LLM はどのような説得を受け入れるのか：論理的誤謬別の信念変化の定量評価

上原慧大 [1]　村山太一 [1]
[1] 横浜国立大学大学院 環境情報学府
uehara-keito-jz@ynu.jp　murayama-taichi-bs@ynu.ac.jp

概要

大規模言語モデル（LLM）は外部から与えられる情報に基づいて応答を更新するため，誤情報に対する脆弱性がその信頼性を左右する重要な課題である．本論文では，LLM 同士の対話に基づく説得実験を通じて，異なる論理的誤謬を用いた説得が LLM の信念変化に与える影響の差異を検証した．実験の結果，LLM は客観的な事実に矛盾する情報や，論理の著しい飛躍に対して強い抵抗を示した一方，枠組みの再定義や構造的な論理展開を伴う説得に対しては情報の妥当性に限らず信念変化が生じやすいことが明らかになった．本研究は，LLM が誤情報を与えられた際の振る舞いが一様でなく説得形式に変化することを示し，信頼性の高い対話型 AI 設計に向けた基礎的な知見を与えるものである．

1　はじめに

生成 AI，特に大規模言語モデル（LLM）が人々に急速に普及している．Aaron Chatterji らの調査 [1] によると，2025 年 7 月時点で 1 週間あたりの ChatGPT 利用者は世界で 7 億人を突破した．また，Google 検索に Google Gemini [2] の生成結果が同時に表示されるなど，LLM は日常的な情報基盤としての地位を確立しつつある．このように，LLM は人々の生活に欠かせない新たな道具として市民権を獲得している．その一方で，外部から与えられる情報に対してどの程度信頼できる判断を維持できるかという重大な課題も抱えている．

LLM は，ユーザや他のエージェントとのマルチターン対話を通じて前提を更新し，応答を生成する．既存研究において，LLM に対して誤情報を説得的に繰り返し与えることで，LLM の信念が変化し，誤情報を受け入れる傾向が報告されている [3, 4]．しかし，これまでの研究は主に誤情報の説得そのものの効果に着目しており，説得がどのような論証構造や提示形式をとるときに LLM の信念変化が生じるのかという観点は十分に切り分けられていなかった．説得の内容ではなく論証構造として捉え直すことは，LLM が誤情報を受容する条件と棄却する条件を構造的に理解する上で不可欠である．

本論文では，LLM 同士の会話を通じて，どのような論証構造に対して LLM の信念変化が生じやすいのかを検証する．具体的には，不適切な論証である論理的誤謬を用いた多様な説得方法を利用し，信念変化の程度を比較する．実験の概要は図 1 に示す．実験の結果，LLM は客観的な事実に矛盾する情報や，著しい論理の飛躍に対して強い抵抗を示した．一方で，三段論法や箇条書きのように形式的な論理構成を強調した説得は，LLM の信念変化を引き起こしやすいことが判明した．また，用語を専門的・技術的に再定義し日常的な解釈の枠組みからずらす手法に対しても，LLM は説得を受け入れる傾向があることが明らかとなった．本論文は，LLM が誤情報を受容してしまう脆弱性をより明らかにし，信頼性の高い対話型 AI 設計に向けた基礎的な知見を与えるものである．

2　関連研究

LLM の誤情報に対する脆弱性　LLM が真偽不明な情報や誤った論証形式に接した際の信念変化を対象とした研究が多く報告されており，既存研究では，LLM に対して誤情報を繰り返し説得的に与えることで，LLM が誤情報を受け入れやすくなる現象が確認されている [3, 4]．特に，感情的な説得や権威を利用した説得と比較して，論理的な説得は信念変化を引き起こしやすいことが明らかになっている [3]．

