

Discord 上における解釈困難なメッセージに対する 大規模言語モデルによる解釈の可能性と課題

大勝友洋 Yin Minn Pa Pa 吉岡克成 森辰則

横浜国立大学大学院

mori-tomohiro-bm@ynu.jp {yinminn-papa-jp, yoshioka, tmori}@ynu.ac.jp

概要

近年, SNS や Discord 等のオンラインチャットにおけるサイバー犯罪が増加している. 犯罪者は隠語や特殊な表現を用い, 検出回避や情報の秘匿を行うため, 犯罪の特定や早期発見が困難になっている. 本研究では, Discord から収集したテキストチャットに基づき, 解釈困難なメッセージの実態を調査する. それらのメッセージを解釈困難性の要因ごとに分類した上で, 人間と大規模言語モデル (LLM) による解釈実験を行う. これにより, LLM が文脈情報の活用によってどの程度メッセージを解釈できるのか, その可能性と課題を検証する.

1 はじめに

近年, SNS やオンラインチャットを利用したサイバー犯罪が増加している [1]. 警察庁サイバー警察局によると, 令和6年のSNS型投資・ロマンス詐欺の被害額は約1,268億円(前年比178.6%増)と急増しており [2], 暗号資産を悪用した詐欺や不正取引, ハッキングツールの売買など, インターネット空間における犯罪行為は年々多様化・巧妙化している.

本研究の対象である Discord も, コミュニティ作成の容易さと高い匿名性から, 犯罪者間の情報交換に利用されている [1, 3]. 特に悪質なコミュニティでは, 専門用語やスラング, 隠語, 検出回避のための特殊表記が多用され, 当該コミュニティの知識を持たない第三者による解釈を困難にしている.

例えば, “I rug pulled yesterday and made bank.” というメッセージは一見意味不明だが, 暗号資産の文脈では「トークンの価格操作による詐欺 (rug pull) で大儲けした」と解釈できる. このように, 犯罪に関連するメッセージは表層的な文理解のみでは捉えられず, 特定分野の知識や文脈理解を必要とする.

これまでにも, 文章の解釈に対する影響について

特定の語に対する認識の有無やコミュニティへの依存性が指摘されてきた [4, 5]. また, スラングや新語について, 辞書の定義と文脈情報を組み合わせた解釈手法や, 大規模言語モデルを用いて動的に意味を推定する試みが報告されている [6, 7]. さらに, 犯罪行為を外部から察知されにくくするために意図的に作られた隠語に着目した調査も行われており, Discord 上の犯罪コミュニティから多数の隠語が抽出されている [8].

一方で, 実際の犯罪コミュニティにおけるメッセージでは専門用語や隠語, 文法誤り等が混在する場合や, 既存研究では言及されていない検出回避を目的とした表現手法が用いられる場合がある. そのため, 既存研究の枠組みのみでは実際のメッセージの解釈困難性を十分に捉えられていない.

そこで本研究では, 解釈困難性の要因を「専門用語・固有名詞」, 「スラング」, 「隠語」, 「単語の綴り・文法誤り」, 「システムを欺く表現」の5種類に分類する. そのうえで Discord 上の悪質なコミュニティから収集したテキストチャットを対象とし, 情報工学を専攻する日本人大学院生がこれらのメッセージをどの程度解釈できるのかを検証しつつ, 各要因が解釈の成否に与える影響を分析する. 加えて, LLM を用いた解釈実験を行うことで, 人間との比較を通じて大規模言語モデルによる解釈の可能性と課題を明らかにし, サイバー犯罪の特定や早期発見への貢献を目指す.

2 解釈困難性の要因分類

本研究では, Discord 上の悪質なコミュニティにおいて観測されるような, コミュニティに属さない人間が意図や内容を正しく理解することが難しい表現を「解釈困難なメッセージ」と定義する. 分析を容易にするため, 解釈困難性の要因を以下の5種類に分類し, 各分類に該当する具体例を表1に示す.

