日本語論理推論ベンチマーク JFLD の提案

森下皓文 [1]　山口篤季 [2]　森尾学 [1]　角掛正弥 [1]　友成光 [1]　今一修 [1]　十河泰弘 [1]
[1] 日立製作所 研究開発グループ　[2] シェフィールド大学
terufumi.morishita.wp@hitachi.com

概要

大規模言語モデル (LLM) はその広汎な課題解決能力が魅力を集め，日本語を含めた様々な言語で開発されてきている．しかしながら LLM は依然として論理推論を苦手としており，今後の研究が求められる．この研究を促進するため，本研究は日本語における論理推論ベンチマーク **JFLD** (**J**apanese **F**ormal **L**ogic **D**eduction) を提案する．**JFLD** を用いた評価の結果，日本語 LLM の論理推論能力は未だ未熟なことが分かった．なお，ベンチマーク・ベンチマーク構築用コード・LLM 評価用コードを公開する[1)]．

1　はじめに

近年登場した大規模言語モデル (LLM) は，広範な課題を巧みに解決し，人工知能＝“人間のように考える機械” [1] の実現を予感させる．人工知能の実現に向けて古くから，**知識**と**推論**という２つの要素が重要視されてきた [2, 3, 4, 5, 6, 7]．自然言語処理の文脈で知識とは，世界に関する事実であり，例えば「質量を持つものは重力場を発生させる」「地球は質量を持つ」等である．一方で推論とは，複数の知識を一定の規則に従って組み合わせることで，新たな知識を得る思考形態である．例えば上述の知識に対して推論規則「“$\forall x, F(x) \rightarrow G(x)$” と “$F(a)$” の２つから，“$G(a)$” が導ける」を適用することにより ($F$=“質量を持つ”, G=“重力場を発生させる”, a=“地球”),「地球は重力場を発生させる」という新たな知識を得る．

最近の観察によると，LLM は推論ではなく “暗記した知識” によって課題を解いていることが示唆される [8, 9, 10]．例えば「過去年度のコーディング試験は解けるが，最新年度は解けない」「算数の有名問題をそのまま出題すれば解けるが，数字を変えると解けない」等，一見推論しているように見えても，実は事前学習コーパス中の類似事例を取り出して解いていたことが分かる．この知識偏重は GPT-4 [11] のような最先端の LLM でまで確認されている [12, 13, 14, 15].

もし LLM が推論を苦手とするのであれば，汎用性の高い人工知能の実現にとって問題となる．なぜならば，知識によって課題を解いている限り，“一度見たことがある” 課題を超えた真に未知なる課題を解くことができないからだ．よって今後，LLM の推論能力向上に関する研究が不可欠である．

研究の促進のためには，良質なベンチマークが必要である．推論の中でも基本的な**論理推論**に関して，これまでも多くのベンチマークが提案されてきた [16, 17, 18, 19]．これらは，各 LLM の性能評価は勿論のこと，創発現象 [20] や反実仮想への脆弱性 [12] といった有用な知見を提供してきた．

しかしながら，これらベンチマークは主に英語が中心であり，日本語 LLM の論理推論能力を評価することはできない．日本語ベンチマークとして JGLUE [21]・JaQuAD [22] 等が有名だが，これらは知識の暗記によっても解ける可能性が高い．また，NLI や RTE [23, 24, 25, 26, 21, 27, 28, 29] もしばしば常識知識を必要とするので，純粋な論理推論能力のみを問うことができない．以上より，日本語における論理推論ベンチマークの構築が必要である．

そこで我々は**日本語の論理推論ベンチマーク JFLD** (**J**apanese **F**ormal **L**ogic **D**eduction) を提案する．**JFLD** は論理推論の基本である演繹推論能力を評価する．即ち，事実群と仮説が与えられ，多段の論理ステップ (多段演繹推論) を経て仮説を証明・反証する (図 1)．**JFLD** の重要な特徴は，1. 知識と切り分けられた純粋な論理推論能力を評価する，2. 多様な演繹規則を問う，ことである (2 節).

