反論生成における LLM の論理構造準拠の向上を目的とした段階的生成手法の提案

天野祥太朗 [1]　尾崎大晟 [1,5]　内藤昭一 [3]　井之上直也 [2,4]
山口健史 [4]　中川智皓 [1,4]　新谷篤彦 [1]
[1] 大阪公立大学大学院　[2] 北陸先端科学技術大学院大学　[3] 株式会社リコー
[4] 理化学研究所　[5] 株式会社松尾研究所
si24037u@st.omu.ac.jp

概要

大規模言語モデル（LLM）は高品質な反論を生成可能であり，教育応用への高いポテンシャルを持つことが報告されている．ディベート教育を想定した場合，LLM には指定された論理構造に沿った反論を生成する能力が重要である．しかし，論理構造のテンプレートを提示する単純なプロンプト手法では，構造への準拠度が低い，あるいは準拠しても反論の品質が劣化するという課題があった．そこで本研究では，論理構造に基づく反論生成プロセスを複数のステップに明示的に分解し，LLM に段階的な指示を与える手法を提案する．本手法は，生成過程における制御性の向上を目的とする．評価実験の結果，提案手法は従来の一括生成アプローチと比較して，論理構造への準拠度を向上させつつ，品質の劣化を抑制できることが確認された．

1　はじめに

近年，大規模言語モデル（Large Language Models; LLM）の急速な発展により，人間と同等，あるいはそれ以上の品質でテキストを生成することが可能となった．最新のモデルは，複雑な文脈理解や論理的な推論能力を示しており，多様な分野での応用が進んでいる．その中でも，教育分野における LLM の活用は大きな注目を集めており，個別の学習者に合わせたフィードバック生成や，対話を通じた学習支援システムへの導入が期待されている．

とりわけ，批判的思考力[1)]や論理的思考力を育成する「ディベート教育」の文脈において，LLM のポテンシャルは高い．ディベート教育では，特定の論題に対して肯定・否定の立場から論理的に主張を構築し，相手の反論に対して再反論を行うプロセスを通じて能力を養う．この際，学習者が自身の主張に対する適切な反論を受け取り，検討することは学習効果を高める上で不可欠である．高品質な反論を即座に生成できる LLM は，理想的なディベートパートナーとしての役割を担い得る．

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].

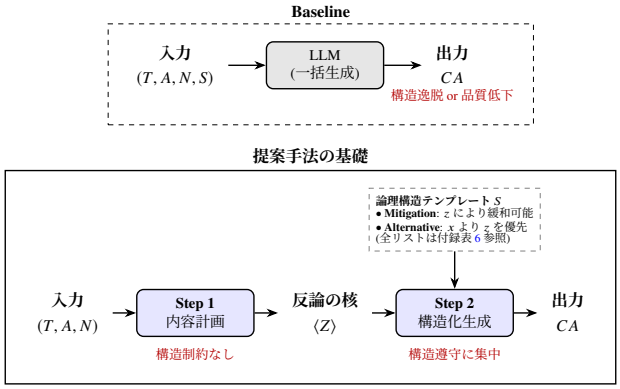

図 1　実験概要．T: Topic（議題），A: Argument（肯定主張），N: Note（$\langle Z \rangle$ を特定するための指針），S: Structure template（論理構造テンプレート），Z: 攻撃点（反論の核），CA: 反論を表す．

2　関連研究

2.1　LLM による反論生成

反論では，相手主張を支える前提を捉え，その弱点を突く反駁が重要である．特に議論文には暗黙の前提が含まれ得るため，反論の説得力は前提理解と攻撃対象の特定に強く依存する．尾崎ら [2] は，LLM による反論生成においても暗黙・批判的前提の特定が反論品質に直結することを指摘している．

一方，教育応用では内容的妥当性に加えて，反論が所望の論理構造テンプレートに安定して従うことが求められる．したがって，構造準拠と内容品質を両立させる生成制御が課題となる．

手法	種別	コール数	要点（入力/共有情報）
Base	1 段階	1	S のみ提示（$\langle Z \rangle$ の指示なし）
1-S	1 段階	1	S+N を提示し内部で $\langle Z \rangle$ 特定と生成を同時実行（反論のみ出力）
1-Z	1 段階	1	1-S に加え $\langle Z \rangle$ を明示出力してから反論生成（明示化の効果）
2-S	2 段階	2	Step1: $(T, A, N) \rightarrow \langle Z \rangle$（+理由），Step2: $(\langle Z \rangle, S) \rightarrow CA$
2-A	2Agent	2	Analyst が $\langle Z \rangle$ 抽出，Debater へ $\langle Z \rangle$ のみ共有して S に沿って生成
2-AR	2Agent	2	2-A に加え $\langle Z \rangle$ 導出理由（Reasoning）も共有（共有情報量の効果）

