MAFT-AH: ヒューマンインザループを備えたマルチエージェント自動ファクトチェックフレームワーク

松永 悠斗 [1]　大磯 秀幸 [1]　柿崎 和也 [1] 宮川 大輝 [1] 古川 諒 [1]
塩原 楓 [2] 山崎 俊彦 [2]
[1] 日本電気株式会社　[2] 東京大学
[1]{yuto-matsunaga, oiso-hideyuki, kazuya1210, miyagawataik, rfurukawa}@nec.com
[2]{shiohara, yamasaki}@cvm.t.u-tokyo.ac.jp

概要

実世界の主張に対する Automated Fact-Checking（AFC）は、入力の曖昧性や証拠の不足、さらに近年増加するマルチモーダルな偽・誤情報による複雑化から、依然として困難な課題である。特に、既存の End-to-End 型マルチモーダル AFC 手法は、実運用において (1) ユーザの入力内の情報欠落や曖昧性の補間、(2) 複雑な調査フローを柔軟に設計・運用するオーケストレーションの二点を十分に扱うことができない。本研究では、ヒューマンインザループによる曖昧性の軽減とマルチエージェントによる複雑なオーケストレーションによりこれらの課題に対処した、ヒューマンインザループを備えたマルチエージェント AFC フレームワークの MAFT-AH を提案する。

1　はじめに

実世界の主張に対する Automated Fact-Checking (AFC) は、人工的な主張に対しての AFC と比べて主張の曖昧性や証拠の不十分さなどから困難であることが知られている [1]。更に、近年のインターネット上の偽・誤情報はテキストに留まらず、画像・動画・音声を含むマルチモーダルなコンテンツとして流通する。このような実世界のマルチモーダル偽・誤情報に対する AFC は、単一モーダルのものよりも主張が複雑化し曖昧性が大きくなり、各モーダルに対しても利用可能な調査プロセスが異なる点で難解である。

マルチモーダルなコンテンツを含む偽・誤情報に対する AFC 手法として、MAFT が知られている。MAFT は、テキスト・画像・動画・音声を含む任意の組合せ入力に対し、各モーダルをテキスト化して統一表現へ変換することで、大規模言語モデル (LLM) を利用したマルチモーダル AFC を行う手法である [2]。しかし、MAFT を含む End-to-End 型の枠組み [3] は、実運用の調査フローに内在する二つの課題を十分に扱えていない。第一に、ユーザの入力は曖昧であったり情報を全て提供しないことがあり、検証可能な主張へ落とし込むための曖昧性解消や追加情報取得が必要となる点である。第二に、マルチモーダルな入力などにより主張が複雑化することで調査のフローが複雑化し、使用する分析ツールも多岐にわたるため、フローの柔軟性やその実行に関するオーケストレーションがボトルネックとなるという点である。

我々は、ヒューマンインザループ (HITL) を備え、マルチモーダル入力に対応したマルチエージェント AFC フレームワークの MAFT-AH を提案する。本手法は MAFT を基に、(1) これまでの AFC 手法では困難であった入力時の情報欠落や曖昧性に対して、LLM により入力からこれらが発生しているかを検知し HITL で補間する、(2) 複雑化する調査フローやオーケストレーションに対しても入力に応じた調査フロー生成とマルチエージェントによる調査によりこれを可能にする。本研究の主な貢献を以下に示す。

- 入力時の情報欠落・曖昧性を LLM で検知し、HITL による補間を組み込んだ Agentic AFC フレームワーク
- 調査フロー生成とマルチエージェントによる柔軟な証拠収集機構
- 実世界のファクトチェックデータセット（AVeriTeC）で既存手法を 5.8%上回る Accuracy を達成（0.784）

2 関連研究

2.1 Multimodal Fact-Checking

マルチモーダルな AFC 手法として CCN が知られている。CCN は画像とキャプションの組みからなる入力に対して Web 上から関連情報を収集し、同一および異なるモダリティ間の整合性を判定する [3]。MAFT も含めこれらは主にマルチモーダルな情報への対応と実世界情報の証拠利用へ焦点を当てる一方、曖昧性やフローの柔軟性に課題が残る。我々の研究では HITL で曖昧性を軽減し、入力に応じたフローの生成で柔軟性を高める。

