Exploring Reasoning Biases in Large Language Models Through Syllogism: Insights from the NeuBAROCO Dataset

Kentaro Ozeki[1,2], **Risako Ando**[1], **Takanobu Morishita**[1], **Hirohiko Abe**[1],
Koji Mineshima[1], **Mitsuhiro Okada**[1]
[1]Keio University, Tokyo, Japan
[2]University of Tokyo, Tokyo, Japan
`kentaro.ozeki@gmail.com` `{risakochaan,morishita,hirohiko-abe}@keio.jp`
`{minesima,okada}@abelard.flet.keio.ac.jp`

Abstract

This paper explores the question of how accurately current large language models can perform logical reasoning in natural language, with an emphasis on whether these models exhibit reasoning biases similar to humans. Specifically, our study focuses on syllogistic reasoning, a form of deductive reasoning extensively studied in cognitive science as a natural form of human reasoning. We present a syllogism dataset called NeuBAROCO, which consists of syllogistic reasoning problems in English and Japanese. This dataset was originally designed for psychological experiments to assess human reasoning capabilities using various forms of syllogisms. Our experiments with leading large language models indicate that these models exhibit reasoning biases similar to humans, along with other error tendencies. Notably, there is significant room for improvement in reasoning problems where the relationship between premises and hypotheses is neither entailment nor contradiction. We also present experimental results and in-depth analysis using a new Chain-of-Thought prompting method, which asks LLMs to translate syllogisms into abstract logical expressions and then explain their reasoning process. Our analysis using this method suggests that the primary limitations of LLMs lie in the reasoning process itself rather than the interpretation of syllogisms.

1 Introduction

The ability to perform accurate deductive reasoning in natural language, once pursued by classical symbolic AI, has recently become a vital issue in the development and evaluation of Large Language Models (LLMs) (Qiao et al., 2022; Huang and Chang, 2022). Regarding humans, empirical research in cognitive science has demonstrated that humans often exhibit various errors and tendencies in reasoning, known as *reasoning biases* (Evans, 1989; Pohl, 2022). Among various forms of reasoning, *syllogistic reasoning* is one of the basic forms of deductive reasoning and has been studied extensively (Johnson-Laird and Steedman, 1978; Evans et al., 1993; Geurts, 2003). However, the evaluation of LLMs and the construction of datasets incorporating these experimental methodologies has not been systematically pursued.

In this paper, we present the evaluation of LLMs with NeuBAROCO, a manually constructed syllogism dataset with diverse properties and annotations designed to evaluate the reasoning abilities and biases of LLMs in syllogistic reasoning tasks. NeuBAROCO is a bilingual (Japanese and English) dataset and includes detailed annotations for the types of reasoning biases associated with each problem. The dataset is based on a problem set used for a series of psychological experiments assessing human reasoning ability with syllogisms (Shikishima et al., 2009, 2015). A subset of the problems in the dataset is aligned with human performance metrics from these experiments. Building on the work of Ando et al. (2023), we have adapted this problem set to evaluate whether language models exhibit reasoning biases similar to those of humans.

The main contributions in this paper are summarized as follows.[1] First, we constructed a dataset containing 95 and 790 syllogistic reasoning problems in the format of Multiple-Choice and Natural Language Inference (NLI) tasks, respectively. This dataset design facilitates comparison with results and insights from human psychological experiments and preliminary studies on machine learning models.

Second, we systematically investigated various reasoning biases observed in LLMs in relation to the form and content of syllogisms, including *belief biases*, *conversion errors*, and *atmosphere effects*, along with other tendencies, across multiple tasks and in both English and Japanese.

Finally, to more precisely identify the reasoning

---

[1] The data and results are available at `https://github.com/kmineshima/NeuBAROCO`.

Findings of the Association for Computational Linguistics ACL 2024, pages 16063–16077
August 11-16, 2024 ©2024 Association for Computational Linguistics

Type	**Sentence Pattern**	**Predicate Logic**	**Set Theory**	**Description**
A (*all*)	All S are P	$\forall x(S(x) \to P(x))$	$S \subseteq P$	Universal Affirmative
E (*no*)	No S are P	$\forall x(S(x) \to \neg P(x))$	$S \cap P = \emptyset$	Universal Negative
I (*some*)	Some S are P	$\exists x(S(x) \wedge P(x))$	$S \cap P \neq \emptyset$	Particular Affirmative
O (*some-not*)	Some S are not P	$\exists x(S(x) \wedge \neg P(x))$	$S \setminus P \neq \emptyset$	Particular Negative

Table 1: Four types of categorical sentences and their translation into logical notations.

errors in LLMs, we introduce a new task called *Translate-and-Explain*. This task requires LLMs to first translate the syllogism into a logical expression (Translate) and then explain the reasoning process (Explain), by leveraging multi-step Chain-of-Thought prompting. Our findings indicate two distinct error patterns: one in the interpretation of the syllogism demonstrated in translation, and the other in the explained reasoning process. However, the primary limitations of these models lie in the reasoning process itself, rather than in the interpretation of syllogisms.

We first provide an overview of syllogistic reasoning in Section 2, followed by a detailed presentation of the dataset in Section 3. Subsequent sections present the evaluation tasks (Section 4) and the experimental setup along with an analysis of the results (Section 5). Additionally, Section 6 reviews related work in the field.

2 Background: Syllogistic reasoning

In this study, we primarily focus on the logical inference task that classifies inferences into three labels: *entailment*, *contradiction*, and *neutral* (neither entailment nor contradiction).

