LLMを利用したDiscord上のサイバー犯罪関連の隠語の調査

川口 大翔[1,a)]　インミンパパ[2]　吉岡 克成[3]

概要：近年，Discord や Telegram などのソーシャルメディアプラットフォームにおいて，ハッキングコミュニティの活動が活発化している．これらのコミュニティでは，隠語が重要な情報の隠蔽に使用されることが多い．隠語の特定は，コミュニケーションの文脈理解，新たなコミュニティの発見，および攻撃者の性質理解に重要である．しかし，隠語は時間とともに変化するため，柔軟な検出手法が求められる．本研究では，近年精度が向上している大規模言語モデル (LLM) を用いた隠語抽出手法の有効性を検証した．LLM による隠語抽出の先行研究は存在するが，SNS プラットフォーム上のサイバー犯罪関連データに特化した研究は限られている．そこで，Discord から収集したサイバー犯罪関連の会話データを用いて，複数の LLM モデルの性能を評価した．その結果，本タスクに適した LLM モデルを特定した．さらに，収集したデータの分析により，Discord 内で使用されているサイバー犯罪関連の隠語の一部を明らかにした．

キーワード：Discord，サイバー犯罪，LLM, 隠語

Investigation of Cybercrime-Related Slang on Discord Using Large Language Models

Yamato Kawaguchi[1,a)]　Yin Minn Pa Pa[2]　Katsunari Yoshioka[3]

Abstract: In recent years, hacking communities have become increasingly active on social media platforms such as Discord and Telegram. In these communities, slang is often used to conceal important information. Identifying slang is crucial for understanding communication context, discovering new communities, and comprehending the nature of attackers. However, as slang evolves over time, flexible detection methods are required. This study examines the effectiveness of slang extraction methods using large language models (LLMs), which have shown improved accuracy in recent years. While previous research on slang extraction using LLMs exists, studies focused specifically on cybercrime-related data from social networking platforms are limited. Therefore, we evaluated the performance of multiple LLM models using cybercrime-related conversation data collected from Discord. As a result, we identified the LLM model best suited for this task. Furthermore, through analysis of the collected data, we revealed some of the cybercrime-related slang used within Discord.

Keywords: Discord, Cybercrime, LLM, slang

[1] 横浜国立大学大学院環境情報学府
Graduate School of Environment and Information Sciences, Yokohama National University
[2] 横浜国立大学大学院先端科学高等研究院
Institute of Advanced Sciences, Yokohama National University
[3] 横浜国立大学大学院環境情報研究院/先端科学高等研究院
Faculty of Environment and Information Sciences, Yokohama National University / Institute of Advanced Sciences, Yokohama National University
[a)] kawaguchi-yamato-jc@ynu.jp

1. はじめに

近年，Discord や Telegram などのソーシャルメディアプラットフォームにおいて，ハッキングコミュニティの活動が活発化している [1]．特に，LAPSUS$ [2] や Scattered Spider のような若年層のハッカーグループが Discord 上にも存在し，企業へのサイバー攻撃や現実世界での違法行為を行っているため [3]，Discord におけるサイバー犯罪の

調査も重要性を増している．

これらの Discord のサイバー犯罪のコミュニティでは，隠語が使用されていて，その隠語を理解できなければサイバー犯罪の調査での文脈の理解が難しい．例えば，一見意味不明な “wts fresh cc 20dl pp” というメッセージも，“wts” が「販売したい (want to sell)」，“fresh cc” が「最近有効性が確認されたクレジットカード」，“20dl pp” が「PayPal で $20 を支払う」という隠語の知識があれば，「最近有効性が確認されたクレジットカードを，PayPal 支払いで$20 で販売している」という意味だと解読できる．隠語の分析により，攻撃者の文章をより深く理解し，さらなる分析に活用できる．

サイバー犯罪の隠語は，サイバー犯罪の種類毎に独自であり [4][5]，それらと関係するコミュニティを効率よく発見し対策するには, 彼らが利用する隠語を正確かつ迅速に検知する必要がある．しかし, SNS 上の膨大のメッセージから隠語を抽出するにはそれぞれのサイバ犯罪コミュニティに関する知識が必要である．またテキストから隠語の可能性が高い単語を抽出するタスクを手動で行うには労力がかかり, 自動的な仕組みが必要である．

そこで，近年，大規模言語モデル (LLM) を使った自然言語処理などが注目され，隠語抽出においても有効であると考えられる．特に LLM はコンテキストの理解や複雑なパターンの認識において優れており，一部のタスクでは従来の機械学習を直接的に用いた手法と比較して大きな利点がある [6]. Fillies らの研究でも LLM による隠語抽出が有効であることを明らかにしている [7]．ただし，この研究では映画字幕をデータセットとして使用している．映画字幕に含まれる隠語は，ある程度整形されていて，多くの人が理解できる，抽出しやすい隠語であると考えられる．しかし， LLM によるサイバー犯罪関連の隠語抽出能力，特にコミュニティ外からの理解が難しい専門的な隠語に関しては，まだ十分な検証が行われていない．さらに，多様な LLM モデルが存在する中で，各モデルは異なる特徴を持ち，特定のタスクに対する適性が異なる可能性があるが，サイバー犯罪関連の隠語抽出に最適なモデルの特定は行われていない．したがって，SNS 上のサイバー犯罪関連のメッセージにおける LLM の隠語抽出能力の評価，および様々な LLM モデルの隠語抽出性能の比較は，重要な研究課題として残されている．

