生成 AI を用いた標的型攻撃メール対応訓練システムの開発

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

多くの組織において，標的型メール攻撃が脅威となっている．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できない．このため，すり抜けた標的型攻撃メールに対する訓練や教育等の人的セキュリティの重要性が高い[1]．近年では，標的型攻撃メールの内容が巧妙化しており，一般的なメール利用者がメールの内容だけで本物と偽物を見分けることは難しくなっている．さらに，生成 AI 技術の急激な発展に伴い，生成 AI を悪用した標的型メール攻撃が行われる可能性も懸念されている[2]．このため，標的型メール攻撃に対する人的セキュリティの重要性は今後さらに高まると考えられる．

生成 AI を悪用した標的型メール攻撃に対応するには，より効果的な訓練が必要となる．文献[3]では高頻度での継続的な教育が必要であることが指摘されている．また，文献[4]では，公開されている注意喚起情報について，攻撃メールに対する教育的アドバイスが抽象的であることや，標的型攻撃に対する具体的記述が無い問題を実態調査により報告している．今後，生成 AI を悪用した標的型メール攻撃が脅威となることを踏まえると，生成 AI が可能な攻撃の限界を見定めた上で，それに対応した具体的で高頻度かつ継続的な対応訓練が必要となる．

そこで本研究では，これらの必要性に応える生成 AI を用いた標的型攻撃メール対応訓練システムを開発することを目的とする．

2 【課題 1】訓練システムの基礎部分の開発

2-1 概要

生成 AI を用いた標的型攻撃メール対応訓練システムを提案する．不審メールによる攻撃として代表的な標的型メール攻撃は，システムセキュリティだけでは防ぐことが難しいことから，標的型メール攻撃に対する訓練や教育といった人的セキュリティの重要性が高い．しかし，高頻度かつ継続的な不審メール対応訓練は組織で許容され難い場合がある．そこで，生成 AI を活用して訓練用のフィッシングメールや訓練用のフィッシングサイトの生成を自動化することにより，訓練対象者が高頻度かつ継続的に自主訓練が可能な生成 AI を用いた標的型攻撃メール対応訓練システムを提案する．本報告では試作したプロトタイプについて述べる．

2-2 はじめに

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に対する訓練や教育等の人的セキュリティの重要性が高い[1]．近年では，標的型攻撃メールの内容が巧妙化しており，一般的なメール利用者がメールの内容だけで本物と偽物を見分けることは難しくなっている．さらに，生成 AI 技術の急激な発展に伴い，生成 AI を悪用した標的型メール攻撃が行われる可能性も懸念されている[2]．このため，標的型メール攻撃に対する人的セキュリティの重要性は今後さらに高まると考えられる．

生成 AI を悪用した標的型メール攻撃に対応するには，より効果的な訓練が必要となる．文献[3]では半年おきの頻度での継続的なセキュリティ教育を行うことを提案している．今後，生成 AI を悪用した標的型メール攻撃が脅威となることを踏まえると，生成 AI が可能な攻撃の限界を見定めた上で，それに対応した高頻度かつ継続的な対応訓練が必要となる．しかし，高頻度かつ継続的な不審メール対応訓練は，本来の業務の圧迫が懸念され，組織で許容され難い場合がある．そこで，生成 AI を用いることで，訓練用のフィッシングメールと訓練用のフィッシングサイトを自動生成することにより，訓練対象者が自身のペースで自習可能な標的型攻撃メール対応訓練システムを提案する．

2-3 システムの設計

（1）訓練用フィッシングメールの生成

フィッシングメールは，ばらまき型とやりとり型に分類される．特に標的型攻撃メールではやりとり型が

公益財団法人電気通信普及財団
研究調査助成報告書 第40号 2025年度

用いられる場合が多い．しかし，やりとり回数を 1 回にするか複数回にするかは実際の攻撃者の判断による．そこでばらまき型については，やりとり回数が 0 の場合のやりとり型として扱うことで両方の型に対応し，やりとり回数の判断は生成 AI に行わせることとする．

