JFLD: A Japanese Benchmark for Deductive Reasoning based on Formal Logic

Terufumi Morishita[1], **Atsuki Yamaguchi**[2*†], **Gaku Morio**[1*], **Hikaru Tomonari**[1], **Osamu Imaichi**[1], **Yasuhiro Sogawa**[1]

[1] Advanced AI Innovation Center. Hitachi, Ltd. [2] The University of Sheffield

Abstract

Large language models (LLMs) have proficiently solved a broad range of tasks with their rich knowledge but often struggle with logical reasoning. To foster the research on logical reasoning, many benchmarks have been proposed so far. However, most of these benchmarks are limited to English, hindering the evaluation of LLMs specialized for each language. To address this, we propose **JFLD** (**J**apanese **F**ormal **L**ogic **D**eduction), a deductive reasoning benchmark for Japanese. **JFLD** assess whether LLMs can generate logical steps to (dis-)prove a given hypothesis based on a given set of facts. Its key features are assessing pure logical reasoning abilities isolated from knowledge and assessing various reasoning rules. We evaluate various Japanese LLMs and see that they are still poor at logical reasoning, thus highlighting a substantial need for future research.

Keywords: Japanese, benchmark, logical reasoning, language model

1. Introduction

Large language models (LLMs) have proficiently solved a broad range of tasks, making significant advancements towards realizing artificial intelligence as "A machine that thinks like humans" (McCarthy et al., 1955). Historically, two critical elements, knowledge and reasoning, have been emphasized for achieving artificial intelligence (McCarthy, 1959; Weizenbaum, 1966; Winograd, 1971; Colmerauer and Roussel, 1973; Shortliffe, 1976; Elkan and Greiner, 1993). In the context of natural language processing, knowledge refers to facts about the world, such as "objects with mass generate gravitational field" and "the Earth has mass." Reasoning, on the other hand, involves combining multiple pieces of knowledge following specific rules to generate new knowledge. For instance, applying the reasoning rule "From '$\forall x, F(x) \rightarrow G(x)$" and "$F(a)$", derive "$G(a)$" to the aforementioned knowledge (where F="has mass", G="generates gravitational field", a="Earth") yields the new knowledge that "the Earth generates gravitational field."

Recent observations suggest that LLMs solve tasks based on "memorized knowledge" rather than reasoning. Such observations include: (i) LLMs can solve past coding exams but not the most recent ones, or (ii) LLMs can solve famous arithmetic problems unchanged but fail when the numbers are altered (Razeghi et al., 2022; Hodel and West, 2023; Dasgupta et al., 2023). These observations reveal that LLMs rely on similar instances in their training corpora to solve the tasks. This tendency towards knowledge reliance has been confirmed even in state-of-the-art LLMs like GPT-4 (OpenAI, 2023) (Liu et al., 2023; Wu et al., 2023; Dziri et al., 2023; Mitchell, 2023).

If LLMs struggle with reasoning, this poses a challenge for achieving versatile artificial intelligence, as they would be limited to solving tasks they have encountered before, unable to tackle genuinely novel challenges. Hence, research for enhancing LLMs' reasoning abilities is essential.

To foster the research on reasoning, high-quality benchmarks are crucial. Indeed, numerous benchmarks have been proposed for the fundamental *logical reasoning*, providing not only performance evaluations of each LLM (Habernal et al., 2018; Niven and Kao, 2019; Clark et al., 2021; Tafjord et al., 2021) but also insights, such as emergent phenomena (Zoph et al., 2022) and vulnerabilities to counterfactuals (Liu et al., 2023).

However, these benchmarks primarily focus on English, lacking in evaluating Japanese LLMs' logical reasoning abilities. While Japanese benchmarks like JGLUE (Kurihara et al., 2022) and JaQuAD (So et al., 2022) are well-known, their problems should often be solved by knowledge. Tasks such as NLI and RTE (Watanabe et al., 2013; Shima et al., 2011; Takumi et al., 2020; Kurihara et al., 2022; Hayashibe, 2020; Yanaka and Mineshima, 2021; Sugimoto et al., 2023) frequently require common-sense knowledge, thus not exclusively testing logical reasoning abilities. Hence, there is a necessity for a Japanese logical reasoning benchmark.

