LLM搭載判断支援型詐欺防止装置

IT未来高等学校　二年次　麻窪眩生

1. 概要

特殊詐欺の被害をなくす！！

本研究は、特殊詐欺による高齢者等の被害を未然に防ぐことを目的とした LLM搭載判断支援型詐欺防止装置の開発である。

従来の対策機器とは異なり、「警察」や「銀行」といったキーワードマッチで判定するのではなく、LLMが通話内容全体の文脈を読み解き、その会話が詐欺特有の要求や誘導を含んでいるかを高精度で判断することを最大の特長とする。

固定電話機の環境に簡単に設置ができ、特殊詐欺かを判定した結果を利用者とその家族に対して判定結果を通知することができることを目標とした。

従来の対策機器の問題点

既知の詐欺の通話内容から抽出した
キーワードを基準に判断
↓
新たな詐欺手法に対応できない

本研究の特徴

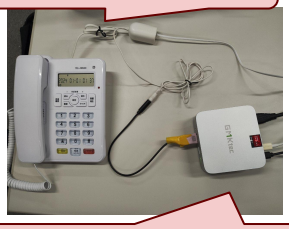
通話内容の文脈をLLMが判断
↓
新たな詐欺手法にも対応できる

2. 機能説明

設置が手軽

固定電話機とモジュラーケーブルの間に挟み込むだけ で設置ができる

電話線と電源ケーブルを挿すだけ！



発信者電話番号を取得

モジュラーケーブル上に流れる
発信者電話番号情報 を取得
↓
取得した番号を自動で Web検索
特殊詐欺履歴があるかを確認

デバイス上で 文字起こし&LLMで詐欺を検知

LLMで通話内容を評価

モジュラーケーブル上に流れる
通話音声をリアルタイム録音
↓
デバイス上で文字起こし
↓
文字データを デバイス上の LLMで評価

詐欺だと本人・家族に通知

発信者電話番号情報・LLMによる評価値で詐欺の可能性が高いと判定
↓
本人・家族にメールで通知

LLMが実際詐欺を検知できるのか？

3. プロンプトエンジニアリング

使用したLLM　Gemma 3n 4B

試行1

「詐欺か詐欺ではないか検知して」と尋ねる
→ 出力形式が安定せず、誤検知も多発し失敗した。

試行2

「出力例: 詐欺の確率 100%」と形式を指定
→ 出力は統一されたが、肝心の判定精度は改善されず、さらなる工夫が必要

試行3

役割の付与、詐欺の定義、出力形式の厳格化を設定
→ このプロンプト設計により、小規模なLLMでもある程度の一貫性と高い精度を引き出すことに成功した。

このプロンプトは本当に最適？
本当にモデルにこのプロンプトは合うの？

4 .LLMとプロンプトの組み合わせ検証

本システムでは、LLMが重要な役割を担っているためLLMの精度の検証実験を行った。
LLMとプロンプトの組み合わせが最も実用的な組み合わせを探す。

実験方法

プロンプトエンジニアリングのみに着目し、複数の著名なLM（GGUF形式）の性能を比較・評価する。各モデルとプロンプトの組み合わせについて3回ずつ実行し、出力結果の質と安定性を評価した。
各LLMに対し、以下の6種類のプロンプト手法（+ 統合プロンプト）を用いて性能を評価した。

1. 独自プロンプト: 研究者がタスクに合わせて用意したプロンプト
2. ゼロショット: タスク指示のみを与える
3. 役割付与: 特定の専門家としての役割を与える
4. ヒューショット(Few-shot): いくつかの例示（入力と出力のペア）を与える
5. CoT (Chain of Thought): 思考の連鎖を促す（例「ステップバイステップで考えて」）
6. 統合プロンプト: 上記の要素（ゼロショット、役割、ヒューショットCoT）を組み合わせたプロンプト

実験環境

Google colab上でllama-cpp-pythonライブラリを使用
temperature=0.1、max_tokens=1500、n_ctx=16384　T4GPU使用

謝辞

本研究は国立研究開発法人情報通信研究機構が実施する、セキュリティイノベーター育成プログラムSecHack365における研究開発の成果である。
若手セキュリティイノベーター育成プログラムSecHack365https://sechack365.nict.go.jp/

対象LLM

- lmstudio-community/OREAL-DeepSeek-R1-Distill-Qwen-7B-Q4_K_M.gguf
- google/gemma-3-4b-it-qat-q4_0-gguf
- NoelJacob/Meta-Llama-3-8B-Instruct-Q4_K_M-GGUF
- ldostadi/Phi-4-mini-reasoning-Q4_K_M-GGUF
- Qwen/qwen2.5-7b-instruct-q4_k_m-*.gguf

