LLM as a Debate Judge: 学習者ディベーターへの自動フィードバック生成

尾崎大晟 [1] 市野敬介 [2] 松田拓 [3] 久保健治 [3] 内藤昭一 [6,7] 山口健史 [7] 天野祥太朗 [1]
井之上直也 [4,5] 中川智皓 [1] 新谷篤彦 [1]
[1] 大阪公立大学大学院 [2] 全国教室ディベート連盟 [3] 全日本ディベート連盟
[4] 北陸先端科学技術大学院大学 [5] 理化学研究所 [6] 株式会社リコー [7] 東北大学
sg23174y@st.omu.ac.jp

概要

本研究は，批判的思考力育成の有効手段とされるディベートにおいて，人的リソースの負担が大きいエキスパートジャッジの役割を，LLM エージェントで代替可能とするシステムを提案する．ディベートエキスパートらと共にマルチ LLM エージェントシステム用のディベート評価指標 (ジャッジ指標) を新たに設計し，それを元に肯定側へフィードバックを提供するジャッジシステムを構築した．さらに，システムが出力したフィードバックと，エキスパートが作成した模範フィードバックを比較評価する実験を行った．その結果，論点の整理などといった観点では LLM によるフィードバックが人間エキスパート以上の品質を示した．

1 緒言

1.1 背景と課題

批判的思考力[1)]の育成は高度情報化社会おいて，国家の重要課題となっている．批判的思考力の育成にはディベートをすること，とりわけディベートを行い，ジャッジからのフィーバックを受けることが有効とされている．しかし，高品質なフィードバックを行うエキスパート評価者が必要となり，人的コストが大きい．そこで本研究グループでは人間のディベートジャッジを代替できる LLM エージェントを用いた高品質ディベートジャッジシステムを開発し，この課題を解決を目指す．しかしマルチ LLM エージェントシステムを用いた学習者ディベーターへのフィードバックを行った研究は少ないため，教育効果の高い高品質フィードバックを行うことができるのか未知数である．さらに LLM エージェントシステムに合わせてデザインされたディベートの評価指標がないという課題もある．

1) 論理的・客観的で偏りのない思考であり，自分の推論過程を意識的に吟味する反省的思考である [1].

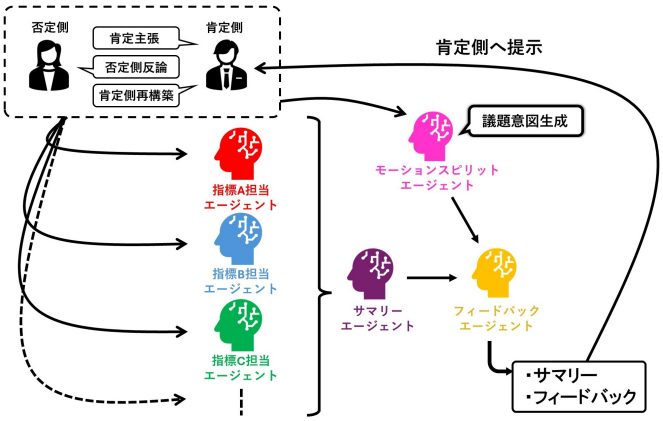

図 1 ディベートジャッジシステムの概要

1.2 本研究の取り組み

本研究ではディベートエキスパートらと協力し，ディベートジャッジとしてのディベート評価指標をマルチ LLM エージェントシステム用に再構築し提案する．同時にマルチ LLM エージェントを用いたディベートジャッジシステム (図 1 参照) を提案し，本システムのディベートジャッジとしてのフィードバック能力を定量的に評価する．本研究を通して，LLM エージェントシステムの教育応用場面の増加，利用ハードルの低減に繋がることを期待する．

1.3 貢献

- マルチ LLM エージェントシステムに合わせた独自のディベート評価指標 (表 1) を作成した．
- 本指標をもとに，学習者が行ったディベートに対してサマリー及びフィードバックを提示するディベートジャッジシステムを構築した．