This paper introduces such a benchmark, **JFLD** (**J**apanese **F**ormal **L**ogic **D**eduction), a deductive reasoning benchmark for Japanese. We showcase an example from **JFLD** in Figure 1, which assesses whether LLMs can generate logical steps to (dis-)prove a given hypothesis based on a given set of facts. Its key features are assessing pure logical reasoning abilities isolated from knowledge and assessing various reasoning rules. We extended a previous framework called **FLD** (Morishita et al., 2023) into Japanese to generate such examples.

Further, we evaluate various Japanese-specialized LLMs and share insights. Most critically, these LLMs

*Contributed equally

†Work done at Hitachi, ltd.

LREC-COLING 2024, pages 9526–9535
20-25 May, 2024.

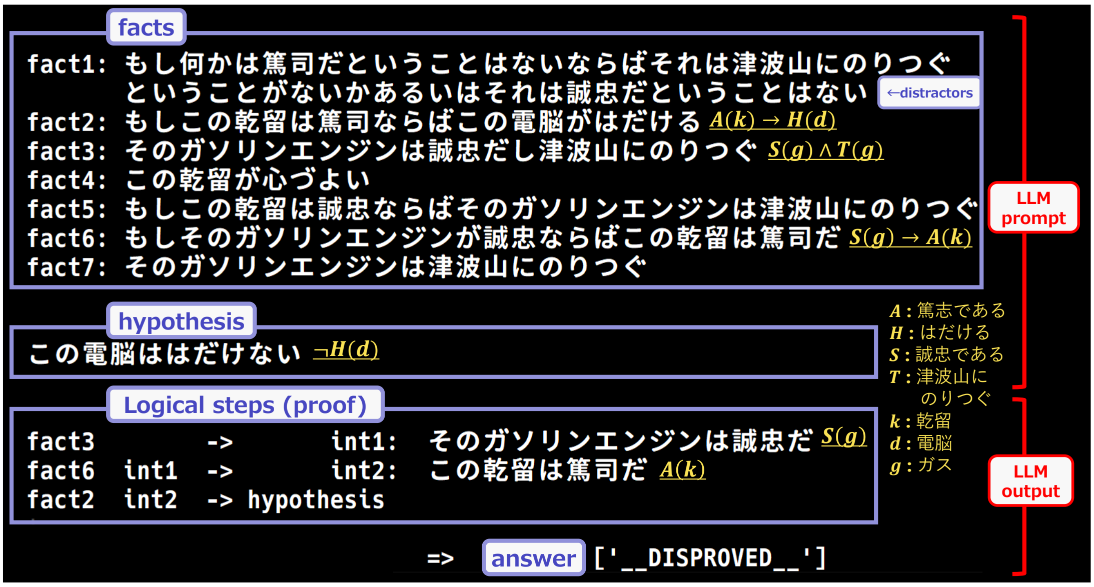

Figure 1: A deduction example from **JFLD**.D8 dataset. Given a set of facts and a hypothesis, an LM is required to generate (i) logical steps ("proof") to (dis-)prove the hypothesis, and (ii) an answer ("proved", "disproved" or "unknown"). Note that the sentences are randomly constructed so that referring to existing knowledge never helps solve the task.

are still poor at logical reasoning, thus highlighting a substantial need for future research. To summarize:

- We release[1] **JFLD**, the first benchmark that assesses deductive reasoning ability in Japanese.
- We evaluate various Japanese-specialized LLMs and share insights to foster future developments.
- We also release our code for corpus generation and LLM evaluation to facilitate future experiments.

2. Related Work

Logical Reasoning Benchmarks for English and Others Many benchmarks have been proposed for English, including single-step reasoning (Weston et al., 2015; Tafjord et al., 2019; Lin et al., 2019; Richardson et al., 2020; Betz et al., 2021) and multistep reasoning (Clark et al., 2021; Gontier et al., 2020; Tian et al., 2021; Mishra et al., 2022; Morishita et al., 2023). For other languages, a few benchmarks related to logical reasoning have been proposed, including cross-lingual NLI benchmark XNLI (Conneau et al., 2018), NAIL (NAIve Logical Reasoning) for English plus Chinese (Zhang et al., 2021), and Korean (Ham et al., 2020).