2.1 専門用語・固有名詞

特定のコミュニティで定着し、そのコミュニティ内では一貫した意味で用いられる語を専門用語・固有名詞として分類する。これらの語は特定のコミュニティで限定的に使用される場合、一般的な語彙知識のみでは意味の推定が難しくなる [5]。表 1 の“Woolworths”はオーストラリアの小売業者，“BNB”は暗号通貨の名称である。

2.2 スラング

口語的・流行語的で、意味が変化しやすい語をスラングとして分類する。スラングには、辞書に未掲載である語や時間とともに意味が変化する語が多く、文脈に強く依存する [7, 9]。そのため、表層的な翻訳や単語置換のみではメッセージの意図を正しく解釈できない場合がある。特定の文脈で表 1 の“fr”は“for real”，“catfish”は 2010 年以降から「なりすまし」という意味で用いられている。

2.3 隠語

専門用語やスラングの中でも、犯罪行為等の内容を秘匿する目的で作られた語を隠語として分類する。一般的な語に別の意味を付与する 경우가多く、その隠語を知らない人間に無害な発言として解釈させる意図がある。表 1 の“Cheeze Pizza”は「児童ポルノ」，“dupe”は「不正コピーアカウント」という意味で用いられる。

2.4 単語の綴り・文法誤り

オンラインチャットでは手軽さや即時性が重視されるため、単語の綴り誤りや文法の誤りが含まれるメッセージが多い [10]。このような言語的ノイズは本来の語義や文構造の理解を妨げ、解釈困難性を高める要因となる。表 1 の“anybdoy”は“anybody”の綴り誤り，“where from”は“where are you from?”の文法誤りの例である。

2.5 システムを欺く表現

自動検出を回避する目的で文字の分割や記号・絵文字の挿入等を行う手法をシステムを欺く表現として分類する。このような表現は、特にルールベースの自動検出システムに対して有効だと考えられる。表 1 は隠語の“Cheeze Pizza”に丸括弧を挿入した例と，“incest porn”に空白文字を挿入した例である。

表 1 解釈困難性の要因とその例

要因	例
専門用語・固有名詞	Woolworths, BNB
スラング	fr, catfish
隠語	Cheeze Pizza, dupe
単語の綴り・文法誤り	anybdoy, where from
システムを欺く表現	(C)heese P)izza), in ces t por n

3 人間による解釈の調査

本章では解釈困難なメッセージを選定し、3人の日本人大学院生による解釈実験を行う。

3.1 使用したデータ

川口ら [11] の情報収集システムを用いて 2023 年 10 月に収集された Discord のメッセージを対象とした。対象となるサーバは“Drug”，“Malicious hacking”，“Scams”など計 196 個のキーワードを使って検索され、取得された 8,476 件のサーバから無作為に 100 件のサーバに参加し、そのうち 55 件のサーバから 1,280,548 件のメッセージを取得している。

本研究では、ここからサイバー犯罪に関連する 23 チャンネルを人手で選定し、それらのチャンネルの中からスラングや隠語等を含むことにより解釈が困難なメッセージを 100 件抽出した。プリンストン大学が公開している英単語の古典的な概念辞書である“WordNet 3.1” [12] と、ウィキメディア財団が運営する共同作業型のオンライン辞書である“Wiktionary” [13] を参照したところ，“WordNet 3.1”では 92 件，“Wiktionary”では 73 件のメッセージにおいて意味が登録されていない言葉が含まれていた。

3.2 実験設定

抽出した 100 件のメッセージに対し、自然言語処理を専門とする研究室所属の 3 人の日本人大学院生が以下の条件で順番に実験を行った。

- メッセージを単体で読み、意味を解釈できる場合にそれを記述する
- 解釈困難なメッセージを前後の文脈を含めて読み、意味を解釈できる場合にそれを記述する
- 外部知識 (Web 検索や LLM に対する質問) を利用し、メッセージの意味を記述する
- すべてのメッセージについて 3 人で相談を行い、一つの解釈を決定する

補足として、条件 B における前後の文脈には解釈

表 2 各条件の正解数と文埋め込み類似度の平均 (N=100)

被験者	正解数			文埋め込み類似度
	条件 A	条件 B	条件 C	条件 C
被験者 1	3	15	70	0.832
被験者 2	3	4	63	0.808
被験者 3	1	11	65	0.810
平均	2.33	10.0	66.0	0.817

対象のメッセージも含めて発言時間と発言者の情報を含んでいる。条件 C までの実験はそれぞれ個別で行い、条件 C での実験後に条件 B の結果を書き直すというような、遡る作業は禁止している。さらに条件 D において 3 人で相談を行って決定した解釈に対して情報セキュリティの専門家である第二著者によるレビューを行い、正解データを構築した。構築したデータの一部については付録 A に記載している。

条件 A~C で被験者が記述した解釈について、意味的な類似度を確かめるため日本語用 Sentence-BERT[14] によって文埋め込みベクトルに基づく類似度を算出し、正解データとの類似度が 0.75 以上であったものを正解として判定した。