更に，**JFLD** を用いて様々な日本語 LLM を評価し洞察を共有する．特に，GPT-4 に比べて日本語 LLM の論理推論能力は大きく劣ること，今後の方向性として大規模な論理推論データセットでの学習が有望であること，等が重要である．

なお，本ベンチマークとコードを公開する[1)]．

1) https://github.com/hitachi-nlp/FLD

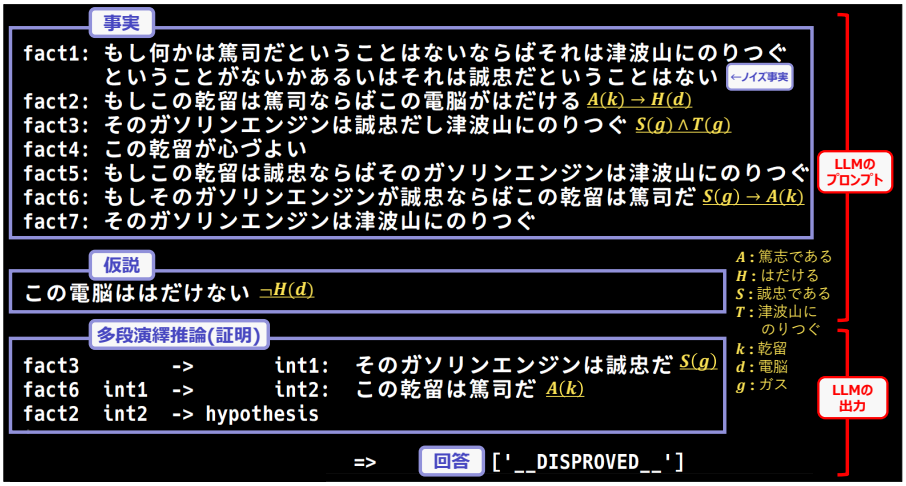

図 1: **JFLD** のサンプル．事実と仮説が与えられ，仮説を証明・反証するための多段演繹推論と回答を生成する．事実が反実仮想的なので (i) 暗記した既存知識では解けず，純粋な論理推論能力のみを評価できる (2 節) (ii) 演繹規則の任意性を学習させることができる (付録 A.2).

2 ベンチマークの設計指針

本節では，数理論理学を訪れつつ純粋な論理推論能力とは何かを考え，ベンチマークの設計指針を定める．まず，以下の一段論理ステップを考えてみる:

$$\frac{\text{1. 地球は太陽を周回している} \qquad \text{2. もし地球が太陽を周回すれば，地球には四季がある.}}{\text{地球には四季がある.}} \tag{1}$$

2 つの前提知識から結論を論理的に導いており，妥当である．次に，別の論理ステップを考える:

$$\frac{\text{1. 地球は太陽を周回している} \qquad \text{2. もし地球が太陽を周回すれば，地球には四季がない.}}{\text{地球には四季がない.}} \tag{2}$$

(地軸が傾いている我々の宇宙では) 前提知識 2 や結論は偽であるが，**もし仮に前提知識が正しかったとしたら結論は論理的に導かれる．** よって (2) も**論理的には妥当**である．最後に:

$$\frac{\text{1. ぴよぴよがある} \qquad \text{2. もしぴよぴよがあれば，ぽよぽよもある.}}{\text{ぽよぽよがある.}} \tag{3}$$

「ぴよぴよ」や「ぽよぽよ」が何かは不明だが，(3) も論理的に妥当であることだけは分かる．以上 (1)–(3) は記号を用いて**演繹規則**に抽象化できる:

$$\frac{\mathcal{F} \qquad \mathcal{F} \to \mathcal{G}}{\mathcal{G}}\ \text{modus ponens (モーダス・ポネンス)} \tag{4}$$

ここまでの考察によれば，演繹規則の論理的妥当性は，$\mathcal{F}$ や $\mathcal{G}$ の内容の正しさ (~ 知識的な正しさ) には依存せず (即ち $\mathcal{F}$ や $\mathcal{G}$ は**任意**)，あくまで前提から結論が「論理的に導かれるか」のみに依存する．**知識的な正しさと論理としての妥当性は別もの**である．

