令和6年度 生成AI に起因するインターネット上の偽・誤情報等への対策技術に係る調査の請負 (実証事業)

AIを活用した情報コンテンツの真偽判別支援技術の開発・実証 成果報告書

2025年3月12日

日本電気株式会社

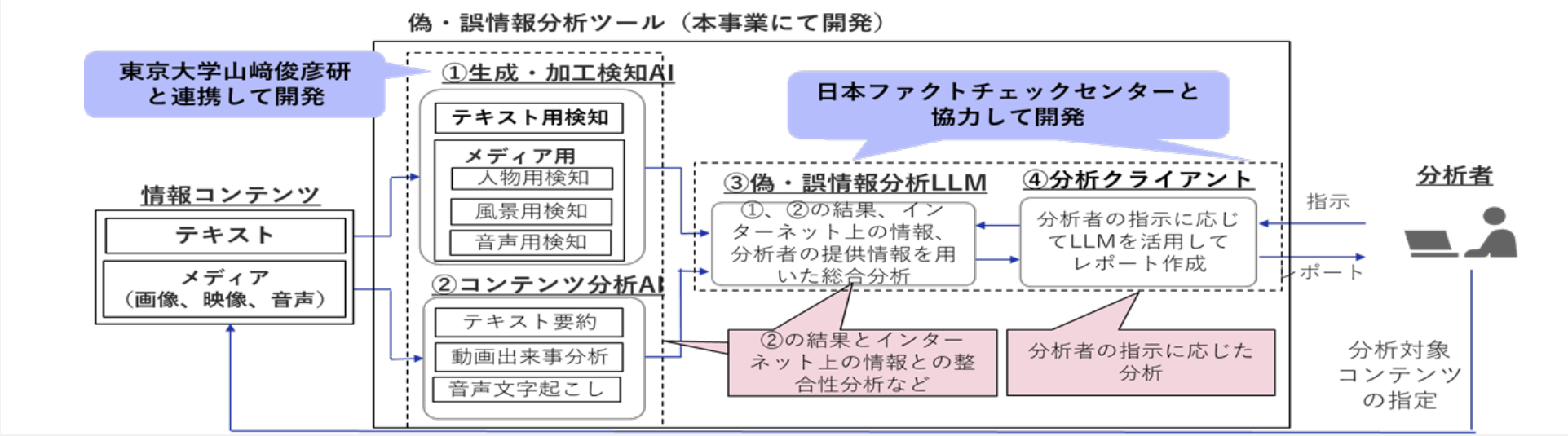
目次

目次

1-1 実証概要のサマリ

実証件名:AIを活用した情報コンテンツの真偽判別支援技術の開発・実証

実施体制 （下線：代表機関）	日本電気株式会社、東京大学山崎俊彦研究室、日本ファクトチェックセンター、放送局	**対象とする偽・誤情報**	➢ 対象とする偽・誤情報の種類を記載 – 情報の対象: 人、物体、風景 – 情報の形態: 画像、映像、音声、テキスト
対策技術	➢ 大規模言語モデル（LLM） ➢ 各種検知ツール、分析サービス	**目標**	➢ 目標：ツールによる工数削減率の定量評価 ➢ 中長期目標：真偽判別支援ツールの社会実装
実証概要	➢ 大規模言語モデル（LLM）などのAIを組み合わせた真偽分析支援ツールによって、専門家の分析工数の削減を目指す。 具体的には、偽・誤情報分析業務においてAIによって支援可能な作業を行うツールをファクトチェック機関の協力の下で開発し、その分析工数削減効果を評価する。 また、放送局の技術実証を実施することでユーザビリティ向上を図る。 ➢ 中長期的には、開発ツールについて、主な利用主体として期待されるファクトチェック機関やマスメディア（放送局等）への社会実装を目指す。		



1-2 事業の目的

事業の目的

＜社会的背景＞
生成AIによって生成されたテキスト・動画像を含んだ**偽・誤情報コンテンツがインターネット上に氾濫**し、国民が日常的に触れる機会が増加している。これにより、AI が生成した偽情報・誤情報が民主主義に不当に介入するなど、**社会を不安定化・混乱させるリスク**が高まっている。また、情報の真偽判定を行っている**ファクトチェック機関**や信頼性の高い情報を発信する義務のある**マスメディア各社(放送局等)**は偽・誤情報コンテンツの判別に多くの人手をかけており**業務負担が急激に増加**している。
＜コンセプト＞
本事業では、**複数種類のデータ（テキスト、画像、動画、音声）で構成されている情報コンテンツが偽・誤情報ではないかをAIで分析評価する支援ツールを開発**する。このツールは、情報コンテンツが生成・加工データでないかを評価した上で、その内容の真偽を分析する。その結果を大規模言語モデルに取り込むことで**ファクトチェック機関が作成する報告書や記事に近い形式での出力**を得ることを可能とする。開発したツールは、**偽・誤情報の判定を行う必要性の高い業種（マスメディア等）での活用も見据え技術検証を実施**する。

偽・誤情報への技術的対策の課題の状況、今後目指す姿

既存の偽・誤情報対策技術については、テキストと動画の真偽分析機能を有するツールや、生成加工の有無を特定のデータ（テキスト、人物、風景、音声）に対して行うAIなどがあるものの、情報コンテンツの分析を行う場合には、分析者が様々なツールを使い分けた上でその結果を統合する必要がある。**本事業においては分析者の負担を減らすため、多面的な評価、総合分析、分析者の意向に応じた分析条件の調整を一手に担うことができるツールを開発**する。

1-3 開発する技術の概要 (1/5)

開発技術

本事業では①生成・加工検知AI、②コンテンツ分析AI、③偽・誤情報分析LLM、④分析クライアントから構成される偽・誤情報分析ツールを開発する。分析では①ー④を以下のように用いる。

1.①生成・加工検知AIによってインプットされたコンテンツが生成AIによって生成、加工されているかを評価する
2.②コンテンツ分析AIによって、内容（テキストの要点、動画内で起こったイベント）を抽出する
3.2.において抽出した内容と、1.の評価結果を加味し、インターネット上の他情報との整合性を③の偽・誤情報分析LLMを用いて評価する
4.④の分析クライアントを用い1.や3.の評価結果をレポートとして出力し、分析者に提示する