また，既存の多くの不審メール対応訓練システムではメールのテンプレートを用いたやりとりを行う場合がある．しかし，同じテンプレートを用いた訓練では標的型攻撃メールの対応訓練にはなり難い．訓練においても，やりとりの内容に応じた返信が行われる必要がある．そこで，提案システムでは生成 AI にやりとりを代行させる．

標的型メール攻撃への防御モデルとして The Unified Kill Chain [4]が参考にされる場合がある．また，文献[5]では，やり取り型の標的型攻撃に対する訓練システムのモデル化が試みられている．しかし，実際の攻撃では，予め想定したモデルのとおりに状態遷移するとは限らない．そこで，偵察，催促，及び攻撃といった状態への遷移条件についても，メールのやりとりの内容を踏まえて生成 AI に判断させる．

文献[3]では，半年おきの頻度での継続的なセキュリティ教育を行うことを提案している．年に 1 回程度の訓練では知識の定着が難しい．この為，生成 AI によりやりとりを自動化することにより，訓練対象者が自分の好きなタイミングでの自主訓練を可能とする．

（２）訓練用フィッシングサイトの生成

文献[6， 7， 8]等の既存の不審メール対応訓練システムにおいては，事前に作成しておいた訓練用のフィッシングサイトへ訓練用フィッシングメールで誘導する手法を採用している場合が多い．しかし，生成 AI が悪用される場合，事前ではなく攻撃時に動的に生成される場合も想定する必要がある．そこで，訓練用の標的型攻撃メールの文面のやり取り内容を考慮して訓練用のフィッシングサイトを生成 AI で自動生成させる．訓練対象者毎の訓練内容に応じた訓練用フィッシングサイトが生成されるため，高頻度かつ継続的に繰り返し様々なパターンで訓練を行うことが可能となり，高い訓練効果が得られることが期待できる．

また，NIST Phish Scale User Guide [1]などのガイドラインに基づいて不審メール対応訓練を行う場合，訓練対象者の振る舞いを記録しておく必要がある．そのため，生成 AI に訓練用のフィッシングウェブサイトを生成する場合において，ガイドラインに基づいた評価を行うためのデータの収集機能も有する必要がある．

