

## GPT-3.5 を用いた議論中に含まれる主張の誤謬検出手法 Method for Detecting Fallacies of Claims in Discussion using GPT-3.5

牟田 真悟<sup>‡</sup> 白松 俊<sup>‡</sup>  
Shingo Muta Shun Shiramatsu

### 1. はじめに

誤謬の含まれた主張や反論は、議論の結論や合意形成を歪ませる可能性がある。誤謬とは「一見正しく見えるものの、実際は誤っている推論」のことをいい、誤謬を意図的に用いたものを詭弁という。よって本研究では、議論に含まれる誤謬を検出する手法の開発を目指す。

塩谷[1]は“クリティカル・シンキングは、議論を理性的に行ううえで基本的な道具として認識されるようになってきた”としたうえで、クリティカル・シンキングを実行するにあたって“「誤謬の議論」をできるだけ回避する努力を常に持つべきだ”と主張している。また、議論中に誤謬を含む主張や詭弁がある場合には、議論参加者は速やかに誤りを指摘することが望まれる。しかし、実際には議論中に参加者たちが誤謬を含まない発言をし続けたり、誤謬を速やかに指摘したりすることは難しい。

そこで本研究では、この問題を解決するために大規模言語モデル GPT-3.5 を用いて誤謬を検出するシステムの開発を目指す。本稿ではその前段階として、ある議論において、誤謬が含まれる主張の有無を示すことで議論の質が向上するか否かを、クラウドソーシングサイト Lancers を用いたアンケート調査によって検証する。

### 2. 関連研究

#### 2.1 オンライン上議論における誤謬の分析

山本ら[2]は、誤謬を見抜く取り組みの必要性を論じると同時に、疑似科学を中心としたオンライン上の議論に着目し、実際の議論の中でどのような誤謬が見られるか分析した。分析の結果、オンライン上のコメントでは意味の曖昧さや多義性に付け込み、受け手に先入観を与える誤謬が多く見られることを明らかにした。

#### 2.2 確証バイアスの検出

安加賀ら[3]は、論理的誤謬とフェイクニュースの性質が似ていることから、フェイクニュースを検出する手法で論理的誤謬が検出できると仮定し、論理的誤謬を検出する手法を提案した。安加賀らの手法では、論理的誤謬の中でも特に確証バイアスに着目し、フェイクニュースを検出する手法による確証バイアスの検出精度の限界を 75%とした。

本研究ではこれらの研究に続き、議論中の誤謬を検出、分析、明示することで、適切な議論の進行を補助するシステムの開発を目指す。

### 3. 提案手法

本研究では、大規模言語モデル GPT-3.5 (ChatGPT) に、特定のプロンプトを与えることで、誤謬を検出する。プロンプトの例を図 1 に、実行例を図 2 に示す。

以下は「学校の授業をオンライン学習にすべきか」についての議論の一部です。この議論に含まれる可能性がある誤謬を簡単に説明してください。ただし、誤謬が含まれているとは限りません。

Aさん: 学校の授業は完全にオンライン学習にすべきだと思います。オンライン学習は経費削減につながり、教材や施設のメンテナンス費用が不要になります。  
Bさん: そうですか、しかし、対面の学校授業では教師と生徒の関係性や相互作用が重要です。オンライン学習ではそれが制限される可能性があります。  
Aさん: それは確かですが、オンライン学習では新たなテクノロジーを導入できるため、より効果的な学習体験が提供できます。例えば、仮想現実技術を使ったバーチャル実験やシミュレーションなどが可能です。  
Bさん: それは興味深いです。オンライン学習では生徒同士や教師との対話やディスカッションが減る可能性があります。それによって学生の学習意欲や理解度に影響が出るかもしれません。  
Aさん: でも、オンライン学習ではチャットやフォーラムを活用して、生徒同士や教師とのコミュニケーションを補完することができます。実際、多くのオンライン教育プラットフォームではそういった機能が提供されています。  
Bさん: そうかもしれませんが、オンライン学習では生徒が自主的に学習する能力や自己管理能力が求められます。一方、対面の学校授業では教師が生徒をサポートし、学習環境を整えることができます。

