大規模言語モデルと知識グラフに基づく AI エージェント型サイバーセキュリティ教育システムの開発

研究代表者　　東野正幸　　国立大学法人鳥取大学 工学部 電気情報系学科 准教授

1 研究目的

大学や企業をはじめ，多くの組織において，標的型攻撃が脅威となっている．標的型攻撃は，攻撃対象の個人や組織を熟知した上で仕掛けられる場合が多いため，セキュリティソフトウェアやファイアウォールといったシステムセキュリティだけでは防ぐことが難しい．システムセキュリティを潜り抜けた標的型攻撃に対応できるか否かは，個人や組織の能力にかかっている．このため，サイバーセキュリティに関する訓練や教育等の人・組織的セキュリティの重要性が高い．

生成 AI の悪用によって高度な攻撃が容易になることが予想される．このため，一般的な組織の構成員であっても，対応するには確かなセキュリティ知識を備える必要がある．しかし，大学などの多くの教育機関において，サイバーセキュリティに関する授業が開講され始めたのは近年のことであり，多くの教職員や会社員はサイバーセキュリティに関する体系的な知識を備えていない．また，変化が極めて激しい分野であるため，継続的に知識を更新しなければならないが，最新の対策方法に追従することも難しい課題がある．

生成 AI の悪用により標的型攻撃が高度化する懸念がある一方で，組織の構成員に対して効率的なサイバーセキュリティ教育を行う手法は確立されていない．

そこで本研究では，この課題を解決するための大規模言語モデルと知識グラフを用いた AI エージェント型のサイバーセキュリティ教育システムを開発することを目的とする．

2 【課題 1】サイバーセキュリティに関する知識グラフの構築とプロトタイプの開発

2-1 概要

サイバーセキュリティ教育においては，一般に組織の構成員の理解度が不均質であるため，個々の理解度に応じた学習内容を提供することが望ましい．組織内の講習や研修では，学習に要する時間が業務負担となる場合があり，短時間で学習できる方法が求められる．そこで本研究では，サイバーセキュリティの攻撃手法と対策手法の知識グラフを構築し，これを活用して個々の理解度と学習コンテンツの難易度に応じた学習コンテンツの回答順序を提供し，さらに共通する知識のうち過去の回答から習得済みと判断できる項目を除外することで，効果的かつ効率的な学習を支援するシステムを開発する．

2-2 はじめに

サイバーセキュリティ攻撃は，システムセキュリティや物理セキュリティだけで完全に防ぐことが困難である．そのため，教育や訓練といった人的セキュリティも重要である．例えば，標的型攻撃メールは，メールゲートウェイやメールソフトウェアのセキュリティ機能といったシステムセキュリティだけでは完全に防ぐことは難しく，すり抜けた標的型攻撃メールへの対応として，訓練や教育等の人的セキュリティの重要性が高い[1]．

また，近年では生成 AI を悪用したサイバーセキュリティ攻撃のリスク[2]や，生成 AI の活用に伴うリスク[3]，[4]，[5]が懸念されている．例えば，攻撃者によるフィッシングメールの作成においては，V-Triad[6]と呼ばれるフィッシングメールを設計するためのルールセットと生成 AI を併用することで，より多くの人を騙すメールを容易に生成できることが示唆されている[7]．そのため，人的セキュリティ対策に必要な知識や技術が高度化及び複雑化することが予想される．

一方で，組織におけるサイバーセキュリティ教育においては，構成員のサイバーセキュリティに関する理解度が一様ではなく，それぞれの役割や習熟度に応じた学習内容を提供することが求められる[8]．また，組織内の研修においては，学習にかかる時間が業務負担となる場合があり，短時間で効果的に学習できる方法が必要とされている．

そこで本研究では，知識グラフを用いた個別適応型のサイバーセキュリティ学習システムを開発する．サイバー攻撃の手法と対策手法を体系化した知識グラフを構築し，学習者の理解度と学習コンテンツの難易度

公益財団法人電気通信普及財団
研究調査助成報告書 第41号 2026年度

に応じた学習順序を提示する．また，学習者が過去の回答からすでに習得済みと判断できる知識を除外して不要な学習を削減する．

提案システムにより，個人に合ったサイバーセキュリティに関する学習が行えるだけでなく，学習に取り組む時間の短縮も可能とし，効果的で効率的なサイバーセキュリティ教育の実現を目指す．

なお，本節ではサイバーセキュリティ攻撃の例として，フィッシング攻撃を題材として取り扱う．

2-3 提案手法

サイバーセキュリティ攻撃の例として，フィッシング攻撃が挙げられる．一般的にフィッシング攻撃の対策としてメールの差出人が正しいことの確認やメールに記載されたURLが正しいものであることの確認などが挙げられる．しかし，そのような対策方法を伝えて対応できる人もいれば，どのように対応すれば良いか分からない人もいる．この差はフィッシング対策に関する基礎的な知識の差だと考えられる．つまり，学習者が学習内容に関する基礎知識を持っていればそれを理解できる．また，攻撃手法への対策に必要な知識の数が多いほど理解が難しくなる．このことから，個人の理解度に合った学習難易度の問題を優先する．

学習においては苦手な問題を優先することで教育効果が高い事例が報告されている[9]．まだ習得できていない基礎知識が多く含まれる問題ほど苦手であり，さらに，1 回の回答でより多くの知識を習得できると考えられる．そのため，未習得の基礎知識が多く含まれる問題を優先する．

以上より，学習者に合った難易度の問題を優先し，さらにより多くの知識を習得できる問題を優先して学習を進める．

（１） 知識グラフの構築

フィッシング対策として何を学習するべきなのかを明らかにし，その知識間の関係性を知る必要がある．そこで，NIST Phish Scale User Guide（TN 2276）[1]を情報源とした知識グラフを構築する．

攻撃手法に関するノードとして，Phish Scaleで定義されている「フィッシングメール自体の観察可能な特徴」の「Cue Type」（手がかりの種類）のノードを作成する．このノードを起点に「Cue Name」（手がかりの名称）のノードを作成して接続する．

NIST Phish Scale User Guide（TN 2276）[1]に記載されている「フィッシングメール自体の観察可能な特徴」のうち，「Technical indicator」（技術的指標）に関する「Cue Type」（手がかり種類）と「Cue Name」（手がかりの名称）を以下に引用する．

- Technical indicator
    - Attachment type
    - Sender display name and email address
    - URL hyperlinking
    - Domain spoofing

これらに続くノードは独自に作成する必要があり，攻撃手法に関する学習項目であるフィッシングに用いられる具体的な攻撃手法のノードを独自に作成して接続する．

さらに，各攻撃手法の学習に必要な基礎知識や攻撃手法ノード間の知識の共有関係を知るために，具体的な攻撃手法のノードの下にサイバーセキュリティに関連する知識ノードを独自に作成して接続する．