A syllogism is an inference that consists of two premises and one conclusion, where the premises and the conclusion are composed of four basic types of quantified sentences: *all*, *no*, *some*, and *some-not*, as shown in Table 1. For example, (1) and (2) are syllogisms composed of sentences with the quantifiers *all* and *no*.

$$(1)\quad \frac{\textbf{P1}\text{: All B are C} \qquad \textbf{P2}\text{: All A are B}}{\textbf{C}\text{: All A are C}} \qquad\qquad (2)\quad \frac{\textbf{P1}\text{: No A are B} \qquad \textbf{P2}\text{: All B are C}}{\textbf{C}\text{: No A are C}}$$

The syllogism in (1) is an instance of *entailment*, where if the premises (**P1** and **P2**) are true, then the conclusion (**C**) is also true. The syllogism in (2) is an instance of *neutral*, where the relationship between the premises (**P1** and **P2**) and the conclusion (**C**) is neither entailment nor contradiction.

Syllogisms are relatively simple logical inferences that can be represented in monadic predicate logic (Łukasiewicz, 1951), a fragment of first-order logic where each predicate can take only one argument. Table 1 shows how to translate each type of categorical sentence into logical notations, that is, into predicate logic and set theory.

Despite being logically simple, syllogisms are known to be challenging inferences for humans. Which types of syllogisms are prone to causing errors in human reasoning, or in other words, involve reasoning biases, is a topic widely studied in the field of cognitive science of human reasoning (Evans et al., 1993; Manktelow, 1999; Geurts, 2003; Stenning and van Lambalgen, 2012; Khemlani and Johnson-Laird, 2012). Our choice to focus on syllogistic reasoning is deliberate and aims to facilitate comparisons with insights from the extensive research on biases and reasoning in cognitive science. Some typical biases of syllogism will be introduced in Section 3.2. Focusing on the possibility of a detailed classification of such human reasoning biases, this study uses syllogistic reasoning in natural language as a testbed to evaluate the logical reasoning capabilities of LLMs.

3 The NeuBAROCO dataset

The NeuBAROCO dataset is based on a syllogism problem set called the BAROCO test (Shikishima et al., 2005, 2009), originally designed for large-scale research on human reasoning abilities. BAROCO includes the so-called belief-bias tasks, which are typical examples involving human reasoning biases (see Section 3.2.1). In addition to linguistic tasks, it also includes Euler diagram tasks to test spatial cognition. These formats of reasoning were used to investigate the correlation and the contributions of genetic and environmental factors through twin studies (Shikishima et al., 2005). Furthermore, studies combining these tasks with experimental tasks in behavioral economics have also been conducted (Shikishima et al., 2015).

Ando et al. (2023) provides a preliminary study preceding this research, aiming to apply the BAROCO problem set for evaluating LLMs by re-

Type	English	Japanese
Symbolic (98)	**P1**: All A are B. **P2**: All B are C. **C**: All A are C. (ENTAILMENT)	**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。 (ENTAILMENT)
Congruent (404)	**P1**: All humans are mammals. **P2**: No mammal is a crocodile. **C**: No crocodile is human. (ENTAILMENT)	**P1**: すべての人間は哺乳類である。 **P2**: どの哺乳類もワニでない。 **C**: どのワニも人間でない。 (ENTAILMENT)
Incongruent (238)	**P1**: Some animals are human beings. **P2**: All animals are tomatoes. **C**: Some humans are tomatoes. (ENTAILMENT)	**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。(ENTAILMENT)

Table 2: Examples of syllogisms in English and Japanese labeled as *Symbolic*, *Congruent*, and *Incongruent*. The numbers in parentheses indicate the number of cases for each type.

Type	English	Japanese
Conversion (66)	**P1**: All B are A. **P2**: All B are C. **C**: All A are C. (NEUTRAL)	**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。(NEUTRAL)
Atmosphere (345)	**P1**: Some canines are pets. **P2**: Some pets are smart. **C**: Some canines are smart.(NEUTRAL)	**P1**: あるイヌはペットである。 **P2**: あるペットは賢い。 **C**: あるイヌは賢い。 (NEUTRAL)

Table 3: Examples of syllogisms labeled as *Conversion* and *Atmosphere*. The numbers in parentheses indicate the number of cases for each type.

casting a portion of the BAROCO problem set to introduce the NeuBAROCO dataset, consisting of 375 syllogistic reasoning problems. In this study, we have expanded and utilized the NeuBAROCO dataset to evaluate the logical reasoning abilities of LLMs more systematically. This expansion includes the addition of new problems, incorporating the Multiple-Choice problems (95 problems) that were originally employed in the psychological experiments of the BAROCO problem set. Additionally, the number of NLI task problems has been expanded from 375 to 790.

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 the rings in this box are gold rings.
4. Some ring inside this box is not a gold ring.
5. None of them.

Correct answer: 3

Figure 1: An example of the Multiple-Choice task

3.1 Overview of the dataset

The original BAROCO problem set consists of two premises and multiple choices that could serve as conclusions. Experiment participants are asked to select a logically valid conclusion from the given choices. The NeuBAROCO dataset was constructed by converting each problem from the BAROCO problem set into a format commonly used for the NLI task. The NeuBAROCO dataset we use in this paper includes 790 problems for the NLI task, namely tasks that classify inferences into entailment, contradiction, and neutral. Of these, 254 problems are classified as *entailment*, 188 as *contradiction*, and 348 as *neutral*. While the original BAROCO problem set is written entirely in Japanese, the NeuBAROCO dataset translates these problems into English, making it available as a Japanese-English parallel reasoning corpus.

