E-025

LLM に基づくディベートジャッジの設計 : 審判の決定プロセスを考慮して
Debate Judge Design based on LLM : A Qualitative Approach

中江 康公* 蜂巣 吉成* 吉田 敦* 桑原 寛明* 野呂 昌満†
Yasuhiro Nakae Yoshinari Hachisu Atsushi Yoshida Hiroaki Kuwabara Masami Noro

1 はじめに

ディベートとは，ある論題に対し，肯定側と反対側が審判という第三者の前で討論し双方の合意を得るプロセスである．世界各国で中高生や大学生向けの競技型英語ディベートが盛んに開催されている．審判は信頼性のある判定を下すために，高度なスキルと長いディベート経験を必要としている．しかし，その人数は限られており，ディベート審判の不足が課題になっている．

Transformer を基盤とした訓練済みの大規模言語モデル LLM(Large Language Model) は文章の質や内容に対し高度な評価能力を示しており，さまざまな分野におけるジャッジとしての応用が期待されている．しかし，ディベート判定というタスクにおいて，次の 3 つの課題が挙げられる．

1. LLM の入力コンテキスト制限
   LLM は Transformer 構造に起因する固定長の入力ウィンドウを持つので，一度に処理できるトークン数に制限がある．ディベートは複数の発表者による長大なスピーチで構成され，その総トークン数は容易に制限を超える．
2. 推論能力の限界
   LLM は文法的整合性のある出力を生成可能である一方で，ディベート判定に必要とされる高次の推論において性能が低下する傾向がある．
3. 価値判断
   ディベートの判定は，単なる事実の正誤判断ではなく，規範的な価値の比較を含む．たとえば，「自由と安全」や「経済成長と環境保護」といった相反する価値間のトレードオフが評価の対象となる．しかし，LLM はこれらの判断を倫理理論に基づく一貫した推論ではなく，学習データに依存して行うことが多い．

本研究は，LLM に基づき，人間審判の評価に近い判定を出すディベートジャッジ AI の設計を目的とする．そのために，審判の判定決定プロセスを取り入れたシステムアーキテクチャの設計を提案する．本研究は LLM 本体の改良ではなく外部のシステムアーキテクチャに焦点を当てる. その理由は，訓練が異なる新たな LLM にも置き換え可能な柔軟性や汎用性を確保するためである．汎用型 LLM の性能は今後も向上し続けると考えられるが，ディベートを含む長大な談話と複雑な構造を持つテキストに対する効果的な処理・評価フレームワークの確立を試みたい．提案アーキテクチャの有効性を検証するために，同一の LLM モデルとベンチマークを用いて，既存アーキテクチャとの比較検証を行い，定量的な評価を行う．実応用を想定し，判定の精度に基づきジャッジとしての実用性を総合的に評価する．本研究で得られた成果は次の 2 点である．

1. ブリティッシュ・パーラメント (BP) 形式のディベートを判定可能な新たなシステムアーキテクチャ DbKai を提案し，それに基づくシステムを実装した．このシステムでは，定性的評価を用いる人間審判の判定決定プロセスに類似した工程を模倣し，複数チーム間の比較評価を通じて勝者を決める枠組みを実現した．
2. DbKai と既存の LLM ベース判定モデルとの比較実験を行い，判定精度や妥当性などの観点に基づいて性能差を分析した．その結果，判定精度において既存手法を上回ることを確認した．

なお, BP 形式は競技型ディベートで広く用いられる形式の一つである．4 つのチーム (Opening Government: OG, Opening Opposition: OO, Closing Government: CG, Closing Opposition: CO) は賛成・反対に分かれて議論を行い，互いに競い合う．審判は 4 チームを順位づけし，最も高い順位を獲得したチームが勝者となる．

第 2 章では先行研究ならび関連技術，第 3 章では提案手法，第 4 章では検証実験および判定結果，第 5 章では考察，第 6 章ではまとめと今後の課題を述べる．

2 先行研究ならび関連技術

2.1 LLM の文章評価能力

本研究では LLM の判定精度を評価するために評価類似度 (agreement) を指標として用いる．評価類似度とは，同一の評価対象に対し，異なる評価者 (LLM また人間) がどの程度一致した判定を下すかを示す．ディベート試合を例に挙げると，3 人のジャッジが 2 つのチームに対し 2-1 の判定を下すと，この 3 人の評価類似度は約 67%である．

