Cognitive Biases in Syllogistic Reasoning: A Comparative Study of Humans and Large Language Models

Stefanie Krause*, Anna-Lena Henk and Frieder Stolzenburg*

Harz University of Applied Sciences, Wernigerode, Germany

Abstract

Understanding human cognitive processes for reasoning is a key step towards developing human-level AI reasoning systems. Syllogistic reasoning — a classical form of deductive reasoning — is a well-established paradigm for investigating human reasoning, known to elicit systematic cognitive biases such as belief bias and atmosphere effects. In this study, we compare the reasoning patterns of humans and three state-of-the-art Large Language Models (LLMs), namely GPT-4, GPT-4o and o4-mini, across multiple syllogism types with belief and atmosphere effects. Using an online questionnaire, we examine the biases in human syllogistic reasoning and compare this with the LLMs responses. We identify biases and differences in accuracy between different LLMs and humans. Belief-inconsistent statements, which contradict generally accepted common knowledge, and the atmosphere effect impact human and LLM answers. Our results show that, although LLMs outperform human participants, they exhibit human-like cognitive effects. We further evaluate the new reasoning model o4-mini, which performs nearly perfectly on the syllogism reasoning tasks and compare logic, human reasoning and LLM reasoning.

Keywords

syllogistic reasoning, large language models, cognitive bias, human-like reasoning

1. Introduction

The advancements of LLMs have garnered significant attention in recent years, particularly following the launch of OpenAI's ChatGPT. LLMs strive to emulate human-like interactions [1]. These skills encompass not only language comprehension but also logical reasoning, which is fundamental to human cognition [2, 3, 4]. Reasoning abilities are vital for structured arguments and deriving valid conclusions. Many real-world applications of LLMs, such as legal reasoning [5, 6], medical diagnostics [7], and automated decision-making [8], rely on the ability to process logical arguments accurately. Furthermore, a recent study has shown that university students have high confidence in using LLM tools for problem-solving and logical reasoning [9]. While studies showcased LLMs' theoretical reasoning capacities on abstract logical problems, they do not entirely capture their practical utility in real-life applications where context drastically affects outcomes [10]. The extent to which LLMs can perform logical reasoning and make their reasoning transparent and understandable to users is still unclear [11]. Our goal is, therefore, to evaluate the current status of LLM reasoning abilities in comparison to human reasoning and provide hints for future directions. Another important aspect that strengthens the need for more efficient LLMs with enhanced reasoning abilities is that models could achieve better performance with less data and fewer computations, leading to more sustainable AI practices with lower energy consumption and environmental impact. Hence, it is a crucial aspect of assessing the abilities of LLMs [12, 13, 14].

Logical reasoning has a long history, dating back to Aristotle, and has been extensively researched [12, 15]. There are various types of reasoning that have been studied so far, e.g., commonsense reasoning [16, 17]. Another fundamental component of logical reasoning are syllogisms, which have modeled human intelligence in philosophy since Aristotle [18]. Logical deductive reasoning abilities, e.g., solving

AIC 2025: 10th International Workshop on AI and Cognition, October 2025, Bologna, Italy
*Corresponding author.
skrause@hs-harz.de (S. Krause); anna.henk@icloud.com (A. Henk); fstolzenburg@hs-harz.de (F. Stolzenburg)
0000-0002-1271-7514 (S. Krause); 0000-0002-4037-2445 (F. Stolzenburg)

syllogisms, are known to be a good measure of an individual's general cognitive ability and well-studied for human reasoning [18]. A syllogism represents a precise form of argumentation where two propositions (the premises) are systematically analyzed to establish a conclusive statement [12, 14, 19], as shown in Example 1:

Example 1. *Syllogism [19]:*
Premise 1: All humans are mortal.
Premise 2: Socrates is a human.
Conclusion: Socrates is mortal.

Syllogistic reasoning is regarded as a natural form of reasoning, i.e., humans naturally tend to conclude patterns similar to syllogisms [12]. Nonetheless, syllogisms fall in the realm of formal logic, which adheres to stringent formal rules and can pose challenges even for trained people [20]. A wide range of research already exists on syllogistic reasoning and show that people tend to have certain biases while reasoning syllogistically [3, 21] (cf. Section 2.2). As LLMs are trained using human data, this raises the question of whether LLMs exhibit similar biases as humans [21]. Even though syllogistic reasoning in LLMs has been researched in recent years, mostly different aspects have been considered, such as the investigation of extended syllogisms instead of traditional ones or the focus on negations. Extended syllogisms have either a boolean inference where conjunctions "and" and "or" appear between terms or premises are a conditional sentence of the form "If P then Q", where P or Q can be a negated sentence [12].

In our study, we analyze traditional syllogisms that include belief and atmosphere effects with different LLMs, including a recently released reasoning LLM and compare these results to a human study. In literature, a direct comparison with human reasoning is rarely drawn [12, 13, 14, 22]. We build on previous research [12, 23] and compares different LLMs: GPT-4 [1], GPT-4o [24] and o4-mini [25]. The comparison focuses on three types of syllogisms: belief-consistent, belief-inconsistent, and symbolic. Syllogisms in which the premises and the conclusion correspond to generally accepted common knowledge are called *belief-consistent*, cf. Example 1 [12]. Accordingly, *belief-inconsistent* statements contradict common knowledge, such as "*All humans are robots.*" [12]. In the *symbolic* type, the syllogisms contain only letters instead of concrete known terms [12]. In these types, humans exhibit different biases [21]. Another bias, namely the atmosphere effect, will be examined more closely as well [12, 15].

This work uses examples from the *NeuBAROCO* dataset [12] to examine how the logical reasoning of LLMs differs from that of humans. This comparison is conducted on the one hand through queries to GPT-4, GPT-4o, and o4-mini, and on the other hand through an online questionnaire for humans. The key contributions of this paper are:

- GPT-4, GPT-4o, and o4-mini are evaluated on a variety of syllogistic reasoning tasks, including belief-consistent, belief-inconsistent, and symbolic syllogisms.
- A human study is conducted to compare human reasoning with LLM reasoning and to assess cognitive biases such as the atmosphere effect, providing insight into similarities between human and LLM syllogistic reasoning.
- Formal logic, LLM reasoning, and human reasoning are compared to identify current differences and outline future research directions.

The paper is structured as follows: First, Section 2 takes a closer look at the theoretical background. This includes an introduction to syllogistic reasoning and LLMs. Furthermore, previous research is discussed in more detail in Section 3, and we derive the research questions in Section 4. Section 5 is dedicated to the methodology. It discusses the selection of the dataset, the questionnaire structure, the implementation of the experiment with the LLMs, and the evaluation methodology. Then, we present and discuss the questionnaire results and the experiment (Section 6) and compare the different types of reasoning (Section 7). Finally, Section 8 concludes the work and provides an outlook.

2. Background

2.1. Syllogistic Reasoning

Syllogisms consist of three *assertions*: the *major premise*, the *minor premise* and the *conclusion* (see Example 2).

Example 2. *Syllogism:*
Premise 1: Some A are B.
Premise 2: All B are C.
Conclusion: Some A are C.

Each assertion relates two categories, with each *category* representing a set of individuals or objects. Like Aristotle, we call the category that appears in both premises *middle term* [3]. In Example 2, category B is the *middle term*. Both other categories appear together in the conclusion and are called *end terms* [3, 26]. Depending on the arrangement of the three categories, i.e., middle term and end terms in the premises and the conclusion, this results in so-called *figures* [3, 20]. Table 1 shows the four possible figures.

