

# セキュリティ訓練済み大規模言語モデルは 攻撃者の認知バイアスを解明できるか？

中井 綱人<sup>1,a)</sup> 梨本 翔永<sup>1</sup> 日夏 俊<sup>1</sup> 大西 健斗<sup>1</sup> 東 拓矢<sup>1</sup>

**概要：**人間の認知や行動、意思決定に関するセキュリティ（コグニティブセキュリティ）の研究は、ソーシャルエンジニアリングなどの攻撃とその対策に関して、その多くが被害者の認知モデルに焦点を当てているが、攻撃者の認知モデルに着目した研究は少ない。攻撃者の認知を理解することは、将来の攻撃を予測し、能動的な防御策を構築するために不可欠である。しかしながら、攻撃者の認知モデルを構築するには、実際の攻撃者（被験者）に関する情報の収集が困難かつサンプル数が少ないという課題がある。そこで、サイバーセキュリティ分野における大規模言語モデル（LLM）活用の進展を背景に、サイバー攻撃について訓練された LLM を攻撃者に見立てる検討を試みた。本稿では、攻撃者の認知モデル構築を目的に、セキュリティ訓練済み LLM の意思決定過程における認知バイアスを初めて調査した。LLM の認知バイアスは、単に学習データ偏りに起因するものではなく、あるタスクの意思決定過程に関するものである。セキュリティ特有の認知モデルの傾向を調査するため、LLM の意思決定過程に対する認知バイアス評価フレームワーク BiasBuster を用いて、セキュリティ訓練済みモデルと訓練前の基盤モデルの 8 つのモデル間で比較評価を行った。実験の結果、訓練前と比較して、セキュリティ訓練済み LLM ではグループ属性バイアスと優先バイアスが顕著に現れることを明らかにした。この結果は、実際の攻撃者も同じバイアスを持つ可能性を示唆しており、攻撃者の特性であるコミュニティの性質、技術への深い関与、問題解決へのアプローチを反映していると考察した。更に、攻撃者の認知バイアスから、より効果的な防御策構築の可能性も議論した。

**キーワード：**コグニティブセキュリティ、大規模言語モデル、認知バイアス、能動的サイバー防御

## Can security-trained large language models elucidate the cognitive biases of attackers?

TSUNATO NAKAI<sup>1,a)</sup> SHOEI NASHIMOTO<sup>1</sup> SYUN HINATSU<sup>1</sup> KENTO OONISHI<sup>1</sup> TAKUYA HIGASHI<sup>1</sup>

**Abstract:** Research on cognitive security, which pertains to human cognition, behavior, and decision-making in relation to security, often focuses on the cognitive models of victims in the context of attacks such as social engineering and their countermeasures. However, there is limited research that focuses on the cognitive models of attackers. Understanding the cognition of attackers is essential for predicting future attacks and developing proactive defense strategies. Nevertheless, constructing a cognitive model of attackers poses challenges, including the difficulty of collecting information about actual attackers (subjects) and the limited sample size. In light of advancements in the utilization of large language models (LLMs) in the field of cybersecurity, this paper considers the possibility of using LLMs trained on cyber attacks as a stand-in for attackers. This paper investigates, for the first time, the cognitive biases in the decision-making processes of security-trained LLMs with the aim of constructing a cognitive model of attackers. The cognitive biases of LLMs are not merely due to biases in the training data, but are related to the decision-making processes for specific tasks. To investigate the tendencies of security-specific cognitive models, we employed the cognitive bias evaluation framework BiasBuster to compare and evaluate eight models, including security-trained models and baseline models prior to training, in terms of their decision-making processes. The results of the experiments revealed that, compared to before training, security-trained LLMs exhibited significant group attribute bias and priority bias. This finding suggests that actual attackers may possess similar biases, reflecting characteristics of attackers such as the nature of their communities, deep engagement with technology, and approaches to problem-solving. Furthermore, the cognitive biases of attackers were discussed in terms of their potential to contribute to the development of more effective defense strategies.

## 1. はじめに

サイバーセキュリティ分野において、人間の認知や行動、意思決定を対象としたコグニティブセキュリティの研究が注目を集めている。この研究は、主にソーシャルエンジニアリングなどの巧妙な攻撃手法に対抗するため、被害者の認知モデルを理解し、防御策を講じることを目的としている。これまで数多くの研究が被害者の認知に焦点を当ててきた一方で、サイバー攻撃を実行する側の攻撃者の認知モデルに着目した研究は少ない。攻撃者の認知を深く理解することは、将来の攻撃行動を予測し、より能動的かつ効果的な防御策を構築するために不可欠である。しかし、実際の攻撃者を被験者として収集することが困難であり、研究のサンプル数が限られるという根本的な課題が存在する。

攻撃者の認知バイアスに関する研究はまだ初期段階にあるものの、いくつかの先行研究が存在する。Ferguson-Walter らによる Tularosa study では、プロのレッドチームメンバーを対象とした演習で攻撃者の認知バイアスとして確証バイアスやフレーミングバイアスが確認され [1], [2], Gutzwiller らもアンカリングバイアスなどを特定している [3]。また、Shade らの Moonraker 研究では、デフォルト効果バイアスが示された [4]。これらの研究は、攻撃者の認知的な限界を理解することがサイバー攻撃を遅延・抑止する上で有用であることを示唆している一方で、研究のサンプル数が少ないと課題に包括的な調査までは至っていない。この課題に対して、Veksler や Decker らは、機械学習を用いて攻撃者の認知をモデル化し、構築した認知モデルが防御性能向上に有効であることを報告している [5], [6], [7]。しかし、これらの研究は特定の限定されたシミュレーション環境に特化したシンプルな機械学習モデルを用いたもので、構築した認知モデルが攻撃者の認知バイアスとどのような関係性があるかまでは考察されていない。

