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

一般財団法人日本科学技術連盟
第 39 年度（2023 年度）ソフトウェア品質管理研究会 成果発表会
研究コース 5人工知能とソフトウェア品質 LLMによる監査チーム
2024 年 3 月 8 日（金）
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）
副主査:栗田 太郎（ソニー株式会社）

目次

自己紹介

ランプの魔人のCM
やってます

- **所属**
- **　TIS株式会社**
- **　品質監査室　チーフ**
- **業務**
  - 社内の提供・利用クラウドサービスのセキュリティ審査
  - 審査を通じて、品質向上を図る
- **経歴**
  - 情報システム部（2003/4～2021/11)
    - 情報系システムの企画・導入・運用管理
  - 品質監査室（2021/12～)

TISのご紹介

■ 会社概要

（2023年7月1日現在）

社　　名	TIS株式会社 （TIS Inc.）
創　　業	1971年4月28日
設　　立	2008年4月1日
資 本 金	100億円
代 表 者	代表取締役社長　岡本　安史
本　　店	東京都新宿区西新宿8丁目17番1号
従業員数	連結：21,946名　単体：5,695名 （2023年3月31日現在）
売 上 高	連結：508,400百万円　単体：238,140百万円 （2023年3月期）
営業利益	連結：62,328百万円　単体：29,450百万円 （2023年3月期）

認定資格

- 総務省「届出電気通信事業者登録」
- 経済産業省「情報セキュリティサービス基準適合サービスリスト」
  「情報セキュリティ監査サービス」
- 経済産業省「システム監査企業台帳登録」
- 情報セキュリティマネジメントシステム(ISMS)(ISO/IEC27001)
- ITサービスマネジメントシステム(ITSMS)(ISO/IEC20000-1)
  [認証対象範囲]
  東京第3センター/東京第4DC/大阪第2DC/大阪第3DC/大阪第4DC
- 品質マネジメントシステム(QMS)(ISO9001)
- プライバシーマーク使用許諾事業者
- 東京都「一般建設業（電気通信工事）」
- 環境マネジメントシステム(ISO14001:2015)

■ 特長

TISは3,000社以上のビジネスパートナーとして「成長戦略を支えるためのIT」を提供

TISの50年の実績が裏付ける、高度な実現力と先進性

TISの提案力と課題解決力を支えるのは200を超えるサービスメニュー

“攻める”“やり切る”現場力、人材力

論文テーマ

**クラウドセキュリティ監査を、LLMに任せることができる
かもしくはクラウドセキュリティ監査の補助ができるか**

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

背景-クラウドセキュリティ監査

ISO/IEC27017とは

- [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] [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
- 要は情報セキュリティマネジメントシステム（ISMS）の仲間でISMSで手薄なクラウドサービス特有のセキュリティに関する規格

所属先での取り組み・課題

- 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
- 特徴：多量の文書（利用約款、サービス仕様書、設計書等）を読む
- 課題：省力化・効率化
  - 設問数100程度の審査を年間約200件、少ない人手で実施するため

用語

- 適合（≒合格）・不適合（≒不合格）

背景ーLLM

- **LLM＝Large Language Model、大規模言語モデル**
  - [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] [2]
  - プロンプトと呼ばれる文章を渡して、結果を得る
- **知ってほしいこと**
  - [ルールや知識に基づいて処理しているわけではない][5]
- **懸念**
  - [数学や論理、事実関係や知識の問題については限界がある][5]
  - [ハルシネーションといって、「もっともらしい嘘」をつくことがある][5]

[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

[5]石川　冬樹、(ChatGPT時代の)AI品質のはじめかた、2023

↓

- **分野別で実際に実験を実施する意義があると考える**

研究課題

下記の研究課題を設け、ChatGPTで実験を実施

(1) 監査性能の評価:

- (仮説)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:

- 2択の回答だけでは評価できないため、根拠を確認
- (仮説)ChatGPT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:

- (仮説)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:

- (仮説)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

実験ー手順①

1. ・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
2. ・同じ設問に対し、適合データと不適合データを用意する。
3. ・監査結果に加え、根拠を質問する
4. ・監査結果が失敗だった場合は、追加質問を行い正しい結果に変化するかを確認

実施の手引きの例

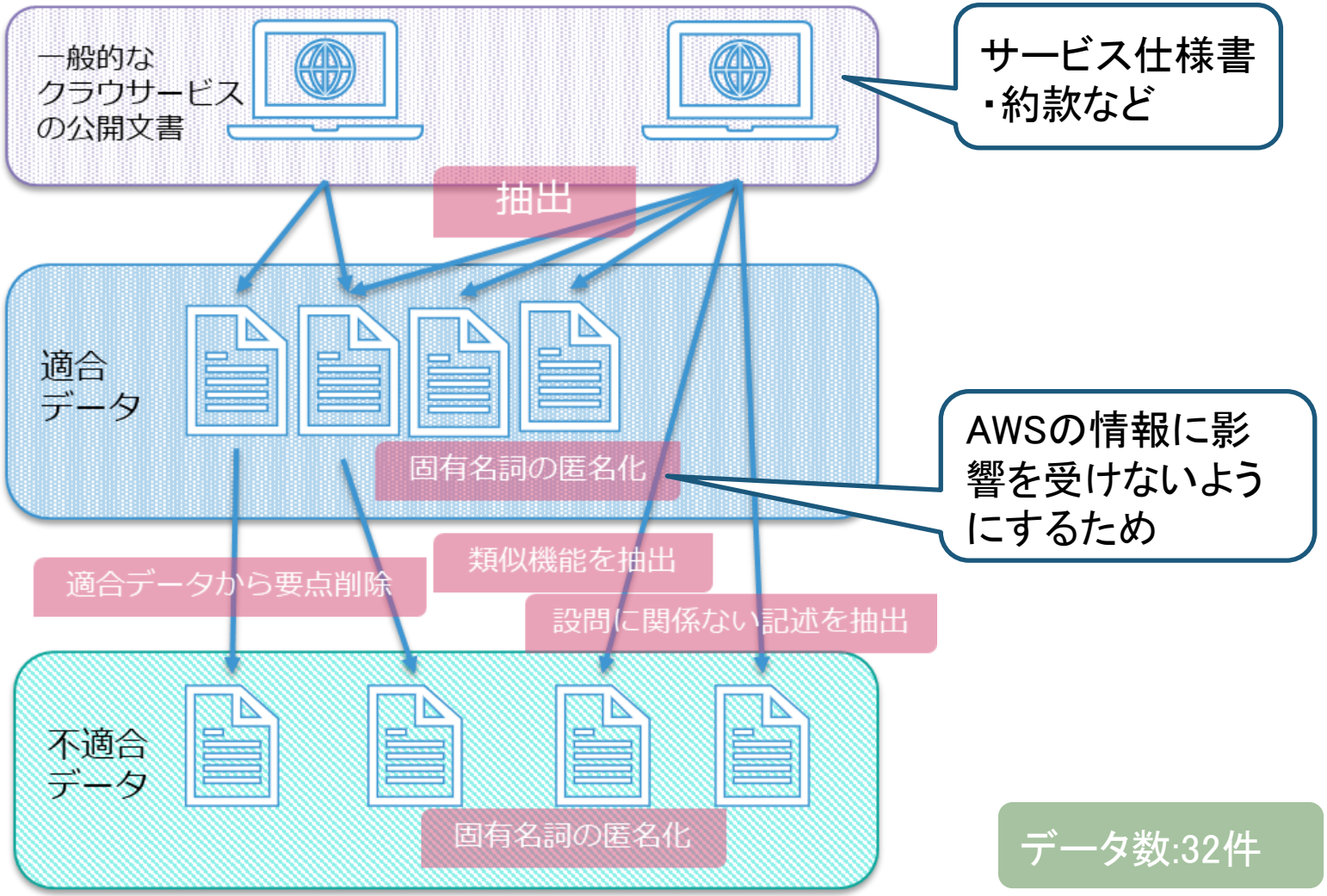
・[CLD9.5.1 仮想コンピューティング環境における分離]
  ・クラウドサービスカスタマ間のリソースの分離や、
  ・クラウドサービスカスタマのリソースからクラウドサービスプロバイダの内部管理の分離