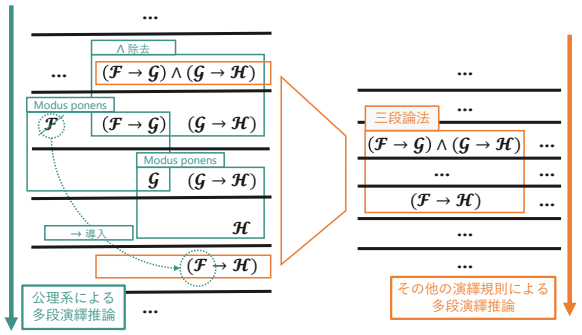

図 2: 公理系による多段演繹推論は，その他の任意の演繹規則を表現できる．

さて，人間にとっては，演繹規則 (4) を念頭に推論 (1)–(3) をこなすことは容易である．おそらく LLM も，(1) の前提を与えて結論を系列生成することはできよう．事前学習コーパスに類似のサンプルが存在し系列尤度が高いからである．しかしこれは，LLM が，演繹規則 (4) を，特に $\mathcal{F}$ や $\mathcal{G}$ の任意性を理解していることを，必ずしも意味しない．LLM が演繹規則 (4) を理解しているかどうかは，(2), (3) のような反実仮想的な前提知識下で結論を論理的に導けるかを問うて初めて浮き彫りになる．そこで:

- **設計指針 1: 反実仮想的な推論サンプルにより，LLM が演繹規則を理解しているのかを問う.**

さて，(4) 以外にも演繹規則は存在する:

$$\frac{(\mathcal{F} \wedge \mathcal{G})}{\mathcal{F}}\ \frac{(\mathcal{F} \wedge \mathcal{G})}{\mathcal{G}}\ \wedge \text{除去} \tag{5}$$

$$\frac{(\mathcal{F} \to \mathcal{G}) \wedge (\mathcal{G} \to \mathcal{H})}{\mathcal{F} \to \mathcal{H}}\ \text{三段論法} \tag{6}$$

以上のように，前提・結論として考えられる論理式の種類は無限なので**演繹規則は無限種類存在する．**

最後に，多段演繹推論 (図 2 左) では，演繹規則を複数ステップ積み重ねて結論を導く．実は三段論法 (6) はより「原子的な」演繹規則による多段演繹推論により表現できる (図 2 右)．実際，**公理系** (図 3) と呼ばれる原子的な演繹規則群が存在し，以下を満たす:

定理 1 (一階述語論理の完全性 Gödel, 1930)**.** 全ての妥当な演繹規則は公理系による多段演繹推論によって表現できる．

さて，LLM が様々な演繹規則を扱えるかを問いたいが，ベンチマークに無限種類の演繹規則を含めることはできない．しかし，定理 1 によると，公理系による多段演繹推論さえ扱えれば，他の演繹規則も実効的に扱えることになる．そこで，

- **設計指針 2: 公理系を用いた多段演繹推論をサンプルとする．定理 1 により，様々な演繹規則を扱えるかを問うことができる．**

3 ベンチマークの構築手法

3.1 推論サンプルの自動生成

推論サンプルの構築方法として，自動生成のアプローチ [18, 30, 19] が研究されてきている．人手構築に対するメリットは: 1. 大量の，2. 演繹規則に厳格に従った，3. 知識に囚われない反実仮想的な，サンプルを生成しやすいことである．

本研究では，著者らが提案した英語の推論サンプル自動生成機構 **FLD** [31, 32, 33] を日本語に拡張する．**FLD** はまず，図 2 左のような，公理系による多段演繹推論サンプル生成する (**設計指針 2**)．この際，各論理ステップで使われる演繹規則を可能なものからランダムに選ぶことにより，多様なサンプルを生成する．次に，サンプル中の各論理式を，テンプレートと語彙割り当てを用いて英語に変換する．語彙割り当てにランダム性を持たせることにより，反実仮想的なサンプルを生成する (**設計方針 1**)．**JFLD** ではテンプレートと語彙割り当てを日本語化する[2)]

3.2 日本語テンプレート・語彙割り当て

各論理式に対する日本語テンプレートを，計約 4,000 件を人手により作成した．例を示す:

表 1: **JFLD** のデータセット群 (低難易度順)．それぞれ train/valid/test=30k,5k,5k サンプルからなる．

	木の深さ	木の分枝	**論理ステップ数**	**ノイズ事実数**
D1$^-$	1	-	1 - 1	0
D1	1	-	1 - 1	0 - 20
D3	1-3	✓	1 - 8	0 - 20
D8	1-8	✓	1 - 13	0 - 20

$$\forall x, F(x) \rightarrow G(x) : F \text{ なものは } G \text{ だ}$$
$$: \text{何かが F なら，それは G だ} \quad (7)$$

次に，(7) 中の F, G, a, b のような各記号に対して，語彙を割り当てる．反実仮想的な例を作成するため，以下のような文法制約は満たしつつ，語彙を多言語 WordNet [34] からランダムに割り当てる:

- 述語 (F や G) は，日本語の述語（動詞・名詞・形容動詞）を割り当てる．
- 定数 (a と b) は，日本語の名詞を割り当てる．

以上の結果得られる割り当ての例を以下に示す：

$$F : \text{“頑健”} \quad G : \text{“腐敗”}$$
$$\forall x, F(x) \rightarrow G(x) : \text{頑健なものは腐敗だ} \quad (8)$$

最終的に図 1 のようなサンプルが得られる．図 1 における「事実」が多段演繹推論の前提であり，「仮説」が多段演繹推論の結論に相当する．

3.3 データセット統計

JFLD には様々な難易度のデータセットを含めた (表 1)．「木の深さ・分枝」は多段演繹推論木[3)] の複雑度合いを制御し，**論理ステップ数**を決める．***ノイズ事実数***は，推論に無関係な事実の数を表す．これが多いほど，LLM が誤った事実を多段演繹推論に含める可能性が高くなるので，難易度が上がる．

4 日本語 LLM の性能評価実験

日本語 LLM(表 5) をまず train set で fine-tuning し，次に，test set 上の性能を評価した．fine-tuning に用いるサンプル数 n を $5 - 10,000$ の範囲で変えた．学習の詳細は付録 A.1．また，GPT-4 の性能 ($n = 5$, 文脈内学習) も評価した．

LLM は，事実と仮説を入力として，多段演繹推論[4)] と回答を生成する (図 1)．性能指標として，回答正解率と証明正解率 [33] を用いる．回答正解率は，最終的な回答 (証明された/反証された/不明) が

2) 論理推論能力を問うなら，論理式のままでもよい，と思われるかもしれないが，推論能力の “日本語上での” 運用能力の学習・評価を行うために日本語化は必要だ (付録 A.2).

3) **FLD** では多段演繹推論を木構造で表現している．

4) “fact12 ->” のような記号も生成する

表 2: LLM の証明正解率．n は学習サンプル数．

	D1-					D1					D3					D8				
	n=5	100	1,000	10,000	30,000	5	100	1,000	10,000	30,000	5	100	1,000	10,000	30,000	5	100	1,000	10,000	30,000
GPT-4	82.1	-	-	-	-	38.6	-	-	-	-	10.9	-	-	-	-	0.9	-	-	-	-
rinna-4B	36.8	51.3	93.3	97.2	99.7	20.2	6.8	16.4	30.8	64.4	3.5	8.9	14.7	31.3	27.3	1.8	9.5	23.3	32.9	32.7
line-4B	31.9	61.1	90.8	95.8	99.7	14.7	11.9	25.3	44.0	81.5	0.0	10.3	14.0	34.1	37.6	1.8	11.3	26.6	34.1	38.1
stablelm-7B	32.2	57.2	94.1	98.9	99.9	19.5	10.5	32.7	77.7	93.1	0.0	5.6	13.7	44.5	68.6	0.0	6.4	18.7	39.6	44.4
calm2-7B	37.6	60.8	93.3	98.9	99.5	26.7	9.3	36.3	77.4	93.2	0.0	5.8	12.7	45.1	69.9	0.0	9.2	20.9	39.2	47.4
weblab-10B	32.9	61.4	94.8	99.7	100	13.4	11.1	37.2	76.1	94.2	0.0	9.9	18.0	45.8	64.9	1.3	8.1	22.6	39.7	43.4
plamo-13B	32.2	57.5	94.7	98.0	100	18.5	11.3	37.0	77.9	93.7	0.0	6.2	18.0	48.4	69.9	0.2	11.7	20.9	39.9	45.8
llmjp-13B	36.6	71.9	95.9	98.8	99.9	19.6	8.3	47.8	74.8	94.3	0.0	7.3	23.3	43.0	66.5	3.5	12.7	16.0	39.3	47.3
stockmark-13B	37.3	66.9	94.0	99.3	100	12.6	12.7	53.2	87.6	96.6	0.0	7.8	28.1	57.6	72.3	0.0	9.5	27.7	41.7	47.7
elyza-13B	35.3	66.4	97.4	99.3	100	4.4	20.6	66.9	90.8	96.9	0.0	9.2	40.4	70.0	82.0	0.0	12.3	31.9	46.9	53.7
swallow-13B	36.3	82.7	98.1	99.9	100	22.3	21.9	71.6	91.0	98.2	0.8	8.3	42.9	69.5	81.6	1.5	8.6	30.1	44.1	54.2
swallow-70B	34.2	91.4	98.0	100	100	9.9	36.2	81.6	97.4	100	0.0	25.7	50.7	82.2	91.4	0.0	13.8	37.5	54.6	65.1