The NeuBAROCO dataset also includes 80 problems for the Multiple-Choice task, along with 15 additional examples and practice problems, following the format of the original BAROCO problem set. For each problem, the number of the correct answer is labeled. Figure 1 shows an example of a problem in the Multiple-Choice task.

3.2 Annotation

We focus on three types of biases in syllogistic reasoning. The three types of biases addressed here can be categorized into two kinds: biases related to *content* (belief bias) and biases related to *form* (conversion errors and atmosphere effects). By investigating these three types of biases, we can systematically evaluate whether LLMs are sensitive to the roles of content words and function words in deductive reasoning.

3.2.1 Labels for bias related to content

To investigate biases caused by content words such as nouns and verbs, we categorized each inference into three types based on whether it is congruent with commonsense beliefs. Table 2 shows examples and the number of instances for each type.

Symbolic When all terms are composed of sentences from abstract symbols, the inference is labeled as *symbolic*. These types of problems are neutral with respect to the beliefs held by humans; that is, the question of whether they agree or disagree with those beliefs does not arise.

Congruent If there is no inconsistency with commonsense beliefs in all premises and conclusions, the inference is labeled as *congruent*.

Incongruent If at least one of the premises or the conclusion does not align with commonsense beliefs, the inference is labeled as *incongruent*. In the example in the bottom row of Table 2, *All animals are tomatoes* and *Some humans are tomatoes* contradict commonsense beliefs.

If it is unclear whether the sentence is consistent with commonsense beliefs, or if it requires specialized knowledge (e.g., *All agnostics are Stoics. Some agnostics are skeptical. Therefore, all Stoics are skeptical*), the inference is classified as *others*. There are 50 instances of this type.

3.2.2 Labels for bias related to form

In addition, we assigned the tags *conversion* and *atmosphere* to types of inferences that are prone to the two major types of reasoning biases. These biases are induced by function words such as *all* and *not*, as well as by grammatical factors such as word order in the premises and conclusions.

Conversion Conversion error is known as a typical reasoning bias in syllogisms (Evans et al., 1993; Geurts, 2003). This error occurs when quantified sentences are misinterpreted by converting the order of two terms: *All A are B* and *Some A are not B* are misinterpreted as *All B are A* and *Some B are not A*, respectively. For instance, interpreting the sentence *All students who score above 90 points receive an A grade* as equivalent to *All students with an A grade score above 90 points* exemplifies this error. Although these two sentences may appear similar, they do differ in logical meaning. Table 3 presents examples of syllogisms where such illicit conversion results in inference being erroneously classified as valid (*entailment*) rather than invalid (*neutral*). We assign the *conversion* label to those inferences where a sentence containing *all* or *some-not* appears in the premises, and the label changes from *neutral* to *entailment* when the order of terms in the sentence is reversed.

Atmosphere The atmosphere effect indicates the tendency to select conclusions that mirror the form of the premises (Woodworth and Sells, 1935). This involves selecting conclusions that superficially resemble the premises in terms of their logical structure (Chater and Oaksford, 1999). For example, a conclusion containing *some* might be preferentially selected if a premise also contains *some*. Similarly, if a premise containing a negation (*no* or *some-not*) tends to promote a negative conclusion. We assign the *atmosphere* label to those inferences with *neutral* labels where either (1) a premise contains *some* and the conclusion is particular (*some* or *some-not*), or (2) a premise contains *some-not* and the conclusion is particular or negative (i.e., *no*, *some*, or *some-not*). Table 3 shows an example that satisfies condition (1).

4 Evaluation Tasks

We introduce three types of tasks for evaluating LLMs using the NeuBAROCO dataset: Multiple-Choice, NLI, and Translate-and-Explain.

Multiple-Choice The Multiple-Choice task is a format widely used in cognitive psychology. In this task, models are presented with two premises and asked to choose the correct conclusion from five options. These options include the *all*, *some*, *no*, and *some-not* sentences, as well as the "none of them" choice. Table 4 presents examples of the prompts used in this task.

NLI The NLI task is a common problem setting in NLP, enabling evaluations on specific instances of reasoning and aligning well with other NLP

Input

```
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:
```

Input

```
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:
```

Table 4: Example prompts for the Multiple-Choice task in English and Japanese.

Input

```
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".

## Input
Premise 1: Some A are B.
Premise 2: All B are C.
Hypothesis: All A are C.

## Translation into predicate logic
Premise 1: ∃x(Ax∧Bx)
Premise 2: ∀x(Bx→Cx)
Hypothesis: ∃x(Ax→Cx)

## Reasoning
[Explain your reasoning for the answer]

## Answer
[Your answer must be one word: "entailment", "contradiction", or "neither"]

## Input
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.

## Translation into predicate logic
```

Table 5: Example prompt used for the Translate-and-Explain task (Translation into predicate logic).

benchmarks. In this task, two premises of a syllogism and one hypothesis are presented. Models are then asked to determine whether the relationship of the hypothesis to the premises is one of *entailment*, *contradiction*, or *neither*. Experiments are conducted in zero-shot and few-shot settings, respectively (Brown et al., 2020). In the zero-shot setting, instructions and problems are provided without examples. Example prompts are shown in Table 10 of the Appendix. In the few-shot (3-shot) setting, three exemplar problems with correct answers are included in the prompt (Table 12 in Appendix). We use abstract symbolic problems as exemplars to avoid potential biases, ensuring that the few-shot examples remain neutral to belief congruence and incongruence.

