何を	患者の病院側への抗議		
	いつ	2022年11月7日		<関連タイムライン>
	どこで			

図 2 評価シートの例

表 3 評価の段階の一覧

段階	説明文
A	正例文にある内容を完全に含意している
A'	正例文にある内容を暗黙的に完全に含意している
B	正例文にある内容を一部含意している
B'	正例文にある内容を暗黙的に一部含意している
C	正例文にある内容を全く含意していない
H	正例文にある内容を誤りを含みながら含意している

ない。不自然な文体の例には、文章の途中で文章が途絶える場合や、前置きもなく一文ごとに話す視点が変化する場面などがある。キュレーション記事として不適切な表現の例として、セキュリティインシデントにより被害を受けた企業の担当者など当事者や関係者視点での語り口調や、日本語以外の言語での記述などがある。評価は、6 段階で行い、各段階の評価の基準を表 3 に示す。表 3 の各説明文は、評価方法を評価者に説明した際に、評価者に提示した文である。A' と B' の説明文にある「暗黙的に」とは、正例文に記載された内容が明記されていないが、その内容を含意していると解釈可能な場合を指す。評価者は、これらの指示に従い、出力に使用された LLM とプロンプトの組み合わせを知らない状態で、評価を行った。作業期間は、2024 年 6 月 24 日から 7 月 23 日までの約 1 ヶ月で、評価者は期間内のいずれかの時間に断続的に作業を実施した。

評価対象とする piyolog の記事は、4 通りのモデルと 5 通りのプロンプトを組み合わせた全 20 通りの入力を実施した際に、入力長の制限によるエラーなどを伴わず入力が行われ、出力が欠損なく得られた記事 7 件である。以降では、記事 7 件に対応する 20 通りの要約文をひとまとめとした session を一単位として、評価対象を取り扱う。各 session をそれぞれ session1~7 と呼称する。一人の作業者は一つの session あたり、1 から 2 時間ほどの時間をかけて作業をした。各 session で取り扱う記事の公開日とタイトル、評価を実施する項目数を表 4 に示す。タイトルには一部特定の製品名や企業名を含むが、既存の公開情報であることから本論文への掲載に伴い、当該製品や団体に対する追加的な影響は極めて低いと判断し、掲載する。

表 4 評価対象の記事一覧

session	公開日	記事タイトル	項目数
session 1	2023-11-27	J:COM のメッシュ Wi-Fi の情報流出についてまとめてみた	19
session 2	2023-02-03	委託先従業員による診療情報の流出についてまとめてみた	40
session 3	2023-01-26	購入者からの問合せになりました EC ショップ管理者を狙うフィッシングについてまとめてみた	40
session 4	2023-01-19	Norton アカウントへのリスト型攻撃についてまとめてみた	36
session 5	2023-01-13	アフラックとチューリッヒの情報流出についてまとめてみた	38
session 6	2022-03-04	メタックスペイメントの情報流出についてまとめてみた	63
session 7	2022-01-17	私物ハードディスク売却後に起きた社内情報の流出についてまとめてみた	41

3.5 評価結果の分析

分析では、評価者間、評価項目間、生成方法ごとの3つの観点から評価結果を比較した。評価者間の比較では、カッパ係数 (kappa coefficient) を算出し、2名の異なる評価者間で評価結果の一致度をそれぞれ測った。カッパ係数は、0~1の値をとり、値が大きいくほど評価者間での評価結果の一致度が高いことを示す。一般にはカッパ係数が0.6以上の場合に、評価者間の一致度が十分高いと判断される。これに倣い、本研究ではカッパ係数が0.6以上の場合に、評価者間での評価結果の一致度が高いと評価する。

