第２回インフォメーション・ヘルスＡＷＡＲＤ『アイデア部門』　応募用紙

○タイトル LLM との対話における意見の偏りの強化対策

○応募者氏名
**　または　グループ名　みたほり班**

○解決したい情報環境をめぐる課題

・生成コンテンツの**政治的偏り**が指摘されており、ユーザーの意見が AI に誘導される可能性があります（https://arxiv.org/abs/2403.18932）。また、ユーザーの意見に過度に同調（Sycophancy）する生成 AI も、ユーザーの視野を狭め、意見の偏りを助長する恐れがあります（https://arxiv.org/abs/2310.13548）。
そのほか、生成 AI による**意見の多様性の減少**も懸念されており、ユーザーが異なる視点に触れる機会が減ることで、同じ意見に固執する傾向が強まります（https://arxiv.org/abs/2302.00560）

○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）

この課題を解決するためのアイデアは、生成 AI が提供するコンテンツの中でユーザーへの過度な同調や偏った情報が見られた場合に、それを検出し、**注意喚起**を行うシステムです。このシステムは、以下のような機能を持ちます。

1. **偏り検出とバランスの提供**

生成 AI が特定の政治的または意見的に偏った内容を提供した際、その偏りを検出します。そして、ユーザーに**逆の視点**を提示することで、多様な意見に触れる機会を提供します。具体的には以下の論文（https://arxiv.org/abs/2404.08699）に基づき政治的偏向をもつ LLM の作成を行ない、様々な政治的立場からの意見を提示します。

2. **ユーザーの意見傾向の追跡**

ユーザーのプロンプトや入力内容の傾向が変化した際に、AI がその変化を認識し、過度な偏りがないかをチェックします。これにより、意見の過激化や一方的な視点への固定化を防ぐことができます。逆にユーザーの意見と AI の意見が一致する傾向が見られた際にも通知および他の意見の提示を行ないます。

このアイデアは生成 AI を使用する全ての人に向けて作ったものですが、特に生成 AI を日常的に使い込んでいる人に使用してもらうことを考えています。生成 AI を普段から使っている人も意見の多様性が減ることは好ましい状況ではないでしょうし、それだけ使っていると生成 AI を信頼してしまい盲信している可能性も高いと考えられます。

普段からこのアイデアの機能を使うことで AI コンテンツをより安全に使用することができるようになり、社会的健全性を高めることができると考えています。特に、LLM との関わりに際する意見強化がもたらす政治的な分極化や、誤情報の拡散が社会に大きな影響を与えるリスクがあるため、LLM が社会に浸透しきっていない今だからこそ早めにやるべきアイデアだと考えています。

○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

このアイデアは現在、未発表のものです。

このアイデアを実現するにあたっては
1. 意見の傾向をどうやって類型化し、その変化を検知するのか
2. 政治的バイアスをもつ LLM をどう開発するか
という 2 点がネックになると考えています。

特に政治的バイアスについては日本では政党ごとの右派左派といった政治的な傾向が曖昧な傾向があり、右派の学習データ、左派の学習データといったものの用意が難しいことが挙げられます。この点について解決することができ、学習データを用意することができればこのアイデアは現実味を帯びることとなると考えています。

○アイデアを思いついたきっかけ

生成 AI の急速な普及とともに、具体的な悪影響は目にしていないものの、そのあまりの成長速度と人間味から、生成 AI を人間以上に信じてしまう未来をひしひしと感じると同時にその怖さを感じ、どうにか安全に生成 AI と関わっていけないかと考えこのアイデアを思いつきました。特に、政治的な分極化や、誤情報の拡散が社会に大きな影響を与えるリスクが高まっているため、AI が提供する情報の信頼性と公平性を確保することの重要性を強く感じています。