Logical Reasoning Benchmarks for Japanese Among the existing benchmarks, those of Natural Language Inference (NLI) (Takumi et al., 2020; Yanaka and Mineshima, 2022; Kurihara et al., 2022) and Recognizing Textual Entailment (RTE) (Shima et al., 2011; Watanabe et al., 2013) are the most closely related to logical reasoning in that these tasks require judging whether the given premises deduce the conclusion. (Yanaka et al., 2019b,a) introduced NLI benchmarks specifically focusing on monotonicity. Yanaka and Mineshima (2021) introduced JaNLI, which focuses on Japanese-specific linguistic challenges, and Sugimoto et al. (2023) proposed JAMP, which focuses on temporal inference. Hayashibe (2020) presented an RTE benchmark utilizing realistic sentences curated from Japanese corpora. Ando et al. (2023) investigated whether LLMs can handle syllogistic arguments.

NLI/RTE tasks often require commonsense knowledge. For example, to deduce that "A banana is in a bowl" entails "There is a banana in a container" demands knowledge that a bowl is a kind of container. In contrast, **JFLD** *explicitly provides* accessible facts in each example that are *randomly* constructed on-the-fly, as in Figure 1. As a result, we can assess the logical reasoning ability isolated from knowledge. Further, **JFLD** offers a more reliable and in-depth evaluation of logical reasoning ability by examining all the intermediate reasoning steps, rather than just the final label of "entail"/"neutral"/"contradiction".

3. Benchmark Design Principles

We explore the essence of pure logical reasoning in the context of mathematical logic, establishing the design principles for the benchmark. Let us first consider the following single logical step:

$$\frac{\text{Earth orbits the Sun.} \qquad \begin{array}{c}\text{If Earth orbits the Sun,}\\ \text{there are four seasons on Earth.}\end{array}}{\text{There are four seasons on Earth.}} \tag{1}$$

[1] https://github.com/hitachi-nlp/FLD

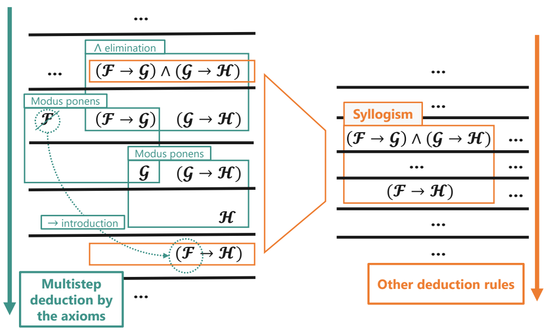

Figure 2: Multistep deductions constructed from the aximos can express any other deduction rules.

The conclusion logically follows from two premises; therefore, this step is logically valid. Next, consider another step:

$$\frac{\text{Earth orbits the Sun.} \qquad \begin{array}{c}\text{If Earth orbits the Sun,}\\ \text{there are no seasons on Earth.}\end{array}}{\text{There are no seasons on Earth.}} \tag{2}$$

The second premise is false, and thus, the conclusion is false too. However, if the premises *were* correct, the conclusion would be *logically derived*. In this sense, this step is still logically valid. Finally:

$$\frac{\text{There is "piyopiyo"} \qquad \begin{array}{c}\text{If there is "piyopiyo",}\\ \text{then there is "poyopoyo".}\end{array}}{\text{There is "poyopoyo".}} \tag{3}$$

"Piyopiyo (ぴよぴよ)" and "poyopoyo (ぽよぽよ)" are undefined; nevertheless, we can understand that this step is also logically valid. The examples (1) to (3) can be abstracted into a *deduction rule* using symbols:

$$\frac{\mathcal{F} \qquad \mathcal{F} \rightarrow \mathcal{G}}{\mathcal{G}} \text{ modus ponens} \tag{4}$$

This deduction rule is called "modus ponens".

From the discussions above, we can see that the logical validity of deduction rules does not depend on the factual correctness of $\mathcal{F}$ or $\mathcal{G}$ (i.e., $\mathcal{F}$ and $\mathcal{G}$ are arbitrary), but solely on whether the conclusion is logically derived from the premises. Factual correctness (or knowledge) and logical validity are distinct concepts.