なお、「納入成果物(報告書)」の作成過程で利用する外部API（OpenAI）やOSS、製作したエンジンなどの著作権は著作物を創作した者（著作者）や団体に帰属するものとする。

本事業全体における当該技術の役割

後述する①-④の技術を内包し、入力された情報コンテンツの真偽・生成加工の有無を評価し、レポートを生成する

1-3 開発する技術の概要 (2/5)

開発技術

①生成・加工検知AI
[機能概要]**ツールに入力される情報コンテンツが生成データなのか、加工されているかを評価する**AIを開発する。生成・加工の検知には、対象データが生成・加工された際の特徴を学習したAIを用いる手法が主流となっているため、本事業においては**対象とするデータの種類や内容に応じて複数種類の検知AIを備える**。具体的には、テキスト用、人物用、風景用、音声用の検知AIを開発する。人物と風景については、学習するべき特徴が大きく異なるため分ける。
[開発方法]個々の検知技術については**活用可能なOSSを選定して利用**することを想定する。**システム導入費用の低減**と、生成・加工技術の進展と検知技術の進展が競争的に行われ、検知技術の発展が望めるためである。**技術の状況に応じて検知AIを置き換えるといった将来の改良を想定し、一部のAIを交換可能とする入出力構造を持ったフレームワークとしての整備に注力する。**
ただし、検知技術の成熟度は対象とするデータによって異なり、テキスト向けの技術や風景向けの技術成熟度は比較的低いためにOSSを直接的に活用することが困難であるケースが想定される。このようなケースを想定して同機能を有するAIを簡易的に開発し、構築手段を報告書に含める。フレームワークの開発においては、技術開発主体であるNECの保有技術を利用することを想定する。
本機能については、人物の合成データの検知技術において多くの業績を保有する東京大学山﨑研究室にアドバイスを求めながら開発を行う。
なお、「納入成果物(報告書)」の作成過程で利用する外部API（OpenAI）やOSS、製作したエンジンなどの著作権は著作物を創作した者（著作者）や団体に帰属するものとする。

1-3 開発する技術の概要 (3/5)

開発技術

②コンテンツ分析AI
[機能概要]
本AIは、**ツールに入力される情報コンテンツの内容をLLMによって分析可能な形式（テキスト）に変換する**。テキスト自体はLLMによって分析可能だが、分析の効率化のため要点を抽出した上で真偽の分析を行う。画像や動画については、その中で映っている人物、物、出来事を抽出する。また、音声についてはその内容を文字起こししてテキスト化し、テキストを要約する。分析者によって信頼性が高い情報が入力された場合にも本機能を通じて要約し、後工程で活用する。
[開発方法]
テキスト要約、動画出来事分析、音声文字起こしのそれぞれについては生成AIのOSSやサービスを選定して利用することを想定する。主としてシステム導入費用の低減を目的としている。**本領域におけるOSSやサービスの進展は著しいため、一部の機能を別サービスに変更可能とすることを想定して開発**するが、本実証期間においてはOSSやサービスを限定し、期間内の変更は行わないものとする。主たる開発対象は、**様々なデータをテキストとして分析して後工程に渡すフレームワーク**であり、その他の費用は有償APIを活用する費用が主となる。フレームワークの開発においては、技術開発主体であるNECの保有技術を利用可能な構成とする。
　なお、「納入成果物(報告書)」の作成過程で利用する外部API（OpenAI）やOSS、製作したエンジンなどの著作権は著作物を創作した者（著作者）や団体に帰属するものとする。

1-3 開発する技術の概要 (4/5)

開発技術

③偽・誤情報分析LLM
[機能概要]
本機能では、①・②**から得られた情報をLLMに入力して、ツールに入力された情報コンテンツの信頼性を評価**する。生成・加工が検知された場合にはその旨を出力し、テキストや音声の要点とインターネットを検索して得られた情報が食い違っていないか、分析者から信頼性が高い情報として与えられた内容と整合しているか、テキストと動画の内容に矛盾がないかを評価したうえで、総合的な評価結果をテキストとして出力する。
[開発方法]
①と②から得られた情報を入力して、分析対象の情報コンテンツの信頼性を評価する仕組みを開発する。生成・加工が検知された場合には、それがコンテンツの主たる主張に関連するかを評価した上で総合判定の結果に加味する。真偽の分析については、情報コンテンツの要約がインターネット検索の結果や分析者から与えられた情報と整合しているかを評価し、簡易的なレポートとして出力する。**偽・誤情報分析のパターンは、日本ファクトチェックセンターのレポートや、インタビューを通じて獲得した専門知識をデータ化することでLLMに導入し、専門家の判断に近い形態の応答に近づける。**
LLMの調整については、独自日本語LLMの開発に付随する知見を活用する。また、**日本ファクトチェックセンターによるレビューに基づく改善を繰り返し、放送局との技術実証結果を必要に応じてフィードバックすることで応答の有用性を高める**計画である。
　なお、「納入成果物(報告書)」の作成過程で利用する外部API（OpenAI）やOSS、製作したエンジンなどの著作権は著作物を創作した者（著作者）や団体に帰属するものとする。

1-3 開発する技術の概要 (5/5)

開発技術

④分析クライアント
[機能概要]
本クライアントは**③の分析条件を分析者の指示に応じて調整するために用いる。③の分析に用いられた情報の一部を除去・修正する・追加するといった手続きによって、所望の分析となるように調整する手続きを実行可能**とする機能を担当する。
[開発方法]
機能①、②から得られた情報の全てが機能③の分析に適していない場合に取り除く、修正する、分析者が独自に保有する追加情報を加えるといった機能を有するクライアントを実装し、分析者とのやり取りの補助を担当する。本クライアントの開発においても、**日本ファクトチェックセンターのレビューを参考に、一部の分析のみ実行できるような設定機能や分析対象をリストで入力可能とするといったバッチ機能などに対応する。**なお、実装機能はレビューにおける要望、分析工数削減効果、開発容易性のバランスを見てNECにて判断し２機能を実装する。実装対象として選定外とした機能については、実装方法の検討を行った上で報告書にまとめる。
　なお、①の検知や②のコンテンツ分析について、それぞれの技術について参考とするOSS相当の性能を想定し、個々の性能向上開発は行わないものとする。本事業における③、④の開発は日本ファクトチェックセンターのレビューや放送局との技術実証の結果を参考に改善するが、あくまで当初の開発規模の範囲内で実施可能なものに限る。
　なお、「納入成果物(報告書)」の作成過程で利用する外部API（OpenAI）やOSS、製作したエンジンなどの著作権は著作物を創作した者（著作者）や団体に帰属するものとする。