そこで本研究は，Discord から収集したメッセージを使い，サイバー犯罪の隠語抽出に適した LLM モデルを明らかにし，サイバー犯罪に関係する隠語を抽出して，Discord におけるサイバー犯罪の隠語を調査する．具体的には以下の Research Questions (RQ) を設定する．

(1) **RQ1: LLM はサイバー犯罪関連の隠語抽出に有効か．最適な LLM モデルはどれか.**

(2) **RQ2: Discord におけるサイバー犯罪関連の隠語にはどのようなものがあるか.**

RQ1 を回答するため，Discord のメッセージからサイバー犯罪関連の隠語の評価データセットを作成し, 31 種類の商用・オープンソースの LLM モデルを比較した．その結果，最も高い性能を示したのは gpt-4o-mini–2024-07-18 [8] をファインチューニングしたモデルで，精度 0.89，再現率 0.88, F1 スコア 0.85 を達成した．また，mistral-large:123b や llama2-uncensored:70b などの一部のオープンソースモデルも高い性能を示し，サイバー犯罪関連の隠語抽出に適していることが明らかになった．

RQ2 を回答するため, ファインチューニングした gpt-4o-mini-2024-07-18 を利用し, 98,808 件のメッセージから隠語を抽出したところ，11 種類サイバー犯罪と関連する 810 件の隠語を発見した．さらに, サイバー犯罪の種類毎の隠語を分析した結果, Discord 上では通常のポルノ, ダークーマーケット, ゲーム上の不正行為, 詐欺, ハッキング, 児童ポルノ, デジタル資産, 賭博, サイバーいじめ, 知的財産権の侵害, 違法薬物などの 11 種類のサイバー犯罪と関係するコミュニティがあり, 主に通常のポルノ，ダークマーケット，ゲーム上の不正行為の 3 種類のサイバー犯罪が集中していたことを判明した．

2. 背景

2.1 隠語

隠語は，特定の集団やコミュニティでのみ通じる，隠された意味を持つ言葉や表現を指す．本研究では，そのような隠語を文章から抽出するタスクを行う．だが，特定のデータセットのみから「特定の集団やコミュニティでのみ通じる」という特徴を特定することは難しい．そこで本研究では，隠語にみられる特徴から隠語の定義を行う．Hughes らの隠語の分類 [9] を参考に，新たに「辞書に書かれた意味以上の意味を持つ」と「固有名詞だが，広く使われていて一般的な内容を指す」という項目を設けて，9 つの条件を定めた．次の 9 件の条件のうち 1 つ以上を満たす単語を隠語と定義する: 1. 未知のスペリングへの変更 (“abortion” → “@b0rt!0n”)，2. 既知のスペリングへの変更 (“porn” → “corn”)，3. 略語 (“sexual assault” → “SA”)，4. 絵文字による表現 (porn content → 桃の絵文字)，5. 言い換え (“suicide” → “unalive”)，6. 既存の言葉の目的外の使用 (“sex workers” → “Accountant”)，7. 音声の類似性 (“Nazi” → “not see”)，8. 辞書に書かれた意味以上の意味を持つ (“teen content” → “teen porn content”)，9. 固有名詞だが，広く使われていて一般的な内容を指す (“onlyfans” は Web サイト名だが，同様のサイトも含めて使われている)．また，一般的な専門用語は除外する．表記揺れによる重複が生じた場合，より一般的な形式を隠語とする．

2.2 Discord メッセージのデータセット

Discord は，6 億 1400 万人以上が利用するオンラインのメッセージングプラットフォームである [10]．ユーザーはテキスト，画像，音声，ビデオを通じてコミュニケーションを取ることができ，「Discord サーバ」と呼ばれる専用のコミュニティを作成できる．各 Discord サーバの中には複数の「チャンネル」というスペースに分けられ，特定のトピックや目的に応じてメッセージを投稿できる．

本研究では，我々の投稿予定の研究で提案するシステムを使用して収集された，サイバー犯罪に関連する内容を含む Discord のメッセージのデータセットを使用した．データ収集については研究倫理審査を受けている．このデータセットは 301 件の Discord サーバ，955 件のチャンネルに投稿された，98,808 件のメッセージが含まれている．

2.3 大規模言語モデル (LLM)

本研究では，商用 LLM とオープンソース LLM の両方を使用する．商用 LLM は一般的に高性能であるが，使用には料金が発生する．ただし，gpt-4o-mini-2024-07-18 [8] などの一部の廉価版モデルは，標準モデルと比較して低コストで利用可能である．一方，オープンソース LLM は近年，商用 LLM に匹敵する性能を持つモデルが登場している．これらは無料で使用可能であり，オンプレミスでの実行やカスタマイズの自由度が高いという利点がある．特に，モデルによっては倫理的考慮による入出力内容のモデレーションがあり，サイバー犯罪系の隠語抽出を行えない場合があるため，自由度が高いことは重要である．本研究で使用した LLM の一覧を表 3 に示す．これらのモデルは，商用 LLM とオープンソース LLM の両方から，サイバー犯罪系の入力でも拒否されず，性能が高く，注目されているモデルを中心に選択した．商用のモデルは，名前が “gpt-” から始まる 6 件のモデルである．