表 3: LLM(weblab-10B-instruct) が生成した，誤った論理ステップ (多段演繹推論のうちの 1 ステップ)．

LLM が選んだ前提事実	LLM が生成した結論
1. その向性は跡見学園女子大学短期大学部を送り届ける **2.** もしあの土管が跡見学園女子大学短期大学部を送り届けるならばこのはたはたは危なっかしい	このはたはたが跡見学園女子大学短期大学部を送り届けるかあるいはそれが段物であるか両方である
1. あの地区は遅谷であるしそれは唱える **2.** 「騒々しいし茶臼台を騒げるということがない」ものがある	あの地区は遅谷である
1. 「この歩兵は安良里であるが退城ということはない」ということは成り立たない **2.** 「この歩兵は安良里であるがそれが退城ということはない」ということが成り立たないならその歩兵はニッコーである	その歩兵はニッコーでない

正しいかを判定する．証明正解率は，最終的な回答が正しく，かつ，多段演繹推論も正しいかどうかを判定する，より厳格な指標である．

5 結果と考察

5.1 定量評価 - 証明正解率

表 2 に各 LLM の証明正解率を示す．まず，GPT-4 の few-shot(n = 5) 性能について，低難易度データセット (D1$^-$, D1) はそこそこ解けるが，高難易度データセット (D3, D8) の正解率は不十分であった．

few-shot 設定における日本語 LLM の性能は GPT-4 より更に低かった．回答正解率によって性能を評価した場合 (付録 A.3)，GPT-4 と差は更に開いた．以上より，日本語 LLM は，事前学習の時点では十分な論理推論能力を獲得できていないことが分かる．

日本語 LLM 同士の比較では，概ね，大きなモデルの方が性能も高かった．

以上を総合すると，日本語 LLM の論理推論能力は，1. モデルの巨大化や事前学習の質・量の向上によってある程度の改善を見込めるが，2. (GPT-4 のように) 完全には届かない，と考えられる．

一方どのデータセットにおいてもサンプル数 n の増加に伴い性能が改善した．よって，より大きな論理推論データセットでの学習が有望である．

elyza の事前学習コーパスのほとんどは英語であり，日本語の量は他の LLM より遙かに小さい．それにも関わらず elyza は，その他の日本語 LLM と同等以上の性能を示した．論理推論能力は言語間で転移可能だからではないかと考える．

5.2 定性評価 - 多段演繹推論の誤り分析

LLM によって生成された誤った論理ステップの例を表 3 に示す．最初の例は「論理的幻覚」とでも言うべきもので，生成された結論は，前提事実から論理的に導けない．第二の例では，結論とは関係ない前提事実 2 が選ばれている．第三の例からは，LLM は否定の論理的意味を理解できていないことが分かる．以上より，日本語 LLM はまだ論理の基礎を理解できていないことが示唆される．

6 結論

日本語論理推論ベンチマーク **JFLD** を提案した．実験により日本語 LLM の論理推論能力が不十分であることを示した．今後は，巨大な論理推論コーパスでの学習や，学習で得られた論理推論能力が広範なタスクに転移可能かどうか，等を調査したい．

謝辞

本研究は，計算機リソースとして，産総研の AI 橋渡しクラウド (ABCI) を用いた．東京工業大学の岡崎直観先生には，内容に関するアドバイスを頂いた．日立製作所の清水正明氏には，社内大規模計算機環境の維持管理をして頂いた．感謝申し上げる．

