卒業論文

LLMマルチエージェント弁証法的推論による質問応答

指導教官　　村上 陽平 教授

立命館大学情報理工学部
先端社会デザインコース

増尾 柚希

2026年1月1日

LLM マルチエージェント弁証法的推論による質問応答

増尾 柚希

内容梗概

大規模言語モデル（LLM）を用いた質問応答システムは，さまざまなアプリケーション，システムで実用化が進んでいる．そして近年では,LLM に複数の価値観や立場を付与したマルチエージェント対話が提案され, 模擬的な議論や意思決定を生成する技術が登場している. 多文化共生社会においては, 質問応答は単なる情報検索にとどまらず, 異なる文化的・宗教的・倫理的背景をもつ人々の相互理解を促進する役割が求められている．しかしながら, 既存の LLM ベースのマルチエージェント対話は, 対立する立場をもとに勝敗を決めるディベート的議論が多く，最終的な結論も表層的な折衷案の提示に留まる．たとえば,「学校清掃を生徒が行うべきか職員が行うべきか」といったトピックにおいて, 現状の LLM は「日常清掃は生徒, 専門清掃は職員」といった妥協案を返す傾向があるが, これは価値観の本質的な対立を踏まえ，清掃を教育活動として再定義した生徒と職員の共同清掃のような止揚的な解決の回答生成は依然として実現できていない．そこで，本研究では，多様な価値観の対立を扱うトピックを対象に，LLM を用いたマルチエージェント対話によって新たな価値・制度・概念の再構成（止揚）を導く弁証法的な対話プロトコルを提案する．具体的には，kido, kurihara[1] らの記号論理に基づく既存の弁証法議論，および統合のアルゴリズムの手続きを参照し，その形式的プロセスを LLM マルチエージェントが自然言語で模倣できるように再構成した段階的対話モデルを設計する．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

議論の収束

質問応答のシステムに適用される推論のプロトコルは，その推論過程の終了を保証できなければばならない．こと弁証法的対話プロトコルにおいては，エージェント同士の議論を収束させることで，推論過程の終了を保証する．記号論理を用いて表現されたエージェントの知識ベースや，それらを用いた論理は手動で定義されるため，有限であり，同じ論理の再使用を許可しない限り，それらを用いて行う議論は必ず収束する．しかし，自然言語を用いて議論を行う LLM のエージェント同士の議論では，意味的に同義である論理の再使用を制限する必要があり，その制御はプロンプトによる指示だけでは難しく，議論の収束を

保証できない.

高次な止揚論証の生成

統合（止揚）は折衷ではなく価値の再構成による高次な概念の創出を意味するが，その成立条件及び構築アルゴリズムは自然言語では明確化されておらず，形式的定義だけでは意味的統合として妥当かを判断できない．また，kido, kurihara[1] の統合アルゴリズムでは，入力としてお互いに defeated された論証のみを入力としていたが，LLM では議論の収束が難しく，議論に決着をつけることがほとんどできないため，統合アルゴリズムをそのまま適用することは実用的に困難である．

本研究では，以上の課題を解決するために，プロトコル内で反復的にエージェントの議論が行われる「反論」フェーズにおいて，同一内容の議論をプロンプトで制限すると共に，反論回数の上限を設けた．上限に達した議論はその時点で終了し，その時点までの対話から対立軸を対話エージェントによって抽出する．これにより，お互いの議論に決着がついていなくても，対立軸の抽出から統合アルゴリズムを適用することができる．止揚論証の構築においては，kido, kurihara[1] らの統合アルゴリズムをプロンプトに落とし込み，入力された論証の特化（Specialization），汎化（Generalization）を行うことで，対立する立場の主張を 1 つに統合することができた．加えて本研究では，提案プロトコル組み込んだ弁証法的推論による質問応答システムを実装し，kido, kurihara[1] らの止揚論証の例との比較を行った．本研究の貢献は以下の通りである．

議論の収束

実装したシステムを用いて 100 個のトピックについて議論を行わせ，対話プロトコルは必ず収束することを確認した.

高次な止揚論証の生成

実装した質問応答システムを用いて kido, kurihara[1] らの対話例を実際に行い，最終的な止揚論証が意味的に一致することを確認した．その上で Anghel[2] らのルーブリック評価を行い,「独創性」,「一貫性」,「弁証法性」において，既存システムよりも高いスコアを示した.

