法人情報の真偽を判別するプロダクト「PIRD」の開発

群馬県立前橋高等学校　2年　岡田 武

背景・仮説

インターネット上では偽情報が多く流布しており，偽情報による様々な被害が近年増加している

特に 法人に関する偽情報では，法人や顧客が甚大な被害を被っている

前回 ChatGPTのオープンモデルを使って開発 ➡ 今回 ファインチューニングさせたLLMを使って開発

仮説

LLMをそのまま使うのではなく**各法人に特化した内容**でファインチューニングすると**真偽判定精度**が向上するのではないか

フロントエンドの開発

判定部

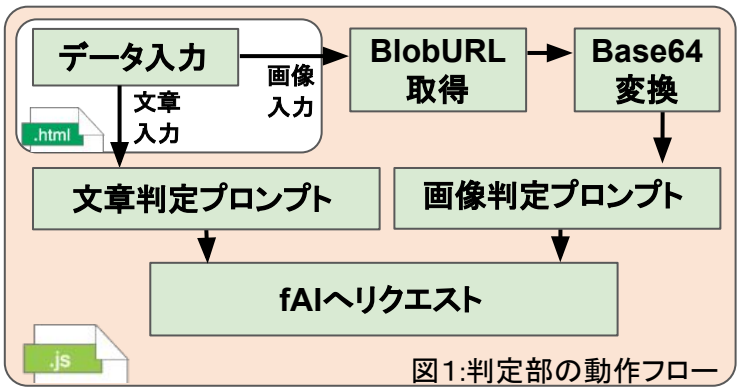

図1:判定部の動作フロー

表示部 総合評価はA～C表記とした（図2）。

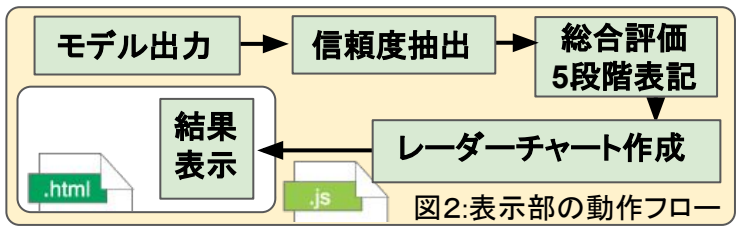

図2:表示部の動作フロー

LLMのファインチューニング

LLMのファインチューニングには，**QLoRA方式**を採用した。
モデルの重みの変化量を2つの小さな行列の積で近似し、学習すべきパラメータ数を大幅に削減した。
QLoRA方式により学習にかかるコストを最小限にしながらfAI (ファインチューニング済モデル)を開発した(図3)。

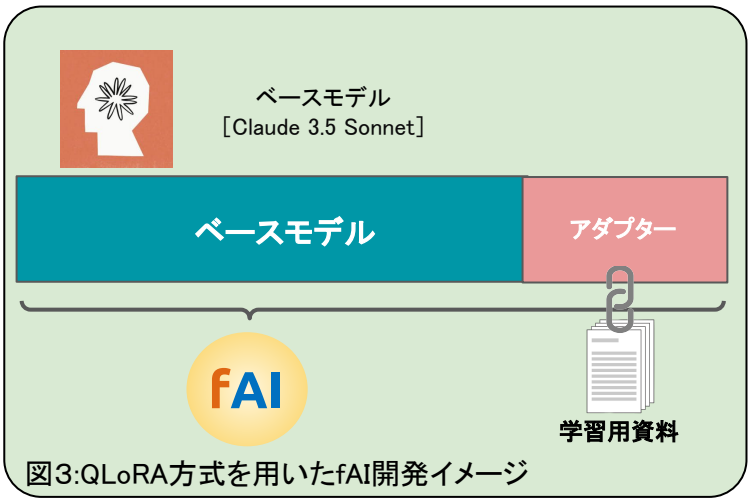

図3:QLoRA方式を用いたfAI開発イメージ

検証：判定精度

今回は 前橋高校を事例とし，前橋高校に関する情報のファクトチェックを行った。
学習ソースは前橋高校のホームページをスクレイピングしたもの使用した。

図4: PIRDのUI（前橋高校用）

文章の検証

ChatGPT：72.5%（58/80）

fAI　：88.8%（71/80）

検証時は ホームページの情報に加え，Wikipedia「群馬県立前橋高等学校」の情報（正しいことを確認済）も含め検証した。

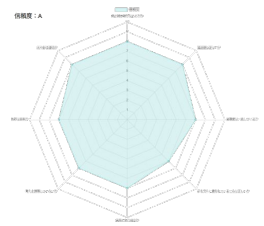

図5: 判定結果の例（信頼度A）

画像の検証

ChatGPT：70.0%（28/40）

fAI　：82.5%（33/40）

画像を言語化することにより，対象の真偽を測定している。
尚，検証用の画像は文章の検証と同様の文を使用し、Gemini（Nano Banana）を用いて生成した。

今後の展望：RAGの実装

RAG（検索拡張生成）をfAIに実装することにより、fAIの再学習サイクルを大幅に減らし，情報の入替えの効率化を図ることができる(図6)。
現在ローカル環境でのRAG機能の構築に成功しており，今後はサーバーでの構築を目指す。

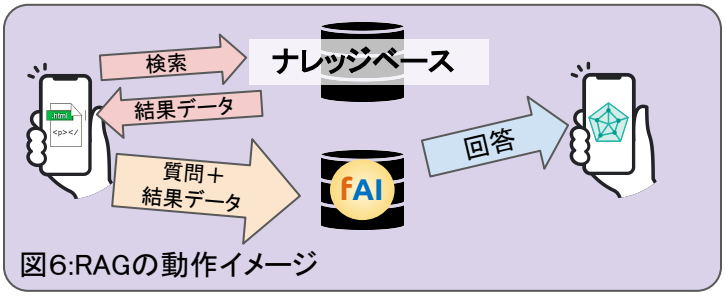

図6:RAGの動作イメージ

参考文献

「偽・誤情報の現状とこれから求められる対策」
https://www.soumu.go.jp/main_content/000867454.pdf
「LoRAとQLoRA」
https://www.redhat.com/ja/topics/ai/lora-vs-qlora
「RAG（検索拡張生成）とは？意味・定義 IT用語集」
https://www.ntt.com/bizon/glossary/e-r/rag.html