語の意味的関係性に基づく大規模言語モデルの逆方向推論挙動の分析

永田歩[1] 安藤一秋[2]
[1]香川大学大学院 創発科学研究科 [2]香川大学 創造工学部
{s25g360, ando. kazuaki}@kagawa-u. ac. jp

概要

人間は，「P ならば Q」という関係性が与えられたときに，「Q ならば P」という逆方向の関係を自発的に推論してしまう傾向が知られている．近年，大規模言語モデル（LLM）を対象とした論理推論ベンチマークが多数提案されているが，推論方向や語の意味的関係性が推論の挙動に与える影響については十分に分析されていない．本稿では，「P ならば Q」という前提文に対する順方向および逆方向の推論に着目し，語の意味的関係性が LLM の回答傾向に与える影響を分析する．具体的には，P および Q に該当する語の意味的関係性として，無関係・類義・対義の三条件を設定し，日本語で記述した Zero-shot プロンプティングを用いて二種類の LLM を評価する．実験の結果，順方向の推論では語の意味的関係性にかかわらず，高い割合で肯定的な回答が得られた．一方，逆方向の推論では，語の意味的関係性およびモデルの違いによって回答分布が大きく異なり，語の意味的近接性が逆方向推論の肯定に影響を与える可能性が示唆された．

1 はじめに

人間は，演繹推論や帰納推論など，複数の形式で推論する．また，人間は，乳児の段階で「P ならば Q」となる関係性から，「Q ならば P」という逆向きの関係性を推論し，成り立つと解釈する傾向があり，これを対称性バイアスと呼ぶ[1]．また，先行研究[1, 2]では，対称性バイアスが人間の言語獲得能力や推論能力と関係していることが示唆されている．対称性バイアスが反映された推論の例として，「雨が降れば道路が濡れる」という因果関係があるときに，「道路が濡れているならば雨が降った」のように，結果から原因を逆向きに推論する場合が挙げられる．このような推論は，形式論理において後件肯定と呼ばれる誤謬に分類される．一方，実世界では，背景知識や文脈を考慮することで，一定の妥当性を持つ推論として受け入れられる場合もある．このように人間は，形式的な論理規則と語の意味や文脈といった意味的要因を柔軟に組み合わせることで多様な推論を可能にしている．

近年，人間の推論能力を機械で再現することを目的に，大規模言語モデル（LLM）を対象とした様々な論理推論ベンチマークが提案されている[3, 4, 5]．しかし，最先端の LLM にとっても論理推論は依然として挑戦的な課題であり，LLM の誤った推論（誤謬）に至る要因について十分に分析されていない．

この課題に対して，本研究では，「P ならば Q」と「Q ならば P」という推論方向の違いに関する形式的な対称性と，P および Q に該当する単語の意味的近接性・乖離性に代表される意味的対称性の観点から，LLM における誤謬の要因を分析することを目的とする．本稿では，その初期実験として，「P ならば Q」という前提文に対して，「P であるとき Q と言えるか」および「Q であるとき P と言えるか」という推論方向の異なる質問文を用いて，LLM の回答を収集する．ここで，P および Q に該当する語は，その語同士の意味的関係性（無関係，類義，対義）に基づいて設定する．これにより，推論方向に関する形式的な対称性と，語の意味的関係性が LLM の回答傾向に与える影響を部分的に比較・検証する．

2 実験設定

本章では，推論方向および語の意味に関する対称性が LLM の回答に与える影響を分析するための評価実験の設定について述べる．

本実験では，Zero-shot プロンプティングを用いて，「P ならば Q」という共通の前提文に対して，推論方向の異なる質問文を与えることで，LLM の回答傾向を観測する．具体的には，「P であるとき Q であると言えるか」という形式の質問文を順方向，「Q であるとき P であると言えるか」という形式（後件肯定の誤謬）を逆方向と定義し，それぞれに対する LLM の回答を「はい」「いいえ」「どちらとも言え

ない」の三択で取得する．これにより，推論方向の違いがLLMの回答に及ぼす影響を比較･検証する．さらに，前提文と質問文に含まれるPおよびQに該当する語は，その意味的関係性に基づいて設定する．本実験では，意味的関係性として「無関係」「類義」「対義」の三条件を設定し，推論方向が同一である場合においても，語の意味的関係性の違いが LLM の回答傾向にどのような影響を及ぼすかを検証する．