表 1　ジャッジ指標

メリット・デメリット評価
議題を肯定するメリットはデメリットより大きいか 議題を肯定する前後で変化が大きいか メリットを得るためのプランは現実的で実行可能か
論理性評価
主張の推論過程に飛躍や矛盾がないか 各論点が適切に関連付けられているか 主張を裏付ける証拠や事例が適切に提示されているか
反論への対応評価
反論の内容を理解し核心を捉えているか 反論に対して直接的に応答しているか 反論の問題や限界を適切に指摘しているか 反論を受けて，主張を強化また修正しているか 再構築された主張は，元主張より説得力が増しているか

- 人間のエキスパートとシステムのフィードバック品質を比較評価することで．本システムのディベートジャッジとしての活用可能性を示した．
- 実験で得たディベートデータに，人間のエキスパートのフィードバックとシステムのフィードバックが付されたデータセットを公開した[2]．

2　関連研究

LLM による論述の評価は人間のエキスパートや素人と比較して高い一貫性があり，特定の評価指標においてはエキスパートの評価と高い一致率を示したことが報告されている [2]．また学生らが作成したエッセイへのフィーバックの同時生成も検討されており，学生の作文スキルの向上に役立つフィーバックを生成した [3].LLM エージェントが注目されて以降は，ディベート評価のように多数の評価指標を要求するタスクにおいてマルチエージェントを用いた評価フレームワークが提案されている [4].

3　提案システム

本研究が提案するジャッジシステム (図 1)[3]は，ディベートデータと評価指標の入力を受け，入力されたディベートの要約と，そのディベートの肯定側に対するフィードバックを出力する．入力する評価指標を表 1 に示す．本評価指標はディベート甲子園で用いられる審判テキストの「議論評価の方法」，および Wachsmuth らの論述評価指標 [5] をベースに，ディベートエキスパートらと独自にデザインした．評価指標は合計 11 の小指標からなり，各小指標を単独で担当する専属エージェントが当該指標につ

2）https://huggingface.co/datasets/DeL-TaiseiOzaki/debate_dataset_with_feedback_ja

3）https://github.com/DeL-TaiseiOzaki/LLM_debate_judge

表 2　フィードバックの作成指針 (概要)

1. フィードバックの目的
試合における選手の議論展開を分析し、優れた点と改善点を明確に提示し，学習者の学習効果が最大化する
2. フィードバックの基本原則
客観性: 個人的な好みや特定の立場によらず，ディベートの基本的技術と論理性に基づいて評価を行う **具体性**: 抽象的な表現を避け，具体的な事例や改善方法を示す **発展性**: 学習者のディベートに向かう意欲の向上と長期的な成長につなげる
3. フィードバックの構造
優れている点 (400 文字): 論証において効果的だった要素，議論の構造で評価できる点など **改善提案 (400 文字)**: 論理展開の改善点，議論構造の最適化案など
4. 評価の観点
(表 1 参照)
5. その他の留意点 (一部抜粋)
肯定側に対するフィードバックであること． メリット評価・反論への対応評価を重視すること．

いて評価を行い，各エージェントが行った評価を要約エージェントがまとめる．同時に議題を受け取ったモーションスピリットエージェントがモーションスピリット (議題意図，付録 A) を生成する．ディベートサマリーとモーションスピリットを受け取ったフィードバックエージェントが学習者に対してフィーバックを送る．またフィードバックエージェントには表 2 に示すフィードバックの作成指針が入力される．本指針はより高品質なフィードバックの生成を目的に，ディベート甲子園の審判テキスト [6] をベースにして，ディベートエキスパートらと独自にデザインした．システム全体としてディベートのサマリーとフィーバックの 2 つが出力される．

