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それを保証できない．

**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，同一内容の主張・反論の再生成をプロンプトによって明示的に制限するとともに，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，それまでに生成された Argument 群から，対立している評価軸をエージェント自身に抽出させる．これにより，お互いの議論に決着がついていない場合であっても，対立構造を明示化した上で統合処理に移行することが可能となる．止揚論証の構築においては，Kido, Kurihara [1] による統合アルゴリズムの形式的手続きを参照し，それを自然言語プロンプトとして再構成した．具体的には，対立する主張の前提（support）を背景知識のもとで拡張する処理，統合可能な評価軸に基づいて論証を特化する処理（Specialization），および両者を包摂する上位概念を導出する汎化処理（Generalization）を段階的に実行させることで，単なる折衷ではなく，価値観の前提そのものを再構成した止揚論証を生成できることを確認した．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 10 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．

# 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

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体の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体の対話エージェントによって相互に行われる．エージェントの議論は，結論とその前提を論理的に1文で表現する argument と，前提部分のみで構成される support を組み合わせた Argument スキーマを通じて行われる．これにより，各フェーズでの発話は，単なる自然言語の主張ではなく,「どの結論が，どの前提に基づいて導かれているのか」が明示された構造化表現として扱われる．その結果，反論フェーズにおいては，相手の結論そのものを否定するのか（rebut），あるいは前提や推論規則を攻撃するのか（undercut）といった攻撃の型を厳密に区別することが可能となる．さらに，本スキーマに基づく表現は，後段の統合フェーズにおいて，対立する論証の前提構造を抽象化・再構成するための入力としても利用される．すなわち，議論フェーズで得られた Argument 群を，結論と根拠の対応関係を保持したまま整理することで，specialization や generalization による止揚論証

（synthesis）の構築を，一貫した形式のもとで実行できる．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，まず一方のエージェント（AG1）が自身の立場に基づく主張を構築する．続く反論フェーズでは，相手のエージェント（AG2）がその主張に対して反論可能かを判断し，可能であれば反論を生成する．AG2 が反論できない場合，AG1 の主張は支持されたものとして議論を終了する．

一方，AG2 が反論を行った場合，AG1 はその反論に対して再度反論可能かを判断し，可能であれば再反論を行う．AG1 が再反論できない場合，AG1 の主張は退けられたものとみなされる．このような反論と再反論のやり取りを 1 回の議論単位とし，システムであらかじめ設定した最大反論回数に達するまで繰り返す．最大回数に達しても，いずれの主張にも決着がつかない場合，その議論は「未決着」の状態として扱う．

AG1 の主張が退けられた場合，あるいは議論が未決着のまま終了した場合には，次に AG2 が自身の立場から新たな主張を構築し，同様の手続きを行う．このようにして，両エージェントの主張に関する議論を一通り行った後，統合フェーズへと移行する．

本来，Kido・Kurihara の統合アルゴリズムは，双方の主張がいずれも反論によって退けられた場合に適用される．しかし本研究では，議論に決着がつかず未決着のまま終了した場合にも，統合処理を行うように拡張した．

その理由は二つある．第一に，LLM エージェントによる自然言語の議論では，意味的に近い主張が繰り返されやすく，形式的な意味での決着がつかないまま議論が継続してしまう場合が多いからである．第二に，実社会の議論においても，必ずしも一方が完全に論破されるとは限らず，むしろ対立点を整理した上で新たな解決策を見出すことこそが，弁証法的推論の本質であると考えられるためである．

統合フェーズでは，これまでの議論をもとに，AG1 が双方の主張の対立点を整理し，統合アルゴリズムに適した形へと再構成する．この再構成は，Kido・Kurihara における Specialization に対応し，両者の立場から受け入れ可能な条件を抽出・明示した形の主張を生成する操作である．続く Generalization では，再構成された二つの主張を入力として，それらが同時に成立しうる制度的枠組みや社会的背景といった上位概念を導出する．

これらの処理により，本手法は，各エージェントが議論を通じて相手の価値観を理解した上で，単なる折衷ではなく，両者の前提を再構成する高次の解決策（止揚論証）を生成することを可能にする．

