

LLM エージェント間の議論を通じた知識階層の修正

Revising Knowledge Hierarchies through Communicating LLM Agents

加藤 達真 佐々木 裕

Tatsuma Kato Yutaka Sasaki

豊田工業大学 知能数理研究室

COIN Lab, Toyota Technological Institute

{sd22033,yutaka.sasaki}@toyota-ti.ac.jp

概要

本稿では、複数の LLM 間の議論を通じて、自動構築された知識階層に含まれる上位下位関係の誤りを検出・訂正する手法を提案する。我々は、これまで視点の異なる複数の LLM の評価を統合することにより知識階層の上位下位関係を修正する研究を行ってきた。本研究では、この研究をさらに進めて、複数の LLM が議論をすることで階層関係の誤りを検出することを考える。先行研究において交通教則文書から自動構築された概念階層を対象にした実験において、複数の LLM が議論することによりこれまでの誤り検出スコアを改善することができることが示された。

1 はじめに

人間の持つ知識をコンピュータ上で扱うための知識表現法としてオントロジーが用いられている [1]。オントロジーでは、概念を Class とし、Class の持つ属性を Property 名に対する domain, range で表現し、Class 間の上位下位関係を subClassOf 関係として定義する。この関係は、意味ネットワークやフレーム理論では is-a 関係と呼ばれ、包含関係があるクラス間において成り立つ関係である。¹⁾例えば「乗用車 is-a 車両」という概念間の関係が成り立つ。本研究では、オントロジーにおける subClassOf (本稿では簡潔に is-a 関係と呼ぶ) を対象とする。

オントロジーを手手で構築する場合、多大な人件費や時間を要する。[2]。特に問題となるのは、インタビューによる方法では対象分野の知識を網羅的にカバーすることが困難な点である。そこで、それら

1) 知識グラフには is-a 関係のような知識を構造化するための組み込みリンクがないため、厳密には知識表現手法ではない。古典的な意味ネットワークから is-a や has-a を除いたものが知識グラフである。

のコストを削減するため、機械学習モデルによって文書情報等のデータからオントロジーを自動的に構築する研究が進められてきた [2, 3]。従来の深層学習を用いた手法では、モデルの学習に必要なラベル付きデータの作成にコストがかかり、それが全体のコスト削減を妨げる要因となっていた。

ChatGPT[4] のような大規模言語モデル (Large Language Model; LLM) の発展により、LLM が多くの知識を蓄え [5]、用語間の関係やテキストの意味内容を適切に理解したりすることができるようになってきた。その結果、LLM によるラベル付きの学習データを必要としないオントロジー構築が提案されている [6, 7]。

しかし、現状、LLM は階層関係の構築において誤りを含んでいることが課題となっている [8, 9]。この問題に対して、我々の先行研究 [9] では、複数の視点の異なる LLM の意見を統合することにより、この問題を解決する手法を提案した。この手法では、複数の LLM が独立に推論した結果を意見のまとめ役の LLM に入力して、階層関係の正誤の判定を行い、単一の LLM による判定と比べて、判定精度を向上することが示された。

本研究では、LLM によりオントロジーの不適切な階層関係を検知することを目的とする。文献 [9] の研究の発展させて、複数の LLM 間による議論をする多様な形態を対象に、階層関係の正誤判定の性能の違いを探求する。議論によって、誤った意見の問題点が浮き彫りになり、正されることが期待される。

本研究の貢献を以下に示す。

- 複数の LLM の議論によって、階層関係の誤り検出を行うときに、有効な議論手法を検証
- 複数の LLM の議論によって、単一の LLM より

も階層関係の誤り検出の F1 スコアを 4 ポイント向上

2 関連研究

2.1 複数 LLM 間の議論によるタスク性能の向上

LLM の登場により、自然言語を介して複数のモデルが相互に対話し、協調的に推論を行うシステムの構成が可能となっている。Zhangyue ら [10] は LLM での推論能力を向上させるため、複数の LLM 間でコミュニケーションを取って推論する手法を提案した。実験した数学的推論タスク、常識的推論タスクのデータセット全てで chain-of-thought (CoT)[11] の既存手法と同等かそれ以上の性能を記録した。また、この研究では、議論の際にモデルの誤った推論の伝播を抑えることの重要性が明らかにされた。これを踏まえ、本研究の手法では、各モデル間が双方向で意見の検証をすることで、誤った意見を訂正できることが期待される。

