SecHack365

LLM搭載判断支援型詐欺防止装置

SecHack365 学習駆動コース 社会実装ゼミ 1Sx 麻窪晄生

1. 概要

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。

高齢者の利用が多い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。

高齢者の特殊詐欺の被害をなくす！！

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

電話線と電源ケーブルを挿すだけ！

デバイス上で文字起こし＆LLMで詐欺を検知

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

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

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

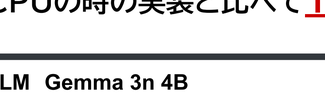
詐欺だと本人・家族に通知

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.電話線を分岐によるノイズと入力音量問題

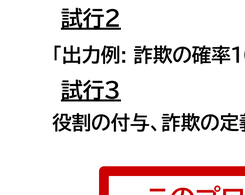
「DRA-L62」というテレフォンピックアップを採用しましたが、これにより、法的な安全性を確保することはできたましたが、
ACアダプタからのノイズが入ったり、もともと採用した市販品は会話をPCにつないで録音するよう設計されていないため、
入力音量が小さくAIが文字起こしできないなどの問題がいくつか発生しました。



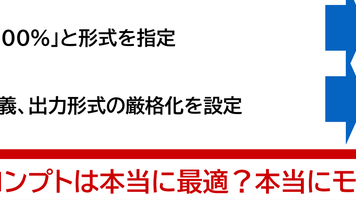
問題解決

ACアダプタからのノイズ対策

機能するか不明
フェライトコア

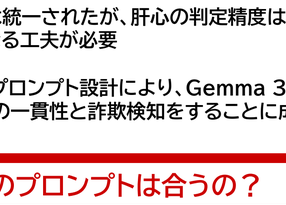


原因不明だけど、ノイズ抑制成功
ACアダプタを電源用挿入口ではなくタイプCに



入力音量の改善

アンプを使って入力音量を増幅



4.処理速度改善問題

文字起こしをする際、CPUで動かす際に処理速度が速くなるよう整備されているfaster-whisperというライブラリを使用した。

↓

結果:数十秒の会話文でも文字起こしから詐欺判定まで数分かかってしまう…

↓

CPUではリアルタイムを実現できない！！GPUを使うしかない！！

↓

AMD版のCUDAと同様の純正プラットフォーム「ROCm」を採用したが、特定のカーネルバージョンや
GPUアーキテクチャへの依存度が非常に高く失敗

どうしてもGPUで動かせるようにしたい！！

↓

Vulkanという3D描画用のグラフィック技術（API）を計算処理に転用し、高パフォーマンスと低オーバーヘッドを実現

↓

本デバイスのGPUで処理できるようになり、CPUの時の実装と比べて10倍程度早くなった。

5.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、Gemma 3n 4Bでもある
程度の一貫性と詐欺検知をすることに成功した。

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？

6 .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LLM（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5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. フューショット (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

【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）

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

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
- Idostadi/Phi-4-mini-reasoning-Q4_K_M-GGUF
- Qwen/qwen2.5-7b-instruct-q4_k_m-*.gguf