評価項目間の比較では、大項目 [前提条件 / 攻撃 / 発覚 / 被害 / 調査 / 対策] 間で、出力の性能に差異がないかを検証するため、有意差検定を実施した。比較する群は、6群の対応のないデータであるため、クラスカル・ウォリス検定 (Kruskal-Wallis test) を実施した。クラスカル・ウォリス検定にて有意差が確認された場合は、スティール・ドゥアスの多重比較分析 (Steel-Dwass test) を実施し、有意差がある群の組を特定した。有意差検定に伴う演算処理では、A から H までの文字列で表現された評価結果を、順に5から0までの整数に変換し、演算を行った。

生成方法ごとの比較では、提案手法において評価の高い記事の生成方法を特定するため、使用した LLM とプロンプトごとに、6段階の評価が出現した割合を求め、有意差の有無を検証した。検証対象は、言語モデルが4種類、プロンプトが5種類であり、両方とも3群以上の対応のないデータである。よって、有意差検定では評価項目間での比較と同様の手順で、言語モデルの4群間とプロンプトの5群間に対して、クラスカル・ウォリス検定とスティール・ドゥアスの多重比較分析を実施した。

なお、session5の被害規模を表す項目 (大項目 [被害]、小項目 [どれくらい]) では、LLM に入力した参考文献のテキストデータと piyolog の記事本文で、被害件数の更新時期が異なったため、記述されていた被害件数に相違が確認された。これにより、LLM への入力文に piyolog の内容とは違う内容が含まれ、LLM の出力に人手評価シートの内容とは異なる内容が含まれた。人手評価シートは、LLM による不正確な情報の誤変換、必要な情報の欠落を評価することが目的であるため、両者には該当しない当該項目は、分析の対象から除外した。

表 5 評価者間で算出したカッパ係数の一覧

	評価者 X-Y 間	評価者 X-Z 間	評価者 Y-Z 間
session 1	0.597	0.341	0.431
session 2	0.694	0.565	0.603
session 3	0.207	0.022	0.052
session 4	0.404	0.218	0.156
session 5	0.811	0.257	0.248
session 6	0.369	0.228	0.249
session 7	0.545	0.291	0.385
全体	0.607	0.314	0.317

表 6 評価者 X-Y 間の評価の段階別の頻度

		評価者 X						評価者 Y の小計
		A	A'	B	B'	C	H	
評価者 Y	A	394(7.112)	44(0.794)	40(0.722)	6(0.108)	61(1.101)	0(0.000)	545(9.838)
	A'	130(2.347)	68(1.227)	3(0.054)	4(0.072)	84(1.516)	0(0.000)	289(5.217)
	B	25(0.451)	16(0.289)	176(3.177)	8(0.144)	52(0.939)	1(0.018)	278(5.018)
	B'	42(0.758)	47(0.848)	57(1.029)	5(0.090)	64(1.155)	0(0.000)	215(3.881)
	C	67(1.209)	31(0.560)	44(0.794)	6(0.108)	4,025(72.653)	2(0.036)	4,175(75.361)
	H	3(0.054)	1(0.018)	12(0.217)	0(0.000)	17(0.307)	5(0.090)	38(0.686)
評価者 X の小計		661(11.931)	207(3.736)	332(5.993)	29(0.523)	4,303(77.671)	8(0.144)	5,540(100.000)

4. 結果

4.1 評価者間での比較

各 session と評価結果の全体に対して、評価者2名ずつの評価結果を用いて算出したカッパ係数の一覧を表5に示す。表5において背景が灰色で強調された箇所は、カッパ係数が0.6以上となり、一致度が高いことを示す。評価者3名の識別には、便宜的にX, Y, Zの記号を用いる。

表5より、評価者Xと評価者Yの間で算出されたカッパ係数は、session1からsession7までの全ての評価結果を用いた場合で0.6以上となり、高い一致度が確認された。特に、session2とsession5において、カッパ係数が0.6以上となり、高い一致度が示された。評価者Yと評価者Zの間では、session2においてカッパ係数が0.6以上となった。評価者Xと評価者Zの間では、いずれの比較においても高い一致度は確認されなかった。