実験ー手順②

プロンプト例

- 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＞

実験ーデータ作成



対象外※図や表のある文書

実験ー用語説明

- **用語説明（正例を不適合）**

LLM予測	正解（実際のデータ）	
	正：不適合	負：適合
正：不適合	TP（True Positive）	FP（False Positive）
負：適合	FN（False Negative）	TN（True Negative）

  - 正解率：全体のうち、正解（実際のデータ）とLLM予測が一致しているもの
  - 適合率：LLM予測が正（不適合）のもののうち、実際に正であった率をさす.
  - 再現率：実際のデータが正（不適合）であったもののうち、LLM予測も正（不適合）とした率をさす
  - 特異率：LLM予測が負（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

実験ー結果(1)監査性能

- **実験結果**
  - 監査性能の評価：
    - （仮説）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
      - ここでは正例を不適合としている
  - 結果
    - 不適合データ（正解が不適合）のパターンで正解率が低い
    - 全体的に適合と判断する傾向にある.

ChatGPTが不適合と判断したものはすべて不適合

正解率	68.8 %	全体のうち、LLM予測と正解が一致した率
適合率	100.0 %	LLM予測が正（不適合）のうち、実際のデータが、正（不適合）だった率
再現率	37.5 %	実際のデータが正（不適合）であったもののうち、LLM予測も正（不適合）とした率をさす
特異率	100.0 %	LLM予測が負（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

本当は不適合のものを60%程度、適合と判断してしまった

実験ー結果(2)根拠の評価①

- **実験結果**
  - (2) 根拠の評価：
    - ChatGPTが根拠とする内容は一定の傾向があるのではないか
  - 結果
    - 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
      - 正解が不適合の場合、根拠が適切なのは66.7%であった。
    - 根拠分類の結果、「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。

表4-根拠分類

	合計
レベル感	1
厳密さの欠如	2
拡大解釈	5
推測	4
専門用語	1
合計	13

実験ー結果(3)失敗事例の分析

- **実験結果**
  - (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
  - 手順・結果
    - 人間と同じ傾向は示さなかった
    - 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