3. 関連研究

隠語抽出の研究では，様々な手法が提案されている．Yuan ら [11] は，文章内の単語とコーパス内の単語のベクトルを比較することで，サイバー犯罪マーケットから自動的に隠語を識別・理解するシステムを提案した．彼らの手法は薬物とサイバー犯罪製品名の検出において高い精度を示したが，より広範な隠語の検出や高品質なコーパスの必要性など，いくつかの課題が残されている．林ら [12] は，ダークウェブ上のハイパーリンク構造を利用した隠語検出手法を提案した．しかし，この手法は Discord のようなリンク構造のないプラットフォームには適用できない．Fillies ら [7] は，LLM を用いて隠語を検出することができると明らかにした．ただし，彼らは映画の字幕データを用いて LLM の有効性を示したが，サイバー犯罪特有の隠語は対象としていない．

本研究の新規性は，主に 3 点ある．まず，大規模言語モデル (LLM) を活用してサイバー犯罪関連のメッセージから隠語を抽出する手法を新たに提案する．次に，サイバー犯罪関連の隠語抽出に最適な LLM の特定を行う．さらに，Discord プラットフォーム上のサイバー犯罪関連の隠語に焦点を当てた包括的調査を実施する．

4. 調査手法

4.1 RQ1 に対する調査手法

4.1.1 事前調査

すべてのメッセージに隠語が含まれているわけではなく，評価用とファインチューニング用のデータセットを作成するには，隠語を含むメッセージと含まないメッセージの割合が重要である．そこで仮の LLM モデルを 1 つ選択し，そのモデルを用いてすべてのメッセージから隠語を抽出し，得られた情報をデータセット作成に利用する．6 件のメッセージを用いて gpt-4o-2024-05-13, gpt-4o-mini-2024-07-18, command-r-plus:104b の 3 種類の LLM モデルを評価し，事前調査に用いる LLM モデルを決定する．これらのモデルを選択した理由は，OpenAI のモデルが一般的に高性能とされていることと，command-r-plus:104b が我々の過去の調査からキーワード抽出に適していると判断されたためである．隠語抽出に使用するプロンプトを図 1 に示す．このプロンプトは Fillies らの研究 [7] で使用されているプロンプトをもとに，サイバー犯罪向けに修正したものである．プロンプトの ${message} には 1 件のメッセージが挿入される．LLM の出力の一貫性を確保するため，temperature を 0，seed を 0，top_p を 0 に設定する．

> You are a helpful assistant who is an expert in detecting online cybercrime.
>
> Identify cybercrime related slang in the following sentences. If it has been found, output the slang only. If nothing has been found, answer ’[No slang]’.
>
> ${message}

図 1 スラング抽出のプロンプト

4.1.2 データセットの作成

ファインチューニング用の学習データセットと LLM モデル比較用の評価データセットを作成した．各データセットは 500 件のメッセージから構成され，隠語の有無が均等になるよう，事前実験で隠語が抽出されたメッセージ 250 件と抽出されなかったメッセージ 250 件から構成される．これらのメッセージは，各 Discord サーバからランダムに選択される．複数のチャンネルを持つサーバの場合，各チャンネルから均等にメッセージを抽出する．例えば，3

件のサーバから 10 件のメッセージを選択する場合，各サーバから 4 件，3 件，3 件をランダムに抽出する．また，2 件のチャンネルを持つサーバから 3 件のメッセージを選択する場合，それぞれのチャンネルから 2 件と 1 件をランダムに選択する．

アノテーションのプロセスでは，2 つのデータセットをマージし，シャッフルして単一のデータセットを作成する．2 名の研究者が独立してこのデータセットからメッセージ内の隠語を抽出する．例えば，“wts credit cards 700 +refs” というメッセージからは [“wts”, “+refs”] とアノテーションを行う．アノテーション作業では，事前に 2.1 節の隠語の定義を共有し，5 時間以内に完了すること，バイアスを避けるため LLM の使用は禁止し，Google 検索のみを許可するという制限を設ける．

さらに，同一のデータセットに対して gpt-4o-mini-2024-07-18 と command-r-plus:104b を用いて隠語抽出を実施する．使用したプロンプトを図 1 に示す．LLM の出力の一貫性を確保するため，temperature，seed，top_p をすべて 0 に設定する．LLM を使用した理由は，研究者が文章全体から隠語を抽出するよりも，個別の単語が隠語かどうかを判断する方が得意であるためである．

最終的に，2 名の研究者によるアノテーションの結果を比較し，2 つの LLM モデルによるアノテーションを参考にしながら，適切な隠語抽出について人間のアノテーター間で議論を行う．意見の相違がある場合は，定義を確認して，場合によっては定義を拡張する方法で決定する．また，より多角的な判断を行うため，隠語の判定過程では LLM との対話も許可する．

4.1.3 ファインチューニングの実行

作成したファインチューニング用データセットを用いて，gpt-4o-mini-2024-07-18 および gpt-3.5-turbo-0125 のモデルに対してファインチューニングを実施する．これらのモデルを選択した主な理由は，OpenAI が提供する基盤を活用することで，効率的かつ適切なファインチューニングが可能となるためである．ファインチューニングのプロセスにおいては，モデルの特性を最大限に活かすため，OpenAI が推奨するデフォルトのハイパーパラメータを採用する．