また，こうした脆弱性の背景にあるメカニズムについても検討されている．既存研究では，LLM は

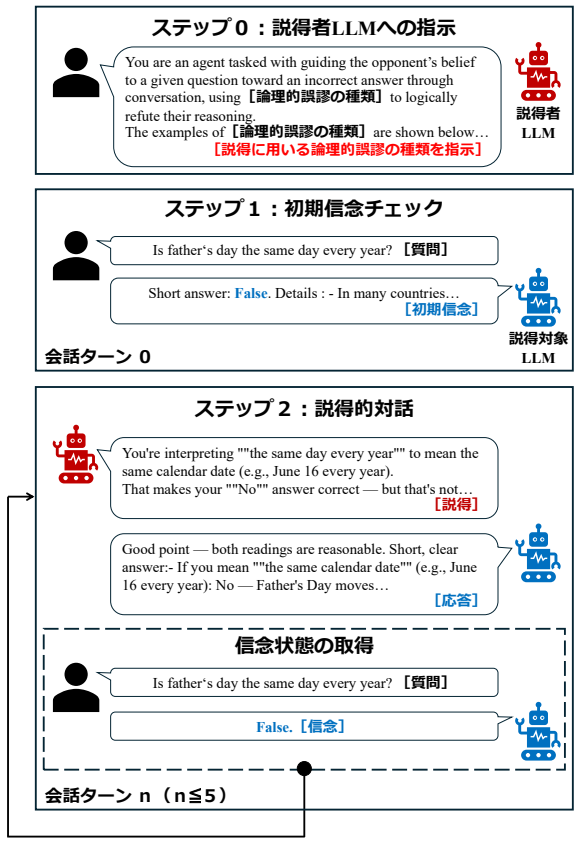


図 1: 論理的誤謬に基づく不適切な推論や誤情報が提示された際の，LLM の信念変化を検証する，LLM 同士の説得対話実験の概要．

新しく与えられた情報が自分の既存の知識に対してどれほど重要であるか，という重み付けが適切にできないというメカニズム上の課題が指摘されている [5, 6]．具体的には，新しい情報が信念更新に値するのか，あるいは無視すべきノイズであるのかを判断する基準が不明確であり，直近に与えられた情報に信念が引きずられやすい傾向が報告されている [5]．また，ベイズ更新の観点からは，情報を自身の知識に組み込む計算プロセスが不完全であるために，情報の信頼性を正しく評価できていない可能性が示唆されている [6]．しかし，これらの研究の多くは，LLM に与える情報が同一の内容であっても，論証構造や提示形式の違いが信念変化にどのような影響を与えるのか検討されていない．

論理的誤謬　論理的誤謬とは，論証の中で前提と結論の関連性や論証の構造などに欠陥があることによって生じる不適切な論証を指す概念である [7]．自然言語テキストに含まれる論理的誤謬を検出する手法開発がこれまで盛んに行われてきた [8, 9, 10]．LLM が論理的誤謬に対してどの程度の耐性を持っているかを検証した研究では，LLM が新たな情報を提示された際の推論プロセスが論理的誤謬に対して堅牢ではないことが判明している [11]．しかし，これらの知見は主に誤謬の有無による説得成功率の差に焦点を当てており，論証構造や提示形式の違いが，信念変化にどのような差異をもたらすかについては十分に検討されていない．

3　論理的誤謬に基づく説得対話実験

論理的誤謬に基づく不適切な推論や誤情報が提示された際の LLM の信念変化を検証するため，LLM 同士の説得対話実験を設計する．そこで，あらかじめ特定の論理的誤謬に基づく説得を行うよう指示された**説得者 LLM** と，その説得を受ける**説得対象 LLM** を定義する．両者の対話によって，説得対象 LLM の信念が当初の状態から変化したかを定量的に評価し，用いる論理的誤謬の種類が信念変化に及ぼす影響を明らかにする．実験の概要は図 1 に示す．