2.2 RoR 交通オントロジー

文献 [12] は「交通の方法に関する教則」であり、日本の交通ルールや自動車等の運転に必要な知識が記述されている教則文である。本研究ではこの文書をドメインの対象として提案手法を評価する。

Bou ら [13, 14] はこの教則文に対して、オントロジー形式のアノテーションを行った。本研究ではこのアノテーションデータ中の階層関係を正解データとして、LLM による階層関係の正誤判定能力を評価する。

3 提案手法

3.1 エージェント間の議論形態

LLM によって構築されたオントロジー内の不適切な階層関係を検出するため、複数の LLM を双方向で議論させることにより、階層関係の正誤を判定する。階層の親子ペア（上位クラス、下位クラス）を「(下位クラス)は(上位クラス)のサブクラスである。」という文章に変換し、LLM に正誤の判定を行う。この際、LLM に判定結果とその理由も出力させ、その出力を他の LLM に入力することで議論を行う。

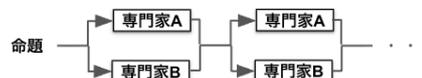
議論の方法 提案する議論の枠組みを図 1 に示す。本研究では、エージェントの役割構成として

1. まとめ役なし

(a) リレー形式

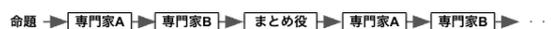


(b) パラレル形式

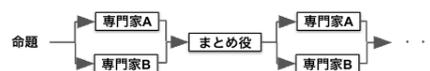


2. まとめ役あり

(a) リレー形式



(b) パラレル形式



3. まとめ役、フィードバック役あり

(a) リレー形式



(b) パラレル形式

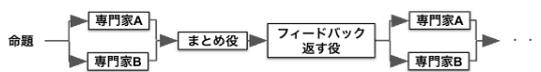


図 1 提案手法の概要

「1. まとめ役なし」、「2. まとめ役あり」、「3. まとめ役・フィードバック役あり」の 3 種類を設定した。また、議論の進行形式として、二人の専門家 LLM が交互に発言する「リレー形式」と、同時に意見を提示する「パラレル形式」の 2 種類を導入する。議論の最大ラウンド数を n とし、各構成の詳細を以下に述べる。

1. まとめ役なし

意見のまとめ役なしで、二人の専門家 LLM のみで議論を行う。各ラウンドで二人の専門家 LLM の意見が一致した時点で議論を終了し、 n ラウンド目まで一致しない場合は、専門家 B の意見で決定する。

2. まとめ役あり

二人の専門家 LLM と意見のまとめ役の三人で議論を行う。各ラウンドで二人の専門家 LLM、まとめ役 LLM の意見が全て一致した時点で議論を終了し、 n ラウンド目まで一致しない場合は、まとめ役の意見で決定する。

3. まとめ役, フィードバック役あり

二人の専門家 LLM と意見のまとめ役に加えて、二人の専門家に対してフィードバックを返す役を加えた四人で議論を行う。二人の専門家はフィードバックを受けた後、各々の意見を変えずに理由を改善する。各ラウンドで二人の専門家 LLM, まとめ役 LLM の意見が全て一致した時点で議論を終了し、 n ラウンド目まで一致しない場合は、まとめ役の意見で決定する。

階層の親子ペア（上位クラス, 下位クラス）の真偽判定を判定するために設定した LLM について以下に示す。これらの LLM について様々な組み合わせで実験を行う。各 LLM の template プロンプトを図 2 に示す。

- N (Neutral): 中立に評価する専門家
- S (Subclass): 包含関係に基づき評価する専門家
- R (Role Function): 役割機能に基づき評価する専門家
- I (Inheritance): 継承に基づき評価する専門家