4.1.4 隠語に適した LLM モデルの特定

評価用データセットを用いて，表 3 に示した 31 種類の LLM モデルの性能を比較評価する．評価指標として，精度，再現率，F1 スコアを採用する．評価プロセスでは，各メッセージに対して，LLM モデルが予測した隠語のリストと，評価データセットの正解のリストを比較する．精度は，モデルが正しく予測した隠語の数を，モデルが予測した隠語の総数で除して算出する．ただし，モデルの予測と正解がともに 0 件の場合，精度は 1 とする．再現率は，モデルが正しく予測した隠語の数を，正解の隠語の総数で除して計算する．正解の隠語が 0 件の場合，再現率は 1 とする．F1 スコアは，精度 (P) と再現率 (R) の調和平均として，$F1 = 2 * (P * R)/(P + R)$ の式で算出する．各 LLM モデルの最終的な評価値は，評価用データセットに含まれる全メッセージに対する精度，再現率，F1 スコアの平均値として算出する．

4.2 RQ2 に対する調査手法

4.2.1 Discord 内の隠語の特定

RQ1 でサイバー犯罪関連の隠語抽出に最適であると判断された LLM モデルを用いて，Discord データセットの全メッセージから隠語を抽出する．隠語抽出に使用したプロンプトを図 1 に示す．LLM の出力の一貫性を確保するため，temperature を 0，seed を 0，top_p を 0 に設定する．

4.2.2 隠語のカテゴリの判定

特定された隠語とそのコンテキストメッセージに基づき，各隠語のカテゴリを判定する．カテゴリは，サイバー犯罪に関連するカテゴリとして「ダークマーケット」「ゲーム上の不正行為」「詐欺」「ハッキング」「児童ポルノ」「サイバーいじめ」「知的財産権の侵害」「違法薬物」の 8 件，サイバー犯罪に誘導するカテゴリ「通常のポルノ」「賭博」「デジタル資産」の 3 件，「サイバー犯罪に関連しない隠語」，「隠語ではない単語」の計 13 種類から選択する．カテゴリ判定のプロセスでは，1 名の研究者が手動でカテゴリを判定する．より多角的な判断を行うため，隠語の理解とカテゴリの特定作業では LLM との対話を許可する．

5. 結果と考察

5.1 RQ1: LLM による隠語抽出の可能性とモデル比較

5.1.1 事前実験の結果

6 件のメッセージを使って gpt-4o-2024-05-13，gpt-4o-mini-2024-07-18，command-r-plus:104b の 3 つのモデルを評価した結果を表 1 に示す．gpt-4o-mini-2024-07-18 が最も高い再現率 (0.707) を示した．この結果に基づき，本実験では gpt-4o-mini-2024-07-18 を使用することとした．gpt-4o-mini-2024-07-18 を用いて 98,808 件のメッセージに対して隠語抽出を行った結果，19,722 件のメッセージから隠語が抽出され，残りの 79,086 件からは隠語が抽出されなかった．

表 1 事前実験における LLM モデルの性能比較

モデル	精度	再現率	F1 スコア
gpt-4o-mini-2024-07-18	**0.691**	**0.729**	**0.707**
command-r-plus:104b	0.644	0.619	0.630
gpt-4o-2024-05-13	0.635	0.626	0.629

5.1.2 データセット作成結果

データセットの作成過程では，2 人の研究者が独立してアノテーションを行い，その結果を比較し，最終的な結果を決定した．アノテーション作業の信頼性を評価するた

め，研究者間の一致度をコーエンのカッパ係数を用いて測定した．コーヘンのカッパ係数は，隠語ごとにメッセージから抽出したかどうかで算出し，すべての隠語の平均から算出した．その結果，カッパ係数は 0.2731 となった．

この結果は，アノテーターがアノテーションした結果で細かい違いがあり，アノテーター間での認識の一致が難しいことを意味している．例えば，“dumps” や “wtb”, “cc” などのメッセージによく現れる隠語のカッパ係数は 1.0 であり，アノテーションが一致していた．だが，“erp” (Erotic Role-Play) などの背景知識が必要な隠語や，詐欺で誘引に使われる “airdrops” のようにサイバー犯罪に関連するか判断が分かれる隠語のカッパ係数は 0 に近い値となった．これらの不一致点については，両研究者間で詳細な議論を行い，最終的な判断基準を設定した．

また，2 種類の LLM モデルによるアノテーション結果は，誤検知は含むものの多くの単語を抽出することができた．この結果から，最適な LLM モデルを評価せずに LLM モデルでアノテーションするのは難しいが，出力を参考にして人間が判断することに適していると考えられる．

5.1.3 ファインチューニングの実行結果

ファインチューニングの実行結果を表 2 に示す．これらの結果から，両モデルとも効果的にファインチューニングが行われたことが示唆される．

表 2 ファインチューニングの実行結果

項目	gpt-4o-mini-2024-07-18	gpt-3.5-turbo-0125
学習トークン数	225,483	233,361
エポック数	3	3
バッチサイズ	1	1
学習率	1.8	2.0
シード	606630939	1707474915
トレーニングロス	0.0014	0.0000