3.3 実験結果

用意した 100 件のメッセージに対する人手による解釈実験の結果を表 2 に示す。メッセージ単体のみを読んだ条件 A では、被験者の平均正解数は 2.33 件と極めて低く、前提知識なしでの解釈は困難であった。次に、前後の文脈を含めて読んだ条件 B では平均正解数は 10.0 件まで増加したが、依然として正確に読めているメッセージは全体の一部にとどまっており、文脈情報だけでは十分な解釈に至らない場合が多いことが分かる。

一方、外部知識を参照した条件 C では平均正解数は 66.0 件と大きく向上した。また、条件 C における解釈結果と正解データとの文埋め込み類似度の平均は 0.817 であり、外部知識を利用することで解釈内容が正解データに近づく傾向が確認された。

これらの結果から、英語の解釈困難なメッセージに対しては、前後の文脈情報のみでは不十分であり、Web 検索や LLM などの外部知識の利用が解釈精度の向上に大きく寄与することが示された。

4 LLM による解釈の調査

現行の LLM を用いて、解釈困難なメッセージをどの程度正確に解釈できるかを検証する。隠語に

表 3 LLM の出力と正解データの類似度 (N=100)

モデル (条件)	文字列類似度	文埋め込み類似度
GPT-OSS-120B(i)	0.443	0.737
GPT-OSS-120B(ii)	0.452	0.763
GPT-OSS-20B(i)	0.426	0.707
GPT-OSS-20B(ii)	0.452	0.722

対する LLM の性能向上を目的とした研究では、プロンプト設計やヒント付与の工夫が提案されている [15]。しかし、本研究の目的はそのような工夫によって性能を引き上げるのではなく、現行の LLM が解釈困難なメッセージに対してどの程度の解釈能力を持つのかを把握することである。そのため、3.2 節で構築した正解データと LLM の出力を比較し、単純な解釈能力の検証を行う。

4.1 実験設定

使用する LLM は GPT-OSS-120B および GPT-OSS-20B の 2 種類のローカル LLM であり、以下の二つの条件によって出力を行う。

- i. 解釈困難なメッセージを単体で入力し、その意味を記述させる
- ii. 解釈困難なメッセージを前後の文脈 (前後各 20 件) を含めて入力し、その意味を記述させる

3.2 節で構築した正解データとの比較はジェネラルパターンマッチング [16] による文字列の類似度と、日本語用 Sentence-BERT[14] による文埋め込みベクトルに基づく類似度の二つの側面から行った。

4.2 実験結果

大規模言語モデルによる解釈結果と人手によって構築した正解データとの類似度を表 3 に示す。本実験では、両モデルとも条件 ii は条件 i と比べて類似度が向上しており、文脈情報の付与により LLM の解釈精度が向上する傾向が確認された。特に GPT-OSS-120B では、条件 ii における文埋め込み類似度が 0.763 と最も高く、文脈情報の付与による効果が顕著であった。

一方で、GPT-OSS-20B と GPT-OSS-120B を比較すると、いずれの条件においてもモデル規模の大きい GPT-OSS-120B の方が高い類似度を示す傾向が見られた。この結果から、モデル規模の違いが解釈困難なメッセージの理解性能に影響を与えることが示唆される。

5 考察

本章では、3, 4章で得られた実験結果を踏まえ、解釈が困難となる理由と、人間および LLM の解釈特性について考察する。

5.1 人間における解釈困難性の分析

3章では、解釈困難なメッセージは日本人大学院生にとって文脈情報があっても外部知識なしでは解釈が困難であった。これは文脈情報が話題の特定に寄与する一方で、用語そのものの意味が不明な場合に文脈だけでは十分な解釈に至らないためだと考えられる。また、解釈対象としたメッセージはほとんど英語であり、スラングや隠語が一般的な英語教育の範囲を超えた知識を要求していたと考えられる。

これに対し、外部知識を参照したことで解釈は大きく改善した。これは外部知識を通じて、不足している専門知識や語義情報を補完できたためだと考えられる。Web 検索を用いた解釈では、一般的なスラングや比較的広く知られた専門用語についてはその言葉に言及した記事に到達することができた。一方で対象語が多義語である場合には、より一般的な意味を持つ語義が優先的に検索結果に表示され、悪質なコミュニティにおける用法に到達することが困難な場合があった。また、ゲームなどの狭いコミュニティでのみ使用される専門用語や口伝的に広まっているスラング・隠語は、Web 上に掲載されている情報が少なく解釈が難しい傾向が見られた。LLM を解釈に利用する際にもこの傾向は見られたが、文脈情報から Web 上の記事にならないような抽象的な概念を推論できる場合があった。

しかし、外部知識を参照してもすべてのメッセージを正しく解釈することはできなかった。例えば、“Lemme google rw” というメッセージは“rw を Google で調べてみて” というような意味だが、この“rw” が何を指しているのかは前後の文脈からも推定することができなかった。このような事例から、外部知識の利用によっても解釈困難性が完全に解消されるわけではないことが示唆される。