LLM をジャッジとして用いる可能性については，Zheng ら [1] は LLM の評価類似度を検証するためのベンチマークを定義した．OpenAI の GPT-4[2] を基盤とした LLM を対象にベンチマークを用いた検証実験では 80%を超える評価類似度を示した．これは人間の評価類似度と同一レベルである．しかし，上記先行研究はあくまで短い文章に対する評価であることに留意したい．

なお，BP 形式を用いる競技型ディベートの世界大会である World Universities Debating Championships (WUDC)[3] において審判の評価類似度は 80%前後であることが報告されている．本研究はこの値を審判の評価に近い類似度と考える．

2.2 Liang ら [4] のシステムアーキテクチャとベンチマーク

Liang ら [4] は，ディベートを評価するアルゴリズムと LLM の動作を制御するシステムプロンプトを組み合わせたシステムアーキテクチャ Debatrix を提案した．その概要を図 1 に示す．LLM の入力長を考慮してディベートスピーチを反復的に分析させ，異なる観点から多面的に評価する．具体的な評価フローおよび判定基準を

* 南山大学 Nanzan University
† 元南山大学 Formerly with Nanzan University

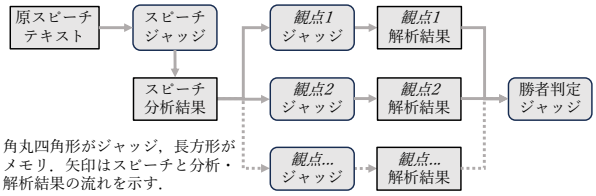

図 1　Debatrix のシステムアーキテクチャ

以下に示す．

1. ディベートにおけるスピーチテキストはスピーチジャッジに入力される．スピーチジャッジはスピーチの内容を要約した分析結果を出力する．このプロセスにより，モデルへの入力長を短縮しつつ，スピーチの内容について大まかな分析を行う．出力結果は出力保存用のメモリに保存され，後続のジャッジがそれを参照する．
2. スピーチ分析結果は観点ジャッジに入力される．観点ジャッジは任意の数に設定可能であり，それぞれ割り当てられた観点に基づいてスピーチを解析する．観点の割り当てはシステムプロンプトによって制御される．Liang らの研究では，BP 形式の試合において argument(議論)，source(根拠)，language(言語) と clash(クラッシュ) の 4 つの観点が用いられている．各観点ジャッジには，対象となるスピーチだけでなく，当該スピーチ以前の解析結果を併せて入力される．それぞれの観点ジャッジはスピーチに対する評価スコアと解析結果を出力し，観点解析結果メモリに保存する．解析結果は前述した分析結果と明確に区別される．分析結果はスピーチを要約のみ含むのに対し，解析結果は観点ごとの意味的分解と評価スコアを含む構造化データである．スピーチの評価スコアは該当チームのスコアとして扱われ，スピーチのたびに更新される．
3. すべてのスピーチが各観点において解析された後，すべての解析結果が勝者判定ジャッジに入力される．各観点にはあらかじめ比重が設定されており，最終的なスコアは全観点スコアにそれぞれ比重を乗じた加重平均として算出される．勝者判定ジャッジはスコアが最も高いチームを勝者として判定する．あわせて，試合全体の総評を出力する．

ディベートを評価する性能について，Liang らは過去に行われたディベート試合の記録をテキスト化したデータセットを用いて，ベンチマーク PanelBench を定義した．このベンチマークには，BP 形式の国際大会における決勝トーナメントラウンドから抽出された 22 試合が含まれている．各試合には審判による判定結果も含まれている．これにより，LLM による判定との一致度を測定することで判定精度を定量的に評価できる．Liang らは，これを用いて GPT-4 と GPT-3.5 turbo[7] をジャッジとする Debatrix などを比較する実験を行った．その結果，人間審判が出した判定を正解とみなした場合，GPT-4 単体の精度は 35%であったのに対し，Debatrix は 52%まで向上した．また，GPT-3.5 turbo 単体の判定完成率は 13.64%に対し，Debatrix は 100%であった．これは長文かつマルチターン形式の対話を評価するにあたり，システムアーキテクチャの有効性を示している．

Deabtrix は第 1 章で述べた入力コンテキストの制限に対する有効な手法を提供している．しかし，人間審判の判定に対する精度は依然として 80%に達していないことに留意したい．これは推論能力と価値判断に関する課題が未解決であることを示している．次の 2 つの要因が考えられる．