表 1　比較手法の概要（S: 論理構造テンプレート，N: 指針，$\langle Z \rangle$: 反論の核となる攻撃点）

2.2　生成過程の分解

LLM の出力を安定化させるため，生成・推論過程を複数ステップに分解し，中間表現を言語化させる手法が広く検討されている代表例として，Chain-of-Thought（CoT）[3] や Zero-shot CoT[4] は推論過程の明示化を促し，さらに自己改善・自己洗練（self-refinement）[5] やマルチエージェントによる役割分担 [6] など，生成過程を分解して中間生成を活用する枠組みが提案されているただし，これらは主として正解率や文章品質の改善を目的とするものが多く，出力が特定の論理構造テンプレートに準拠することを保証する設計や，その制御性評価は十分に整理されていない．

2.3　本研究の位置づけ

以上を踏まえ，本研究は反論の核となる要素を中間生成し，生成過程を段階的に分離することで，反論品質の劣化を抑えつつ，所望の論理構造への準拠度と制御性を高める手法を提案する．本手法は，反論の「攻撃点の同定」や一般的な品質改善に加え，教育応用で重要となる「型に沿った反論生成」を生成プロセス設計として扱う点に特徴がある．

3　提案手法

3.1　手法概要: プロセス分離による論理構造制御

本研究の目的は，教育的ディベートにおいて重要となる「論理構造（型）への準拠」と「反論内容の質」を両立させることである．単一プロンプトで両者を同時に満たそうとすると，(i) テンプレート逸脱，あるいは (ii) テンプレートを守る代わりに内容が希薄化する，といった失敗が生じやすい．

そこで本研究では，反論生成を **内容計画** と **構造化生成** の 2 段階に分離する．まず構造的制約を課さずに，反論の核となる攻撃点（以下 $\langle Z \rangle$）を特定する．次に，確定した $\langle Z \rangle$ を入力として与え，論理構造テンプレートへ埋め込むことに集中させて反論文を生成する．この分離により，後段でテンプレート準拠を強めても内容の質を保ちやすくなる．

なお，本研究で用いる論理構造テンプレート S は，内藤ら [7] によって提案された反論構造モデルに基づいている．

3.2　タスク定義

入力はトピック T，肯定側主張 A，反論が従うべき論理構造テンプレート S，および $\langle Z \rangle$ を特定するための指針（Note）N である．モデルは，N に合致する $\langle Z \rangle$ を含み，かつ S の構造要件を満たす反論テキスト CA を生成する．

3.3　比較手法の設計

プロセス分解（1 段階/2 段階）および役割分担（単一/2 エージェント），さらに $\langle Z \rangle$ と導出理由の明示・共有が，構造準拠度と品質に与える影響を分析するため，表 1 の 6 手法を比較する．

4　実験設定

4.1　実験条件とモデル

提案手法の有効性と汎用性を検証するため，30 の多様なディベート議題を用意した．各議題に対し，第 3 節で定義した 6 つの生成手法を適用する．

実験に使用する LLM として，OpenAI モデル (gpt-3.5-turbo, gpt-4.1-mini, gpt-4.1) および Qwen2.5 シリーズ (7b/32b/72b-instruct) の計 6 モデルを選定した。

また，特定の論理構造に対する依存性を排除し，多様な反論の型に対する制御性を検証するため，10 種類の異なる論理構造テンプレートを用意した．各条件（議題・手法・モデル）につき，これら 10 種のテンプレートをすべて適用して生成を行い，総生成数は 10,800 件（30 議題 × 6 手法 × 6 モデル × 10 テンプレート）となる．

4.2 人手評価の設計とデータセット構築

論理構造の厳密な評価，特に反論における「主張・理由・論拠」の整合性判定は，既存の自動評価指標や LLM による自動採点では依然として困難な課題である．そこで本研究では，評価の信頼性と網羅性を担保し，自動評価との整合性を検証するための正解データセットを構築することを目的として，5 名の評価者による人手評価を実施した．

全生成データ（10,800 件）からの層別サンプリングにより，以下の 2 種類の評価データセットを構築した．アノテーションの信頼性を担保するため，1 サンプルにつきランダムに選定された 3 名の評価者が独立して評価を行い，その多数決を採用した．なお，本評価に先立ち，別途用意した議題を用いて評価者間のキャリブレーションを実施した．

