NeuBAROCO データセットによる大規模言語モデルの推論能力の検証

森下 貴允 [1]　安東 里沙子 [1]　阿部 裕彦 [1]　小関 健太郎 [1,2]　峯島 宏次 [1]　岡田 光弘 [1]
[1] 慶應義塾大学　[2] 東京大学
{morishita,risakochaan,hirohiko-abe}@keio.jp　kentaro.ozeki@gmail.com
{minesima,okada}@abelard.flet.keio.ac.jp

概要

この論文では、「現在の大規模言語モデルが論理推論をどれくらい正確に行うことができるのか」という問いを、特に大規模言語モデルが人間と同様の推論バイアスを示すかどうかという点に着目して探究する。論理推論として、論理学だけでなく認知科学の中で人間が自然に行う演繹推論の形式として広く研究されている三段論法に注目し、NeuBAROCO という三段論法データセットを導入する。このデータセットは、もともと三段論法を用いて人間の推論能力を評価する心理実験のために設計されたもので、英語と日本語の三段論法推論を含んでいる。現在の代表的な大規模言語モデルで実験を行った結果、現行のモデルは人間と同様の推論バイアスを示し、特に含意関係が含意でも矛盾でもない推論問題で大きな改善の余地があることが示唆された。

1　はじめに

大規模言語モデル（Large Language Model, LLM）が一定の成功を収め、社会の変化に直結するような基盤技術として期待が高まる一方で、それが従来の AI に期待されてきたような正確な推論を自然言語上でどれくらい行うことができるのかはいまだ明らかではない。この論文では、人間の自然な演繹推論の形式として広く研究されている**三段論法**に注目し、LLM の推論能力を検証するための NeuBAROCO データセットを導入する。このデータセットは、もともと三段論法を用いて人間の推論能力を評価する心理実験のために設計されたもので [1]、これを転用して、LLM が人間と同様のエラー傾向、すなわち、**推論バイアス**を示すかどうかを検証する。

推論の認知科学研究 [2, 3, 4] では、人間が様々な推論バイアスを示すことが多様な実験を通して明らかにされているが、その知見を活用して機械学習モデルの評価にも利用できる推論データセットを構築する試みは十分に進展していない。NeuBAROCO は、日本語と英語のパラレル推論データセットであり、各問題についてどのような推論バイアスが関連しているのか、詳細なタグが付与されている。またその一部には一連の心理実験 [5, 1] に基づいて、人間の正解率が対応付けられている。

以下では、言語処理での三段論法推論の関連研究を紹介し（2 節）、NeuBAROCO データセットの特徴について説明した上で（3 節）、このデータセットを用いて、代表的な LLM の推論能力を検証したベースライン実験の結果について報告する（4 節）。

2　背景

2.1　三段論法

自然言語推論 (Natural Language Inference, NLI) とは、前提文と結論文（仮説文とも呼ばれる）の間に含意関係が成り立つか否かを判定するタスクである。本研究では、**含意** (entailment)、**矛盾** (contradiction)、含意でも矛盾でもない、すなわち、**中立** (neutral) という 3 つのラベルに分類する NLI タスクに着目する。三段論法とは、推論の中でも、2 つの前提と 1 つの結論から成るものであり、前提と結論は表 1 の 4 つの形の基本的な量化文から構成される。

タイプ	文	説明
A	すべての S は P である	全称肯定文
E	どの S も P でない	全称否定文
I	ある S は P である	特称肯定文
O	ある S は P でない	特称否定文

表 1　三段論法を構成する 4 つのタイプの文

例えば、次の三段論法は**含意**の例であり、前提文 (**P1**, **P2**) が真であれば結論文 (**H**) もまた真である。

P1: すべての A は B である
P2: すべての B は C である

H: すべての A は C である

次の例は、含意でも矛盾でもない**中立**の例である。

P1: すべての A は B である
P2: すべての C は B である

H: すべての A は C である

三段論法は、形式的には単項述語論理で表現可能な比較的単純な論理推論であるが、人間にとって難しい推論であることが知られている [6]。どのようなタイプの三段論法が人間の推論エラーを引き起こしやすいか、つまり、推論バイアスを伴うかは認知科学の分野で広く研究されている [2, 3, 4]。

