第３回インフォメーション・ヘルスアワード『社会実装部門』　応募用紙

○アイデア名　Think Critical: 認知の罠を見抜く LLM　― 批判的思考を育む次世代リテラシーツール

○応募者氏名
**　または　グループ名　株式会社 TDAI Lab**

○解決したい情報環境をめぐる課題

私たちが解決したい情報環境における課題は、偽情報・誤情報の拡散による社会問題です。偽情報・誤情報は災害時や選挙などの局面で拡散されてしまうと、陰謀論を助長したり、社会的分断を引き起こします。
特にソーシャルメディアの情報空間においては、短く感情に訴える強いメッセージほど拡散されやすい傾向があります。そうした投稿は利用者の直感的な反応を誘い、冷静な判断を妨げます。その結果、誤情報や偽情報が広がり、社会問題を引き起こす要因となっています。こうした状況を改善するには、情報リテラシー・批判的思考を強化する取り組みが不可欠です。

○アイデアの具体的な内容（どんなもので、どんな人が、どう使うと、課題が解決できるのか）

・どんなものか?
X の投稿に含まれる感情に強く訴える技法（論理的誤謬）を LLM(大規模言語モデル)で自動検出するツール。検出結果を「①ブラウザ拡張機能」及び「②X 上のボット」としてユーザーに提供します。
①ブラウザ拡張機能の場合：
論理的誤謬に関する３つの情報（論理的誤謬の種類、該当箇所、分類理由）について、ブラウザ拡張機能では画面スクロールに合わせて各投稿ごとに即時判定します。各投稿内で用いられている論理的誤謬の分析結果を１枚の「解説画像」の形式にまとめ、ハイライトした誤謬箇所にマウスオーバーした際にポップアップとして提示し、ユーザーの注意を促します。
②X ボットの場合：
リプライ欄でボットがメンションされた投稿、あるいはコミュニティノートが付与された投稿について、分析した結果をリプライとして自動送信します。ブラウザ拡張と同様に解説画像を投稿することで、ユーザーの注意を促します。

・誰が使うのか？
一般 SNS ユーザー：日常的にニュースや投稿を閲覧する際の情報精査ツールとして利用

・どのように使うと課題が解決されるか?
1. 要注意投稿の見逃し防止
SNS 投稿を見ている最中にハイライト箇所がユーザーの注意を喚起
2. 批判的思考の誘導
「なぜこの投稿が要注意か」を熟慮して判断（System 2 レベルで理解）でき、批判的思考を促進

○アイデアは未発表のものかどうか。すでに「試作」「試行」している場合は、新たに付け加えたいアイデア（ブラッシュアップするポイント）など

本アイデアの一部の機能については、SNS データ分析の中ですでに試行しています。例えば「原発」のテーマについて、論理的誤謬を用いた投稿割合やその拡散傾向などについて分析を行いました。「参院選」のテーマでは、ヒーロー・敵・犠牲者を定めたナラティブが特定の政党支持者の間で顕著に用いられていたことを突き止めた。また類似技術の検証事例としては、「東京都知事選」「総裁選」の 2 テーマについて、情報空間の可視化（各種候補者に対する賛否の投稿量・支持理由）に関する分析記事を複数メディア寄稿しており、技術的にも実用レベルであることが示されている。
本事業においては、これまでの弊社の知見をシステム実装しユーザー体験に変換することで、偽・誤情報に迅速に対抗できるツールを提供したいと考えています。

○アイデアを思いついたきっかけ

私は SNS やネット討論番組を見ていると、しばしば強い違和感を覚える瞬間がありました。
ある時、ネットの討論番組で人気のインフルエンサーが、相手の発言を歪めて単純化し、それを攻撃する――いわゆる「藁人形論法」を用いていました。番組のコメント欄や SNS ではその主張に賛同する声が一気に広がり、しばしば強い違和感を感じることがありました。
そんな中偶然、「藁人形論法」以外にも「恐怖訴求」や「滑り坂論法」といった論理的誤謬と呼ばれる分類法があることを知りました。冷静に見れば過激な表現だとわかると思いますが、自分もそうだったように多くの人にとっては感情的に反応してしまうと思います。
これらの体験を通じて、「こうした論理的誤謬を自動で検出し、ユーザーに注意を促してくれる仕組みがあれば、SNS の情報空間はもっと健全になるのではないか」と考えるようになりました。

○アイデアを実現するための技術的裏付けや学術的裏付け

本アイデアは、自然言語処理技術と大規模言語モデル（LLM）の進展を基盤としています。私たちはこれまでに、SNS 投稿に含まれる論理的誤謬（誇張、因果の単純化、感情的訴求など）を分類・抽出し、その拡散傾向を分析する研究を継続的に行ってきました。その結果、投稿に潜む認知操作の特徴を高精度に検出できることを確認しています。
近年の LLM の進化により、テキストの意味構造を解釈し、理由づけとともに提示することが可能になりました。
私たちはすでに国際会議や学術誌、メディア寄稿を通じて成果を発表しており、大規模 SNS データを並列処理する実務的知見も蓄積しています。これらの実績は、本提案のツールの技術的実現性を裏付けています。
さらに、東京大学鳥海研究室との連携により、計算社会学・情報的健康の学術的知見を取り入れています。学術研究の裏付けと、弊社が培ってきた生成 AI 活用の実務経験を組み合わせることで、社会実装可能なレベルでのサービス提供が可能です。

○アイデアを実現するための期間および資金、具体的な工程表

プロジェクト全体はおよそ 9 ヶ月で完遂を目指します。
初期段階から正式リリースまでの開発に必要な総予算は約 2,600 万円を想定しています。内訳は以下の通りです。
人件費（約 2,300 万円）：フロントエンド／バックエンドエンジニア各 1 名、プロジェクトマネージャー 1 名をフルタイムでアサインし、計 27 人月×月額約 85 万円で算出
LLM API 利用料（約 100 万円）：開発・テスト・初期運用フェーズを通じた OpenAI API 呼び出しにかかるトークン費用
インフラ／ツール費用（約 200 万円）：CI/CD 環境、ホスティングなど

工程は以下のスケジュールで進行します。
1. MVP 開発（1～3 ヶ月目）
ブラウザ拡張機能の基本 UI 開発
SNS 投稿上で論理的誤謬を検出し、ハイライト＋ポップアップで提示する機能を実装
X ボットの試作（投稿に対する自動リプライ形式での解説画像提示）
社内 PoC で技術検証

2. α版改良（4～6 ヶ月目）
誤検知フィードバック機能、キャッシュ処理の導入
ブラウザ拡張とボットを統合し、複数環境での安定稼働を検証
論理的誤謬分類の精度向上
クローズドβテストを実施し、実際の SNS ユーザーの利用体験と批判的思考効果を確認

3. β版テスト（7～9 ヶ月目）
複数 SNS（X、Facebook など）および主要ニュースサイトへの対応拡大
批判的思考スコア、利用履歴ダッシュボードなどユーザーエンゲージメント機能を追加
パブリックβを通じて利用データを収集し、最終的なバグ修正・UX 改善・ドキュメント整備を行い正式リリース

○アイデアを実現するための応募者の実施体制・能力（専門的技能、知識や経験など）があれば付記

私たちのチームは、技術的なスキルと学術的支援を活用し、プロジェクトを進める強固な体制を整えています。
弊社は計算社会学を専門とする東京大学鳥海不二夫教授を技術顧問としており、鳥海教授からの直接的なサポートを受けることで、最先端の知見や技術的アドバイスを得られる環境にあります。鳥海教授は情報的健康分野の第一人者であり、研究室の支援は、ツールの学術的裏付けを強化します。さらに、鳥海研究室で博士号を取得した福馬が弊社の代表をしており、自然言語処理技術を活用したソーシャルメディア解析の研究の実務に精通しており、ネット上の議論構造の解析に関する専門的知識・実績を有しています。すでに試行しているアプローチであり、技術的な検証も進んでいます。
技術的な実装おいては、弊社CTO野田がアプリケーション開発の設計・実装を担当しています。彼はこれまでAI搭載システムの開発に5年間携わっており、インフラからフロントエンド、バックエンドまで一気通貫して対応が可能です。昨今では生成 AI の利用についても取り組んでおり、特に LLM を用いたテキストデータの議論構造分析に強みを持っています。

「第3回インフォメーション・ヘルス AWARD」社会実装部門

Think Critical: 認知の罠を見抜くLLM
— 批判的思考を育む次世代リテラシーツール

2025年9月30日

株式会社TDAI Lab

TDAI Lab

Confidential © TDAI Lab All right reserved.

背景①：アテンションエコノミーに支配されるSNS

SNSでは「即座に理解でき、感情を刺激する短い言葉」が目立ちやすい。

- SNSは**アテンションエコノミー** に基づいて設計
- 人間は複雑な情報よりも、**短く単純で感情を伴うメッセージ** に直感的・感情的に反応しやすい (→ いいね・シェア)
- SNSではこれらが多くの人に見られた結果、社会的な分断などを助長

Confidential © TDAI Lab All right reserved.

背景②：論理的誤謬は拡散されるための文法

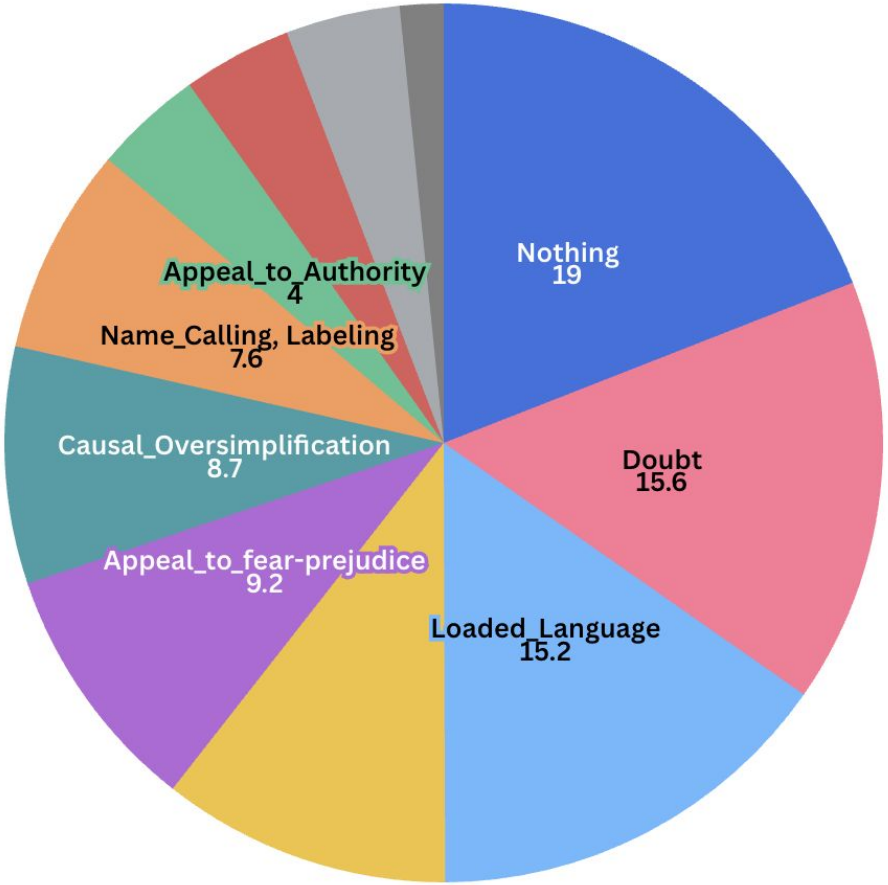
拡散に適した表現は、経験の積み重ねによって
思考誘導に作用するレトリックとして研ぎ澄まされてきた

- SNSは**拡散に適した文章テクニック** が繰り返し利用される場
  - 論理学の**「論理的誤謬」** や 社会心理学の**「プロパガンダ技法」**と重なる
  - これら表現が「嘘をつこう」としてではなく「こう書けば広まりやすい」と経験的に学ばれ、無自覚に使われている
- 引き起こす悪影響
  - ミクロレベル：**熟慮的思考の麻痺**
  - マクロレベル：**偽情報の拡散や社会的分断の加速**

Confidential © TDAI Lab All right reserved.

補足： 論理的誤謬はどれくらい使われているのか？

弊社独自調査では約8割の投稿に何かしら論理的誤謬が使われている

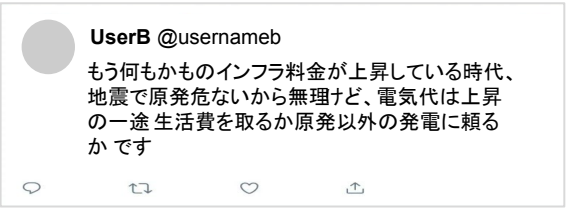


Confidential © TDAI Lab All right reserved.

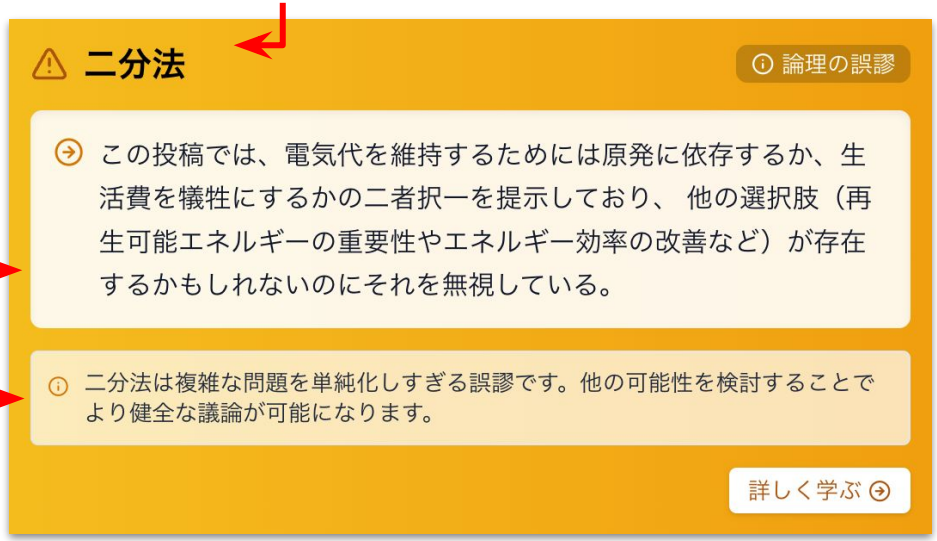
提案：LLMで誤謬検知と警告の表示

LLMの判定による「警告」で批判的思考を促す

- Xの投稿に対して、LLMを用いた論理的誤謬検知を行う
  - (機能1) 14種類の分類ラベル
  - (機能2) 判定理由



14種類の分類ラベル



判定理由

誤謬説明

Confidential © TDAI Lab All right reserved.