LLM Multi-agent Dialectical Reasoning for Question Answering

Yuzuki MASUO

Abstract

LLM マルチエージェント弁証法的推論による質問応答

目次

第 1 章　はじめに　1

第 2 章　関連研究　4

2.1　記号論理に基づく弁証法的議論とその限界 4

2.2　大規模言語モデルによる弁証法的議論とその限界 5

第 3 章　弁証法的推論による質問応答システム　7

3.1　概要 7

3.2　システム構成 7

3.2.1　入力 8

3.2.2　議論制御 8

3.2.3　LLM Agents 8

3.2.4　出力 9

3.3　弁証法プロトコル 9

3.3.1　概要 9

3.3.2　論証構築 11

3.3.3　反論 11

3.3.4　統合 12

第 4 章　評価手法　14

4.1　評価方針 14

4.2　評価データ 14

4.3　実験設定に関する補足 15

4.4　評価方法 16

4.4.1　既存研究との意味的比較 16

4.4.2　自然言語表現の確認（補助的評価） 16

第 5 章　評価結果　17

5.1　評価結果の概要 17

5.1.1　既存研究との比較結果 17

5.1.2　ルーブリック評価結果 18

第 6 章　考察 **19**
6.1　考察 19
6.2　展望 19
第 7 章　おわりに **20**
謝辞 **21**
参考文献 **22**

第1章　はじめに

近年，大規模言語モデル（LLM）を用いた質問応答は，さまざまなアプリケーション，システムで実用化が進んでいる．同時に世界では，多文化共生がますます進んでおり，さまざまな価値観を持つ人々が共に暮らす社会ができている．質問応答には情報検索としての役割に加えて，そのような社会に適応するために，異なる価値観を持つ人々の相互理解を促進する，合意可能な回答を生成する役割も求められている．この要請に対し，複数の役割や価値観を付与したLLMエージェント同士の対話により結論を導くマルチエージェント議論の枠組みが提案されている．しかし既存手法の多くは，対立する立場をもとに勝敗を決めるディベート的議論，あるいは双方の主張を並列した表層的な折衷案に収束しやすい．その結果，対立を保持したまま共通項を抽出し，新たな価値・制度・概念の再構成として解決を導く止揚的な回答生成には至りにくいという課題がある．

そこで本研究では，記号論理で形式化された弁証法的推論の枠組みに基づき，複数のLLMエージェントが論証構築・相互反論・統合を段階的に実行する対話プロトコルを提案する．具体的には，Kido, Kurihara らの弁証法議論形式 [1] を参照し，記号論理で示された対話プロトコルと統合の手法を，2人のLLMエージェントを用いた弁証法的推論による質問応答システムに適用するために拡張する．プロトコルはフェーズは論証構築フェーズ，反論フェーズ，統合フェーズから構成され，論証構築とそれに対する反論は，2人の対話エージェントによって相互に行われる．エージェントの論証は結論（Consequent）と，それを導くための前提（Premises）によって成り立つ．そしてこのプロトコルを適用したシステムは，エージェントの論証の状態や，プロトコルの制御を行う．具体的に次のようにプロトコルは遷移する．

論証構築フェーズではまず片方のエージェント（AG1）が主論証（main argument）を構築し，反論フェーズでは相手のエージェント（AG2）が反論可能かを判断し，可能であれば相手の論証に対する反論（defeating argument）を構築する．反論を行うことができない場合，AG1 の main argument は正当化され（main argument justified），システムは終了する．AG2 が反論できる場合，AG2 の反論後，AG1 は同様に再反論の可否判断を行い，可能であれば相手の反論に対する再反論を構築する．AG1 が再反論を行うことができない場合，AG1

の main argument は敗北する（main argumnet defeated）．この反論・再反論を 1 回のエポックとし，システムで設定された最大反論回数に達するまで繰り返し行われる．最大反論回数に達しても main argument が justified・defeated どちらかの状態に遷移していない場合，その議論は保留状態（pending）に遷移する．AG1 の main argument の状態が defeated，もしくは議論が pending の状態の時，次に AG2 の論証構築フェーズに遷移し，同様の議論を行う．お互いの main argument に関して議論が行われると、kido，kurihara らの Specialization，Generalization を用いつつ，以下のように工夫を行うことで統合を行う．

本来 kido, kurihar の統合アルゴリズムは，お互いの main argument がどちらも defeated されている論証のペアに対して適用されるが，本研究では pending の議論に対しても統合アルゴリズムを適用することにした．理由は 2 つあり，1 つは LLM エージェントによる議論の性質上，議論の決着がつかず，pending の状態で統合フェーズに進んでしまうことである．2 つ目は，実世界での議論を考慮した時，議論の決着がつかないことは容易に考えられ，むしろその議論の中から対立点を抽出し，統合することが，本来の議論を通じた止揚を達成する弁証法的推論と考えられるからである．

では統合の手法だが，以上の状態の議論を kido, kurihara の統合アルゴリズムの入力にするために，AG1 は相互の議論収束後，これまでの議論を振り返り，それぞれの議論の対立軸を明らかにし，アルゴリズムの入力となるようにそれぞれの AG の論証を再構築する．この再構築は既存手法の Specialization にあたり，お互いのエージェントの立場でも受け入れられる条件を付与した形の論証を構築する．Generalization では再構築した論証のペアに対し，それらが共に共存する制度や，社会的な背景を提案する．これらの操作により，各エージェントの主張が議論を通してお互いの価値観を理解した上で妥協案を提案し，それらが共存する高次な解決策を生成する．

本研究の実装上の課題は，(1）議論の収束と，(2）高次な止揚論証の生成である．課題（1）に対しては，反論フェーズにおいて反論の最大回数を設定することによって、議論の収束を保証した．課題（2）に対しては，kido, kuriara ら [1] の Specialization，Generalization をプロンプトに落とし込むことで，記号論理で定義された既存手法と同様の止揚論証を生成できることを確認した．以降では，関連研究と課題設定を整理した上で，システムのアーキテクチャと提案プロトコルの詳細設計を述べ，評価と考察を通じて有効性と限界を明らかにし，最

後に結論と今後の課題をまとめる.