2-4 システムの実装

システムの概要を図 1 に示す．

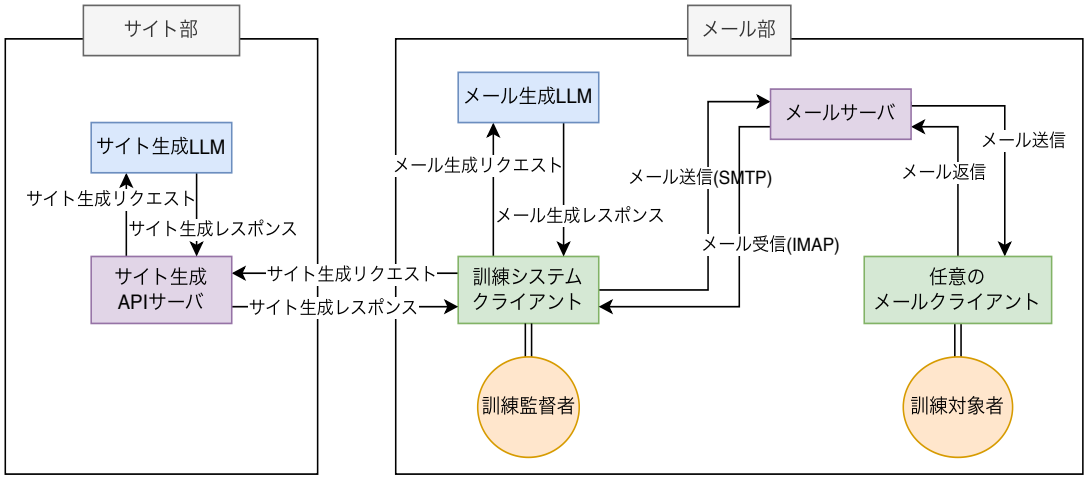

図 1: システムの概要

メール部は訓練用メールのやり取りを管理・実行する．サイト部は訓練用の攻撃時に添付するリンク先の訓練用のフィッシングサイトを作成するとともに訓練状況のデータを管理する．訓練監督者は，訓練システムクライアントを操作し，訓練対象者に対する訓練全体を管理・進行する．

訓練を行う際は，訓練監督者が訓練クライアントにシナリオの概要を入力することから始まる．訓練システムクライアントは，入力されたシナリオをプロンプトに組み込みメール生成 AI を用いて偵察メールを作成し，訓練対象者に送信する．訓練対象者からの返答に基づき，システムは訓練対象者の信頼度を計測し，一定の閾値を下回る限り偵察メールのやり取りを続ける．信頼度が閾値を超えた際，訓練は攻撃メールの送信に移行し，悪意のあるサイトへのリンクを含むメールを訓練対象者に送る．サイト生成 API サーバは，やり取りに基づいて訓練用フィッシングサイトを生成し，その URL を攻撃メールに含める．最終的に，訓練対象者の行動を評価し，必要に応じて注意喚起を行う．

訓練システムクライアントは Go 言語を用いて CLI アプリケーションとして実装した．訓練監督者は手元の計算機から本 CLI の操作を行い一連の訓練を監督する．また，メール生成 LLM は，ChatGPT API [9]によるクラウド型 LLM や Ollama によるローカル型 LLM を LangChain で利用する Python スクリプトとして実装した．サイト部については，文献[10]で OSS として公開しているフィッシング対応訓練システムに，提案する生成 AI による訓練用 Web サイトの自動生成機能を追加した．サイト生成 LLM は OpenAI の ChatGPT によるクラウド型 LLM または Ollama によるローカル型 LLM を LangChain で呼び出す方式で実装した．

信頼度については，やり取り回数を用いて定義した．やり取り回数の m 倍が信頼度であり，閾値 n を超えた場合に偵察から攻撃へ移る判断を行う．なお，m と n は正の整数とする．m と n は訓練開始時に設定可能とし，これにより訓練監督者は設定した任意のやり取り回数で訓練を実施できる．

2-5 実験

開発したシステムが複数の訓練対象者に対して滞りなく訓練が実施できることを確認する．

（1）環境

本実験では，開発した訓練システムを実際に利用し，電子メールを用いたやり取り型フィッシング攻撃の訓練を行う．本システムは，電子メールを通じてフィッシング攻撃を模擬し，実際の攻撃に近い状況での訓練を可能にする．実験参加者は，各自のメールアドレスと普段使用しているメールクライアントを用いて，個々に訓練を受ける．実験の日時および参加者の選定は，我々の研究室のメンバーと公募により決定した．

（2）手順

本実験では学生と教員のべ 12 名が実験に参加した．訓練テーマについては，今回は所属する大学の学生と教員を対象とすることを考慮して「学務支援システムの仕様変更に伴う再ログインのお知らせ」に相当するものとした．

実験では，はじめに各グループに異なるやり取り回数（0 回，1 回，及び 2 回）で訓練を実施するため，実験前に参加者を 3 グループ（A，B，及び C）に分ける．次に，訓練システムからの電子メールを介したフィッシング攻撃訓練を参加者に実施する．参加者は各自で返信するか，URL をクリックするか，あるいは破棄するか等を判断し，結果をアンケート形式で収集する．終了後，実験で得られたパラメータとアンケート結果から考察を行う．

