生成 AI を用いた詐欺メッセージ判定支援システムの設計

Design of a Scam Message Judgment Support System Using Generative AI

森 優希斗，小渡 悟
Yukito MORI，Satoru ODO
沖縄国際大学産業情報学部
Department of Industry and Information Science, Okinawa International University
Email: 22DB123@okiu.ac.jp

あらまし：近年，生成 AI の発展により，自然な文章を容易に生成できるようになった一方で，その悪用によるフィッシング詐欺やなりすましメッセージが問題となっている．従来の詐欺検知手法は電子メールを主な対象としており，SMS や SNS など多様な媒体への対応には課題がある．本研究では，大規模言語モデル（LLM）の推論能力を活用し，複数媒体に対応した詐欺メッセージ判定支援システムを構築した．提案システムは，入力されたメッセージを分析し，詐欺の可能性とその根拠を自然言語で提示する．詐欺メッセージ 90 件および正規メール 50 件を用いた評価実験の結果，一定の判定精度が確認された．また，利用者調査の結果から，判定根拠を提示することが防犯意識の向上に寄与する可能性が示唆された．
キーワード：生成 AI，詐欺メッセージ判定，防犯リテラシー

1. はじめに

近年，ChatGPT に代表される生成 AI の発展により，人間が作成したものと区別が難しい自然言語文章を容易に生成できるようになった．この技術は業務効率化や教育支援など多方面で活用されている一方で，フィッシング詐欺やなりすましといった不正なメッセージの作成に悪用される事例も報告されている．フィッシング対策協議会の報告によれば，フィッシング報告件数は増加傾向にあり，利用者が不審なメッセージに接触する機会は今後も継続して増大すると予測される[(1)]．

従来の詐欺検知手法としては，特定キーワードに基づくフィルタリングや送信元ドメインの検証，悪意ある URL のブラックリスト照合などが用いられてきた．しかし，これらの多くは電子メールを主対象としており，SMS や SNS のダイレクトメッセージなど，短文や口語表現を含む媒体への対応には課題がある．また，多くの検知システムでは判定結果のみが提示され，利用者が「なぜ詐欺と判断されたのか」を理解しにくいという問題も指摘されている[(2)(3)]．

本研究では，大規模言語モデル（LLM）の推論能力と自然言語による説明生成能力に着目し，複数媒体に対応した詐欺メッセージ判定支援システムを設計・構築する．提案システムは，入力されたメッセージを複数の観点から分析し，詐欺の可能性とその根拠を自然言語で提示することを特徴とする．これにより，単なる自動判定に留まらず，利用者自身が不審な点に気づくための注意喚起支援を行うことを目指す．

2. 提案システム

本研究で構築する詐欺メッセージ判定支援システムは，利用者が受信した任意のメッセージを入力した際に，詐欺の可能性とその根拠を自然言語で提示することを目的とする．本システムは，電子メールに加え，SMS や SNS のダイレクトメッセージ等，多様な媒体を介して届く短文メッセージを対象とする点に特徴がある．

システムの基盤には，大規模言語モデル（LLM）を基盤とした GPTs を採用する．GPTs に対しては，「詐欺メッセージ判定の専門家」としての役割を定義し，入力文に対して一貫した分析を行うようプロンプトを設計する．判定プロセスは，①送信元の不自然さ，②緊急性の強調，③個人・金銭情報の要求，④リンクの信頼性，という 4 つの観点から段階的に実施する．各観点に基づく分析結果を踏まえ，システムは詐欺の可能性を「高い」「低い」「判断が難しい」の 3 段階で判定する．加えて，判定の根拠を自然言語で記述し，利用者が取るべき行動に関する注意喚起を生成する．これにより，単なる自動分類に留まらず，利用者が詐欺メッセージの特徴を学習し，将来的な判断能力を向上させることを支援する設計とする．図 1 に提案システムの処理の流れを示す．

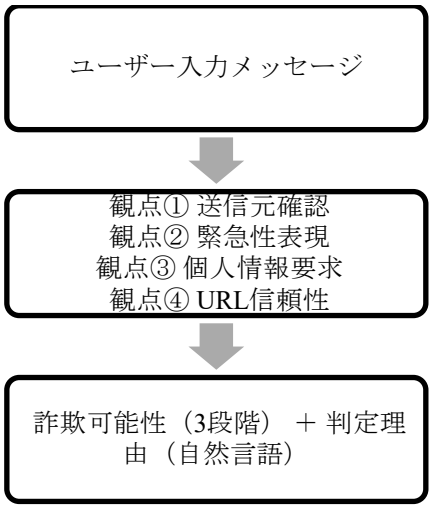

図 1　提案システムの判定フロー

3. 実証実験

3.1 実験方法

提案システムの有用性を検証するため，詐欺メッセージおよび正規メッセージを用いた判定実験を行った．評価対象は，インターネット上で公開されている詐欺事例および実際に受信したメッセージから収集した詐欺メッセージ 90 件と，大学や正規サービスから送信された正規メール 50 件の計 140 件であった．詐欺メッセージには，金銭要求，個人情報要求，偽サイトへの誘導を含むものを選定した．