Figure 1	**Figure 2**	**Figure 3**	**Figure 4**
A - B	B - A	A - B	B - A
B - C	C - B	C - B	B - C
A - C	A - C	A - C	A - C

Table 1
The four syllogistic figures (adapted from: [3, p. 3])

The three syllogistic assertions can take four different forms, these are called *mood* or *mode*, and vowels have been assigned to the traditional moods [27, 28]. The moods are distinguished by the quantifiers "*all*" (**A**), "*none*" (**E**), "*some*" (**I**), and "*some...not*" (**O**), which are each marked with vowels. However, these four quantifiers are only the traditional quantifiers. Various studies have also used syllogisms with quantifiers such as "*most*" or "*few*", but extensions of the traditional four moods also exist [29, 20]. In [20], for example, a fifth mood **U** is described, which serves as an extension of the moods **A** and **I**. **A** means the classical universal quantifier and therefore also implies **I** (if the domain is not empty). On the other hand, the extended mood **U** starts here by stating: "*some, but not all...*" [20, p. 4].

Two of the four traditional moods are *universal*, and two are *particular*. Similarly, two are *affirmative*, and two are *negative* [3]. Table 2 shows examples of this, with the vowels assigned directly to the moods:

All A are B.	affirmative universal (abbreviated as "A")
Some A are B.	affirmative particular (abbreviated as "I")
No A are B.	negative universal (abbreviated as "E")
Some A are not B.	negative particular (abbreviated as "O")

Table 2
Syllogistic moods (adapted from: [3, p. 5])

Based on the quantifiers used in the premises and the conclusion, the mood of the whole syllogism can be expressed by a sequence of three vowels [26]. In Example 2, the mood is **IAI**. Each premise and conclusion can take one of the four moods, and depending on the arrangement of the categories, the syllogism corresponds to one of the four figures. This results in 256 possible syllogisms [27, 26]. From the point of view of psychologists, though, this number doubles to 512 syllogisms when also considering that the premises can be arranged in different orders [26, 3, 20]. In classical logic, the conjunction (AND-operator) is commutative, which means that $A \wedge B$ is equivalent to $B \wedge A$ [19]. Therefore, the order of the premises is not relevant in logic [19]. Psychologists believe that the order affects human reasoning, as the presentation of premises can influence how humans process and interpret information [26, 30].

A syllogism is considered (logically) *valid* if the conclusion logically follows from the premises. This means that if the premises are true, the conclusion must also be true. A syllogism is *invalid* if the conclusion does not logically follow from the premises, even if the premises and conclusion might separately be true. Note that validity is determined by the form and structure here, not by the actual content or truth of the assertions involved.

2.2. Human Biases

Research on syllogistic reasoning has shown that human thinking is often influenced by various biases. Humans can be influenced by different aspects, such as content and mood, or assumptions that mislead them [3, 26]. Well-known examples of biases are: *conversion errors* where the reversal of an assertion mistakenly is considered equivalent with the original version, e.g., "*All A are B.*" and "*All B are A.*" [31]; *figural effects* where the preferred arrangement of the assertions is essential; *content effects* where the content and thus the believability of the assertions influence humans; *atmosphere effect* where the general mood of the assertions influences the acceptance of the conclusion [3, 12, 21]. In this work, the atmosphere and content effects are considered more deeply. These two biases show how information is processed, both in terms of the content assumptions and the structure of the premises, which is why they are appropriate for comparing the thought processes of humans and LLMs.

2.2.1. Content Effects

As the name suggests, content effects concern the content of a syllogism [32]. Numerous studies have demonstrated that the content of a logical task influences how people respond. For example, humans tend to regard a conclusion as formally valid if it is consistent with their knowledge [21, 33]. This form of the content effect is called *belief bias* [12, 21]. An example of this, which was examined in more detail by [34]:

Example 3. *Syllogism:*
Premise 1: All flowers have petals.
Premise 2: All roses have petals.
Conclusion: All roses are flowers.

Even if the content of the syllogism is belief-consistent, it may still not be formally valid. Nevertheless, humans tend to recognize the example as valid [34]. Conversely, it is also more difficult for people to draw valid conclusions if a syllogism does not correspond to their knowledge [21]. In research studies, assertions are therefore divided into two groups: *belief-consistent* and *belief-inconsistent* [32] (sometimes called believable or congruent and unbelievable or incongruent, respectively). Syllogisms are described as belief-consistent if the content of all three assertions corresponds to the commonsense beliefs of humans. Belief-inconsistent assertions, conversely, do not correspond to common sense or are contrary to human acceptance [12]. However, it is essential to note that it is sufficient for only one of the three assertions to be belief-inconsistent for the entire syllogism to be considered belief-inconsistent [12]. Belief biases are often examined in connection with the acceptance of formally valid or invalid syllogisms [32, 34]. Table 3 presents examples of such an examination.

The purpose of experiments with syllogisms, such as those in Table 3 is to determine the extent to which human belief can lead to bias [32]. Content effects are not only related to belief but also include effects that arise in connection with content in general [32]. In observations that concentrate on formal logical correctness, symbolic categories are often used for assertions, as shown in Example 2. A *symbolic* assertion (also known as abstract) means that the categories are replaced by numbers or letters so that they no longer have any reference to real knowledge [12]. At latest since the famous Wason selection task [35], it is well-known that it makes a significant difference whether people have to solve an abstract reasoning task or a concrete task with concrete categories. Other forms that have been used to evaluate content effects are, for example, fantasy words or arbitrarily realistically selected content, where the content has no relation to knowledge [32, 21].

		belief-consistent
valid	**P1**	All friends of Paul are German.
	P2	All friends of Taro are friends of Paul.
	C	All of Taro's friends are German.
invalid	**P1**	Every sparrow is a bird.
	P2	No stone is a sparrow.
	C	No bird is a stone.
		belief-inconsistent
valid	**P1**	All tomatoes are plants.
	P2	Some animals are tomatos.
	C	Some animals are plants.
invalid	**P1**	All metals are liquids.
	P2	Some metals are stones.
	C	All stones are liquid.

Table 3
Examples of the four types of syllogism from the NeuBAROCO dataset [12].

2.2.2. Atmosphere Effect

The atmosphere effect describes the tendency of people to be influenced by the mood of the premises when it comes to accepting a conclusion [36, 37]. This effect was first defined more precisely by [36] already in 1935. According to this, people are more likely to accept a conclusion with the same quantifiers as one of the premises [3]. People are, therefore, guided by the atmosphere or global impressions [38, 37]. This effect is important as it highlights the extent to which the atmosphere shapes the acceptance of invalid inferences [3]. In contrast to other human biases, which are concerned with the content or formal structure of syllogisms, the atmosphere effect is only concerned with the quantifiers used and, thus, the mood [26, 39]. In the following example, the atmosphere of the premises creates an impression of credibility that makes it easier to accept the conclusions as valid, even though only the first conclusion is truly valid [39]:

Example 4. *Syllogism:*
Premise 1: Some A are B.
Premise 2: No B are C.
Conclusion 1: Some A are not C.
Conclusion 2: Some C are not A.

2.3. Large Language Models

Inspired by the human brain, artificial neural networks are information-processing systems that consist of neurons linked in network-like structures [40]. The transformer architecture [41] has been introduced to overcome weaknesses in the previous models, such as sequential computation. The architecture has an encoder-decoder structure and uses attention mechanisms. This parallel structure is used to address problems that arose due to memory limitations, particularly with longer text sequences.

Generative Pre-Trained Transformer (GPT) is the primary name of a series of language models developed by the company OpenAI [4]. In 2023, the variant GPT-4 was introduced, which was trained on large amounts of data [1]. Various studies show that the model is better at solving more complex tasks than its predecessors, and its performance in connection with human problem solving has increased [4]. The GPT-4o model was introduced in May 2024. The *o* stands for *omni*. It is capable of solving complex tasks and processing images, audio, and text in real time, achieving a human-like speed of communication [24]. Despite the numerous stages of development, the models continue to reach their limits; for example, the models tend to hallucinate (generate content that appears factual but is

ungrounded [42]) and make mistakes when solving reasoning tasks [1].

To solve these problems, so-called reasoning models have been developed. The idea is just as a person takes time to think deeply before answering a difficult question, those models have increasing computational resources during inference to improve output quality (called inference-time scaling). Other methods are applied as well to improve the reasoning quality, such as reinforcement learning [43, 44], meta chain-of-thought [45] and search and optimization algorithm, such as Monte Carlo tree search, to systematically explore and evaluate reasoning paths [46]. Reasoning models learn to break down complex problems into simpler steps and learn to try a different approach if the current one is not effective. In early 2025 reasoning LLMs, such as OpenAI's o3 and o4-mini, have been introduced. Those models reason more deeply and offer more reliable answers by formulating novel hypotheses and critically assessing them [47]. This claim has been shown by two empirical studies that demonstrate a strong tendency for exploratory behavior, which is evident in the formulation of novel hypotheses and the pursuit of alternative solution paths [48, 49]. According to [45], slow-thinking models engage in a latent generative process, which is especially evident when predicting subsequent tokens. A more detailed analysis of reasoning models with their core methods for advanced reasoning capabilities is presented in [46].

3. Related Work

The extent to which LLMs show similarities to humans, particularly in logical reasoning, is a topic of great interest. For this purpose, various logical tasks, such as syllogisms, have been used to evaluate the abilities of LLMs in recent literature. Two studies examined various transformer-based pre-trained LLMs to determine the extent to which they are prone to belief biases, conversion errors, or the atmosphere effect during syllogistic reasoning with the NeuBAROCO dataset [12]. GPT-3.5, RoBERTa and BART are evaluated, which showed that the models are prone to similar biases and errors like humans [12]. In the other study, they compared different LLMs using a new chain-of-thought prompting method, which forces LLMs to translate syllogisms into abstract logical expressions and afterwards explains their reasoning process [23].

Parmar et al. [50] aim to answer whether LLMs can reason about natural language. To this end, the work evaluates 25 different reasoning patterns from the areas of first-order, propositional, and non-monotonic logic. For this purpose, a dataset called LogicBench was generated using GPT-3.5. The dataset contains two types of tasks: multiple-choice, in which the permissible conclusion must be selected, and binary-choice, in which a decision must be made as to whether the specified conclusion is valid. Several LLMs of different sizes are evaluated. The study shows that LLMs have difficulties in processing complex logical patterns and negations and often miss contextual information. The results, therefore, show that it is relevant to evaluate where LLMs show weaknesses, like in negations.

Like [50], Espejel et al. [51] also examine a number of logical tasks, including deductive reasoning such as syllogisms. The work concentrates on the manual evaluation of ChatGPT-3.5, ChatGPT-4, and BARD. They use different prompting techniques in the tasks, which has shown that the wording of the tasks can lead to differences in performance. Their results also show that, on average, GPT-4 performed better on the tasks than the other models.

Bertolazzi et al. [52] examine the impact of chain-of-thought reasoning, in-context learning (ICL), and supervised fine-tuning (SFT) on syllogistic reasoning. Their study considers syllogisms whose conclusions either align with or contradict world knowledge, as well as cases involving multiple premises. The findings show that both ICL and SFT enhance model performance on valid inferences; however, only SFT substantially reduces reasoning biases without compromising model consistency. ICL fails to improve models' ability to detect when no valid conclusion can be drawn. They analyse the Atmosphere effect as well and conjecture that the reason zero-shot LLMs seldom generate "nothing follows" is because they rely on the quantifiers in the premises as guidance.

Liu et al. [53] evaluate ChatGPT, GPT-4, and RoBERTa on multi-choice reading comprehension and natural language inference tasks. It was noticeable that the models performed better on older datasets

than on newly released and out-of-distribution datasets. Even though the comparison with humans was not the focus of this work, the work contains the average results of humans when solving the tasks of the well-known datasets LogiQA [54] and ReClor [55]. However, LLMs performed worse than humans in all datasets in their work in early 2023.

Lampinen et al. [21] contain a direct comparison with humans. The research evaluated different LLMs and humans in three tasks: syllogisms, the Wason selection task [35], and natural language inference. Different contents are tested, from realistic to unrealistic to nonsensical. In this way, the extent to which the content of the tasks influences the LLMs, as is the case with humans, is examined. The models use similar instructions for the evaluation to explain the tasks as they did with humans. The tasks with humans are carried out as part of an online experiment. The results of this study, especially concerning the syllogistic experiment, show that the LLMs are prone to similar biases as humans. In particular, humans and LLMs are more likely to accept a conclusion as valid if the content matches their beliefs. In addition, the non-sense category also led to increased acceptance of non-valid assertions.

Similarities to our research has [56]. It focuses on comparing humans and the LLMs of the PaLM 2 family [57]. For this purpose, they use the dataset from [58], which was developed for human research in the field of syllogisms. Syllogisms are used in which the categories bear no realistic connection to one another. The LLMs are given eight possible conclusions or the choice that no conclusion is possible. The results of this work indicate that the larger models show higher accuracy in solving the tasks than the smaller models and perform better overall than humans. Nevertheless, the LLMs exhibited systematic errors similar to humans.

Due to the continuous release of new LLMs, the evaluation of models remains a current research topic. In particular, the comparison of new research results with previous ones in order to observe developments more closely and to test new models appears to be popular. Our work, therefore, focuses on the evaluation of the currently latest GPT models. A direct comparison with humans provides a deeper insight to evaluate more precisely which biases humans and LLMs show in the same tasks.

4. Research Questions and Hypotheses

This work investigates two research questions (RQs) concerning syllogistic reasoning in LLMs and humans. We focus on syllogism because it has been thoroughly examined by logicians and psychologists, and it serves as the fundamental building block for more complex forms of reasoning. Given that the question arises whether LLMs reason similarly to humans, a direct comparison between humans and LLMs is particularly important. To this end, we are focusing on the following RQs:

- RQ 1: Does the performance of LLMs and humans differ in syllogistic reasoning?
- RQ 2: What impact do well-known human biases such as the content or atmosphere effects have on LLMs?

Given that LLMs are trained on human-generated data, we assume these models are biased similarly to humans. We expect the LLMs to be affected by belief-inconsistent syllogisms similar to the research results of [23, 12, 21, 52]. Further, we hypothesize that the GPT models GPT-4, GPT-4o and o4-mini are affected by the atmosphere effect. The atmosphere effect is a well-documented human cognitive bias during syllogistic reasoning [52]. Studies show that large language models such as GPT-3.5, RoBERTa, and others also display this bias [52, 23, 12]. We also assume that neutral syllogisms are the most difficult type of syllogism for LLMs, as we are aware that while entailment and contradiction can often be identified via syntactic or lexical cues, neutral cases require deeper semantic understanding [59]. We further hypothesize that the latest o4-mini reasoning model performs better than the GPT-4.0 and GPT-4o versions. All our hypotheses are summarized below:

- H1: LLMs solve belief-inconsistent syllogisms less accurately than belief-consistent ones.
- H2: LLMs solve symbolic syllogisms more accurately than belief-based (belief-inconsistent or belief-consistent)

- H3: The atmosphere affects the accuracy of LLMs to solve syllogisms.
- H4: LLMs solve neutral syllogisms less accurately than entailment or contradiction syllogisms.
- H5: The newest reasoning model o4-mini has a higher overall accuracy than the versions GPT-4 and GPT-4o.

5. Method

This section first describes the selection of a suitable dataset for the comparison of humans and LLMs on syllogism. Afterwards, the questionnaire for our human study and the experiments with the LLMs are described.

5.1. Dataset

We chose to utilize the dataset *NeuBAROCO* presented in [12]. It is based on the original Japanese dataset *BAROCO* [60] developed in 2004 for a study with humans. For the study in [12], the dataset was translated from Japanese into English and adapted for the evaluation of LLMs. In total, the dataset contains 375 syllogisms. For our study, we did not consider extended syllogisms, therefore, we used only 318 syllogisms of the NeuBAROCO dataset.

Depending on whether the premises support the conclusion, every syllogism is labeled as *entailment*, *contradiction*, or *neutral* example. These labels represent the possible answers in the experiments and serve as the ground truth (called gold answer), which defines the correct classification for each syllogism. Hereby, it follows the traditional assumption that a universally quantified category is not empty [12]. However, note that the logical principle that false premises lead to arbitrary (*ex falso quod libet*) is not applied here. The same is true for human reasoning – people do not have a rule like "from contradiction, infer anything". Cognitive psychology has shown that people do not treat contradictions as explosively as formal logic. Instead, contradictions often lead to confusion, rejection of the task, or suspension of judgment, rather than deriving arbitrary conclusions. The mental models theory [61] explains this by claiming that people construct possible models of a situation, and contradictory premises lead to the failure of model construction.

As mentioned earlier, validity is determined solely by the form and logical structure, not by the actual content or truth of the assertions involved. The dataset contains our three types of syllogisms: belief-consistent, belief-inconsistent, and symbolic. Capital letters are used for the symbolic type. To solve belief-(in)consistent syllogism (logically) correctly, i.e., with respect to their validity, one needs to replace the terms with symbols (such as A, B, C) to solve the symbolic syllogism. The syllogisms are labeled as to whether they support a conversion error or an atmosphere effect. These human biases are explained in Section 2.2. The atmosphere label has been given to all syllogisms where the correct answer is *neutral* and contains a premise of type **O** with conclusion of type **E**, **I** or **O** or a premise of type **I** with conclusion of type **I** or **O**.

5.2. Questionnaire

We conducted an online questionnaire to evaluate the syllogistic reasoning of humans compared to LLMs. In our questionnaire, we use multiple-choice questions. To further enhance the quality of our research, we included questions about educational qualifications and language level. The latter is particularly crucial to gauge the potential impact of language barriers on participants' responses. In this question, we also emphasized the desirability of at least an A2 level of English proficiency. In the questionnaire, we are not aiming for logic experts but people with diverse higher educational backgrounds (at least a secondary school certificate up to a master's degree). A huge number of participants (88 %) had at least a university entrance qualification, which is sufficient to have a basic logical understanding. We aimed for young people with a sufficient educational background to compare our results to a previous study on young Japanese on the same syllogism dataset (with tasks in Japanese) [18].

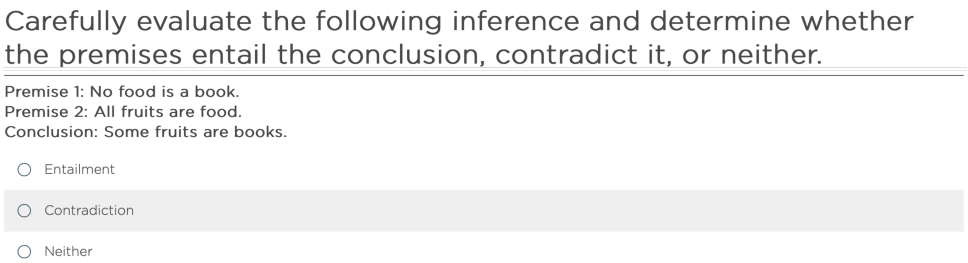


Figure 1: Example from the questionnaire (correct answer: contradiction)

Prior to undertaking the syllogism tasks, participants received a detailed explanation of syllogisms through an illustrative example, and the interpretation of the various answer options was explicitly defined. Participation in the study was entirely voluntary. We did not collect any personally identifiable information, such as gender. Based on the research questions, a selection of syllogisms was chosen from the dataset described above. Questions from each of the three different syllogism types and two questions that lead to the atmosphere effect are selected. We did not choose more questions due to potential fatigue issues. When choosing the syllogisms, we also considered which figure the syllogisms belong to, asking each figure at least twice. The questions are designed so that participants are asked to evaluate a given syllogism consisting of two premises and a conclusion and to indicate whether the conclusion is entailed by the premises, represents a contradiction, or neither. Accordingly, each question has three possible answers. Figure 1 shows a question from the questionnaire.

A random order of syllogisms was selected for each participant to avoid any patterns emerging from the syllogism types. We utilized a software program approved by our university called Tivian [62] to create the questionnaire. We conducted a pre-test with a small group of participants to identify any misunderstandings that might arise from the way questions are formulated. Based on the feedback from the pre-test, we revised the example on the introduction page and the explanation of the answer options. The questionnaire was then made available online and distributed through various channels, such as LinkedIn or via email distribution lists. The responses of 188 participants out of 217 are considered for the analysis. Participants who stated that their language level was A2 or lower are filtered out. Other participants are not included due to short processing times, as the results cannot be considered reliable.

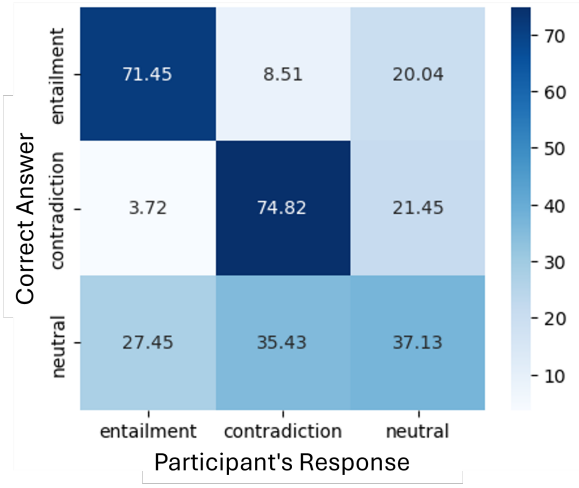
5.3. Experiment using GPT models

For the evaluation with LLMs, the adapted dataset was queried in standardized prompts. This evaluation includes the GPT-4, GPT-4o, and o4-mini models, enabling a direct comparison among the three LLMs. Relevant known specifications of these models are presented in Table 4. It is important to mention that GPT-4 and GPT-4o learn reasoning purely from deep learning, and there is no explicit reasoning module, other than o4-mini. We chose GPT-4o as compared with ten top LLMs (using metrics such as throughput, response time, and latency) GPT-4o was the best model [63], and we want to compare its capabilities with its predecessor. Further, we aim to evaluate OpenAI's new reasoning model 04-mini. The reasoning model o4-mini is trained with reinforcement learning to perform reasoning. Reasoning models reason before they answer, producing a long internal chain of thought before responding to the user. They introduce reasoning tokens in addition to input and output tokens. The models use these reasoning tokens to "think", breaking down the prompt and considering multiple approaches to generating a response. After generating reasoning tokens, the model produces an answer as visible completion tokens and discards the reasoning tokens from its context.

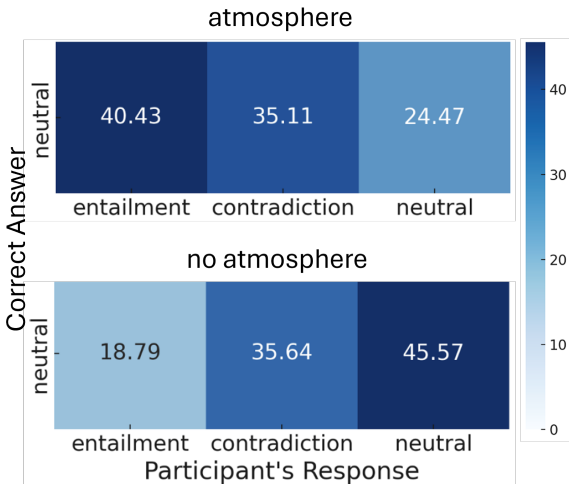
The three LLMs have been used via OpenAIs API. We used model versions o4-mini-2025-04-16, gpt-4o-2024-08-06 and gpt-4-06-13 with default parameters. We ran every experiment three times to check the consistency of the LLM output. Another important factor is that the NeuBAROCO dataset we use for our study was not available at the time of training for all three evaluated LLMs. The instruction prompt for the LLMs was designed similarly to the questionnaire: the motivation, purpose, and an

	GPT-4	GPT-4o	o4-mini
Initial Release Date	March 14, 2023	May 13, 2024	April 16, 2025
Knowledge Cutoff	Dec 01, 2023	Oct 01, 2023	Jun 01, 2024
Reasoning token	no	no	yes
Context window	8192	128000	200000
Parameters	1.76 trillion	YTD (yet-to-disclose)	YTD

Table 4
Comparison of GPT-4, GPT-4o and o4-mini specifications.



(a) Confusion matrix of human study in %.



(b) Confusion matrix of human study in % for neutral syllogism with and without atmosphere effect.

example are first explained. It was also pointed out that the tasks should only be answered with one word: *entailment*, *contradiction* or *neutral*. The procedure was carried out similarly for the three LLMs, and 318 syllogisms were queried in a random order.

6. Results and Discussion

We first analyze the outcomes of the human study before proceeding to the evaluation involving LLMs. In our human study, we evaluated eleven carefully selected syllogisms by 188 participants. The overall accuracy in our human study is 57 %. The confusion matrix in Figure 2a compares the accuracy of the syllogism types entailment, contradiction and neutral. It shows that neutral syllogisms are very hard to solve for humans, with only 37 % accurate answers. The distribution clearly shows that humans are rather uncertain about neutral syllogisms, with false responses of type entailment with 27 % and contradiction 35 %. In Figure 2b we compare neutral syllogisms with and without the atmosphere effect to analyze its impact on human reasoning. Neutral syllogisms with atmosphere effect have a low accuracy of 24 %. Most syllogisms with atmosphere effect are with 40 % considered as entailment. When comparing the neutral syllogisms without the atmosphere effect, it is noticeable that 46 % of the syllogisms are correctly classified as neutral. By comparing the two confusion vectors, it can be assumed that the atmosphere effect influences humans. However, it must be taken into account that due to the small number of syllogisms, there are only three syllogisms with the atmosphere effect in the questionnaire.

To directly compare human performance to LLMs, we evaluated the eleven syllogisms from the questionnaire with three LLMs. As reported in Table 5 o4-mini answered all the syllogisms included in the questionnaire correctly, while GPT-4 made two errors and GPT-4o one error for the content type neutral. We already recognized that neutral syllogisms are difficult to answer, especially for humans.

In the following section, we compare the results of our human study and three different LLMs. We

No.	Gold Answer	Content Type	A	Human Majority	GPT-4	GPT-4o	o4-mini
S1	contradiction	inconsistent	no	contradiction	contradiction	contradiction	contradiction
S2	contradiction	consistent	no	contradiction	contradiction	contradiction	contradiction
S3	contradiction	symbol	no	contradiction	contradiction	contradiction	contradiction
S4	entailment	inconsistent	no	entailment	entailment	entailment	entailment
S5	entailment	consistent	no	entailment	entailment	entailment	entailment
S6	entailment	symbol	no	entailment	entailment	entailment	entailment
S7	neutral	inconsistent	no	neutral	contradiction	neutral	neutral
S8	neutral	consistent	yes	entailment	neutral	entailment	neutral
S9	neutral	symbol	no	neutral	neutral	neutral	neutral
S10	neutral	consistent	yes	entailment	neutral	neutral	neutral
S11	neutral	consistent	yes	contradiction	contradiction	neutral	neutral

Table 5
Evaluation of 11 Syllogism form our questionnaire. Comparison of LLM answers and Human Majority Decisions. A stands for atmosphere present. Incorrect answers are marked in red. Abbreviation: inconsistent= belief-inconsistent, consistent= belief-consistent.

	Humans	**GPT-4**	**GPT-4o**	**o4-mini**
All	56.53 (sd 0.50)	73.79 (sd 0.30)	77.67 (sd 0.68)	95.60 (sd 0.00)
Syllogism Types				
Entailment	71.30 (sd 0.45)	84.00 (sd 0.00)	86.33 (sd 1.89)	92.00 (sd 0.82)
Contradiction	74.69 (sd 0.43)	85.42 (sd 0.00)	94.44 (sd 1.94)	93.06 (sd 0.98)
Neutral	36.79 (sd 0.48)	64.51 (sd 0.55)	67.84 (sd 1.82)	98.43 (sd 0.55)
Content Types				
belief-consistent	49.09 (sd 0.50)	78.63 (sd 1.08)	80.66 (sd 1.90)	93.38 (sd 0.36)
belief-inconsistent	56.51 (sd 0.50)	58.50 (sd 0.46)	60.78 (sd 2.12)	94.77 (sd 0.46)
Symbolic	68.98 (sd 0.46)	83.75 (sd 0.00)	92.92 (sd 1.56)	100 (sd 0.00)
Atmosphere				
No Atmosphere	63.76 (sd 0.48)	80.06 (sd 0.22)	86.45 (sd 1.14)	94.55 (sd 0.22)
Atmosphere present	24.06 (sd 0.43)	60.90 (sd 0.45)	59.62 (sd 4.15)	97.76 (sd 0.45)

Table 6
Comparison of the answers accuracy in % for our human study and three different GPT versions on different syllogism and content types. Results are the average over three runs, including the standard deviation (sd).

ran each LLM three times over 318 syllogisms and calculated the mean accuracy and standard deviation (sd). Table 6 shows the response accuracies of humans, GPT-4, GPT-4o and o4-mini in comparison.

Overall, humans performed worst at 57 %, the GPT-4 and GPT-4o model perform rather similar around 75 % and o4-mini performed best at 96 % mean accuracy. The newer o4-mini model outperformed its larger predecessor models, although the exact size of the models is not publicly disclosed. This contrasts with the findings of [56], which claim that larger models demonstrate higher accuracy in solving tasks compared to smaller models. This discrepancy might be due to the new functionalities of the reasoning model. Humans and the LLMs GPT-4 and GPT-4o have in common that the response accuracy is lowest for neutral syllogisms and highest for contradiction syllogisms. In our questionnaire analysis, we found that the majority of humans answered three out of four neutral syllogisms incorrectly. This trend can not be found with the newest reasoning model o4-mini. The neutral syllogisms are easy to solve for this reasoning LLM with a nearly perfect accuracy of 98 %. When considering different content types, humans perform worse than the LLMs in all syllogism types but have a rather similar accuracy to GPT-4 and GPT-4o for belief-inconsistent syllogisms. For both humans and LLMs, symbolic syllogisms yield the highest levels of accuracy. In contrast to the LLMs, humans have the lowest accuracy for belief-consistent syllogisms, whereby the value is negatively influenced by the syllogisms with atmosphere effect, as these are only queried as belief-consistent syllogisms. For the LLMs GPT-4

Hypothesis	Supported?	Interpretation
H1	Yes, Chi² = 8.33, p= 0.0039 < 0.05	Belief bias has a significant effect on LLMs.
H2	Yes, Chi² = 7.59, p=0.0059 < 0.05	Symbolic syllogisms are easier to solve for LLMs.
H3	Yes, Chi² = 20.21, p = 6.93e-06 < 0.05	Atmosphere significantly affects LLMs.
H4	Yes, Chi² = 61.95, p = 3.52e-15 < 0.05	Neutral syllogisms are more difficult than entailment or contradiction.
H5	Yes, Chi² = 89.06, p = 3.83e-21 < 0.05; Yes, Chi² = 45.91, p = 1.24e-11 < 0.05	o4-mini outperforms GPT-4. o4-mini outperforms GPT-4o.

Table 7
Significance analysis of hypotheses.

and GPT-4o, it is noticeable that belief-inconsistent syllogisms are more difficult to solve compared to belief-consistent syllogisms. There is a huge drop in accuracy of 19 % for GPT-4 and 20 % for GPT-4o. This so-called content effect was detected in other LLMs as well [21, 52]. In comparison to a Japanese study with young people, they had a similar overall accuracy of 53 %, however, they were better at solving belief-inconsistent and worse at symbolic syllogisms [18]. This might be explained by the different language structure and cultural background. We further analyze the performance of syllogisms that favor the atmosphere effect. The results show that the atmosphere misleads human, GPT-4 and GPT-4o. Especially, GPT-4o and GPT-4 lose around 20 % accuracy. The o4-mini model is not affected by the atmosphere. Overall, we can recognize that the *sd* values of all LLMs as well as humans are rather low in general. This indicates a high consistency in answers. The only outlier with $sd = 4.15$ occurs for GPT-4o when atmosphere is present in syllogisms.

To test our hypotheses (see Section 4), we use the Chi² test for significance with significance level $\alpha = 0.05$. The results of our analysis are presented in Table 7. We could verify all our five hypothesis and found the LLMs perform significantly differently on belief-inconsistent versus belief-consistent syllogisms, therefore we can conclude that the belief bias effects LLMs. This is further supported by the finding that, symbolic syllogisms are significantly easier compared to belief-based syllogisms. Besides, the atmosphere significantly affects the difficulty for LLMs to solve syllogisms. Another finding is that, as already supported by the literature [59, 52], neutral syllogisms are more difficult than entailment or contradiction. We further found that the new reasoning model o4-mini performs better than GPT-4 and GPT-4o. Especially for neutral syllogisms and present atmosphere. Therefore, we should try to understand more about how this new reasoning model works.

7. Comparing Formal Logic, LLM Reasoning and Human Reasoning

Formal logic (see e.g. [64]) provides a foundational, deterministic framework for evaluating syllogisms based on precise logical rules. Derived from classical Aristotelian logic, it enables the systematic assessment of arguments through strict adherence to defined logical structures. Each conclusion drawn from a syllogism is ensured to be valid if the premises are true, offering a clear, explainable pathway of reasoning. In contrast, LLMs such as GPT-4 or GPT-4o generate inferences through statistical association, learned from vast textual datasets rather than formal logic. While mimicking human-like responses to logical prompts, LLMs lack an intrinsic understanding of logical principles. They do not apply the strict rules of formal, mathematical logic on their own initiative. Their outputs are generated based on statistical patterns rather than formal deductive reasoning. As a result, their conclusions often reflect plausible reasoning patterns without ensuring deductive soundness. Human reasoning as well does not rely on formal rules. Instead, it involves envisioning possibilities based on initial perceptions, assertions, memory, or a combination of these [65]. We create mental models of each possibility and draw conclusions from them. This approach predicts systematic errors in reasoning, supported by evidence. Thus, reasoning simulates the world using our knowledge rather than manipulating logical sentence structures by exactly defined inference rules.

A significant limitation of both humans and LLMs in syllogistic reasoning tasks is their susceptibility to cognitive biases. For example, the belief bias, i.e., the tendency to endorse conclusions based on their perceived plausibility rather than logical validity, and the atmosphere effect, wherein the quantifier mood of premises influences the expected structure of the conclusion, have consistently led to systematic reasoning errors [66]. LLMs, due to their training on natural language texts written by humans and statistical nature, are prone to reproducing these same biases [67]. Our findings support this alignment: tasks involving semantically neutral symbolic syllogisms were solved more reliably by both humans and LLMs, suggesting that symbolic abstraction may reduce the influence of belief-driven errors. To address such biases and enhance logical accuracy, recent work has proposed intermediate translation steps that convert natural language syllogisms into predicate logic. By formalizing arguments symbolically, LLMs are nudged toward more structured logical pathways and away from shallow linguistic correlations. A recent study [23] found that enforcing LLMs to first express syllogisms using predicate logic and to explicitly justify their reasoning led to increased logical validity and transparency in model explanations.

According to [65], humans try things, may fail, explore, and eventually settle into a strategy. Different people develop different strategies. As individuals explore problems, they build their strategies using existing inferential tactics, such as adding information to possible scenarios. Once a strategy is developed for a specific problem type, it often guides their reasoning. This strategy has been developed by thinking aloud and using paper and pencil while reasoning. Newer reasoning models enhanced with reasoning-specific adaptations, such as reasoning tokens, incorporate chain-of-thought prompting, forcing step-by-step reasoning, such as human think-aloud, verifier modules akin to metacognitive checking and search/planning augmentations (e.g., tree-of-thoughts, Monte Carlo tree search). This makes them closer to human problem solvers. They simulate strategies, check consistency, and sometimes backtrack. These advanced LLMs, such as OpenAI's o4-mini, demonstrate improved performance on syllogistic reasoning tasks and outperform both older LLM variants and human participants. However, those reasoning models lack the explicit logical reasoning structure that formal methods such as predicate logic can enforce. While formal logical calculi, such as the resolution calculus [64], precisely specify both the permissible inferences and thus what constitutes a reasoning step, there is no well-defined account of what reasoning occurs within LLMs, let alone what would qualify as a reasoning step.

Given the strengths and limitations of each method, a hybrid approach – combining symbolic translation (e.g., into predicate logic) with newer LLM architectures optimized for reasoning tasks – could be part of future research. This may leverage the structural precision of formal logic and the linguistic flexibility of LLMs. Such an integrative design could improve both the reliability and explainability of LLM outputs in deductive tasks.

Finally, while human reasoning is strongly influenced by cognitive heuristics and context-dependent beliefs, LLMs replicate similar patterns through learned data distributions. Importantly, recent generations of LLMs show significantly improved performance on syllogistic reasoning, suggesting the potential for LLMs to surpass average human logical reasoning. Nevertheless, without mechanisms to enforce logical correctness, such outputs remain probabilistic approximations rather than formal deductions. In the end, the question remains whether humans, machines with artificial intelligence, or the validity of formal logic will remain the measure of all things.

8. Conclusion

In this work, we compare humans and LLMs in syllogistic reasoning based on the *NeuBAROCO* dataset. Different focal points are analyzed, as we evaluate diverse syllogism types and content effects, as well as the atmosphere effect. GPT-4, GPT-4o and o4-mini are presented with prompts to classify different syllogisms. Each LLM was tasked to answer 318 syllogisms from the *NeuBAROCO* dataset using OpenAI's API. Further, we conducted a human study in the form of an online questionnaire consisting of eleven syllogisms. The results from 188 participants are used for the analysis. Humans performed overall less accurately than the GPT models in classifying syllogisms. Both humans and models had difficulties with neutral syllogisms. Symbolic syllogisms were the easiest to classify correctly. We also

found that the LLMs and humans are both influenced by the atmosphere effect. The latest reasoning model, o4-mini, showed nearly perfect accuracy and a huge improvement compared to the previous versions GPT-4 and GPT-4o. The results illustrate that the development of LLMs is progressing greatly and can outperform human logical reasoning abilities. At the same time, the results also show that LLMs sometimes exhibit similar behavioural patterns to those of humans, especially when influenced by content or the atmosphere effect.

Our study has certain limitations that future research could explore. One limitation is the different amounts and distributions of syllogisms for humans and LLMs. In addition, the atmosphere effect was only analyzed to a limited extent due to the distribution of the syllogisms in the dataset. In order to ultimately confirm the effect, a more comprehensive assessment should be carried out. The incorporation of more diverse and open-source LLMs and reasoning models could help to improve the study.

This provides several reference points for future research. Subsequent studies could adopt a more homogeneous experimental setup and examine a smaller, more targeted participant group using a larger set of syllogisms to improve statistical power. Including specific populations, such as individuals with autistic traits, would further enable detailed analysis of content bias effects. Further, a combination with formal logic structures could be included in a more extensive study. More detailed analyzes could also help to identify the current status of LLMs and thus evaluate the extent to which logical thinking patterns are similar in humans and LLMs. In summary, our study shows that while LLMs outperform human syllogistic reasoning accuracy, they show human-like biases.

Acknowledgments

This publication was funded by the European Social Fund (ESF+) and the state of Saxony-Anhalt as part of the InterGrad-EGD project [funding reference: ZS/2023/11/181808].

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] OpenAI, J. Achiam, et al., GPT-4 Technical Report, arXiv.org, 2024. doi:10.48550/arXiv.2303.08774.

[2] Z. Aghahadi, A. Talebpour, Language-based syllogistic reasoning using deep neural networks, Cognitive Semantics 8 (2022) 210–239. doi:10.1163/23526416-bja10026.

[3] S. Khemlani, P. N. Johnson-Laird, Theories of the syllogism: A meta-analysis., Psychological Bulletin 138 (2012) 427–457. doi:10.1037/a0026841.

[4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A Survey of Large Language Models, arXiv.org, 2023. doi:10.48550/arXiv.2303.18223.

[5] Y.-A. Lu, H.-Y. Kao, SemEval-2024 Task 5: Enhancing legal argument reasoning with structured prompts, in: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 385–390. doi:10.18653/v1/2024.semeval-1.60.

[6] X. Yang, Z. Wang, Q. Wang, K. Wei, K. Zhang, J. Shi, Large language models for automated q&a involving legal documents: a survey on algorithms, frameworks and applications, International Journal of Web Information Systems 20 (2024) 413–435. doi:https://doi.org/10.1108/IJWIS-12-2023-0256.

[7] T. Savage, A. Nayak, R. Gallo, E. Rangan, J. H. Chen, Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine, NPJ Digital Medicine 7 (2024) 20. URL: https://doi.org/10.1038/s41746-024-01010-1.

[8] A. Y. Tsai, A. Kraft, L. Jin, C. Cai, A. Hosseini, T. Xu, Z. Zhang, L. Hong, E. H. Chi, X. Yi, Leveraging LLM Reasoning Enhances Personalized Recommender Systems, arXiv.org, 2024. doi:10.48550/arXiv.2408.00802.

[9] S. Krause, A. Dalvi, S. K. Zaidi, Generative AI in Education: Student Skills and Lecturer Roles, arXiv.org, 2025. doi:10.48550/arXiv.2504.19673.

[10] W. Hua, K. Zhu, L. Li, L. Fan, S. Lin, M. Jin, H. Xue, Z. Li, J. Wang, Y. Zhang, Disentangling logic: The role of context in large language model reasoning capabilities, arXiv.org, 2024. doi:10.48550/arXiv.2406.02787.

[11] S. Krause, Explainable artificial intelligence and reasoning in the context of large neural network models, in: Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence, Valetta, Malta, 2024. URL: https://ceur-ws.org/Vol-3793/paper_51.pdf.

[12] R. Ando, T. Morishita, H. Abe, K. Mineshima, M. Okada, Evaluating Large Language Models with NeuBAROCO: Syllogistic Reasoning Ability and Human-like Biases, Technical Report, 2023. doi:10.48550/arXiv.2306.12567.

[13] D. Brand, M. Mittenbühler, M. Ragni, Generalizing syllogistic reasoning: Extending syllogisms to general quantifiers, Proceedings of the Annual Meeting of the Cognitive Science Society 44 (2022). URL: https://escholarship.org/uc/item/2sr4v07m.

[14] Y. Wu, M. Han, Y. Zhu, L. Li, X. Zhang, R. Lai, X. Li, Y. Ren, Z. Dou, Z. Cao, Hence, Socrates is mortal: A Benchmark for Natural Language Syllogistic Reasoning, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023, pp. 2347–2367. doi:10.18653/v1/2023.findings-acl.148.

[15] L. S. Moss, Natural Logic, in: S. Lappin, C. Fox (Eds.), The Handbook of Contemporary Semantic Theory, 1 ed., Wiley, 2015, pp. 559–592. doi:10.1002/9781118882139.ch18.

[16] S. Krause, F. Stolzenburg, Commonsense reasoning and explainable artificial intelligence using large language models, in: ECAI 2023 International Workshops, CCIS 1947, Springer Nature Switzerland, 2024, pp. 302–319. doi:10.1007/978-3-031-50396-2_17, part 1.

[17] S. Krause, F. Stolzenburg, From data to commonsense reasoning: the use of large language models for explainable AI, arXiv.org, 2024. URL: https://doi.org/10.48550/arXiv.2407.03778.

[18] C. Shikishima, K. Hiraishi, S. Yamagata, Y. Sugimoto, R. Takemura, K. Ozaki, M. Okada, T. Toda, J. Ando, Is g an entity? A Japanese twin study using syllogisms and intelligence tests, Intelligence 37 (2009) 256–267. doi:10.1016/j.intell.2008.10.010.

[19] I. Copi, C. Cohen, K. McMahon, Introduction to Logic, Routledge, 2016. doi:10.4324/9781315510897.

[20] F. Stolzenburg, R. Lüderitz, Syllogistic reasoning in seven spaces, in: C. Beierle, G. Kern-Isberner, M. Ragni, F. Stolzenburg (Eds.), Proceedings of the KI 2017 Workshop on Formal and Cognitive Reasoning, CEUR Workshop Proceedings 1928, 2017, pp. 77–88. URL: https://ceur-ws.org/Vol-1928/paper7.pdf.

[21] A. K. Lampinen, I. Dasgupta, S. C. Y. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, J. L. McClelland, F. Hill, Language models show human-like content effects on reasoning tasks, arXiv.org, 2024. doi:10.48550/arXiv.2207.07051.

[22] M. Ye, T. Kuribayashi, J. Suzuki, G. Kobayashi, H. Funayama, Assessing Step-by-Step Reasoning against Lexical Negation: A Case Study on Syllogism, arXiv.org, 2023. doi:10.48550/arXiv.2310.14868.

[23] K. Ozeki, R. Ando, T. Morishita, H. Abe, K. Mineshima, M. Okada, Exploring reasoning biases in large language models through syllogism: Insights from the NeuBAROCO dataset (2024). doi:10.48550/arXiv.2408.04403.

[24] OpenAI, Hello GPT-4o, 2024. URL: https://openai.com/index/hello-gpt-4o/, accessed: 2024-07-17.

[25] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, et al., GPT-4o system card, arXiv.org, 2024. doi:10.48550/arXiv.2410.21276.