同じ評価対象について評価者間で評価が異なる場合、評価の段階がどれほど異なるのかを明らかにするため、クロス集計の結果を2名の評価者ごとにそれぞれ表6, 表7, 表8に示す。表中の括弧内の数値は、session1から7までを含む全評価対象5,540件に対する該当件数の割合を100分率で表したものである。各表の対角線上のマスを確認すると、評価対象の約60%以上では異なる評価者間で評価が一致していたことが明らかになった。そのほとんどは、異なる評価者が両者ともC(含意関係が全くない)と評価した場合であった。一方、同じ評価対象について、A(完全に含意関係がある)と評価した評価者と、H(含意関係が全くない)と評価した評価者が存在する場合は、約0.5%から約1.6%あることが明らかになった。この他、評価者間で評価結果が異なった割合が高かったケースは、正例文との含意関係が完全あるいは部分的かの評価であった。

表 7 評価者 X-Z 間の評価の段階別の頻度

		評価者 X						評価者 Z の小計
		A	A'	B	B'	C	H	
評価者 Z	A	161(2.906)	25(0.451)	28(0.505)	1(0.018)	69(1.245)	0(0.000)	284(5.126)
	A'	0(0.000)	0(0.000)	0(0.000)	0(0.000)	0(0.000)	0(0.000)	0(0.000)
	B	398(7.184)	133(2.401)	184(3.321)	15(0.271)	708(12.78)	2(0.036)	1,440(25.993)
	B'	4(0.072)	1(0.018)	3(0.054)	3(0.054)	35(0.632)	0(0.000)	46(0.830)
	C	89(1.606)	48(0.866)	100(1.805)	10(0.181)	3,436(62.022)	4(0.072)	3,687(66.552)
H	9(0.162)	0(0.000)	17(0.307)	0(0.000)	55(0.993)	2(0.036)	83(1.498)	
評価者 X の小計		661(11.931)	207(3.736)	332(5.993)	29(0.523)	4,303(77.671)	8(0.144)	5,540(100.000)

表 8 評価者 Y-Z 間の評価の段階別の頻度

		評価者 Z						評価者 Y の小計
		A	A'	B	B'	C	H	
評価者 Y	A	155(2.798)	0(0.000)	302(5.451)	5(0.090)	75(1.354)	8(0.144)	545(9.838)
	A'	37(0.668)	0(0.000)	197(3.556)	1(0.018)	52(0.939)	2(0.036)	289(5.217)
	B	21(0.379)	0(0.000)	159(2.870)	2(0.036)	84(1.516)	12(0.217)	278(5.018)
	B'	13(0.235)	0(0.000)	131(2.365)	3(0.054)	66(1.191)	2(0.036)	215(3.881)
	C	55(0.993)	0(0.000)	642(11.588)	33(0.596)	3,397(61.318)	48(0.866)	4,175(75.361)
H	3(0.054)	0(0.000)	9(0.162)	2(0.036)	13(0.235)	11(0.199)	38(0.686)	
評価者 Z の小計		284(5.126)	0(0.000)	1,440(25.993)	46(0.830)	3,687(66.552)	83(1.498)	5,540(100.000)