2.2 関連研究

三段論法に着目した機械学習・深層学習モデルの研究は近年盛んに行われている。三段論法推論の学習・評価用データセットとして、Dong et al. [7] や Gubelmann et al. [8] では、WordNet 等の言語資源や独自の単語リストを用い、三段論法のタイプに基づくテンプレートベースの自動生成手法によってデータセットを構築している。Gubelmann et al. [8] のデータセットは格ラベルおよび含意、矛盾、中立の 3 値の正解ラベルを含むが、その他の情報は付加されていない。Avicenna [9] は、クラウドソーシングによって構築されたデータセットであり、2 つの文を前提として何らかの三段論法の関係が成り立つかどうかの 2 値の正解ラベル（成り立つ場合はさらに結論の文）を含むが、格ラベルなどは含まれていない。これらはいずれも英語のデータセットである。

Wu et al. [10] は、SYLLOBASE というデータセットを構築している。SYLLOBASE は、自然言語で書かれた 5 種類のタイプの三段論法を含んでおり、既存の知識ベースから自動で作成された 50,000 のデータからなる。また、テストセットとして 1,000 データが人手でアノテーションされている。Wu et al. [10] では、zero-shot および few-shot の例示と、生成タスクと選択タスクが実施されている。生成タスクでは、学習済みのモデルの正解率がより高い傾向にあり、全体的に、結果の評価で使用する指標に応じてモデルの正解率が大きく異なっている。選択タスクでは、全体の正解率は 70%程度で、定言三段論法での正解率が低かった。

Dasgupta et al. [11] および Lampinen et al. [12] は、三段論法の結論の内容が推論の妥当性の判断に影響することを明らかにしている。これらの研究では、結論の内容が私たちの信念に相反する場合、相反しない場合、結論の文が意味のない単語から構成される場合に分類し、それぞれの場合で 2 つの前提と 1 つの結論の組み合わせが三段論法の推論として妥当か非妥当かを LLM に 2 択で判定させている。これらの実験では、結論の内容が信念に相反せず、推論が妥当な場合、あるいは、結論の内容が信念に相反し、推論が非妥当な場合に、正解率が高いことが明らかになっている。なお、Lampinen et al. [12] では、LLM の生成する確率分布と人間の解答時の反応時間を確信度の尺度として比較している。LLM では、結論の内容が信念に整合する場合に、確信度がより高いことが明らかにされている。

3 NeuBAROCO データセット

3.1 BAROCO プロジェクト

本研究の背景にある初期の三段論法問題集は BAROCO と呼ばれ、大規模な被験者推論能力の研究のために設計・開発された。BAROCO は、人間の推論バイアスを伴う典型例の一つであるいわゆる信念相反課題（3.2 節を参照）を含む。また、標準的な言語的課題のほかに、空間的認知をテストするオイラー図課題も含み、これらの推論形式の相関及び双生児法による遺伝要因と環境要因の調査に用いられた [13, 1]。また、行動経済学の実験課題と組み合わせた研究も行われている [5]。

Ando et al. [14] は、本研究に先行する予備的な研究であり、BAROCO 問題集を LLM の評価に応用することを目的として BAROCO 問題集の一部を転用して、375 問の三段論法推論から構成される推論データセット NeuBAROCO を導入した。各推論課題には、含意、矛盾、中立の 3 値のラベルが人手で付与され、さらに後述の信念相反などの推論バイアスを引き起こす推論課題を含んでいる。本研究では Ando et al. [14] の NeuBAROCO 三段論法データセットをさらに整備・拡張し、同時に、被験者実験で一般的な 5 択選択課題を追加し、現在の代表的な LLM で評価実験を行う。