4　評価実験

4.1　評価実験概要

本研究では，評価用のミニマムディベートデータセット 35 セットに対して，システムが生成したフィードバック (以下，**システムフィードバック**と呼ぶ) と，人間のディベートエキスパートが作成したフィードバック (以下，**模範フィードバック**と呼ぶ)，それぞれに対して表 3 に示す評価指標での人手のスコアリング評価を行い，比較した．加えてシステムの生成したディベートサマリーの評価も行った (付録 C 参照)．実験にはすべてのエージェントに gpt-4o-2024-08-06[4]を使用した．人手の評価は 2 名

4）主要内部パラメータである temperature は 0.7，max_tokens は 1024 に固定し生成した．

のディベート甲子園での審判経験あるエキスパートが作業を行った．評価時は生成 AI サービスやインターネットサイトの閲覧，及びエキスパート間での相談を禁止した．エキスパートには全 35 件のシステムフィードバックのうち 5 件を用いてキャリブレーションを行った．

4.2 評価用データセットの構築

本研究で評価対象とするミニマムディベート (図 1 参照) は議題に対して肯定側がまず肯定主張を行い，それに対し否定側が否定側反論を行い，それを受けた肯定側が肯定側再構築を行う 1 ターンの特殊ルールのディベートである．このミニマムディベートデータを gpt-4o を用いた自動生成により構築した．構築したデータセットの概要を表 5 に示す．過去に PDA[5] の全国大会やディベート甲子園[6] で用いられた議題に独自の議題を加えた 12 議題に対して，肯定主張，否定側反論，肯定側再構築を順に生成した．この作業を各議題に対して複数回行い，合計 35 セットの評価用データセットを構築した．またデータセットに対して付録 B に示す評価実験を行い，評価に十分な品質のデータセットであることを確かめた．

4.3 フィードバック評価指標の構築

本研究でフィードバック評価に用いる評価指標 (表 3) は，Beatriz らの研究 [7] で提案されているフィードバックの制御観点とディベート甲子園審判テキストをベースに，ディベート審判経験のある方 7 名へのインタビュー結果を反映して，ディベートエキスパートらと独自の指標を構築した．

4.4 模範フィードバックの作成

35 セットのディベートデータに対して，2 名のディベート甲子園での審判経験のある 2 名のエキスパートが手分けをして，各セットに 1 つの模範フィードバックを作成した．このとき各エキスパートは表 1 と表 2 の内容に準ずるフィードバックを作成するよう指示されており，生成 AI サービスやインターネットサイトを利用することを禁止とした．また各エキスパートは事前に 35 セット中 5 セットのデータセットを用いてキャリブレーションを行った．

5) 一般社団法人パーラメンタリーディベート人財育成協会 (https://pdpda.org/)
6) 全国中学・高校ディベート選手権 (https://nade.jp/koshien/)

表 3 フィードバック評価指標

No.	適用可能性
Q1	メリット・デメリットの具体的な改善方法が示されているか
Q2	論理展開の具体的な改善方法が示されているか
Q3	反論対応の具体的な戦略が提示されているか
	構造評価
Q4	内容が不明確な指摘が多く含まれていないか
Q5	抽象的な改善方法の提示が多く含まれていないか
Q6	長期的な成長視点の包含しているか
	実行可能性
Q7	改善提案は実行可能であるか
Q8	指示を受容すべき説明がなされているか
	評価基準の明確性
Q9	各評価観点の具体的基準に基づいているか
Q10	指針にない独自の評価指標を提案していないか
	価値
Q11	肯定的評価と改善提案のバランスは適切か
Q12	モーションスピリットを掴んでいるか
Q13	対象が表現していない有効な視点を提供しているか
	スタイル
Q14	専門的かつ理解しやすい表現であるか
Q15	建設的なトーンであるか

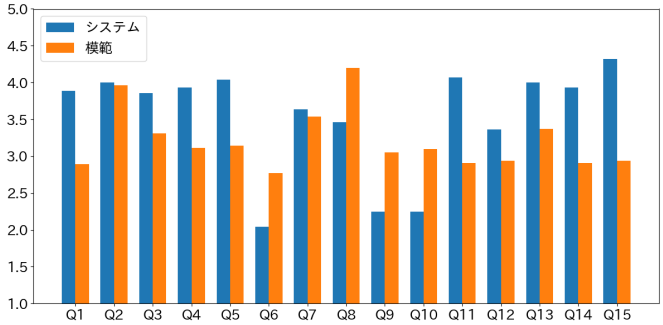

図 2 フィードバックの評価結果 (30 セット)