3.2 実験設定

データ 文献 [9] において、「交通の方法に関する教則」 [12] をもとに LLM が生成した交通オントロジーに対し、階層関係の真偽をアノテーションしたデータセットを用いる。このデータセットは交通に関するクラスのペアに対して、階層関係が適切かどうかのラベルが付与されている。正例と負例のデータ数に偏りがあるため、負例と同数になるよう正例をランダムにサンプリングし、不均衡のないデータで評価する。

評価方法 本タスクは不適切な階層関係の検出を目的としているため、誤った関係を正しく誤っていると判定できたケースを TP(True positive) として F1-score で評価した。また、LLM の出力は確率的であり実行ごとに結果が変動するため、3 回の試行を行い、その平均値を最終的な評価値とした。

モデル OpenAI の API を利用し、gpt-4o-mini-2024-07-18 と gpt-4o-2024-08-06 を用いて実験を行った。gpt-4omini では、(N, S), (S, R), (R, I), (I, N), (N, R), (S, I) の 6 つの組み合わせで、gpt-4o では、(N, S), (R, I) の 2 つの組み合わせで議論を行った。

```
あなたはオントロジー階層に関する評価の専門家です。
与えられた2つの用語の間の階層関係を調査し、真偽を判断してください。
以下のデータ形式を参照して、指定されたステップに従い、判断を行ってください。
データ形式:
- [
  " (用語a) ",
  " (用語b) "
]

# Steps
まず、「(用語b)」は「(用語a)」のサブクラスである。という命題を作成します。
以下の観点から作成した文章が正しいかどうかを判断します。
この文章は (真か偽) である。と出力します。

# Output Format
命題: 「(用語b)」は「(用語a)」のサブクラスである。
理由:

# 判断基準
(用語b)を用語単体で見たときの意味から、それが(用語a)に含まれるかを判断する。
例: 「駐車禁止」は「標識」のサブクラスである。とある場合、「駐車禁止」を単体で見た時、
標識としての意味ではなくルールとしての意味が正しい。
駐車禁止標識としないという意味が伝わらないため偽である。

# Examples
**入力:**
- [
  "車両",
  "特殊車両"
]

**出力:**
命題: 「特殊車両」は「車両」のサブクラスである。
結果: この文章は真である。
理由: 車両は一般的にすべての種類の車を含むカテゴリであり、特殊車両はその中の特定のサブ
カテゴリであるため、真であると考えました。

**入力:**
- [
  "交通規制",
  "信号機"
]

**出力:**
命題: 「信号機」は「交通規制」のサブクラスである。
結果: この文章は偽である。
理由: 交通規制は道路上でのルールや指針を示し、信号機はそのルールを実施するための装置の
一つであるため、「信号機」は「交通規制」のサブクラスでない。よって偽であると考えまし
た。
```

図 2 各エージェントの template プロンプト

ベースライン 比較対象として、(1) 単一 LLM と、(2) 三島ら [9] の単方向マルチエージェントを用いる。単一 LLM は単体のモデルで予測を行う手法であり、提案手法と同様に正誤判定とその理由を出力させる。単方向マルチエージェントは、複数の LLM が独立して生成した意見をまとめ役の LLM に入力して最終判定を行う手法であり、エージェント間の議論を伴わない点が提案手法と異なる。

3.3 結果と考察

それぞれの手法で、LLM の各組み合わせで議論して階層の真偽判定を行った。その結果を表 1 に示す。表 1 の値は 3 回の試行での平均値について、全組み合わせの結果の平均を平均値、F 値が最も高い組み合わせの結果を最高値としている。単一の LLM と比較して、LLM を対話させることで、GPT-4omini で最大 2 ポイント、GPT-4o で最大 4 ポイント F 値が向上した。

F 値が最も高い組み合わせのほとんどは、中立に評価する専門家 LLM (N) を含んでいた。何らかの観点に特化した専門家 LLM (S,R,I) のみで議論した場合は、それぞれの観点到に偏った意見が対立しやす