Translate-and-Explain To provide a finer-grained analysis of the reasoning ability of the models, we design the Translate-and-Explain task, a variant of the NLI task with a dedicated Chain-of-Thought (CoT) prompt. In this task, we emulate the translation between natural language sentences and formal expressions of reasoning before the actual reasoning step, identifying whether errors and biases stem from the process of interpreting sentences or from the process of reasoning.

CoT prompting is a technique of having LLMs perform intermediate reasoning steps by using few-shot examples or other means, and has been reported to improve the reasoning ability of LLMs (Wei et al., 2022). This enables us to apply a method analogous to the protocol analysis in psychological experiments (Evans et al., 1983) to the evaluation of LLMs.

The Translate-and-Explain CoT prompt is a 1-shot structural prompt that instructs LLMs to perform (i) a translation step, (ii) an explanation step, and (iii) an answer step for each problem. For the translation step, LLMs are instructed to translate the given syllogism into abstract expressions. In this study, we compare translations into *predicate logic* in formal language and into *set-theory* in natural language (see Table 1 in Section 2). We conducted experiments in three setups: (a) explanation without translation, (b) predicate logic translation + explanation, (c) set-theoretic translation + explanation. We manually checked the correctness of the translation outputs of LLMs.

Language	Model	Overall	Symbolic	Congruent	Incongruent
English	GPT-3.5	53.75	45.00	72.50	25.00
	GPT-4	83.75	75.00	90.00	80.00
	Llama-2-13B	26.25	20.00	40.00	5.00
	Swallow-13B	25.00	20.00	30.00	20.00
	Llama-2-70B	56.25	65.00	62.50	35.00
	Swallow-70B	60.00	45.00	80.00	35.00
Japanese	GPT-3.5	42.50	55.00	40.00	35.00
	GPT-4	95.00	85.00	97.50	100.00
	Llama-2-13B	21.25	20.00	22.50	20.00
	Swallow-13B	30.00	20.00	37.50	25.00
	Llama-2-70B	66.25	60.00	62.50	80.00
	Swallow-70B	50.00	45.00	52.50	50.00
	Human	53.00	49.10	51.20	59.70

Table 6: Accuracy (%) on the Multiple-Choice task (80 problems).

Table 5 shows an example of the prompt in English. Full examples are listed in Appendix A.3. Note that we do not provide specific examples of reasoning in the prompt to avoid leading LLMs to adopt similar methods. This allows us to analyze LLMs' "free" reasoning without conditioning.

5 Experiments

5.1 Experimental Setup

We conducted experiments on the three types of tasks both in English and in Japanese. In the Multiple-Choice task, we evaluated the following models:

- **GPT-3.5** (Ouyang et al., 2022) and **GPT-4** (OpenAI, 2023). The GPT models used were gpt-3.5-turbo-1106 and gpt-4-0613 available via OpenAI's API. The number of parameters for the GPT models has not been disclosed.
- **Llama-2** (Touvron et al., 2023) with 13 billion (13B) and 70 billion (70B) parameters. For Llama-2, detailed model information, including the number of parameters, is publicly available.
- **Swallow** (Fujii et al., 2024) with 13B and 70B parameters. Swallow is a model family based on Llama-2 with state-of-the-art Japanese language capability, enhanced through continual pre-training using a dedicated corpus.

In the NLI and Translate-and-Explain tasks, the Llama-2 and Swallow models failed to produce output with a valid answer in most cases. Therefore, we focus on the evaluation results of the GPT models, which consistently produced valid outputs.

The default values were used for the hyperparameters, except for the maximum output token length. This was set to 10 for the Multiple-Choice and NLI tasks, and 2,048 for the Translate-and-Explain task, which are sufficiently long given the design of the tasks.

5.2 Results and Analysis

5.2.1 Multiple-Choice Task

Table 6 shows the experimental results for the Multiple-Choice task. The row labeled **Human** presents the average scores of 440 participants based on data from the psychological experiment conducted by Shikishima et al. (2009). Note that there are some terminological differences between Shikishima et al. (2009) and our study.

In terms of the biases related to content, *incongruent* cases are generally harder for the LLMs than *congruent* cases in English. In contrast, no similar trend is clearly observed in Japanese problems.

In terms of model scale, while the overall accuracy of the smaller (13B-parameter) models ranged from 20% (chance level) to 30%, the 70B-parameter models and GPTs achieved an overall accuracy of 42% to 95%, with some surpassing human overall accuracy in Japanese (53%). The strong performance of large-parameter LLMs can be partly attributed to the nature of the Multiple-Choice task. As we will see in Section 5.2.2, LLMs (especially those with larger parameter) mark high accuracy particularly in identifying *entailment* and *contradiction* over *neutral* cases. In this task, the correct choice is always an *entailment* of the given premises, unless the answer is "none of them." Following the original BAROCO problem set, the NeuBAROCO dataset does not include any problems where the correct answer is "none of them."

Language	Model	Overall	E	C	N	Symbolic	Cong	Incong	Conv	Atmos
English	GPT-3.5	49.75	84.25	42.02	28.74	55.10	56.19	36.55	18.18	32.67
	(Few-Shot)	47.09	88.58	27.66	27.30	47.96	50.50	39.92	13.64	29.70
	GPT-4	71.77	85.04	93.62	50.29	76.53	76.24	61.76	40.91	50.99
	(Few-Shot)	77.47	90.55	88.83	61.78	79.59	82.92	67.23	50.00	59.41
Japanese	GPT-3.5	40.00	82.28	51.06	3.16	35.71	48.02	28.57	4.55	1.98
	(Few-Shot)	40.00	90.94	35.64	5.17	35.71	46.53	31.09	4.55	5.45
	GPT-4	70.38	87.40	95.74	44.25	71.43	76.73	60.92	43.94	38.12
	(Few-Shot)	78.61	92.13	88.30	63.51	82.65	81.44	73.95	74.24	57.43

Table 7: Accuracy (%) on the NLI task (790 problems). **E** = *entailment*, **C** = *contradiction*, **N** = *neutral*, **Cong** = *Congruent*, **Incong** = *Incongruent*, **Conv** = *Conversion*, **Atmos** = *Atmosphere*.