1-4 社会実装のための実証の詳細 (1/2)

社会実装のゴール

本事業では、情報コンテンツを構成する主要なデータタイプであるテキスト、画像、動画、音声を対象に、生成・加工の有無と事実との整合性を総合的に評価するツールを開発する。これにより、多様な偽・誤情報の分析評価を一手に引き受ける枠組みを提供する。
このような枠組みが社会実装されれば、ファクトチェック機関の業務速度が向上し、タイムリーに偽・誤情報の分析とレポート化が可能になる。これにより、国民は対象情報の真偽を把握したうえで情報コンテンツに触れることができ、社会を不安定化・混乱させるリスクを大きく低減できる。また、日本のメディア等から発信される情報コンテンツの品質も保たれ、国民に正しい情報コンテンツが届く状態が維持できる。

社会実装に向け、必要な検討事項

<ファクトチェック機関>
①偽・誤情報分析LLM/分析クライアントにおいての必要機能明確化・開発方式検討(本事業で実証予定)
②継続的に運用できるような仕組みづくりの検討

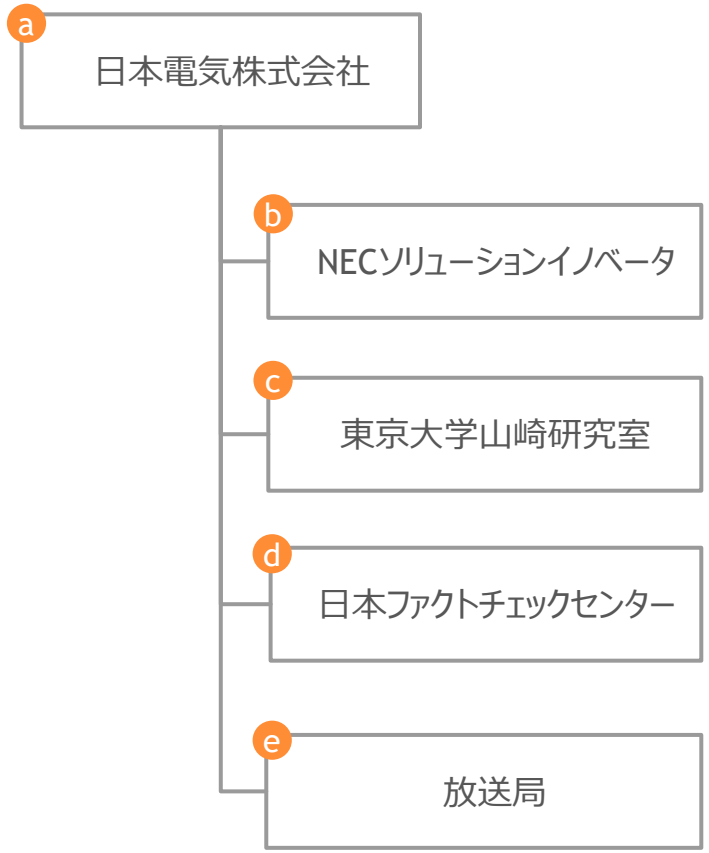
<放送局>
①要件・課題の洗い出し(本事業で実証予定)
②偽・誤情報分析LLM/分析クライアントにおいての必要機能明確化・開発方式検討
③継続的に運用できるような仕組みづくりの検討

1-4 社会実装のための実証の詳細 (2/2)

実証内容	実証パートナー	想定ゴール
本事業で開発したものがファクトチェック機関での ファクトチェック作業効率化につながることを確認する	日本ファクト チェックセンター	以下の観点で現状の作業よりも効率化されることを確認できる ・情報コンテンツの内容確認時間の削減 ・情報コンテンツの関連情報探索時間の削減 ・分析レポート作成時間の削減
利用主体が放送局とした場合の ファクトチェック作業効率化に向けた要件・課題を 確認する	放送局	放送局で必要とされる要件・課題が明確になる