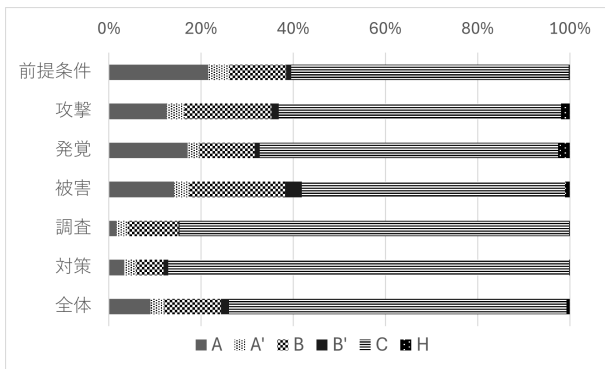


図 3 項目ごとに各評価段階が出現した割合

4.2 評価項目間の比較

評価項目ごとに、評価の段階別の内訳を図 3 に示す。図 3 より、[調査]ならびに[対策]の項目では、C と評価された割合が約 80% と非常に高く、A と評価された割合はその他の比較群に比べて低くなるのが明らかになった。また評価項目間でクラスカル・ウォリス検定を行った結果、統計値は 1373.9、 p 値は $6.09e-295$ となった。 p 値が 0.05 を下回ったことから、評価項目間には評価結果に統計的に有意な差が認められた。スティール・ドゥアスの多重比較分析を行った結果を表 9 に示す。表 9 中の灰色で強調された箇所は、 p 値が 0.05 を下回り、有意な差が確認されたことを示す。「 <0.0005 」という表記は、 p 値が 0.0005 以下であることを表す。表 9 より、[前提条件]と[被害]、[攻撃]と[発覚]、[対策]と[調査]では有意な差が確認されなかったが、それ以外の 12 組では有意な差があることが示された。

4.3 生成方法の比較

4.3.1 モデルの比較

モデルごとに評価の段階別の内訳を図 4 に示す。4 種類の全てのモデルにおいて、最も出現した割合が高かったのは C で、各モデルにおいて約 60% 以上の割合を占めた。また完全に含意関係があると評価されたことを示す A の割合が model 4 で最も高くなった。モデル間でクラスカル・ウォリス検定を行った結果、統計値は 341.11、 p 値は $1.26E-73$ となった。 p 値が 0.05 を大幅に下回ったこと

表 9 項目間での有意差検定の結果

比較対象	統計値	p 値
前提条件：攻撃	3.165	0.019
前提条件：被害	1.203	0.836
前提条件：発覚	3.781	0.002
前提条件：調査	12.283	<0.0005
前提条件：対策	19.810	<0.0005
攻撃：被害	3.319	0.012
攻撃：発覚	1.240	0.817
攻撃：調査	11.230	<0.0005
攻撃：対策	22.044	<0.0005
被害：発覚	4.307	<0.0005
被害：調査	14.973	<0.0005
被害：対策	34.427	<0.0005
発覚：調査	9.009	<0.0005
発覚：対策	16.573	<0.0005
対策：調査	1.784	0.476

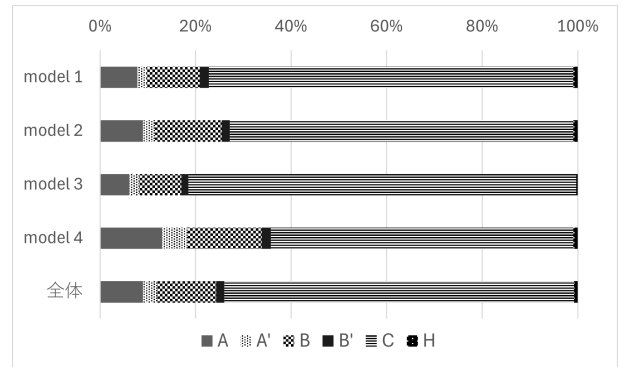


図 4 モデルごとに評価段階が出現した割合

表 10 モデル間での有意差検定の結果

比較対象	統計値	p 値
model 1 : model 2	4.482	<0.0005
model 1 : model 3	4.179	<0.0005
model 1 : model 4	13.137	<0.0005
model 2 : model 3	8.678	<0.0005
model 2 : model 4	8.804	<0.0005
model 3 : model 4	17.247	<0.0005

から、モデル間には統計的に有意な差が認められた。有意な差が確認された群の組み合わせを明らかにするため、スティール・ドゥアスの多重比較分析を行った結果を表 10 に示す。表 10 内の灰色で強調された箇所と「 <0.0005 」の意味は、表 9 で説明したとおりである。表 10 より、全ての組み合わせで有意な差が確認された。モデル間での出力を比較するため、session3 の prompt 4 を用いた出力を例として付録の表 A.1 に示す。

4.3.2 プロンプトの比較

プロンプトごとに評価の段階別の内訳を図 5 に示す。5 種類のプロンプトにおいて、評価の段階別の割合に大きな差は見られなかった。評価項目間、モデルごとの比較と同様に、最も出現した割合が高かったのは C であった。ク

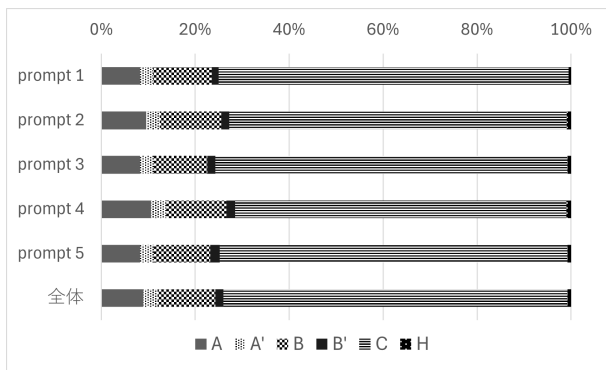


図 5 プロンプトごとに評価段階が出現した割合

表 11 プロンプト間での評価結果の比較

比較対象	統計値	p 値
prompt 1 : prompt 2	1.858	0.340
prompt 1 : prompt 3	0.800	0.931
prompt 1 : prompt 4	3.074	0.018
prompt 1 : prompt 5	0.035	1.000
prompt 2 : prompt 3	2.632	0.065
prompt 2 : prompt 4	1.216	0.742
prompt 2 : prompt 5	1.887	0.324
prompt 3 : prompt 4	3.833	0.001
prompt 3 : prompt 5	0.762	0.941
prompt 4 : prompt 5	3.099	0.017

ラスカル・ウォリス検定の結果、統計値は 20.377、 p 値は 0.0004206 となった。 p 値が 0.05 を大幅に下回ったことから、モデルごとの比較と同様に、プロンプト間には有意な差が認められた。有意な差が見られる組み合わせの特定のため、スティール・ドゥアスの多重比較分析を行った結果を表 11 に示す。表 11 内の灰色で強調された箇所は、表 9 と表 10 の灰色の箇所と同様の意味を持つ。表 11 より、3 組の群間で p 値が 0.05 下回り、そのほかの群間では有意差が確認されなかった。有意差が確認された組み合わせはいずれも prompt4 を片方に含み、prompt1, prompt3 ならびに prompt5 との間で有意な差が確認された。

5. 考察

5.1 評価結果に対する考察

評価者間の比較では、一部の評価者間で高い一致度が確認された。これは、評価者間で評価のずれが少ないことを示し、評価手法の妥当性が部分的に認められたといえる。評価者間での一致度が比較的高かったのは session2 で、評価対象全体では高い一致度があると認められなかった評価者間においても中程度以上の一致度が確認された。session2 では他の session に比べ、各評価者が含意関係が全くないと評価した割合が高かった。この原因には、入力文の情報不足が考えられる。session2 で LLM へ入力された文章は、セキュリティインシデントの発生と経緯を報じた新聞記事のうち、無料で閲覧可能な冒頭の 190 文字であった。その

ため、他の session の入力文に比べて、session2 の入力文字数が少なく、piyolog の記事本文と比べても情報量が不足していた。情報量が十分ではない入力によって得られた出力に対し、含意関係が全くないという共通した評価がされたことから、ハルシネーションが起きた可能性は低いと考えられる。他方、一致度が比較的低かった session3 では、同じ評価対象に対し、正例文の内容を部分的に含意していると評価する評価者と、全く含意していないと評価する評価者が存在し、判断が分かれる場合が多く確認された。session3 は、セキュリティインシデントの発生によって被害を被った企業が顧客に向けて注意喚起を行う文章を入力としている。そのため、LLM による出力は、企業側の立場で問い合わせ先を案内しているように読み取れる文章になっている場合が多く、この案内に含まれる内容に対し、判断が分かれたと推察される。

評価項目での比較では、ほぼすべての項目間で有意な差が確認された。このことから、LLM の出力は項目ごとに情報の誤変換の度合いや必要な情報の欠落の頻度にばらつきがあると考察される。有意な差が確認されなかった [攻撃] と [発覚] の項目は、他の項目に比べ、ハルシネーションが起きている可能性を示す H と評価された割合が高く、出力は誤解を与える表現が多い傾向にあると考えられる。同じく有意差が確認されなかった [対策] と [調査] の項目では、完全に含意されないことを示す C と評価された割合が他の項目に比べて高く、模範としたキュレーション記事には記述があるが、LLM の出力には調査や対策に関する記述が含まれない傾向があることが明らかになったといえる。

生成方法の比較では、モデル間においても、プロンプト間においても、評価結果に有意な差が確認された。付録の表 A-1 の例に示すように、出力はモデル間で取捨選択された情報の量や文章の語り手の視点などが大きく異なる。モデル間では、全てのモデル間で有意差が確認されたことで、モデルによる性能の違いを裏付ける結果が得られた。評価の 6 段階の出現割合の比較において、完全に含意関係があると評価されたことを示す A の割合が model 4 で最も高くなったことから、moneyforward が作成した LLM が最も高く評価されたといえる。プロンプト間では prompt 4 と他の prompt の間でのみ有意差が確認された。prompt 4 は、補足的な情報も踏まえてより詳しい出力を要求する命令である。prompt 4 で、完全に含意関係があると評価された割合が最も高くなったことから、出力は命令に従い、多くの情報を入力から抽出したと考えられる。よって、より詳しい出力を要求することは、必要な情報の欠落を防止する方法として有用であると考察される。

5.2 提案手法の実運用への応用可能性に対する考察

提案手法で使用した人手評価シートのように、セキュリティインシデントの発生から収束に至るまでに起こりうる

事象ごとにデータを構造化することで、セキュリティインシデントについて新たな情報が見つかった場合に、情報の突合と記事の更新にかかる作業が容易になることが期待される。一方、評価手法の妥当性の評価するには、本研究で検証の対象とした7つ以外のデータセットを用い、更なる分析が必要である。また評価シートの作成と評価作業には時間的および人的コストが多くかかることから、実運用への応用に向けては、これらの工程の一部自動化が期待される。評価シートの作成の自動化では、攻撃の手法や被害の規模といった項目ごとに、収集した情報から情報を抽出するように LLM へ指示することで、自動化の実現が検討されている。また評価作業の自動化に向けては、既存の含意判定分析の手法の応用実験の実施を検討している。含意判定分析とは、自然言語処理において、二文間で意味の包含関係を自動的に判断する手法で、人手での含意判定を支援することが期待される。

6. おわりに

セキュリティインシデントが発生した際、被害状況などを纏めたキュレーション記事の需要は高く、記事の迅速な配信には自動化技術が欠かせない。自動化技術の開発に向けては、LLM の応用が期待される。しかし、LLM によって自動生成された記事の評価手法は確立されていない。本研究では、LLM によって作成された出力に対し、記事に含むべき情報を必要十分に誤りなく含むかを評価する手法を提案した。提案手法では、セキュリティインシデントの発生から収束に至る事象に即し、キュレーション記事に記載される必要のある情報を構造化した人手評価シートを作成し、評価実験を行った。評価者間、評価項目間、生成方法ごとの比較を行った結果、異なる評価者間で高い一致度が確認され、特に必要な情報が欠落していると評価する場合に一致度が高かった。また評価項目間、生成方法ごとの比較では、各比較群間で統計的に有意な差が確認され、LLM の出力には情報の誤変換の度合いや必要な情報の欠落の頻度にばらつきがあることが示された。これらの成果により、LLM をキュレーションに実用するための評価手法の構築と出力方法の確立に貢献した。今後の課題には、評価項目の追加、プロンプトの改良、評価シートの自動作成、既存の含意判定分析の適応実験の実施などが挙げられる。