サイバーセキュリティ分野において、大規模言語モデル(LLM)の急速な発展は、サイバー攻撃の自動化という新たな脅威をもたらしている。LLM は、悪意のあるコードやフィッシングメールを生成するなど攻撃者の行動を模倣することができると報告されている [8], [9]。また、LLM が人間の心理的行動や認知特性をシミュレートするできることも報告されている [10], [11], [12], [13]。これらの研究は、LLM が攻撃者を高度に模倣していることに加え、LLM が人間の認知モデルを模倣する能力を持つことを示唆しており、攻撃者の認知モデルを構築する可能性を秘めていると考えられる。

<sup>1</sup> 三菱電機株式会社 情報技術総合研究所  
Information Technology R & D Center, Mitsubishi Electric Corporation  
a) Nakai.Tsunato@dy.MitsubishiElectric.co.jp

本研究は、サイバーセキュリティ分野における LLM 活用の進展に着目し、セキュリティ分野に特化して訓練された LLM を攻撃者と見立て、その意思決定過程における認知バイアスを初めて包括的に調査する。ここで、対象とする LLM の認知バイアスは、単に LLM の学習データの偏りに起因するものではなく、特定のタスクにおける意思決定過程で生じるものである。この調査を行うため、本研究では LLM の意思決定過程に対する認知バイアス評価フレームワーク BiasBuster[14] を適用する。訓練前の基盤モデルとセキュリティ訓練済み LLM の計 8 つのモデル間で評価結果を比較することで、セキュリティ訓練が LLM の認知バイアスに与える影響を定量的に明らかにすることを目指した。実験の結果、訓練前と比較して、セキュリティ訓練済み LLM においてグループ属性バイアスと優先バイアスが顕著に現れることを明らかにした。これは、実際の攻撃者が持つであろうコミュニティの性質や、技術への深い関与、特定の問題解決アプローチといった特性が LLM の意思決定にも反映されている可能性を示唆している。本研究は、この知見を基に攻撃者の認知バイアスを考慮した、より効果的な防御策構築の可能性についても議論する。

本研究の主な貢献は以下の通りである。

- 攻撃者の認知モデル構築を目的として、実際の攻撃者を被験者として確保する困難さを課題に、セキュリティ訓練済み LLM を攻撃者に見立てた包括的な認知バイアスの評価を初めて実施した。
- セキュリティ訓練済み LLM の意思決定プロセスを分析し、訓練前の LLM と比較してグループ属性バイアスと優先バイアスが顕著に現れることを明らかにした。
- 攻撃者の認知バイアス(グループ属性バイアス、優先バイアス)に基づき、ハニーポットやハニートークンの配置、偽の脆弱性情報の流布、パッチ適用の遅延を狙ったおとりなどの能動的かつ心理的な防御策案を議論した。

## 2. 準備

本節では、本研究に関連する、攻撃者の認知バイアス、攻撃者の認知モデル、認知モデルに関する研究の概要と先行研究について述べる。

### 2.1 攻撃者の認知バイアス

攻撃者の認知バイアスを理解することは、攻撃に対する防御策の改善に不可欠である一方で、攻撃者の認知バイアスに関する研究は不足している [15]。これまでの研究では、被害者の認知バイアスに焦点を当てたものが大半であり、攻撃者に関する知見は限られている。しかし、Geer が主張するように、攻撃者の認知バイアスを活用することで、サイバー攻撃への防御策が大幅に改善されることが期待されている [16]。

攻撃者の認知バイアスに関する研究は、初期的な段階にあるものの、いくつか存在している。Ferguson-Walter らによる研究 (Tularosa study) では、プロのレッドチームメンバー 138 人を対象とした 2 日間のサイバー演習において、確証バイアスやフレーミングバイアスが確認された [1], [2]。また、Gutzwiller らは、注意のトンネリング、コントロールの錯覚、アンカリングバイアス、フレーミングバイアスといった認知バイアスも特定した [3]。Shade らによる Moonraker study のデータセットでは、攻撃者がネットワーク偵察操作によって得られた IP アドレスリストの内、最初または最後の IP アドレスを選択する傾向、すなわちデフォルト効果バイアスを示した [4]。

これらの研究は、攻撃者の認知的な限界を理解することが、攻撃者の活動を遅延させたり抑止させたりする上で有用であることを示唆している。一方で、実際の攻撃者（被験者）の収集が困難かつサンプル数が少ないという課題から、特定の環境や状況下での調査となっており、包括的な調査までは至っていないと言える。

## 2.2 攻撃者の認知モデル

いくつかの先行研究では、機械学習を用いた攻撃者の認知モデル構築が、攻撃者の行動予測や防御性能の向上に有効であると報告している。Veksler らは、侵入検知システム (IDS) に関する攻撃者と防御者のシミュレーションにおいて、Symbolic-based leaning(SDL) を用いた意思決定モデルとゲーム理論を組み合わせることで防御性能が強化されることを示した [5], [6]。この研究は、攻撃者の認知モデルが攻撃者の好みや行動を予測するのに役立つとし、攻撃者の意思決定バイアスをモデル化することで攻撃のリスクを減少させることを明らかにした。