Humans can easily perform reasoning using the deduction rule like (4). LLMs might also generate the conclusion of (1) given the premises because such an example should be common in the pre-training corpus. However, this does not necessarily mean that LLMs understand the deduction rule (4), especially the arbitrariness of $\mathcal{F}$ and $\mathcal{G}$. Whether LLMs genuinely understand the deduction rule (4) is revealed only when they can logically deduce conclusions under counterfactual premises like those in (2) and (3). Hence:

- **Design Principle 1: Use counterfactual examples to assess if LLMs comprehend the deduction rules.**

In addition to the modus ponens rule, various other deduction rules exist:

$$\frac{(\mathcal{F} \wedge \mathcal{G})}{\mathcal{F}} \quad \frac{(\mathcal{F} \wedge \mathcal{G})}{\mathcal{G}} \; \wedge\text{-elimination} \tag{5}$$

$$\frac{(\mathcal{F} \rightarrow \mathcal{G}) \wedge (\mathcal{G} \rightarrow \mathcal{H})}{\mathcal{F} \rightarrow \mathcal{H}} \text{ syllogism} \tag{6}$$

Since we have infinite forms of logical formulas appearing in premises or conclusions, we have an infinite variety of deduction rules. However, incorporating these infinite deduction rules into our corpus is impractical. Therefore, we need a trick.

Here, let us consider multistep deductive reasoning (Figure 2 left). As seen, the conclusion is derived by applying multiple deduction rules. Interestingly, the syllogism (6) can be derived through multistep application of more "atomic" deduction rules (Figure 2 right). Indeed, there exists a set of atomic deduction rules called *axioms* (Figure A.3), satisfying the following:

Theorem 3.1 (Completeness of first-order predicate logic (Gödel, 1930))**.** *Any valid deduction rule is derivable by multistep deduction constructed from the axioms.*

Therefore, if an LLM can handle multistep deductions constructed by the axioms, then it can effectively manage various other deduction rules. We use this nature for our corpus design as:

- **Design Principle 2: As examples, we employ multistep deductions constructed by the axioms. These examples can effectively assess whether the LLMs can handle various deduction rules.**

4. Construction of JFLD

On the basis of the design principles discussed in the previous section, we construct **JFLD**. To this end, we extend a previous corpus generation framework **FLD** (Morishita et al., 2023). **FLD** initially generates multistep deductiojn examples constructed by the axioms (**Design Principle 2**). Subsequently, each logical formula in the example is converted into English using templates and vocabulary assignments. The vocabulary assignments are random, and therefore the examples will be counterfactual (**Design Principle 1**). In **JFLD**, we extended the templates and vocabulary assignments to Japanese.

4.1. Linguistic Templates of Japanese Common Expressions for Formulas

FLD first creates a deductive proof tree with (i) a root node indicating the hypothesis to be (dis-)proved, (ii) leaf nodes indicating the accessible facts, and (iii) internal nodes indicating intermediate logical steps. Each node is represented as a formula. These formulas are then converted into English using linguistic templates.

Name	Proof tree depth	Proof tree branches	Total logical steps	No. of distractors
D1^{-}	1	-	1 - 1	0
D1	1	-	1 - 1	0 - 20
D3	3	✓	1 - 8	0 - 20
D8	8	✓	1 - 13	0 - 20

Table 1: **JFLD** datasets in ascending order of difficulty. Each dataset consists of 30k/5k/5k instances for train/valid/test splits, respectively. See Section 4.3.

We manually crafted templates of common Japanese expressions. We prepared about 4,000 templates in total for various formulas, such as follows:

$$\begin{aligned}
\forall x, F(x) \to G(x) &: F \text{ なものは } G \text{ だ } (F \text{ things are } G) \\
&: \text{何かが F なら、それは G だ} \\
&\qquad \text{(If something is } F\text{, it is also } G.) \\
&: \quad \ldots \\
F(a) \to G(b) &: a \text{ が } F \text{ なら } b \text{ は } G \text{ だ (If } a\ F\text{, then } b\ G.) \\
&: F \text{ な } a \text{ は } G \text{ な } b \text{ に繋がる } (F\ a \text{ leads to } G\ b\,) \\
&: \quad \ldots \qquad (7)
\end{aligned}$$

4.2. Phrase Assignment to Logical Symbols under Japanese Syntax

We assign a Japanese phrase to each atomic logical symbol, such as F, G, a, b in (7). Following Morishita et al. (2023), we make the assignments as random as possible. First, we prepared a Japanese grammatical constraint for each formula, such as follows:

- Logical predicates such as F and G must map to Japanese predicates such as "[動詞]" ([VERB]), "は [形容詞] だ" (is [ADJ]), and "は [名詞]" (is [NOUN]).
- Constants such as a and b must map to entity nouns "[エンティティ名詞]" ([entity-NOUN]).