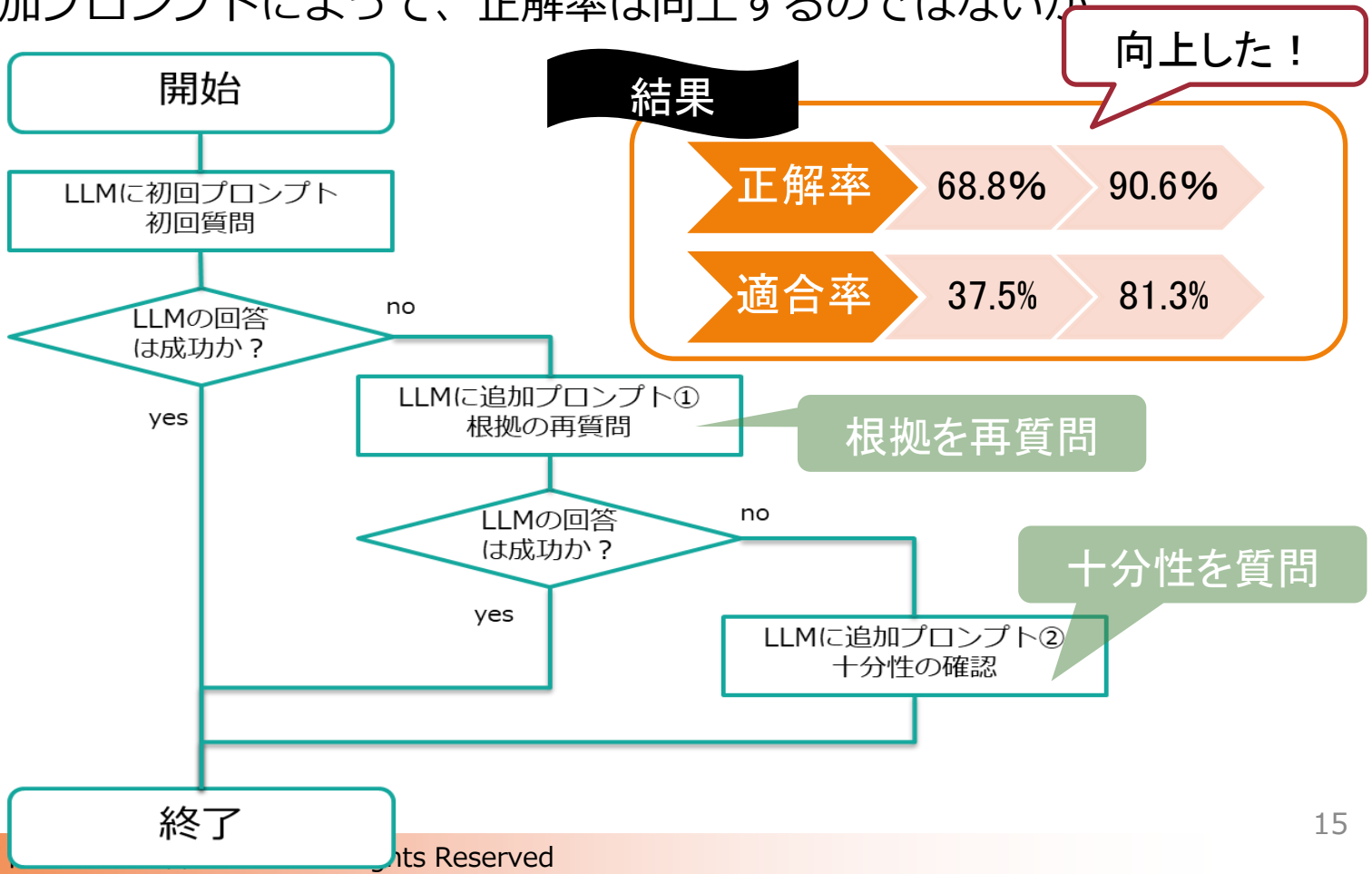
大きい数字の方が人間は不適合と判断しやすい

表5-正解：不適合のデータ分類別適合／不適合

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

実験ー結果(4) 改善評価

- **実験結果**
  - (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



考察①

(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」
  - 適合率100%、再現率が37.5%
  - 傾向としては不適合を見抜く力が低かった

(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」
  - 「拡大解釈」「推測」が全体の69.2%
  - ポジティブに適合ととらえる傾向
- 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、より広く捉えすぎていないかを注意する必要性がある

考察②

(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」
  - 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
  - 「拡大解釈」や、「推測」により少しでも文章が設問に触れると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までは見つけられなかった

(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」
  - 初回プロンプトでは68.6%の正解率が、90.6%に向上
  - 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

考察③

総論

- 「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」
  - 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

課題・制約

- 実験外での課題
  - 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
- 効率上の制約
  - 現在、ChatGPTで扱える文書量が制限がある（事前に渡すデータにあたりをつける必要がある）

考察④

将来課題

- より実践的な利用手法の検討・提案
- 実験の精査
  - ペルソナの設定是非に応じた正確性への影響確認等
- 利用者の主観的受け止め方の検討
  - [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]

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

謝辞

- 本論文の作成にあたり、終始多大なご指導を賜った石川冬樹主査、徳本晋副主査、栗田太郎副主査には大変感謝申し上げます。
  実験実施、論文作成に不慣れな私に、基本的なことから丁寧にご指導いただき大変感謝しております
- 査読いただいた先生方も大変ありがとうございます。

ご挨拶

- ご清聴ありがとうございました！