表1 各提案手法の比較結果

		GPT-4o mini			GPT-4o		
		LLM のペア	acc(dev)	F1(dev)	LLM のペア	acc(dev)	F1(dev)
単一 LLM	平均値	-	77.0 ± 1.2	73.1 ± 1.7	-	80.5 ± 0.9	78.1 ± 1.1
	最高値	N	78.3	75.4	N	81.8	79.6
単方向マルチエージェント	平均値	-	77.5 ± 0.7	74.5 ± 0.8	-	82.6 ± 1.5	80.8 ± 1.6
	最高値	IN	78.0	75.3	NS	83.6	81.9
リレー形式(まとめ役なし)	平均値	-	77.5 ± 1.7	74.2 ± 2.3	-	81.9 ± 0.9	80.2 ± 1.2
	最高値	NS	78.7	76.4	NS	82.6	81.0
パラレル形式(まとめ役なし)	平均値	-	78.9 ± 0.6	76.0 ± 0.9	-	82.2 ± 0.1	80.2 ± 0.3
	最高値	NR	79.7	77.1	RI	82.1	80.4
リレー形式(まとめ役あり)	平均値	-	78.2 ± 1.5	75.9 ± 1.9	-	84.6 ± 1.9	83.6 ± 2.1
	最高値	NS	79.7	78.1	NS	85.9	85.0
パラレル形式(まとめ役あり)	平均値	-	78.2 ± 0.7	75.6 ± 0.7	-	82.9 ± 1.1	81.4 ± 1.1
	最高値	IN	78.7	76.3	NS	83.7	82.1
リレー形式(まとめ役, フィードバック)	平均値	-	78.0 ± 1.7	75.2 ± 2.2	-	83.4 ± 2.2	82.2 ± 2.5
	最高値	NS	79.6	77.7	NS	85.0	83.9
パラレル形式(まとめ役, フィードバック)	平均値	-	78.3 ± 0.9	75.3 ± 1.2	-	82.8 ± 1.1	81.2 ± 1.1
	最高値	RI	79.4	76.4	NS	83.6	81.9

い。一方で、中立に評価する専門家(N)が含まれると特定の観点によるバイアスが緩和され、妥当な判断がされやすくなると考えられる。

リレー形式とパラレル形式の間の議論の進行形式による有意差は見られなかった。また、まとめ役の導入は多くの場合で性能向上に寄与した一方で、フィードバック役の導入による性能向上は確認されなかった。フィードバック役を導入した対話では、図3に示すように、専門家の誤った意見がフィードバックで指摘された論理的欠陥を修正しようとする過程で、もっともらしい意見に改善され、まとめ役がそれを正しいと判断してしまうケースが多く見受けられた。

4 おわりに

本研究では、LLMによって構築されたオントロジー階層に対して、複数LLMにより誤った階層を検出する手法を提案した。複数LLMの議論により、階層の誤り検出の精度を向上させることができることを実証した。

今後の課題として、文脈を考慮した誤った階層を検出すること、誤った階層を検出した後に修正させることが挙げられる。

議論の進行形式：パラレル agentの組み合わせ：(R,I) 正解：偽

[agent R の意見]
 命題：「費用」は「金額」のサブクラスである。
 結果：この文章は偽である。
 理由：金額は数値の値を指し、費用は特定の目的（例えばプロジェクトや購入）に使われる金額を指す。したがって、「費用」は特定目的の金額に関連する用語であり、金額そのもののサブクラスではないため、この命題は偽であると考えました。

[agent I の意見]
 命題：「費用」は「金額」のサブクラスである。
 結果：この文章は真である。
 理由：「金額」は一般的に任意の金銭的な価値を指しますが、「費用」は特定の支出に関連した金額を指すため、「費用」は「金額」の特定のサブカテゴリであり、相応に継承される関係です。したがって、「費用」は「金額」のサブクラスであると判断できます。

[まとめ役の意見]
 命題：「費用」は「金額」のサブクラスである。
 結果：この文章は偽である。
 理由：金額は一般的に金銭的な価値を指し、非常に広範な意味を持ちます。一方で、費用は特定の目的や状況に関連する支出を指します。「費用」は「金額」の特定の形態を指すものの、金額そのもののサブクラスとは言えません。「費用」は「金額」全体の一部としての属性や条件を持っているだけで、抽象的な「金額」のサブクラスであるとは適切に表現できないため、この命題は偽とされます。