図 2-1 に構築した知識グラフを示す．Phish Scaleの「Cue Type」のうち，茶色の「Technical Indicator」を起点として，黄色の「Cue Name」にリンクする．さらに薄い赤色の具体的な攻撃手法に細分化し，その周囲に濃い赤色の基礎知識をリンクする．Phish Scaleで定義されているものは黄色の「Cue Name」までであり，それ以降は独自に作成した内容に基づいている．

（２） 学習項目選択関数

式1の目的関数を最小にするのに最も効果的な学習項目から順に学習する．

$$\sum_{n=1}^{N} a_n \cdot \frac{\alpha - \left|L_{a_n} - L_i\right|}{\alpha} \cdot \frac{1 + K_n + \sum_{m=1}^{N} \alpha_m P_{nm}}{\beta} \quad (1)$$

α は学習項目に設定する難易度の最大値，β は $1 + K_n + \sum_{m=1}^{N} \alpha_m P_{nm}$ の取り得る値の最大値とする．Nは学習項目の個数とする．a_n,$(1 \leq n \leq N)$は学習状況を表し，a_n=1であれば学習が終わっていない項目（初期値）であり，a_n=0であれば学習が完了した項目を表す．K_n は学習項目 a_n のみに関係する知識の個数とする．P_{nm} は2つの学習項目 a_n と a_m が共通して持っている知識を1つ経由して繋がる経路の数とする．n=mのときは0とする．L_{an} は学習項目 a_n の難易度とし，必要な知識の数が少ないものから順番に 1,2,3,･･･と付番す

る．L_i(1 ≤ L_i ≤ α)は学習者の習熟度とする．

i を学習回数とし，R_i を学習結果とする．学習を行うたびに，習熟度 $L_{i+1} = L_i + R_i$ により更新する．学習回数が i 回目の学習項目について理解できたと判断したとき，$R_i = 1$ とし，習熟度を $L_{i+1} = L_i + R_i$ により更新し，学習状況を $a_n = 0$ として学習済みとする．学習回数が i 回目の学習項目について理解できていないと判断したとき，$R_i = -1$ とし，習熟度を $L_{i+1} = L_i + R_i$ により更新し，学習状況を $a_n = 1$ として学習済みとする．

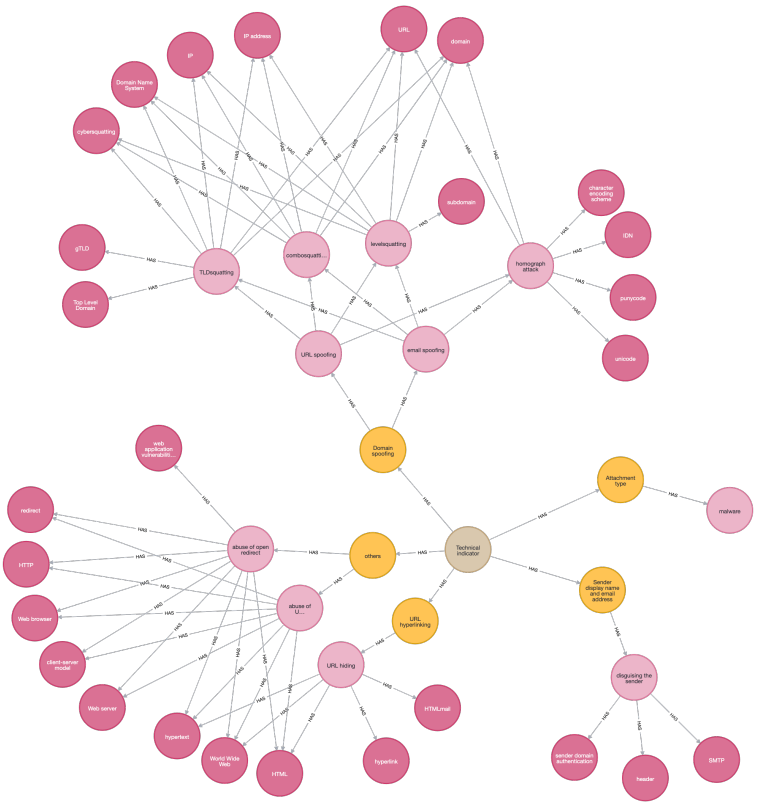


図 2-1: NIST Phish Scale User Guide (TN 2276)を情報源とし，Technical Indicator を起点とする知識グラフ

2-4 実験

提案手法における個人適応に関する評価として，実験参加者ごとの学習項目の回答順を記録し，実験参加者の習熟度（L_i）と出題した学習項目の難易度（L_{an}）がどの程度一致しているかを確かめる．また，アンケートによる評価として，実験終了後に「難易度が調整される感覚はあったか.」等を項目とする主観評価を実施する．

（１） 実験の方法

実験はゲームブック形式による筆記試験で行う．提案手法に基づき，各学習項目の番号，問題文，正答，及び解説文を作成する．また，提案手法の知識グラフと評価関数が実装されたプログラムを用意する．実験参加者はこのプログラムが提示する番号の問題に取り組み，その回答をプログラムに入力すると，正誤判定が行われ，その結果に応じて次の問題の番号が提示される．各学習項目の問題に全て正答するまで回答を繰り返す．

プログラムは,学習項目として定めた攻撃手法に関する問題を評価関数によって選ばれた順番に出題する．出題された問題に正解すると，その学習項目について学習が完了したとみなし，それ以降の出題候補から除外する．間違えた場合は，その問題の解説文と関係する知識項目の説明文を提示することでフィードバックとする．また，間違えた問題は学習がまだ完了していないとみなし，それ以降の出題候補として残す．このように学習を繰り返し，全ての問題に正解することで学習終了とする．また，学習終了後にはアンケートを行う．

（２）　　実験の結果と考察

提案手法を用いて 10 名の実験参加者が学習を行った．提案手法を用いたゲームブックによる学習結果について，実験参加者の習熟度（Proficiency）と学習項目の難易度（Question Difficulty）の平均と標準偏差の推移を**図 2-2** に示す．学習の序盤から中盤については，出題された学習項目の難易度が実験参加者の習熟度に近い値を維持していることから，個人適応した学習ができていると言える．また，学習の終盤については，本実験では全ての学習項目について正答するまで学習を続けることとしていたため，最後まで残った難易度の低い学習項目が出題されている．このように，実験参加者の習熟度に対して，難易度の低い学習項目しか残っていない場合，それ以降の学習を打ち切ることで，総学習時間を短縮する判断も可能であると考えられる．また，各問題の難易度と回答に要した時間の平均と標準偏差を**表 2-1** に示す．問題と難易度と回答に要した時間には相関がなく，簡単な問題であっても長時間要する場合と要しない場合が観察された．