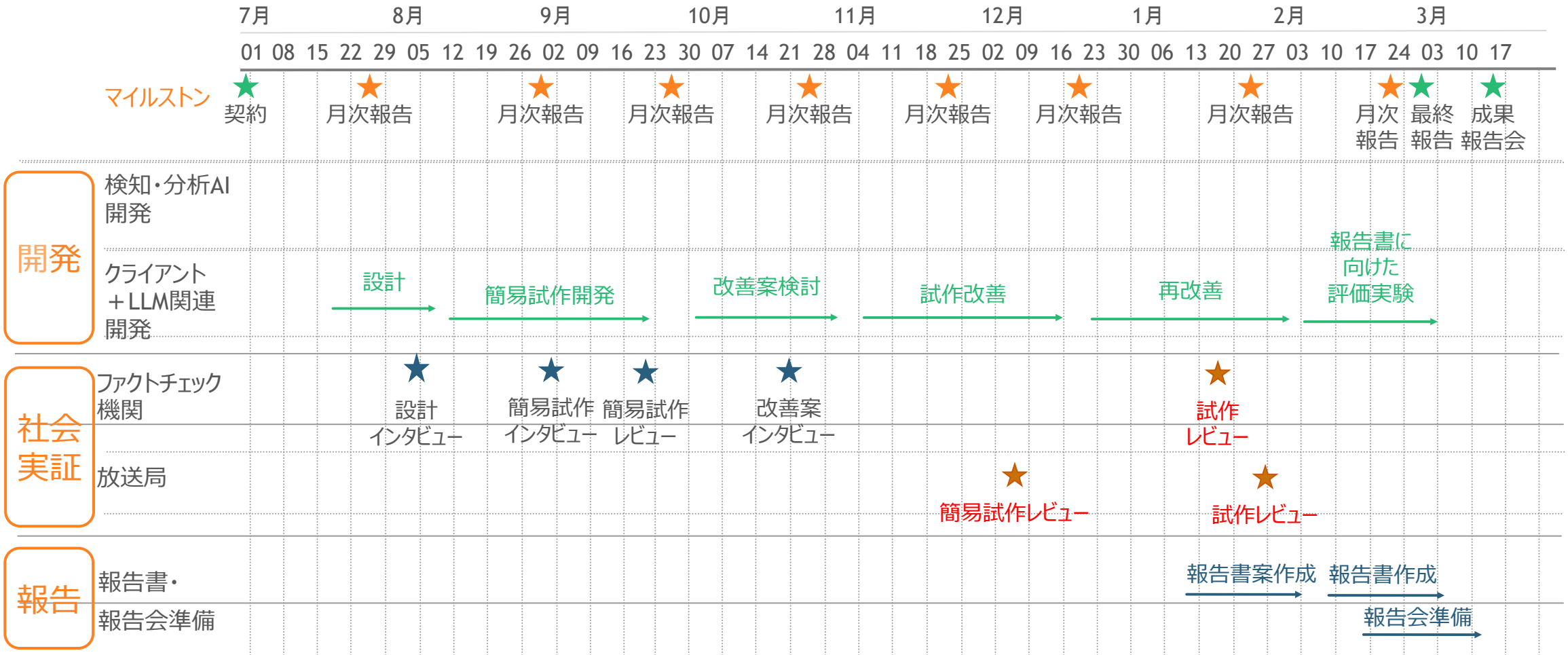
1-5 実証の実施体制

実施体制図



団体名	役割及び責任
a 日本電気 株式会社	実証事業代表者、実証事業の全体とりまとめ 事業計画の立案、報告書の作成をはじめとする事業全般の管理・統括業務 偽・誤情報分析ツールの開発
b NECソリューション イノベータ	ネットワーク・実証環境構築
c 東京大学 山崎研究室	生成加工検知AIの開発に関する技術アドバイザー
d 日本ファクトチェック センター	ファクトチェックに関するアドバイザー
e 放送局	実証事業を通して利用主体として課題・要望の情報提供

1-5 実施計画 (スケジュール)



※検知・分析AI：生成・加工検知AIおよびコンテンツ分析AI
※クライアント＋LLM：偽・誤情報分析LLMおよび分析クライアント

1-5 実証スケジュール (月次マイルストーン　1/2)

	計画（当該期間に実施予定の事項）
～7/22週	• 検知・分析AIに関する調査を行い、利用する技術候補を選定する • LLMサービスの調査を行い、利用する技術候補を選定する • システムの全体設計のドラフトを作成する
～8/26週 (夏季休暇想定で5w)	• 利用技術候補である検知・分析AIの技術調査を行い、簡易試作品の仕様を確定 • 利用技術候補であるLLMを評価、簡易試作で利用するLLMを選定 • クライアントの簡易試作の仕様を確定
～9/23週	• 検知・分析AI開発、クライアント+LLM関連の簡易試作の開発完了 • 簡易試作のファクトチェック機関によるレビュー実施
～10/21週	• 簡易試作の放送局によるレビュー実施 • 簡易試作レビュー結果に基づく改善案検討 • 改善案に関するファクトチェック機関インタビューの実施

1-5 実証スケジュール (月次マイルストーン　2/2)

	計画（当該週に実施予定の事項
～11/18週	• 検知・分析AI開発、クライアント+LLM関連の簡易試作の改善仕様を確定・開発開始
～12/16週	• 検知・分析AI開発、クライアント+LLM関連の改善版試作の開発完了 • ファクトチェック機関、放送局によるレビュー実施
～1/20週 (年末年始想定で5w)	• レビュー結果に基づくクライアントの再改善開発の着手 • 報告書案ドラフトの作成
～2/17週	• クライアント再改善開発の完了 • 報告書案の作成完了 • 報告書作成開始
最終報告	• 報告書の作成完了 • 報告会の実施

目次

2-1 開発した技術・ツールの詳細 (1/11)

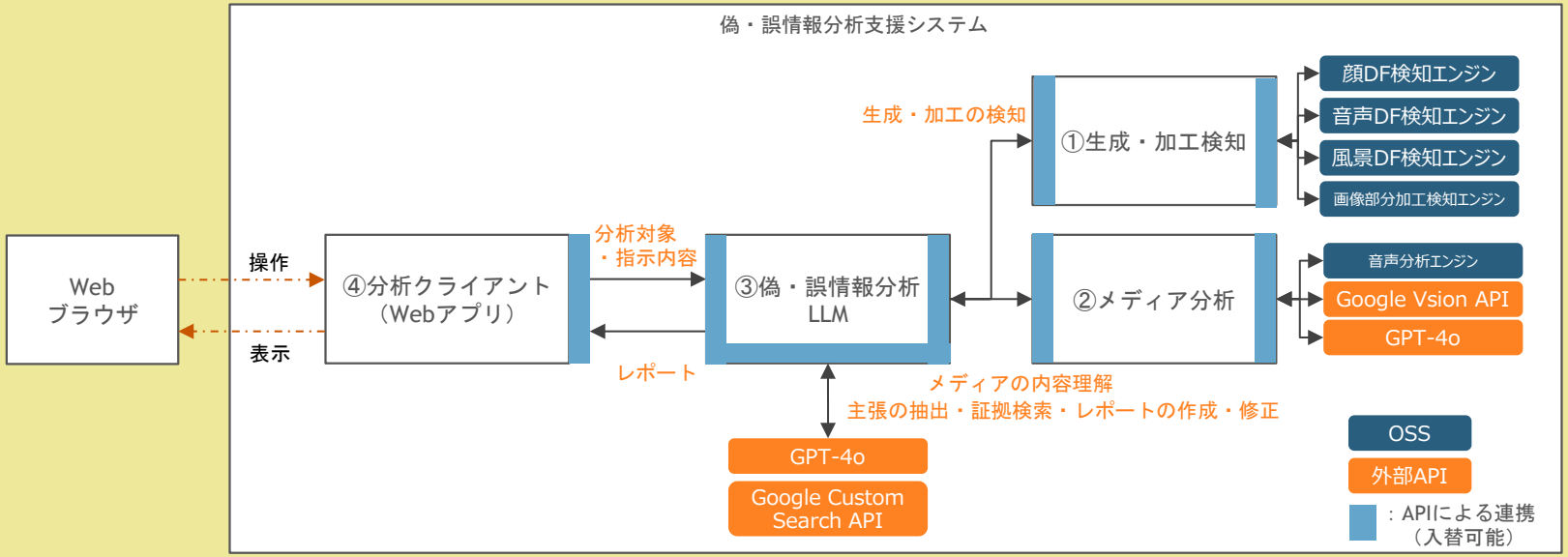
開発技術の概要

偽・誤情報分析支援システム

①生成・加工検知機能、②メディア分析機能、③偽・誤情報分析LLM機能、④分析クライアント機能
により構成される。それぞれの機能が連動し、指定された分析対象に対して以下のようにファクトチェックを行い結果をレポートとして発行する機能を提供する。