参考文献

[1] J McCarthy, ML Minsky, and N Rochester. A proposal for the dartmouth summer research project on artificial intelligence. 1955.
[2] John W. McCarthy. Programs with common sense. In *Proc. Tedding Conf. on the Mechanization of Thought Processes*, pp. 75–91, 1959.
[3] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, Vol. 9, No. 1, pp. 36–45, 1966.
[4] T Winograd. Procedures as a representation for data in a computer program for understanding natural language, mit ai technical report 235. 1971.
[5] A. Colmerauer and P Roussel. The birth of prolog. *The ALP Newsletter*, 1973.
[6] eh Shortliffe. Computer based medical consultations: Mycin. *Elsevier*, 1976.
[7] Charles Elkan and Russell Greiner. Building large knowledge-based systems: Representation and inference in the cyc project: Db lenat and rv guha, 1993.
[8] Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854, 2022.
[9] Damian Hodel and Jevin West. Response: Emergent analogical reasoning in large language models, 2023.
[10] 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.
[11] OpenAI. Gpt-4 technical report. *ArXiv*, Vol. abs/2303.08774, , 2023.
[12] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4, 2023.
[13] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks, 2023.
[14] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2023.
[15] Melanie Mitchell. Can large language models reason? *blog*, pp. https://aiguide.substack.com/p/can–large–language–models–reason, 2023.
[16] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1930–1940, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
[17] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics.
[18] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3882–3890, 2021.
[19] Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, August 2021. Association for Computational Linguistics.
[20] Barret Zoph, Colin Raffel, Dale Schuurmans, Dani Yogatama, Denny Zhou, Don Metzler, Ed H. Chi, Jason Wei, Jeff Dean, Liam B. Fedus, Maarten Paul Bosma, Oriol Vinyals, Percy Liang, Sebastian Borgeaud, Tatsunori B. Hashimoto, and Yi Tay. Emergent abilities of large language models. *TMLR*, 2022.
[21] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.
[22] ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. JaQuAD: Japanese Question Answering Dataset for Machine Reading Comprehension, 2022.
[23] Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Shuming Shi, Teruko Mitamura, Noriko Kando, et al. Overview of the recognizing inference in text (rite-2) at ntcir-10. In *Ntcir*. Citeseer, 2013.
[24] Hideki Shima, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi, and Koichi Takeda. Overview of ntcir-9 rite: Recognizing inference in text. In *Ntcir*, 2011.
[25] Yoshikoshi Takumi, Kawahara Daisuke, and Kurohashi. Sadao. https://nlp.ist.i.kyoto-u.ac.jp/?2020.
[26] Hitomi Yanaka and Koji Mineshima. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, Vol. 10, pp. 1266–1284, 2022.
[27] Yuta Hayashibe. Japanese realistic textual entailment corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6827–6834, Marseille, France, May 2020. European Language Resources Association.
[28] Hitomi Yanaka and Koji Mineshima. Assessing the generalization capacity of pre-trained language models through japanese adversarial natural language inference. In *Proceedings of the 2021 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP2021)*, 2021.
[29] Tomoki Sugimoto, Yasumasa Onoe, and Hitomi Yanaka. Jamp: Controlled Japanese temporal inference dataset for evaluating generalization capacity of language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 57–68, Toronto, Canada, July 2023. Association for Computational Linguistics.
[30] Gregor Betz, Christian Voigt, and Kyle Richardson. Critical thinking for language models. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pp. 63–75, Groningen, The Netherlands (online), June 2021. Association for Computational Linguistics.
[31] 森下皓文, 森尾学, 山口篤季, 十河泰弘. 形式論理学に基づく演繹コーパスによる言語モデルに対する演繹推論能力の付与. 言語処理学会予稿集, 2023.
[32] 森下皓文, 森尾学, 山口篤季, 十河泰弘. 人工演繹推論コーパスによる学習は言語モデルをどのように強化するか? 人工知能学会全国大会論文集, 2023.
[33] Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Learning deductive reasoning from synthetic corpus based on formal logic. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 25254–25274. PMLR, 23–29 Jul 2023.
[34] Francis Bond and Ryan Foster. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1352–1362, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.
[35] Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12284–12314, Toronto, Canada, July 2023. Association for Computational Linguistics.
[36] Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data, 2022.
[37] Zhangdie Yuan, Songbo Hu, Ivan Vulić, Anna Korhonen, and Zaiqiao Meng. Can pretrained language models (yet) reason deductively? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1439–1454, 2023.
[38] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.
[39] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023.

A 参考情報

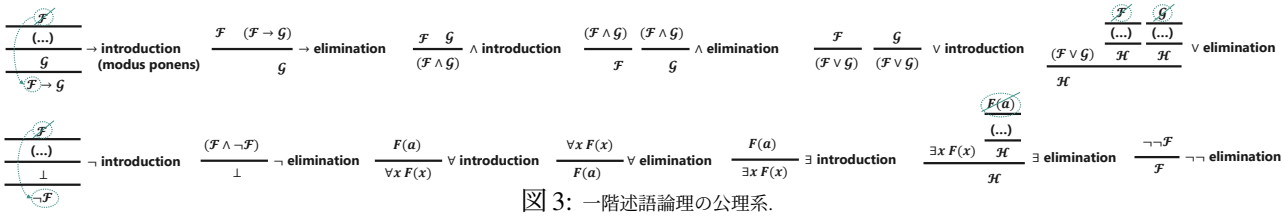

図 3: 一階述語論理の公理系.