また，アンケートの結果については，「難易度が調整されている感覚は有ったか.」という設問に対して，4 名からは「有った.」，6 名からは「無かった.」という趣旨の回答が得られた．また，「有った.」群と「無かった.」群との学習項目の正誤率に大きな違いは無かった．また，「特に URL とドメインに関連する問題が難しく感じた.」という意見も得られた．提案手法では，学習項目の難易度として，その学習項目の習得に必要な基礎知識の数を用いている．専門知識については十分簡単な基礎知識まで分解していることを前提としているが，例えば URL やドメインは基礎知識といっても複雑な仕様が定められており，初学者にとっては難しい内容であった可能性がある．このため，基礎知識についてはさらに細分化を行う必要があると考えられる．

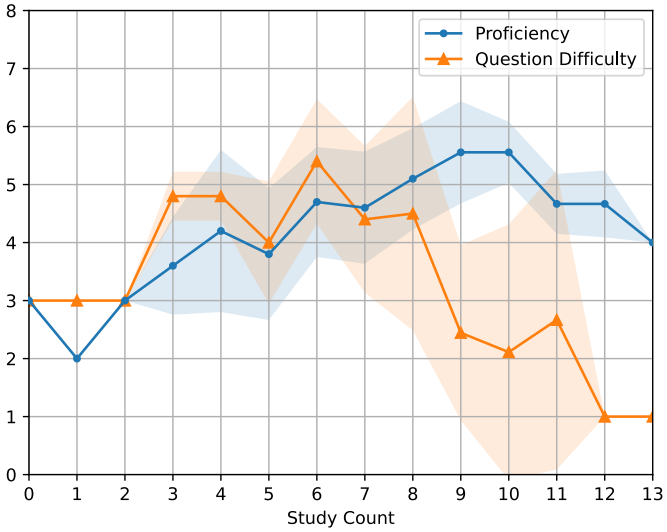


図 2-2: 実験参加者の習熟度（Proficiency）と学習項目の難易度（Question Difficulty）の平均と標準偏差の推移

表 2-1: 問題の難易度と回答に要した秒数の平均と標準偏差

問題番号	難易度	平均	標準偏差
1	1	31.2	26.6
2	2	19.2	8.9
3	5	13.3	10.4
4	3	21.5	36.1
5	4	14.9	10.9
6	3	14.8	10.8
7	5	21.5	11.6
8	6	28.9	23.6

2-5 おわりに

本節では，知識グラフを用いた個人適応型のサイバーセキュリティ学習システムの検討を行った．標的型攻撃メールを題材としてNIST Phish Scale User Guide（TN 2276）[1]からフィッシングに関する攻撃手法と防御手法に関する知識グラフを構築し，それらの知識グラフに基づき作成した学習項目の正答に必要な基礎知識の数から学習項目の難易度付けを行なった．また，学習者の習熟度，学習項目の難易度，及び関連する基礎知識の関係から学習項目の学習優先度を算出し，それらに基づき回答する学習項目の順番を決定する手法を提案した．

提案手法のプロトタイプとしてゲームブックを作成し，実験参加者 10 名による学習を実施した．その結果，実験参加者の習熟度に応じた難易度の問題を出題できることを確認した．また，実験参加者の習熟度が十分に高く，かつ全ての未着手の学習項目の難易度が低い場合に，それ以上の学習を打ち切ることで，総学習時間を短縮するための判断材料を示した．

今後は，提案手法を用いた場合と用いなかった場合での比較実験を行う予定である．また，今後の課題として，本節では扱わなかったNIST Phish Scale User Guide（TN 2276）[1]におけるフィッシングメール判別の評価基準（Criteria for Counting Cues）の他の手がかりの種類（Cue Type）や，標的型攻撃メールにおけるフィッシング以外の脅威に対する学習に提案手法を適用可能かどうか検証することが挙げられる．また，大規模言語モデルを用いて知識グラフと問題，回答，及び解説の自動生成を行うことで，本節では扱わなかった他の攻撃手法と対応手法にも容易に対応可能とすることを検討している．

3 【課題 2-1】大規模言語モデルを用いた組織規定に基づく役割別情報セキュリティ学習コンテンツ選択手法の検討

3-1 概要

組織の情報セキュリティ教育においては，組織が定める情報セキュリティ規定に基づき役割ごとに求められる知識と技能の範囲が異なることから，一律的な教育では各役割を割り当てた構成員が必要となる力量を効率的に確保することが難しい．このため，各役割に必要な学習範囲を特定し，役割に応じた適切な学習コンテンツを提示することが望ましいが，教育を実施する側の負担が大きく，十分に実施できていない課題がある．そこで本研究では，組織が定める情報セキュリティに関する規定と役割の定義に基づき，既存の学習コンテンツのスーパーセットを対象として，LLM-as-a-Judge 手法を用いて役割と学習コンテンツとの適合性を判定し，役割に応じた学習コンテンツを選択して提示する手法を検討する．さらに，実際の規則，役割，及び学習コンテンツのスーパーセットを用いて，本手法による選択精度を評価し，その結果を報告する．

3-2 はじめに

現在，サイバー攻撃は増加の一途をたどっており，その手口は巧妙化・多様化している．Check Point Research の記録によると，2025 年第 1 四半期の全世界のサイバー攻撃の週平均攻撃件数は，2024 年同期から 47 %増加している[10]．

サイバー攻撃の脅威に備えるために，個人，組織ともにセキュリティ対策する必要がある．情報セキュリティには，人的セキュリティ，技術的セキュリティ，及び物理的セキュリティの 3 つがあり，どれも欠けてはいけない[11]．その中でも本研究では，人的セキュリティに焦点を置いている．

情報セキュリティマネジメントシステム（ISO/IEC 27002:2022）では，情報セキュリティに関する教育や訓練を計画的に実施することが必要とされている．そして，これらの教育を通して，組織の構成員の役割に応じて必要な力量を保有することが求められている[12]．

そのため，情報セキュリティに関する教育において，個人の役割に適した内容の教育を提供することが求められているが，役割ごとに必要となる問題を提示することは大きなコストを要することから容易ではない[13]．

そこで本研究では，既存の情報セキュリティに関する教材から，学習者が役割を達成するために必要な学習コンテンツを判定・選択する手法を提案する．提案手法により，役割に必要な知識に絞った情報セキュリティ教育が可能となることが期待される．

3-3 提案手法

（1） 概要

情報セキュリティに関する知識の範囲は広く，それぞれの役割において全ての知識が必要というわけではない．そのため，組織全体で共通の学習コンテンツを用いたとしても，人によっては無駄な問題を出題して

しまっている場合もある．また，組織の構成員が業務を行いながら，情報セキュリティに関する十分な学習時間を確保することは難しい．このことから，役割に応じて効率よく学習できる教材を選択する手法が重要である．