- AIによる生成・加工の有無のチェック
- メディアに含まれる内容を理解し、検証すべき主張を抽出
- 主張に関係する証拠情報の検索と主張の真偽分析
- ファクトチェックレポートの作成

各機能の詳細については別ページにて述べる。



2-1 開発した技術・ツールの詳細 (2/11)

R5補正における技術開発の成果

偽・誤情報分析支援システム

- 実現した要件
  - 偽の映像・画像や音声を見分けることができる。
  - 映像・画像や音声の内容をテキスト情報にまとめることができる。
  - テキスト・映像・画像・音声の組み合わせが**主張している内容を抽出**できる。
  - 検証する対象の内容の検証に必要な情報を検索できる。
  - 検索情報を組合せて情報内容の真偽について**分析者の判断を支援**できる。

- 想定する利用者
  - 偽・誤情報を分析し、判定する業務に従事している担当者
    例）ファクトチェック機関、報道機関など

- その他特徴・精度など
  - 日本ファクトチェックセンターが公開しているレポート61件に対して評価を行い、**レポートと同等の判定をした割合として85%を達成。**
  - 検証対象としてテキスト・音声・画像・動画およびそれらの組み合わせに対応。
  - 複数対象物から分析した結果をLLMを活用し、分析しまとめてフェイク検知を実施。
  - **使用するAPI/サービスの入れ替えが出来るように設計・構築されており、最新かつ最適なAPI/サービスを利用することが可能。**
    今回は社会実装時の採用ハードルを下げるため、OSSを中心に費用面を考慮した。

2-1 開発した技術・ツールの詳細 (3/11)

R5補正における技術開発の成果②

偽・誤情報分析支援システム

- ファクトチェックの現場からのフィードバックの反映
  - 日本ファクトチェックセンターに対してシステムのデモを３回行い、得られたフィードバックを機能に反映することで、現場で活用しやすいシステムを実現

フィードバックを反映し開発した機能

項目	機能の内容
主張の抽出と選択	様々な観点で検証を行うためにシステムで複数の主張を抽出し、それを分析者が選択できるようにした
証拠情報の表示項目	証拠情報の確認を効率的に行うために証拠情報URLおよび判断に強く関係した原文を引用表示するようにした
証拠情報のカテゴライズ	証拠情報の重要性を容易に確認するために、証拠情報が主張に対して肯定的か否定的かといった立場の観点や、1次情報・２次情報といった情報を分析し示すようにした
証拠検索先のホワイトリスト設定	証拠情報をある程度信頼できるソースや分野に関係するソースから収集するために、検索先をホワイトリストとして分析者が設定できるようにした
生成・加工検知AIの確信度の表示	生成・加工検知AIが検知を行った際の確信度を0-100%の間で表示するようにし、どの程度結果を信頼できるかの指標を示すようにした
総合判断の結果	当初は誤り/正解を断定的に分析する傾向があったが、証拠が少ない場合や真偽の判断に不十分な場合などに対し、断定的な判断結果を出さないようにした

2-1 開発した技術・ツールの詳細(4/11)

開発技術の詳細内容

①生成・加工検知機能：インプットされたコンテンツが生成AIによって生成、加工されているかを評価

機能: 静止画、動画、音声を入力として、それぞれが四種類の生成・加工物であるか判定した結果と生成・加工物である確率を出力する

- 顔ディープフェイク検知器: 動画・静止画に含まれる顔がディープフェイクであるか判定
- 音声ディープフェイク検知器: 音声がディープフェイクであるか判定
- 画像部分加工検知器: 静止画が部分加工されているか判定
- 風景画像ディープフェイク検知器: 静止画が生成AIにより生成されたものか判定

顔ディープフェイク

Face Swap: 顔の入れ替え
(図は[1]より引用)

Reenactment: 表情の操作
(図は [1]より引用)

音声ディープフェイク

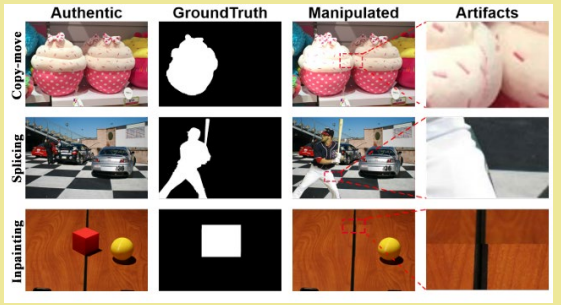
“文章”⇒“音声”

Text to speech: テキストを音声化

“音声”⇒“音声”

Speech to speech: 他者の声に変換

部分加工



Copy-move: オブジェクトを複製
Inpainting: オブジェクトの除去
Splicing: 他画像の貼り付け
(図は [2]より引用)

風景画像ディープフェイク

GAN modelによる生成
(図は[3]より引用)

Diffusion modelによる生成
(図は [4]にて生成)

[1]: Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, Applied intelligence, 2023
[2]: IML-ViT: Benchmarking Image Manipulation Localization by Vision Transformer, AAAI, 2024
[3]: Large Scale GAN Training for High Fidelity Natural Image Synthesis, ICLR, 2019
[4]: https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

2-1 開発した技術・ツールの詳細(5/11)

R5補正における技術開発の成果

①生成・加工検知機能：静止画像、動画、音声に対するあらゆる加工・生成を多面的かつ包括的に評価するツールを開発

本ツールにより実現できること:

- 各メディアに対する加工や生成の有無を明示し、情報コンテンツが偽・誤情報でないか評価するために証拠となる情報を抽出
  - 抽出された証拠情報は、他の機能により抽出された証拠情報と統合され、総合的に偽・誤情報の評価に活用

検知技術の選出:

- 商用利用可能なライセンスでソースコードおよびモデルパラメータが公開された検知技術の中から、論文の報告値や実測値を参考に、性能が優れているものを選定
  - 候補の選出から評価、最終的な選定に至るまで、東京大学大学院情報理工学研究科 山崎研究室の知見を効果的に反映

各検知器技術特徴:

- 顔ディープフェイク: 顔のアイデンティティではなく偽造の痕跡に注目して学習し未知のフェイクに対する検知性能を向上 [1,2] (ライセンス: Apache-2.0)
- 音声フェイク: 膨大な音声で学習された音声基盤モデルをフェイク音声で追加学習し未知のフェイクに対する検知性能を向上[3,4] (ライセンス: MIT)
- 画像部分加工: 大域的特徴と局所的特徴を合わせて学習することで、様々な大きさの改ざんに対応 [5,6] (ライセンス: MIT)
- 風景画像ディープフェイク: 膨大な画像で学習した基盤モデルを用いた特徴量抽出により多様な方法で生成される画像を検出 [7,8] (ライセンス: MIT)

各検知技術精度:

- 評価指標
  - 検知率: 加工・生成コンテンツを正しく加工・生成コンテンツとして判定した割合
  - 誤検知率: 非加工・非生成コンテンツを誤って加工・生成コンテンツとして判定した割合
  - AUC: 加工・生成コンテンツと非加工・非生成コンテンツとの識別能力の度合(検知率と誤検知率の関係を表すROC曲線の下の面積、高いほど良い)
- 評価結果:
  - 顔ディープフェイク: 顔フェイクベンチマークデータ[9]においてAUC 99.78%, [10]においてAUC 93.08% (論文[1]報告値)
  - 音声フェイク: 音声フェイクベンチマークデータ ASVspoof2021 LA[11]において誤検知率0.82%で検知率99.18% (論文[3]報告値)
  - 画像部分加工: 部分加工ベンチマークデータ [12]において誤検知率7.7%で検知率20.1% (実測値)
  - 風景画像ディープフェイク: 生成AIサービス・ツール[13,14,15]で生成した災害画像と実災害画像[16] において誤検知率3.6%で検知率21.2%(実測値)

[1]: Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization, CVPR, 2023
[2]: https://github.com/megvii-research/CADDM?tab=readme-ov-file
[3]: Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation, Odyssey Workshop, 2022
[4]: https://github.com/TakHemlata/SSL_Anti-spoofing
[5]: PSCC-Net: Progressive Spatio-Channel Correlation Network for Image Manipulation Detection and Localization, IEEE Transactions on Circuits and Systems for Video Technology, 2022
[6]: https://github.com/proteus1991/PSCC-Net
[7]: Towards Universal Fake Image Detectors that Generalize Across Generative Models, CVPR,2023
[8]: https://github.com/WisconsinAIVision/UniversalFakeDetect
[9]: Faceforensics++: Learning to detect manipulated facial images, ICCV, 2019
[10]: Celeb-df: A large-scale challenging dataset for deep fake forensics, CVPR, 2020
[11]: https://www.asvspoof.org/index2021.html
[12]: https://staff.utia.cas.cz/novozada/db/
[13]: https://www.bing.com/images/create
[14]: https://huggingface.co/alkzar90/ppaine-landscape
[15]: https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
[16]: https://github.com/ckyrkou/AIDER/tree/master

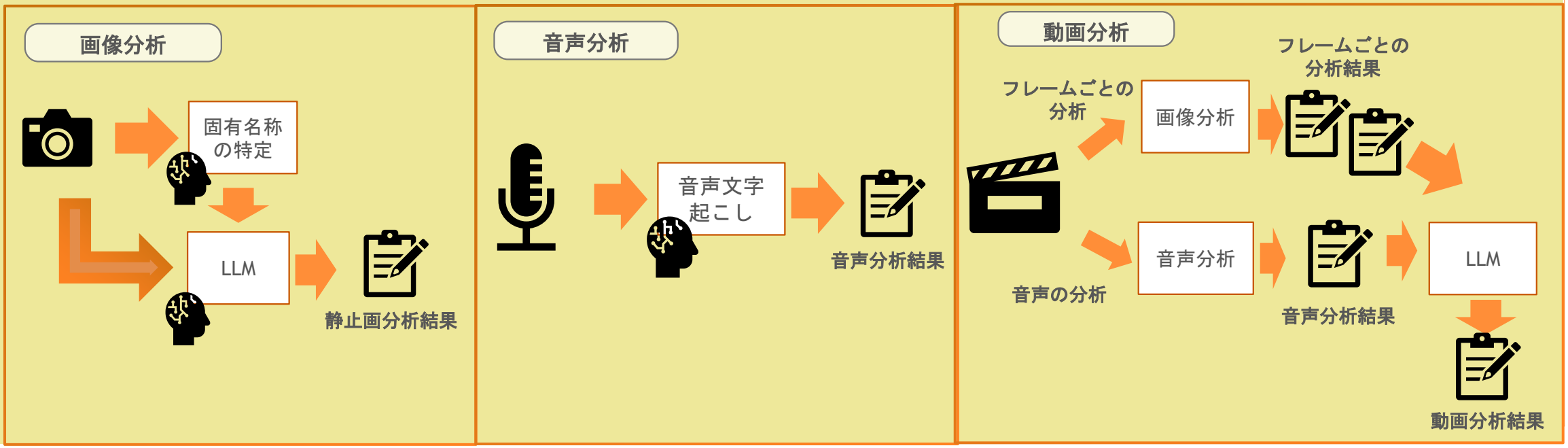
2-1 開発した技術・ツールの詳細 (6/11)

開発技術の詳細内容

②メディア分析機能

ファクトチェック対象として入力されたマルチメディアコンテンツの内容を分析し、テキスト情報に変換する機能
画像・音声・動画の内容をそれぞれに対応した分析機能により分析を行い、それぞれ内容を説明するテキスト情報に変換する。

- 画像分析：画像の内容を理解した分析結果をテキストで出力する
- 音声分析：音声を文字起こしした結果をテキストで出力する
- 動画分析：動画の内容を理解した分析結果をテキストで出力する



2-1 開発した技術・ツールの詳細 (7/11)