1. スコア集約型の勝者判定に伴う構造的乖離
   勝者判定が，個別観点ごとに付与されたスコアの合算によって行われている．一方，人間審判は各観点間の相互作用やチーム全体の説得構造を総合的に判断して勝者を決定している．そのため，Deabtrix と人間審判との間には，判定決定プロセスそのものに本質的な乖離があると考えられる．
2. 定量スコアと定性的評価の乖離
   Debatrix の議論ジャッジが出力した解析結果の一例を表 1 に示す．OG と CG の議論についてのスコア（Score）はそれぞれ 8 と 9 であるが，両者の評価理由にはいずれも「complete, coherent, and logically sound argument framework」という共通の評価を含んでおり，記述上の差異が極めて小さい．このように，定性的な記述と数値スコアとの対応関係が曖昧である．

表 1　Debatrix の議論ジャッジの解析結果の一例．OG と CG はチーム名．下線部分が共通している．

Argument Judge
Score of OG: 8
OG presented a complete, coherent, and logically sound argument framework, providing a critical perspective on the implications of global reliance on the dollar.
Score of CG: 9
CG provides a unique perspective on the redistribution of financial benefits among countries and the potential positive impacts of a decline in global reliance on the dollar. Their argument framework is complete, coherent, and logically sound, offering a comprehensive analysis of the benefits of reducing reliance on the dollar.
... 省略...

3　提案手法

以上の課題をふまえ，本研究では数値スコアに依存しない，定性的な比較判断を重視した評価アーキテクチャ DbKai を提案する．特に，各チームの説得構造を相対的に比較する過程をモデル化することで，人間審判に近い判定過程の再現を目的とする．

本研究のアプローチは次の通りである．

1. 人間審判の判定決定プロセスの定義
   人間審判がどのように複数観点を統合，比較して勝者を決定しているかを整理する．
2. 定性的な評価をもとに判定を下すシステムアーキテクチャの設計
   観点ジャッジの設計を再構成し，観点間の相互作用を評価判断に反映させる処理構造を導入する．

DbKai の設計・実装に加えて，PanelBench で判定精度を検証する．その後，出力された判定が定性的な比較評価に基づいているかを考察する．なお，本研究は競技型ディベートの判定に限定されるので，PanelBench に含まれる BP 形式の試合データのみを使用する．

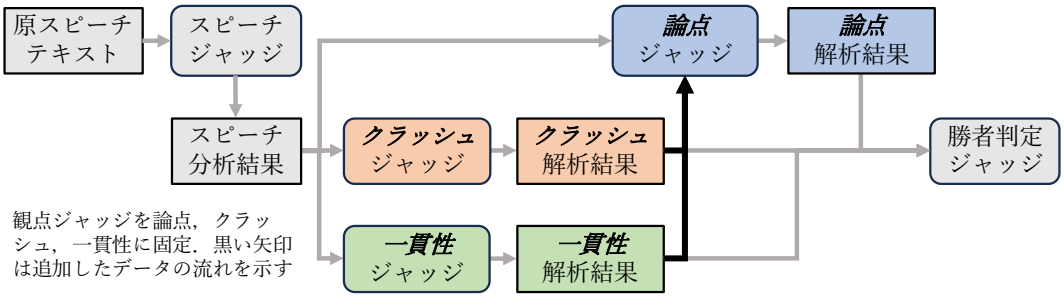

図 2　DbKai のシステムアーキテクチャ

3.1　人間審判の判定決定プロセスの定義

ディベートにおける人間審判の判定は，審査哲学（judging paradigm）と呼ばれる，複数の評価観点を総合的かつ比較的に捉える知的枠組みに基づいて行われる．松本 [5] によると，特に近年では，物語審査哲学が主流となりつつある．この枠組みでは，一般市民 (ordinary person) の視点から，どのチームの主張が最も納得できるかを判断することが求められる．WUDC の公式ジャッジマニュアル [6] はこれについて言及し，全体の議論に最も貢献したチームを勝者とする比較的評価 (comparative adjudication) を行うことが推奨されている．このプロセスでは，論点，クラッシュと一貫性が互いに関連しながら勝者決定に統合される．本研究では，人間審判による比較的かつ統合的判断の特徴をふまえ，システムアーキテクチャとシステムプロントの設計に反映させる．具体的に，観点ジャッジ間の情報を段階的に統合し，観点間の相互作用や議論全体の対立構造と展開を重視する処理機構を設計する．