図 1 プロンプトの例

この議論には特定の誤謬は見られません。それぞれの発言は個別の観点や利点を強調しており、両者が相手の主張に反論や補足を加えています。ただし、この議論は主にオンライン学習と対面の学校授業の比較に焦点を当てており、他の重要な要素や視点が欠落している可能性があります。例えば、教育の質や効果に関する証拠、異なる学習スタイルや生徒のニーズの考慮、教師のトレーニングや技術的なサポートなど、さらなる要素を考慮することが重要です。

図 2 実行例

例のように、プロンプトは「議題」「議論のログ」「誤謬の検出を促す文言」の 3 つで構成されている。このうち「誤謬の検出を促す文言」では「この議論に含まれる可能性がある誤謬を...」や「ただし、誤謬が含まれているとは限りません。」のように、議論中に誤謬が含まれていると断定することを避けており、正当な主張が誤謬と判定されることを抑制している。

### 4. 実験と結果

#### 4.1 実験設定・評価指標

本実験の目的は、ある議論において、誤謬が含まれる主張の有無を示すことで、議論の質が向上するか否かを検証することである。実験では、被験者はある議題（以下「議題 1」とする）に関する議論の例を読み、アンケートに回答する。その後、別の議題（以下「議題 2」とする）の議論の例を読み、再びアンケートに回答する。1 回目の議論の例（以下「システム無し」という）では誤謬の有無を示さず、2 回目の議論の例（以下「システム有り」という）では、誤謬の有無を、システムからの発言という形で議論例の最後に示した。実験は 4 つのグループに分けて行い、クラウドソーシングサイト Lancers から募ったのべ 399 名を対象とした。グループの詳細を表 1 に示す。

<sup>‡</sup> 名古屋工業大学 Nagoya Institute of Technology

表 1 被験者のグループ分け

	A (N=100)	B (N=99)	C (N=100)	D (N=100)
システム無し	議題 1 誤謬なし	議題 1 誤謬あり	議題 2 誤謬なし	議題 2 誤謬あり
システム有り	議題 2 誤謬あり	議題 2 誤謬なし	議題 1 誤謬あり	議題 1 誤謬なし

アンケートでは以下の項目について調査した。

- システム無し、有りで共通
  - この議論中の主張に納得感がありますか？  
(納得感—システム無し、システム有り)
  - この議論は有意義であると感じますか？  
(有意義さ—システム無し、システム有り)
- システム有りのみ
  - システムによる誤謬の指摘内容は適切だと感じましたか？  
(指摘の適切さ)
  - 誤謬の指摘があることによって、主張や議論全体に対する納得感が高まりましたか？  
(納得感の向上)

これらの質問は 1~7 までの 7 段階で評価する質問で、数字が小さいほど程度が低くなっている。アンケートでは、「この議論中の主張に納得感がありますか?」「この議論は有意義であると感じますか?」「誤謬の指摘があることによって、主張や議論全体に対する納得感が高まりましたか?」によって誤謬の指摘の有無による議論の質の変化を明らかにし、「システムによる誤謬の指摘内容は適切だと感じましたか?」によってそれらを支える GPT-3.5 による誤謬検出の精度を明らかにする。

## 4.2 結果

実験の結果を図 3 に示す。アンケート調査の結果、被験者のうち約 64%が「指摘の適切さ」において 5 以上を選択し、約 59%が「納得感の向上」において 5 以上を選択した。また、システムの有無による有意義さの違いを見ると、「有意義さ—システム無し」では 5 以上の選択肢を選んだのは全体の約 66%であるのに対し、「有意義さ—システム有り」では約 69%に増加している。しかし、同様に納得感を比較すると、「納得感—システム無し」では約 66%が 5 以上の選択肢を選んでいるのに対し、「納得感—システム有り」では約 54%に減少した。