また、単語の綴り・文法誤りは、隠語やスラングと同時に発生することで解釈困難性をさらに高める要因となった。例えば“how do u loot ppl need to be a group memb” というメッセージは文法が崩れており、“loot ppl” が「人を奪う」という意味なのか「人から物を奪う」という意味なのかで解釈が分かれた。

5.2 LLM における解釈困難性の分析

4章では、文脈情報の付与により LLM の解釈精度が向上する傾向が確認された。しかし隠語に見られるような強い隠喩表現を含みつつ、無害な発言としても解釈できるメッセージに対しては LLM であっても誤った解釈を出力することがあった。例えば、“isnt that the child pizza video?” というメッセージは「それは児童ポルノの動画ですか」という意味だが、LLM は文脈情報があっても「子どもがピザを食べている動画ですか？」という解釈を出力してしまった。

またスラングの中でも、同じ話題で複数の意味が想定される略語に対して誤った解釈をすることがあった。“How much djs left?” というメッセージでは、“djs” は“david jones” という百貨店の略語であったが、LLM は“discounts” の略語の誤記として捉えてしまった。このような略語は明示的な語義や文法構造に依存せず、特定のコミュニティ内の共有知識に基づいて意味が成立することがあるため [5]、それらの知識が学習データや入力した文脈情報に十分に含まれていない場合には LLM による推論が難しくなると考えられる。そのため、今回使用したモデルよりも最新のスラング・隠語等を学習データに多く含むモデルではさらに解釈精度が向上する可能性がある。

一方で、“(C)heese P)izza)” のように文字の分割や記号挿入によってシステムを欺く表現については、文脈情報と合わせて元となる語を推定できた場合、比較的正しい解釈が得られた。これは LLM が表層的な文字列変形に対して、一定の頑健性を持つことを示唆している。また本研究では、モデル規模の大きい LLM がより高い類似度を示した。これは規模の大きいモデルほど蓄積されている概念の解像度が高く、深い推論ができるためだと考えられる。

6 おわりに

本研究では、Discord 上の悪質なコミュニティにおける解釈困難なメッセージを対象とし、解釈困難性の要因を分類しつつその特徴と解釈の難しさを明らかにした。

今後はより多様な領域を対象とした分析や、本研究で得られた知見を活用した解釈支援手法の検討を通じて、サイバー犯罪の特定や早期発見に資する技術の発展に貢献したいと考えている。

謝辞

本研究の一部は、NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託事業「経済安全保障重要技術育成プログラム／先進的サイバー防御機能・分析能力強化」（JPNP24003）によるものである。

また、実験に協力いただいた新村柊月氏、高田尚輝氏、吉田健人氏に深く感謝いたします。

参考文献

- [1] Cognyte. The rise of cybercrime on telegram and discord and the need for continuous monitoring, 2020. <https://www.cloudsek.com/blog/the-rise-of-cybercrime-on-telegram-and-discord-and-the-need-for-continuous-monitoring>.
- [2] 警察庁サイバー警察局. 令和6年におけるサイバー空間をめぐる脅威の情勢等について, 2025. <https://www.npa.go.jp/publications/statistics/cybersecurity/index.html>.
- [3] INTEL471. How discord is abused for cybercrime, 2024. <https://www.intel471.com/blog/how-discord-is-abused-for-cybercrime>.
- [4] Ulrich Ludewig, Nicolas Hübner, Sascha Schroeder. Vocabulary, text coverage, word frequency and the lexical threshold in elementary school reading comprehension. **Reading and Writing**, Vol. 36, p. 2409–2431, 2023.
- [5] Abhinav Bhandari, Caitrin Armstrong. Community characteristics and semantic shift of high affinity terms on red-dits. In **Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)**, 2019.
- [6] Zhewei Sun, Richard Zemel, Yang Xu. Semantically informed slang interpretation. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, 2022.
- [7] Yiwei Wang Baolong Bi Xueqi Cheng Lingrui Mei, Shenghua Liu. Slang: New concept comprehension of large language models. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, 2024.
- [8] 川口大翔, Yin Minn Pa Pa, 吉岡克成. Llm を利用した discord 上のサイバー犯罪関連の隠語の調査. コンピュータセキュリティシンポジウム 2024 論文集, pp. 222–229, 2024.
- [9] Dong Nguyen, Barbara McGillivray, Taha Yasserir. Emo, love and god: making sense of urban dictionary, a crowd-sourced online dictionary. **Royal Society Open Science**, Vol. 5, p. 172320, 2018.
- [10] Bo Han, Timothy Baldwin. Lexical normalisation of short text messages: Mkn sens a #twitter. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, 2011.
- [11] 川口大翔, Yin Minn Pa Pa, 吉岡克成, 松本勉. Discord 上のサイバー犯罪に対する chatgpt を利用した情報収集システム. 暗号と情報セキュリティシンポジウム 2024, 2024.
- [12] Princeton University. About WordNet, 2010. <https://wordnet.princeton.edu/>.
- [13] ウィキメディア財団. Wiktionary, 2025. https://en.wiktionary.org/wiki/Wiktionary:Main_Page.
- [14] Isao Sonobe. sonoisa/sentence-bert-base-ja-mean-tokens-v2, 2024. <https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2>.
- [15] Xu Ji, Jianyi Zhang, Ziyin Zhou, Zhangchi Zhao, Qianqian Qiao, Xiali Hei Kaiying Han, Md Imran Hossen. Can't say cant? measuring and reasoning of dark jargons in large language models. In **Security and Privacy in Communication Networks**, 2025.
- [16] John W. Ratcliff, David Metzener. Pattern matching: The gestalt approach. **Dr. Dobb's Journal**, p. page 46, 1988.