評価方法は次のとおりとした．

- アンケート項目： 実験後に参加者に対してウェブフォームでアンケートを実施し，得られた結果を考察とする．アンケートでは，「メールに返信した/しなかった理由」，「メールに記載された URL をクリックした/しなかった理由」，「自動生成されたフィッシングメールに対する感想・意見」を質問項目とする．
- クリック率： 各グループについてメールに記載された URL のクリック率を算出する．グループにおいて URL をクリックした参加者の人数を a，グループ全体の人数を b とし，URL クリック率 x は x = a / b で算出する．

（3）結果

開発したシステムは問題なく動作し，不具合が発生することはなかった．生成したメール文面に対するフィードバックについて，アンケートを集計した結果，9 名から回答があった．「一般的なメールの文面として違和感がなかった」という旨の回答が 77.8 %，「部分的に違和感があった」という旨の回答が 22.2 %となった．

また，追加のフィードバックとして，攻撃フェーズに移行していないにもかかわらず，メールの文面がやり取りを締めくくる形になっており，訓練対象者側からアクションを起こす必要がないと判断し，返信を打

ち切った例があった．また，URL のクリック率は表 1 のとおりとなった．

表 1: 各グループにおける URL クリック率

グループ	やりとり回数	URL クリック率（%）
A	0	25
B	1	0
C	2	0

（4）考察

システムは不具合なく動作した．しかし，メール文面のゆらぎによって実質的に訓練が中断される場合があった．一般的なメールの文面として違和感が無いという点では実験参加者のうち過半数の評価を得られたが，やりとり型の対策訓練に用いる文章としては不十分である．

また，ばらまき型（グループ A）に比べてやり取り型（グループ B，グループ C）の URL のクリック率が低くなった原因として，前述の偵察フェーズ中のメール文面におけるやり取りの中断が挙げられる．このような問題を防ぐためには，生成 AI によって生成したメール文面が訓練において適切であるかを定量的に評価する方法を検討し，不適切な場合には再生成を繰り返すことで対応可能できると考えられる．

2-6 おわりに

本研究では，不審メールに対する人的セキュリティ対策のために，生成 AI を用いた不審メール対応訓練システムを提案し，そのプロトタイプを開発した．開発したシステムは正しく動作したものの，生成 AI がメールによる会話を締めくくる回答をすることで，やり取りが中断してしまう場合があった．

今後の課題として，生成 AI によって生成したメール文面がやり取りを継続する出力となっているかを定量的に評価する方法の検討が必要である．また，NIST Phish Scale User Guide [1]や The Unified Kill Chain [4]といった不審メール対応訓練に関連するガイドラインやホワイトペーパを Retrieval-Augmented Generation (RAG)に活用することや，それぞれの訓練対象の訓練結果に応じて訓練の難易度を自動調整する手法の検討を予定している．

3 【課題 2-1】訓練用メール・サイト・アドバイスの自動評価：訓練内容の自動選定

3-1 概要

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．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．また，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検討する．

3-2 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[11]や，生成 AI の活用に伴うリスク[12, 13, 14]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 [15]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[2]．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[16]．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検討する．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．