While Swallow is a Japanese-enhanced model based on Llama-2, the comparative results for the two are not straightforward. For a detailed discussion, see Appendix B.

5.2.2 NLI Task

Table 7 shows the results of the NLI task. The few-shot setting improves the overall accuracy of GPT-4 both in English and Japanese.

In terms of gold labels, The models achieve higher scores on problems labeled as *entailment*, while those labeled as *neutral* are typically the most challenging. Even GPT-4, which performed the best, scored approximately 30 points lower on *neutral* problems compared to other labels. As noted above, when comparing the Multiple-Choice and NLI tasks, the Multiple-Choice task includes only problems that correspond to the *entailment* problems in the NLI task. Consequently, the scores for the Multiple-Choice task are similar to those for the *entailment* problems in the NLI task.

In terms of bias-related labels, the results suggest that LLMs are influenced by the content and form of syllogisms. Among the *symbolic*, *congruent*, and *incongruent* cases, the accuracies in the incongruent cases are generally lower than those in the other cases and the overall score for each model. Regarding conversion errors, the accuracies for the problems labeled *conversion*, which may cause conversion errors, are significantly lower than the overall accuracies. A similar trend is observed for the problems labeled *atmosphere* in the context of atmosphere effects.

5.2.3 Translate-and-Explain Task

Table 8 shows the experimental results for the Translate-and-Explain task. Table 9 provides an example output from GPT-4.

English With GPT-4, the translations from English to predicate logic are highly accurate, correctly translating 87/90 problems. However, translating the same problems into set theory is more challenging, with 82/90 problems correctly translated. It is observed that GPT-4 often responds with *contradiction* when the correct answer to an inference is *neither*. As a typical error, GPT-4 interpreted *All animals are tomatoes* as *The set of animals is identical to the set of tomatoes* rather than *The set of animals is a subset of the set of tomatoes*.

With GPT-3.5, the translations into predicate logic are almost accurate, correctly translating 82/90 problems. As a typical error, GPT-3.5 interprets *A certain police officer is not a public servant* as $\neg\exists x(Px \wedge Sx)$ instead of $\exists x(Px \wedge \neg Sx)$, failing to capture the correct scope of negation. The number of problems translated correctly to set theory is 79/90. GPT-3.5 mistakenly interprets *A certain police officer is not a public servant* as *The set of police officers does not intersect with the set of public servants*, which is logically equivalent to the wrong predicate logic formula, $\neg\exists x(Px \wedge Sx)$.

Japanese With GPT-4, 86/90 problems are correctly translated from Japanese to predicate logic. Translations into set theory are more error-prone than those into predicate logic, with only 69/90 problems correctly translated. GPT-3.5 has trouble in translation into predicate logic, with 64/90 problems correctly translated. Similar to the mistakes often made by beginners in logic, the model incorrectly interprets a Japanese sentence corresponding to *Some A are B* as $\exists x(A(x) \rightarrow B(x))$, while the correct interpretation is $\exists x(A(x) \wedge B(x))$.

With GPT-3.5, 28/90 problems are correctly translated from Japanese to set theory. The translation of A-type (*all*-type) sentences in Japanese is generally accurate, but the translations of the

Language	Model	Explanation Only	Pred + Explanation		Set + Explanation	
		Answer	Translation	Answer	Translation	Answer
English	GPT-3.5	58.89	91.11	52.22	87.78	54.44
	GPT-4	77.78	96.67	83.33	92.22	80.00
Japanese	GPT-3.5	57.78	71.11	57.78	33.33	60.00
	GPT-4	70.00	95.56	75.56	76.66	74.44

Table 8: Accuracy (%) on the Translate-and-Explain task (90 problems). **Pred** = *Predicate logic*, **Set** = *Set theory*.

Output (GPT-4)

```
Premise 1: ∃x((Friend(Taro, x) ∧ Friend(Paul, x))
Premise 2: ∀x((Friend(Paul, x) → ¬German(x))
Hypothesis: ∃x(Friend(Taro, x) ∧ German(x))

## Reasoning
Given that one friend of Taro is also a friend of
Paul’s and none of Paul’s friends are German, it
follows that this mutual friend cannot be German.
Thus, the hypothesis that one of Taro’s friends is
German is not contradictory to any of the premises,
because while we know one friend is not German, it
is still possible for Taro to have other friends who
are German.

## Answer
neither
```

Table 9: Example response from GPT-4 in the Translate-and-Explain task (Translation into predicate logic).

sentences of other types tend to include mistakes. Especially, O-type (*some-not*) and E-type (*no*) sentences are frequently confused. For example, a Japanese sentence corresponding to *Some robots are not commodities* is incorrectly interpreted as *The set of products and the set of robots do not have any common elements*.

6 Related Work

Recent years have seen active research on machine learning and deep learning models that focus on logical reasoning in natural language, including syllogistic reasoning.