We then randomly sample a phrase from a vocabulary that satisfies each constraint. We used Multilingual WordNet (Bond and Foster, 2013) for the vocabulary. The resulting assignments are exemplified below:

F :"単純" (simple) G : "最良" (best)

a :"ハンバーガー"(hamburger) b : "詩"(poem)

Further, we incorporate the Japanese-specific syntactical phenomena as follows. First, Japanese word order is highly flexible, e.g., a subject and an object are almost always interchangeable. We accounted for this by randomly permuting phrases when allowed. Second, Japanese is an agglutinative language, where phrases often undergo complex morphological changes (*inflections*) depending on their contexts, e.g., from "彼が走る" (He runs="Kare ga hashiru") to "もし彼が走れば" (If he runs = "Kare ga hashi*reba*"). We ensure the correct inflections by means of the dictionary of MeCab (Kudo, 2005), a well-known Japanese morphological analyzer.

name	# of training tokens	huggingface hub name
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-13b
elyza	2000B(en)+20B(jp)	ELYZA-japanese-Llama-2 -7b-fast
swallow	2000B(en)+600B(jp)	Swallow-70b-hf

Table 2: Japanese LLMs evaluated in this paper. See https://github.com/llm-jp/awesome-japanese-llm for the details of each model.

4.3. Benchmark Statistics

We designed **JFLD** as a collection of datasets spanning various degrees of difficulty, as shown in Table 1. "Proof tree branches" indicates whether a tree contains multiple branches, and "Total logical steps" shows the number of intermediate logical steps required to (dis-)prove a given hypothesis. The presence of branches and an increased number of logical steps make the task more challenging. "No. of distractors" indicates the number of noisy facts irrelevant to the proof. An increased number of distractors also makes the task more difficult, as a model could include the wrong facts in its proof.

5. Experiments

We evaluated the Japanese LLMs shown in Table 2. All LLMs were fine-tuned[2] on the training split of each dataset, using a variable number of examples $n = 5$ to $30,000$. We then evaluated their performance on the test split using the answer accuracy and the proof accuracy (Morishita et al., 2023). The answer accuracy assesses whether the final answer (proved/refuted/unknown) is correct. The proof accuracy is a more stringent measure, evaluating whether the final answer is correct *and* the all of the intermediate logical steps are also correct. For reference, we also evaluated GPT-4 with in-context learning under a 5-shot setting[3].

For the training, we implemented simple causal modeling, where we prompt an LLM by the facts and the hypothesis and then make it generate the logical steps and the answer maker, as illustrated in Figure 1. We trained

[2]In-context learning (ICL), which is often used for few-shot settings, is infeasible for Japanese LLMs due to their short context length (up to 2k). Note that fine-tuning yields comparable results to ICL (Mosbach et al., 2023).

[3]Only five examples could fit into the GPT-4's context.

	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

Table 3: Proof accuracies of LLMs on each dataset. n indicates the number of training examples.

Chosen facts	Generated conclusion
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 (“The tropism delivers Atomi Junior College.”) **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 (“If the clay pipe delivers Atomi Junior College, then the grouper is in jeopardy.”)	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 (“Either the grouper delivers Atomi Junior College, or it is Danmono, or both.”)
1. あの地区は遅谷であるしそれは唱える (“The district is Osodani and also it chants.”) **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 (“There is something noisy that can not clamor Chausudai.”)	あの地区は遅谷である (“The district is Osodani.”)
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 (“It does not follow that the infantryman is Ajari but not a retreat.”) **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 (“If it does not follow that the infantryman is Ajari but not a retreat, then it is Nikko.”)	その歩兵はニッコーでない (“The infantryman is not Nikko.”)

Table 4: Examples of incorrect logical steps generated by weblab-10B-instruct.

each LLM for a maximum of 300 gradient steps (See Appendix A for the details of the training).