本研究では，学習者が学習を始めるまでに，学習者の役割に適応した学習コンテンツを選択して提供することを実現する．そのために，組織規定，役割，スーパーセット，及びサブセットを次のとおり定義する．

- **組織規定**：組織ごとに異なる情報セキュリティに関する規定文．組織が規定する情報セキュリティに関する規則，規約，及びガイドライン等を指す．
- **役割**：情報セキュリティマネジメントシステム（ISO/IEC 27002:2022）[12]が定義する「role」を指す．
- **スーパーセット**：既存の情報セキュリティに関する学習コンテンツのセットを指す．本研究では，独立行政法人情報処理推進機構（IPA）が公表している情報処理技術者試験の過去問題[14]のうち，情報セキュリティに関連する選択問題をスーパーセットとして選定済みとする．
- **サブセット**：スーパーセットから役割に関連する問題を選択して作成された学習コンテンツのセットを指す．

提案手法では，まず組織規定，役割，及びスーパーセットを準備する．次に，組織規定，役割，及びスーパーセットを含むプロンプトにより，組織規定，役割，及びスーパーセットの問題の関連性に基づき適合度とよぶスコアを付与するように指示を大規模言語モデル(LLM)に対して行う．これらのLLM-as-a-Judge（LaaJ）[15]により適合度を付与し，適合度の高い問題を抽出することで，その役割向けのサブセットを構成する．

（２） LaaJ によるスコアリング

提案手法では，大規模言語モデル（LLM）を用いたLLM-as-a-Judge（LaaJ）により，スーパーセットに含まれる個々の学習コンテンツが指定した役割に対して学習すべき内容として相応しい学習コンテンツであるかを数値で示す適合度により評価させる．

スーパーセットに含まれる学習コンテンツと指定した役割と，その役割に関連する説明文の組に対して 0 から 100 の数値を用いて評価し，これを適合度とする．適合度が高いほどサブセットへ含めるに適している．また，評価基準として 100 から 76，75 から 51，50 から 26，25 から 1，及び 0 の 5 段階で区分を設ける．

（３） スコアに基づく問題の選択

LaaJ によるスコアリング結果をまとめたスコア表を作成する．スコア表には，役割に対応した組織規定と学習コンテンツの適合度が記載されている．そして，閾値を指定して閾値以上の評価値を持つ学習コンテンツを選択したサブセットを作成する．なお，使用するプロンプトについては，**図 3-A1**，**図 3-A2**，及び**図 3-A3** に示す．

3-4 評価実験

（１） データセット

組織規定として，鳥取大学情報システム運用管理要項[16]を使用する．また，組織の役割として，鳥取大学情報システム運用管理要項に記載されている部局情報技術責任者を使用する．さらに，同要項に記載されている役割として部局情報技術責任者に関する規則項目のうち 20 項目を無作為抽出する．なお，原文は PDF 形式で公開されているが，前処理として機械可読性のある Markdown 形式[17]へ変換する．

学習コンテンツのスーパーセットとして，独立行政法人情報処理推進機構（IPA）が公表している情報処理技術者試験の過去問題[14]のうち，2025 年度分の情報セキュリティに関連する全ての選択問題 119 問を使用する．なお，原文は PDF 形式で公開されているが，前処理として機械可読性のある JSON 形式[18]へ変換する．

精度評価に使用する適合度の正解データについては，人手で作成する．部局情報技術責任者に関連する組織規定から無作為抽出した 20 規則項目と，学習コンテンツのスーパーセットを構成する 119 問について，各規則項目の内容について，各学習コンテンツの問題を解くべきかどうかを，適合度として合計で 2380 件（20 規則項目×119 問）のアノテーションを作成する．

（２） 大規模言語モデル

評価に使用する大規模言語モデル（LLM）として，gpt-5.2-2025-12-11 [19]，gemini-3-pro [20]，及び gpt-oss-120b [21]を使用する．gpt-5.2-2025-12-11 については，以下単に gpt-5.2 と表記する．

なお，LLM による推論で指定する Reasoning Effort のパラメータについては，gpt-5.2 では low と xhigh を使用し，gemini-3-pro と gpt-oss-120b では low と high を使用する．パラメータの値が大きい方については，それぞれのモデルで最も大きく設定できる値とする．

（３）　　多クラス分類の評価結果：混同行列

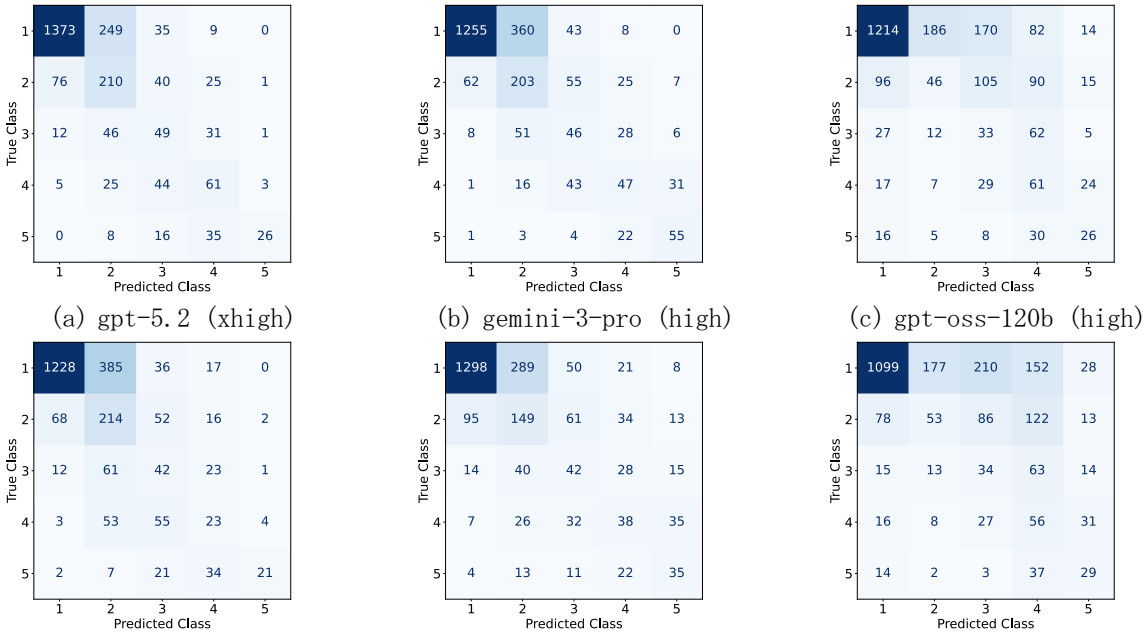
提案手法を用いて 1 から 5 の 5 段階評価による多クラス分類を行いその精度を評価した．