3.2　システムアーキテクチャの設計

DbKai 全体の構造を, Liang ら [4] が定義した Debatrix にならい，図 2 に示す．Debatrix との主な 3 つの違いを以下に示す．

1. 観点ジャッジの構成の固定化
   観点別解析を議論に特化するために，観点ジャッジを「論点」,「クラッシュ」と「一貫性」の 3 種に固定する．図 2 の色付きの 3 つのジャッジ部とメモリに該当する．
2. 観点間の情報統合処理の導入
   論点ジャッジはクラッシュと一貫性の解析結果を参照できるよう変更を行う．他ジャッジの解析結果をフェッチできる構造を実装する．図 2 の黒色の矢印に該当する．
3. 勝者判定方法の再構成
   システムプロンプトにおいて，勝者判定ジャッジにおける判定方式をスコアによる定量的評価から論点の正しさと重要さに基づく定性的評価へ変更する．これにより，説得構造全体に着目した比較的・構造的判断が可能となる．

論点ジャッジにクラッシュと一貫性の解析結果を参照させるように変更を行った理由は，人間審判が判定において各論点を単独観点で評価するのではなく，それらが試合全体の中でどのように支持され，反論され，維持されているかを総合的に判断していることに基づくからである．特に，ある主張が他チームから十分に反論されているか（クラッシュ）や，同一チーム内で主張間に矛盾がないか（一貫性）は，その論点の説得力を大きく左右する．したがって，これらの情報を論点評価に統合することにより，人間審判に近い評価が可能になると考えた．

3.3　プロトタイプの実装

DbKai は Debatrix の既存実装をもとに上述設計を実現するための部分のみを変更して実装した．変更点 1 と 3 はシステムプロンプトが書かれている yml ファイル群を編集して実現する．Debatrix と異なる観点ジャッジと評価基準を用いるので，既存のシステムプロンプトをすべて書き直す必要がある．各観点ジャッジに与えるプロンプトの概要を次に示す．

- クラッシュジャッジ：主要な対立点に対し反論と反駁がどれほど効果的であったかを評価する．
- 一貫性ジャッジ：同一チームによるスピーチが試合全体を通して論理的に整合しているかを評価する．
- 論点ジャッジ：クラッシュ・一貫性ジャッジの出力を参照しつつ，試合の進行につれ各論点の妥当性と重要性の増減を総合的に比較・評価する．

プロンプトファイル群の変更後のサイズは 20 キロバイト, 英単語にして約 2300 語となる．

変更点 2 はジャッジのフェッチメソッドとメモリの書き込みメソッド，main におけるジャッジ処理の同期管理や勝者情報の取得管理を変更して実現する．変更した行数は全部で 68 行である．

4　検証実験

DbKai の人間審判評価への類似度を調べるために，PanelBench の BP 形式の 22 試合をに対しそれぞれ 3 回独立した判定を行う．人間審判による判定を正解ラベルとし，DbKai の判定との一致率を判定精度として測定する．ただし，対象試合に勝者が 2 チームとなる場合もあるので，判定結果がそのいずれかと一致すれば正解とみなす．比較実験の公平性を保つため，実行環境はの python3.11 とし，ジャッジとして使用する LLM モデルは gpt-3.5-turbo-0125 とし，Debatrix の実験条件と揃える．

また，DbKai と Debatrix では観点ジャッジの構成が異なるので，その影響を評価する補助実験も実施する．具体的には，Debatrix の観点ジャッジを DbKai と同様に変更したモデル Debatrix-Qualitative を用い，同一のベンチマークで精度を検証する．

表 2 には，本研究で評価した 2 種のアーキテクチャ（DbKai，Debatrix-Qualitative）に加え，Liang ら [4] の実験結果を含めた合計 5 構成の比較結果を示す．各構成の概要を次に示す．