第2章　関連研究

2.1　記号論理に基づく弁証法的議論とその限界

対立する主張を形式的に扱う枠組みとして，弁証法議論に関する研究が行われてきた．

Sawamura，Umeda ら [2][3] は，Prakken, Sartor ら [4] の拡張論理プログラミングに基づき，弁証法議論の枠組み（Computational Dialectics） を提案している．この枠組みでは論証の構築と，相手の論証に対する攻撃（rebut, undercut）を段階的に行い，複数のエージェントが対話を通じて弁証法的に統合を行う議論フレームワークを形式定義している．ただし Sawamura，Umeda らは文献の中で，具体的な止揚論証の導出方法は神託（a sort of oracle）であると述べており、形式定義される止揚論証の意味的な実態や，それらの構築方法は未実装であった．

kido, kurihara ら [1] はこれらの課題を解決するため，形式定義された止揚論証を，Specialization 及び Generalization を用いて構築する手法を提案した．これにより，弁証法的な対話を通じて，止揚を行う形式的な操作が明らかになった．しかしながら，これらの文献における論証の表現は記号論理を用いて行われており，自然言語の解釈を行うことは困難である．特に，高度な止揚論証を記号論理で導出しても，それが実際自然言語でどう表され，どのような意味を持つのかを表現することは，ユーザーの解釈を必要とし，未だ課題である．

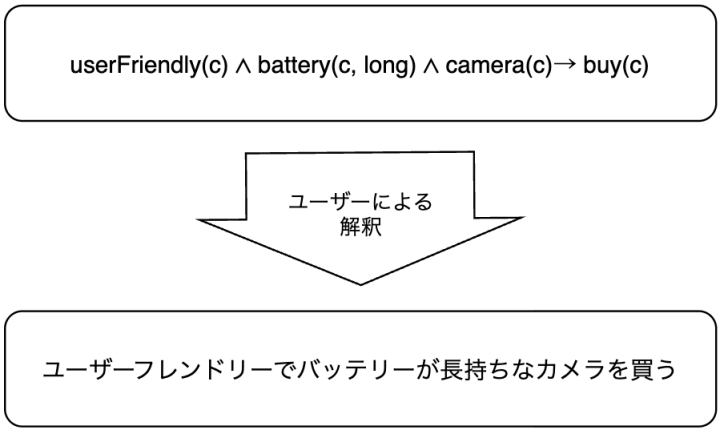

図 1: 記号論理で表された止揚論証の解釈

2.2　大規模言語モデルによる弁証法的議論とその限界

そこで近年では，大規模言語モデル（LLM）を複数のエージェントに分担させ，異なる視点から推論を進めるマルチエージェント対話が注目されている．

Anghel ら [5] は，複数の LLM を用いて弁証法的推論を模倣する枠組みを提案し，(1) 主張（thesis）の生成，(2) 反論（antithesis）の生成，(3) 統合（synthesis）の生成という三段階の推論フローを設計している．同手法では，主張生成エージェントと反論生成エージェントを分離し，さらに生成された推論過程や最終的な統合結果に対して，別の評価エージェントがルーブリックに基づく評価を行う点に特徴がある．

一方で，このような LLM を用いた弁証法的推論では，対立する立場を明確に定義すること自体が容易ではないという課題がある．例えば，賛否を問う形式のトピックにおいても，立場の違いが価値の否定としてではなく，責任の範囲や強度の差として解釈される場合が多く，明確な対立軸を構成しにくい．その結果，立場間の関係は「賛成か反対か」という二値的な整理に還元されやすく，弁証法的議論として重要な，互いに排他的な主張同士の衝突を前提とした対話を十分に構成できない場合がある．

また，反論生成の過程においても，反対の結論を導くための前提が列挙される一方で，それらが主張側のどの前提を否定しているのかが明示されないことが多い．このような場合，反論は独立した懸念の提示にとどまり，前提間の衝突関係や攻撃対象が不明確なままとなる．そのため，多ターンにわたる相互反論を通じて争点を掘り下げることが難しく，議論の過程が対話的に発展しているかどうかを判断することも困難となる．

さらに，このような議論構造の不明確さは，最終的に生成される統合（synthesis）にも影響を及ぼす．すなわち，どの反論がどの理由で却下され，どの論点が保持されたのかを，議論過程から再構成することが難しく，止揚がどのような根拠に基づいて成立しているのかが不透明になりやすい．Anghel らも，固定的な三段フローに基づく弁証法的推論には限界があることを指摘しており，より自然な多ターンの対話的交換へ拡張する必要性を述べている．

以上より，止揚（統合）を弁証法的に妥当な形で生成するためには，統合に先立って多ターンの議論を通じて争点を明確化し，どの前提がどのように衝突しているのかを構造的に扱う枠組みが必要である．次章ではこの課題に対し，議

論フェーズと統合フェーズを明確に分離し，対立点の明確化を経て止揚論証へ接続するプロトコルと，それを適用するシステムのアーキテクチャについて述べる．

第3章　弁証法的推論による質問応答システム

3.1　概要

本システムは，ユーザから与えられたトピックを入力とし，複数のLLMエージェントによる対話的推論を通じて，最終的に止揚論証（統合された回答）を生成するマルチエージェント推論システムである．本研究では，弁証法的推論の規則や進行手順を「弁証法プロトコル」として抽象化し，それを実行するための計算基盤としてシステムアーキテクチャを設計する．本章では，提案プロトコルの内容には立ち入らず，システムを構成する各コンポーネントと，それらの間の情報の流れに焦点を当てる．

3.2　システム構成

本システムは，図2に示すように，入力，議論制御，弁証法プロトコル，状態 / 動作，LLM Agents，出力の6つの主要コンポーネントから構成される．各コンポーネントは明確な責務分離に基づいて設計されており，議論プロトコルの変更や拡張が，システム全体の構造に影響を与えにくい構成となっている．

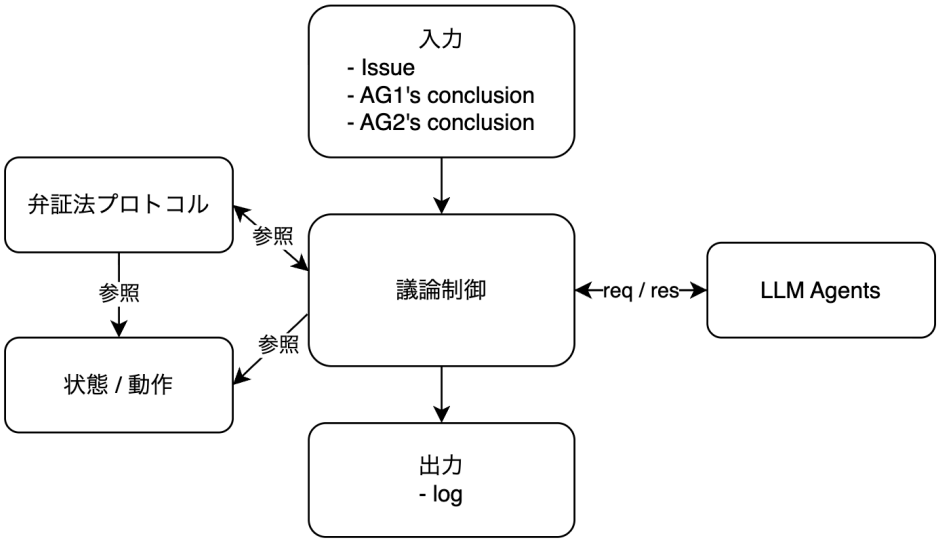

図 2: 提案システムの全体アーキテクチャ

3.2.1 入力

入力コンポーネントは，ユーザから与えられる議論トピックや初期条件を受け取り，議論制御コンポーネントに引き渡す役割を担う．

入力には，以下の 4 つが含まれる．

- Issue（議論対象となるトピック）
- AG1's Conc（エージェント 1 の結論）
- AG2's Conc（エージェント 2 の結論）

Issue はエージェントがそれぞれの立場に基づいて議論を行うためのトピックである．AG1's Conc，AG2's Conc はトピックに対するエージェントのスタンスである．本研究では弁証法的に議論を行い，さまざまなトピックに対して統合を達成する対話プロトコルを提案することがゴールであるため，エージェントのトレーニング・チューニングなど Issue 依存の操作は行わないものとする．したがって，対立はユーザーの入力によって意図的に形成しつつ，結論を導く論理や前提はエージェントに推論させることで，多角的な議論を促進する．

以下に本システムの入力例を示す

- Issue: 「学校清掃は生徒が行うべきか」
- AG1's Conc: 「学校清掃は生徒がすべき」
- AG2's Conc: 「学校清掃は生徒がすべきでない」

3.2.2 議論制御

議論制御コンポーネントは，弁証法プロトコルに従って議論全体の進行を管理する．このコンポーネントは，入力された Issue，各エージェントの結論を元に弁証法議論を実行し，定義されたフェーズごとの発話を LLM Agents に要求する．適用するプロトコルによって終了状態に遷移したとき，システムは対話の履歴をログに出力し，終了する．

議論の終了状態は以下の 2 つである．

- main argument が正当化されたとき
- AG1's Conc: 「学校清掃は生徒がすべき」
- AG2's Conc: 「学校清掃は生徒がすべきでない」

3.2.3 LLM Agents

LLM Agents は以下のモデルを使用する．

- モデル: GPT-4.1-nano

3.2.4　出力

システムが終了するとき，議論制御コンポーネントによって対話ログと，存在すれば最終回答が出力される．

3.3　弁証法プロトコル

これからは本システムで適用される弁証法プロトコルについて紹介する．本章では，プロトコル全体の概要と，各フェーズへの遷移条件，終了条件，そしてフェーズ内の状態に対応する LLM Agents への指示内容を説明する．

3.3.1　概要

弁証法プロトコルは主に論証構築，反論，統合の 3 つのフェーズから構成され，各エージェントの論証に対して，お互いに反論を述べ合い，最終的に 1 つの止揚論証を構築する．

論証構築では，エージェントはシステムの入力で与えられた各エージェントのスタンスをもとに，それらを形成するための前提を肉付けする．

反論フェーズではその論証に対して相手のエージェントが反論可能かを判断し，可能であれば反論，そうでなければ，最初の論証が正当化されたとして，議論を終了する．反論フェーズでは相手の反論に対する再反論の構築可否も判断させ，最終的に最初の論証が defeated 状態になるか，反論の掛け合いが設定した最大回数に達するまで，各エージェントによる反論は繰り返し行われる。

この議論によってお互いの主張を明らかにし，論証を片方のエージェントがもう一度お互いの main argument を再構築した上で，統合を行う．統合のアルゴリズムは kido, kurihara[1] の Generalization を用いる．プロトコルの全体像を図 2 に示す.

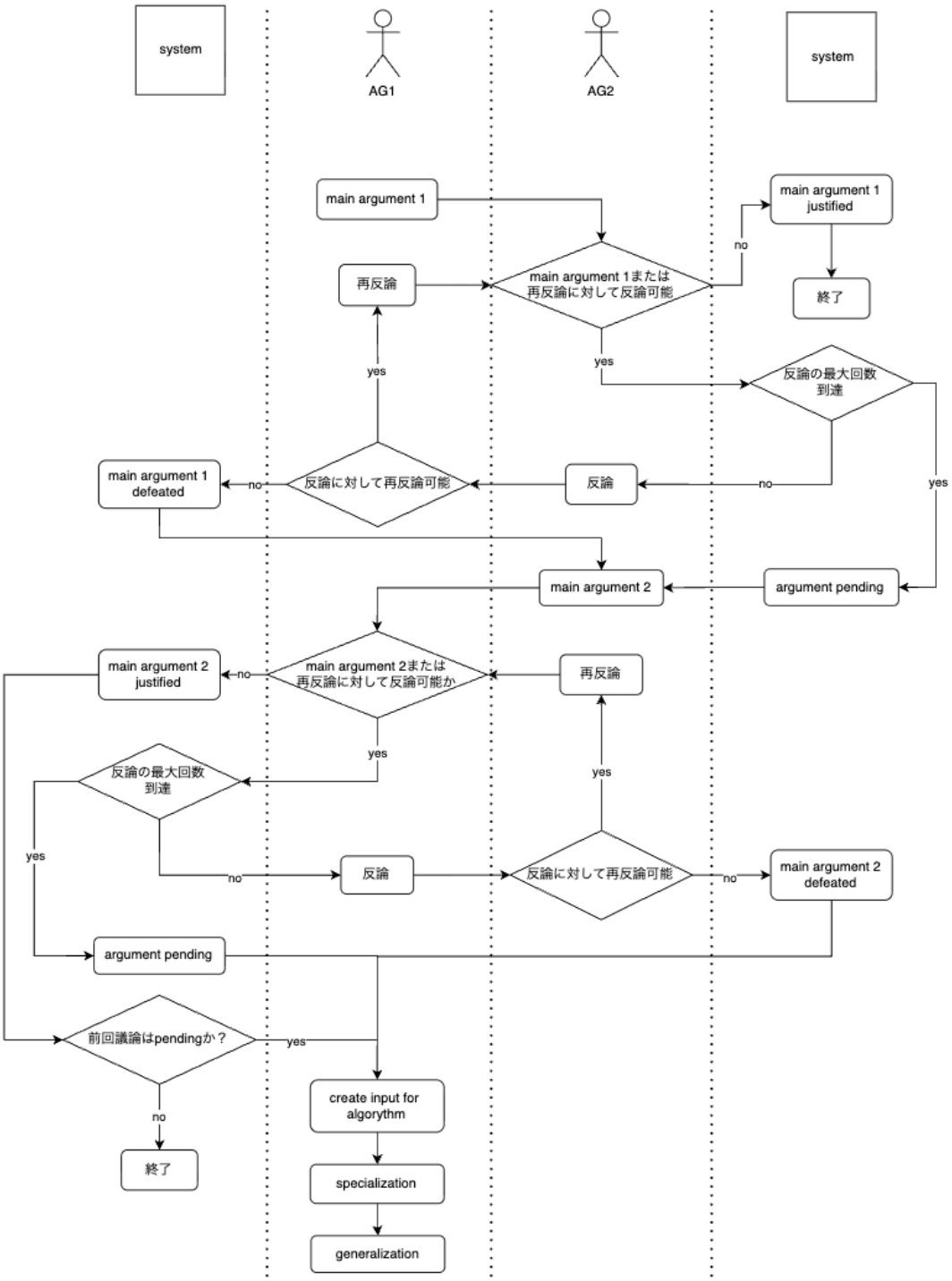
system
AG1
AG2
system
main argument 1
main argument 1 justified
no
再反論
main argument 1または
再反論に対して反論可能
終了
yes
yes
反論の最大回数
到達
main argument 1
defeated
no
反論に対して再反論可能
反論
no
yes
main argument 2
argument pending
main argument 2
justified
no
main argument 2または
再反論に対して反論可能か
再反論
yes
yes
反論の最大回数
到達
yes
no
反論
反論に対して再反論可能
no
main argument 2
defeated
argument pending
前回議論はpendingか？
yes
no
終了
create input for
algorythm
specialization
generalization

図 3: 提案プロトコル

3.3.2 論証構築

論証構築フェーズでは，各 LLM エージェントがトピックに対する自身のスタンスに基づき，初期論証を構築する．本研究では，論証（Argument）を，結論を導出するための前提集合および推論規則（warrant）からなる推論単位として形式化する．Argument は，各エージェントに与えられた理論（事実と strict rules の集合）に基づき構築され，例えば，

$$buy(b) \leftarrow affordable(b) \wedge inStock(b) \wedge camera(b)$$

のように，単一の結論とそれを支持する前提・規則を明示的に含む．本プロトコルでは，このように形式化された Argument を，反論および統合フェーズにおける操作対象とする．

各エージェントは，システムから与えられたスタンスに従い，その立場を支持する結論を明示した上で，当該結論が成立すると考える理由を前提として列挙する．この際，前提は事実記述または推論規則として表現される．本実装では具体的に，以下のようにエージェントに出力させる．

```
[ARGUMENT]
camera(a), compact(a), light(a).
buy(a) <- compact(a) & light(a) & camera(a).
[NATURAL_LANGUAGE]
We should buy a camera because it is compact, light.
```

図 4: 論証のスキーマ

本フェーズでは，相手エージェントの主張や想定される反論は考慮せず，反論や防御，統合的主張の生成は行わない．ここで生成された論証は，後続の反論フェーズにおいて攻撃および防御の対象として用いられる．

3.3.3 反論

反論フェーズでは，論証構築フェーズで提示された main argument に対して，相手エージェントがそれを defeat する Argument（defeating argument）を構築できるかを判定し，可能な場合はそれを提示する．

defeating argument は，相手の主張を成立不能にする Argument であり，結論を直接否定する形（例：$\neg buy(a) \leftarrow outOfStock(a)$）や，前提・規則の適用

を阻止する形（undercut）として表現される．文献の対話例（カメラ選択）では，$outOfStock(a) \rightarrow \neg buy(a)$ や $overTheBudget(b) \rightarrow \neg buy(b)$ のように，結論 $buy(\cdot)$ を否定する Argument により defeat が与えられる．

相手が defeating argument を構築できない場合，当該 main argument は（その時点では）defeat されないため justified とみなし，議論を終了する．一方，双方が新たな main argument を提示できず，かつ main argument が justified されない場合には，対立する主張（特に warrant）を統合対象として抽出し，統合フェーズへ移行する．

なお LLM 実装では，反論の反復が停止しない可能性があるため，実装上の安全策として反論回数に上限を設け，上限到達時は議論を打ち切って統合フェーズへ移行する．

3.3.4　統合

統合フェーズでは，反論フェーズを通じて明らかになった対立構造を入力として，止揚論証（synthesized argument）を生成する．本フェーズの目的は，既存の論証を単に折衷するのではなく，対立する判断根拠（warrant）を再構成することで，より高次の観点から新たな main argument を構築することである．

本研究では，統合の手法として，Kido および Kurihara によって提案された Generalization に基づくアルゴリズムを用いる．具体的には，反論フェーズにおいて相互に defeat 関係にある複数の main argument を参照し，それらに含まれる warrant を統合対象として抽出する．

例えば，以下のような 2 つの warrant が対立している場合を考える．

$$compact(X) \wedge light(X) \wedge camera(X) \rightarrow buy(X)$$

$$affordable(X) \wedge inStock(X) \wedge camera(X) \rightarrow buy(X)$$

Generalization では，これらの warrant に含まれる共通要素（$camera(X)$）と，対立を生じさせている要素（$compact(X) \wedge light(X)$ と $affordable(X) \wedge inStock(X)$）を再構成し，

$$userFriendly(X) \wedge inStock(X) \wedge camera(X) \rightarrow buy(X)$$

のような，新たな warrant を導出する．

今回，統合アルゴリズムについては，参考にしている文献とは少し異なるア

プローチをとる．既存アプローチでは議論フェーズがなく，specialization による妥協的な調整をアルゴリズム上で実施する必要があるが，本研究では議論フェーズを通じて対立点を明確化できるため，specialization の探索的ステップを省略できると考えている．本研究では，generalization アルゴリズム（文献中の Algorithm 4）を適用する前段として，議論を通じて各エージェントの主張を整理するが，Kido and Kurihara が定義する探索的な specialization（Algorithm 3）は明示的には実装していない．代わりに，議論フェーズにおいて対立点が明確化された warrant を Algorithm 4 の入力として直接用いる．