図 3-1 に LLM のモデルと Reasoning Effort のパラメータごとに出力した混同行列を示す．対角成分は正解数を示し，非対角成分は誤った分類数を示す．

全てのモデルに共通して，クラス 1 の正解数が最も多く，分類性能も最も高い結果となった．一方で，クラス 2，クラス 3，クラス 4 及びクラス 5 の正解数は多いとは言えず，誤分類が多い結果となった．

この結果より，役割と関係の無い学習コンテンツを取り除くタスクにおいては有効であると考えられる．一方で，他のクラスの正解数は多いとは言えず誤分類が多い傾向があり，これらのクラスは，どちらとも一概には言えない中間的な基準により分類していることから，評価がより曖昧になることで誤分類が多く発生している可能性が考えられる．

Reasoning Effort とクラスに着目すると，Reasoning Effort を最大にすることで gpt-5.2 と gemini-3-pro は全てのクラスにおいて精度の改善が見られるが，gpt-oss-120b では必ずしもそうでない場合が見られた．例えば，クラス 5 に対して，gpt-oss-120b（low）では正解数が 29 となったが，gpt-oss-120b（high）では正解数が 26 と減少している．



(a) gpt-5.2（xhigh）　(b) gemini-3-pro（high）　(c) gpt-oss-120b（high）

(d) gpt-5.2（low）　(e) gemini-3-pro（low）　(f) gpt-oss-120b（low）

図 3-1：LLM-as-a-Judge によるスコアリング結果の制度評価

（４）　　多クラス分類の評価結果：分類性能指標（全クラス平均）

表 3-1 に，全てのクラスを対象とした正解率（Accuracy，Acc.），適合率（Precision，Prec.），再現率（Recall，Rec.），及び F1 スコア（F1）の平均を，LLM のモデル（Model）および Reasoning Effort（RF）のパラメータ別に示す．

各モデルで設定可能な Reasoning Effort を最大にした場合において，gpt-5.2 は最も高い正解率 0.889，適合率 0.562，及び F1 スコア 0.502 を示した．また，gemini-3-pro は最も高い再現率 0.530 を示した．なお，gpt-oss-120b は全指標において他の 2 モデルより低い値を示した．F1 スコアに基づく総合性能においては gpt-5.2（xhigh）が最も高い指標を示す結果となった．

Reasoning Effort に着目すると，全クラス平均で見れば，全てのモデルにおいて low よりも high または xhigh の方が高い指標を示す結果となった．改善幅で見れば gpt-5.2 が最も顕著となる一方で，gpt-oss-120b ではあまり大きな改善は見られず，本研究における多クラス分類においては，モデルに応じて改善幅が異なる結果となった．

gpt-5.2 は，適合率が最も高い一方で，再現率は gemini-3-pro よりも低い結果となった．このことから，gpt-5.2 は誤検出を抑制する傾向があり，gemini-3-pro は取りこぼしを抑制する傾向があると推察される．

表 3-1 分類性能指標（全クラス平均）

Model	RF	Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑
gpt-5.2	xhigh	0.889	0.562	0.504	0.502
	low	0.857	0.478	0.412	0.404
gemini-3-pro	high	0.870	0.485	0.530	0.496
	low	0.863	0.403	0.438	0.414
gpt-oss-120b	high	0.832	0.332	0.369	0.332
	low	0.814	0.317	0.360	0.312

（５） 分類性能指標（クラス別平均）

表 3-2 に，クラス別の適合率（Precision, Prec.），再現率（Recall, Rec.），及び F1 スコア（F1）の平均を，LLM のモデル（Model）および Reasoning Effort（RF）のパラメータ別に示す．

クラス 1（**表 3-2a**）の分類においては，全てのモデルにおいて他のクラスの分類よりも高い性能を示した．また，モデル間の差は比較的小さい結果となった．F1 スコアにおいては gpt-5.2（xhigh）が最も良い結果となった．

クラス 2（**表 3-2b**），クラス 3（**表 3-2c**），及びクラス 4（**表 3-2d**）の分類においては，全体的に性能が低下する結果となった．

クラス 5（**表 3-2e**）の分類においては，モデル間の差が比較的大きい結果となった．F1 スコアにおいては，gemini-3-pro（high）が最も良い結果となった．

これらの結果より，クラス 1 の分類においては，F1 スコアが最大で 0.877 と高い値を得られていることから，役割に関係の無い問題を取り除くタスクにおいて一定の有効性があるものといえる．また，クラス 5 の分類においては，F1 スコアが 0.598 となり，高い値とまでは言えないものの，役割に関係のある問題を選択するタスクにおいて一定の有効性を示したといえる．一方で，クラス 2，クラス 3，及びクラス 4 については，全体的に性能が低く改善が必要である．しかしながら，クラス 1 とクラス 5 の両端のクラス分類結果を使用することで，役割に関係の無い問題の除去や，役割に関係の有る問題の選択を比較的安全に行えると言える．

表 3-2a クラス 1 の分類性能指標（平均）

Model	RF	Prec. ↑	Rec. ↑	F1 ↑
gpt-5.2	xhigh	0.937	0.824	0.877
	low	0.935	0.737	0.824
gemini-3-pro	high	0.946	0.753	0.839
	low	0.915	0.779	0.842
gpt-oss-120b	high	0.886	0.729	0.800
	low	0.899	0.660	0.761

表 3-2b クラス 2 の分類性能指標（平均）

Model	RF	Prec. ↑	Rec. ↑	F1 ↑
gpt-5.2	xhigh	0.390	0.597	0.472
	low	0.297	0.608	0.399
gemini-3-pro	high	0.321	0.577	0.412
	low	0.288	0.423	0.343
gpt-oss-120b	high	0.180	0.131	0.151
	low	0.209	0.151	0.175

表 3-2c クラス 3 の分類性能指標（平均）

Model	RF	Prec. ↑	Rec. ↑	F1 ↑
gpt-5.2	xhigh	0.266	0.353	0.303
	low	0.204	0.302	0.243
gemini-3-pro	high	0.241	0.331	0.279
	low	0.214	0.302	0.251
gpt-oss-120b	high	0.096	0.237	0.136
	low	0.094	0.245	0.136

表 3-2d クラス 4 の分類性能指標（平均）

Model	RF	Prec. ↑	Rec. ↑	F1 ↑
gpt-5.2	xhigh	0.379	0.442	0.408
	low	0.204	0.167	0.183
gemini-3-pro	high	0.362	0.341	0.351
	low	0.266	0.275	0.270
gpt-oss-120b	high	0.188	0.442	0.263
	low	0.130	0.406	0.197