4.2.1 データセット A：モデル間比較

手法およびモデル間の性能差を比較するため，30 議題の中から代表的な 1 議題を固定し，全条件（6 手法 × 6 モデル）の出力を抽出した．さらに，各条件について 10 種類の論理構造テンプレートをすべて適用した出力を評価対象とし，合計 360 件（36 条件 × 10 テンプレート）をデータセット A として構築した．これにより，議題の内容や難易度によるバイアスを排除し，生成手法の違いが論理構造の準拠度や反論の質に与える影響を直接的に比較・分析することを目的とする．

4.2.2 データセット B：汎化性能検証

特定の議題に過学習していないか，あるいは議題の性質によらず手法が有効であるかを検証するため，残りの 29 議題からデータセット A と同様に 360 件を抽出した．ここでは，多様なトピックに対する提案手法の頑健性と汎化能力を評価することを目的とする．

4.3 評価指標

本実験では，生成された文章の流暢さや内容の精度よりも，ユーザーが設計した論理構造を，忠実に実行できたかを最優先の評価項目とする．一方，反論としての品質評価や論理の整合性については，構造制約による著しい劣化が生じていないかを検証するため，データセット A およびデータセット B の双方において人手評価を実施する．

1. **論理構造準拠度**: システムがテンプレートの構造制約を遵守しているかを判定する．大規模な検証のため全生成データ（10,800 件）に対して自動評価を行うとともに，上記セット A・B に対しては人手による判定を行い，自動評価の妥当性を検証する．
2. **反論の品質**: 構造制約による著しい内容の劣化が生じていないかを確認するため，人手評価セット A・B を用いて「論理的一貫性」および「反論としての適切性」を評価する．

5 実験結果

本節では，自動評価および人手評価によって得られた，論理構造への準拠と反論の品質に関する分析結果を報告する．なお，準拠スコアは 0-1 の範囲で高いほど望ましく，品質スコア（論理的一貫性・適切性）は 1-3 の範囲で高いほど望ましい．

5.1 評価者間一致率

評価の信頼性推定として Gwet の AC1 を用いた [8]．準拠判定の AC1 は 0.675 であったため，人手評価は 3 名の多数決を採用した．一方，品質（論理的一貫性・適切性）の AC1 は 0.241，0.256 と低かったため，MACE[9] により真値推定を行った．また，全生成データへの適用に向けて自動評価を検証したところ，データセット A における人手（多数決）との一致は正解率 0.783，AC1 0.705 であった．以降では自動評価結果も併記する．

5.2 論理構造への準拠

表 2 に，データセット A（モデル間比較セット）における各手法の論理構造準拠スコアを示す．全体として，Baseline と比較して提案手法は準拠が高い傾向を示し，特に Baseline で準拠が低下する条件において改善が顕著である．例えば Qwen2.5-72b では，Baseline が 0.60 と低い一方で，1-S/1-Z/2-S/2-A はいずれも 1.00（完全準拠）に達している．

一方で，全ての条件で一様に改善するわけではなく，例えば gpt-3.5 では 1-Z が 0.60 となり Baseline（0.70）を下回るなど，手法とモデルの組合せによっては準拠が低下する事例も確認された（表 2）．以上より，生成プロセスを段階化し，論理構造に関する制約処理を明示的に分離する設計は，少なくとも本実験で扱った複数モデルにおいて準拠を安定化させる可能性が示唆される．加えて，全生成デー

表 2　論理構造への準拠スコア（データセット A）

Model	Base	1-S	1-Z	2-S	2-A	2-AR
gpt-3.5-turbo	0.70	**1.00**	0.60	0.80	0.70	0.50
gpt-4.1-mini	0.90	**1.00**	**1.00**	**1.00**	**1.00**	0.90
gpt-4.1	**1.00**	0.90	**1.00**	**1.00**	0.90	**1.00**
Qwen2.5-7b	0.90	**1.00**	0.90	0.80	0.70	**1.00**
Qwen2.5-32b	**1.00**	0.90	**1.00**	0.80	**1.00**	**1.00**
Qwen2.5-72b	0.60	**1.00**	**1.00**	**1.00**	**1.00**	0.80

表 3　自動評価による論理構造準拠スコア

Model	Base	1-S	1-Z	2-S	2-A	2-AR
gpt-3.5-turbo	0.59	0.74	0.62	**0.76**	0.68	0.70
gpt-4.1-mini	0.75	0.88	**0.89**	0.88	0.83	0.86
gpt-4.1	0.85	0.87	**0.91**	0.88	0.80	0.84
Qwen2.5-7b	0.80	**0.89**	0.80	0.62	0.70	0.76
Qwen2.5-32b	0.71	0.87	**0.89**	0.82	0.75	0.78
Qwen2.5-72b	0.62	**0.87**	**0.87**	0.81	0.69	0.77