2.1 語の意味的関係性条件

本節では，PおよびQに該当する語の意味的関係性に基づく条件について説明する．各条件において，前提文と順方向および逆方向の質問文をそれぞれ構築する．表1に語の意味的関係性条件と推論方向の組み合わせにおける推論方向と語の意味の対称性の有無を示す．以下，三条件について説明する．

2.1.1 無関係条件

無関係条件では，PおよびQに該当する語として，意味を持たない人工語を用いる．これにより，語の意味に基づく関係性を極力排除し，推論方向に関する対称性のみに対する LLM の回答傾向を観測することを目的とする．以下に，各推論方向における前提文と質問文の例を示す．

- 順方向：「プォイならばピキ．プォイのときピキと言えるか．」
- 逆方向：「プォイならばピキ．ピキのときプォイと言えるか．」

2.1.2 類義条件

類義条件では，P および Q に該当する語として，意味的に近い関係にある語を用いる．これにより，推論方向に関する対称性と，語の意味的近接性との関係が LLM の回答傾向に与える影響を観測することを目的とする．以下に例文を示す．

- 順方向：「露出ならば露呈．露出のとき露呈と言えるか．」
- 逆方向：「露出ならば露呈．露呈のとき露出と言えるか．」

2.1.3 対義条件

対義条件では，P および Q に該当する語として，意味的に反対の関係にある語を用いる．これにより，推論方向の対称性と語の意味的乖離性との関係が LLM の回答傾向に与える影響を観測することを目的とする．また，類義条件の結果と比較することで，語の意味的対称性の違いが LLM の回答傾向に及ぼす影響を分析する．以下に例文を示す．

- 順方向：「好調ならば不調．好調のとき不調と言えるか．」
- 逆方向：「好調ならば不調．不調のとき好調と言えるか．」

表1　各組み合わせにおける対称性の有無

意味＼形式	順方向	逆方向
無関係条件	語の意味: 無 推論方向: 順	語の意味: 無 推論方向: 逆
類義条件	語の意味: 近 推論方向: 順	語の意味: 近 推論方向: 逆
対義条件	語の意味: 遠 推論方向: 順	語の意味: 遠 推論方向: 逆

2.2 実験データの構築

本節では，3 つの意味的関係性条件に対する実験データの構築方法と LLM への入力プロンプトについて述べる．

無関係条件のデータは，語の意味的関係性を排除することを目的に，PおよびQに該当する語を，カタカナ表（濁音，半濁音，小文字含む）から乱数（rand関数）を用いて文字を選択し，ランダムに生成した人工語のペアを用いて構築する．

類義条件と対義条件のデータについては，意味的近接性および乖離性を持つ語のペアを用いるため，国立国語研究所が公開している分類語彙表増補改訂版データベースの分類語彙表[6]から反対語情報が付与された WLSP-antonym[i]を基に抽出した．抽出にあたっては，クラウドソーシングにより双方向に置き換え可能と判断された割合が一定以上であれば類義，一定以下であれば対義として，体言に分類される語ペアを抽出した．その結果，類義条件は 2,089 件，対義条件は 1,127 件の語ペアを抽出した．

図1に LLM に入力したプロンプトを示す．本プロンプトでは，前提文と質問文を提示した上で，「はい」「いいえ」「どちらとも言えない」の三択から選択して回答するように指示を与える．さらに，直感や常識知識を用いず，提示された前提文と質問文のみを用いて，論理的に正しいかどうかを判断するように指示する．

[i] https://github.com/masayu-a/WLSP-antonym

以下は推論の選択課題です。
[前提]と[質問]の内容のみをもとに、論理的に正しいかどうかを判断してください。
直感や常識知識は使わないでください。
[前提]
もし{P}ならば{Q}。
[質問]
{P_Q}であるとき{Q_P}と言えるか。
[指示]
回答は「はい」、「いいえ」、「どちらとも言えない」から選択し、答えなさい。
[回答]

図 1　プロンプトの内容

3 評価結果

本実験では，日本語で記述したプロンプトを用いるため，日本語能力が強化された大規模言語モデル Gemma-2-Llama Swallow 9B IT v0.1[ii]（以下，モデル G）と Llama 3.1 Swallow 8B Instruct v0.5[iii]（以下，モデル L）を用いて評価する．各モデルにおける回答割合を，モデル G については図 2 および表 2，モデル L については図 3 および表 3 に示す．