表 3-2e クラス 5 の分類性能指標（平均）

Model	RF	Prec. ↑	Rec. ↑	F1 ↑
gpt-5.2	xhigh	0.839	0.306	0.448
	low	0.750	0.247	0.372
gemini-3-pro	high	0.556	0.647	0.598
	low	0.330	0.412	0.366
gpt-oss-120b	high	0.310	0.306	0.308
	low	0.252	0.341	0.290

3-5 おわりに

本研究では，組織規定に基づいて役割別の学習コンテンツを選択するための LLM-as-a-Judge 手法を提案した．組織規定，役割，及び学習コンテンツのスーパーセットから，役割に関連する規則項目と学習コンテンツの組に対する適合度を LLM によりスコアリングすることで，役割に応じたサブセットを構成する枠組みを示した．

評価では，著者らが所属する公開規定から指定した役割に関連する 20 個の規則項目を無作為に抽出し，独立行政法人情報処理推進機構（IPA）が公表している情報処理技術者試験の 2025 年度の情報セキュリティに関連する過去問題 119 問との組み合わせ 2380 件の正解データを作成し，複数の LLM による 5 段階の多クラス分類の精度を比較した．その結果，クラス 1（非関連）とクラス 5（高関連）において相対的に良好な F1 スコアが得られ，役割に関係のない問題の除去および高関連問題の抽出に一定の有効性が示された．

一方で，クラス 2 からクラス 4 の中間クラスの多クラス分類の精度は十分ではなく，閾値設計や評価基準の詳細化などによる精度の改善が今後の課題である．

また，本研究は単一の組織，単一の役割，及び単一の年度の問題での評価にとどまっており，汎用性の検証が不十分である．規則の階層構造を明示的に扱うスコアリングや，役割間の相対的重要度を考慮した重み付け，および複数組織・複数役割による評価を行うことで，実運用での信頼性を高める必要がある．さらに，学習者の到達度や業務内容に応じた個別最適化を組み込むことで，教育効果と運用負荷の両立を図りたい．

また，多くの高等教育機関でサンプルとして採用されている「高等教育機関の情報セキュリティ対策のためのサンプル規程集」[23]などを用いた評価を行うことや，提案手法を用いて，実際に e-Learning システムとして使用するとともに，ユーザによる客観評価を行うことも検討している．

3-6 付録

本節で使用したプロンプトを**図 3-A1**, **図 3-A2**, 及び**図 3-A3** に示す.

```
- あなたの役割は、与えられた「役割」を達成するために必要な学習コンテンツを「ルールブック」を参照して、「スーパーセット」を評価する専門家です。
- 「役割」は ISO/IEC 27000:2018 で定義されている「role」であり、user_prompt で指定する。
- 「ルールブック」を参考にして、その役割が持つ責務について考察する。
- 「スーパーセット」は学習コンテンツであり、user_prompt で指定する。
- 「question」と「responseLabel」の文脈を読み取り、問題を解くことで得られる知識・技術について考察する。
- 「ルールブック」は組織の規則、規約であり、user_prompt で指定する。
- 項目（章、条、項…）が階層で構成され、評価対象となる項目を「項目の単位」とし、user_prompt で指定する。
- ある項目が他の項目との関係性を示している場合は、そのほかの項目の内容を考慮して文脈を読み取る。
- 項目から階層構造と「役割」に関する内容を読み取り、「役割」に求められる知識・技術について考察する。
- 数値評価の際に「適合度」を用いる。
- 「適合度」は「ルールブック」を参考にして、「役割」に求められる学習内容としてどれだけ適しているかを 0〜100 で表したものとする。
- 「適合度」の評価には文脈理解を用いた評価を行ってください。
- 「適合度」の評価には「役割」という観点を重視した評価を行ってください。
```

図 3-A1: システムプロンプト

```
- 「役割」は「全学情報総括責任者」、「全学情報実施責任者」、「部局情報総括責任者」、「部局情報技術責任者」、「部局情報技術担当者」、「アカウントを管理する者」、「教職員等」。
- 今回の「役割」は「部局情報技術責任者」とする。
- 「スーパーセット」は {superset}
- 「ルールブック」は下の規則全文を用いる。
# 「ルールブック」本文
{rule_text}
- 「適合度」は、user_prompt で指定された「スーパーセット」の問題 1 問と「ルールブック」の指定された項目 1 条に対してそれぞれ 1 つ算出する。
- 「適合度」の値の基準を以下に示す。
  - 前提条件:「ルールブック」の項目が、user_prompt で指定された「役割」に関するものではない場合は「適合度」を 0 とする
  - 100〜76:「役割」の学習内容に不可欠な知識・技術（前提となる基礎知識・技術も含む）の学習コンテンツ。
  - 75〜51:「役割」の学習内容に必要とする知識・技術（前提となる基礎知識・技術も含む）の学習コンテンツ。
  - 50〜26:「役割」の学習内容に直接は扱わない知識・技術（前提となる基礎知識・技術も含む）の学習コンテンツ。
  - 25〜1:広い意味では知識・技術の接点はありうるが、「役割」の学習内容にはほとんど結び付かない知識・技術（前提となる基礎知識・技術も含む）の学習コンテンツ。
  - 0:「ルールブック」の項目・「役割」の責務とつながりがない知識・技術。
- 出力
  - 「適合度」20 個（問題 1 問 * 指定する規則項目 20 項目（clauses）を計算する。
    - 規則項目:{', '.join(clauses)}
  - 「適合度」の計算結果を必ず以下の例（json_template）のような JSON 配列のみで出力してください。
  - 余計な文章・説明・改行前文は禁止。
```

図 3-A2: ユーザプロンプト 1

```
# 出力例（余計な文章・説明・改行前文は禁止）
[
  {
    "第 5 条": 0-100 の整数,
    "第 10 条": 0-100 の整数,
    ...
    "第 98 条": 0-100 の整数
  }
]
```

図 3-A3: ユーザプロンプト 2

4 【課題 2-2】情報セキュリティ教育における個人の役割に適応した学習コンテンツの自動生成

4-1 はじめに

サイバー攻撃は世界規模で増加し続けている．情報セキュリティ対策において，技術的セキュリティ対策だけで防ぐことは難しく，人的セキュリティ対策が重要となっている．組織における人的セキュリティ対策として組織内での情報セキュリティ教育を実施する際に，情報セキュリティに関する知識の範囲は非常に広い一方で，組織の規則で定められた役割に応じて必要な知識の範囲が異なる．役割に応じた内容に絞った学習コンテンツや研修を実施することが必要となるが，自組織で定めた規則と役割に応じた学習コンテンツの内製は負担が大きいことから実施が難しいという問題がある[22]．