		詐欺通話		非詐欺通話		
モデル名	プロンプト	詐欺1	詐欺2	セールス	警察	友達
DeepSeek	1. 独自	3	5	1.5	1	1
	2. ゼロショット	0	3	1	0	1
	3. 役割	4	4	1	4	1
	4. ヒューショット	5	5	2	3	2
	5. CoT	4	4	1	1	1
	6. 統合	5	4.5	5	5	1
Gemma	1. 独自	4	4	3	3	2
	2. ゼロショット	3.5	4.5	4	3	3
	3. 役割	4	4	4	4	3
	4. ヒューショット	4	5	3	4	2
	5. CoT	4	4	4	4	2
	6. 統合	4	5	4	5	3
Llama	1. 独自	4	4	4	4	1
	2. ゼロショット	4.5	3.8	3	3.5	1.8
	3. 役割	4	4	4	4	2
	4. ヒューショット	4	4	1	2	1
	5. CoT	5	4	4	4	1
	6. 統合	4	4	3	4	1
Phi	1. 独自	5	エラー	4	5	エラー
	2. ゼロショット	5	4.5	4	4	4
	3. 役割	5	5	エラー	4	2
	4. ヒューショット	4	5	4	5	3.5
	5. CoT	5	5	4	4	2
	6. 統合	5	4	4	5	エラー
Qwen	1. 独自	5	5	4	4	1
	2. ゼロショット	4	4	4	4	1
	3. 役割	4	4	4	4	2
	4. ヒューショット	4	5	4	2	1
	5. CoT	4	4	3	2	1
	6. 統合	4	4	1	1	1

【通話内容】
詐欺事例1
千葉県警PDF[2]
詐欺事例2
千葉県警PDF[3]
非詐欺事例3
警察官からの被害確認電話例
セールス事例4
ChatGPT生成
日常会話事例5
ChatGPT生成

【評価基準】
5: 特殊詐欺であると強く疑われる（または断定できる）
4: 特殊詐欺の可能性が非常に高い
3: 疑わしい点があり、特殊詐欺の可能性がある
2: 特殊詐欺の可能性は低い
1: 特殊詐欺の可能性は極めて低い（または、完全に問題ない）

5. 結果

赤色＝誤検知　エラー＝評価値を数字として出さない
黄色＝評価値を返すがエラーの場合もある　で塗っている。

各LLMで非詐欺事例でセールスや警察を示したプロンプトの際結果が良かった手法は以下の通りであった。

- Qwen (qwen2.5-7b):
  - 統合プロンプト
- Llama (Meta-Llama-3-8B):
  - ヒューショット
- DeepSeek (OREAL-DeepSeek-R1):
  - 独自プロンプト
  - CoT

DeepSeekは詐欺判定に関してはすべての事例に対して最も正しい評価値を返すことができる。だが、評価値を出さないエラーケースが多いため実用性に欠けるため除外する。よって、結果が評価値を必ず返し、警察やセールスの非詐欺事例でも正しい評価値を返すLlamaのヒューショット、Qwenの統合プロンプトが実用的という結果になった。

6. 考察

- 今回の実験はモデルを評価するためにプロンプトを固定化し、同一プロンプトをすべてのモデルに対して実行した。結果モデルによって正しい評価値を返すプロンプトが異なることが分かった。そのためプロンプトとモデルはセットで考えなければならない。
- 公式発表のLLMと比べて有志によって量子化された非公式モデルはエラーが出力され不安定になるので使用は難しい。

7. まとめ

本研究成果を基に、私は AIによる「判断支援型 詐欺防止装置」の実用化を目指す。現状、LLMをプロンプトエンジニアリングのみで詐欺の判定をするには、検知が精度が不十分である。他にも検知に限らず製作するデバイスに正しく文字起こしができないなどの問題点がある。

今後の展開として、デバイスの機能面にある文字起こしや LLMの判定処理遅延や通話内容を家族に送る際の同意取得などの問題がなくなり、 LLMの精度を上げるだけの状態になったらファインチューニングを行い精度を高めていく。

本ポスターで参照した先行研究とサイト

・AIと犯罪心理学を活用し特殊詐欺を未然に防ぐ日本初の共同研究を尼崎市で開始　（富士通）
URL: https://pr.fujitsu.com/jp/news/2022/03/24.html
・シャープの AIで安心！ 迷惑電話対策で快適スマホライフ　（シャープ株式会社 ）
URL: https://k-tai.sharp.co.jp/dash/lp/meiwakuai/index.html
・『詐欺電話に延々と応対し時間を浪費させる　AIおばあちゃん、O2が開発』(note / Google Pixelの機能にも言及)
URL: https://note.com/monika_wa/n/nd8dd0c87e5d8
・生成 AIで詐欺電話を再現、高齢者の訓練に　富士通など、被害防止へ新技術　(ITmedia NEWS)
URL: https://www.itmedia.co.jp/news/articles/2312/01/news113.html
・AIが詐欺電話を解析するサービスを　NTTが開始、警察の意見も参考に　(Ledge.ai)
URL: https://ledge.ai/articles/ntt-scam
・警察庁「令和6年上半期における特殊詐欺の状況について」
URL: https://action.digipolice.jp/files/15cc1537a0df7b599107a81e202984a3.pdf
・消費者庁「令和 5年版消費者白書　第1部 第2章 第2節 (2) 特殊詐欺の被害状況」
URL: https://www.caa.go.jp/policies/policy/consumer_research/white_paper/2023/white_paper_1_02_02.html
・警察庁「国際電話対策の強化に向けた連携」
URL: https://www.npa.go.jp/bureau/criminal/souni/tokusyusagi/sagi_keihatsu2024.pdf