[26] J. S. B. T. Evans, S. E. Newstead, R. M. Byrne, R. M. J. Byrne, Human reasoning: the psychology of deduction, Lawrence Erlbaum Associates, Inc., 1993. URL: https://psycnet.apa.org/record/

1993-98609-000.
[27] T. Ebert, U. Nortmann, Aristoteles: Analytica Priora. Buch I. übersetzt und erläutert., Aristoteles Werke in deutscher Übersetzung, Akademie Verl, Berlin, 2007. URL: https://philpapers.org/rec/EBEAAP.
[28] K. Stenning, M. v. Lambalgen, Human reasoning and cognitive science, A Bradford book, MIT press, 2008. URL: https://philpapers.org/rec/STEHRA-5.
[29] N. Chater, M. Oaksford, The probability heuristics model of syllogistic reasoning, Cognitive Psychology 38 (1999) 191–258. doi:10.1006/cogp.1998.0696.
[30] P. N. Johnson-Laird, M. Steedman, The psychology of syllogisms, Cognitive psychology 10 (1978) 64–99. URL: http://www.modeltheory.org/papers/1978syllog.pdf.
[31] S. E. Newstead, Interpretational errors in syllogistic reasoning, Journal of Memory and Language 28 (1989) 78–91. URL: https://www.sciencedirect.com/science/article/pii/0749596X89900296. doi:https://doi.org/10.1016/0749-596X(89)90029-6.
[32] J. S. B. T. Evans, Bias in human reasoning: causes and consequences, Essays in cognitive psychology, reprinted in paperback ed., Erlbaum, Hove, 1994. URL: https://psycnet.apa.org/record/1989-98394-000.
[33] J. S. B. T. Evans, J. L. Barston, P. Pollard, On the conflict between logic and belief in syllogistic reasoning, Memory & Cognition 11 (1983) 295–306. doi:10.3758/BF03196976.
[34] D. Trippas, D. Kellen, H. Singmann, G. Pennycook, D. J. Koehler, J. A. Fugelsang, C. Dubé, Characterizing belief bias in syllogistic reasoning: A hierarchical Bayesian meta-analysis of ROC data, Psychonomic Bulletin & Review 25 (2018) 2141–2174. doi:10.3758/s13423-018-1460-7.
[35] P. C. Wason, Reasoning about a rule, Quarterly Journal of Experimental Psychology 20 (1968) 273–281. doi:10.1080/14640746808400161.
[36] R. S. Woodworth, S. B. Sells, An atmosphere effect in formal syllogistic reasoning, Journal of Experimental Psychology 18 (1935) 451–460. doi:10/cn7tqm.
[37] R. Revlis, Two models of syllogistic reasoning: Feature selection and conversion, Journal of Verbal Learning and Verbal Behavior 14 (1975) 180–195. doi:10.1016/S0022-5371(75)80064-8.
[38] I. Begg, J. P. Denny, Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors., Journal of Experimental Psychology 81 (1969) 351–354. doi:10.1037/h0027770.
[39] P. C. Wason, P. N. Johnson-Laird, Psychology of Reasoning: Structure and Content, volume 86, Harvard University Press, 1972.
[40] R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, M. Steinbrecher, Computational Intelligence: A Methodological Introduction, Springer Nature, 2022. doi:10.1007/978-3-030-42227-1.
[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is All you Need, Advances in Neural Information Processing Systems (2017). doi:10.48550/arXiv.1706.03762.
[42] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, A. Das, A comprehensive survey of hallucination mitigation techniques in large language models, Technical Report, 2024. doi:10.48550/arXiv.2401.01313.
[43] T. OpenAI, Neu: OpenAI o3 und o4-mini, OpenAI Blog (2024). URL: https://openai.com/de-DE/index/introducing-o3-and-o4-mini/.
[44] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, arXiv.org, 2025. doi:10.48550/arXiv.2501.12948.
[45] V. Xiang, C. Snell, K. Gandhi, A. Albalak, A. Singh, C. Blagden, D. Phung, R. Rafailov, N. Lile, D. Mahan, et al., Towards system 2 reasoning in LLMs: Learning how to think with meta chain-of-thought, arXiv.org, 2025. doi:10.48550/arXiv.2501.04682.
[46] Z.-Z. Li, D. Zhang, M.-L. Zhang, J. Zhang, Z. Liu, Y. Yao, H. Xu, J. Zheng, P.-J. Wang, X. Chen, et al., From system 1 to system 2: A survey of reasoning large language models, Technical Report, 2025. doi:10.48550/arXiv.2502.17419.
[47] T. OpenAI, Reasoning mit großen Sprachmodellen (LLMs) lernen, OpenAI Blog (2024). URL:

https://openai.com/de-DE/index/learning-to-reason-with-llms/.
[48] H. Luo, Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, D. Zhang, WizardMath: Empowering mathematical reasoning for large language models via reinforced evol-instruct, Technical Report, 2023. doi:10.48550/arXiv.2308.09583.
[49] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al., DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, Technical Report, 2024. doi:10.48550/arXiv.2402.03300.
[50] M. Parmar, N. Patel, N. Varshney, M. Nakamura, M. Luo, S. Mashetty, A. Mitra, C. Baral, Towards Systematic Evaluation of Logical Reasoning Ability of Large Language Models, arXiv.org, 2024. doi:10.48550/arXiv.2404.15522.
[51] J. L. Espejel, E. H. Ettifouri, M. S. Y. Alassan, E. M. Chouham, W. Dahhane, GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts, Natural Language Processing Journal 5 (2023) 100032. doi:10.1016/j.nlp.2023.100032.
[52] L. Bertolazzi, A. Gatt, R. Bernardi, A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences, arxiv.org, 2024. doi:10.48550/arXiv.2406.11341.
[53] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, Y. Zhang, Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4, arXiv.org, 2023. doi:10.48550/arXiv.2304.03439.
[54] J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, Y. Zhang, Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, arXiv.org, 2020. doi:10.48550/arXiv.2007.08124.
[55] W. Yu, Z. Jiang, Y. Dong, J. Feng, Reclor: A reading comprehension dataset requiring logical reasoning, arXiv.org, 2020. doi:10.48550/arXiv.2002.04326.
[56] T. Eisape, M. H. Tessler, I. Dasgupta, F. Sha, S. van Steenkiste, T. Linzen, A Systematic Comparison of Syllogistic Reasoning in Humans and Language Models, arXiv.org, 2024. doi:10.48550/arXiv.2311.00445.
[57] R. Anil, A. M. Dai, O. Firat, et al., Palm 2 technical report, 2023. URL: https://arxiv.org/abs/2305.10403. arXiv:2305.10403.
[58] M. Ragni, H. Dames, D. Brand, N. Riesterer, When does a reasoner respond: Nothing follows?, in: Proceedings of the 41st Annual Conference of the Cognitive Science Society, 2019, pp. 2640–2546. URL: https://nicolasriesterer.net/publications/Ragni2019.pdf.
[59] A. Williams, N. Nangia, S. R. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, Technical Report, 2017. doi:10.48550/arXiv.1704.05426.
[60] J. Ando, C. Shikishima, K. Hiraishi, Y. Sugimoto, R. Takemura, M. Okada, At the crossroads of logic, psychology, and behavioral genetics, in: D. Andler, M. Okada, I. Watanabe (Eds.), Reasoning and Cognition, 2006, pp. 9–36. URL: https://philpapers.org/rec/ANDATC-2.
[61] P. N. Johnson-Laird, Mental models: Towards a cognitive science of language, inference, and consciousness, 6, Harvard University Press, 1983.
[62] Tivian, Tivian Academic Edition – Marktforschungssoftware, 2024. URL: https://www.tivian.com/de/feedback-software/marktforschung-software/academic-edition/, accessed: 2024-07-17.
[63] R. Islam, O. M. Moushi, GPT-4o: The cutting-edge advancement in multimodal llm, Authorea Preprints (2024). URL: 10.36227/techrxiv.171986596.65533294/v1.
[64] S. Russell, P. Norvig, Artificial Intelligence – A Modern Approach, 4th ed., Prentice Hall, Englewood Cliffs, NJ, 2020. URL: https://elibrary.pearson.de/book/99.150005/9781292401171, part III: Knowledge, Reasoning, and Planning.
[65] P. N. Johnson-Laird, Mental models and human reasoning, Proceedings of the National Academy of Sciences 107 (2010) 18243–18250. URL: https://doi.org/10.1073/pnas.1012933107.
[66] J. S. B. T. Evans, J. L. Barston, P. Pollard, On the conflict between logic and belief in syllogistic reasoning, Memory & cognition 11 (1983) 295–306. doi:https://doi.org/10.3758/BF03196976.
[67] A. K. Singh, B. Lamichhane, S. Devkota, U. Dhakal, C. Dhakal, Do large language models show human-like biases? Exploring confidence—competence gap in AI, Information 15 (2024) 92. URL: https://doi.org/10.3390/info15020092.