ISO27017 に基づくクラウドセキュリティ監査業務に対する LLM の性能評価

Assessing LLM's audit performance in ISO 27017-based cloud security audit engagements

研究員：多田 麻沙子（TIS 株式会社）
主査：石川 冬樹（国立情報学研究所）
副主査:徳本 晋（富士通株式会社）
アドバイザー:栗田 太郎（ソニー株式会社）

研究概要

クラウドセキュリティ監査を生成 AI の LLM(Large Language Model,大規模言語モデル)に任せられるかをテーマとした.不適合が正解であるパターンで失敗が多いのではないかとの仮説の下,ChatGPT (GPT4) を用いた実験で監査性能を評価した.併せて根拠の評価,失敗事例の分析,追加プロンプトによる正解率の向上を確認した.結果,やはり不適合が正解であるパターンでの正しい回答を導けないケースが多かったが,全体としてはクラウドセキュリティ監査を補助することは可能と考える.具体的な監査性能は,正解率 68.8%,適合率 37.5%,再現率 100%,特異率 100%であった.（不適合を正例とする.）傾向としては想定通り不適合を見抜く力が低く,傾向拡大解釈や推測などをして,ポジティブに適合と判断する傾向にあった.前述の監査性能は心元ない数字だが,重ねての質問で,正解率 90.6%,適合率は 81.3%まで上昇したため,補助能力ありと考える.

1. はじめに

クラウドサービスの情報セキュリティ管理策のガイドライン規格である ISO/IEC 27017 について,内部監査等を LLM に任せられるかをテーマとした.

筆者の携わる業務は ISO/ICE 27017:2015(以下,ISO/ICE 27017) の内部監査や,該当規格をベースとした点検結果の審査を実施している. 年間 200 件程度の審査を数名という少ない人手で実施するため, いかに業務効率を高めるかは大きな課題である. クラウドセキュリティに関する監査業務は多量の文書（利用約款,サービス仕様書,設計書等）を読み,合致しそうな箇所を確認し,判断する.生成 AI の LLM が文章読解にたけていることより,業務効率化のため,監査自体や監査の補助をすることを見出したい.

本論文では不適合が正解であるパターンで失敗が多いのではないかとの仮説の下,ChatGPT (GPT4)を用いた実験で監査性能を評価した.また,根拠の評価,失敗事例の分析,追加プロンプトによる正解率の向上を確認した.

結果,やはり不適合が正解であるパターンでの正しい回答を導けないケースが多かったが,全体としてはクラウドセキュリティ監査を補助することは可能と考える.

具体的な監査性能は,正解率 68.8%,適合率 37.5%,再現率 100%,特異率 100%であった.（不適合を正例とする）適合率は正（不適合）と ChatGPT が判断したもののうち,実際に正であった率をさす.再現率は実際のデータが正（不適合）であったもののうち,正（不適合）と ChatGPT が判断した率をさす.特異率は ChatGPT が負（適合）と判断したもののうち,実際のデータが負（適合）であったものをさす.傾向としては想定通り不適合を見抜く力が低く,ChatGPT が回答した根拠を分析すると,傾向拡大解釈や推測などをして,ポジティブに適合と判断する傾向にあった.

前述の監査性能は心元ない数字だが,重ねての質問で,正解率 90.6%,適合率は 81.3%まで上昇したため,補助能力ありと考える.

2.　背景

2.1.ISO/IEC27017 に基づくクラウドセキュリティ監査

ISO/IEC 27017 は, クラウドサービスに関する情報セキュリティ管理策のガイドライン規格である. [ISO/IEC 27017:2015 は,クラウドサービス分野の ISMS を確立するための分野別規格である.][1] 一般にアドオン認証と言われ ISMS で手薄なクラウドサービス特有のセキュリティリスクに対応している. ISO/IEC 27017 は, ISMS 構築の実践的なセキュリティ管理策を定めた ISO/IEC 27002 に対し, 特にクラウドサービスに関連した「管理策」と「実施の手引き」を追加したものである. 本論文では最も具体的な「実施の手引き」で実験を行い,「設問」と呼ぶこととする.

筆者は ISO/ICE 27017 の認証取得サービスの内部監査の実施や,社内の全クラウドサービスについて ISO/ICE 27017 ベースの点検を義務付け,点検結果の審査を実施している. フォローアップ監査を含めると年間 200 件程度の審査を数名という少ない人手で実施するため,いかに業務効率を高めるかは大きな課題である.