参考文献

- [1] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, Text Summarization Branches Out, Barcelona, Spain, Association for Computational Linguistics, pp. 7481 (online), available from <https://aclanthology.org/W04-1013> (2004).
- [2] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y.: Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).
- [3] Devlin J., Chang M., Lee K., and Toutanova K.: BERT:

Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pp. 4171 - 4186 (2019).

- [4] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., and Polosukhin I.: Attention is all you need, Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), pp. 6000 - 6010 (2017).
- [5] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation, Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318 (2002).
- [6] Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth: Towards question-answering as an automatic metric for evaluating the content quality of a summary. Transactions of the Association for Computational Linguistics, Vol.9, pp.774-789 (2021).
- [7] 宮本耕平, 藤田彬, 松田美慧, 古本啓祐, 井野毅也, Piyokango, 井上大介, "サイバーセキュリティ関連情報のキュレーションに向けた大規模言語モデルによる要約生成の定量評価," 情報処理学会研究報告, vol.2024-CSEC-106, no.36, 8 pages, 2024年7月.
- [8] piyokango: piyolog, <https://piyolog.hatenadiary.jp/about>.

付 録

表 A-1 出力文の例: session 3 の prompt 4 での出力

model	出力
model 1	BASE を装った不審な URL にはアクセスしないようお願いいたします。重要な情報のためご確認いただくと幸いです。
model 2	BASE のショップページからのお問い合わせは、以下の方法で行うことができます。注文控えメールや商品発送通知メール下部の「ご注文に関するお問い合わせはこちら」ショップページの「CONTACT(お問い合わせ)」フォームショップページの「特定商取引法に基づく表記」記載の電話番号※ショップ運営者の住所、電話番号が非公開設定となっている場合がございます。ショップの「メッセージ」機能「メッセージ」機能でのお問い合わせショップにてメッセージ App をインストールしている場合、ショップへメッセージを送信し、チャット形式でやり取りすることが可能です。ショップページの右下にメッセージアイコンが表示されておりますので、こちらをクリックして、ショップへのメッセージを入力してください。なお、メッセージはショップからすぐに返答がない場合があります。ショップページを離れたり、ショップから返信があった際に通知を希望する場合には、ご希望の通知先のメールアドレスや電話番号 (SMS) を登録する必要があります。— こちらの事案ですが、メッセージ App 経由でも同様の BASE 管理画面のログイン情報を不正に入手することを目的とした、なりすましの事案が確認されました。BASE を装った不審な URL にはアクセスしないようお願いいたします。重要な情報のためご確認いただくと幸いです。
model 3	ショップページ (Web) からのお問い合わせ注文控えメールや商品発送通知メール下部の「ご注文に関するお問い合わせはこちら」ショップページの「CONTACT (お問い合わせ)」フォームショップページの「特定商取引法に基づく表記」記載の電話番号ショップの「メッセージ」機能「Pay ID アプリ」からのお問い合わせショップ情報の「ショップへのお問い合わせ」フォームショップ情報の「特定商取引法に基づく表記」記載の電話番号※ショップ運営者の住所、電話番号が非公開設定となっている場合がございます。ショップにてメッセージ App をインストールしている場合、ショップへメッセージを送信し、チャット形式でやり取りすることが可能です。
model 4	ショップページ (Web) からのお問い合わせ、「Pay ID アプリ」からのお問い合わせ、「メッセージ」機能でのお問い合わせがあるが、BASE を装った不審な URL にはアクセスしないようにと呼びかけている。