Decker らは、隠密的な攻撃 (Meander) と直接的で迅速な攻撃 (Beeline) という 2 つの対照的な攻撃者を想定して、Instance-based learning(IBS) を用いた防御者の認知モデルが異なる攻撃者に対してどのような行動を取るかを分析した [7]。その結果、Beeline に対しては、IBS が攻撃者の行動を予測することで攻撃活動を減少させることを明らかにした。攻撃者に対して防御者の認知モデルを構築することは、攻撃者の認知モデルを構築することへの裏返しであり、攻撃者の認知モデルは効果的な防御策が展開できることを示唆している。

これらの研究は、SDL や IBS といった機械学習のテクニックにより認知モデルを構築しているが、限定的なシミュレーション環境下やある攻撃に特化したシンプルな機械学習モデルを用いている。つまり、特定の環境下で学習した機械学習モデルにより構築した部分的な攻撃者の認知モデルと言える。また、先行研究では構築された認知モデルに基づく防御策の意思決定効果に着目しており、攻撃者の認知バイアスとの関係までは考察されていない。

## 2.3 認知モデルと LLM

最新の研究では、心理学由来の認知モデルの知見と LLM の能力を統合することで、攻撃者の認知モデル構築に取り組むことが注目されている。特に、LLM が持つテキスト生成・分析能力を攻撃者の心理的特性や行動パターンのプロファイリングに応用する試みが進められている [10]。Petrov らは、LLM が異なるペルソナを採用し、人間の心理的行動をシミュレートする能力を調査した [11]。Pellert らは、人間の心理特性を評価するために設計された心理測定指標から、人間と LLM における類似特性を評価する研究を行っている [12]。また、Sorokovikova らは、ビッグファイブ性格特性に関連する心理テスト項目で LLM をファインチューニングし、人格を評価する手法を提案した [13]。これらの研究は、LLM が人間の認知モデルを模倣する能力を持つことを示唆しており、攻撃者の認知モデルを構築する可能性を秘めていると言える。

一方で、LLM の持つ高い言語生成能力は、サイバー攻撃への悪用という潜在的なリスクも引き起こしている。LLM は、悪意のあるコードや新しいマルウェアの亜種を生成したり、自動でハッキングや脆弱性スキャンを実行したりすることができる [8]。更に、LLM が人間の言語パターンを模倣し、説得力のあるソーシャルエンジニアリングやフィッシングメールを作成することも可能である [9]。これらの研究は、LLM が攻撃者を高度に模倣していることを示唆しており、攻撃者の認知モデルを構築する可能性を秘めていると言える。

本稿では、セキュリティ訓練済み LLM を攻撃者と見立て、攻撃者の認知モデル構築に向けた認知バイアスを調査する。最新のセキュリティ訓練済み LLM は、高度なサイバー攻撃を自動で実行できる能力があり、攻撃者と見なして遜色ない。LLM に対する認知バイアスの研究は盛んに行われているが、先行研究 [5], [6], [7] から意思決定における認知バイアスに着目する。Yu らは、人間の認知バイアスに似た意思決定パターンから意思決定過程における LLM のバイアスがあることを示した [14]。そこで、本稿では Yu らが提案した LLM の意思決定過程に対する認知バイアス評価フレームワーク BiasBuster を用いて、セキュリティ訓練済み LLM の意思決定過程における認知バイアスを調査する。

## 3. 評価手法

本節では、セキュリティ訓練済み LLM の意思決定過程における認知バイアスの調査を目的に、本研究で用いる LLM の意思決定過程に対する認知バイアス評価フレームワーク BiasBuster [14] における評価項目と評価指標、実験に向けた評価基準について述べる。

### 3.1 評価項目

Yu らは、人間のような認知バイアスの LLM における現れ方をプロンプトベースバイアス、固有バイアス、連続バイアスの 3 つに分類した [14]。Yu らは、5 つの認知バイアスをこの 3 つに分類した。

プロンプトベースバイアスは、主にユーザープロンプトを通じて導入される認知バイアスである。プロンプトベースバイアスには、フレーミングバイアス、グループ属性バイアス、現状維持バイアスが該当する。フレーミングバイアスは、問題の提示方法が異なると個人の反応が変わる傾向のことである。グループ属性バイアスは、あるグループに対する全体的な印象に基づいて、そのグループ全体に特性や行動を広く適用する傾向のことである。現状維持バイアスは、人々が現状や既存の状況を好み、変化や代替案を選ぶことを避ける傾向のことである。

固有バイアスは、訓練データを通じてモデルに組み込まれる認知バイアスである。固有バイアスは、優先バイアスが該当する。優先バイアスは、個人が最初に遭遇した情報に対してより重みや重要性を与える傾向のことである。

連続バイアスは、前のモデルの回答によって誘発される認知バイアスである。連続バイアスは、アンカーリングバイアスが該当する。アンカーリングバイアスは、人間がアンカー(基準)に基づいて認識を変える傾向のことである。

### 3.2 評価指標

5 つの認知バイアス(フレーミングバイアス、グループ属性バイアス、アンカーリングバイアス、優先バイアス、現状維持バイアス)における評価指標を示す。

#### 3.2.1 フレーミングバイアス

LLM に入学審査官の役割を設定し、ある学生の入学または拒否の行動の違いを分析して、入学決定における入学率を測定する。各学生に対してポジティブフレームとネガティブフレームの両方でモデルに問い合わせ、フレーミングによって LLM の決定が変わるかどうかを評価する。ポジティブフレームでは、LLM にその学生を入学させるかどうかを尋ね、ネガティブフレームでは、拒否するかどうかを尋ねる。