そこで本研究では，組織の規則で定めた役割に適応した学習コンテンツを大規模言語モデル（LLM）で自動生成する手法を提案し，その生成精度を評価する．

4-2 提案手法

（1） 概要

本研究において役割とは，情報セキュリティマネジメントシステム（ISO/IEC 27001:2022）[12]において，組織の中で割り当てられる職務や責任のことを指す．役割は，セキュリティ規則（以下規則）で定義されており，管理責任者や監督者はこれを遵守させる．規則は複数の規則項目から構成されており，役割を与えられた個人は，役割に該当する規則項目を理解して遵守する必要がある．

提案手法では，まず規則に記載されている役割ごとに必要な規則項目群を抽出し，その規則項目群から学習コンテンツ（問題文，選択肢，解答，及び解説）を生成する．

提案システムの構成を**図 4-1** に示す．本システムでは規則から必要な規則項目を抽出するエージェント，抽出された規則項目群から学習コンテンツを生成するエージェント，及び全体を管理するシステムから構成される．

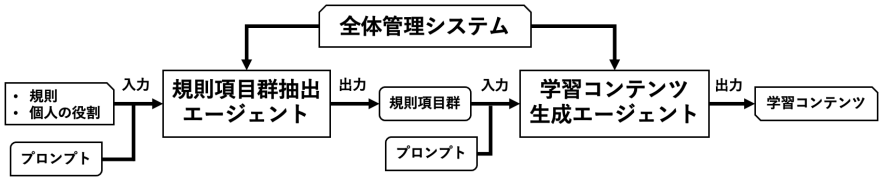


図 4-1: 学習コンテンツの自動生成システムの構成

（2） 規則項目群の抽出

組織で定められている規則は，機械可読性の低い形式のテキストかつ組織によって様式も異なる場合が多い．このような不定形なテキストを整形するとともに必要な部分を抽出するタスクにおいて LLM は有用である．

そこで提案手法では，規則項目群抽出エージェントが，個人の役割，規則，及び抽出方法を記述したプロンプトを入力として，LLM により推論を行い，個人の役割において遵守する必要のある規則項目群を出力する．プロンプトにおいては，規則の条や項等の規則の基本的な構成要素について説明するとともに，規則項

目に対して一意な識別子の付与を指示する．また，役割と関連する規則項目の抽出を指示する．

（３）　学習コンテンツの生成

規則においては様々な専門用語が記載されているが，その用語の辞書的な意味の理解を確認しても，実際の情報セキュリティインシデントの発生時に必要な知識を有しているかどうかは分からない．このため，規則項目が実際に適用される具体的な状況を考えた上で，その状況で必要となる知識を問う学習コンテンツの作成が必要となる．このようなタスクにおいて LLM は有用であると考えられる．

そこで提案手法では，学習コンテンツ生成エージェントが，規則項目群抽出エージェントにより抽出された規則項目群とプロンプトを入力として，LLM により推論を行い，学習コンテンツを出力する．プロンプトにおいては，学習コンテンツを出力する生成過程において，まず規則項目が適用される具体的な状況を推論させ，その具体的な状況を元に学習コンテンツを出力する指示を行う．これにより，規則に記載されている用語の暗記問題ではなく，規則の適用が必要となる具体的な状況での知識の理解を問う学習コンテンツの生成を行う．

4-3 実験

（１）　条件

実験では，LLM のモデルとして gpt-oss-120b と gpt-5.2-2025-12-11（以下単に gpt-5.2 と記載）を使用する．Reasoning Effort について，gpt-oss-120b では low，medium，及び high を使用する．gpt-5.2 では low，medium，及び xhigh を使用する．

gpt-5.2 は，gpt-oss-120b には存在しない機能として，Responses API やファイル添付による拡張検索を有する．このため，gpt-5.2 に特有の機能を使用する場合と使用しない場合についても実験を行う．なお，temperature と top_p については，gpt-oss-120b では 0，gpt-5.2 では非対応のため未設定とする．

規則として鳥取大学情報システム運用管理要項[16]を使用する．役割として全学情報総括責任者，部局情報技術担当者，及び教職員等を使用する．

（２）　規則項目群の抽出

役割に応じて必要な規則項目群の抽出を行い，その性能を評価した．抽出結果を分析したところ，モデル，Reasoning Effort，及び gpt-5.2 に特有の機能の使用有無の違いは抽出性能の違いに大きな影響は与えなかった．

一方で，役割の違いは抽出性能に大きな影響を与えた．役割別に抽出性能を評価した結果を**表 4-1** に示す．これらは役割以外の条件について平均値を算出したものである．

役割に応じて F1 スコアが低くなる場合があることが確認できる．これは，役割に応じて必要な規則項目かどうか判断する難易度が大きく異なることや必要な規則項目の判断基準を明確にプロンプトで与えなかったことが原因として考えられる．

表 4-1 規則項目群の抽出性能評価（役割別）

役割	正解率	適合率	再現率	F1 スコア
全学情報総括責任者	0.939	0.953	0.552	0.725
部局情報技術担当者	0.785	0.501	0.683	0.518
教職員等	0.871	0.797	0.913	0.811

（３）　学習コンテンツの生成

役割別に人手で用意した規則項目群を対象に，学習コンテンツの生成を行い，その性能を LLM-as-a-Judge（LaaJ）により評価する．

学習コンテンツの生成方法として，1 回の LLM での推論において，複数の規則項目を入力として複数の学習コンテンツを出力する場合は n/n とし，1 つの規則項目から 1 つの学習コンテンツを生成する場合は 1/1 とし，複数の規則項目から 1 つの学習コンテンツを生成する場合は n/1 とする．

LaaJ による評価において，生成した学習コンテンツ 1 問に対して，次に示す 7 つの評価観点毎に 0 から 100 の範囲の点数で評価する指示をプロンプトで行い，LLM の推論により得られた評価観点毎の点数の和をその学習コンテンツの品質の総合評点とする．また，総合評点を満点で除した値を評点率とし，これを 1 問の評価値として学習コンテンツ全体で 1 問の評価値の平均，中央値，及び標準偏差を算出した．

- 高次の能力を測る問題では受験者にとって新奇な素材を用いること
- 問いたいことは何か，問題を解くために必要な能力は何かが明確であること
- いずれの選択肢ももっともらしいこと
- 正答枝と誤答枝が明確に区別できること
- 問題文，選択肢，解説の内容が事実や規則に基づき正確であること
- 学習者の学習に役立つ情報を解説に記載していること
- 学習コンテンツの内容が規則と役割に合致していること

学習コンテンツの生成性能の評価結果を**表 4-2** に示す．評価結果を分析したところ，モデル，Reasoning Effort,役割,及び gpt-5.2 に特有の機能の使用有無の違いは抽出性能の違いに大きな影響は与えなかった．

一方で，学習コンテンツ生成方法の違いは評価結果に大きな影響を与え，1/1 と n/1 の方が n/n よりも良い評価値となった．これは，LLM による 1 回の推論において，1 つのみ学習コンテンツを出力させることで，LLM の注意が分散しないことが理由として考えられる．