このようにして得られた止揚論証は，反論フェーズではいずれのエージェントからも導出できなかった新規の main argument として位置づけられる．本研究では，この止揚論証を，最終的な統合結果として自然言語に変換し，ユーザに提示する．

第4章　評価手法

本章では，提案する弁証法プロトコルの有効性を検証するための評価手法について述べる．本研究における評価の主目的は，提案プロトコルによって生成される議論過程および止揚論証が，既存の計算的弁証法モデルに基づく形式例と意味的に整合しているかを検証することである．

従来の質問応答システムの評価では，単一の正解を持つタスクに対する正解率が主な指標として用いられることが多い．しかし，弁証法的議論に基づく推論では，複数の立場が相互に反論し合う過程そのものが重要であり，単一の正解を仮定した評価は必ずしも適切ではない．そこで本研究では，既存研究との意味的一致を中心とした評価を行う．

4.1　評価方針

本研究における評価方針は，以下の二点に基づく．

第一に，提案プロトコルによって生成される議論構造が，Kido および Kurihara によって定義された計算的弁証法の形式例と意味的に一致しているかを検証する．ここでは，論証の構築順序，反論関係（defeat 関係）の形成，および相互に正当化されない状態を経た後に新たな main argument が生成される過程に着目する．

第二に，生成される止揚論証が，既存の論証の単なる折衷や再表現ではなく，対立構造を踏まえた新たな判断根拠（warrant）に基づく論証として構成されているかを確認する．

なお，本研究で用いる評価は，提案プロトコルの形式的妥当性および意味的一貫性の検証を目的としており，多様なトピックに対する汎用的性能評価は本章の対象外とする．

4.2　評価データ

評価には，Kido および Kurihara によって提案されたカメラ購買問題に関する対話例（文献 [1] 第 6 章）を用いる．この対話例では，複数のカメラ候補に対して異なる評価基準が与えられ，それらが相互に反論関係を形成する典型的な弁証法的対立構造が示されている．

文献中では，各エージェントが順に main argument を生成し，それらが相互

に defeat 関係を形成することで，いずれの論証も正当化されない状態に到達した後，既存の論証の単なる組み合わせではない新たな main argument が生成される．本研究では，この一連の推論過程を評価基準として用いる．

4.3 実験設定に関する補足

提案システムの本来の設計では，エージェントはユーザから与えられたスタンスのみに基づき，前提および論証を自律的に構築することを想定している．

一方，本研究の評価実験では，既存研究との意味的整合性を厳密に検証するため，各エージェントに理論（事実および strict rules からなる集合）を明示的に入力として与えている．これは，文献に示された形式例を忠実にトレースし，提案プロトコルが計算的弁証法の要請を満たしているかを確認するための実験的手法である．

したがって，理論を入力として与える本設定は，提案システムの前提条件ではなく，あくまで検証のための補助的手段である．

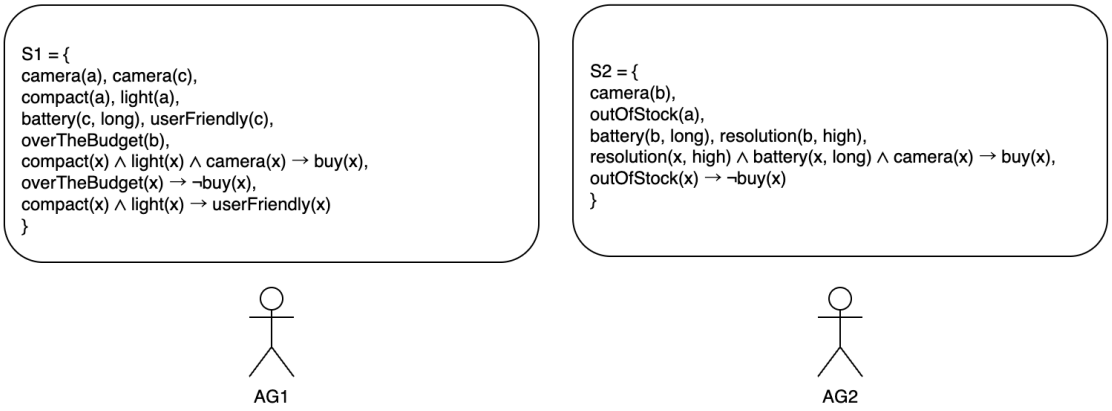

図 5: 各エージェントの理論

図中の S_1 は AG1 が論証構築に用いる理論であり，事実として $camera(a)$, $camera(c)$, $compact(a)$, $light(a)$, $battery(c, long)$, $userFriendly(c)$, $overTheBudget(b)$ を含む．規則は $compact(x) \wedge light(x) \wedge camera(x) \rightarrow buy(x)$, $overTheBudget(x) \rightarrow \neg buy(x)$, $compact(x) \wedge light(x) \rightarrow userFriendly(x)$ であり，「コンパクトで軽いカメラを買う」ことを主張できる構成にしている．同様に，S_2 は AG2 の理論であり，事実として $camera(b)$, $outOfStock(a)$, $battery(b, long)$, $resolution(b, high)$