4.5 評価結果

図 2 にフィードバックの評価結果を示す．また実際の生成例，及び模範例を表 4 に示す．またこのときエキスパート間の評価一致率 Gwet'sAC1[8] の値は全体で 0.527 を示し，一定の信頼性を得た．結果として Q6,8,9,10 以外の指標においてはシステムフィードバック品質が模範フィードバック品質を上回り，LLM のフィードバック品質が人間のエキスパートと比較しても遜色ない，あるいはエキスパート以上の品質であることが確かめられた．

模範フィードバックは特に，Q8「指示を受容すべき説明がなされているか」，Q9「各評価観点の具体的基準に基づいているか」，Q10「指針にない独自の評価指標を提案していないか」の点においてシステムフィードバックに対して優れた．これは模範

表 4 フィードバックの改善提案例: 表 5 のディベートデータに対して

人間: 優れていた点は、主張の一つ目の観点において、外国の事例ではなく日本における事例の資料を証拠として引用した点は、日本という環境において週休三日制がどのような効果を齎すのかを証明できるため、メリットの発生確率が高いことを証明できていてとても効果的でした。(省略) 改善点としては，再構成の二点目の再反論は、企業ごとに休みにする日をずらすことで、社会全体は上手くという観点でのものだったため、否定側からの一つの企業が困難に陥るという反論の核心を捉えておらず、反論として機能していませんでした。改善策としては、一つの企業内で従業員ごとに休む日を変えるなど、一つの企業内の負担にフォーカスした反論をするべきです。(省略)
システム: 優れていた点：肯定側は議論の構造を明確にし、メリットの提示を多角的に行った点が非常に効果的でした。特に労働者の幸福度向上、経済活性化、少子高齢化対策という 3 つの主要な軸を設定したことで、議論の方向性に一貫性が生まれ、説得力を高めていました。(省略) 改善提案：肯定側が提示したメリットをさらに説得力のあるものにするため、証拠資料の幅を広げるべきです。例えば、他の業種や多国間の実験結果を補足的に引用することで、「一企業の限定的な事例」以上の汎用性を示すべきです。(省略)

フィードバックが改善提案の実行可能性，及び評価基準の明確性において優れていると言える．表 4 の模範例においては，改善提案に繋がる説明が明確で，かつ具体的にどのジャッジ指標においてどう評価されるかを明言しており，明確性が高い．一方でシステムフィードバックは Q4「内容が不明確な指定が多く含まれていないか」や Q5「抽象的な改善方法の提示が多く含まれていないか」，さらに価値やスタイルの観点において優れており，表 4 の生成例においては，追加すべき証拠資料を具体的に指示しており，学習者ディベーターが挙げた主張の根拠の限定性を指摘し，汎用性を高めるよう提案している．従って指示がより具体的で学習者にとって新たな学習観点を提示した価値が高いフィードバックになっていると言える．

4.6 定性的分析

システムフィードバック 本システムはまずディベート内容をサマリーすることで論点の全体像を整理し，フィードバックを生成している．議論の要点を見失わずに改善点を列挙できるのは，システム構造に依るところが大きいと考えられる．また LLM のフィードバックには学習者が提示した主張に対し，より多面的な根拠を添えるべきという指摘を積極的に行うという特徴がある．具体的には「国際的な指標やランキング」，「著名な再審事例」といった客観性や信頼性の高い情報を推奨するため，学習者の論証を証拠の観点から幅広く強化するヒントが豊富であった．