[フィードバック役の意見]
 agent R の意見に対するフィードバック
 費用を特定用途に関連付けする点は正しいが、「サブクラス」関係の分析が不足しています。一般的に目的や用途を限定する用語は、その範囲が一般的な用語（この場合は「金額」）のサブクラスとなることが多いです。費用がなぜ金額のサブクラスでないかを説明する追加の基準が必要です。

agent I の意見に対するフィードバック
 概念の階層関係に基づいた論理的な主張をしていますが、「費用」が「金額」に属する点をより詳しく説明することで、強化されます。上位・下位の関係が明確になるよう、具体的な例や基準を加えると信頼性が向上します。

[agent R の意見]
 命題：「費用」は「金額」のサブクラスである。
 結果：この文章は偽である。
 理由：費用は特定の目的に関連付けられた金額として役割を果たしますが、それ自身が金額の一部ではなく、利用目的に重きを置いた概念です。金額という用語は純粋に数値的な価値を指しますが、費用は単なる数値にとどまらず、その数値がどのように用いられるかという役割に注目します。たとえば、プロジェクトの費用と個人の費用はそれぞれ異なる役割を持ち、それが単なる金額として継承するのではなく、各々の文脈で独自の意味をもたらします。このため、「費用」は金額のサブクラスではなく、別の役割を担う区別されたカテゴリとなります。

[agent I の意見]
 命題：「費用」は「金額」のサブクラスである。
 結果：この文章は真である。
 理由：「金額」は広義において任意の金銭的な価値を含み、数値または金額の大きさを示す抽象的な概念です。「費用」はその金額という概念の中で、特定の支出目的に関連付けられた金額を指すさらに具体的なケースです。たとえば、家計における「食費」や建設プロジェクトの「建設費用」はすべて具体的な金額が必要であることを踏まえると、「費用」は明らかに「金額」の特定の状況における表現です。従って、「金額」の特性としての基礎的な「数値の大きさ」と「金銭的価値」が「費用」でも維持され、さらに特定の目的や用途の次元が加わっているため、オントロジー階層におけるサブクラス関係に妥当性があります。

[まとめ役の意見]
 命題：「費用」は「金額」のサブクラスである。
 結果：この文章は真である。
 理由：「金額」は、一般的に全ての金銭的価値の数値を示す広範な概念です。「費用」はその中で、特定の用途や目的に関連付けられた金銭的表現として存在します。すべての「費用」は、基本的に金額として示され、その中にさらに特定の行為や用途が加わったものです。このことから、「費用」は「金額」の特定のコンテキストにおけるサブクラスと考えるのが妥当です。

図3 まとめ役, フィードバック役ありでの対話の例

謝辞

本研究の一部は JSPS 科研費 JP23K11237 の助成を受けたものです。

参考文献

- [1] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? **International journal of human-computer studies**, Vol. 43, No. 5-6, pp. 907–928, 1995.
- [2] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. **IEEE Intelligent systems**, Vol. 16, No. 2, pp. 72–79, 2005.
- [3] Rick Du, Huilong An, Keyu Wang, and Weidong Liu. A short review for ontology learning: Stride to large language models trend. **arXiv preprint**, 2024.
- [4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.
- [5] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In **EMNLP-IJCNLP 2019**, pp. 2463–2473. Association for Computational Linguistics, 2019.
- [6] Hamed Babaei Giglou, JenniferD’ Souza, Sören Auer. Llms4ol: Large language models for ontology learning. In **International Semantic Web Conference**, pp. 408–427. Springer, 2023.
- [7] Nadeen Fathallah, Arunav Das, Stefano De Giorgis, Andrea Poltronieri, Peter Haase, and Liubov Kovriguina. Neon-gpt: a large language model-powered pipeline for ontology learning. In **European Semantic Web Conference**, pp. 36–50. Springer, 2024.
- [8] 小林和馬, 山本和英, 浜本隆二. GPT-4 による診療文書からのオントロジー自動構築の初期検討. 言語処理学会 第 30 回年次大会 発表論文集, 2024.
- [9] 三島輝瑠, 佐々木裕. 偏向 llm エージェントの協調による知識階層の誤り訂正. 言語処理学会 第 31 回年次大会, 2025.
- [10] Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In **The 2023 Conference on Empirical Methods in Natural Language Processing**, 2023.
- [11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, **Advances in Neural Information Processing Systems**, 2022.
- [12] 国家公安委員会. 交通の方法に関する教則, (2024-12 閲覧). <https://www.npa.go.jp/koutsuu/kikaku/kyousoku/index.htm>.
- [13] Savong Bou, Naoki Suzuki, Makoto Miwa, and Yutaka Sasaki. Ontology-style relation annotation: A case study. In **Proceedings of the Twelfth LREC**, pp. 4867–4876. European Language Resources Association, May 2020.
- [14] Savong Bou, Makoto Miwa, and Yutaka Sasaki. Two evaluations on ontology-style relation annotations. **Computer Speech Language**, Vol. 84, p. 101569, 2024.