各メッセージをシステムに入力し，出力された判定結果と実際の分類を比較することで，正解率，偽陽性および偽陰性の発生状況を確認した．本実験は，大規模な精度評価を目的とするものではなく，提案システムの基本的な判定性能を確認することを目的として実施した．

3.2 判定結果

実験の結果を表 1 に示す．詐欺メッセージ 90 件のうち 87 件（真陽性）を詐欺と判定し，正規メール 50 件のうち 45 件（真陰性）を非詐欺と判定する結果を得た．全体の正解率は約 94.3%であり，提案システムが一定の判定精度を有していることが示された．

一方で，正規メールを詐欺と誤判定する偽陽性が 5 件，詐欺メッセージを見逃す偽陰性が 3 件確認された．偽陰性は，送信元情報やリンク先 URL など，判定に必要な情報が十分に含まれていないメッセージにおいて発生する傾向が見られた．これらの結果から，入力情報の量や内容が判定精度に影響を与える可能性が示唆された．

表 1　システムの判定結果

実際の分類 / 判定結果	詐欺 (不審)	非詐欺 (安全)	合計
詐欺メッセージ	87 (真陽性)	3 (偽陰性)	90
正規メール	5 (偽陽性)	45 (真陰性)	50
合計	92	48	140

4. 考察

実証実験の結果から，提案システムは，電子メールに加えて SMS や SNS のダイレクトメッセージといった複数媒体における詐欺メッセージ判定において，一定の有効性を有することが確認された．特に，詐欺の可能性を提示するだけでなく，判定に至った理由を自然言語で説明する点は，従来の自動判定型システムとは異なる特徴である．

利用者調査の結果からは，判定結果に加えて根拠が提示されることで，利用者がメッセージ内容をより慎重に確認する意識を持つ可能性が示唆された．アンケートでは，「判定理由の理解しやすさ」や「警戒心の向上」に関して比較的肯定的な評価が得られており，AI による説明が利用者の納得感や注意喚起に一定程度寄与していると考えられる．これらの結果は，単なる自動判定に留まらず，利用者の判断を支援する仕組みとしての有用性を示すものである．

一方で，送信元情報やリンク先 URL などの情報が不足しているメッセージに対しては，判定が難しくなる傾向が見られた．また，正規の通知であっても緊急性を強調する表現を含む場合には，詐欺と誤判定される可能性がある．利用者調査においても，AI の判定結果をそのまま信頼することへの不安が一部で示されており，本システムはあくまで利用者の判断を補助する支援ツールとして位置付けることが適切である．

5. まとめ

本研究では，大規模言語モデル（LLM）の推論能力と自然言語による説明生成能力を活用し，複数媒体に対応した詐欺メッセージ判定支援システムを設計・構築した．提案システムは，電子メールに限定されがちであった従来の詐欺対策を，SMS や SNS のダイレクトメッセージへと拡張し，詐欺の可能性とその根拠を自然言語で提示する点に特徴がある．

実証実験の結果から，提案システムは一定の判定精度を有していることが確認された．また，利用者調査により，判定理由を提示する仕組みが，利用者の理解や警戒心の向上に寄与する可能性が示唆された．これらの結果は，単なる自動判定に留まらず，利用者の判断を支援する注意喚起ツールとしての有用性を示すものである．

一方で，入力情報が不足しているメッセージに対する判定の難しさや，正規メッセージを詐欺と誤判定する可能性といった課題も明らかとなった．今後は，追加情報を対話的に取得する仕組みの導入や，判定結果の提示方法の改善を通じて，実用性の向上を図る必要がある．

謝辞

本研究は JSPS 科研費 25K04261, 25K04260 の助成を受けたものです．

参考文献

(1) フィッシング対策協会:“月次報告書”，https://www．antiphishing．jp/report/monthly/202503．html （参照日 2026 年 1 月 10 日）．

(2) NTT Security：“ChatSpamDetector：生成 AI によるフィッシングメール検知”，https://jp．security．ntt/tech_blog/chatspamdetector-ai （参照日 2026 年 1 月 10 日）．

(3) Mahendru，S．and Pandit，T．：“A Comparative Study of DeBERTa and Large Language Models for Phishing Detection”，2024 IEEE 7th International Conference on Big Data and Artificial Intelligence (BDAI)，pp.160-169 (2024)