3-3 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．

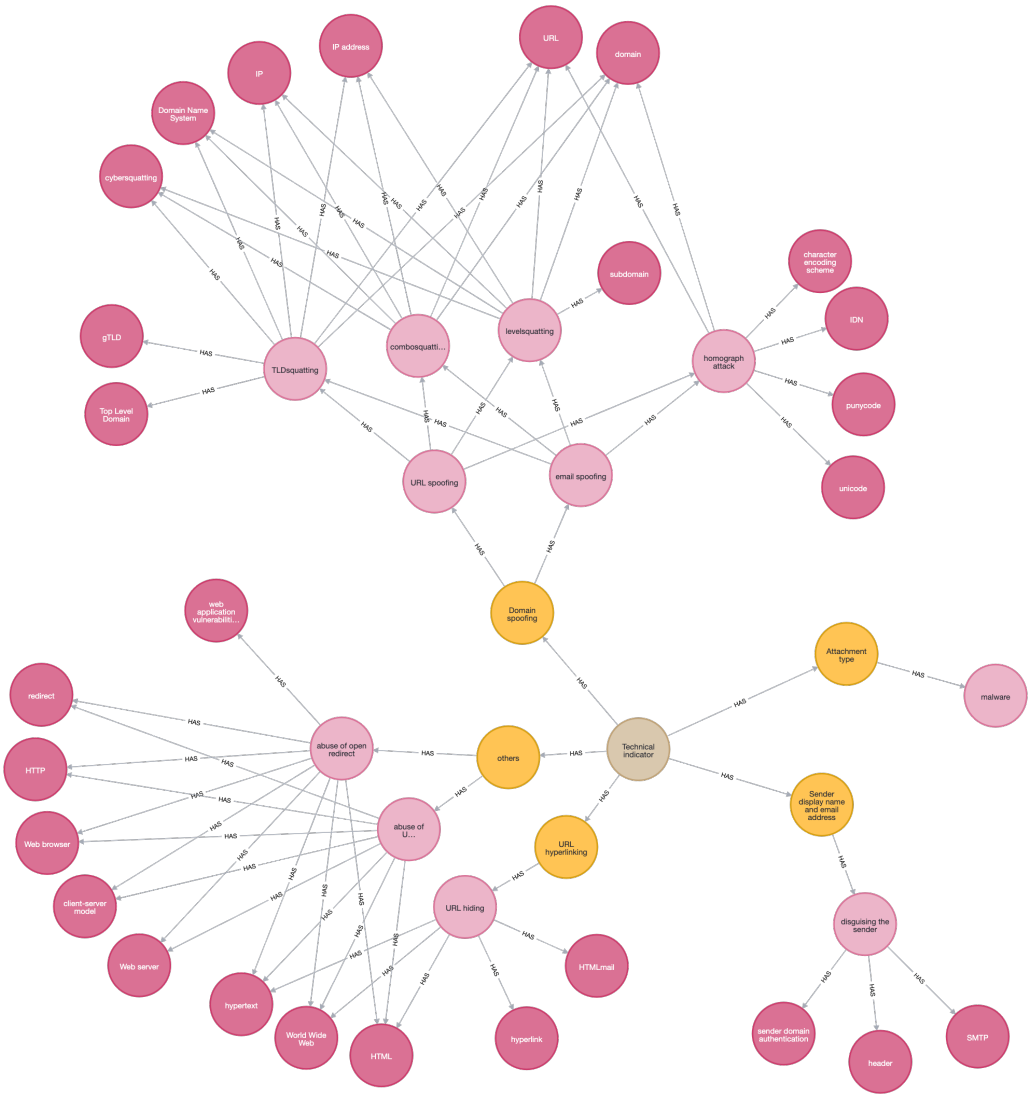

図 2: NIST Phish Scale User Guide (TN 2276)を情報源とし，Technical Indicator を起点とする知識グラフ

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[17]．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．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」ノードを起点に「Cue Name」ノードを作成し，その下に学習項目であるフィッシングに用いられる攻撃手法ノードを作成する．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所有関係を知るために，攻撃手法ノードの下に知識ノードを作成する．

図 2 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．

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

α は学習項目に設定する難易度の最大値，β は $1+K_n+\sum_{m=1}^{N} a_m P_{nm}$ 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．$L_i(1 \leq L_i \leq \alpha)$ 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．

3-4 実験

（1）実験の目的

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解順を記録し，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．

（2）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．

（3）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 3 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．

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．」と言う意見も得られた．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．

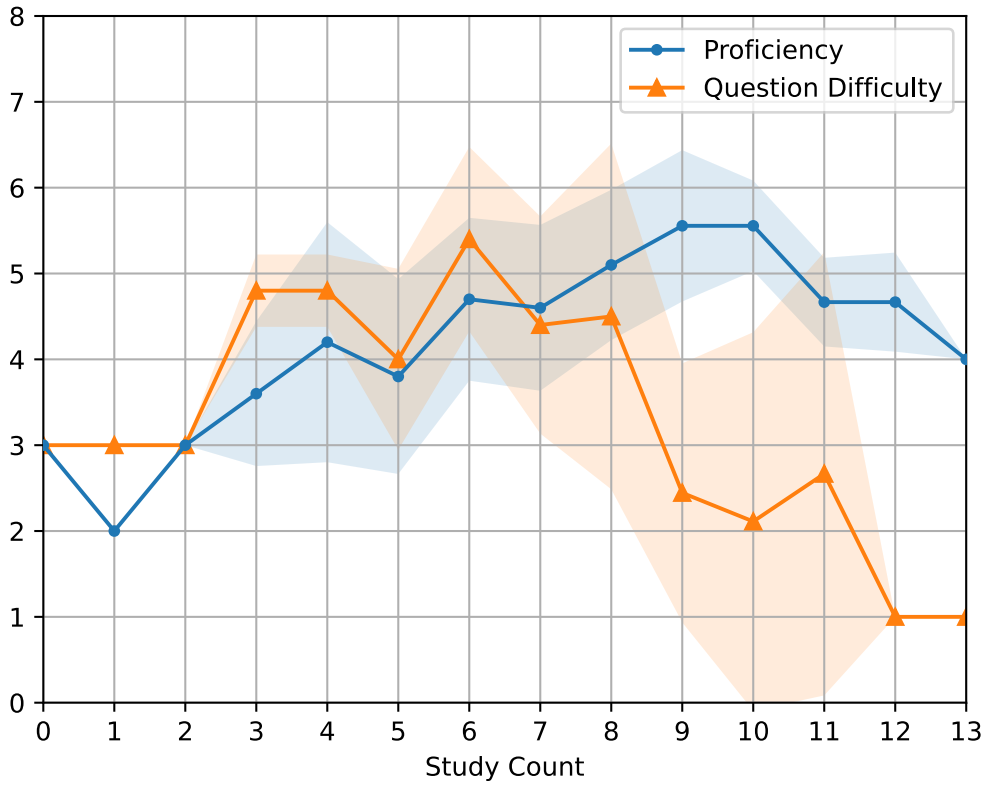

図 3: 実験参加者の習熟度（Proficiency）と学習項目の難易度（Question Difficulty）の平均と標準偏差の推移

3-5 おわりに

本研究では，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，本稿では扱わなかった 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）の適用可能性を検証することが挙げられる．

4 【課題 2-2】訓練用メール・サイト・アドバイスの自動評価：LLM による自動評価

4-1 概要

本研究では，生成 AI を用いて学習者の習熟度に合わせた訓練用メール文とフィードバック文の自動生成手法を検討する．

4-2 設計と実装

（1）システムの構成

提案システムの構成を図 4 に示す．本研究では，学習者の習熟度に応じて訓練用メール文とフィードバック文を生成するために，生成 AI によるテキスト生成に，知識グラフ組み合わせる．

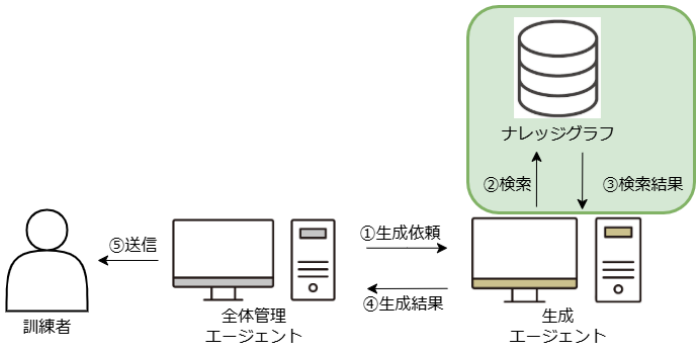

図 4: システムの構成

（2）知識グラフの利用

知識グラフに与える情報は，訓練用メールの内容，訓練者の学習状況，フィードバックの内容の３つに分けられる．これらの情報を整理し，知識グラフには，学習項目をまとめるための学習テーマ，学習項目，学習項目に対するポイントと対策，及び学習項目が用いられる事例の 4 つのノードを設定する．学習項目には NIST Phish Scale User Guide TN2276 [1]の視覚的なフィッシングメールの特徴から「言語と内容」の 7 項目を設定する．

（3）訓練用メール文の生成

本研究では，提案訓練システムにおいて，訓練対象者の習熟度に応じた訓練用メール文を生成する．従前のシステムはプロトタイプであり，訓練内容のシナリオが 1 つしか設定できない．そこで，訓練対象者毎に学習項目を設定可能とし，訓練対象者の習熟度に応じて異なる訓練用メールを生成する．

フィッシングメールは視覚的な特徴が多いほど脅威に気づきやすい[1]．訓練対象者にとって最も難易度が低く分かりやすいフィッシングメールから始めるために，視覚的な特徴の学習項目を全て含む訓練用メールを最初の学習内容とする．

フィッシングであると検知できた場合は，学習者の習熟度から最も重要でない学習項目を既習とする．訓練用メールは常に，未習とされている学習項目を組み合わせて生成する．学習項目が全て既習となった時，学習を終える．本研究では，NIST Phish Scale User Guide TN2276[1]における視覚的なフィッシングメール

の特徴「言語と内容」の 7 項目を用いる．この場合の学習項目の組み合わせは 2^7=128 通りとなり，これらの組み合わせを基に訓練用メールを生成する．

（3）フィードバック文の生成

フィードバックとは，現在の自分の理解とこれから学習すべき内容との差を埋めるための，学習課題や学習過程に関連した情報である．フィッシング対応訓練で個人毎に異なる訓練を行う場合，フィードバックも個々に対応した内容を生成する．

フィードバック文の生成方法として訓練用メール文からフィードバック文を生成する手法と，知識グラフの情報のみからフィードバック文を生成する手法を検討する．本研究においては，実験において 2 つの手法を比較し，どちらの方が訓練用メール文に対して適切なフィードバック文であるかを調査する．

4-3 実験

（1）実験目的

システムが正しく動くかを確認する．訓練用メール文とフィードバック文の生成をすべての組み合わせ（2^7=128 通り）で行い，適切に生成されているかを判定する．

（2）実験手法

適切に生成されているかを生成 AI により自動判定する．生成と評価には OpenAI の gpt-4o を用いる．

（3）訓練用メールの自動評価結果

生成された訓練用メール全体に対して，適切であると判定されたものの比率を適正率とし，表 2 に結果を示す．全体の平均は 94.20 %となり，「緊張感を感じさせる」という学習項目以外では，90 %以上で学習項目に沿った生成がされていると判定された．

表 2: 訓練用メールの評価

学習項目	適正率[%]
1. 法的文書/免責事項が含まれる	100.00
2. 気を散らす詳細がある	93.75
3. 機密情報を要求する内容がある	97.66
4. 緊急感を感じさせる	75.78
5. 脅迫的な言葉がある	94.53
6. 一般的な挨拶などがない	99.22
7. 署名の情報の不足	98.44

（4）フィードバック文の自動評価結果

フィードバック文の生成手法は 2 種類検討しており，訓練用メール文から生成する手法と，知識グラフに与えた情報のみから生成する手法を比較している．全体に対して適切であると判定されたものの比率を適正率とした．その結果，生成 AI による評価では，訓練用メール文の内容から生成した場合の適正率は 94.53 %となり，知識グラフの情報のみから生成した場合の適正率は 87.50 %となり，訓練用メール文の内容から生成する方がより良い結果となった．

4-4 考察

適正率が低い学習項目は，他の学習項目により，選択されていない時に訓練用メールの学習項目として組み込まれていると誤判定されていた．目視で確認したところ，想定通りの訓練用メールが作成されていたため，評価用の生成 AI に与えるプロンプトを改良する必要があると考えられる．また，選択された学習項目に沿った訓練用メールを送るために，適切に生成されているかを判定するアルゴリズムを導入する必要があると考えられる．

知識グラフから生成する手法によるフィードバック文が適切でないと判断された理由は，知識グラフには生成された訓練用メールの内容について詳しく言及されていないというものが多かった．実験結果から，フィードバック文は，知識グラフだけを基に生成するよりも，知識グラフから生成された訓練用メール文からそのフィードバック文を生成する手法の方が具体性のあるフィードバック文が生成できると考えられる．

4-5 おわりに

本研究では，生成 AI を用いた不審メール対応訓練システムにおいて，学習者の習熟度に応じた訓練内容の選択手法と，それに基づく訓練用メール文とフィードバック文を生成する手法を提案した．生成文を生成 AI により評価した結果，おおよそ実用的な精度で出力を得ることができた．

今後の課題として，システム全体を統合した実用性の検証を行うことが挙げられる．また，生成 AI による評価の精度が一部低い部分があった．生成 AI により自動修正できるようになれば，学習者に適切な内容を自動で送ることができると考える．

【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] 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).

[3] Reinheimer, B., Aldag, L., Mayer, P., Mossano, M., Duezguen, R., Lofthouse, B., von Landesberger, T. and Volkamer, M.: An investigation of phishing awareness and education over time: When and how to best remind users, Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020), USENIX Association, pp. 259–284 (online), available from ⟨https://www.usenix.org/conference/soups2020/presentation/reinheimer⟩ (2020).

[4] Pols, P.: The Unified Kill Chain, White paper (2023).

[5] 西川弘毅, 山本 匠, 木藤圭亮, 河内清人：やり取り型の標的型メールを自動で行うシミュレータの検討, 情報処理学会研究報告， Vol. 2017-CSEC-76, No. 33, pp. 1-7 (2017).

[6] Wright, J.: Gophish - Open-Source Phishing Framework, (online), available from ⟨https://getgophish.com/⟩ (accessed 2024-08-23).

[7] PentestGeek: PhishingFrenzy: Ruby on Rails Phishing Framework, (online), available from ⟨https://github.com/pentestgeek/phishing-frenzy⟩ (accessed 2024-08-23).

[8] Higashino, M.: A Design of an Anti-Phishing Training System Collaborated with Multiple Organizations, Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services, iiWAS2019, Association for Computing Machinery, p. 589–592 (online), DOI: 10.1145/3366030.3366086 (2020).

[9] OpenAI: OpenAI Platform, OpenAI (online), available from ⟨https://platform.openai.com/docs/guides/⟩ (accessed 2024-08-23).

[10] Higashino, M., Kawato, T., Ohmori, M. and Kawamura, T.: An Anti-phishing Training System for Security Awareness and Education Considering Prevention of Information Leakage, Proceedings of the 5th International Conference on Information Management, pp. 82–86 (online) DOI: 10.1109/INFOMAN.2019.8714691 (2019).

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

[12] 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).

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

[14] 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).

[15] Vishwanath, A.: The Weakest Link: How to Diagnose, Detect, and Defend Users from Phishing, The MIT Press (2022).

[16] 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).

[17] 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).

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
生成 AI を用いた不審メール対応訓練システムの試作	コンピュータセキュリティシンポジウム 2024 論文集	2024 年 10 月
生成 AI を用いた不審メール対応訓練システム	コンピュータセキュリティシンポジウム 2024, デモンストレーションセッション (※デモンストレーション及びポスター資料のみ)	2024 年 10 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検討	情報処理学会研究報告	2025 年 3 月