タ（10,800 件）に対する自動評価結果を表 3 に示す．全体傾向として，人手評価（表 2）と同様に，提案手法は Baseline より高い準拠を示し，大規模データにおいても段階化の有効性が概ね確認された．一方で，モデル・手法の組合せによっては差異も見られ，例えば Qwen2.5-7b では 2-S が 0.62 と低下している（表 3）．なお，自動評価が判定するのは論理構造テンプレートへの準拠であり，反論内容の妥当性や説得力そのものは評価しない点に注意が必要である．

5.3　反論の品質評価

次に，論理構造への準拠だけでなく，生成された反論の内容品質について検証する．表 4 に「論理的一貫性」，表 5 に「適切性」のスコアを示す．

品質面では，準拠ほど一様な改善は見られず，モデルおよび手法により挙動が異なる．例えば gpt-3.5 では，論理的一貫性スコアにおいて Baseline（2.29）に対して 1-S が 1.80 へ低下しており，単に構造制約を強めるだけでは内容品質が損なわれる場合があることがわかる（表 4）．

一方で，反論の核となる要素 Z を明示する設計（1-Z）や，エージェント間で情報を整理して共有する設計（2-A/2-AR）は，品質向上に寄与する傾向が確認された．例えば Qwen2.5-7b では，適切性スコアが Baseline の 1.78 から 1-Z の 2.44 へ改善している（表 5）．また gpt-4.1 では，2-A により論理的一貫性スコアが 1.50 から 2.78 へ，適切性スコアが 2.10 から 2.67 へと改善しており，論理構造を維持しつつ説得力の高い反論が生成されることが示唆される．

以上より，段階化は論理構造準拠の改善に有効である一方，品質を安定して向上させるには，Z の明示や（マルチエージェントの場合）共有情報の粒度設計が重要であるといえる．

表 4　論理的一貫性スコア（データセット A）

Model	Base	1-S	1-Z	2-S	2-A	2-AR
gpt-3.5-turbo	**2.29**	1.80	2.17	1.88	**2.29**	2.20
gpt-4.1-mini	2.22	2.20	1.90	2.10	2.40	**2.67**
gpt-4.1	1.50	2.22	2.40	1.90	**2.78**	2.30
Qwen2.5-7b	1.44	1.30	2.44	2.00	2.43	**2.50**
Qwen2.5-32b	2.00	2.11	**2.90**	2.50	2.50	2.70
Qwen2.5-72b	2.33	2.10	2.10	**2.70**	2.00	2.25

表 5　適切性スコア（データセット A）

Model	Base	1-S	1-Z	2-S	2-A	2-AR
gpt-3.5-turbo	2.43	**2.50**	2.17	1.88	1.86	2.20
gpt-4.1-mini	**2.67**	2.50	2.60	2.30	2.00	2.44
gpt-4.1	2.10	2.44	2.60	1.90	**2.67**	2.20
Qwen2.5-7b	1.78	1.60	**2.44**	2.00	2.00	2.30
Qwen2.5-32b	2.50	2.33	**2.80**	1.88	2.30	2.00
Qwen2.5-72b	2.33	**2.60**	2.10	2.40	2.00	2.25

5.4　汎化性能の検証（データセット B）

多様な議題に対する提案手法の頑健性を検証するため，データセット B（汎化性能検証セット）における評価を実施した．詳細な結果は付録に示す．

分析の結果，データセット A と同様に提案手法の有効性が確認されたが，Baseline の挙動が不安定であった．データセット A では Qwen2.5-72b の準拠が 0.60 と低迷したのに対し，データセット B では Qwen2.5-32b の Baseline が 0.60 まで低下しており，議題や相性によって Baseline の性能が大きく変動することが確認された（付録表 7）．これに対し，提案手法（特に 2-S や 2-AR）は，両データセットにおいて概ね高い準拠と品質を維持している．一方で，一部条件では手法により準拠が低下する例も見られるため（付録表 7），共有情報の設計が安定性に与える影響については今後の検討課題である．

これらの結果は，提案手法が特定のデータセットに過度に依存せず，論理構造の制御に関して一定の汎用性を有する可能性を示唆している．

6　結論

本研究では，反論生成を「攻撃点の特定」と「構造化生成」に分離する段階的生成手法を提案した．多様なモデルを用いた評価実験の結果，提案手法は一括生成に比べて論理構造への準拠を改善し，かつ反論品質の劣化を抑制できることを確認した．

謝辞

本研究は JSPS 科研費 22H00524 の助成を受けたものです.