解釈対象テキスト	レビュー済み正解データ	メモ	Ollama-GPT-OSS-120B(+前後40発話)
If you fish react me I will drop meth recipe	もし私にリアクションをくれたら、メタンフェタミン(覚せい剤)のレシピを共有するよ。	fishは魚のリアクション。 meth=メタンフェタミン	私にリアクションしてくれたら、メタンフェタミン(覚せい剤)の製造手順を教える
Binance ma ENA pr agr 100 fdusd stake krta hu to koi loss to nhi hga?	Binance(暗号通貨取引所)でENA(分散型金融プロトコル)に100 FDUSD(ステーブルコイン)をステーキング(一定期間預けること)したら、損は出ませんか?	ヒンディー語 ma=?,pr=? ア.agr=?,krta=?,hu=?,koi loss to nhi hga=損失は出ませんか?	BinanceでENA(エナ)に対して100 FDUSD(ステーブルコイン)をステーキングしたら、損は出ませんか?
Markets trolling ngl. high liq. Today	相場が今日は意図的にかからかうように動いていると感じる。流動性が高く荒れやすい状態だ	ngl=not gonna lie,liq=liquidite	相場がかからかうように(意図的に)動いていると正直に言える。今日は流動性が高く、価格が大きく変動しやすい。
Great bro sry couldn't reply, wbu?	よかったよ、冗責。返事できなくてごめん。君はどう?	wbu=what about you	ごめん、返信できなかったんだ。君はどう?
Sum ppl catfish	一部の人はなりすましをしている	ppl=people.catfish=なりすまし	一部の人は(他人を)キャットフィッシュ(偽の身分で騙す)ことをする
i mean if anyb is tryna bless me up w 5\$ rq thru cashapp	「誰かがCash Appで5ドルほど送ってくれたら助かります」と金銭的な支援を求めている	tryna=trying to. thru=through	もし誰かが5ドルを私に送ってくれる(="bless me up")なら、Cash Appでリクエストしてください
\$10 PayPal to chat w me n my mommy I send media too!	「PayPalで10ドル支払えば、私と私のママとチャットできるよ。メディア(画像・動画)も送ります!」という宣伝。	w=with,n=and	\$10(米ドル)をPayPalで支払えば、本人とその母親とチャットでき、さらに画像・動画などのメディアを送ってくれる
I used to buy QPs of shrooms from some chick named <name4>	以前は<name4>という子から、マジックマッシュルームを大量に買っていた	QPsはquantity packs(大量パック)またはquarter-packs(4分の1パック)	<name4>という女性から幻覚作用のあるマジックマッシュルーム(shrooms)を、一定量(QPs)購入していた

図1 実験で構築したデータの一部

A 実験で構築したデータ

3, 4章の実験で構築したデータの一部を図1に示す。より詳細なデータは、[GitHub](#)で公開している¹⁾。

B LLM への入力プロンプト

システムプロンプト

「あなたは英語文の意味を日本語で噛み砕いて伝えるアシスタントです。元の意図やニュアンスが伝わるように自然な日本語で解釈を返してください。入力文を繰り返したり英語を返す必要はありません。回答は短く端的にまとめてください。」

プロンプト

「上記の★で囲まれた文の解釈を教えてください。回答は端的に、一文でまとめてください。」

1) <https://github.com/Tiny-Colony/annotation-2025>