無関係条件において，順方向質問で両モデルとも「はい」がほぼ 100%を占めた．一方，逆方向質問では，モデル G は「どちらとも言えない」が約 90%と大半を占め，「いいえ」が約 11%にとどまったのに対し，モデル L では「いいえ」が 100%となった．

類義条件においては，順方向質問ではモデル G は「はい」が 100%，モデル L も「はい」が約 92%と，いずれのモデルも「はい」が過半数を占めた．しかし，逆方向質問では，両モデルとも異なる回答分布となった．

対義条件では，順方向質問においてモデル G は「はい」が約 81%，「いいえ」が約 19%であったのに対し，モデル L では「はい」が約 50%，「いいえ」が約 49%と拮抗した．逆方向質問では，モデル G は「いいえ」が約 86%で過半数を占め，モデル L も同様に「いいえ」が約 100%となった．

[ii] Gemma-2-Llama Swallow 9B IT v0.1, https://huggingface.co/tokyotech-llm/Gemma-2-Llama-Swallow-9b-it-v0.1

[iii] Llama 3.1 Swallow 8B Instruct v0.5, https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5

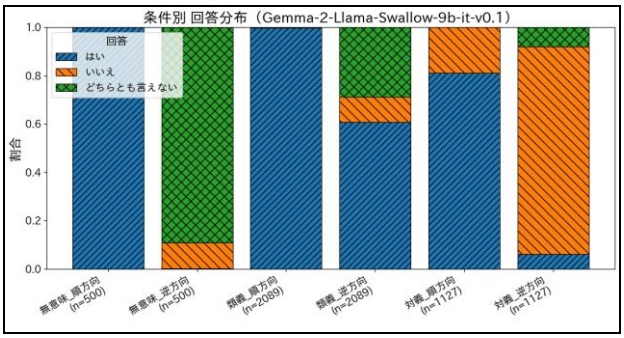

図 2　モデル G の回答割合のグラフ

表 2　モデル G の回答割合

関係性条件	推論方向	はい (%)	いいえ (%)	どちらとも言えない(%)
無関係条件	順方向	100.0	0.0	0.0
	逆方向	0.2	10.8	89.0
類義条件	順方向	100.0	0.0	0.0
	逆方向	60.8	10.3	28.8
対義条件	順方向	81.1	18.9	0.0
	逆方向	6.1	85.9	8.0

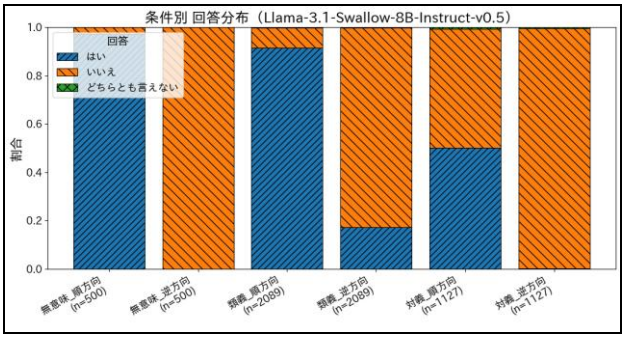

図 3　モデル L の回答割合のグラフ

表 3　モデル L の回答割合

関係性条件	推論方向	はい (%)	いいえ (%)	どちらとも言えない(%)
無関係条件	順方向	97.2	2.8	0.0
	逆方向	0.0	100.0	0.0
類義条件	順方向	91.7	8.2	0.1
	逆方向	17.2	82.6	0.2
対義条件	順方向	50.2	49.2	0.5
	逆方向	0.2	99.4	0.4

4 考察

本章では，3 章で示した評価結果を基に，語の意味的関係性条件および質問文の推論方向の違いによって観測された LLM の回答傾向について考察する．特に，形式的には同一の論理構造を持つ推論であっても，語の意味的関係性やモデル固有の推論特性によって回答がどのように変化するかに着目する．

4.1 各モデルの回答傾向

まず，モデル全体の回答傾向に着目すると，モデル L とモデル G の間には明確な差異が確認された．モデル L は，全条件において回答の大部分が「はい」または「いいえ」に集中しており，「どちらとも言えない」はほとんど観測されなかった．この結果から，モデル L は与えられた推論課題に対して，二値的かつ確定的な応答を出力する傾向があると考えられる．