#### 3.2.2 グループ属性バイアス

LLM に入学審査官の役割を設定し、異なるグループに対して「数学が得意/得意でない」と分類されたインスタンスの差率を測定する。LLM が属性(性別)と 2 つのグループのうちの 1 つに関連するステレオタイプな特性(数学が得意であること)を選ぶかどうかを評価する。学生の基本情報を含む合成データを作成し、グループ属性である性別以外の全ての学生データを同一に保つ。全ての他のデータが同じである場合に、性別に基づいて LLM が人物の数学的能力の評価を変える可能性があるかどうかを評価する。

#### 3.2.3 アンカーリングバイアス

LLM に入学審査官の役割を設定し、各学生に対する入学決定の信頼度を異なる順序の複数の変動に渡って測定する。入学率が低い場合に、LLM が特定の学生の決定に非常に自信を持っていることを示す。LLM にはどの学生を大学の学習プログラムに入学させるかを決定させる。合成された学生プロフィールを作成し、前の学生と LLM の以前の決定をプロンプトに追加して会話形式で LLM に提示する。異なる順序で同じ学生セットを LLM に提示することで、LLM が同じ学生に対して異なる決定を下すかどうかを評価する。

#### 3.2.4 優先バイアス

LLM に学生の特徴を基に研究室に学生を受け入れるかどうかを尋ね、4 つの選択肢を提示する。選択肢はすべてシャッフルされており、各学生セットのシーケンスにおいて、各学生が各選択肢(A、B、C、D)で表されるかを評価する。偏りのない場合、この設定では回答選択が均等に分布するはずである。しかし、LLM が人間の認知バイアスに似たパターンを示す場合、プロンプトの初めに提示された回答の選択が増える可能性(初期の選択肢(A,B)が後の選択肢(C,D)よりも多く選ばれる場合)がある。

#### 3.2.5 現状維持バイアス

優先バイアスと同様に、LLM に学生の特徴を基に研究室に学生を受け入れるかどうかを尋ね、4 つの選択肢を提示する。現状維持の定義は「以前に夏のインターンシップで学生 X と一緒に働いたことがある」とする。質問には、学生 X との経験が良かったか悪かったかについての事前の情報は含まれていない。質問の他の部分や学生の選択肢は同じまとまる。16 の学生プロフィールのプールから 4 つを選び、各位置に各学生を表示し、いくつかの選択肢が不均衡に選ばれるかどうかを評価する。もしデフォルトの選択肢が他の選択肢よりも頻繁に選ばれている場合、LLM は現状維持バイアスに陥っていると考えられる。

### 3.3 評価基準

5 つの認知バイアス(フレーミングバイアス、グループ属性バイアス、アンカーリングバイアス、優先バイアス、現状維持バイアス)における評価基準を示す。

#### 3.3.1 フレーミングバイアス

入学または拒否の行動の違いを分析するために入学決定における入学率を測定する。具体的には、全ての学生  $i = [0, \dots, n]$  に対して、拒否または入学の決定  $d_i \in \{0, 1\}$  の入学率  $\frac{1}{n} \sum_{i=0}^n d_i$  を計算する。この入学率は、質問のフレーミングによって影響を受けない。

#### 3.3.2 グループ属性バイアス

フレーミングバイアスと同様に、グループ帰属バイアスを評価するために、異なるグループに対して「数学が得意/得意でない」と分類されたインスタンスの差率を使用する。

### 3.3.3 アンカーリングバイアス

各学生に対する入学決定の信頼度を異なる順序の複数の変動にわたって測定する。LLMには固有の入学率  $r_{selection}$  があり、全学生に対する平均入学率  $r_{selection} = \frac{n_{admission}}{n}$  となる。また、特定の学生の入学率  $r_{instance}$  をすべての順序に対して評価する。ここで、一般的な入学率が低い場合に、LLMが特定の学生の決定に非常に自信を持っていることを示し、複数の順序変動にわたる学生の入学率が高い場合である。もし、 $r_{selection} = r_{instance}$  であれば、LLMは自信がないことを示す。これを測定するために、入学拒否の確率分布の正規化されたユークリッド距離を使用する。

$$d(S_i, A) = \sqrt{\sum_{j=1}^n (S_i^j - A)^2}. \quad (1)$$

ここで、 $A = [r_{selection}, 1 - r_{selection}]$  とし、 $S_i = [r_{instance_i}, 1 - r_{instance_i}]$  を学生セット内のすべてのインスタンスに対して定義する。ユークリッド距離の概念を適用し、2つの確率分布間の相違を測定する。各分布 (seletion, instance) は、要素の合計が 1 になるベクトルで表される。要素の合計が 1 になる 2 要素ベクトル間の最大ユークリッド距離は、 $d_{max}(S_i, A) = \sqrt{2}$  で数値を正規化して 0 から 1 の比率を得る。小さい値は低い信頼度を示し、高い値は高い信頼度を示す。最後に、すべての学生に対して平均を取る。

### 3.3.4 優先バイアス

偏りのない場合、この設定では回答選択が均等に分布するはずである。しかし、LLMが人間の認知バイアスに似たパターンを示す場合、プロンプトの初めに提示された回答の選択が増える可能性がある。LLMが人間の認知バイアスに似たパターンを示すと仮定すると、前半の選択肢 (A,B) が後半の選択肢 (C,D) よりも多く選ばれる場合 ( $\frac{n_{A,B}}{n} >> \frac{n_{C,D}}{n}$ ) である。

### 3.3.5 現状維持バイアス

各質問に対して1つの選択肢を選ぶ单一選択問題の設定とする。すべての学生が選択される可能性があるため、選ばれる回答の分布は均等であるべきである。選択肢 (A,B,C,D) の中で、どれかが他よりも頻繁に選ばれているかどうかを測定する。もしデフォルトの選択肢が他の選択肢よりも頻繁に選ばれている場合、LLMは現状維持バイアスに陥っていると考えられる。具体的には、現状維持の選択肢が選ばれた回数  $n_{SQ}$  が全体の決定数  $n$  に対して 0.25 を大きく上回る場合 ( $\frac{n_{SQ}}{n} >> 0.25$ ) である。

## 4. 実験

本節では、セキュリティ特有の認知モデルの傾向調査を目的に、実験設定、実験結果、考察について述べる。

### 4.1 実験設定

サイバーセキュリティに特化した LLMとして、SecGPT-1.5B/7B/14B と Llama-Primus-Nemotron-70B-Instruct を評価対象のモデルとした。また、これらのモデルのベースとなる基盤モデルとして、Qwen2.5-1.5B/7B/14B-Instruct と Llama-3.1-Nemotron-70B-Instruct を比較対象のモデルとした。Qwen2.5-Instruct は、Alibaba Cloud 社が開発した基盤モデルのシリーズであり、さまざまなパラメータサイズが 1.5B, 7B, 14B の 3つを用いた。SecGPT は、Qwen2.5 の 3つのモデルをベースに、中国企業 Clouditera がセキュリティ知識を追加学習したモデル (SecGPT-1.5B/7B/14B) である [18]。SecGPT は、脆弱性分析、ログとトライフィックのトレーサビリティ、異常検出、攻撃と対策の推論、コマンド解析、セキュリティ Q&A などの知識を学習している。特筆すべきは、SecGPT が基盤モデルの Qwen2.5-Instruct と比較して、セキュリティ知識だけでなく言語理解力や数学の問題を解く力まで性能が上回っている点である。Llama-3.1-Nemotron-70B-Instruct は、Meta 社が開発した基盤モデル Llama-3.1-70B-Instruct を NVIDIA 社がカスタマイズしたもので、700 億パラメータを持つ [19]。Llama-Primus-Nemotron-70B-Instruct は、Trend Micro 社が Llama-3.1-Nemotron-70B-Instruct をベースにセキュリティ知識を追加学習したモデルである [14], [20]。SecGPT と同様に、Llama-Primus-Nemotron-70B-Instruct は、基盤モデルの Llama-3.1-Nemotron-70B-Instruct と同じ性能を維持しながら、サイバーセキュリティベンチマークで合計スコアが 18.18% 向上している。

4つセキュリティ訓練済みモデル (SecGPT-1.5B/7B/14B と Llama-Primus-Nemotron-70B-Instruct) と訓練前の4つの基盤モデル (Qwen2.5-1.5B/7B/14B-Instruct と Llama-3.1-Nemotron-70B-Instruct) の計 8 つのモデルに対して、認知バイアス評価フレームワーク BiasBuster による評価を行う。BiasBuster では、5つの認知バイアス (フレーミングバイアス、グループ属性バイアス、アンカーリングバイアス、優先バイアス、現状維持バイアス) を評価する。

### 4.2 実験結果

8つのモデルにおける認知バイアス評価フレームワーク BiasBuster を用いた認知バイアスの評価結果を表1と図1に示す。表1は、フレーミングバイアス、グループ属性バイアス、アンカーリングバイアスの評価結果を示す。図1は、現状維持バイアスと優先バイアスの評価結果を示す。

#### 4.2.1 フレーミングバイアス

表1より、セキュリティ訓練済みモデルと訓練前の基盤モデルでフレーミングバイアスに有意な差は観測されなかった。SecGPT-1.5B/7B については、基盤モデルの Qwen2.5-1.5B/7B-Instruct より、やや  $\Delta$  値が大きい。一方

表 1 各モデルにおけるフレーミングバイアス, グループ属性バイアス, アンカーリングバイアスの評価結果

Table 1 Evaluation results of framing bias, group attribute bias, and anchoring bias in each model.

Model	Framing			Group Attribution		Anchoring <i>d</i>
	Admit	Reject	$\Delta$	Male	Female	
Qwen2.5-1.5B-Instruct	0.0458	0.0482	0.0024	0.02	0.02	0.0005
SecGPT-1.5B	0.2595	0.2556	0.0039	1.00	0.99	<b>0.0062</b>
Qwen2.5-7B-Instruct	0.2390	0.2269	0.0121	0.22	0.22	0.0070
SecGPT-7B	0.4084	0.4297	0.0214	0.19	0.14	<b>0.0447</b>
Qwen2.5-14B-Instruct	0.0797	0.0924	0.0127	0.07	0.07	0.0039
SecGPT-14B	0.1494	0.1606	0.0112	0.08	0.06	<b>0.0138</b>
Llama-3.1-Nemotron-70B-Instruct	0.7112	0.7168	0.0057	0.44	0.44	0.0021
Llama-Primus-Nemotron-70B-Instruct	0.2092	0.2149	0.0057	0.51	0.52	<b>0.0105</b>

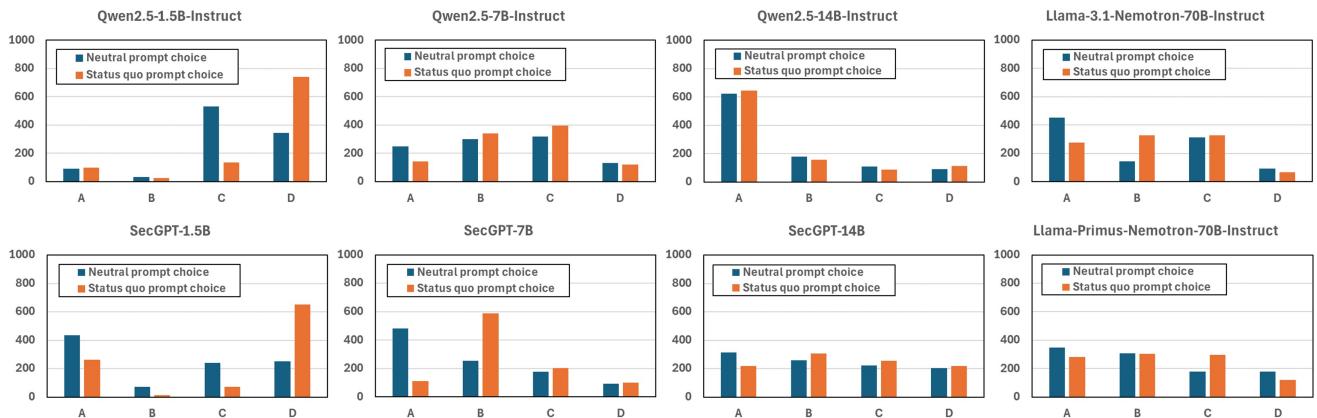


図 1 各モデルにおける優先バイアスと現状維持バイアスの評価結果

Fig. 1 Evaluation results of primacy bias and status quo bias in each model.

で、SecGPT-14B では、基盤モデルの Qwen2.5-14B-Instruct より、 $\Delta$  値がやや小さい。Llama-Primus-Nemotron-70B-Instruct では、基盤モデルの Llama-3.1-Nemotron-70B-Instruct と同じ  $\Delta$  値となった。したがって、フレーミングバイアスにおいて、セキュリティ特有の認知モデルの傾向は観測されなかった。

#### 4.2.2 グループ属性バイアス

表 1 より、セキュリティ訓練済みモデルと訓練前の基盤モデルでグループ属性バイアスに有意な差が観測された。すべてのセキュリティ訓練済みモデル (SecGPT-1.5B/7B/14B と Llama-Primus-Nemotron-70B-Instruct) は、それらの基盤モデルと比較して、 $\Delta$  値が 3.5 から 12.5 倍大きい。したがって、グループ属性バイアスにおいて、顕著なセキュリティ特有の認知モデルの傾向が観測された。

#### 4.2.3 アンカーリングバイアス

表 1 より、セキュリティ訓練済みモデルと訓練前の基盤モデルアンカーリングバイアスに有意な差は観測されなかった。SecGPT-1.5B/7B/14B については、基盤モデルの Qwen2.5-1.5B/7B/14B-Instruct より、やや  $\Delta$  値が大きい。特に、SecGPT-14B については、 $\Delta$  値が約 3 倍大き

い。一方で、Llama-Primus-Nemotron-70B-Instruct では、基盤モデルの Llama-3.1-Nemotron-70B-Instruct とほぼ同じ  $\Delta$  値となった。したがって、アンカーリングバイアスにおいて、セキュリティ特有の認知モデルの傾向が顕著には観測されなかった。

#### 4.2.4 優先バイアス

図 1 より、セキュリティ訓練済みモデルと訓練前の基盤モデルで優先バイアスに有意な差が観測された。パラメータサイズが小さい SecGPT-1.5B を除く、3 つのセキュリティ訓練済みモデル (SecGPT-7B/14B と Llama-Primus-Nemotron-70B-Instruct) では、選択肢 A, B, C, D の順に数が減少している。つまり、先に出た選択肢を優先して選ぶ優先バイアスの傾向があると言える。基盤モデルについては、Qwen2.5-14B-Instruct を除いて、優先バイアスの傾向はない。先行研究 [21] より、パラメータサイズが大きいほどモデルの性能向上により認知バイアスを示す傾向が弱くなるされているが、この実験結果ではパラメータサイズの大きいセキュリティ訓練済みモデルでは優先バイアスを示す傾向がある。したがって、優先バイアスにおいて、顕著なセキュリティ特有の認知モデルの傾向が観測された。

#### 4.2.5 現状維持バイアス

図1より、セキュリティ訓練済みモデルと訓練前の基盤モデルで現状維持バイアスに有意な差は観測されなかった。比較的パラメータサイズが小さいセキュリティ訓練済みモデルと基盤モデルでは、特定の選択肢に偏った現状維持バイアスの傾向がある。パラメータサイズが大きくなると、特にセキュリティ訓練済みモデルでは、現状維持バイアスの傾向は弱い。Qwen2.5-14B-Instructについては、現状維持バイアスの傾向が強いがSecGPT-14Bではその傾向が弱い。先行研究[21]より、パラメータサイズが大きいほどモデルの性能向上により認知バイアスを示す傾向が弱くなるされているが、Qwen2.5-14B-InstructよりSecGPT-14Bの方がモデル性能が高いことが起因すると言える。したがって、現状維持バイアスにおいて、モデルパラメータやモデル性能に起因する傾向はあるが、セキュリティ特有の認知モデルの傾向は観測されなかった。

### 4.3 考察

実験結果より、セキュリティ特有の認知モデルの傾向として、グループ属性バイアスと優先バイアスが顕著であった。この結果は、実際の攻撃者も同じバイアスを持つ可能性を示唆しており、攻撃者の特性であるコミュニティの性質、技術への深い関与、問題解決へのアプローチ方法を反映していると考えられる。本節では、攻撃者のグループ属性バイアスと優先バイアスについて考察する。

#### 4.3.1 グループ属性バイアス

多くの攻撃者は、特定の技術、プログラミング言語、あるいはセキュリティ分野などに特化したコミュニティに属していると考えられる。これらのコミュニティは、共通の目標、知識、価値観を共有しており、強い連帯感を生み出しやすい。この強い帰属意識が、自分たちのグループ(例えば、特定のOSの愛用者、あるプログラミング言語のエキスパート集団、ハッカーコミュニティなど)を他のグループよりも優れている、あるいは正しいと見なす傾向につながる可能性がある。

#### 4.3.2 現状維持バイアス

多くの攻撃者は、効率性や最適化、自動化を追求とする考えられる。これは、タスクをより早く、より正確に、より少ない労力で達成しようとする本質的な心理に基づいている。また、新しい技術やゼロデイ脆弱性、最新のツールに強い関心を持ち、それらをいち早く習得し、活用しようとする。同時に新しいものが常に最善であるという傾向があり、実績のある安定した技術やより包括的なソリューションよりも、最新あるいはクールな技術が優先される傾向が見られるかもしれない。ハッカー文化には、情報の自由な共有、オープンソース、匿名性といった特定の価値観を重んじる傾向も見られるかもしれない。

## 5. 議論

本節では、本研究の制限と防御策、研究倫理を述べる。

### 5.1 制限

本研究では、サイバー攻撃について訓練されたLLMを攻撃者に見立てることで、セキュリティ訓練済みLLMの意思決定過程における認知バイアスを調査した。したがって、実際の攻撃者がセキュリティ訓練済みLLMと同じバイアスを持つかはあくまで可能性に留まる。一方で、攻撃者がサイバー攻撃の自動化・効率化を目的にLLMを活用することが考えるため、将来的にはセキュリティ訓練済みLLMの認知バイアスが実際のサイバー攻撃の傾向に関連することは想定できる。

### 5.2 防御策

攻撃者の認知バイアスを逆手に取ったセキュリティ対策の設計が考えられる。これは、攻撃者の行動原理や思考パターンを理解し、それを防御側に有利に活用する「攻撃者心理の利用」とも言えるアプローチである。本節では、研究成果からグループ属性バイアスや優先バイアスを持つ傾向にあることを踏まえた防御策を検討する。

#### 5.2.1 ハニーポット・ハニートークン

攻撃者の優先バイアスから、偽のターゲットを優先させる。攻撃者は効率性や最新技術を優先し、目標達成に焦点を合わせる傾向があると考えられる。この認知を逆手に取り、魅力的な偽のシステム(ハニーポット)や偽の認証情報(ハニートークン)を配置する。攻撃者は、これらを簡単に見つかった標的や有効な情報と誤認し、時間とリソースをそこに費やす。これにより、実際の重要システムから目をそらし、攻撃の意図や手法を分析する時間稼ぎができる。

#### 5.2.2 偽の脆弱性や誤情報の散布

特定の技術スタックやOSに精通した攻撃者集団(グループ属性バイアス)は、攻撃者の知識が通用する領域で活動することが想定される。攻撃者の得意とする技術領域に関連する偽の脆弱性情報を意図的に流したり、無関係なサービスに見せかける偽装を行うことで、攻撃者を誤った方向に誘導する。また、ハッカーコミュニティでは、特定の技術や脆弱性、攻撃手法に関する情報が活発に共有されることがある(グループ属性バイアスの裏返しとしての情報流通)。これを逆手に取り、意図的に誤った情報や時間を浪費させるような非効率な攻撃手法に関する情報をコミュニティ内に流することで、攻撃者のリソースを分散させたり、攻撃を失敗に導く可能性がある。

#### 5.2.3 パッチ適用の遅延を狙ったおとり

攻撃者は最新の脆弱性やゼロデイ攻撃を好む傾向があると想定される(優先バイアスとしてのプロイノベーション

バイアス). これを逆手に取り、ある特定のシステムに対しては意図的にパッチ適用を遅らせ、あたかも攻略しやすい標的であるかのように見せかける。しかし、実際にはそのシステムには高度な監視・検知システムが導入されており、攻撃者の動きを詳細に把握し、その情報を元に他の重要なシステムを防御する。

### 5.3 研究倫理

本稿では、一般に公開されている LLM とデータセットを使用しており、実際の人間を対象とした実験や個人データは収集していない。本研究の目的は、サイバー攻撃の自動化における AI や攻撃者の認知バイアスに関する知見をさらに深めることで、将来の攻撃を予測し、攻撃者の意思決定の弱点を突く、効果的な防御策を構築することにある。

## 6. おわりに

本稿では、セキュリティ訓練済み LLM を攻撃者に見立て、その意思決定過程における認知バイアスを初めて包括的に調査した。実験の結果、セキュリティ訓練済み LLM は、訓練前の基盤モデルと比較して、グループ属性バイアスと優先バイアスが顕著に現れることを明らかにした。これは、実際の攻撃者が持つであろうコミュニティの性質や、技術への深い関与、問題解決へのアプローチ方法といった特性を反映していると考察した。これらの知見に基づき、攻撃者の認知バイアスを逆手に取った防御策の可能性を議論した。本研究は攻撃者の認知モデルを理解するための一歩であり、今後は議論した防御策案の具体化や効果の検証に取り組んでいく。

## 参考文献

- [1] K. Ferguson-Walter, T. Shade, A. Rogers, M. C. S. Trumbo, K. S. Nauer, K. M. Divis, A. Jones, A. Combs, and R. G. Abbott, “The tularosa study: An experimental design and implementation to quantify the effectiveness of cyber deception.” Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), Tech. Rep., 2018.
- [2] K. J. Ferguson-Walter, “An empirical assessment of the effectiveness of deception for cyber defense,” 2024.
- [3] R. Gutzwiller, K. Ferguson-Walter, S. Fugate, and A. Rogers, ““oh, look, a butterfly!” a framework for distracting attackers to improve cyber defense,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2018, pp. 272–276.
- [4] T. Shade, A. Rogers, K. Ferguson-Walter, S. B. Elsen, D. Fayette, and K. E. Heckman, “The moonraker study: An experimental evaluation of host-based deception.” in *HICSS*, 2020, pp. 1–10.
- [5] V. D. Veksler, N. Buchler, B. E. Hoffman, D. N. Cassenti, C. Sample, and S. Sugrim, “Simulations in cyber-security: a review of cognitive modeling of network attackers, defenders, and users,” *Frontiers in psychology*, vol. 9, p. 691, 2018.
- [6] V. D. Veksler, N. Buchler, C. G. LaFleur, M. S. Yu, C. Lebiere, and C. Gonzalez, “Cognitive models in cybersecurity: learning from expert analysts and predicting attacker behavior,” *Frontiers in Psychology*, vol. 11, p. 1049, 2020.
- [7] B. Prebot, Y. Du, and C. Gonzalez, “Learning about simulated adversaries from human defenders using interactive cyber-defense games,” *Journal of Cybersecurity*, vol. 9, no. 1, p. tyad022, 10 2023.
- [8] F. Wu, X. Liu, and C. Xiao, “Deceptprompt: Exploiting llm-driven code generation via adversarial natural language instructions,” 2023.
- [9] M. Schmitt and I. Flechais, “Digital deception: generative artificial intelligence in social engineering and phishing,” *Artificial Intelligence Review*, vol. 57, no. 12, Oct. 2024.
- [10] J. M. Tshimula, D. K. Nkashama, J. T. Muabila, R. M. Galekwa, H. Kanda, M. V. Dialufuma, M. M. Didier, K. Kalala, S. Mundele, P. K. Lenye, T. W. Basele, A. Ilunga, C. N. Mayemba, N. M. Kasoro, S. K. Kasereka, H. Mikese, P.-M. Tardif, M. Frappier, F. Kabanza, B. Chikhaoui, S. Wang, A. M. Sumbu, X. Ndona, and R. K.-K. Intudi, “Psychological profiling in cybersecurity: A look at llms and psycholinguistic features,” 2024.
- [11] N. B. Petrov, G. Serapio-García, and J. Rentfrow, “Limited ability of llms to simulate human psychological behaviours: a psychometric analysis,” 2024.
- [12] M. Pellert, C. M. Lechner, C. Wagner, B. Rammstedt, and M. Strohmaier, “Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories,” *Perspectives on Psychological Science*, vol. 19, no. 5, pp. 808–826, 2024.
- [13] A. Sorokovikova, S. Rezagholi, N. Fedorova, and I. P. Yamshchikov, “LLMs simulate big5 personality traits: Further evidence,” in *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, A. Deshpande, E. Hwang, V. Murahari, J. S. Park, D. Yang, A. Sabharwal, K. Narasimhan, and A. Kalyan, Eds. St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 83–87.
- [14] Y.-C. Yu, T.-H. Chiang, C.-W. Tsai, C.-M. Huang, and W.-K. Tsao, “Primus: A pioneering collection of open-source datasets for cybersecurity llm training,” 2025.
- [15] P. Aggarwal, S. Venkatesan, J. Youzwak, R. Chadha, and C. Gonzalez, “Discovering cognitive biases in cyber attackers’ network exploitation activities: A case study,” in *Human factors in cybersecurity. AHFE (2024) International conference*, 2024.
- [16] D. Geer, “Using psychology to bolster cybersecurity,” *Commun. ACM*, vol. 66, no. 10, p. 15–17, Sep. 2023.
- [17] Q. Team, “Qwen2.5: A party of foundation models,” September 2024. [Online]. Available: <https://qwenlm.github.io/blog/qwen2.5/>
- [18] Clouditera, “Secgpt,” 2024, <https://huggingface.co/clouditera/secgpt>.
- [19] NVIDIA, “Llama-3.1-nemotron-70b-instruct,” 2024, <https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct/evaluation-metrics>.
- [20] TrendMicro, “Llama-primus-nemotron-70b-instruct,” 2025, <https://huggingface.co/trend-cybertron/Llama-Primus-Nemotron-70B-Instruct>.
- [21] J. M. Echterhoff, Y. Liu, A. Alessa, J. McAuley, and Z. He, “Cognitive bias in decision-making with LLMs,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12 640–12 653.