対話型 AI の判断過程の可視化を目的とした視点分離型プロンプトの設計

Multi-Perspective Prompting with Separated Evaluation Criteria for Visualizing Decision-Making in Conversational AI

安全 22-0035　上杉　弥生
Miki UESUGI
指導教員：河野　和宏

This study proposes a multi-perspective prompt based on separated evaluation criteria in conversational AI. Our prompt provides a visualization of the decision-making process in conversational AI. It also enables users to recognize several biases present in both user input and AI output. Our results show that our prompt makes it easier for users to understand multiple perspectives than the existing template-based prompt. We also observe that our prompt provides an opportunity to reflect on their own thinking.

Key Words: *conversational AI, prompt engineering, bias, fairness, neutrality*

1.　はじめに

近年，大規模言語モデルの発展に伴い，ChatGPT をはじめとする対話型 AI は，学習や業務支援など多様な場面で活用されている．一方で，新たな課題も生じている．そのひとつが，対話型 AI の出力内容に見られる偏りである．

生成 AI を基盤とした対話型 AI の出力は，一見すると公平性・中立性かつ客観的な情報を提示しているかのように受け取られやすい．しかしながら，実際には学習データの分布やモデル設計上の方針，さらには利用者の問いの立て方によって，特定の価値観や判断基準に偏った内容が生成される場合がある．その結果，利用者が複数の立場や評価軸を十分に検討しないまま，提示された結論を妥当なものとして受け入れてしまう危険性が指摘されている．この問題は，対話型 AI の利用が拡大するにつれて，公平性や中立性の観点からますます重要な課題となっている．

そこで，対話型 AI を正しく安全かつ有効に活用するために出力精度の向上や作業効率化を目的としたプロンプトエンジニアリング技法が数多く提案されている．代表的なものとして，タスク構造を整理したテンプレート型プロンプトがあげられる[1]．これは，

1）明確な質問：曖昧性のない明確な質問をする
2）具体性：トピックや要求を具体的・詳細に記載する
3）プロンプトの構造：質問を構造化して欠落をなくす
4）文脈の提供：重要な文脈や背景情報を提示する
5）複数の質問：複数の質問を連続して投げかける
6）ステップバイステップ指示：段階的に考えさせる
7）校正とフィードバック：結果を評価して校正させる

のように，役割設定や出力形式の指定により，望ましい回答を引き出す技法となる．また，バイアス対策としては，学習データやアルゴリズムに内在する偏りを AI 自身に点検させる手法も提案されている[2]．しかし，これらの方法は，プロンプト自体が複雑になりやすく，対話型 AI の利便性を低下させるという問題がある．

本研究では，この課題に対応するため，人間側と AI 側双方に存在するバイアスや思考の偏りに気づくための，簡便なプロンプトの枠組みを提案する．提案手法では，正しい答えを得ることよりも，どのような基準や視点を経てその答えに至ったのかを利用者が意識できる出力が得られるようになっている．加えて，専門的知識を必要とせず，一般の利用者でも簡単に実践できるようになっている．

2.　公平と中立の概念の定義づけ

本研究では，生成 AI における「公平」と「中立」という概念を，研究上の操作概念として以下のように定義する．

公平：特定の属性や価値判断が入力段階において不当に固定化されず，複数の解釈可能性が保持された状態で応答が生成されること

中立：対立する立場のいずれかを自動的に正当化する前提を含まず，問いの立て方自体に内在する価値判断を相対化し得る応答が生成される状態

これらの定義は，入力と出力の関係全体を分析対象とする本研究の立場を反映したものである．

3. 提案手法：視点分離型プロンプト

本研究で用いるプロンプトは，「視点の明示的分離」と「判断基準の外在化」の二つの要素を組み合わせており，これを視点分離型プロンプトと呼ぶ．具体的な手順は，次の通りである．

① 公平と中立の定義を事前に入力して記憶させる
② 対話型 AI に聞きたい問いを入力(出力は求めない)
③ 問いの後に視点分離のため，「#賛成/反対/中立の立場，社会的影響，個人への影響，実行可能性など，複数の評価観点を分けて説明せよ」と入力