参考文献

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

[2] Taisei Ozaki, Chihiro Nakagawa, Naoya Inoue, Shoichi Naito, and Kenshi Yamaguchi. LLM DEBATE OPPONENT : Counter-argument generation focusing on implicit and critical premises. In Abteen Ebrahimi, Samar Haider, Emmy Liu, Sammar Haider, Maria Leonor Pacheco, and Shira Wein, editors, **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)**, pp. 456–465, Albuquerque, USA, April 2025. Association for Computational Linguistics.

[3] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[4] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

[5] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

[6] Li Zhang and Kevin D. Ashley. Mitigating manipulation and enhancing persuasion: A reflective multi-agent approach for legal argument generation, 2025.

[7] Shoichi Naito, Wenzhi Wang, Paul Reisert, Naoya Inoue, Camélia Guerraoui, Kenshi Yamaguchi, Jungmin Choi, Irfan Robbani, Surawat Pothong, and Kentaro Inui. Designing logic pattern templates for counter-argument logical structure analysis. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 11313–11331, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[8] Werner Vach and Oke Gerke. Gwet's ac1 is not a substitute for cohen's kappa – a comparison of basic properties. **MethodsX**, Vol. 10, p. 102212, 2023.

[9] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1120–1130, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

A 付録 (Appendix)

表 6 反論文テンプレート.

テンプレート
T1: 緩和策 [x] promotes the bad outcome [y], but only to a limited degree. There exits a methods <Z> that mitigates the promotion of [y].
T2: 代替案 [x] promotes the bad outcome [y], but only to a limited degree. Another factor <Z> also promotes [y], and it is better to deal with <Z>.
T3: 根拠不十分 [x] does not promote [y]. There is not sufficient evidence that [x] promotes [y].
T4: 別の真因 [x] does not promote [y]. Another true factor <Z> that promotes [y] exists or appears.
T5: 見落とされた機序 #1 [x] does not promote [y]. [x], on the contrary, suppresses [y]. [x] promotes factor <Z>, which suppresses [y].
T6: 見落とされた機序 #2 [x] does not promote [y]. [x], on the contrary, suppresses [y]. [x] suppresses factor <Z>, which promotes [y].
T7: 対処不要 [y] has no value or is sufficient and does not need to be addressed.
T8: y による好影響 [y] is not a bad outcome. [y] causes a good outcome <Z>.
T9: y とは別観点のメリット #1 [x] promotes a good outcome <Z> from a different point of view than [y].
T10: y とは別観点のメリット #2 [x] suppresses a bad result <Z> from a different point of view than [y].

表 7 論理構造追従性スコア（データセット B）

Model	Base	1-S	1-Z	2-S	2-A	2-AR
gpt-3.5-turbo	1.00	1.00	0.70	1.00	0.90	1.00
gpt-4.1-mini	1.00	1.00	0.90	1.00	1.00	1.00
gpt-4.1	0.90	1.00	1.00	1.00	1.00	1.00
Qwen2.5-7b	1.00	1.00	1.00	1.00	0.90	1.00
Qwen2.5-32b	0.60	**1.00**	0.90	**1.00**	**1.00**	**1.00**
Qwen2.5-72b	1.00	0.90	1.00	1.00	0.50	1.00

表 8 論理的一貫性スコア（データセット B）

Model	Base	1-S	1-Z	2-S	2-A	2-AR
gpt-3.5-turbo	**3.00**	2.30	2.29	1.00	2.89	2.70
gpt-4.1-mini	**3.00**	2.40	2.44	2.40	2.10	2.80
gpt-4.1	1.89	**3.00**	2.40	2.50	2.80	2.50
Qwen2.5-7b	2.00	1.90	**3.00**	1.70	2.67	2.30
Qwen2.5-32b	**3.00**	2.90	2.22	**3.00**	2.50	**3.00**
Qwen2.5-72b	**3.00**	1.89	**3.00**	2.50	**3.00**	2.80

表 9 適切性スコア（データセット B）

Model	Base	1-S	1-Z	2-S	2-A	2-AR
gpt-3.5-turbo	**2.70**	2.30	2.00	1.00	2.11	2.30
gpt-4.1-mini	**2.80**	2.30	2.00	2.20	2.00	2.40
gpt-4.1	2.00	2.50	2.30	2.10	**2.60**	2.30
Qwen2.5-7b	2.10	2.10	**2.80**	2.00	2.33	2.10
Qwen2.5-32b	2.17	2.30	2.22	**2.40**	2.20	2.30
Qwen2.5-72b	**2.70**	2.00	2.60	2.10	2.00	2.50