2.2 Claim Refinement for Fact-Checking

AFC における主張抽出において、入力のノイズや文脈依存などを減らすことで曖昧性の低い主張抽出を行い、ファクトチェックの精度を向上させる手法として CACN や AVeriTeC-DCE が知られている。CACN は複雑でノイズの多いソーシャルメディア上の投稿をより単純で理解しやすい形式に正規化することで曖昧性を軽減している [4]。AVeriTeC-DCE は文書全体から検証すべき主張を抽出し、これを文脈なしでも理解できるように整形することで曖昧性を軽減している [5]。これらの手法は、入力の曖昧性を軽減し検証可能な形にする重要性を示すがオフライン変換として扱うため、入力の欠落や曖昧性が本質的に残る場面に対応できない。我々の研究では HITL によりユーザが既知の情報に関しては取得可能であり、ユーザも不明な点はそれも含めてファクトチェックの対象として調査が可能になる。

2.3 Clarifying Questions

ユーザ入力が曖昧な場合、適切な「明確化質問（clarifying questions）」を行うことで、意図や不足情報を補完し、後段の検索・推論の品質を上げる手法として STYLE や CLAMBER が知られている。STYLE は、LLM ベース会話エージェントがいつ・どのように明確化質問をするかを戦略として学習し、未知ドメインへの転移性を改善することを狙う [6]。CLAMBER は、曖昧性のタイプ分類に基づき、LLM が曖昧性を識別・質問生成できるかを評価する [7]。これらの手法は主に一般的な QA タスクにおける曖昧性解消を対象とし、ファクトチェックまでは踏み込まない。我々の研究ではファクトチェックにおける曖昧性解消のために、QA タスクで扱わないメディアコンテンツのフェイク検知や入力そのものの矛盾検知など、入力されてきた情報に対しての分析も考慮して明確化質問を行う。

2.4 Agentic AFC

近年は LLM のツール利用能力を背景に、AFC を単一モデルの判定問題としてではなく、LLM エージェントにより検索・証拠抽出・推論・要約といった複数工程からなるワークフローをオーケストレーションする手法が提案されている。ClaimCheck は、テキストで入力された主張を、エージェントにより検証可能なサブタスクに分割するようなプランニングをすることで実世界での検索であっても十分な証拠情報を収集できる [8]。他にも、ファクトチェックにおける各工程をそれぞれ専門エージェントに任せ、全体をオーケストレータが統括する multi-agent 型の AFC も提案されている [9]。これらの手法は、エージェントにより各工程を分解し検証フローを柔軟にすることで性能や説明可能性を向上させるが、主張が既に確定している前提であったり、入力の欠落や曖昧性が残る状況には対応できない。我々の研究では入力に対して主張の抽出から行う End-to-End 型の手法であり、HITL で入力の欠落や曖昧性を軽減する。

3 提案手法

本手法は、図 1 に示すように 4 つの工程から構成される。

3.1 Process1: Multimodal Textualization

本プロセスでは MAFT に基づき、マルチモーダルな入力コンテンツを理解するために入力の中から、動画、画像、音声の入力に対してはテキスト化を行い、マルチモーダルな情報をテキスト空間に落とし込むことで、どのような組み合わせのマルチモーダルな入力に対しても対応を可能にしている [2]。本手法は、画像のテキスト化に GPT-4o、音声のテキスト化に Whisper、動画のテキスト化は等間隔でサンプリングされた 8 枚のフレーム画像を GPT-4o でテキスト化し、このテキスト化結果と Whisper でテキスト化された音声内容と合わせて GPT-4.1-mini で動画としての内容要約を行う。

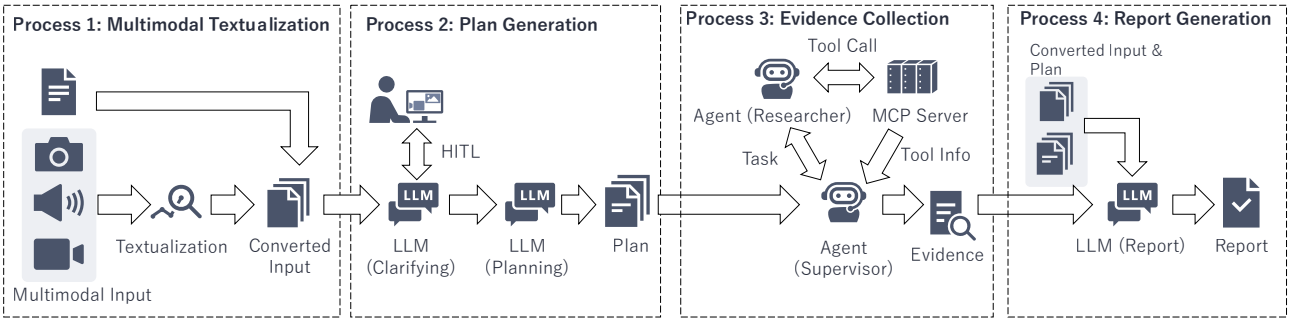

図 1　提案手法

3.2　Process2: Plan Generation

本プロセスでは、テキスト化されたマルチモーダル情報を元に検証のプランを生成する。ユーザからの入力に対して、既に十分な情報があるか、 不足情報が何か、 検証対象が明確であるかを Clarifying LLM（本研究では GPT-4.1 を使用）に判定させこれらが必要な場合にのみ HITL でユーザから追加の情報入力を求めることで入力情報と検証対象を明確化する。

Clarifying LLM により問題が無いと判断された後、Planning LLM（本研究では GPT-4.1 を使用）を用いて検証すべき対象（主張）と調査方針を生成する。調査方針は以降の調査での実行可能性を重視し、検証に必要な調査タスク一覧となるように生成される。

3.3　Process3: Evidence Collection

本プロセスでは、Web 上の関連情報収集やメディアコンテンツに対する DeepFake や加工の検知器、逆画像検索による初出調査といった分析ツールを必要に応じて組み合わせて証拠収集を行う。本プロセスは Supervisor Agent と Researcher Agent、MCP サーバから構成される。

Supervisor Agent（本研究では GPT-4.1 を使用）では、調査方針内の調査タスクをより具体的な調査項目に落とし込むために、調査方針の内容と使用可能なツールの一覧から使用するツール名と使用方法のペアを生成することで、調査項目の生成を行う。この時調査に使用可能なツール一覧は MCP サーバから提供される。その後、調査項目内の順番に従い Researcher Agent を呼び出し、Supervisor Agent はその結果とこれまでの調査結果、調査方針を入力とし、調査方針の内容を満たしたかの判断を行う。調査終了後、Supervisor Agent は各証拠の重要度に応じた粒度で要約することで情報の取捨選択を行い、無関係な情報や信頼性の低い情報の影響を軽減し冗長な情報を圧縮する。

Researcher Agent（本研究では GPT-4.1-mini を使用）は、ツール名と使用方法、各種ツールの仕様を入力とし、ツールに適した引数の生成を行う。その後ツールを呼び出し、ツールからエラーが返ってきた場合は引数の再生成と再呼び出しを行う。

MCP サーバでは調査で使用する様々な分析ツールを持ち、各種ツールの仕様を Supervisor/Researcher Agent に提供することでエージェント内にツールを持つよりも高い拡張性を実現している。本研究で MCP サーバは Web 検索、逆画像検索、Deepfake や加工検知などの検知器といった分析ツール群を管理する。

3.4　Process4: Report Generation

本プロセスでは、これまでのプロセスの結果を入力として Report LLM（本研究では GPT-5.1 を使用）を用いて偽・誤情報かの判定とその理由をレポート形式で生成する。判定はファクトチェックの専門家の分類に倣い、True / Almost True / Baseless / Inaccurate / False の 5 段階で行う [10]。本研究では偽・誤情報である理由の基準として、外部情報との不整合、フェイクの痕跡、モーダル間の不整合を挙げている。

本手法では判定を行うものの、実運用では最終判定を人手で行うことを想定しているため、判定理由がユーザに解釈しやすい形式で提示される必要がある。そのため、本手法ではレポート形式でユーザに提示する。このレポートは判定ロジックが明確になるよう調査過程と各過程での結果が章立てでまとめられ、ユーザが追検証できるように使用した証拠の一覧が別途提示される。

4 検証

4.1 設定

本検証では Plan Generation の効果を分離して評価するため AVeriTeC（テキスト入力）を用い、マルチモーダルツール群は今後の課題とした。AVeriTeC は Web 上の証拠を用いた主張検証を目的としたデータセットである [1]。提案手法の他に比較手法として、ClaimCheck [8] を用い、ベースラインとして LLM 単体 (GPT-5.1) による判定を行う場合を用いた。入力がテキストのみの場合、ClaimCheck は提案手法の Planning LLM 以降の処理と類似しているため、Clarifying LLM による不足情報や曖昧性の補間の効果とそれによる調査フロー生成の明確化の効果を評価することができる。

本検証では AVeriTeC の dev セットを用い、入力データは AVeriTeC の主張と日付情報のみを用いた。提案手法の Process2 において入力主張に曖昧であると判断された場合には、HITL の処理として AVeriTeC から主張の発信された媒体、発信者、国を不足情報として取得し機械的に返答する。これは本検証が人との対話の評価ではなく、不足情報や曖昧性の補間が後段に与える影響を分離して評価するための設定である。

AVeriTeC のクラスラベルは Supported, Refuted, Not Enough Evidence, Conflicting Evidence/Cherrypicking で構成される。提案手法においては True と Almost True を Supported として、Inaccurate と False を Refuted として、Baseless を Not Enough Evidence として扱った。また、提案手法は Refuted と Conflicting Evidence/Cherrypicking を区別していないため、クラスラベルが Conflicting Evidence/Cherrypicking の場合のみ Inaccurate と False を Conflicting Evidence/Cherrypicking として扱った。評価は Accuracy に加え、AVeriTeC のクラス不均衡の影響を考慮して macro-F1 スコアも算出した。

本検証ではテキスト入力のみであるため、逆画像検索や Deepfake・加工検知などの検知器といったメディアデータに対する分析ツールは使用しない。また、Web 検索として Google Custom Search JSON API [11] を使用した。

表 1 検証結果

Method	Accuracy	macro-F1
Baseline (GPT-5.1)	0.690	0.554
ClaimCheck	0.726	0.436
MAFT-AH (Ours)	0.784	0.623

4.2 結果と考察

検証の結果を表 1 に示す。提案手法の Accuracy が他の 2 手法を上回る結果となった。特に本検証条件における提案手法と ClaimCheck との主な差異は Clarifying LLM による明確化であり、この結果は、実世界の発信者・媒体・地域などの欠落情報を補完し、調査方針を具体化し適切な証拠収集の実現が精度向上に寄与したことを示唆する。また macro-F1 も大きく向上しており、曖昧性補間による明確化が少数クラスに対してより効果があったと考えられる。

本検証は、HITL の返答を疑似的に行っているため実運用を想定した場合には人手による入力とは異なるという限界が存在する。また、提案手法のマルチモーダルな入力に対する評価やマルチエージェントによるオーケストレーションの評価が不十分であるため、Mocheg データセット [12] や NewsCLIPpings データセットなど [13] マルチモーダルなデータセットを用いて単一フローのマルチモーダル AFC 手法である MAFT などと比較する必要がある。

5 まとめと今後の課題

マルチモーダル情報に対する AFC の実運用では、入力の情報欠落・曖昧性への対応と、複雑な調査フローを管理するオーケストレーションが課題となる。本研究では、これらに対して HITL を備えたマルチエージェント AFC フレームワーク MAFT-AH を提案した。

提案手法は入力の情報欠落・曖昧性を LLM で検知し、必要に応じて HITL により補完を行う。また、入力内容から調査方針を生成し、Supervisor/Researcher からなるマルチエージェントで証拠を収集することで入力に応じた調査フローの実行を可能にする。AVeriTeC による評価では、既存手法を上回る Accuracy を達成し、曖昧性補間による有効性を確認した。今後はマルチモーダル入力に対するオーケストレーションの定量的評価を行う。

謝辞

本研究は、総務省のインターネット上の偽・誤情報等への対策技術の開発・実証事業において実施されました。

参考文献

[1] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, Vol. 36, pp. 65128–65167. Curran Associates, Inc., 2023.

[2] K. Kakizaki andY. Matsunaga and R. Furukawa. Maft: Multimodal automated fact-checking via textualization. In *AAAI, vol. 39, no. 28,*, pp. 29646–29648, 2025.

[3] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14920–14929, 2022.

[4] Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. From chaos to clarity: Claim normalization to empower fact-checking. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6594–6609, Singapore, December 2023. Association for Computational Linguistics.

[5] Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. Document-level claim extraction and decontextualisation for fact-checking. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11943–11954, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] Yue Chen, Chen Huang, Yang Deng, Wenqiang Lei, Dingnan Jin, Jia Liu, and Tat-Seng Chua. STYLE: Improving domain transferability of asking clarification questions in large language model powered conversational agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10633–10649, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[7] Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10746–10766, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[8] Akshith Reddy Putta, Jacob Devasier, and Chengkai Li. ClaimCheck: Automatic fact-checking of textual claims using web evidence. In Weijia Shi, Wenhao Yu, Akari Asai, Meng Jiang, Greg Durrett, Hannaneh Hajishirzi, and Luke Zettlemoyer, editors, *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pp. 303–316, Albuquerque, New Mexico, USA, May 2025. Association for Computational Linguistics.

[9] Tam Trinh, Manh Nguyen, and Truong-Son Hy. Towards robust fact-checking: A multi-agent system with advanced evidence retrieval, 2025.

[10] 日本ファクトチェックセンター. Jfc ファクトチェック指針, 2026. https://www.factcheckcenter.jp/guidelines/.

[11] Google. Custom search json api, 2025. https://developers.google.com/custom-search/v1/overview.

[12] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, p. 2733–2743, New York, NY, USA, 2023. Association for Computing Machinery.

[13] Grace Luo, Trevor Darrell, and Anna Rohrbach. Newsclippings: Automatic generation of out-of-context multimodal media. 2021.