また，学習コンテンツ生成方法の違いは，必要な総トークン数にも大きな影響を与え，n/1 が最も消費するトークン数が大きい結果となった．トークン数は gpt-oss を使用する場合は推論時間に影響し，gpt-5.2 を使用する場合は使用料金に影響する．このことから，1/1 がコストとパフォーマンスのバランスが良いと考えられる．

表 4-2 学習コンテンツの生成性能評価（生成方法別）

生成方法		LaaJ 評価値	トークン数	
入力	出力	平均±標準偏差	入力	出力
n	n	0.777 ± 0.009	5k	20k
1	1	0.825 ± 0.006	59k	44k
n	1	0.828 ± 0.006	230k	48k

4-4 おわりに

本研究では，個人の役割に適応した学習コンテンツを LLM により自動生成する手法を提案し，規則からの規則項目群の抽出精度と学習コンテンツの生成精度を評価した．

規則からの規則項目群の抽出においては，役割によっては F1 スコアが 0.725 や 0.811 と良好なスコアとなったが，0.518 と低くなる場合もあることが分かった．学習コンテンツの生成においては，LLM-as-a-Judge による評価値は 0.777～0.828 となり一定の有効性を示したと考えられるが，今後の課題として，その LaaJ 評価値の信頼性の定量的評価が挙げられる．

【参考文献】

[1] Dawkins, S. and Jacobs, J.: NIST Phish Scale User Guide, Nist series Technical Note (NIST TN) 2276, National Institute of Standards and Technology (NIST) (2023).

[2] Neupane, S., Fernandez, I. A., Mittal, S. and Rahimi, S.: Impacts and Risk of Generative AI Technology on Cyber Defense (2023).

[3] Humphreys, D., Koay, A., Desmond, D. and Mealy, E.: AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business, AI Ethics, Vol. 4, pp. 791–804 (online), DOI: 10.1007/s43681-024-00443-4 (2024).

[4] The Open Worldwide Application Security Project (OWASP): OWASP Top 10 for LLM Applications 2025. Version 2025, Guide (2024).

[5] The Open Worldwide Application Security Project (OWASP): Agentic AI - Threats and Mitigations: OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative. Version 1.0, Guide (2025).

[6] Vishwanath, A.: The Weakest Link: How to Diagnose, Detect, and Defend Users from Phishing, The MIT Press (2022).
[7] Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J. and Park, P. S.: Devising and Detecting Phishing: Large Language Models vs. Smaller Human Models, Black Hat USA 2023 Whitepaper, (online), DOI: 10.48550/arXiv.2308.12287 (2023).
[8] Ministry of Economy, Trade and Industry (METI) and Information-technology Promotion Agency, Japan (IPA): Cybersecurity Management Guidelines for Japanese Enterprise Executives Ver. 3.0, Guide (2023).
[9] 和久友親, 田村哲嗣, 川瀬真弓: 学習者の理解度に応じた自動問題生成 AI システムの開発, 日本教育工学会論文誌(2024).
[10] Check Point Research: Global Cyber Attacks Surge 21% in Q2 2025—Europe Experiences the Highest Increase of All Regions（online）, available from ⟨https://blog.checkpoint.com/research/global-cyber-attacks-surge-21-in-q2-2025-europe-experiences-the-highest-increase-of-all-regions/⟩（accessed 2026-06-30）.
.
[11] 齋藤孝道: マスタリング TCP/IP 情報セキュリティ編, オーム社, 第 2 版 (2022).
[12] 中尾康二: ISO/IEC27001:2022（JIS Q 27001:2023）情報セキュリティマネジメントシステム要求事項の解説, 日本規格協会 (2023).
[13] 独立行政法人情報処理推進機構（IPA）:「2024 年度中小企業における情報セキュリティ対策に関する実態調査」報告書 (2025).
[14] 独立行政法人情報処理推進機構（IPA）: 過去問題, 独立行政法人情報処理推進機構（IPA）(online), available from ⟨https://www.ipa.go.jp/shiken/mondai-kaiotu/index.html⟩（参照 2026-02-19）.
[15] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L. and Guo, J.: A Survey on LLM-as-a-Judge, arXiv, (online), DOI: abs/2411.15594 (2025).
[16] 国立大学法人鳥取大学: 鳥取大学情報システム運用管理要項, 国立大学法人鳥取大学(online), available from ⟨https://www.oism.tottori-u.ac.jp/about/rules⟩(参照 2026-02-19).
[17] Penango, S. L. and Angeles, F. L.: The text/markdown Media Type, Technical Report RFC 7763, Internet Engineering Task Force (IETF) (2017).
[18] Bray, T.: The JavaScript Object Notation (JSON) Data Interchange Format, Technical Report RFC 8259, Internet Engineering Task Force (IETF) (2017).
[19] OpenAI: Update to GPT-5 System Card: GPT-5.2, OpenAI (online), available from ⟨https://openai.com/index/gpt-5-system-card-update-gpt-5-2/⟩ (accessed 2026-02-19).
[20] Google LLC: Gemini 3 Pro Model Card, Google LLC (online), available from ⟨https://deepmind.google/models/model-cards/⟩ (accessed 2026-02-19).
[21] OpenAI: gpt-oss-120b & gpt-oss-20b Model Card, OpenAI (online), available from ⟨https://arxiv.org/abs/2508.10925⟩ (accessed 2026-02-19).
[22] 独立行政法人情報処理推進機構（IPA）.「2023 年度 SECURITY ACTION 宣言事業者における情報セキュリティ対策の実態調査」報告書, 2024.
[23] 国立情報学研究所: 高等教育機関の情報セキュリティ対策のためのサンプル規程集 (online), available from ⟨https://www.nii.ac.jp/service/sp/⟩ (accessed 2026-06-30).

〈発表資料〉

題名	掲載誌・学会名等	発表年月
知識グラフを用いた個人適応型サイバーセキュリティ学習システムの開発	コンピュータセキュリティシンポジウム 2025 論文集, 情報処理学会, pp.24-27, 2025. https://ipsj.ixsq.nii.ac.jp/records/2008775	2025 年 10 月
大規模言語モデルを用いた組織規定に基づく役割別情報セキュリティ学習コンテンツ選択手法の検討	情報処理学会研究報告, 情報処理学会, Vol.2026-CSEC-112, No.73, pp.1-6, 2026. https://ipsj.ixsq.nii.ac.jp/records/2008594	2026 年 3 月
Development of a Personalised Cybersecurity Learning System with Knowledge Graph	Information Management, Springer CCIS series, pp.?-?, 2026. (※to appear) https://link.springer.com/conference/icim (※to appear)	国際会議発表 2026 年 3 月 国際会議論文 2026 年?月(※to appear)