この手順では，一つの立場から結論を出させるのではなく，賛成・反対・中立といった複数の立場から同一の問いを，同程度の分量で説明するよう指示する構造が採用されている．さらに，良い・悪いという総合評価を求めるのではなく，社会的影響，個人への影響，実行可能性などの複数の評価観点を分けて並列に提示させるとともに，表が作成されるため，判断基準を可視化することができる．

4. 実験結果と考察

大学生 10 名を対象に ChatGPT-4o を用いて実験を行った．対話形式は以下の 3 パターンである：①自由入力(各々の普段通りでの対話)；②既存のテンプレート型プロンプトによる入力（役割と制約条件を具体的に入力)；③提案手法による入力（視点分離型プロンプト)．問いを「生成 AI サービスに年齢制限を設けるべきか」とした際の，入力プロンプトと出力結果の一例(抜粋済)を以下に示す．

・ 自由入力

問：「生成 AI サービスに年齢制限を設けるべきか．」

返答例：「未成年が不適切な情報に触れる可能性があるため，一定の年齢制限は必要であると考えられる一方で，学習などに役立つ面もあるため，完全な禁止ではなく保護者の管理のもとでの利用が望ましい．」

・ テンプレート型プロンプトによる入力

問：「あなたは教育学者です．生成 AI サービスに年齢制限を設けるべきかについて，メリット･デメリットを挙げ，最後に結論を述べてください．」

返答例：「メリットとして有害情報への接触防止や依存の抑制が挙げられ，デメリットとして学習機会の制限とデジタル格差の拡大のため，教育的配慮は必要だが，年齢制限自体は導入すべきである．」

・ 提案手法（視点分離型プロンプト）による入力

問：「公平性・中立性を鑑み，生成 AI サービスに年齢制限を設けるべきかについて賛成の立場，反対の立場，中立的立場の 3 つからそれぞれ同程度の分量で論じたうえで，社会的影響，個人への影響，実行可能性など，複数の評価観点を分けて明示してください．」

返答例：「それぞれの問題点として賛成の立場では，未成年保護や誤情報拡散防止．反対の立場では，学習機会の不平等や技術理解の遅れ．中立の立場では，年齢制限よりも利用環境の整備やリテラシー教育の重要性が大事であるといえる．」

この返答例に続き，複数の立場と評価軸を基にした表が出力され，判断過程が可視化された出力構造となった．

これらの結果について，主観評価に基づき，各手法に対する印象や有用性を被験者から得た．

まず被験者 10 名のうち 9 名が，役割付与や制約条件を明示する既存のテンプレート型プロンプト手法そのものを知らなかったことが確認された．そのため，既存手法の時点で，「普段はそのまま質問していたので，枠組みを指定するという発想がなかった」といった感想が見られた．

提案手法である視点分離型プロンプトでは，「違う立場の意見もまとめられているため理解しやすく，～すべきだといった断定の言葉が書かれていないため自分自身で考えることの大切さを痛感した」といった感想が多く得られた．提案手法は，論点が整理されやすく，異なる立場の意見が明確に区別された構造で出力されるため，被験者は，判断に至る過程を追いやすくなったと考えられる．さらに，出力結果が特定の結論を断定的に示すものではないことから，被験者は出力結果を単なる結論として受け取るのではなく，前提条件や価値基準の違いに着目しつつ内容を再考するきっかけを得やすくなったといえる．

5. さいごに

本研究では，まず生成 AI における公平性や中立性を定義した．そのうえで，対話型 AI の判断過程を，異なる視点から可視化できる簡易なプロンプトを提案した．

参考文献

[1] 野口竜司：ChatGPT 時代の文系 AI 人材になる －AI を操る 7 つのチカラ，東洋経済新報社，2023 年 10 月発刊．
[2] Y. Lyu, S. Ren, Y. Feng, Z. Wang, Z. Chen, Z. Ren, and M. Rijke: Cognitive Debiasing Large Language Models for Decision Making, arXiv, https://arxiv.org/abs/2504.04141v3 (2026/1/1 確認)