3.2 データセットの概要

オリジナルの BAROCO 問題集は、2 つの前提と、結論となりうる複数の選択肢で構成されている。BAROCO 問題集の各問題を NLI モデルの評価で一般

的に使用される形式に変換することで NeuBAROCO データセットを構築した。NeuBAROCO は現在、870 問の 3 択判定課題、すなわち、推論を含意・矛盾・中立に分類する課題を含む。その内訳は、含意に分類される問題が 282 問、矛盾に分類される問題が 204 問、中立に分類される問題が 384 問となっている。BAROCO 問題集はすべて日本語で書かれているが、NeuBAROCO データセットでは、それらの問題を英語に翻訳し、日本語・英語の対訳推論コーパスとして利用可能である。

NeuBAROCO データセットは、元の BAROCO 問題集の様式を踏襲し、2 つの前提と、結論となりうる 5 つの選択肢からなる 80 問の問題（5 択選択課題）を含む。各推論に対し、正解となる選択肢の番号がラベル付けされている。5 択選択課題の例を付録 A.3 に示す。この問題の正解は 4 の選択肢である。

3.3 アノテーション

推論バイアスを伴うタイプの問題を分類するため、NeuBAROCO データセットの各推論を、記号、信念相反あり、信念相反なし、Conversion の 4 つの異なるタイプに分類してラベル付けを行った。それぞれの件数と例を表 2 に示す。

記号　すべてのタームが抽象的な記号（アルファベット）の文から構成されている場合、推論は「記号」とラベル付けされる。人間にとって、これらは信念に対して中立であると考えられる。

信念相反あり　前提もしくは結論の少なくとも一つが常識的な信念と一致しない場合、推論は「信念相反あり」とラベル付けされる。表 2 の例では、「すべての動物はトマトである」と「ある人間はトマトである」が常識に反する内容となっている。

信念相反なし　前提と結論のすべてで信念相反が生じていない場合、推論は「信念相反なし」とラベル付けされる。

Conversion　三段論法の代表的な推論バイアスとして、Conversion error が知られている [2, 6]。例えば、表 2 の全称肯定文が含まれている例では、「すべての B は A である」を「すべての A は B である」と読み替えると、正解は「含意」となる。また、特称否定文が含まれている例では、「ある動物は霊長類でない」を「ある霊長類はチンパンジーでない」と解釈すれば、正解が「含意」となる。Conversion error とはこのようにタームを誤って置換することで起こるエラーのことである。このタイプの推論を区別するため、「中立」とラベル付けされている推論のうち、前提に全称肯定文もしくは特称否定文が含まれており、少なくとも一つの前提のタームを入れ替えて解釈した場合にラベルが中立から含意に変わるような推論を、Conversion とラベル付けした。

タイプ	例
記号 (170 件)	**P1**: すべての A は B である **P2**: すべての B は C である **C**: すべての A は C である
信念相反あり (244 件)	**P1**: ある動物は人間である **P2**: すべての動物はトマトである **C**: ある人間はトマトである
信念相反なし (406 件)	**P1**: 太郎のある友人はポールの友人である **P2**: ポールのすべての友人はドイツ人である **C**: 太郎のある友人はドイツ人である
Conversion (68 件)	(全称肯定文が含まれている例) **P1**: すべての B は A である **P2**: すべての B は C である **C**: すべての A は C である (特称否定文が含まれている例) **P1**: すべてのチンパンジーは動物である **P2**: ある動物は霊長類でない **C**: ある霊長類はチンパンジーでない

表 2　NeuBAROCO データセットで記号、信念相反あり、信念相反なし、Coversion とラベル付けされた推論の例

言語	モデル	正解率
英語	GPT-3.5	48.75
	GPT-4	83.75
日本語	GPT-3.5	36.25
	GPT-4	95.00

表 3　5 択選択課題における各モデルの正解率 (%, n=80)

4 実験

LLM の推論能力の評価のため、NeuBAROCO データセットを用いて以下の実験を行った。

4.1 実験設定

2 種類の課題について、問題の解答方法の指示と、本データセットから生成した問題 1 題を含むテキストを 1 回の試行の入力（プロンプト）として、問題数分の試行における言語モデルの出力を収集し、全体および分類ごとの正解率を尺度として評価した。実験には、OpenAI 社が API を提供する GPT-3.5 (gpt-3.5-turbo-1106) および GPT-4 (gpt-4-0613)[1] を言語モデルとして用いた。解答の長さを制限するため、最大出力トークン長を 10 に設定し、その他のハイパーパラメータは既定値を用いた。