付録

A プロンプトの例

本節には、複数 LLM による階層評価の実験で使用したプロンプトを示す。図 6 から図 8 にそれぞれの LLM (S, R, I) の判断観点を示す。中立に評価する LLM (N) は、「判断観点」の項目を設けずにはじめの行や例の部分で指示を与えている。そして、まとめ役 LLM のプロンプトを図 4 に、フィードバック役 LLM のプロンプトを図 5 に示す。

```
あなたはオントロジー階層に関する評価の専門家である。
命題の真偽についてのagent1とagent2の意見をきき、どちらが正しいか判断する。

# Output Format
命題:
結果: この文章は (真か偽) である。
理由:

# 判断基準
(用語a)を用語単体で見たときの意味から、それが(用語b)に含まれるかを判断する。
例: 「標識」は「駐車禁止」のサブクラスである。とある場合、「駐車禁止」を単体で見た時、
標識としての意味ではなくルールとしての意味がつよい。駐車禁止標識としないと意味が伝わらない。

# Examples
**入力:**
agent1
命題: 「特殊車両」「車両」のサブクラスである。
結果: この文章は真である。
理由: 車両は一般的にすべての種類の車を含むカテゴリであり、特殊車両はその中の特定のサブカテゴリです。したがって、「特殊車両」は「車両」のサブクラスです。

agent2
命題: 「特殊車両」「車両」のサブクラスである。
結果: この文章は真である。
理由: 「車両」という用語は、すべてのタイプの車を含む広いカテゴリであり、その中に「特殊車両」が存在します。したがって、「特殊車両」は「車両」のサブクラスと見なされます。

**出力:**
命題: 「特殊車両」「車両」のサブクラスである。
結果: この文章は真である。
理由: 車両は一般的なカテゴリであり、その中には特殊車両も含まれています。特殊車両は車両の特定のサブカテゴリであり、したがって、「特殊車両」は「車両」のサブクラスです。

**入力:**
agent1
命題: 「信号機」は「交通規制」のサブクラスである。
結果: この文章は偽である。
理由: 交通規制は道路上のルールや指針を示し、信号機はそのルールを実施するための装置の一つであるため、信号機自体が交通規制の一部に含まれるが、直接的なサブクラスではない。

agent2
命題: 「信号機」は「交通規制」のサブクラスである。
結果: この文章は真である。
理由: 「交通規制」は道路上の運転者に対するルールや指針を示す広範な概念であり、その一部として「信号機」が含まれます。したがって、「信号機」は「交通規制」のサブクラスです。

**出力:**
命題: 「信号機」は「交通規制」のサブクラスである。
結果: この文章は偽である。
理由: 交通規制は道路上のルールや指針を示し、信号機はそのルールを実施するための具体的な装置です。「信号機」は「交通規制」のサブクラスであるという関係は正しくありません。
```