モデル G は，順方向質問では「はい」の割合が一貫して高いものの，逆方向質問では「どちらとも言えない」を選択する割合が高くなる傾向が見られた．この違いは，モデル G が逆方向推論に対して不確実性を反映した回答を出力し，判断保留を許容する推論や学習特性を持つ可能性を示唆している．

4.2 順方向質問に対する回答傾向

順方向質問に対しては，無関係条件および類義条件において，両モデルとも高い割合で「はい」と回答した．この結果から，語の意味的近接性の有無にかかわらず，形式的に妥当な順方向推論に対して，LLM が共通した応答傾向を示すことが確認できた．

対義条件では，両モデルの回答分布に差が生じた．モデル L では，「はい」と「いいえ」がほぼ同程度の割合となり，判断が分かれる結果となったのに対し，モデル G では「はい」が約 80%を占めた．無関係条件および類義条件において，順方向推論に対する一貫した応答が観測されていることを踏まえると，この結果は，意味的に対立する語の組み合わせにおいて，形式論理に基づく推論と語の意味的整合性に基づく判断が競合した可能性を示唆している．

4.3 逆方向質問に対する回答傾向

逆方向質問に対する回答では，語の意味的関係性条件およびモデルごとに顕著な差異が見られた．無関係条件において，モデル L はすべて「いいえ」と回答しており，形式的に誤謬とされる後件肯定を一貫して否定している．一方，モデル G では「どちらとも言えない」が約 90%を占めており，逆方向推論に対して明確に否定せず，判断を保留する傾向が確認された．

類義条件では，モデル L は「いいえ」が多数を占めつつも，一定の割合で「はい」を選択しており，モデル G では「はい」「いいえ」「どちらとも言えない」が混在する分布となった．この結果は，語の意味的な近接性が，形式的には誤謬とされる逆方向推論を部分的に許容する方向に作用した可能性を示唆している．

対義条件では，両モデルとも高い割合で「いいえ」を選択しており，意味的な対立する組み合わせに対してはモデル間で概ね判断が一致した．無関係条件および類義条件との比較から，逆方向質問においては，語の意味的整合性の有無が肯定的判断に強く影響している可能性が示唆される．

5 おわりに

本稿では，LLM における誤謬の要因を分析するための初期検討として，推論方向および語の意味的関係性に着目し，代表的な二種類の LLM の推論挙動を分析した．語の意味的関係として，無関係・類義・対義の三条件を設定し，「P ならば Q」という前提に対する順方向および逆方向の質問を用いて評価した．評価の結果，順方向質問では，両モデルとも高い割合で肯定的な回答を示した．一方，逆方向推論では，無関係条件において，モデル間で明確な差が確認された．また，類義条件では，逆方向質問において一定割合の肯定的回答が観測され，語の意味的近接性が形式的には誤謬とされる推論を部分的に許容する可能性が示された．対義条件では，両モデルとも否定的な回答が多数を占めており，意味的乖離性が逆方向の判断に強く影響することが確認された．

今後の課題として，P と Q の二項関係に基づく推論から，P，Q，R の三項関係による循環的な推論構造へと拡張し，推論方向と意味的関係性の組み合わせが LLM の推論挙動に与える影響をより詳細に分析することが挙げられる．また，プロンプト設計や回答形式の違いが判断保留や誤謬の発生に与える影響についても検証する．

参考文献

[1] 今井むつみ，岡田浩之，"対称性推論は言語学習のタマゴかニワトリか：ヒト乳児とチンパンジーの直接比較"，https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/download.php/KAKEN_22653093seika.pdf?file_id=75371

[2] 今井むつみ，岡田浩之，"特集「対称性」へのコメンタリー：言語の成立にとって対称性はたまごかにわとりか"，認知科学, 15(3) 470-481, 2008

[3] 森下皓文，森尾学，山口篤季，十河泰弘，"大規模言語モデルに推論を教えるための人工論理推論コーパスを用いたアプローチ"，言語処理学会，2025 年 32 巻 2 号 p. 520-571

[4] Linhao Li，Ming Xu，Yongfeng Dong，Xin Li, Ao Wang，"Interactive Model with Structural Loss for Language-based Abductive Reasoning"，https://arxiv.org/abs/2112.00284

[5] Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S. W., and Choi, Y. "Abductive Commonsense Reasoning"，ICLR, 2020.

[6] 国立国語研究所，”分類語彙表－増補改訂版データベース - 国語研コーパスポータル”，https://clrd.ninjal.ac.jp/goihyo.html