本ツールで分類する論理的誤謬一覧

分類名	説明	例文
Appeal to Authority	権威ある人物や機関が支持しているからという理由で、主張が正しいとみなす手法	「世界保健機関が、新薬はこの病気に最も効果的な治療法だと述べた。」
Appeal to fear-prejudice	不安・恐怖・偏見を植え付け、支持を得ようとする手法	「あの難民たちを止めろ。彼らはテロリストだ。」
Bandwagon, Reductio ad hitlerum	多数派に便乗・嫌悪対象との関連付けによる正当化・否定	「クリントンを大統領に選びますか？ 57%が『はい』と答えています。」
Black-and-White Fallacy	他の選択肢を隠し、二択しかないように見せかける手法	「あなたは共和党員か民主党員のどちらかでなければならない。」
Causal Oversimplification	複雑な原因を単純化して説明する手法	「もしフランスがドイツに宣戦布告しなければ、第二次世界大戦は起こらなかった。」
Doubt	信頼性に疑念を投げかける手法	「彼は市長になる準備ができているのか？」
Exaggeration, Minimisation	誇張・過小評価で印象操作を行う手法	「彼女と喧嘩していたんじゃない。ただ遊んでいただけだよ。」
Flag-Waving	愛国心・帰属意識に訴えて行動を正当化する手法	「この戦争に参加すれば、我が国の未来はもっと良くなる。」
Loaded Language	感情を煽る強い言葉で聞き手に影響を与える手法	「一人の議員の子どもじみた怒鳴り声。」
Name Calling, Labeling	好悪を誘うラベルを人・団体に貼る手法	「小ブッシュ（ブッシュ・ジュニア）。」
Repetition	同じメッセージを繰り返して刷り込む手法	「我らの偉大な指導者は知恵の象徴だ。彼の決断は常に賢明で公正である。」
Slogans	印象的な短いフレーズで感情に訴える手法	「アメリカを再び偉大な国に！」
Thought-terminating Cliches	議論や批判的思考を遮断する決まり文句の使用	「それはそれで仕方ない。」
Whataboutism, Straw Men, Red Herring	直接反論せず、矛盾や他の問題を持ち出して信用を落とす手法	「彼らはFBIの評判を守りたいだけだ。」

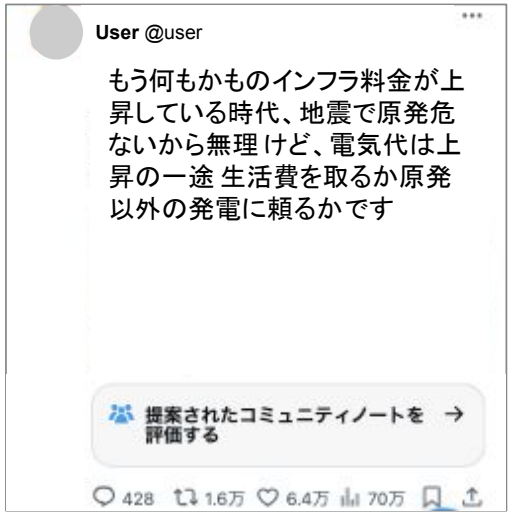
Confidential © TDAI Lab All right reserved.

提供形態① ブラウザ拡張機能

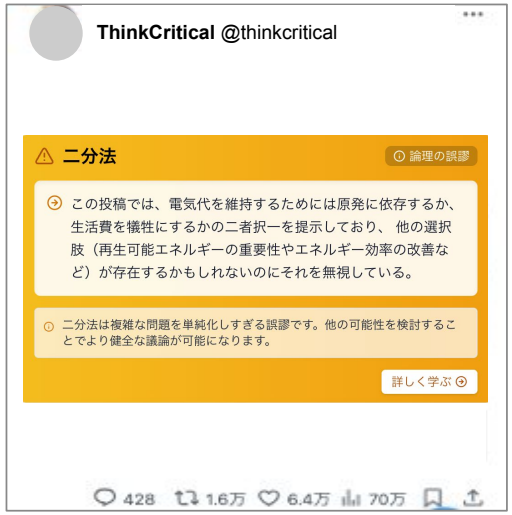
- Xのサイトのアクセスする際に自動で起動
- 画面スクロールに合わせて自動で各投稿を読み込み、論理的誤謬が使われている箇所を**即時ハイライト**
- ハイライト箇所に**マウスオーバー**することで詳細情報を**ポップアップ表示**

提供形態② Xボット

- リプライ欄でボットが メンションされた投稿、あるいは コミュニティノート が付与された投稿について分析した結果を リプライとして自動送信
- 投稿のどの箇所がどのような論理的誤謬なのかをハイライトして解説した画像をリプライとして投稿



リプライ



Confidential © TDAI Lab All right reserved.