開発技術の詳細内容

②メディア分析機能

- メディア分析機能は以下の分析を実現し、結果をテキストで出力する
  - 画像分析
    - 画像内に写っている人物や物、建物などの固有名称を特定できる
    - 固有名称を考慮した画像全体の内容を分析できる
  - 音声分析
    - 音声ファイル内の発話内容を分析できる
  - 動画分析
    - 動画からフレーム画像を切り出して上記画像分析機能で分析できる
    - 動画から音声を切り出して上記音声分析機能で分析できる
    - フレームごとの分析結果および音声分析結果をまとめて動画内容として分析できる

- その他特徴・精度など
  - 画像分析に利用したツール
    - 固有名称の特定: Google Vision API
    - 画像全体の内容分析: Open AI GPT-4o
  - 音声分析に利用したツール
    - 発話内容分析: Whisper (OSS)

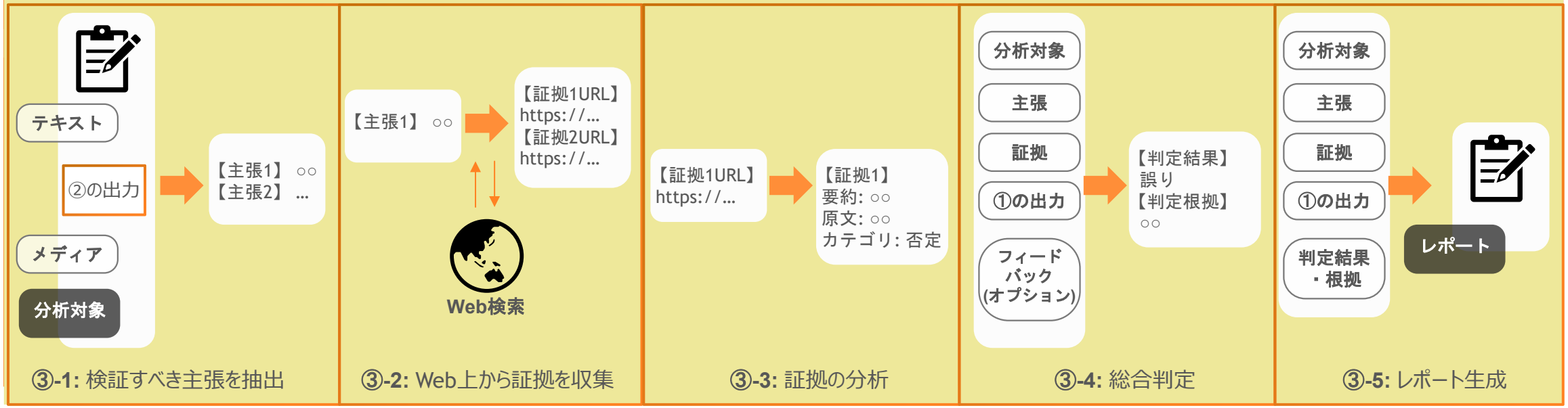
2-1 開発した技術・ツールの詳細 (8/11)

開発技術の詳細内容

③偽・誤り情報分析LLM機能

分析対象に対して①生成加工検知機能と②メディア分析機能と連携し以下のファクトチェックとレポート生成を行う。
本機能は以下のプロセスで行われる。

- ③-1: 検証すべき主張を抽出
- ③-2: Web上から証拠を収集
- ③-3: 証拠の分析
- ③-4: 総合判定
- ③-5: レポート生成



2-1 開発した技術・ツールの詳細 (9/11)

開発技術の詳細内容

③偽・誤り情報分析LLM機能

- 偽・誤り情報分析LLM機能は以下のサブ機能を持ち、ファクトチェックとその判断過程をレポートとして出力することができる
  - 検証すべき主張を抽出
    - 検証対象がどのような内容を主張しているのかを抽出することができる
  - Web上から証拠を収集
    - 主張に対して適切なキーワードを生成し、信頼のおける検索先リストから証拠情報を収集することができる
  - 証拠の分析
    - 収集されたwebページの内容を解析し、内容の理解やその内容のカテゴライズ(肯定・否定, 1次情報・2次情報等)をすることができる
  - 総合判定
    - 収集された証拠情報や①生成加工検知機能の結果を活用し、主張に対してのファクトチェックと根拠の生成ができる
  - レポート生成
    - これまでの情報をユーザが分かりやすい形で記述することができる

- その他特徴・精度など
  - 使用LLM: GPT-4o
  - 検索ツール: Google Custom Search JSON API
  - 精度:日本ファクトチェックセンターが公開しているレポート61件に対して評価を行い、レポートと同等の判定をした割合として85%を達成

2-1 開発した技術・ツールの詳細 (10/11)