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，これらの課題に対して，形式定義された止揚論証を Specialization および Generalization によって構築する手法を提案した．これにより，対話を通じて，双方の立場を包含する統合を行うための形式的操作が明確化された．一方で，これらの手法は記号論理による知識表現を前提としており，統合可能な概念や評価軸は，あらかじめ定義された述語や規則の範囲に強く制約される．その結果，現実の社会的・倫理的議論に見られるような，文脈依存的で曖昧な価値観や，新たな概念の創発を伴う高次の止揚を十分に扱うことは困難である．さらに，形式的に導出された止揚論証が，実際にどのような意味的解釈を持ち，どのような制度や価値の再構成を示すのかは，論理式そのものからは直接的に読み取ることができず，依然として課題として残されている．

加えて，これらの記号論理で表現される議論のフレームワークでは，議論に用いられる事実や規則，およびそれらの組合せがあらかじめ明示的に定義された知識ベースに強く依存する．そのため，扱える概念や評価軸は限定的であり，現実の社会的・倫理的問題に見られるような，曖昧で文脈依存的な価値観や，新たな観点の創発を柔軟に取り込むことが難しい．また，統合（止揚）においても，論理式として記述可能な範囲での形式的な一般化にとどまり，意味的に妥当な再解釈や，高次の概念形成を伴う統合を十分に表現できないという課題がある．

## 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と、対話を行うエージェントのスタンスを入力とし，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 1 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．

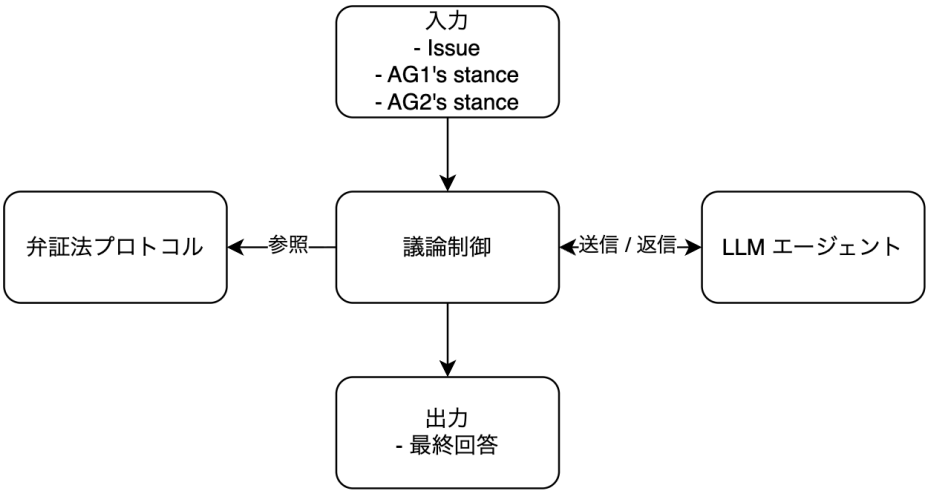


図 1: 提案システムの全体アーキテクチャ

### 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 stance（エージェント 1 の結論）
- AG2's stance（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 stance，AG2's stance 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 stance:
  「あなたは議論エージェント AG1 です。
  あなたは、学校清掃は教育に含まれるべきであり、生徒が行うのが望ましいと考えています。
  清掃は単なる作業ではなく、責任感や協力、公共心を身につけるための実践であり、教室の外で学べる重要な教育活動だと捉えています。
  あなたは社会性や主体性、共同体の一員としての意識を特に重視します。
  もし清掃を教育に含めないと、学校生活の行動から学びを切り離してしまい、生徒が共同体を支える経験を得る機会を失うと主張できます。」
- AG2's stance:
  「あなたは議論エージェント AG2 です。
  あなたは、学校清掃は教育に含まれず、生徒にやらせるべきではないと考えています。
  清掃は本質的に衛生管理や施設管理の領域に属する仕事であり、教育目的を理由に正当化すると、安全面や責任の所在が曖昧になりやすいと捉えています。
  あなたは安全性と衛生水準、責任の明確さ、そして学習時間の確保を重視します。
  教育だと言うのであれば、本来は目標・指導・安全基準などが明確に設計

されるべきですが、現実には運営の負担を生徒に移しているだけになりがちだ、と反論できます。」

#### 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．具体的なプロトコルの終了状態に関しては，次章で説明するが，本システムでは必ずしも最終回答が出力されるとは限らないことに注意したい．

また本システムでは，対話用の LLM エージェントのモデルとして GPT-4.1-nano を使用する．

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 エージェント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 stance をもとに，議論の対象となる最初の主張を行う．