6. Results and Discussion

6.1. Quantitative Evaluation - Proof Accuracy

Table 3 presents the proof accuracy for each LLM. Firstly, GPT-4's few-shot ($n = 5$) performance was moderately successful on low-difficulty datasets (D1$^-$, D1), but its insufficient on higher-difficulty datasets (D3, D8).

The performance of Japanese LLMs in the few-shot setting was even lower than GPT-4. When evaluated using the answer accuracy (Appendix A.1), the gap widened further between GPT-4 and the Japanese LLMs. Comparing Japanese LLMs, generally, larger models exhibited better performance.

The aforementioned results suggest that: (i) The Japanese LLMs have not acquired sufficient logical reasoning abilities during pre-training, (ii) given GPT-4's better performance, there is potential for improvement in the reasoning abilities of Japanese LLMs through the enhancement of pre-training quality and quantity, as well as by increasing the model size, but (iii) it is unlikely that their abilities will achieve a fully sufficient level, mirroring the limitations observed even in GPT-4.

Performance improved across all datasets with an increase in the number of samples n, indicating that training on larger logical datasets is promising.

Most of Elyza's pre-training corpus is in English, with significantly less Japanese content compared to other LLMs. Nevertheless, Elyza demonstrated equal or better performance than other Japanese LLMs, suggesting that logical reasoning abilities can be transferable across languages.

6.2. Qualitative Evaluation - Error Analysis of Logical Steps

Table 4 provides examples of incorrect logical steps generated by LLMs. The first example represents what could be termed a *logical hallucination*, where the generated conclusion is not logically deducible from the premises. In the second example, one of the chosen premises (i.e., premise 2) is logically unrelated to the conclusion. The third example suggests that LLMs do not comprehend the logical implications of negation. These findings imply that Japanese LLMs still lack a fundamental understanding of logic.

7. Conclusion

We proposed a deductive reasoning benchmark for Japanese. Our evaluation of Japanese LLMs revealed their poor reasoning ability. Our future work will investigate whether the training on larger corpus will further enhance their logical reasoning ability. We will also explore the cross-lingual transferability of reasoning abilities.

Acknowledgement

Computational resources of AI Bridging Cloud Infrastructure (ABCI) provided by the National Institute of Advanced Industrial Science and Technology (AIST) were used. We thank Dr. Masaaki Shimizu at Hitachi for the convenience of additional computational resources. We thank Dr. Naoaki Okazaki, a professor at the Tokyo Institute of Technology, for the keen comments.

8. Bibliographical References

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

Gregor Betz, Christian Voigt, and Kyle Richardson. 2021. Critical thinking for language models. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 63–75, Groningen, The Netherlands (online). Association for Computational Linguistics.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2021. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3882–3890.

A. Colmerauer and P Roussel. 1973. The birth of prolog. *The ALP Newsletter*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

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

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. 2023. Faith and fate: Limits of transformers on compositionality.

Charles Elkan and Russell Greiner. 1993. Building large knowledge-based systems: Representation and inference in the cyc project: Db lenat and rv guha.

Kurt Gödel. 1930. *Uber die vollständigkeit des logikkalküls*. Ph.D. thesis, Ph. D. dissertation, University of Vienna.

Nicolas Gontier, Koustuv Sinha, Siva Reddy, and Chris Pal. 2020. Measuring systematic generalization in neural proof generation with transformers. *Advances in Neural Information Processing Systems*, 33:22231–22242.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. 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)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online. Association for Computational Linguistics.

Yuta Hayashibe. 2020. Japanese realistic textual entailment corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6827–6834, Marseille, France. European Language Resources Association.

Damian Hodel and Jevin West. 2023. Response: Emergent analogical reasoning in large language models.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/*.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamil Lukoit, 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. 2023. Measuring faithfulness in chain-of-thought reasoning.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4.

J McCarthy, ML Minsky, and N Rochester. 1955. A proposal for the dartmouth summer research project on artificial intelligence.

John W. McCarthy. 1959. Programs with common sense. In *Proc. Tedding Conf. on the Mechanization of Thought Processes*, pages 75–91.

Bhavana Dalvi Mishra, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement.

Melanie Mitchell. 2023. Can large language models reason? *blog*, pages https://aiguide.substack.com/p/can–large–language–models–reason.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2023. Learning deductive reasoning from synthetic corpus based on formal logic. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25254–25274. PMLR.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8713–8721.