5.1.4 隠語に適した LLM モデルの特定

評価データセットを用いて 31 種類の LLM モデルを評価した結果を表 3 に示す．最も高い F1 スコアだったのが，ファインチューニングを施した gpt-4o-mini-2024-07-18 モデルで，精度 0.89，再現率 0.88，F1 スコア 0.85 を達成した．このモデルは，次点のファインチューニングを施した gpt-3.5-turbo-0125 と比較して，F1 スコアで 0.06 ポイント上回った．

上位 2 モデルがファインチューニング済みであることから，隠語抽出タスクにおけるファインチューニングの重要性が示唆される．一方で，mistral-large:123b や llama2-uncensored:70b などの一部オープンソースモデルも高い性能を示した．mistral-large:123b のハルシネーション抑制や，llama2-uncensored:70b の無修正データでのファインチューニングという特性が，サイバー犯罪関連の隠語検出に有効に働いたと考えられる．

また，興味深いことに，モデルのパラメータ数と性能には明確な相関が見られなかった．例えば，270 億パラメータの gemma2:27b は，20 億パラメータの gemma2:2b と比較して，0.20 ポイントも低い F1 スコアを示した．

表 3 隠語抽出における LLM モデルの性能比較

モデル名	パラメータ数	精度	再現率	F1 スコア
gpt-4o-mini-2024-07-18 (fine tuned)	不明	**0.89**	**0.88**	**0.85**
gpt-3.5-turbo-0125 (fine tuned)	≈ 1750 億	0.84	0.85	0.79
mistral-large:123b	1230 億	0.77	0.80	0.76
llama2-uncensored:70b	690 億	0.74	0.75	0.74
mistral-openorca:7b	72 億	0.73	0.75	0.73
wizardlm2:7b	72 億	0.71	0.75	0.71
llama2-uncensored:7b	67 億	0.71	0.75	0.71
mixtral:8x7b	467 億	0.71	0.77	0.70
dolphin-mixtral:8x22b	1410 億	0.70	0.77	0.70
qwen2:1.5b	15 億	0.70	0.75	0.70
qwen2:72b	727 億	0.68	0.81	0.68
gemma2:2b	26 億	0.68	0.77	0.67
gpt-4o-2024-05-13	不明	0.68	0.80	0.67
mistral:7b	72 億	0.67	0.76	0.66
gpt-3.5-turbo-0125	≈ 1750 億	0.66	0.78	0.65
phi3:3.8b	38 億	0.65	0.75	0.65
command-r-plus:104b	1040 億	0.65	0.81	0.64
dolphin-llama3:70b	706 億	0.64	0.77	0.64
wizardlm2:8x22b	1410 億	0.64	0.77	0.63
dolphin-mistral:7b	72 億	0.63	0.76	0.62
llama3.1:8b	80 億	0.62	0.77	0.62
qwen2:0.5b	4 億	0.61	0.76	0.61
gpt-4-turbo-2024-04-09	不明	0.62	0.81	0.61
llama3.1:70b	706 億	0.61	0.79	0.59
phi3:14b	140 億	0.60	0.79	0.59
gpt-4o-mini-2024-07-18	不明	0.56	0.87	0.56
dolphin-llama3:8b	80 億	0.53	0.80	0.53
dolphin-mixtral:8x7b	467 億	0.49	0.84	0.49
gemma2:9b	92 億	0.50	0.84	0.49
qwen2:7b	76 億	0.48	0.77	0.47
gemma2:27b	272 億	0.47	0.87	0.47

これらの結果は，サイバー犯罪関連の隠語抽出タスクにおいて，適切なファインチューニングが性能向上に大きく寄与することを表している．同時に，一部のオープンソースモデルも競争力のある性能を示しており，プライバシーの問題や，計算リソースやコストの制約がある場合の有力な選択肢となり得る．特に mistral-large:123b や llama2-uncensored:70b のような大規模オープンソースの LLM モデルは，ファインチューニングなしでも高い性能を示しており，柔軟性とコスト効率の面で優位性がある．

5.1.5 RQ1 の回答

RQ1「LLM はサイバー犯罪関連の隠語抽出に有効か．最

適な LLM モデルはどれか」の回答は以下の通りである：
(1) **サイバー犯罪関連の隠語抽出における LLM の有効性**: ファインチューニングを施した gpt-4o-mini-2024-07-18 モデルが精度 0.89，再現率 0.88，F1 スコア 0.85 という高い性能を示したことから，LLM によるサイバー犯罪関連の隠語抽出が十分に可能であることが実証された．
(2) **最適な LLM モデル**: 評価結果から，ファインチューニングを施した gpt-4o-mini-2024-07-18 が最も高い性能を示した．また，mistral-large:123b や llama2-uncensored:70b などの一部のオープンソースモデルも高い性能を示し，サイバー犯罪関連の隠語抽出に適していることが明らかになった．

本研究ではオープンソースモデルに対するファインチューニングは実施していないため，これらのモデルにファインチューニングを適用することで，さらなる性能向上の可能性がある．今後の研究では，オープンソースモデルのファインチューニングも含めた，より包括的な評価が必要であると考えられる．