反論フェーズでは、以下のようにエージェントが相互に発言をし，議論を反復的に行う．

1. 最初の主張，もしくは相手の再反論に対し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，相手の論証が正当化されたとして，議論を終了する．
2. 反論ができた場合、次に最初に主張を述べたエージェントが同様にその反論に対して再反論可能かを判断し，可能であれば再反論，そうでなければ，

敗北した（defeated）として，議論を終了する．

3. これらを設定する反論の最大回数まで反復して行う．

統合フェーズでは，反論フェーズで得られた議論履歴（Argument 群）を入力として，以下のステップで最終回答（止揚論証）を構築する．

1. 議論履歴から統合の前提（背景知識）を整理する．これまでの主張・反論で実際に使われた事実関係や因果関係を抽出し，双方が共有している前提と，どの条件下で各主張が成立するのか（成立条件・制約）を明確化する．ここでは議論に現れていない新しい知識は持ち込まず，あくまで履歴に基づいて土台を作る．
2. それぞれの立場の論証を，背景知識に沿って補強し，統合可能な形に整える．各論証の結論は変えずに，背景知識から一貫して導ける範囲で前提（根拠）を補い，どの条件を満たすならその主張が妥当かをより具体化する．これにより,「単なる言い分のぶつけ合い」ではなく,「成立条件を明示した主張」として両者を比較できる状態にする．
3. 補強された 2 つの論証を踏まえ，両者を包摂する上位概念を導出して最終結論（止揚論証）を構築する．背景知識の中に，双方の価値（評価軸）を同時に満たしうる第三の選択肢や枠組みが見いだせる場合は，それを優先して採用する．そうでない場合も，単に両者の理由を並べる折衷ではなく，対立していた価値の前提を捉え直し，両者が納得可能な形で再構成した結論として提示する．

### 3.3.2 メッセージスキーマ

本研究では、Kido, Kurihara の統合アルゴリズム及び議論形式に準拠させつつ，LLM を用いた幅広い議論を可能にするため，以下のようなメッセージスキーマを定義する

```
{
  "Argument": {
    "type": "main",
    "argument": "結論とその根拠を含んだ自然言語の推論文",
    "support": [
      "推論文を構成する根拠 1",
      "推論文を構成する根拠 2",
      ...
```

```
    ]
  }
}
```

これにより，エージェントは論証の構造を把握し，反論のフェーズにおいて，rebut では argument を，undercut では support を明示的に対象にした上で，反論を行うことができる．