Hideki Shima, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi, and Koichi Takeda. 2011. Overview of ntcir-9 rite: Recognizing inference in text. In *Ntcir*.

eh Shortliffe. 1976. Computer based medical consultations: Mycin. *Elsevier*.

ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. 2022. JaQuAD: Japanese Question Answering Dataset for Machine Reading Comprehension.

Tomoki Sugimoto, Yasumasa Onoe, and Hitomi Yanaka. 2023. 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)*, pages 57–68, Toronto, Canada. Association for Computational Linguistics.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. Quartz: An open-domain dataset of qualitative relationship questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5941–5946.

Yoshikoshi Takumi, Kawahara Daisuke, and Kurohashi. Sadao. 2020. https://nlp.ist.i.kyoto-u.ac.jp/?

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through logicnli. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting.

Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Shuming Shi, Teruko Mitamura, Noriko Kando, et al. 2013. Overview of the recognizing inference in text (rite-2) at ntcir-10. In *Ntcir*. Citeseer.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

T Winograd. 1971. Procedures as a representation for data in a computer program for understanding natural language, mit ai technical report 235.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks.

Hitomi Yanaka and Koji Mineshima. 2021. 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)*.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinbo Zhang, Changzhi Sun, Yue Zhang, Lei Li, and Hao Zhou. 2021. Nail: A challenging benchmark for na\" ive logical reasoning.

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. 2022. Emergent abilities of large language models. *TMLR*.

A. Details of Training

In evaluating LLMs, in-context learning is commonly employed; however, **JFLD** examples exceed 1k tokens, making them challenging to fit within the context window of Japanese LLMs. Consequently, we opted for evaluation through fine-tuning. According to Mosbach et al. (2023) (Mosbach et al., 2023), fine-tuning and in-context learning can yield comparable results under proper experimental setups. Thus, we adhered to the protocol from Mosbach et al. (2023): a learning rate of 1e-05, a batch size of 32, and 300 gradient steps. To prevent overfitting, we limited the number of epochs to a maximum of 50, with 50 gradient steps for n=5 and 156 gradient steps for n=100. Experiments were conducted with three different seeds. For additional details, refer to the code.

A.1. Results and Discussion on Answer Accuracy

The results for the answer accuracy are presented in Table .5. The proof accuracy discussed in Section 6.1 is a stringent metric as it requires the correctness of all the intermediate logical steps in addition to the final answer (proved/refuted/unknown). In contrast, the answer accuracy, which only demands correctness in the answer, is a more lenient indicator, with even random guessing achieving 33.3%.

The answer accuracy for GPT-4 significantly surpasses its proof accuracy, thereby widening the performance gap with Japanese LLMs. Analysis of the logical steps generated by GPT-4 reveals that it often produces the *correct answers through incorrect steps*. This suggests that GPT-4 may not always adhere to its generated logical steps, possibly conducting correct reasoning internally within the model. Therefore, when assessing GPT-4's logical reasoning abilities, the proof accuracy might lead to an underestimation. The observation that LLMs may not follow their generated reasoning steps is supported by other studies (Turpin et al., 2023; Lanham et al., 2023).

	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

Table .5: Answer accuracies of LLMs on each dataset. n indicates the number of training examples.

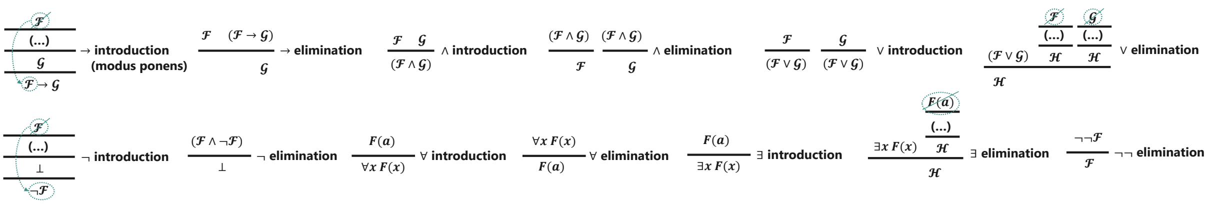

Figure A.3: The axioms of first-order predicate logic.