クラウドセキュリティに関する監査業務は多量の文書（利用約款,サービス仕様書,設計書等）を読み,合致しそうな箇所を探し出し,判断する.例えば,「CLD9.5.1 仮想コンピューティング環境における分離」では,クラウドサービスカスタマ間のリソースの分離や,クラウドサービスカスタマのリソースからクラウドサービスプロバイダの内部管理の分離が求められる.それにはシステム構成・ネットワーク構成,仮想環境の分離方式などを設計書等から把握した上で判断することとなる.

設問の要求に対して十分な対応が証跡より確認できている場合は「適合」,確認できない場合を「不適合」と表現する.

生成 AI の LLM が文章読解にたけていることより,業務効率化のため,監査自体や監査の補助をすることを見出したい.

2.2.LLM(Large Language Model,大規模言語モデル)

LLM(Large Language Model,大規模言語モデル）とは,[大量のデータとディープラーニング（深層学習）技術によって構築された言語モデルである.言語モデルは文章や単語の出現確率を用いてモデル化したものであり, 文章作成などの自然言語処理で用いられている. 大規模言語モデルと従来の言語モデルでは,「データ量」「計算量」「パラメータ量」が大幅に増加したことで,精度が格段に向上した][2]違いがある.

LLM に指示を与えるための入力をプロンプトといい, プロンプトにはテキスト（ここでは日本語による文章）を利用する.ChatGPT は人間の会話に近い形でコミュニケーションを行える.[言語モデル（LMs）を効率的に使用するためのプロンプトを開発および最適化するためのプロンプトエンジニアリングという学問分野がある][3].その中で, [ペルソナパターンとテンプレートパターン][4]を本実験では利用した.[ペルソナパターンは, LLM が特定の視点や視野を常に持って出力することを期待し, LLM に「ペルソナ」を与え, どのようなタイプの出力を生成し, どの詳細に焦点を当てるべきかを選択するのに役立てる][4]. 例えば,「あなたが上級エンジニアで,初級エンジニアにアドバイスをすることを想像してください」等である. [テンプレートパターンは LLM の出力構造の面で正確なテンプレートに従うことを保証するため][4], 出力形式を指定することをいう.

LLM には懸念や難しさもある.[ルールや知識に基づいて処理しているわけではない][5]ため,[数学や論理,事実関係や知識の問題についてはどこかに限界がある][5]し, [ハルシネーションといって, 一般性のありそうな回答など「もっともらしい嘘」をつくことがある][5]これより,実際に該当分野で評価を行うことは重要と考える.

3.　目的

クラウドセキュリティ監査（前述の内部監査および，該当規格をベースとした点検結果の審査の総称としてクラウドセキュリティ監査とする）を,LLM に任せることができるか,もしくはクラウドセキュリティ監査の補助ができるかを研究の目的とする.ただし,どちらにせよ最終責任は人間が負うものとし,知識のある人間のチェックは想定する.

LLM による監査は，特に不適合と判断するべき内容を不適合と判断できないのではないか,との仮説を立て,まずは適合/不適合の監査結果の評価をする.だが，適合/不適合の結果のみでは実際の成否は判断できず,人間が最終判断をするために根拠の提示が重要である.そのため,ChatGPT が出力した根拠の評価を行い併せて根拠の傾向を確認する.次に ChatGPT の失敗する傾向をしれば,人間の最終判断時に失敗しがちな根拠に注目できるため,失敗事例の分析を行う.最後は初回で正確な結果を出せなくとも追加プロンプトで失敗が減るかを考察する.

まとめると,本論文が答えようとする研究課題は以下である.

(1) 監査性能の評価：不適合が正解であるパターンで失敗が多いのではないか
(2) 根拠の評価：ChatGPT4 が根拠とする内容は一定の傾向があるのではないか
(3) 失敗事例の分析：ChatGPT と人間で失敗の傾向に違いがあるのではないか
(4) 改善評価：追加プロンプトによって,正解率は向上するのではないか

4.　実験

4.1. 実験内容

ISO/IEC27017 のクラウドサービスプロバイダ（サービス事業者）の「実施の手引き」の設問で,ChatGPT を用いて監査を行う.

同じ設問に対し,適合データと不適合データを用意し,適当/不適合を正しく判断できるか確認する.また,その判断の根拠も記載させ,その妥当性も確認する.

4.1.1. ツール

ChatGPT(GTP4)を利用した.

4.1.2. プロンプト

プロンプトは下記の要領で作成した.

(1) ISO/IEC27017 のクラウドサービスプロバイダの実施の手引きの設問で,ChatGPT を用いて監査を依頼する.

(2) 同じ設問に対し,適合データと不適合データを用意し,正しく適合/不適合を判断できるか確認する.

(3) ChatGPT に判断の根拠も記載させ,根拠の妥当性も確認する.

(4) 監査の結果,不正解だった場合は,追加質問を行い正しい結果に変化するかを確認する.

その他の条件は下記の通り.

(5) 各設問と適合/不適合の組み合わせでそれぞれは,新たな対話として質問する.これは，プロンプトを共通にすると，それまでの質問に ChatGPT の回答が影響を受ける可能性があるため，それを避ける狙いである．逆に追加プロンプトは前の質問を受けての回答を期待するため，該当質問の回答に対する返答とする.

(6) ペルソナとして監査員であることを伝える.これは監査員の立場でより厳密に監査結果を出すことを期待するためである.

(7) プロンプトに渡した文章からのみ判断するよう伝える.その理由は,前述の通り LLM は事実に基づかず,一般性がありそうな回答をする懸念があるためである.

(8) 実験結果を評価しやすいよう,出力テンプレートを定義した.

具体的なプロンプトは,下記の通りである.

あなたは IT 分野やクラウドサービス,セキュリティに詳しい監査員です.
とあるクラウドサービスについて,監査をしてください.
下記の【文章】から文末までで,【管理策】に続く文章に適合しているかを回答し,根拠を記載してください.
以下は回答フォーマットです.
◆適合・不適合：
◆根拠：
条件は以下です.
・【文章】の文からのみ判断してください.
・【文章】の内容は該当クラウドサービスから提供されている文書です.
【管理策】
＜設問を記載＞
【文章】
＜データを記載＞

4.1.3 データ

データは下記の通りである.

適合データは一般に公開されている利用約款や,サービス仕様書,Web ページなどから適合と判断できる文章を抽出して適合データを作成した. これは実験において, クラウドサービスの実際の文書を利用することで, より実践的な監査結果に近づけるためである. 出典元は付録「3.出典元」に記載した.

不適合データは下記の複数の方法で作成した.これらは実務でよくあるパターンを用意することでより実際に近づける狙いがある.

(1) 適合データから要点を削除する.
(2) 適合データを利用せず, 設問に似た機能の記載とする.
(3) 全く設問に関わらない内容の記載を抽出する.

(2)(3)は適合データと同じく, 一般に公開されている利用約款や,サービス仕様書,Web ページから抽出した.不適合データは設問の内容に近さごとに3つのレベルに分類した.

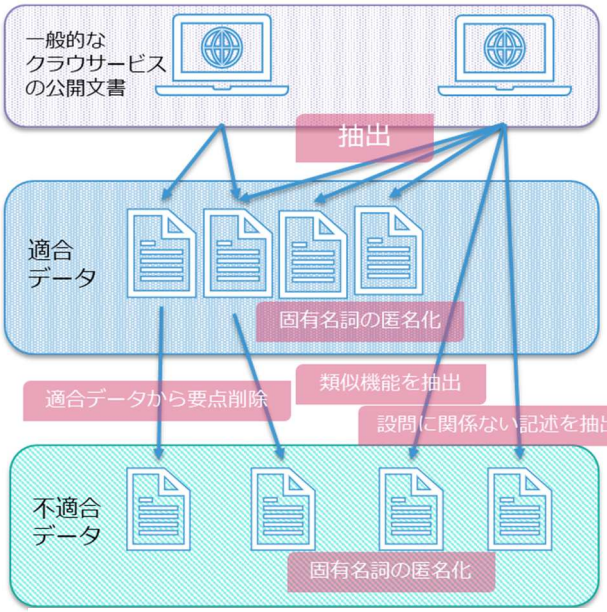


図-1. データ作成方法

データ作成時の考慮として,実在のサービスから流用しているため,判断が影響されないよう,固有名詞はダミー名称に変更した.前述の通り,LLM は事実に基づかず,一般性がありそうな回答をする懸念があるためである.

監査基準となる設問は ISO/IEC 27017:2015 で定義されている「実施の手引き」を利用する.ここでは参考文献[1]を参照した.極力,適合データと不適合データともに実際の利用約款・サービス仕様等を利用する方針としたため,設計書等の一般公開されておらず,社内でも情報資産管理上, 利用できないといったデータの準備が困難な設問は対象外とした.

4.2. 実験結果

4.2.1. 監査性能の評価

想定通り,不適合データ（正解が不適合）のパターンで,正解率が低い傾向にあった.
初回プロンプト時の結果は表 1 の通りである.
不適合データを研究課題にしているため,不適合を正例として記載する.
縦の「LLM 予測」は ChatGPT の結果,横の「正解」は実際のデータがどちらであったかを

示す.

表 1-監査結果①初回プロンプト

		正解		
LLM 予測		正：不適合	負：適合	合計
	-	TP (True Positive)	FP (False Positive)	-
	正：不適合	6	0	6
	-	FN (False Negative)	TN (True Negative)	-
	負：適合	10	16	26
	合計	16	16	32

正解率・適合率・再現率・特異率は表２の通り.

表 2-正答率・適合率・再現率・特異率②追加プロンプト

正解率	68.8 %	(TP+TN)/(TP+FP+FN+TN)
適合率	100.0 %	TP/(TP+FP)
再現率	37.5 %	TP/(TP+FN)
特異率	100.0 %	TN/(FP+TN)

正解率も 68.8%と高くないが,再現率が 37.5%となっており,不適合データを適合と判断しやすい傾向にある.また,適合率が 100%,特異率も 100%となっているため,全体的に適合と判断する傾向にある.

4.2.2. 根拠の評価

根拠は筆者自身が判定した.「拡大解釈」や「推測」をしてポジティブに適合と判断するという,一定の傾向があった.

根拠の評価結果は表 3 の通りである. 成功パターン（正解・LLM 予測ともに不適合,もしくは正解・LLM 予測ともに適合）は,正解が不適合の場合,根拠が適切なのは 66.7%,正解が適合の場合は 93.8%であり,監査結果が成功していても根拠が不適切であるケースが確認された.

表 3-根拠の評価

正解	LLM 予測	-	適切	不適切
不適合	不適合	TP	4	2
	適合	FN	0	10
適合	不適合	FP	0	0
	適合	TN	15	1
合計		-	19	13

次に失敗事例の根拠分類は下表の通りである.

表 4-根拠分類

	TP	FN	TN	合計
レベル感	0	1	0	1
厳密さの欠如	1	1	0	2
拡大解釈	0	4	1	5
推測	1	3	0	4
専門用語	0	1	0	1
合計	2	10	1	13

設問に対し機能的に合致していないが,広く捉え適合と判断する「拡大解釈」の傾向や,この機能があるならば,おそらく設問の機能もあるはず,という「推測」が,全体の 69.2%を占め,ポジティブに適合ととらえる傾向が見えた.

その他は,厳密さが欠如しているケース,専門用語の理解を誤っているケース, 実際的な記述はないが見出しがあるためレベル感を気にせず適合と失敗しているケースがあった.

4.2.3. 失敗事例の分析

不適合データの失敗事例については,筆者の判断により不適合データを分類し結果を集計した.分類について,レベルを付与し数字が大きいほど,不適合と判断しやすいデータ,つまり人間にとっての間違えにくいと考えて作成している.分類は下記通りである

1.一部不足：設問の求める内容の一部は満たしているが,一部は満たしていない
2.隣接機能：設問の求める内容に近い機能のデータだが, 明確に該当機能ではない
2.レベル違い：設問の求める内容に即した見出し一文の記載があるが本文詳細はない
3.包括概論：設問の求める内容・機能をごく少し含む全体的な説明をするデータである
3.内容乖離：設問の求める内容から乖離した内容のデータである

不適合データの失敗事例のデータ分類に傾向があるかを示したものが表 5 である.

表 5-正解：不適合のデータ分類別適合／不適合

	データ数	適合	不適合	適合/データ数
1.一部不足	4	2	2	50 %
2.隣接機能	5	3	2	60 %
2.レベル違い	1	1	0	100 %
3.包括概論	2	2	0	100 %
3.内容乖離	4	2	2	50 %

データ数に対し,失敗した割合が多いのは,「隣接機能」,「レベル違い」,「包括概論」だった.

人間と同じく分類の数値が大きい順になると考えていたが,ChatGPT は傾向が異なった.根拠分類であげた,「拡大解釈」や,「推測」により少しでもデータが設問に掠ると判断すれば失敗である適合と判断したようである.隣接機能や,レベル違い, 包括概論を用いた例が失敗事例について多かったものの,傾向と言い切るほどではなかった.

4.2.5. 改善評価（追加プロンプト）

正解率をあげるため,監査結果の失敗事例のうち不適合データ（正解が不適合）であるケースに対し,追加プロンプトを試行した.その結果は表 6,表 7 の通りである.

68.8%から 90.6%へ正解率は上昇, 適合率 37.5%からも 81.3%まで上昇した.

なお,1 件について不適合データ（正解が不適合）に「判断できない」という結果がある.失敗例と考え,「正解：不適合」「LLM 予測：適合」の扱いとする.

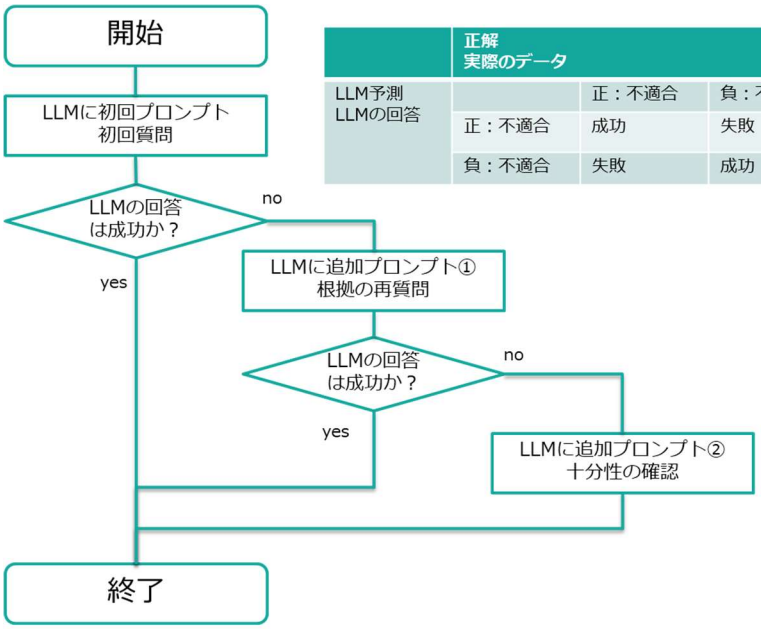
表 6-監査結果②追加プロンプト

		正解		
		正：不適合	負：適合 ※1	合計
LLM 予測	―	FP	TP	―
	正：不適合	13	0	13
	―	TN	FN	―
	負：適合	2	16	18
	（判断できない）	1	-	1
	合計	16	16	32

表 7-正答率・適合率・再現率・特異率

正解率	90.6 %
適合率	100.0 %
再現率	81.3 %
特異率	100.0 %

追加プロンプトの内容は,以下を実施した.



	正解 実際のデータ		
LLM予測 LLMの回答		正：不適合	負：不適合
	正：不適合	成功	失敗
	負：不適合	失敗	成功

図 2-追加プロンプトのフローチャート

(1) 初回プロンプトで失敗したケースについて，根拠と判断した記載箇所の再質問を行った．初回プロンプトでも根拠は確認しているが,改めて判断した箇所の回答を依頼した．

(2)(1)でも失敗となったパターンに対し，十分性の確認を依頼して再質問を行った．ただし十分性の確認は 80%を十分と考えるか 100%を十分と考えるかなど，考え方次第で不適合にできてしまう懸念があるため,積極的な活用は非推奨と考える．

根拠の再質問は，これまでの経験より改めて論点にフォーカスして再質問すると考えを変更するケースがあったため採用した．十分性については実務においても部分的に適合している箇所が見受けられ検討することが多いため，追加プロンプトとして採用した．

4.2.6. 実験結果からの考察

(1) 監査性能の評価として,「不適合が正解であるパターンで失敗が多いのではないか」の研究課題の結果は，不適合を正例とする実験で初回プロンプトでは適合率 100%，再現率が 37.5%となっており,傾向としては想定通り不適合を見抜く力が低かった．

(2) 根拠の評価として,「ChatGPT4 が根拠とする内容は一定の傾向があるのではないか」の研究課題の結果は，設問に対し機能的に合致していないが,広く捉え適合と判断する「拡大解釈」の傾向や,この機能があるならば,おそらく設問の機能もあるはず,という「推測」が,全体の 69.2%を占め,ポジティブに適合ととらえる傾向が見えた.監査で活用する上で,適合と判断した根拠が，監査対象文書に記載していないことをより広く捉えすぎていないかを注意する必要性がある．本論文では未実施だが，推測や拡大解釈をプロンプトで禁じてみる，といったプロンプトの工夫の余地はあるかもしれない．