を含む．規則は $resolution(x, high) \land battery(x, long) \land camera(x) \rightarrow buy(x)$，$outOfStock(x) \rightarrow \neg buy(x)$ であり，「高解像度かつ電池が長持ちするカメラを買う」ことを主張できるようにしている．

4.4 評価方法

本研究では，提案プロトコルの評価を以下の二段階で行う．

4.4.1 既存研究との意味的比較

まず，提案プロトコルによって生成された議論過程および止揚論証が，Kido および Kurihara による形式的な弁証法的推論例と意味的に一致しているかを検証する．

具体的には，以下の観点から比較を行う．

- 各エージェントが Issue に対する ground instance として初期の main argument を構築しているか
- 生成された論証同士が defeat 関係を形成し，相互に正当化されない状態に到達しているか
- 両エージェントが新たな main argument を生成できない状態に至った後，既存の論証の単なる再表現ではなく，新たな warrant に基づく main argument が構築されているか
- 最終的に正当化される止揚論証が，文献中の形式例と同様の意味内容を持っているか

これらの観点から，提案プロトコルが計算的弁証法モデルと同等の意味構造を再現できているかを評価する．

4.4.2 自然言語表現の確認（補助的評価）

次に，生成された各論証および止揚論証について，自然言語としての一貫性および理解可能性を確認する．本研究では，Anghel らによって提案されたルーブリック評価手法を参考にし，論証の論理的明瞭性や記述の一貫性といった観点から，定性的な確認を行う．

ただし，本研究における主たる評価基準は，既存の計算的弁証法モデルとの意味的一致であり，ルーブリック評価は，生成された自然言語表現が破綻していないかを確認するための補助的な評価として位置づける．

第5章　評価結果

5.1　評価結果の概要

本章では，第4章で述べた評価手法に基づき，提案プロトコルによって生成された議論過程および止揚論証の評価結果を示す．評価は，(1) 既存の計算的弁証法モデルとの意味的一致，(2) 生成された自然言語表現の妥当性，の二点から行った．

5.1.1　既存研究との比較結果

Kido および Kurihara によるカメラ購買問題の対話例を入力とし，提案プロトコルを適用した結果，文献に示された弁証法的推論過程と意味的に整合した議論構造が生成されることを確認した．

具体的には，各エージェントが Issue に対する ground instance として自身のスタンスに基づく main argument を構築し，それらが相互に defeat 関係を形成する過程が再現された．反論フェーズにおいては，いずれの main argument も最終的に正当化されず，両エージェントが新たな反論を生成できない状態に到達した．この状態は，文献において示される “mutually defeating” な議論状態と意味的に一致している．

その後，統合フェーズにおいて，対立する main argument の warrant を入力とし，既存の論証の単なる折衷ではない新たな warrant に基づく main argument が生成された．この止揚論証は，文献中の形式例と同様に，先行する論証では導出不可能であった判断を可能にするものとして位置づけられる．

以上より，提案プロトコルは，計算的弁証法モデルに基づく形式的推論構造を保持したまま，その推論過程を LLM エージェントによる対話として再現できていることが確認された．

図6に，対話例の抜粋を示す．

図 6: 対話例（抜粋）

図6では，論証構築から相互反論を経て統合に至るまでの発話系列を示している．各ターンで結論と前提が明示され，どの前提がどの結論を支持または攻撃しているかが追跡できるため，弁証法的推論の推移を確認できる．

5.1.2　ルーブリック評価結果

次に，生成された各論証および最終的な止揚論証について，自然言語としての妥当性を確認した．本評価では，論証の一貫性，結論と前提の対応関係，および止揚論証の理解可能性といった観点から，定性的な確認を行った．

その結果，生成された論証はいずれも，前提と結論の関係が明示されており，自然言語として破綻のない形で表現されていることが確認された．また，止揚論証についても，対立する価値基準を単に並列するのではなく，より高次の判断基準として再構成する形で記述されており，人間にとって意味的に解釈可能な表現となっていた．

ただし，本研究におけるルーブリック評価は，生成結果の品質を定量的に比較することを目的としたものではなく，既存の計算的弁証法モデルとの意味的一致を前提とした上で，自然言語表現が破綻していないかを確認するための補助的な評価である．

第6章　考察

6.1　考察

6.2　展望

第7章　おわりに

謝辞

本研究の遂行にあたり，ご指導・ご助言を賜った関係者の皆様に深く感謝する．

参考文献

[1] Sawamura, H. and Umeda, Y.: Computational Dialectics for Argument-based Agent Systems, *Proc. ICMAS 2000*, pp. 271–278 (2000).

[2] Kido, H. and Kurihara, M.: Computational Dialectics Based on Specialization and Generalization, *Proc. JSAI 2008*, pp. 228–241 (2008).

[3] Anghel, C. et al.: Multi-Model Dialectical Evaluation of LLM Reasoning Chains, *Informatics*, Vol. 12, No. 3, p. Art. 76 (2025).