フェイク情報を簡単に早く判別するWebアプリ「フェイクチェッカー」の制作

群馬県立前橋高等学校　1年　岡田 武　鏑木 友斗

背景

フェイク情報が増加：対策は **個人のリテラシーに依存** *1

ディープフェイクの判別に不安：**84%** *2

真偽判定ツール URL：一般向け複数あり　**画像：専門性大**

動機・目的

ChatGPTでフィッシングサイト判定：**98%で判定** 可能*3

➡ **ChatGPT4V** なら、画像の真偽判定も可能？しかし、日本人の**9.1%しか生成AIを利用していない** *4

目的

URLや画像を**簡単に**ChatGPT4Vに真偽判定させる仕組みをつくり、リテラシーの補助ツールにする

検証1：フェイクチェッカーのプロトタイプ開発

システム概要

ChatGPT4VでのURLや画像の真偽判定を可視化するWebアプリ「**フェイクチェッカー**」を制作した(図1)。

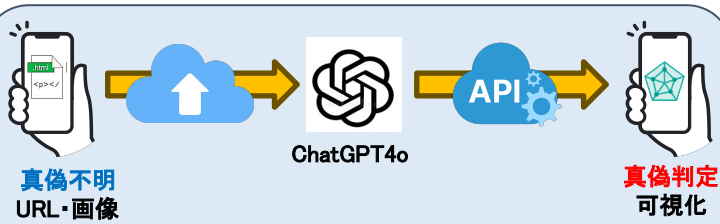

図1:「フェイクチェッカー」の操作イメージ

フェイクチェッカーは判定部と表示部から構成され、それぞれHTMLとJavaScriptで制作した。

判定部 データとプロンプトは**JS上に表記**した（図2）。

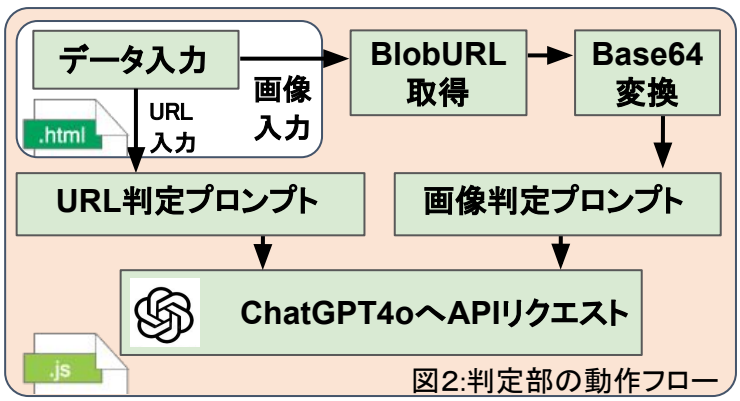

図2:判定部の動作フロー

表示部 総合評価は1～10表記をA～E表記とした（図3）。

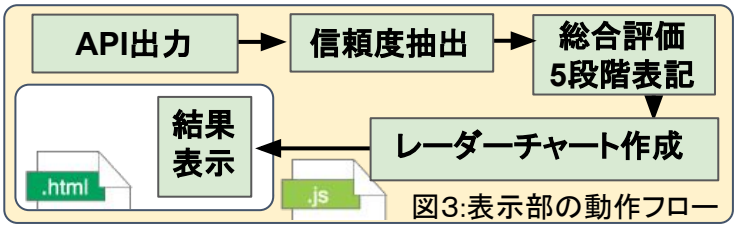

図3:表示部の動作フロー

検証1の結果・考察

フェイクチェッカーのUIを示す(図4)。

フェイクチェッカーは MonacaEducationで制作 Webアプリとして公開可

図4:「フェイクチェッカー」のUI

URLの検証　文献4と同じ題材で検証:「偽」判定

図5:PayPayの偽サイト(左)
フェイクチェッカーの判定(右)

画像の検証　YouTube上の題材 *5で検証：「偽」判定

図6: YouTube上の題材画像*5(左)フェイクチェッカーの判定(右)

課題

・人のディープフェイク加工画像判定は、今回設定したプロンプトでは判定できなかった。
・JavaScript上にAPIキーを記載したため、APIキーの漏洩の可能性がある。➡ **サーバー経由で判定を実施**

検証2：NodeRedサーバー経由での判定

前橋高校内のLinuxパソコンをNodeRedサーバーとし、APIキーとプロンプトをサーバー上で管理した（図7）。

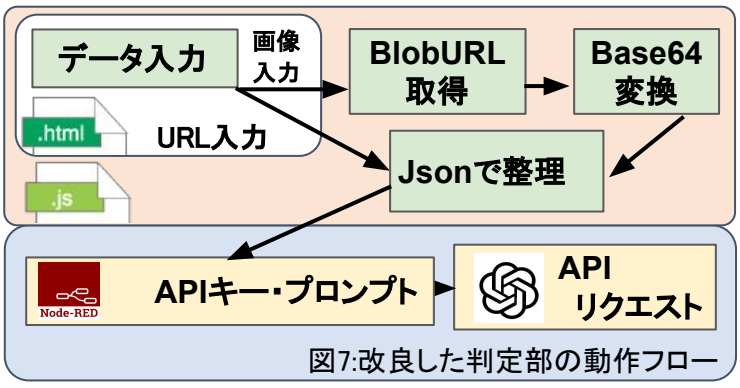

図7:改良した判定部の動作フロー

考察　検証1と同じく限定的だが画像判定できる。**課金が発生**し、複数アクセスが対応できない。

展望　自サーバー上にAIモデルを構築し、**無課金化**。顔を判定できるプロンプトが存在する *6ため、複数のプロンプトで切り替えられる仕様を検討。

参考文献

[1] 総務省「 ユーザのフェイクニュースに対する意識調査」　(2020年)
[2] トレンドマイクロ「ディープフェイクに関する国内実態調査　2024」(2024年)
[3] NTTセキュリティ・ジャパン「 Detecting Phishing Sites Using ChatGPT」(2023年)
[4] 総務省「デジタルテクノロジーの高度化とその活用に関する調査研究」　(2024年)
[5] https://www.youtube.com/shorts/vLRDcyeU5bo(2024年7月閲覧)
[6] Shan Jia(University at Buffalo, State University of New York)ら「Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics」(2024年)