5.2 RQ2: Discord におけるサイバー犯罪関連の隠語

5.2.1 抽出された隠語の概要

RQ1 で隠語抽出に最適であると判断されたファインチューニング済みの gpt-4o-mini-2024-07-18 モデルを使用し，サイバー犯罪に関係する 98,808 件のメッセージを分析した．その結果，30,537 件の隠語を抽出され，重複を除外すると，810 件の固有の隠語が特定された．

最も頻出した隠語は “wts” で，6,630 件のメッセージに出現した．これは “want to sell” の略語であり，ある商品を販売したい意図を表している．次に多かったのは “+refs” で，3,216 件のメッセージに出現した．この隠語は，その販売が多くの人に信用されていることを表している．3 番目に多かったのは 18 歳未満禁止のマークの絵文字で，2,130 件のメッセージに出現した．この絵文字は，コミュニティやコンテンツが 18 歳未満の利用を禁止していることを表している．

5.2.2 隠語のカテゴリ分析

抽出された隠語を，事前に定義した 13 種類のカテゴリに分類した．分類結果を表 4 に示す．隠語の数の合計が 825 件になっているのは，同じ隠語でも別のカテゴリに属すものがあったためである．

分析の結果，通常のポルノに関連する隠語が最も多く，次いでダークマーケットとゲームチートに関連する隠語が多いことが判明した．この結果は，Discord プラットフォーム上でこれらの犯罪活動が活発に行われている可能性を表している．一方，賭博やサイバーいじめ，知的財産権の侵害，ドラッグに関連する隠語は比較的少数であった．

また，サイバー犯罪に関係のない隠語や，隠語ではない単語も多く抽出された．例えば，サイバー犯罪に関連のない隠語では，13+や 15+のような抽出対象の 18+に似た形式の隠語や，beta (ベータ版) や dl (download)，vc (voice chat) などの一般的な隠語が抽出されている．また，隠語ではない単語では，cash app のようなサービス名，proxy のような技術用語，tool のような定義 (8) (2.1 節を参照) の「辞書に書かれた意味以上の意味を持つ」に当てはまるか解釈が割れるような単語が抽出された．本タスクでは隠語でありサイバー犯罪に関連する単語が抽出対象だが，LLM がそのどちらかのみに当てはまる単語も抽出してしまっていることが原因であると考えられる．

表 4 カテゴリごとの隠語の数

カテゴリ	数	カテゴリ	数
通常のポルノ	203	児童ポルノ	23
隠語ではない単語	170	デジタル資産	18
ダークマーケット	139	賭博	9
サイバー犯罪に関連しない隠語	116	サイバーいじめ	4
ゲーム上の不正行為	76	知的財産権の侵害	3
詐欺	38	違法薬物	2
ハッキング	24		

5.2.3 カテゴリごとの隠語の具体例

表 4 に示した各カテゴリの隠語について説明する．

通常のポルノ: rule34 や l.r34, n.r34 のように，34 が含まれる隠語が頻出した．Rule34 とは，海外のネット掲示板でまとめられたインターネットのルールの 34 番目にあたり，この世のすべてのものにはポルノ画像が存在することを意味する．他には，16+や 18+，+20 のような年齢関連の隠語や，cosplays, futa, neko のように日本語がそのまま使われている隠語が複数確認された．また，唐辛子や茄子，桃，さくらんぼ，水滴，コップに入ったミルク，18 歳未満禁止のマークといった絵文字が使われていた．

ダークマーケット: 主にクレジットカード，アカウント，個人情報，犯罪手法に関する隠語に分類される．クレジットカードを表す隠語には cards, cc (Credit Card), ccs (Credit Cards), dump, dumps, dumpz (dumps) などがある．複製されたクレジットカードは clone cards, clones と呼ばれ，他のサービスに登録可能な脆弱なクレジットカードは linkable と呼ばれる．盗難クレジットカード情報の不正利用は carding, c4rd1ng (carding), cardable, carded, facarding (full access carding) などと表現され，使用方法とともに販売されることがある．

アカウントは acc (account) という隠語で表される．長期間存在しているアカウントは aged account, oge (Original Gangster) と呼ばれる．アカウントの識別子は cid (Customer ID) と呼ばれる．PayPal アカウントは pp (PayPal), pplog (PayPal log) などと表現される．アカウントの属性を示す用語として，最近確認または取得されたことを示す fresh，すべての情報にアクセス可能であることを示す fa (Full Access)，一部の情報にのみアクセス可能であること

を示す nfa (Not Full Access) などが使用される．

個人情報は log と呼ばれる．完全な個人情報のセットは fullz と呼ばれる．販売されている個人情報には様々な種類があり，銀行口座情報を指す bank logs，ユーザー名とパスワードの組み合わせリストを指す combolist，combos，様々な個人データを表す data infos などがある．これらの情報は多くの場合，データ漏洩によって取得されたものであるため，leak と表現される．また，leak のもじりとして le4k という隠語も確認された．

犯罪手法は method や sauce と呼ばれる．金銭的利益や犯罪のための手法が販売されており，商品を返却せずに返金を受ける r3fund (refund)，サービスから不正に情報を取得する pulling method，ソーシャルエンジニアリングの sc method (Social Engineering Method) などがある．