モデル	プロンプト	Precision（適合率）	Recall(再現率)	F1 Score	判定の傾向
DeepSeek	1. 独自	1	1	1	完璧
	2. ゼロショット	1	0.5	0.67	見逃しあり
	3. 役割	0.67	1	0.8	バランス型
	4. フューショット	0.5	1	0.67	誤検知多め
	5. CoT	1	1	1	完璧
	6. 統合	0.4	1	0.57	誤検知非常に多い
Gemma	1. 独自	0.5	1	0.67	過剰反応(中)
	2. ゼロショット	0.4	1	0.57	全判定(過剰)
	3. 役割	0.4	1	0.57	全判定(過剰)
	4. ヒューショット	0.5	1	0.67	過剰反応(中)
	5. CoT	0.5	1	0.67	過剰反応(中)
	6. 統合	0.4	1	0.57	全判定(過剰)
Llama	1. 独自	0.4	1	0.57	過剰反応
	2. ゼロショット	0.4	1	0.57	過剰反応
	3. 役割	0.4	1	0.57	過剰反応
	4. フューショット	1	1	1	完璧
	5. CoT	0.4	1	0.57	過剰反応
	6. 統合	0.4	1	0.57	過剰反応
Phi	1. 独自	0.33	0.5	0.4	不安定・見逃し
	2. ゼロショット	0.4	1	0.57	全判定(過剰)
	3. 役割	0.67	1	0.8	意外と良好
	4. ヒューショット	0.4	1	0.57	全判定(過剰)
	5. CoT	0.5	1	0.67	過剰反応(中)
	6. 統合	0.5	1	0.67	過剰反応(中)
Qwen	1. 独自	0.4	1	0.57	過剰反応
	2. ゼロショット	0.4	1	0.57	過剰反応
	3. 役割	0.4	1	0.57	過剰反応
	4. フューショット	0.67	1	0.8	良好
	5. CoT	0.67	1	0.8	良好
	6. 統合	1	1	1	完璧

【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

実験で使用した使用したプロンプトの例（フューショット）

以下の例のように、会話内容が特殊詐欺かどうかを判定してください。
例1（詐欺）:
会話:「もしもし、俺だよ、俺。事故っちゃってさ、すぐにお金が必要なんだ。」
判定: 特殊詐欺の可能性が高い
例2（非詐欺）:
会話:「こんにちは、〇〇銀行です。来月の引き落としについてのご連絡です。」
判定: 特殊詐欺の可能性は低い

フューショットの手法として、例として二つ、詐欺と非詐欺の会話文と回答の組み合わせ、与えるプロンプトにした。

7. 結果

実験はF1スコアで表現した
・Precision（適合率）:どれだけ結果が正確だったか
・Recall（再現率）:どれだけ詐欺を取りこぼさなかったか
・F1スコア: PrecisionとRecallのバランスを示す総合点

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):
  - フューショット
- DeepSeek(OREAL-DeepSeek-R1):
  - 独自プロンプト
  - CoT

Qwen2.5-Math-7Bモデルに対して DeepSeekで蒸留したものを
OREALという効率よく学習させる手法を使ったモデル

・最適な組み合わせは Llama+フューショットの組み合わせと Qwen+統合の組み合わせとなりましたが、本デバイスに導入できる LLMのパラメータ数が 7B未満のため導入が不可能だった。

・現在はGemma＋役割の組み合わせで実装している。

・LLMごとに得意なタスクがあり、詐欺判定の精度を上げるためには、文脈の理解に特化した LLMを選定する必要がある。

・LLMとプロンプトの最適な組み合わせは LLMごとに異なる。

・判定に使う出力を制限するには、役割プロンプトを指示する必要がある。

8.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。
- 推論回数を増やしたり、他の4Bモデルの LLMを利用、プロンプト文の変更などを行い、実験を増やして最適なモデルとプロンプトの組み合わせを見つける必要がある。

9.今後の展開

- RAGを用いた詐欺判定の向上とその参考資料をアップデートできる機能の追加をする。
- 高齢者の方に本デバイスを日常的に使用してもらい実証実験を行う。
- 通知機能の自動音声を家族の声が録音された音声データに変更できる機能をつける。
- 通知に視覚を用いてもっと危険を伝える。
- 商品化を睨んでデバイスのコンパクト化をする。
- LLMとプロンプト選定の検証件数を増やす。
- Gemini検索機能を使わなくてもいいように外部からデータベースに危険な電話番号を追加できるようにする。

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、02が開発』（note / Google Pixelの機能にも言及）
URL: https://note.com/morika_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

謝辞

SecHack365の期間を通して、トレーナーの方々の作品に対するコメントや成果発表会の資料作りに関するアシスタントの方々の助言は大変参考になりました。トレーニーや運営の方々を含めて、多くのご支援をいただいたことを、この場を借りて深く感謝を申し上げます。

NICT National Institute of Information and Communications Technology