Richardson et al. (2020) investigated the abilities of NLI models in handling various logical inferences involving boolean operators, quantifiers, conditionals, and negation using synthetically generated data. Yanaka et al. (2019) studied monotonicity inferences. Monotonicity inferences are simpler than syllogistic inferences in that they only have single premises, whereas syllogisms involve multiple premises with challenging combinations of quantifiers and negation. Schlegel et al. (2022) conducted an empirical study to explore the detection of formally valid inferences within controlled fragments of natural language, designed to increase satisfiability problem complexity. These studies combine pre-training and fine-tuning with relatively large datasets for logical reasoning and are not aimed at evaluating current LLMs based on in-context learning.

As datasets for learning and evaluating syllogistic reasoning, Dong et al. (2020) and Gubelmann et al. (2022) have constructed datasets using linguistic resources such as WordNet and their own lists of words, employing template-based automatic generation methods based on types of syllogisms. The dataset of Gubelmann et al. (2022) includes three-class labels (entailment, contradiction, and neutral) but does not include additional information related to reasoning biases. AVICENNA (Aghahadi and Talebpour, 2022) is a crowdsourced dataset containing binary labels that indicate whether the conclusion of a syllogism follows from two given sentences (and if so, the conclusion sentence is also provided), but it does not contain information about reasoning biases. SYLLOBASE (Wu et al., 2023) contains five types of syllogisms. It consists 50,000 syllogism samples automatically generated from existing knowledge bases, with 1,000 of these samples manually annotated as a test set. Experiments were conducted in zero-shot and few-shot settings, covering both generation and selection tasks. All of the datasets above are in English.

As a framework for generating deduction datasets, FLD (Morishita et al., 2023) and its Japanese version, JFLD (Morishita et al., 2024), have been proposed. Morishita et al. (2023) empirically verifies that language models trained with FLD demonstrate enhanced generalizable deductive reasoning abilities.

Dasgupta et al. (2022) and Eisape et al. (2023) are the works closely related to ours. Dasgupta et al. (2022) focused on belief biases in syllogistic reasoning and showed that the alignment of conclusions with human beliefs affects the performance of LLMs. In their studies, scenarios were classified based on whether the content of the conclusion contradicts our beliefs, does not contradict, or consists of meaningless words. LLMs were then tasked with determining whether a combination of two premises and one conclusion constitutes valid or invalid syllogistic reasoning in a binary choice format. This approach differs from the three-class classification used in our work. Our research aims to further explore reasoning biases in LLMs by examining various types of biases, particularly those related not only to the content of reasoning, as highlighted by belief bias, but also to the form of reasoning. Additionally, we investigated these biases in experimental settings, including the Multiple-Choice and Translate-and-Explain tasks. Eisape et al. (2023), focusing on syllogisms, investigates how language models perform logical reasoning compared to humans. Experiments were conducted using the PaLM 2 models, and the results indicate that larger models perform more accurately than smaller models and humans. It was also confirmed that even the largest models tend to make errors that reflect human reasoning biases in certain types of syllogisms.

Ando et al. (2023) is a preliminary work that precedes our research, reporting only the zero-shot performance of GPT-3.5 across 375 NLI tasks. In contrast, our study reports results on 790 NLI tasks, 80 Multiple-Choice tasks using Chain-of-Thought prompts, and 90 Translate-and-Explain tasks. Furthermore, we evaluated a broader range of models, including GPT-4, Llama-2 (13B and 70B), and Swallow (13B and 70B), in both zero-shot and few-shot settings. The findings related to the Multiple-Choice and Translate-and-Explain tasks represent entirely new contributions of this study.

7 Summary and Conclusion

In this paper, we evaluated the logical reasoning ability of LLMs by developing the NeuBAROCO dataset, which consists of syllogisms in both Japanese and English, annotated with information on reasoning biases and results from large-scale human evaluation data. The results of experiments using the NLI task demonstrated that some LLMs, particularly GPT-4, achieved high accuracy in both English and Japanese, especially in the few-shot setting. However, significant room for improvement remains in problems where the entailment relation is neutral (neither entailment nor contradiction). The results suggest that, particularly in problems labeled as *Conversion* and *Atmosphere*, there is a tendency to exhibit the same reasoning biases as humans.

Moreover, in the experiments using the Multiple-Choice task, a method commonly used in psychological studies of syllogisms, some models achieved accuracies surpassing those of human participants in large-scale experiments. It is important to note that the methodology for comparing human accuracy with LLMs is not yet fully established, and a detailed comparison with human performance remains a future challenge.

Finally, in the Translate-and-Explain task, which requires providing explanations of reasoning along with translations of syllogisms into logical expressions, many models showed improved accuracy, with translation accuracy nearly reaching 100%. However, reasoning errors remain, suggesting that the source of these errors is not the misinterpretation of the premise sentences but rather the reasoning process itself.

While syllogisms represent one of the basic forms of logical reasoning used in psychology, expanding our evaluation to include more diverse and complex natural language inferences that induce reasoning biases remains an essential challenge. This includes boolean propositional inferences (Evans, 1989), *if-then* conditionals (Johnson-Laird and Byrne, 2002), and spatial inferences involving polyadic relations (Byrne and Johnson-Laird, 1989). The biases we addressed are not specific to syllogistic reasoning and, therefore, may potentially be generalized to tasks beyond syllogistic reasoning. Addressing these issues in future research is crucial for advancing our understanding of the reasoning capabilities of LLMs.

Limitations

In this study, we used closed-source LLMs, particularly noting that the GPT models we used lack precise information about the models, such as the number of parameters, and scale and distribution of the training data. Although the NeuBAROCO dataset has not been publicly released so far, it cannot be denied that some parts of syllogistic reasoning, especially symbolic syllogisms themselves, may be included in the training data. Also, there is a risk associated with using closed-source models in scientific research due to the lack of reproducibility.