1. null[4]
   GPT-3.5 turbo 本体のみで判定を行う．すべてのスピーチを一度に入力し，試合全体に対して総合的に判断するよう設計された専用のシステムプロンプトを用いる．
2. Debatrix[4]
   Debatrix のもとに判定を行う．観点ジャッジは議論，根拠，言語とクラッシュであり，比重は 1:1:1:1 である．
3. Debatrix-Chronological[4]
   Debatrix のもとに判定を行う．ただし観点別に評価せず，各スピーチを時系列順に処理し，総合的に判定する．
4. Debatrix-qualitative
   Debatrix のもとに判定を行う．ただし観点ジャッジは本研究が考案した論点，クラッシュと一貫性とする．比重は 1:1:1 である．
5. DbKai
   本研究が提案するシステムアーキテクチャのもとに判定を行う．観点ジャッジは論点，クラッシュと一貫性である．比重によるスコア合算は行わず，定性的な比較判断に基づいて勝者を決める．

表 2　PanelBench を部分的に用いた検証結果

アーキテクチャ名	精度
null[4]	0.00
Debatrix[4]	51.52
Debatrix-Chornological[4]	30.30
Debatrix-qualitative	48.48
DbKai	69.70

5　考察

表 2 から，DbKai の判定精度は既存の各手法を大きく上回ることがわかり，本手法の有効性が確認された．判定理由の出力を比較すると，ディベートを観点別に，抽象的に評価する Debatrix に対し，DbKai は議論の文脈や反論に注目し，表 3 のように論点を主体とした具体的な評価の出力に成功している．特に，各論点がそれぞれの反論を経て妥当性と重要性がどのように増減しているかが大きく判定に影響している．これは，人間審判の判定プロセスに近似していると言える．また，システムプロンプトの影響を検証するために導入した debatrix-qualitative と元の debatrix の精度に大きな差が見られなかった．これは，本研究の精度向上はシステムプロンプトの設計ではなく，観点間の情報統合処理によるものと考えられる．

しかし，DbKai の判定精度は目標の 80% に達しておらず，改善の余地がある．その要因として，次の 3 つが考えられる．

1. システムアーキテクチャの最適化
   現在のアーキテクチャでは観点間の情報統合は行われているものの，議論の文脈をより構造的に分解し，論点の展開過程を追跡する処理が十分でない可能性がある．
2. システムプロンプトの最適化
   定性的な評価を促すプロンプトは導入されているが，評価観点の解像度や指示の粒度が十分に定義されていない．
3. 議題の種類
   実験に用いた試合の議題の中に「この議会は，知的な宇宙人が存在することを望む」や「この議会は，すべての人が Ubuntu という考え方を強く信じる世界を望む」など価値判断や文化的背景の強い議題が含まれている．このような議題において，アライメントによる制限がある LLM は人間のように一貫性をもつ判断を下すことが困難である．

表 3　DbKai の勝者判定出力の一部

Performance Summary
... 省略...
While OG raises valid concerns about systemic risks and the need for autonomy in currency choices, OO's focus on the benefits of the dollar as a stable and reliable currency carries more weight in terms of persuasiveness in this debate.
... 省略...

6　まとめ

本研究では以下を行った．

1. LLM に基づき人間審判の評価プロセスに倣いディベートを定性的に評価するシステムアーキテクチャの提案
2. 1 の判定精度とその理由付けの観点から評価

その結果，検証実験で示したように，既存モデルと比べより人間審判による評価に近い成果を出すことができた．今後の課題や展望を以下に示す．

- BP 形式ディベート以外の即興型や準備型ディベートに対する有効性の調査
- ジャッジ AI の実用に向けて，LLM の軽量化やディベートジャッジに特化したファインチューニングの実施

謝辞

本研究の一部は JSPS 科研費 23K11359，2025 年度南山大学パッヘ奨励金 I-A-2 の助成を受けた．

参考文献

[1] Lianmin Zheng et al., “Judging llm-as-a-judge with mt-bench and chatbot arena”, Advances in Neural Information Processing Systems 36 (2023).
[2] OpenAI, “GPT-4” (2023). https://www.openai.com/gpt-4
[3] World Universities Debating Championships. https://www.worlddebating.org/
[4] Jingcong Liang et al., “Debatrix: Multi-dimensinal Debate Judge with Iterative Chronological Analysis Based on LLM”, In Findings of ACL 2024, pages 14575–14595 (2024).
[5] 松本 茂, 鈴木 健, 青沼 智, “英語ディベート 理論と実践”, 玉川大学出版部 (2009).
[6] World Universities Debating Council, “World Universities Debating Championships Debating & Judging Manual,” (2023).
[7] OpenAI, “GPT-3.5-turbo” (2022). https://platform.openai.com/docs/models/gpt-3.5-turbo