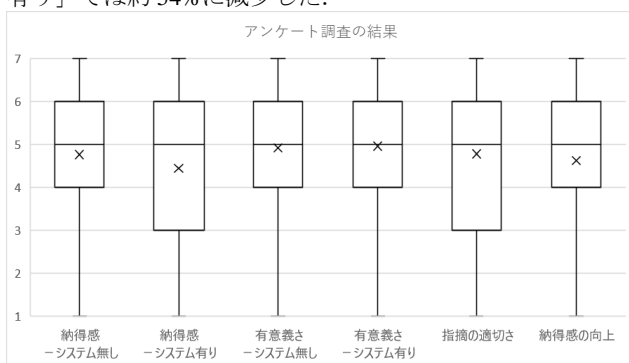


図 3 アンケート調査の結果 (N=399)

## 5. 考察

アンケートの結果から、全体の約 64%が「システムによる誤謬の指摘内容が適切である」と感じているため、GPT-3.5 にはある程度の誤謬検出能力があるといえる。

また、「誤謬の指摘による、主張や議論全体に対する納得感の高まり」を約 59%が感じているにもかかわらず、システムの有無による納得感の違いを見ると、実際には納得感が低下しているという結果が得られた。この結果は次の 2 通りに解釈できる。

1. マジョリティである「納得感が向上した」層の向上度合いがわずかであり、それに比べてマイノリティである「納得感が低下した」層の低下度合いが顕著である。
2. 誤謬が指摘されたことで主張自体に対する納得感は低下しているものの、システムによる誤謬検出までが議論の一部であるため、議論全体に対する納得感は向上した。

1 が主な原因である場合、誤謬の検出によって議論の質が向上したとは言いが、2 が主な原因である場合、誤謬の検出によって議論の質が向上したと言え、今後データを詳細に分析することで原因を明らかにする。

## 6. 結論

### 6.1 まとめ

本研究では、議論中の主張に含まれる誤謬を検出するため、大規模言語モデル GPT-3.5 を利用したシステムの開発を提案した。そこで本稿では、議論中における誤謬の指摘の有効性を検証するため、被験者を 4 つのグループに分けてアンケート調査を行った。その結果、約 64%が誤謬の指摘は適切だと感じており、GPT-3.5 にはある程度の誤謬検出能力があることが分かった。また、約 59%が誤謬の指摘によって主張や議論全体に対する納得感の高まりを感じたものの、それと同時に主張に対する納得感が低下していたため、議論中における誤謬の指摘の有効性を検証することはできなかった。

### 6.2 今後の展望

まずはデータを詳細に分析することで原因を明らかにし、誤謬の検出によって議論の質が向上したと言えるか検証する。また、重要な課題として出力が不安定であることが挙げられる。この問題は、プロンプトエンジニアリングおよび GPT-4 の利用によって改善できる可能性がある。今後はこれらの問題を考慮しつつ、システムの開発を進める。

### 謝辞

本研究の一部は、JST CREST (JPMJCR20D1) および NEDO (JPNP20006) の支援を受けた。

### 参考文献

- [1] 塩谷 英一郎, “言語学とクリティカル・シンキング—誤謬論を中心に”, 帝京大学総合教育センター論集, 3, 79-98 (2012).
- [2] 山本 輝太郎, 石川 幹人, 菊池 聡, “疑似科学を中心としたオンライン上議論における誤謬の分析”, 日本科学教育学会年会論文集, 42(0), 341-342 (2018).
- [3] 安加賀 元基, 大沢 英一, “言語ベースおよびユーザベースアプローチによる確証バイアスの検出”, 人工知能学会全国大会論文集, JSAI2022(0), 3P4GS202-3P4GS202(2022).