開発技術の詳細内容

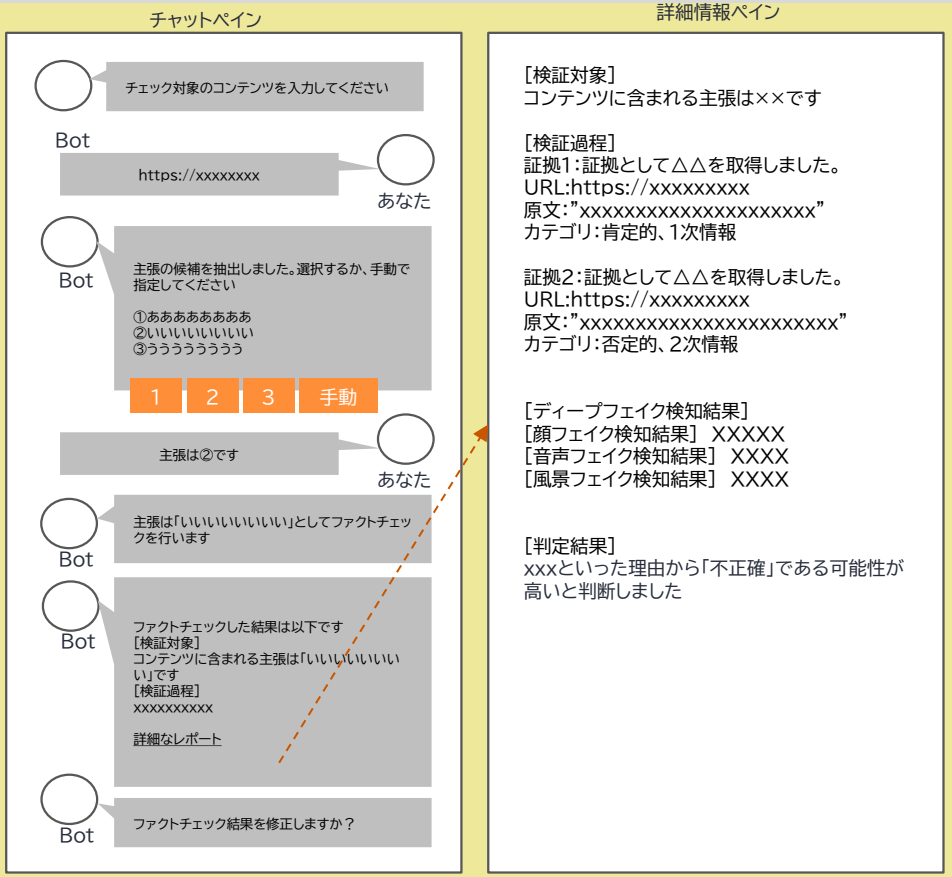
④分析クライアント機能

ユーザによる検証対象入力や、ファクトチェック結果取得を行うためのクライアント
一般的なブラウザでアクセスし、分析者がファクトチェックを行うためのUIを提供する。
UIとして提供する機能は以下である。

- ファクトチェック対象の入力
- ファクトチェックの実行
- ファクトチェックする主張の選択
- メディア分析結果の表示
- ファクトチェック結果（レポート）の表示
- ファクトチェック結果へのフィードバック（証拠の追加、修正指示）

これらのUIはchat bot形式で提供され、インタラクティブにファクトチェックを
実施可能である。

分析クライアント画面イメージ



2-1 開発した技術・ツールの詳細 (11/11)

開発技術の詳細内容

④分析クライアント機能

- 分析クライアント機能は以下を実現する
  - ユーザからの検証対象の入力を受け付けることができる
  - システムが出力したファクトチェックを行う主張一覧からユーザが任意のものを選択することができる
  - メディア分析結果を表示し、ユーザが独自にファクトチェックを行う主張を決定したい場合の判断要素として利用できる
  - ファクトチェックの結果をサマリとして確認することができ、必要に応じて詳細な内容を閲覧することができる
  - 必要に応じて追加の指示や証拠情報の追加を行うことができる
  - 生成されたファクトチェックのレポートをダウンロードすることができる
  - 信頼のおける検索先リストや検索サイト数、使用するLLMをユーザが設定することができる

- 利用者・その他特徴
  - chatbot形式でインタラクティブにファクトチェックが可能

2-2 社会実装のための実証の結果 (1/3)

R5補正における社会実装の成果

本事業で開発したものがファクトチェック機関でのファクトチェック作業効率化につながることへの確認

- 日本ファクトチェックセンターの作業従事者に偽・誤情報分析支援システムを触ってもらい、概ね業務時間の短縮には寄与できるとのコメントを得られた。
  - 現在、証拠出展情報の収集は手作業で行っており、対象物によるが**２時間程度要する。**
    当該システムを使用すれば５分から１０分になり大幅な改善が見込める。
  - **分析レポート作成時間の削減については、効果はあると考える。**ただし、具体的に何分という数字は示しづらい。
    生成されたレポートをそのまま使うのは厳しいが、文章の組み立て方の参考になる。
    また、動画や画像検証でそれぞれのモデルごとの判定が何%であるかの数字は証拠情報として引用できるようにすると時間削減につながる。
    ※試作レビュー後に対応済み
  - 対象記事の主張の選択はあったほうが良い。１ツイートに複数の主張があるため、選ばせるのは良い。
    複数の検証を同時に行うことはできていないが、個別に検証できるので問題ない。

2-2 社会実装のための実証の結果 (2/3)

R5補正における社会実装の成果

利用主体が放送局とした場合のファクトチェック作業効率化に向けた要件・課題の確認

下記業務フロー図をもとに東京３局、大阪１局の偽・誤情報分析を行う部署にヒアリングを実施。

＜業務フロー（想定）＞

2-2 社会実装のための実証の結果 (3/3)

課題及び今後の深化の観点

社会実装に向けた課題

（ファクトチェック機関向け）

- 出展情報の取得方法のさらなる工夫
  - 他社記事を引用した２次、３次情報から１次情報に辿りつき、証拠出展情報の優先度をつけて、分析精度を高めたい。
  - 有料サイトや論文、社内情報を加味した真偽分析ができるか。

（放送局）

- 速報用途での偽・誤情報の分析ニーズへの対応
  - 災害報道への対応
  - 位置情報に関する他マップツールなどとの風景照合や過去画像の流用確認による偽・誤情報検知ができるか。

（その他企業）

- 判定を目的とした用途への対応
  - ファクトチェック機関や放送局と異なり、分析よりも判定に重きをおいている。
  - 一方ファクトチェック機関からは、偽・誤情報に対してのリテラシーの低いまま、本システムを使用することで真偽結果を誤ってしまう懸念もあることから対象者へのレクチャーや、よりわかりやすいUIや判断理由の説明ロジックが必要となる。

2-3 本実証後の展望

2025年度	2026年度	2027年度以降
➢ 一部ユーザーでの実用化 ➢ プラットフォーム検討/整備/構築 ➢ 適用ユーザ拡大検討	➢ 実用化ユーザ拡大 ➢ プラットフォームの高度化	➢ ビジネスとしての自走化 ➢ 更なる適用ユーザ拡大検討
優先度が高い領域で専門ユーザーが業務利用 • 一部ユーザの利用が進む プラットフォーム構築 • 総務省24年度実証の他案件成果物（NEC以外）等との組み合わせによる総合的な偽・誤情報分析・検知プラットフォームの構築 適用ユーザーの拡大に向けた実証 • 適応ユーザー拡大のための必要機能整理・設計	適用範囲が拡大、専門ユーザーの業務利用の範囲が拡大 • 感度の高い / 親和性の高い利用主体(メディア・自治体等)での活用が広がる プラットフォームの高度化 • 広い事業者等で専門ユーザーが業務利用を行えるプラットフォームへ成長	継続的な技術開発・投資に向けてビジネスモデルが確立 • 先行事例の確立に伴い、幅広いユーザでの活用が進み専門ユーザーの業務利用が拡大 更なる適用ユーザ拡大 • 非専門ユーザへの適応拡大検討・実証