模範フィードバック 模範フィードバックはジャッジやオーディエンスを実際にどう納得させるかを重視しており，説得力を増すための表現方法や論理構成を丁寧に具体的に指摘していた．とりわけ，否定側反論に対する対応力を重視する傾向があり「同じ主張を繰り返しているだけの再構築は弱い」や「否定側の指摘を覆すための根拠が不足している」といった，肯定側再構築に対するフィードバックが特徴的であった．全体的にディベートの前提条件や前後の社会文化的要素といった暗黙的な要素など実際の競技ディベートでも勝敗を分けやすい要素をどれだけ言及できているかを重視していた．

システムフィードバックは広範なエビデンスの提示や論点整理の明快さに強みがあり，学習者の視野を広げる効果が期待できる．一方，エキスパートの模範フィードバックは，特定の反論や事例にどう対応すべきかをより具体的に示し，説得力の高い競技的手法を伝授するという強みを持っていた．

5 結言

本研究では，ディベートエキスパートとの協働によりマルチ LLM エージェントを用いたディベートジャッジシステムを提案し，その教育応用可能性について検証した．具体的には，ディベート評価基準を再設計し，本システムに適用することで，ディベートのサマリーおよび肯定側へのフィードバックを自動生成させ，さらに生成されたフィードバックを人間のエキスパートが作成したフィードバックと比較・評価した．その結果，論点整理や提示されるエビデンスの多面性といった観点でシステムフィードバックが十分に効果的であることが示された一方，改善提案を説得力ある形で実行可能性や競技的視点に結びつける点などでは，人間のエキスパートの強みが引き続き示唆された．

本システムは，競技ディベートだけでなく，初学者や実践的な議論スキル獲得を目指す場面においても効率的な学習支援を提供する可能性がある．今後は，提案したジャッジ指標の有効性や単一の LLM エージェントによるフィードバック品質などを検証する．加えて，学習者のディベート能力の改善への効果測定を目指す．これらの課題に取り組むことで，より教育現場への LLM 活用が進み，ディベートのみならず批判的思考力育成へ向けた対話型学習の領域でも活用が期待できると考える．

謝辞

本研究は，科研費 基盤研究 (A)22H00524「深い論述理解の計算モデリングと論述学習支援への応用」（代表：東北大学 乾健太郎教授）の支援を受けました．ここに謝意を表します．
また本研究におけるディベートエキスパートとして，一般社団法人パーラメンタリーディベート人財育成協会のディベートスタッフの方々，及び上智大学弁論部の方々，ディベート甲子園関係者の方々にご協力頂きました．ここに謝意を表します．

参考文献

[1] 楠見孝. 現代の認知心理学 3: 思考と言語. 北大路書房, 2010.

[2] Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. Are large language models reliable argument quality annotators? **ArXiv**, Vol. abs/2404.09696, , 2024.

[3] Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. Exploring llm prompting strategies for joint essay scoring and feedback generation. **ArXiv**, Vol. abs/2404.15845, , 2024.

[4] Jingcong Liang, Rong Ye, Meng Han, Ruofei Lai, Xinyu Zhang, Xuanjing Huang, and Zhongyu Wei. Debatrix: Multi-dimensinal debate judge with iterative chronological analysis based on llm. **ArXiv**, Vol. abs/2403.08010, , 2024.

[5] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**, pp. 176–187, Valencia, Spain, April 2017. Association for Computational Linguistics.

[6] 特定非営利活動法人 全国教室ディベート連盟 NADE（National AssociationEducation）. ディベート甲子園審判テキスト. 2022.

[7] Beatriz Borges, Niket Tandon, Tanja Kaser, and Antoine Bosselut. Let me teach you: Pedagogical foundations of feedback for language models. In **Conference on Empirical Methods in Natural Language Processing**, 2023.

[8] Nahathai Wongpakaran and et al. A comparison of cohen’s kappa and gwet’s ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. **BMC Medical Research Methodology**, Vol. 13, No. 61, pp. 1471–2288, 2013.