Comparing the accuracy of LLMs to human accuracy obtained from psychological experiments presents an intriguing challenge for research in both NLP and cognitive psychology regarding human reasoning with natural languages. However, to our knowledge, there is not yet a well-established methodology for comparing the accuracy of LLMs with that of humans yet. In the case of the BAROCO project, which involves more than 400 participants, further discussion is necessary on how to systematically and meaningfully compare the accuracy of LLMs with that of humans.

Research on English and Japanese LLMs is rapidly advancing even as our study progresses. While evaluations are being conducted using a variety of current representative models, both open and closed, it is naturally impossible to cover all of them comprehensively. Particularly, the number of Japanese LLMs is increasing, and conducting a systematic comparison of Japanese and English LLMs is one of the important challenges for the future.

8 Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions, which have greatly improved the paper. We also extend our gratitude to Professors Juko Ando and Chizuru Shikishima, as well as the members of their KTS-KoTReC project, for their collaboration with the last author in developing the original BAROCO test set and for providing us with the associated human subject data. This work is partially supported by JST, CREST Grant Number JPMJCR2114 and JSPS Kakenhi Grant Numbers JP24K00004, JP21K00016, JP21H00467, JP23K20416, and JP21K18339.

References

Zeinab Aghahadi and Alireza Talebpour. 2022. Avicenna: a challenge dataset for natural language generation toward commonsense syllogistic reasoning. *Journal of Applied Non-Classical Logics*, 32(1):55–71.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ruth MJ Byrne and Philip N Johnson-Laird. 1989. Spatial reasoning. *Journal of memory and language*, 28(5):564–575.

Nick Chater and Mike Oaksford. 1999. The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, 38(2):191–258.

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

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

Tiwalayo Eisape, MH Tessler, Ishita Dasgupta, Fei Sha, Sjoerd van Steenkiste, and Tal Linzen. 2023. A systematic comparison of syllogistic reasoning in humans and language models. *arXiv preprint arXiv:2311.00445*.

J St BT Evans, Julie L Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306.

Jonathan St.B. T. Evans. 1989. *Bias in Human Reasoning: Causes and Consequences*. Lawrence Erlbaum Associates, Inc.

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

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*.

Bart Geurts. 2003. Reasoning with quantifiers. *Cognition*, 86(3):223–251.

Reto Gubelmann, Christina Niklaus, and Siegfried Handschuh. 2022. 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)*, pages 38–50.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Philip N Johnson-Laird and Mark Steedman. 1978. The psychology of syllogisms. *Cognitive psychology*, 10(1):64–99.

PN Johnson-Laird and Ruth MJ Byrne. 2002. Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109(4):646–678.

Sangeet Khemlani and Philip N Johnson-Laird. 2012. Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, 138(3):427.

Jan Łukasiewicz. 1951. *Aristotle's Syllogistic From the Standpoint of Modern Formal Logic*. Oxford University Press.

Ken Manktelow. 1999. *Reasoning and Thinking*. Psychology press.

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*, pages 25254–25274. PMLR.

Terufumi Morishita, Atsuki Yamaguchi, Gaku Morio, Hikaru Tomonari, Osamu Imaichi, and Yasuhiro Sogawa. 2024. JFLD: A Japanese benchmark for deductive reasoning based on formal logic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9526–9535, Torino, Italia. ELRA and ICCL.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Rüdiger F Pohl. 2022. *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*, 3 edition. Routledge.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

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.

Viktor Schlegel, Kamen Pavlov, and Ian Pratt-Hartmann. 2022. Can transformers reason in fragments of natural language? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11184–11199, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chizuru Shikishima, Juko Ando, Pierre Grialou, Ryo Takemura, and Mitsuhiro Okada. 2005. 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*, volume 1, pages 69–85. Keio University Press Tokyo.

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

Chizuru Shikishima, Kai Hiraishi, Shinji Yamagata, Yutaro Sugimoto, Ryo Takemura, Koken Ozaki, Mitsuhiro Okada, Tatsushi Toda, and Juko Ando. 2009. Is g an entity? a Japanese twin study using syllogisms and intelligence tests. *Intelligence*, 37(3):256–267.

Keith Stenning and Michiel van Lambalgen. 2012. *Human Reasoning and Cognitive Science*. MIT Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Robert S Woodworth and Saul B. Sells. 1935. An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, 18(4):451–460.

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

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

A Examples of Prompts

A.1 Multiple-Choice task

Table 4 in Section 5 shows examples of prompts for the Multiple-Choice task.

A.2 NLI task

Table 10 shows examples of the prompts we used for the NLI task in English and Japanese. We employed the most effective prompt pattern from among several we evaluated, including those listed in Table 11. Table 12 shows examples of few-shot prompts for the task.

A.3 Translate-and-Explain task

Table 13 and Table 14 show examples of the prompts we used for the Translate-and-Explain task in English and Japanese.

B Llama-2 and Swallow comparison

In the Multiple-Choice task, Llama-2-13B performs better in English than in Japanese, while Llama-2-70B performs better in Japanese than in English. Conversely, Swallow-13B excels in Japanese, while Swallow-70B excels in English. These results indicate that performance trends can change significantly within the same model family depending on the model scale, at least for inference tasks. Furthermore, when comparing the two, Swallow-13B outperforms Llama-2-13B in Japanese but underperforms in English, whereas Swallow-70B underperforms Llama-2-70B in Japanese but outperforms it in English. Although Swallow is a model family that improves Llama-2 through continual pre-training on a large-scale Japanese corpus, continual pre-training in a specific language does not always enhance performance in that language for inference tasks, indicating a complex effect.