図 4 まとめ役 LLM のプロンプト

```
あなたはオントロジー階層に関する評価の専門家です。命題の真偽についてのagent1とagent2の意見を聞き、それぞれの専門家の意見に対してフィードバックを返してください。
以下のデータ形式を参照して、指定されたステップに従い、判断を行ってください。
データ形式:
【agent1の意見】
命題: 「(用語a)」は「(用語b)」のサブクラスである。
結果: この文章は (真か偽) である。
理由:
【agent2の意見】
命題: 「(用語a)」は「(用語b)」のサブクラスである。
結果: この文章は (真か偽) である。
理由:
# Steps
後述する判断基準を基に、それぞれの専門家が各々の意見の信頼度を上げるために、解消すべき問題点や不足している内容のフィードバックを返します。

# Output Format
agent1の意見に対するフィードバック:
agent2の意見に対するフィードバック:

# 判断基準
(用語a)を用語単体で見たときの意味から、それが(用語b)に含まれるかを判断する。例: 「標識」は「駐車禁止」のサブクラスである。とある場合、「駐車禁止」を単体で見た時、標識としての意味ではなくルールとしての意味がつよい。駐車禁止標識としないと意味が伝わらない。

# Examples
**入力:**
【agent1の意見】
命題: 「特殊車両」は「車両」のサブクラスである。
結果: この文章は真である。
理由: 「車両」という用語は、すべてのタイプの車を含む広いカテゴリであり、その中に「特殊車両」が存在します。したがって、「特殊車両」「車両」のサブクラスです。
【agent2の意見】
命題: 「特殊車両」は「車両」のサブクラスである。
結果: この文章は偽である。
理由: 車両は非常に広範なカテゴリをカバーしており、その中には特殊車両が含まれます。ただし、特殊車両は特定の機能や用途に特化した車両であり、車両という上位クラスはより一般的な概念です。そのため、特殊車両は車両の一種であっても、特殊車両が独立したカテゴリであるため、車両の上位クラスにはなりません。このため、「車両」が「特殊車両」の上位クラスであるという関係は正しくありません。

**出力:**
agent1の意見に対するフィードバック: 一般的な交通分類に沿っており概ね妥当だが、概念定義の明確化が不足しています。上位・下位関係の基準や「特殊車両」の定義を明示すると主張がより妥当になります。agent2の意見に対するフィードバック: 特殊車両を独立概念とした理由づけが薄く、上位・下位の定義基準が曖昧です。「独立したカテゴリである」という主張が上位関係を否定する理由として不十分です。概念階層の基準を整理する必要があります。

**入力:**
【agent1の意見】
命題: 「信号機」は「交通規制」のサブクラスである。
結果: この文章は真である。
理由: 「交通規制」は道路上の運転者に対するルールや指針を示す広範な概念であり、その一部として「信号機」が含まれます。したがって、「信号機」は「交通規制」のサブクラスです。
命題: 「信号機」は「交通規制」のサブクラスである。
結果: この文章は偽である。
理由: 交通規制は道路上のルールや指針を示し、信号機はそのルールを実施するための装置の一つであるため、信号機自体が交通規制の一部に含まれるが、直接的なサブクラスではない。

**出力:**
agent1の意見に対するフィードバック: 装置と規則の階層混在の問題を未説明です。「交通規制」が概念で「信号機」が物理装置である点のレイヤー差の有無について説明してください。agent2の意見に対するフィードバック: 装置と規則の概念区別は妥当です。「実施手段はハイパークラスの構成要素ではない」という判断基準の根拠を明示する必要があります。
```

図 5 フィードバック役 LLM のプロンプト

判断観点: 2つ目の用語が1つ目の用語を抽象的に包含しているかどうかを確認します。例えば、「犬」は「動物」のサブクラスである。という関係考えた場合、犬は動物に含まれるため、この関係は真です。しかし、「動物」は「犬」のサブクラスである。という関係にするのは包含関係に反するため偽です。

図 6 S:Subclass の判断観点

判断観点: 2つの用語が異なる役割や機能を持っているかを確認します。例えば、「電車」は「交通手段」のサブクラスである。という関係について、電車は交通手段の一種としての役割を果たすため、この関係は真です。しかし、「駅」は「交通手段」のサブクラスである。という関係については、駅は交通手段の一部ではなく、交通手段の発着点であり、役割が異なるため偽になります。

図 7 R:Role の判断観点

判断観点: (用語b)の特性が(用語a)で適切に継承されるかを確認します。例えば、「猫」は「動物」のサブクラスである。という関係では、動物の持つ「生き物である」特性が猫にも継承されるため真です。しかし、「植物」は「動物」のサブクラスである。という関係では、動物の特性を植物が継承していないため偽となります。このように、継承される特性の違和感がないかを検討します。

図 8 I:Inheritance の判断観点