表 4: LLM の回答正解率. n は学習サンプル数.

	D1-					D1					D3					D8				
	n=5	100	1,000	10,000	30,000	5	100	1,000	10,000	30,000	5	100	1,000	10,000	30,000	5	100	1,000	10,000	30,000
GPT-4	83.2	-	-	-	-	60.4	-	-	-	-	39.6	-	-	-	-	37.6	-	-	-	-
rinna-4B	38.8	53.0	94.8	98.9	99.9	33.9	43.8	55.1	72.5	82.6	31.9	39.5	49.6	44.2	55.4	26.4	38.7	40.0	37.6	38.3
line-4B	35.9	64.2	92.5	98.0	99.7	37.6	40.2	59.2	72.4	89.8	30.4	37.0	44.6	39.3	58.1	25.9	37.0	40.7	36.8	40.6
stablelm-7B	32.2	59.5	94.6	99.3	99.9	33.9	41.0	62.3	83.4	94.2	30.5	37.1	48.1	59.2	73.4	28.4	40.5	40.6	40.4	45.2
calm2-7B	38.4	63.5	94.6	99.7	99.5	32.1	48.1	63.8	85.1	93.8	32.6	40.0	51.5	62.3	73.9	35.7	38.0	46.2	40.3	48.9
weblab-10B	35.5	64.0	95.6	99.8	100.0	36.1	45.8	64.2	81.1	95.0	31.9	39.5	47.1	54.3	68.3	27.0	37.8	42.6	41.2	43.9
plamo-13B	37.1	60.2	95.7	98.1	100.0	34.1	37.7	61.3	83.6	94.1	28.6	38.2	50.2	59.3	75.7	19.6	47.5	43.7	40.5	46.4
llmjp-13B	37.3	75.0	96.5	99.8	99.9	33.4	40.5	65.8	82.1	95.5	35.4	38.6	57.8	57.8	74.4	28.4	40.2	48.4	40.6	50.7
stockmark-13B	42.8	69.4	94.9	99.3	100.0	36.5	52.1	69.7	89.9	97.2	33.1	39.4	56.7	67.5	75.5	28.6	42.4	48.1	42.5	49.5
elyza-13B	36.6	68.9	97.8	99.3	100.0	36.8	50.9	74.1	91.9	98.0	36.3	48.3	64.2	77.2	84.9	30.8	41.8	49.7	47.8	55.0
swallow-13B	39.6	84.7	98.2	99.9	100.0	34.6	49.9	80.3	92.3	98.6	33.0	38.6	65.4	75.2	84.3	25.7	41.1	50.1	45.4	55.2
swallow-70B	34.2	92.8	99.3	100.0	100.0	34.2	59.2	82.9	98.0	100.0	46.7	42.1	66.4	83.6	92.8	32.2	40.8	53.9	55.9	67.8

A.1 実験設定

表 5: 日本語 LLM の詳細.

名称	事前学習トークン数	huggingface hub 上の名前
rinna	300B	japanese-gpt-neox-3.6b-instruction-ppo
line	- (600GB)	japanese-large-lm-3.6b
stablelm	750B	japanese-stablelm-base-alpha-7b
calm	1300B	open-calm-7b
weblab	600B	weblab-10b
plamo	1500B (en+jp)	plamo-13b
llmjp	300B	llm-jp-13b-v1.0
stockmark	200B	stockmark/stockmark-13b
elyza	2000B (en) + 20B (jp)	ELYZA-japanese-Llama-2-7b-fast
swallow	2000B (en) + 600B (jp)	Swallow-13b-hf/Swallow-70b-hf

LLM の評価では文脈内学習がよく使われるが，**JFLD** のサンプルは 1k トークン以上と日本語 LLM の文脈に収まりづらい．そこで，fine-tuning による評価を行った．[35] によると，fine-tuning と文脈内学習は適切な実験設定において，同等の結果をもたらす．そこで我々も [35] に従う：学習率は 1e-05，バッチサイズは 32，300 勾配ステップ．ただし過学習を防ぐためエポック数を最大 50 に制限し，n=5 では 50，n=100 では 156 勾配ステップとした．実験は 3 つの seed で行った．その他はコード[1)]参照のこと．