Input

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:

Input

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:

Table 10: Example prompts for the NLI task.

Input

Carefully evaluate the following inference, and determine whether the premises entail or contradict the conclusion. Answer exactly with entailment, contradiction, or neither. The answer may be entailment even if it contradicts common sense. For example, the answer to the following inference is entailment.

Premise 1: All people are watermelons.
Premise 2: All watermelons are delicious.
Conclusion: All people are delicious.

Input

Carefully evaluate the following inference, and determine whether the premises entail or contradict the conclusion. Answer exactly with entailment, contradiction, or neither. Let’s think step by step.

Table 11: Examples of other tested prompts not adopted.

Input

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: 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

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:

Input

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: 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

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:

Table 12: Example prompts for the NLI task in the few-shot setting.

Input

```
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".

## Input
Premise 1: Some A are B.
Premise 2: All B are C.
Hypothesis: All A are C.

## Translation to predicate logic
Premise 1: ∃x(Ax∧Bx)
Premise 2: ∀x(Bx→Cx)
Hypothesis: ∀x(Ax→Cx)

## Reasoning
[Explain your reasoning for the answer]

## Answer
[Your answer must be one word: "entailment", "contradiction", or "neither"]

## Input
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.

## Translation to predicate logic
```

Output (GPT-4)

```
Premise 1: ∃x((Friend(Taro, x) ∧ Friend(Paul, x))
Premise 2: ∀x((Friend(Paul, x) → ¬German(x))
Hypothesis: ∃x(Friend(Taro, x) ∧ German(x))

## Reasoning
Given that one friend of Taro is also a friend of Paul's and none of Paul's friends are German, it follows that this mutual friend cannot be German. Thus, the hypothesis that one of Taro's friends is German is not contradictory to any of the premises, because while we know one friend is not German, it is still possible for Taro to have other friends who are German.

## Answer
neither
```

Input

```
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。

## 入力
前提1: あるAはBである。
前提2: すべてのBはCである。
仮説: すべてのAはCである。

## 述語論理への翻訳
前提1: ∃x(Ax∧Bx)
前提2: ∀x(Bx→Cx)
仮説: ∀x(Ax→Cx)

## 論証
[答えを導く論証を説明しなさい]

## 答え
[「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]

## 入力
前提1: 太郎のある友人はポールの友人である。
前提2: ポールのどの友人もドイツ人でない。
仮説: 太郎のある友人はドイツ人である。

## 述語論理への翻訳
```

Output (GPT-4)

```
前提1: ∃x(Fxt∧Fpx)
前提2: ∀x(Fpx→~Gx)
仮説: ∃x(Fxt∧Gx)

## 論証
前提1と前提2から、「x」を「ポールの友人であり太郎の友人でもある人」に置き換えると、その人はドイツ人ではない（~Gx）と結論付けることができる。これは仮説、すなわち「太郎のある友人はドイツ人である」と矛盾している。

## 答え
矛盾
```

Table 13: Example prompts used for the Translate-and-Explain task (translation to predicate logic).

Input

```
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".

## Input
Premise 1: Some A are B.
Premise 2: All B are C.
Hypothesis: All A are C.

## Translation to set theory
Premise 1: The set of As overlaps with the set of
Bs.
Premise 2: The set of Bs is a subset of the set of
Cs.
Hypothesis: The set of As is a subset of the set of
Cs.

## Reasoning
[Explain your reasoning for the answer]

## Answer
[Your answer must be one word: "entailment",
"contradiction", or "neither"]

## Input
Premise 1: One friend of Taro is a friend of Paul's.
Premise 2: None of Paul's friends are German.
Hypothesis: One of Taro's friends is German.

## Translation to set theory
```

Output (GPT-4)

```
Premise 1: The set of Taro's friends overlaps with
the set of Paul's friends.
Premise 2: The set of Paul's friends is disjoint
from the set of Germans.
Hypothesis: The set of Taro's friends overlaps with
the set of Germans.

## Reasoning
The first premise tells us that Taro and Paul
share at least one friend. The second premise
tells us that none of Paul's friends are German.
Therefore, since the friend they share can't be
German, the hypothesis that one of Taro's friends is
German doesn't necessarily follow from the premises
nor contradicts them, since Taro could have other
friends that aren't friends with Paul.

## Answer
Neither
```

Input

```
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。

## 入力
前提1: あるAはBである。
前提2: すべてのBはCである。
仮説: すべてのAはCである。

## 集合論への翻訳
前提1: Aの集合とBの集合は共通部分を持つ。
前提2: Bの集合はCの集合の部分集合である。
仮説: Aの集合はCの集合の部分集合である。

## 論証
[答えを導く論証を説明しなさい]

## 答え
[「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]

## 入力
前提1: 太郎のある友人はポールの友人である。
前提2: ポールのどの友人もドイツ人でない。
仮説: 太郎のある友人はドイツ人である。

## 集合論への翻訳
```

Output (GPT-4)

```
前提1: 太郎の友人の集合とポールの友人の集合は共通
部分を持つ。
前提2: ポールの友人の集合はドイツ人の集合と共通部
分を持たない。
仮説: 太郎の友人の集合とドイツ人の集合は共通部分を
持つ。

## 論証
太郎のある友人がポールの友人であり、ポールのすべて
の友人がドイツ人でないという前提から、太郎のすべて
の友人がドイツ人でないという結論を導くことができ
る。したがって、仮説は前提と論理的に矛盾している。

## 答え
矛盾
```

Table 14: Example prompts used for the Translate-and-Explain task (translation to set theory).