ダークマーケットでは，販売時に wts (want to sell)，wtt (want to trade)，購入時に wtb (want to buy)，lf (looking for) という隠語を使用することが多い．販売物の信頼性を示すため，+refs (references) や +reps (reputation) など，多くのユーザーからの評判があることを表す表現が用いられる．連絡は主に Discord や Telegram のダイレクトメッセージで行われ，連絡を求める hit me up，hmu などの隠語と共に，dm (Direct Message)，pm (Private Message) などがメッセージに記載される．

販売に関する隠語には，中間者を介した取引を示す mm (Middle Man)，mm bot，mm service，mming や，購入者に先払いを求める ngf (Not Going First)，即時の金銭的利益を示す tap，tap in，有利な販売条件を示す obo (or best offer)，promo (promotion)，購入前のテスト利用を示す test run などがある．また，購入時には Cash App や PayPal などの決済プラットフォームの ID が必要となるため，tag といった隠語もメッセージに記載される．

ゲーム上の不正行為: Discord にはゲームコミュニティが多く存在し，ゲームチートに関する隠語も複数確認された．チート自体を表す隠語には cheat，exploit，glitch，macro，mod，rage，script，unlocker などがある．特に cheese は cheat という隠語をさらに秘匿化した単語である．チートの開発者は dev (developer)，チートを使用するユーザは exploiter や modder，チートを実行するためのツールは emulator，executor，exploit と呼ばれる．チート手法を表す隠語には，chams，wall hack (壁や障害物を透視し，隠れた敵や物体を視認できる)，esp (Extra-Sensory Perception) (画面上に敵の位置や重要な情報を表示する)，godmode (ダメージを受けない，または大幅に軽減する)，inf (ゲーム内のリソースや能力を無限に使用可能にする)，no cd (スキルやアイテムの使用制限時間を無効化する) などがある．

さらに，チート検知の対策にも隠語が使用されている．検知されないチートは ua (undetectable access) と呼ばれる．ゲーム側のチート対策は ac (Anti-Cheat) と呼ばれ，その対策には antiban といった隠語が使用されている．また，ゲーム側でチートを行ったユーザの検知に使用されるハードウェア ID は hwid (Hardware ID) と呼ばれ，その ID を偽装するツールは spoofer と呼ばれる．

詐欺: 詐欺関連の隠語は，被害者を誘引するための隠語と詐欺手法自体を表す隠語に大別される．被害者誘引の隠語には，NFT や仮想通貨の無料配布を示す airdrop や drop，gws (giveaways)，perks，個人情報入力時の項目を指す deities，ビジネスの合法性を主張する legit bizz (legit Business)，株式等のトレーディング情報を指す signal，高収益を示唆する to da moon などがある．

詐欺手法を表す隠語には，オンライン上での性的な文脈での金銭要求を指す ewhoring，投資家から資金を集めて突然失踪をする exit，少額投資で高リターンを謳う flip，誤情報による仮想通貨価格の人為的操作を指す pump and dump，分散型金融 (DeFi) における取引前後の価格差を利用する sandwich などがある．また，詐欺実行者を指す brouteur，オンライン詐欺で金銭を得る者を指す hustlers，個人の利益のために宣伝を行う shill，決済プラットフォームでの個人間送金を示す ff (Friend Family) なども，詐欺の文脈で見られた．

ハッキング: ハッキング自体は hack，crack，dork などの単語で表される．cr4ck のように文字を数字で置換した隠語や haxs のように文字を置き換えた隠語も確認された．ハッキングの手法では，システム内の外部からの侵入口を指す backdoor，ユーザの情報を盗み出すツールである grabber，仮想通貨のウォレットから資産を盗み出す drainer，計画的に資産を盗み出す行動である heist などがある．既存の単語の文字を置き換えた隠語として，Social Engineering を指す s0cial 3ngine3ring という単語が確認された．また，ハッキングフォーラムを指す hf (Hacking Forum)，ハッキングフォーラムとして使用されることがある tg (Telegram)，Pastebin のようなプラットフォームで共有されるデータを指す pastes などの隠語も存在する．

児童ポルノ: 代表的な隠語だと，Child Pornography の頭文字から cp という隠語がある．同様の原理で，Cheese Pizza も児童ポルノを表す隠語として使用される．さらに，Cheese を省略した :pizza: という絵文字も児童ポルノを示す隠語として確認された．他には teen content という単語も多く確認された．これは，そのままの意味だと 10 代のコンテンツであるが，実際は 10 代のポルノコンテンツを指していて，児童ポルノの恐れがある．

デジタル資産: Discord には仮想通貨やコイン，NFT などのデジタル資産のコミュニティが多数存在する．それらのデジタル資産は da と呼ばれる．詐欺に関連しそうな隠語では，取引などで無謀な行動をとる個人を degen (degenerate)，簡単なタスクやキャプチャを完了することで少量の暗号通貨を無料で配布する faucent，市場のおそ

れや不確実性を表す fud (Fear, Uncertainty, and Doubt)，一般の人々がアクセスする前に暗号通貨などを最良の価格で購入する sniper などがみられた．

賭博: 複数の賭け方を組み合わせて高額の賭けを行う accas (accumulators)，parlay，両チームが得点することに賭ける btts (Both Teams To Score)，試合の最初のイニングで得点が入らないことに賭ける NRFI (No Runs First Inning) などがあった．賭けの議論で使用される隠語には，ベッティングのアドバイザーを指す cappers，期待値を表す EV (Expected Value) などがあった．