A.2 裏設計指針

- **疑問 1:** 論理推論能力の評価が目的なら，論理式で書かれたサンプル (例: 図 2) だけで十分で，日本語化は不要では？
- **疑問 2:** 日本語化するとしても「WordNet からランダムな語彙を割り当てることで反実仮想的な...」などと大仰な仕掛けは必要無い．例えば「赤いぷよぷよは優しい」「優しいモンスターは強い」というように限られた語彙で十分では？その方が学習が簡単だし，人間にとっても分かりやすい．

疑問 1 への回答: 我々の最終目的は，日本語において，知識と推論を統合運用できる機械を作ることである．この第一歩として本研究では，純粋な推論に限ったベンチマークを作成した．**最終目標から考えて，ベンチマークには，「論理推論能力の “日本語上での” 運用能力を学習・評価できること」が求められる．よって，各サンプルを日本語で書き下している．**この回答には**「LLM が仮に論理式での推論を遂行できても，日本語でできるとは全く限らない．よって日本語での推論能力を問うためには，日本語のサンプルが必要である．**」という前提がある．

真に論理を理解していれば，即ち，例えば modus ponens(4) と $\mathcal{F}, \mathcal{G}$ の任意性を理解していれば，それが日本語で書かれていようが，推論を実行できるであろう (論理記号 → から “ならば” への変換のような自明な処理は当然実現できるとする)．実際これは，人間にとっては容易である．

しかし，LLM (Transformer) は帰納的な系列学習器なので，学習時と似ている条件付系列を生成することのみしかできない．学習データ次第では，論理式での推論は遂行できるが日本語での推論は遂行できない，ということは十分に起こりうる．

実際，このような挙動が，我々の予備実験で確認されている．適当な LLM を持ってきて，論理式で書かれた **JFLD** で追加学習する．この LLM に，論理式で書かれた推論問題を与えると，よく解ける．しかし，それと等価な日本語の推論問題を与えても，全く解けない．生成結果を見てみると，論理式データセットの学習する前とほとんど変化が無い: 論理とは程遠い，(おそらく事前学習の時点で学習していた) 系列を生成する．**結局，論理式データセットでの学習で得られたのは，$\mathcal{F}$ や $\mathcal{G}$ といったアルファベットが入力された際にのみ論理的に振る舞う機械**である．

なぜこのような現象が起こるというと，論理式で書かれた modus ponens(4) を見せられても，それを日本語の空間まで汎化させて良いかどうかを判断する材料は，そのサンプル中には存在しない．人間は演繹的な心を持っているので**勝手**に汎化させてしまうが，帰納的な機械はそうはしない．尚，このような論理推論における分布外への非汎化性は [36, 37] でも確認されている．

疑問 2 の回答の回答は，上記考察から従う．JFLD を日本語での論理推論能力を獲得するための学習データとして活用することを考える．$\mathcal{F}$ や $\mathcal{G}$ に入れる語彙を「ぷよぷよ」「赤い」などに限った場合，学習された LLM は「ぷよぷよ」「赤い」(とせいぜいその類義語) が入ってきた場合にのみ論理的に振る舞うようになる．**$\mathcal{F}$ や $\mathcal{G}$ の任意性を学習させるためには，実際に $\mathcal{F}$ や $\mathcal{G}$ に任意の日本語を入れて見せる必要があるのである．**このため本研究では，大語彙からランダムに選んだ日本語を割り当てている．

A.3 証明正解率の結果と考察

表 4 に回答正解率の結果を示す．5.1 節で論じた証明正解率は，回答 (証明/反証/不明) に加えて多段演繹推論の正しさも要求するため，厳格な指標であった．一方，回答正解率は回答の合致のみを要求するため，random guess でも 33.3% と緩い指標である．GPT-4 の回答正解率は，証明正解率を大幅に上回っており，結果として日本語 LLM との性能差もより大きい．GPT-4 にが生成した多段演繹推論を分析すると，GPT-4 は**間違った推論で正しい回答を生成していることがある**ことが多い．これは，GPT-4 が自ら生成した “推論” に必ずしも忠実に従わず，モデルの内部で完結して正しい推論を実現している可能性を示唆する．よって，GPT-4 の論理推論能力を計測する場合，証明正解率は過小評価に繋がっている可能性がある．尚，LLM が自ら生成した思考系列に従わない，という観察は別の研究 [38, 39] でも得られている．