最後に，弁証法プロトコルの全体像をシーケンス図で表す（図 2）．この図では，それぞれの最初の主張をそれぞれ A1，A2 で示す．

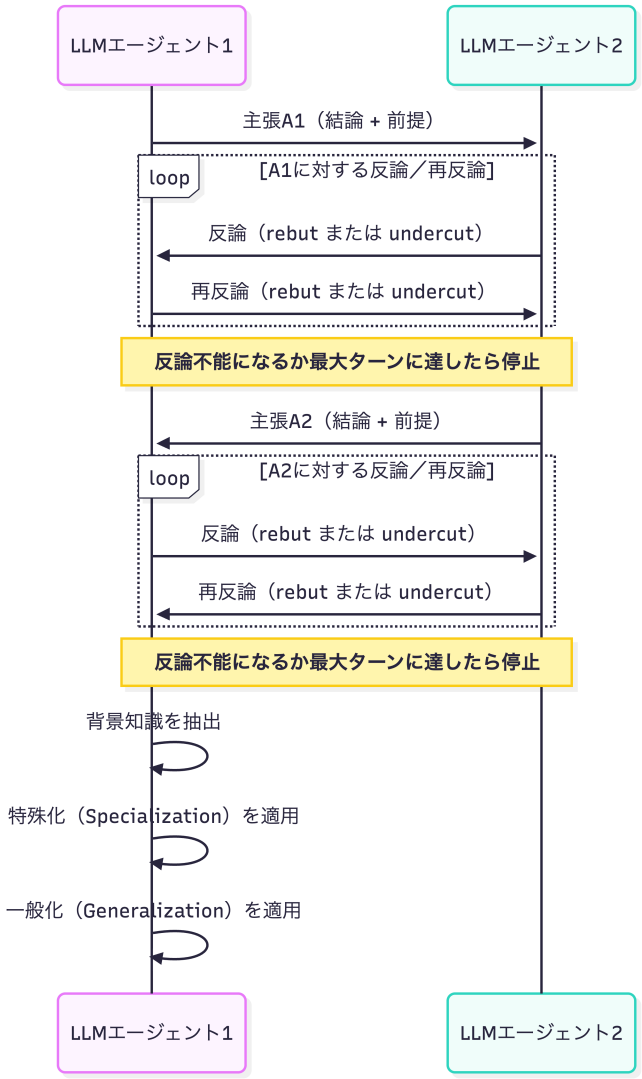


図 2: 提案システムの全体アーキテクチャ

# 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，議題に対するただ1つの回答を定義することが難しい．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，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，および議論の終了後に，統合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，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,「どのカメラを買うか」という議題に対して，それぞれの知識，理論をもった2体のエージェントが議論を行うものである．以下に，それぞれのエージェントの知識と理論を，実際にプロンプトとして与える形で表現した図を示す．

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．

本研究では，カメラ c に関する事実を AG1 の初期プロンプトに埋め込むのではなく，統合フェーズ直前に構築される背景知識へ追加する設計を採用した．この理由は次の通りである．

第一に，初期プロンプトは「立場（価値観・評価軸）」を与えるための入力として位置づけ，環境に関する具体的事実（各カメラの属性や在庫・価格など）は統合処理における参照対象として別管理することで，**立場情報と世界知識の役割を分離**した．これにより，エージェントの人格（何を重視するか）を固定したまま，事実条件のみを変更する実験が可能となり，実装と評価の再現性が高まる．

第二に，カメラ c の情報を初期プロンプトに含めると，反論フェーズの早期から「第三の選択肢」へ議論が逃げやすくなり，対立構造の掘り下げが起こらないまま議論が収束する危険がある．本研究の目的は，まず対立する評価軸（warrant）の衝突を明示化し，その上で統合（止揚）に移行することであるため，**第三の選択肢は統合フェーズで初めて参照可能**とする方がプロトコルの意図に整合する．

第三に，既存研究では知識ベース $S_1$ にカメラ c の事実が含まれているが，LLM においてそれを初期プロンプトへ直接埋め込むと，議論中にどの知識が参照されたかが不透明になりやすい．そこで本研究では，統合フェーズにおいて背景知識（facts/rules）を明示的に構築し，さらに「追加情報適用後」としてカメラ c の事実を別枠で付与することで，**止揚に用いた知識をログとして追跡可能**にした．

以上より，本研究では初期プロンプトを立場提示に限定し，カメラ c の情報は背景知識として統合直前に追加する設計とした．これにより，対立の明示化→統合という段階構造を保ったまま，既存研究と同様に第三の選択肢を含む止揚論証の生成を可能にした．

## 4.3 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．

### 4.3.1 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, 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．

- main argument の構築：各エージェントが入力されたスタンスに基づいて初期の main argument を構築しているか
- 反論の構築：各エージェントが入力されたスタンスに基づいて反論（rebut, undercut）を行うか
- 議論の終了後，既存の論証の単なる再表現ではなく，拡張された論証（warrant）に基づく止揚論証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．

#### 4.3.2 ルーブリック評価（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，あくまで補助的な評価として位置付ける．

# 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，文献とはことなる挙動が得られた．

それは，文献で想定されているように，一方の論証が相手を反論不能にし，justified / defeated の状態が明確に確定して議論が収束する状況が，LLM ベースの対話では観察されなかった点である．

Kido, Kurihara の議論的用例では，有限回の応酬ののち，いずれかのエージェントが相手の論証に対してこれ以上有効な反論（defeating argument）を構成できなくなり，その時点で論証の正当化 / 敗北が決定されることが前提となっている．しかし，本研究の実装では，エージェントが自然言語で反論を生成するため，反論の焦点（攻撃対象）が前提・結論・評価軸の間で揺れやすく，反論の形式も言い換えや観点の追加によって容易に継続してしまう．その結果,「反論不能」状態が発生しにくく，議論が最大回数に達するまで終了しなかった．

その後，統合フェーズにおいて，対立する main argument と構築した背景知識を入力とし，既存の論証の単なる折衷ではない新たな拡張論証に基づく 止揚論証 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3に，対話例の抜粋を示す．


図 3: 対話例（抜粋）

#### 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).