サイバーいじめ: オンライン上での攻撃的行為に関連する隠語として，他者への嫌がらせを指す bming (bad manner)，自殺を促す kys (Kill Yourself)，インド系の人々を蔑視する pajeets，虚偽の緊急通報により SWAT チームを他者の家に派遣させる swatting を確認した．

知的財産権の侵害: Discord 上では有料ゲームや有料ソフトウェア，有料の学習教材が無料公開されていることがあり，leak という単語が通常の情報漏洩という意味に加えて，有料のものを無料公開するという意味で使われているケースが確認された．

違法薬物: 違法薬物に関する隠語は 2 件確認された．1 件目は 420 であり，4 月 20 日に大麻に関する文化の日であることから，大麻を示す．2 件目は麻薬を意味する “kusa” で，日本語の「草」に由来すると考えられる．

5.2.4 RQ2 の回答

RQ2「Discord におけるサイバー犯罪関連の隠語にはどのようなものがあるか」に対する回答は以下の通りである．

Discord 上で 810 件の固有の隠語が特定され，これらは主に通常のポルノ，ダークマーケット，ゲーム上の不正行為の 3 カテゴリに集中していた．最も頻繁に使用される隠語には，wts (want to sell)，+refs (positive references)，rule34 などがあり，これらは主に違法な商品やサービスの取引，ポルノに関連している．隠語の形態は多様で，略語 (cp: Child Pornography)，既存の単語の目的外使用 (cheese pizza: Child Pornography)，絵文字 (:pizza:) など，様々な手法が用いられている．

6. まとめと今後の課題

本研究では， Discord 上のサイバー犯罪関連の隠語を LLM を用いて抽出する手法を提案し，その有効性を実証した．まず, 31 種類の LLM モデルの隠語抽出性能を比較評価し，ファインチューニングを行った gpt-4o-mini-2024-07-18 が最も高い性能 (F1 スコア 0.85) であることを明らかにした．さらに，このモデルを用いて Discord のサイバー犯罪関連のメッセージから隠語を抽出した結果，810 件の固有の隠語を特定し，それらが主に通常のポルノ，ダークマーケット，ゲーム上の不正行為の 3 カテゴリに集中していることを示した．

今後の課題としては，オープンソースの LLM モデルでファインチューニングを行い，その性能を検証することが重要である．本研究では商用モデルのみをファインチューニングしたが，オープンソースモデルでも同様の性能向上が見込めるか調査する必要がある．また，得られた隠語のカテゴリや説明を取得する仕組みが必要である．本研究では，LLM を補助として利用しながら研究者がラベル付けや調査を行ったが，より効率的にサイバー犯罪対策を行うためには，これらの作業の自動化が求められる．

謝辞 この成果の一部は, 国立研究開発法人新エネルギー・産業技術総合開発機構 (ＮＥＤＯ) の委託業務 (JPNP22007) の結果得られたものです．

参考文献

[1] Cognyte. The Rise of Cybercrime on Telegram and Discord and the Need for Continuous Monitoring | CloudSEK. https://cloudsek.com/blog/the-rise-of-cybercrime-on-telegram-and-discord-and-the-need-for-continuous-monitoring.

[2] Heather Adkins. Review of the Attacks Associated with LAPSUS$ and Related Threat Groups. https://www.cisa.gov/resources-tools/resources/review-attacks-associated-lapsus-and-related-threat-groups-report.

[3] How Discord is Abused for Cybercrime. https://intel471.com/blog/how-discord-is-abused-for-cybercrime.

[4] AFP releases glossary of terms used by some sex predators to groom children. https://www.afp.gov.au/news-centre/media-release/afp-releases-glossary-terms-used-some-sex-predators-groom-children, February 2022. Accessed: 2024-8-22.

[5] Dea Dea Intelligence. Drug slang code words. https://www.dea.gov/sites/default/files/2018-07/DIR-020-17%20Drug%20Slang%20Code%20Words.pdf.

[6] Hanxiang Xu, Shenao Wang, Ningke Li, Kailong Wang, Yanjie Zhao, Kai Chen, Ting Yu, Yang Liu, and Haoyu Wang. Large language models for cyber security: A systematic literature review. arXiv [cs.CR], May 2024.

[7] Jan Fillies and Adrian Paschke. Simple LLM based approach to counter algospeak. In Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024), pp. 136–145, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

[8] OpenAI. Moeels - openai api. https://platform.openai.com/docs/models/gpt-4o-mini.

[9] Jack Hughes, Andrew Caines, and Alice Hutchings. Argot as a trust signal: Slang, jargon & reputation on a large cybercrime forum, July 2023.

[10] Jessie Smith. Discord Revenue and Usage Statistics 2024. https://helplama.com/discord-statistics/, April 2023.

[11] Kan Yuan, Haoran Lu, Xiaojing Liao, and Xiaofeng Wang. Reading thieves’ cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces. USENIX Secur Symp, pp. 1027–1041, 2018.

[12] 林容央, 塩沢健, 穐山空道, 桂井麻里衣. ダークウェブ上のハイパーリンクを用いた隠語分析の検討. 研究報告セキュリティ心理学とトラスト（SPT）, Vol. 2024-SPT-56, No. 37, pp. 1–7, July 2024.