A　モーションスピリット

特定の議題から想定される議論の論点や文脈.「暴力的なゲームの未成年販売を禁止すべきである.是か非か」という議題があったときに，議題設定背景には未成年の健全な発育や規制の社会的影響などがあるはずであり，「暴力的なコンテンツが子供に与える影響とはどういったものがあるのか」や「規制するのは保護者なのかゲーム会社なのか，あるいは政府機関なのか」といったことがモーションスピリットとして挙げられる．議題設定背景には未成年の健全な発育や規制の社会的影響などがあるはずであり，それを前提とするフィーバックを行うことが重要とされる．

B　データセットの評価

本研究で評価対象として生成したディベートデータセットに対して，ディベートとして最低限の品質を保つものかを評価した．表 6 に評価指標を示す．各肯定主張，否定側反論，肯定側再構築の各論述の繋がりがディベートの体裁を保っており，かつ定義通りの論述になっているかを確かめた．各論述の定義は表 5 に付す通りである．評価はディベートエキスパート 3 名による，各指標をディベート全体と各論述毎で満たすかの二値分類評価で行った．結果を表 7 に示す．結果としてディベート全体として十分な品質であると判断した．このときエキスパート間の評価一致率を示す Gwet'sAC1 の値は 0.87 を示し一定の信頼性を得た．一方，否定側反論の品質が低いサンプルが多かったことから人手で作成した高品質サンプルを 5 セット追加し，合計 35 セットのデータセットを構築した．

表 5　データセットの例

議題: 日本は週休 3 日制を導入すべきである。
肯定主張 (議題に対して肯定する側が最初に立てる論のこと): 私たちは週休 3 日制の導入を強く主張します。その最も重要な理由は、働くことの本質的な目的を取り戻すためです。人々が働く目的は、生活の質を向上させ、人生をより豊かにすることにあります。しかし、・・・
否定側反論 (議題に対して否定する側が肯定主張に対して行う否定のこと．立論ではない): 肯定側は週休 3 日制による健康増進とワークライフバランスの改善を主張していますが、その論理には重大な欠陥があります。まず、・・・
肯定側再構築 (議題に対して肯定する側が反論を受けた上で肯定主張を改めた論のこと): 反論で指摘された問題点を踏まえ、私たちの主張を以下のように修正します。確かに単純な休日増加だけでは、労働密度の上昇や収入減少という問題を招く可能性があります。そこで・・・

表 6　データセットの評価指標

	ディベート全体
Q1	各論述が有機的に繋がっているか
Q2	議論されている論点が認知できるか
Q3	論点に一貫性があるか
	各論述
Q4	各論述がその定義通りの内容になっているか
Q5	推論過程に破綻がないか
Q6	明らかな社会倫理に反する内容ではないか
Q7	明らかな嘘を元に主張を

表 7　データセットの評価結果 (全 30 セットの平均)

No.	ディベート全体		
Q1	1.0		
Q2	1.0		
Q3	0.93		
No.	肯定主張	否定側反論	肯定側再構築
Q4	0.87	0.87	0.63
Q5	0.93	0.13	0.67
Q6	1.0	1.0	0.97
Q7	1.0	1.0	1.0

C　システムのサマリー評価

ジャッジが述べるサマリーは高い網羅性と真実性が求められる．そこでシステムが出力するディベートサマリーを前述の観点から評価した．表 8 に評価指標を示す. 評価はディベートエキスパート 3 名により行われ，Q16 については 5 段階のリッカート尺度で，Q17 については二値分類で行った．

表 8　サマリー評価指標

Q16	ディベート内の議論内容が網羅的にまとめられている
Q17	ディベート内で発話されていない内容を含めてない

評価結果を表 9 に示す．結果として網羅性，真実性ともに高評価と言える．本システムは複数のエージェントを回答を要約する形でサマリーを出力するが，肯定主張，否定側反論，肯定側再構築の 3 論述すべての重要要素に確率高く言及することができていた．このとき 2 者のエキスパート間での Gwet'sAC1 の値は 0.97 を示し一定の信頼性を得た．

表 9　サマリー評価結果

Q16 (1-5)	4.69
Q17 (0,1)	0.97