(3) 失敗事例の分析として「ChatGPT と人間で失敗の傾向に違いがあるのではないか」の研究課題の結果は，人間が判断しやすい不適合データ分類と ChatGPT が不適合としやすいデータ分類は一致しなかった．特に「包括概論」は一般的なセキュリティ全般の概要を記載した文書であり，それを適合判断をするのは，人間の判断と大きく異なるところと考える．これは「拡大解釈」や,「推測」により少しでも文章が設問に関係があれば適合と判断しやすいのではないかと推察した．ただ，不適合データ分類での全体的な傾向を言い切るほどの傾向は見つけられなかった．

(4) 改善評価として「追加プロンプトによって,正解率は向上するのではないか」の研究課題の結果は初回プロンプトでは 68.6%の正解率で数値がよくないものの,追加プロンプトで 1，2 回すれば 90.6%は正解にたどり着いた．監査員の技能に頼らず，単純に改めて

根拠を問い直すことで，ある程度の正解率の改善が見込まれた

今回の実験では ChatGPT の文字数制限を考慮し筆者が該当しそうな箇所を探し,LLM に投入している．それは業務効率化点では好ましくないため，監査自体でなく,監査を補助する活用を目指す方がよいように考える.だが，実業務では適合とすべき内容を見落とすことが怖いが，ChatGPT は広く適合といえる可能性のある記述を根拠として絞って提示してくれるので，その中で最終判断を人間が下すという意味では省力化になる部分はあると考える．

半面,実験外の課題が 1 点,効率上の大きな制約が 1 点ある．課題は実際の監査対象データを ChatGPT に渡せるかのセキュリティポリシー上での課題,制約は文書量がやはり ChatGPT で扱える以上に多い点(事前に渡すデータにあたりをつける必要がある)である．

将来課題は 3 点ある．1 点目はより実践的な利用手法の検討・提案である．実験外の課題や効率上の制約等から，現時点で実業務上での活用が即座にできない内容にとどまっている．2 点目はペルソナの設定是非に応じた正確性への影響確認等の実験の精査である．3 点目は参考文献[6]のように主観的な受け止め方を評価することである．主観的な観点の調査事例として[IT 技術 QA サイトとの比較で ChatGPT の回答は 52%が誤りで 77%が冗長だが，利用者は 39%の確率で誤情報を見逃すが，35%の確率で ChatGPT を好む][6]と報告されているため，主観的な受け止め方についても検討の価値がある．

4.3. 妥当性への脅威

メガクラウドなどの整備されたサービス文書から不適合データを作成しているため,実は不適合時の推測は的を射ている.より適合の可能性が考えられる文章であることから,不適合の正解率が落ちた可能性がある.小規模サービスや自社サービスのレビュー過程における未完成のものを利用すればより実務に即した評価になる．

5.おわりに

「クラウドセキュリティ監査を LLM に任せることができるか,もしくは監査の補助ができるか」については,不適合を見抜く力が低く,拡大解釈や推測などをして,ポジティブに適合と判断する傾向にあることを留意した上で，根拠を確認する追加プロンプトを与えながら，監査の補助として使用すればよいと考える．

今回,データは公開データを利用したが秘密度の高いデータを使える環境が作れれば,より実務に即した監査性能を把握できると考えられる．

また本論文は専門知識なく活用できる可能性から汎用的な AI をそのまま利用した．だが，専用 AI の作成や ChatGPT にファインチューニング（追加データを与えて訓練）し，クラウドセキュリティ監査を行う方式で可能性もあると考える．

6.参考文献

[1]羽田 卓郎 (著，編集) 山崎 哲 (著) 間形 文彦 (著)中尾 康二 (監修)，ISO/IEC 27017 クラウドセキュリティ管理策と実践の徹底解説,2017

[2] 株式会社　日立ソリューションズ・クリエイト，大規模言語モデル（LLM）とは？仕組みや種類・用途など,https://www.hitachi-solutions-create.co.jp/column/technology/llm.html

[3] Prompt Engineering Guide，https://www.promptingguide.ai/jp

[4] Jules White,A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT,2023

[5]石川　冬樹,(ChatGPT 時代の)AI 品質のはじめかた,2023

[6]Samia Kabir,″Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions″, arXiv 2308.02312