2 種類の課題はそれぞれ以下の形式であり、英語と日本語の各言語で問題とプロンプトを生成した。

1）https://platform.openai.com/docs/models/

言語	モデル	全体	含意	矛盾	中立	記号	信念相反なし	信念相反あり	Conversion
英語	GPT-3.5	51.61	83.33	40.20	34.38	55.29	58.62	36.07	22.06
	(few-shot)	48.39	88.65	27.45	29.95	47.06	53.69	38.11	13.24
	GPT-4	69.89	81.21	89.71	51.04	77.65	74.14	55.74	38.24
	(few-shot)	75.29	83.69	87.75	62.50	80.00	80.05	63.93	48.53
日本語	GPT-3.5	38.16	82.27	46.57	1.30	38.82	45.32	25.82	0.00
	(few-shot)	40.23	90.43	33.82	6.77	39.41	46.80	31.15	2.94
	GPT-4	70.11	88.65	95.59	42.97	70.59	77.09	59.84	41.18
	(few-shot)	79.31	92.20	89.22	64.58	83.53	81.77	74.18	72.06

表 4　3 択判定課題における各モデルの正解率 (%, n=870)

言語	モデル	正解率
英語	GPT-3.5	48.75
	GPT-4	83.75
日本語	GPT-3.5	36.25
	GPT-4	95.00

表 5　5 択選択課題における各モデルの正解率 (%, n=80)

3 択判定課題　三段論法の大前提・小前提と仮説（結論の候補）を提示し、「含意」、「矛盾」、「どちらでもない」のいずれかを一語で解答させる。この課題では、(i) 指示と問題のみを与えるパターン（例示なし、zero-shot プロンプト）と、(ii) 正しい三段論法の例数例を指示と問題の間に挿入したパターン（例示あり、few-shot プロンプト）の 2 パターンで実験を行った。パターン (i) のプロンプトの例を付録 A.1、パターン (ii) で挿入した例示を付録 A.2 に示す。

5 択選択課題　三段論法の大前提・小前提を提示し、結論の候補 4 つと「どれでもない」の 5 つの選択肢から正しい選択肢の番号を解答させる。プロンプトの例を付録 A.3 に示す。

4.2　結果と分析

3 択判定課題と 5 択選択課題の評価結果をそれぞれ表 4 と表 5 に示す。3 択判定課題について、以下では例示なし（zero-shot プロンプト）の場合について述べるが、GPT-4 では特に例示の追加（few-shot プロンプト）によって英語と日本語のいずれにもおいても全体正解率の向上が見られ、英語で 5.40 ポイント、日本語で 9.20 ポイント高くなっている。

3 択判定課題　英語では、GPT-4 が 69.89%の全体正解率で GPT-3.5 を 18.28 ポイント上回った。他方で、正解ラベルが「中立」である問題の正解率は、GPT-3.5 から GPT-4 で向上しているものの、GPT-4 でも 51.04%と、他の正解ラベルでの正解率に比べて 30 から 40 ポイント程度低くなっている。

日本語では、GPT-3.5 の全体正解率が 38.16%に留まっているが、GPT-4 では 70.11%と英語の場合と同等以上に改善している。いずれのモデルでも、正解ラベルが「含意」である問題の正解率は高い。一方、英語・日本語ともに、正解ラベルが「中立」である問題の正解率は他の正解ラベルでの正解率に比べて GPT-3.5 および GPT-4 ともに低く、few-shot プロンプトを追加した GPT-4 でも 6 割台に留まる。

問題のタイプごとの正解率では、記号タイプに分類される問題の正解率が多くのケースで全体正解率を上回っている。また、GPT-4、GPT-3.5 とも、信念相反ありの問題は信念相反なしの問題よりも正解率が低いが、GPT-4 ではその差は少ない。Conversion error を引き起こす問題（Conversion タイプに分類される問題）の正解率は、ほとんどのケースで全体正解率より顕著に低くなっている。