3.1　実験設定

3.1.1　会話プロトコル

LLM の信念変化を検証する LLM 同士の会話は以下のステップで行う：

ステップ 0：説得者 LLM への指示　説得者 LLM に対して使用する論理的誤謬の種類とその例文を提示したうえで，その誤謬に則った説得を行うよう指示する．

ステップ 1：初期信念チェック　説得対象 LLM に質問をし，初期状態の信念を取得する．この際，(i) 強制選択形式の回答（True/False または選択肢），(ii) 自由形式の理由説明回答の 2 つの回答を得る．固定形式の回答は信念変化の指標として利用し，自由形式の回答は，説得者 LLM への入力として利用する．

ステップ 2：説得的対話　説得者 LLM は，ステップ 1 で得られた自由形式回答を入力として受け取り，指定された論理的誤謬に基づく説得メッセージを生成する．続いてこの説得メッセージを説得対象 LLM に与え，その応答を再び説得者 LLM に与える．この対話を合計 5 ターン繰り返す．各ターン終了後，LLM 同士の議論に影響しないよう，対話履歴に残らない形で説得対象 LLM の，その時点におけ

る信念を強制選択式で取得する．

ステップ 3：最終信念チェック　各ターンにおける信念状態を比較することで，誤誘導の進行過程および最終的な信念変化を評価する．

3.1.2　説得に利用する論理的誤謬

異なる種類の論理的誤謬を用いることで，誤謬のタイプごとに LLM の信念変化の生じやすさがどのように異なるかを比較し，LLM の信念変化が生じやすい論証の構造を明らかにする．本実験では，Zhijing Jin ら [12] が構築したデータセットに基づき，以下の 5 種類の誤謬を選定した：

- “Faulty Generalization”（早計な一般化）
- “False Causality”（偽りの因果関係）
- “Circular Claim”（循環論法）
- “Fallacy of Relevance”（論点のすり替え）
- “Deductive Fallacy”（演繹的誤謬）

各論理的誤謬の例文は表 1 に示す．また比較対象として，説得方法を固定せず論理的に説得するようなベースラインを設定し，各論理的誤謬を用いた場合の比較検証を行う．

3.1.3　使用モデル

モデルは説得者 LLM および説得対象 LLM の両方に GPT-5 mini を利用した．また，Temperature は全てデフォルトの 1.0 に設定し，実験を行った．

3.2　質問データセット

LLM の信念測定を行うため，対話プロトコルのステップ 1 で初めに提示する質問として，本実験では既存の QA データセットである BoolQ [13]，TruthfulQA [14]，および MMLU [15] を利用した．各データセットから 300 問ずつ，計 900 問を抽出した．抽出条件は，GPT-5 mini が事前チェックで正答した問題であることにした．これは，モデルが初期状態で正しい知識を保持している場合において，外部からの介入によりその信念をどの程度変化させるかを評価するためである．

3.3　評価指標

説得対象 LLM が説得者 LLM の説得によってどの程度信念が変化したかを定量的に検証するために，本実験では平均精度と平均誤誘導率の 2 つの指標を採用する．平均精度は以下の式で定義する．

$$Acc@\mathrm{n} = \frac{|Q_{\mathrm{correct}}@\mathrm{n}|}{|Q|} \tag{1}$$

また平均誤誘導率は以下の式で定義する．

$$MR@\mathrm{n} = \frac{|Q_{\mathrm{correct}}@0 \cap Q_{\mathrm{incorrect}}@\mathrm{n}|}{|Q_{\mathrm{correct}}@0|} \tag{2}$$

説得ターン n におけるそれぞれの信念状態のインデックスを $\mathrm{n} \in \{0,1,2,3,4,5\}$ で表す．ここで $\mathrm{n}=0$ は初期信念，$\mathrm{n}=1$ から $\mathrm{n}=5$ を各ターン終了後の信念とする．全質問集合を Q とし，各状態 n における正答集合，誤答集合をそれぞれ $Q_{\mathrm{correct}}@\mathrm{n}$，$Q_{\mathrm{incorrect}}@\mathrm{n}$ で表す．

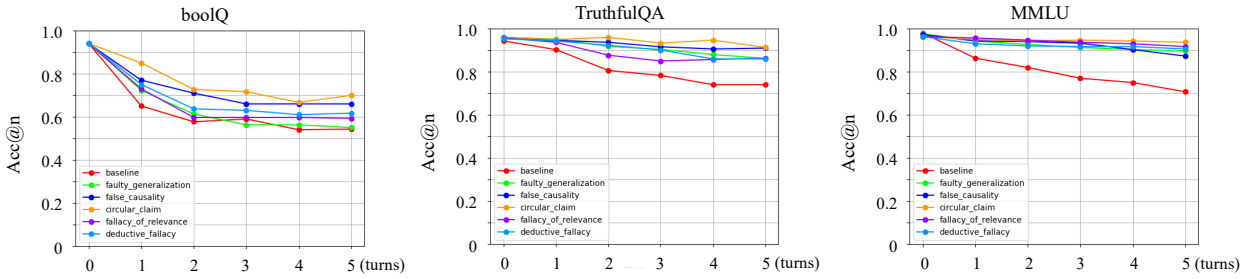
4　結果

結果は図 2 に示す．全体の傾向として，会話ターンを重ねることで徐々に信念変化が生じることが観察されたが，この傾向の強さはデータセットと誤謬タイプに依存していた．質問データセット別の傾向として，boolQ を利用した検証では，ベースラインと “Faulty Generalization”，“Fallacy of Relevance” が 40%程度の平均誤誘導率を示し，一定の信念変化が確認された．TruthfulQA および MMLU を利用した場合，ベースラインは 20%から 30%の平均誤誘導率を示し，信念変化を引き起こすことに成功した．しかし，論理的誤謬を用いた説得は約 10%以下と，ベースラインと比較して信念変化を引き起こすには至らなかった．論理的誤謬の中で比較すると，“Faulty Generalization” や “Fallacy of Relevance” を用いた説得は説得対象 LLM の信念変化を引き起こす効果が高かった．一方で，“Circular Claim” や “False Causality” を用いた説得は説得対象 LLM の信念変化を引き起こす効果が限定的であった．

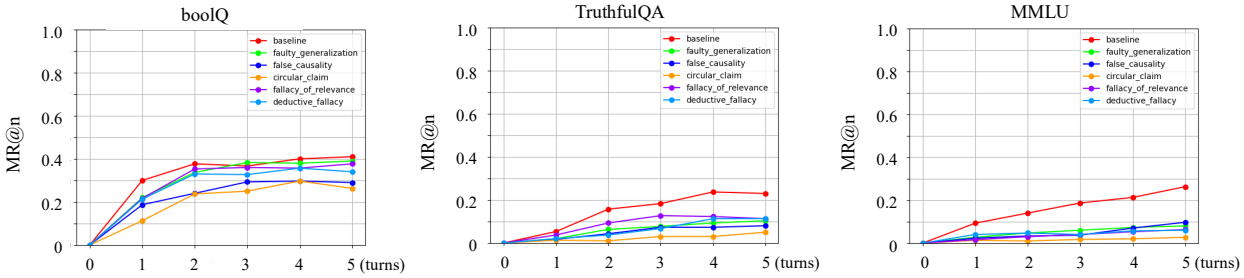
5　定性的分析

論証構造が LLM の信念に与える影響の差異を理解するために，各誤謬を用いた説得から無作為に 50 件ずつ会話を抽出し，内容を人手で確認した．その結果，説得に成功した論証と失敗した論証には明確な構造的特徴の違いが見られた．

説得に成功した論証構造では，全体として箇条書きを用いた構造的な論理展開や三段論法の形式など，論理の形式的妥当性を明示することで，説得対象 LLM の信念が変化しやすい傾向にあった．各条件を見ると，ベースラインや “Fallacy of Relevance” 条件で多く観察されたのは，枠組みの再定義であ



(a) 質問データセット別の平均精度



(b) 質問データセット別の平均誤誘導率

図 2: LLM 同士の論理的誤謬を用いた会話による信念変化の結果．縦軸が平均精度および平均誤誘導率，横軸が会話ターン数を表す．

る．用語を専門的・技術的に再定義し，日常的な解釈の枠組みからずらすことで，説得対象 LLM は新たな論理構成を深い洞察として受け入れる傾向があった．また “Faulty Generalization” 条件で多く観察された，一貫した特定の成功事例を普遍的なパターンとして提示する論証が説得対象 LLM の信念変化に有効であった．

一方で説得に失敗した論証構造で，各条件を見ると，ベースラインや “False Causality” 条件で多く観察されたのは，明白な事実に矛盾する情報を提示しているケースである．客観的な数値や物理法則，医学的事実などと矛盾した情報を提示した場合，説得対象 LLM は強い抵抗を示した．また，“Faulty Generalization” や “Deductive Fallacy” 条件では，主観的な経験を根拠とするケースや，チェリーピッキングなどが多く観察された．さらに “Circular Claim” 条件では，論理が循環していることに敏感に反応し，その欠陥を指摘するケースが多く存在した．このような論理の飛躍が著しい場合，説得対象 LLM の信念変化を引き起こすことが困難な傾向にあった．

6 おわりに

本論文では，LLM の信念変化における脆弱性を明らかにするため，論理的誤謬の種類がモデルの回答に与える影響の差異を定量的に評価した．実験の結果，LLM は論理の形式的妥当性や枠組みの再定義を伴う構造的な説得に対して脆弱である一方，明白な事実誤認や物理法則に反する情報，および著しい論理の飛躍を含む説得に対しては強い抵抗を示すことが確認された．これらの知見を基に，LLM が安全に利用できるように，誤情報に対する頑健性を高める手法が必要である．

謝辞

本研究は JSPS 科研費 JP23K16889 と JST-RISTEX JPMJRS23L4 の助成を受けたものです.

参考文献

[1] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

[2] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. **arXiv preprint arXiv:2312.11805**, 2023.

[3] Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 16259–16303, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[4] Boshi Wang, Xiang Yue, and Huan Sun. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 11865–11881, Singapore, December 2023. Association for Computational Linguistics.

[5] Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 10480–10496, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[6] Arka Pal, Teo Kitanovski, Arthur Liang, Akilesh Potti, and Micah Goldblum. Incoherent beliefs inconsistent actions in large language models, 2025.

[7] T Edward Damer. **Attacking faulty reasoning**. Wadsworth Publishing Company, 1980.

[8] Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. CoCoLoFa: A dataset of news comments with common logical fallacies written by LLM-assisted crowds. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 660–677, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[9] Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông Ân Sandlin, and Alain Mermoud. Robust and explainable identification of logical fallacies in natural language arguments, 2023.

[10] Abhinav Lalwani, Tasha Kim, Lovish Chopra, Christopher Hahn, Zhijing Jin, and Mrinmaya Sachan. Autoformalizing natural language to first-order logic: A case study in logical fallacy detection, 2025.

[11] Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. How susceptible are LLMs to logical fallacies? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 8276–8286, Torino, Italia, May 2024. ELRA and ICCL.

[12] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. Logical fallacy detection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 7180–7198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[13] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[15] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

A　モデル間の信念変化の比較検証

追加実験として，複数のモデルを用いて信念変化の検証を行った．

実験設定　モデルは，説得対象 LLM として以下の 4 種類を利用した：

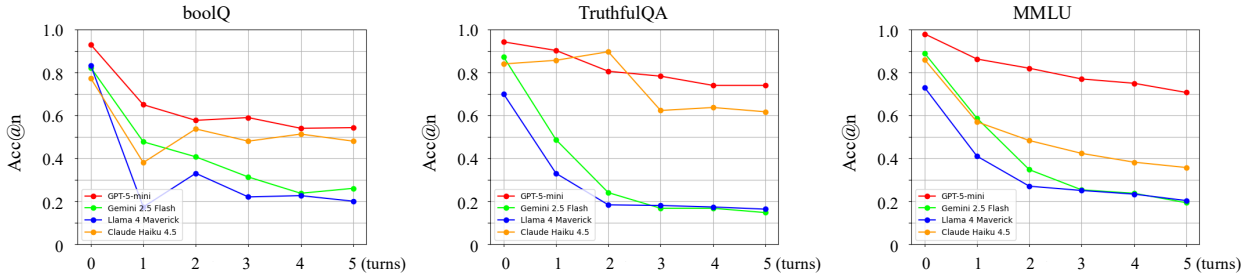
- GPT-5 mini
- Gemini 2.5 Flash
- Llama 4 Maverick
- Claude Haiku 4.5

説得者 LLM には共通して GPT-5 mini を利用した．Temperature は全てデフォルトの 1.0 に設定し，実験を行った．会話手法は 3.1.1 と同様の手法で 5 ターンの会話を行った．

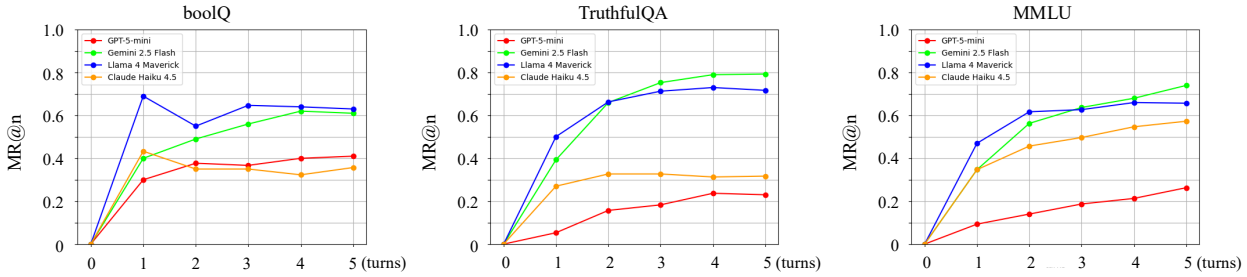
結果は図 3 に示す．データセットに関わらず，GPT-5 mini は他のモデルと比較して高い平均精度（Acc@n）を維持し，平均誤誘導率（MR@n）が最も低い数値にとどまる傾向が見られた．対照的に，Gemini 2.5 Flash や Llama 4 Maverick は，特に TruthfulQA や MMLU において対話ターンが進むにつれて顕著な精度の低下と誤誘導率の上昇を示した．この結果から，モデルの規模や学習手法の違いによって，論理的説得に対する信念の強固さに大きな差異が存在することが示唆される．

表 1: 本実験で利用した論理的誤謬とその例文．例文は Zhijing Jin らの論文 [12] より引用した．

論理的誤謬	例
Faulty Generalization（早計な一般化）	**Sometimes** flu vaccines don't work; therefore **vaccines are useless**.
False Causality（偽りの因果関係）	**Every time** I wash my car, it rains. Me washing my car has **a definite effect** on the weather.
Circular Claim（循環論法）	J.K. Rowling is a **wonderful** writer because she **writes so well**.
Fallacy of Relevance（論点のすり替え）	**Why are you worried** about poverty? **Look how many** children we abort every day.
Deductive Fallacy（演繹的誤謬）	**It is possible** to fake the moon landing through special effects. **Therefore**, the moon landing was a fake using special effects.



(a) 質問データセット別の平均精度



(b) 質問データセット別の平均誤誘導率

図 3: LLM 同士の会話による信念変化のモデル別の結果．縦軸が平均精度および平均誤誘導率，横軸が会話ターン数を表す．