5 択選択課題　GPT-4 の正解率は英語で 83.75%、日本語で 95.00%であり、GPT-3.5 をそれぞれ 35.00 ポイントと 58.75 ポイント上回り高い性能を示した。また、3 択判定課題、5 択選択課題のいずれにおいても、GPT-4 は英語より日本語で正解率が高い傾向が見られた。5 択選択課題の正答率は、3 択判定課題の含意のケースの正答率と同様に高い傾向を示している。これは、5 択選択課題では基本的に前提文と正解の結論文の関係は含意であり、中立の問題は含まれないことが影響していると考えられる。

5　おわりに

本稿では LLM の論理推論能力の検証のため、日本語・英語の三段論法からなる NeuBAROCO データセットを導入した。ベースライン実験の結果、現状の LLM は含意ラベルが「中立」の問題、特に人間の推論バイアスとして知られている Conversion error を引き起こす問題に関して大きな改善の余地があることが示唆された。オリジナルの BAROCO 問題集（5 択選択課題）では、各タイプの三段論法に対して大規模な被験者実験による正答率が付与されている。人間の正答率との詳細な比較は今後の課題の一つである。

5.1 謝辞

BAROCO 問題集の初期バージョンについて情報提供をいただいた敷島千鶴氏、佐藤有理氏に感謝します。本研究は、JST、CREST、JPMJCR2114、JSPS 科研費 JP21H00467、JP21K18339、JP21K00016 の助成を受けたものです。

参考文献

[1] Chizuru Shikishima, Kai Hiraishi, Shinji Yamagata, Yutaro Sugimoto, Ryo Takemura, Koken Ozaki, Mitsuhiro Okada, Tatsushi Toda, and Juko Ando. Is g an entity? a Japanese twin study using syllogisms and intelligence tests. **Intelligence**, Vol. 37, No. 3, pp. 256–267, 2009.

[2] Jonathan St.B. T. Evans, Stephen E. Newstead, and Ruth M. J. Byrne. **Human Reasoning: The Psychology of Deduction**. Psychology Press, 1993.

[3] Ken Manktelow. **Reasoning and Thinking**. Psychology press, 1999. K. マンクテロウ『思考と推論: 理性・判断・意思決定の心理学』服部雅史・山祐嗣（監訳）, 北大路書房, 2015.

[4] Keith Stenning and Michiel Van Lambalgen. **Human Reasoning and Cognitive Science**. MIT Press, 2012.

[5] Chizuru Shikishima, Kai Hiraishi, Shinji Yamagata, Juko Ando, and Mitsuhiro Okada. Genetic factors of individual differences in decision making in economic behavior: A Japanese twin study using the Allais problem. **Frontiers in Psychology**, Vol. 6, p. 1712, 2015.

[6] Bart Geurts. Reasoning with quantifiers. **Cognition**, Vol. 86, No. 3, pp. 223–251, 2003.

[7] Tiansi Dong, Chengjiang Li, Christian Bauckhage, Juanzi Li, Stefan Wrobel, and Armin B Cremers. Learning syllogism with euler neural-networks. **arXiv preprint arXiv:2007.07320**, 2020.

[8] Reto Gubelmann, Christina Niklaus, and Siegfried Handschuh. A philosophically-informed contribution to the generalization problem of neural natural language inference: Shallow heuristics, bias, and the varieties of inference. In **Proceedings of the 3rd Natural Logic Meets Machine Learning Workshop (NALOMA III)**, pp. 38–50, 2022.

[9] Zeinab Aghahadi and Alireza Talebpour. Avicenna: a challenge dataset for natural language generation toward commonsense syllogistic reasoning. **Journal of Applied Non-Classical Logics**, Vol. 32, No. 1, pp. 55–71, 2022.

[10] Yongkang Wu, Meng Han, Yutao Zhu, Lei Li, Xinyu Zhang, Ruofei Lai, Xiaoguang Li, Yuanhang Ren, Zhicheng Dou, and Zhao Cao. Hence, socrates is mortal: A benchmark for natural language syllogistic reasoning. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 2347–2367, 2023.

[11] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. **arXiv preprint arXiv:2207.07051**, 2022.

[12] Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. Language models show human-like content effects on reasoning tasks, 2023.

[13] Chizuru Shikishima, Juko Ando, Pierre Grialou, Ryo Takemura, and Mitsuhiro Okada. A behavioural genetic study of syllogism solving using linguistic and graphical representations: A preliminary report. In **Images and reasoning: Interdisciplinary conference series on reasoning studies**, Vol. 1, pp. 69–85. Keio University Press Tokyo, 2005.

[14] Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. Evaluating large language models with NeuBAROCO: Syllogistic reasoning ability and human-like biases. In Stergios Chatzikyriakidis and Valeria de Paiva, editors, **Proceedings of the 4th Natural Logic Meets Machine Learning Workshop**, pp. 1–11, Nancy, France, June 2023. Association for Computational Linguistics.

A 付録

A.1 3 択判定課題のプロンプト例

Determine the correct logical relationship between the given premises and the hypothesis.
- Answer “entailment” if the hypothesis follows logically from the premises.
- Answer “contradiction” if the premises and the hypothesis are logically incompatible with each other.
- Answer “neither” if the relationship is neither “entailment” nor “contradiction”.
Your answer must be one word: “entailment”, “contradiction”, or “neither”.

Premise 1: One friend of Taro is a friend of Paul.
Premise 2: All of Paul’s friends are German.
Hypothesis: All of Taro’s friends are German.
The answer is:

与えられた前提と仮説の正しい論理的関係を判定しなさい。
- 仮説が前提から論理的に導かれる場合は「含意」と答えなさい。
- 前提と仮説が論理的に両立しない場合は「矛盾」と答えなさい。
- その関係が「含意」でも「矛盾」でもない場合は「どちらでもない」と答えなさい。
「含意」「矛盾」「どちらでもない」のいずれか一語で回答しなさい。

前提 1: 太郎のある友人はポールの友人である。
前提 2: ポールのすべての友人はドイツ人である。
仮説: 太郎のすべての友人はドイツ人である。
答えは:

A.2 Few-shot プロンプトで指示と課題の間に挿入した例示（3 択判定課題）

Premise 1: Some X are Y.
Premise 2: All Y are Z.
Hypothesis: All X are Z.
The answer is: neither

Premise 1: Some X are Y.
Premise 2: All Y are Z.
Hypothesis: Some X are Z.
The answer is: entailment

Premise 1: Some X are Y.
Premise 2: All Y are Z.
Hypothesis: No X are Z.
The answer is: contradiction

前提 1: ある X は Y である。
前提 2: すべての Y は Z である。
仮説: すべての X は Z である。
答えは: どちらでもない

前提 1: ある X は Y である。
前提 2: すべての Y は Z である。
仮説: ある X は Z である。
答えは: 含意

前提 1: ある X は Y である。
前提 2: すべての Y は Z である。
仮説: どの X も Z でない。
答えは: 矛盾

A.3 5 択選択課題のプロンプト例

Select one statement from the five options provided that logically follows as a conclusion from the two premises presented in each problem. Answer by providing the number of your choice.

Premise 1: All the rings in this box are Yuki’s rings.
Premise 2: None of Yuki’s rings are gold rings.

1. All the rings inside this box are gold rings.
2. Some of the rings inside this box is a gold ring.
3. none of them.
4. None of the rings in this box are gold rings.
5. Some ring inside this box is not a gold ring.

The answer is:

各問題にある 2 つの前提の結論として成り立つ文を、5 つの選択肢の中から 1 つだけ選んでください。番号で回答してください。

前提 1: この箱の中のすべての指輪はユキの指輪である。
前提 2: ユキのどの指輪も金の指輪でない。

1. この箱の中のすべての指輪は金の指輪である。
2. この箱の中のある指輪は金の指輪である。
3. どれでもない。
4. この箱の中のどの指輪も金の指輪でない。
5. この箱の中のある指輪は金の指輪でない。

答えは: