Categorical Syllogisms Revisited: A Review of the Logical Reasoning Abilities of LLMs for Analyzing Categorical Syllogisms

Shi Zong, Jimmy Lin
David R. Cheriton School of Computer Science
University of Waterloo
{s4zong, jimmylin}@uwaterloo.ca

Abstract

There has been a huge number of benchmarks proposed to evaluate how large language models (LLMs) behave for logic inference tasks. However, it remains an open question how to properly evaluate this ability. In this paper, we provide a systematic overview of prior works on the logical reasoning ability of LLMs for analyzing categorical syllogisms. We first investigate all the possible variations for categorical syllogisms from a purely logical perspective and then examine the underlying configurations (i.e., mood and figure) tested by existing datasets. Our results indicate that compared to template-based synthetic datasets, crowdsourcing approaches normally sacrifice the coverage of configurations (i.e., mood and figure) of categorical syllogisms for more language variations, thus bringing challenges to fully testing LLMs under different situations. We then summarize the findings and observations for the performance of LLMs to infer the validity of syllogisms from the current literature. The error rate breakdown analyses suggest that the interpretation of quantifiers seems to be the current bottleneck that limits the performance of the LLMs and is thus worth more attention. Finally, we discuss several points that might be worth considering when researchers plan to release categorical syllogism datasets. We hope our work will provide a timely review of the current literature regarding categorical syllogisms, and motivate more interdisciplinary research between communities, specifically computational linguists and logicians.

1 Introduction

Large language models (LLMs) have achieved remarkable performance on a variety of tasks (Brown et al., 2020; Wei et al., 2022; Bubeck et al., 2023). Over the years, a large number of benchmarks have been proposed that try to evaluate the different abilities of LLMs, many of which are designed for measuring logical reasoning ability using a variety of tasks. Habernal et al. (2018) propose an argument reading comprehension task to test deductive reasoning. CLUTRR (Sinha et al., 2019) tests inductive reasoning capabilities by requiring to infer kinship relations between characters in short stories. ReClor (Yu et al., 2020), MMLU (Hendrycks et al., 2021), and LogiQA (Liu et al., 2020) contain multiple-choice reading comprehension questions to evaluate diverse forms of logical reasoning. Datasets such as SylloBase (Wu et al., 2023) and FOLIO (Han et al., 2022) require LLMs to conduct inferences using syllogism logic or first-order logic. Among these datasets, many consist of questions that are directly taken from exams. For example, MMLU (Hendrycks et al., 2021) contains practice questions from tests such as the Graduate Record Examination (GRE), and ReClor (Yu et al., 2020) collects problems from the Law School Admission Test (LSAT).

A fundamental question behind these datasets is: *how to design a benchmark to ensure a fair and comprehensive evaluation of logic reasoning abilities?* This question is particularly important when the test questions are self-generated, instead of directly collected from established examinations for humans mentioned above. Problems in those human examinations are developed over decades and are designed in support of theories such as psychometrics and measurement in education. Thus, having rigorous analyses of current benchmarks designed for LLMs would ensure that we can track the development progress of LLMs accurately.

In this work, we make progress in answering the above question for a specific task: categorical syllogisms.[1] Besides the reason that to the best of our knowledge, there is no prior work on analyzing categorical syllogism datasets from a designing principle's perspective, we note some other compelling

[1] Unless specified, the term "categorical syllogisms" is also directly written as "syllogisms" (due to space issues).

Proceedings of the 1st Workshop on NLP for Science (NLP4Science), pages 230–239
November 16, 2024 ©2024 Association for Computational Linguistics

reasons for choosing this task. (1) Syllogisms are inarguably the most basic building block in logical reasoning abilities. Having a deeper understanding of syllogism inference is thus beneficial when designing models for solving more complex reasoning tasks. (2) Categorical syllogisms have a finite number of situations (discussed in Section 2.1), which could enable a complete check of all the possible cases for LLMs. (3) How to properly solve categorical syllogisms has been studied by logicians over decades. There is a huge literature that we can draw inspiration from to help understand how LLMs behave or make LLMs more efficient.

To sum up, our intention is not to propose new models to achieve the start-of-the-performance on certain datasets, nor introduce new benchmarks. Rather, we hope to take a step back and systematically review all existing work to understand where we are right now. Our goal is to check missing pieces and identify areas that are worth clarifying or need future research. Specifically, in this paper, we make the following contributions:

- We investigate all existing categorical syllogism datasets in literature along with their properties in Section 3. A checklist covering all the variations of categorical syllogisms from a purely logician's perspective is provided and we then examine the coverage of different cases for existing benchmarks.
- We summarize all prior findings related to the performance of LLMs for checking the validity of syllogisms in Section 4. By presenting an error rate breakdown by the mood and figure of syllogisms, we highlight the importance of enhancing the abilities of LLMs for interpreting quantifiers.
- We provide suggestions for the future releases of categorical syllogism datasets in Section 5, including clarifying certain issues such as existential import, providing complete annotations, and building datasets containing ordinary arguments.

2 A Concise Introduction to Syllogisms

In this section, we provide a brief introduction to categorical syllogisms from a logician's perspective. We will show in Sections 3.2 and 4.3 that these preparations will help us evaluate current syllogism datasets and better understand the bottleneck of the performance of LLMs.

Major Premise: All Greeks are humans.
Minor Premise: All Athenians are Greeks.
Conclusion: Therefore, all Athenians are humans.

Table 1: An example of a standard-form categorical syllogism (mood AAA, figure 1, configuration AAA-1).

Proposition	Type	Gen. quant.
All S are P.	Universal Affirmative (A)	$S \subseteq P$
No S is P.	Universal Negative (E)	$S \cap P = \varnothing$
Some S is P.	Particular Affirmative (I)	$S \cap P \neq \varnothing$
Some S is not P.	Particular Negative (O)	$S - P \neq \varnothing$

Table 2: Types of propositions with corresponding expressions using generalized quantifier theory.

2.1 Categorical Syllogisms

Categorical Propositions. A categorical proposition relates two classes, or categories. In practice, we care most about a categorical proposition in its standard form, which can be written as: `Quantifier (Subject) Copula (Predicate)`. There are only 4 kinds of standard-form categorical propositions, listed in Table 2.

Terms. A syllogism contains three terms: the predicate term (P), the middle term (M), and the subject term (S). The middle term never occurs in the conclusion but always appears in both premises. The term that occurs as the predicate and the subject of the conclusion is called the major term and minor term, respectively.

Standard-Form Categorical Syllogisms. A categorical syllogism in its standard form must meet the following two requirements: (1) Its premises and conclusion are all standard-form categorical propositions (A, E, I, or O; see Table 2); and (2) Propositions are arranged in standard order (major premise, then minor premise, then conclusion). Table 1 is an example of a standard-form syllogism.

Mood and Figure. The *mood* of a categorical syllogism consists of the letter names of the propositions it contains. For example, the mood for the syllogism presented in Table 1 is AAA. The *figure* of a categorical syllogism is determined by the location of the two occurrences of the middle term in the premises. As shown in Table 3, there are 4 possible figures. To accurately determine the mood and figure of a categorical syllogism, it must be in standard form (defined above). Any standard-form syllogism is completely described when we specify its mood and figure. To simplify the terminology,

Figure	1	2	3	4
Major Premise	**M** - P	P - **M**	**M** - P	P - **M**
Minor Premise	S - **M**	S - **M**	**M** - S	**M** - S
Conclusion	S - P	S - P	S - P	S - P

Table 3: Categorical syllogisms have 4 different figures.

in this paper, we define the combination of mood and figure as the *configuration* of this syllogism.

Valid Inference Types. Since there are 4 kinds of categorical propositions and 3 categorical propositions in a categorical syllogism, there are 64 possible moods ($4^3 = 64$). As each mood can occur in each of the four figures, in total we have $4^4 = 256$ different syllogisms. Among these, only 24 are valid forms, which are extensively studied by logicians.[2] Thus, we have the following fact: *the validity of the standard syllogism can be determined by checking the configuration (mood and figure) against a list of valid syllogistic forms*.

2.2 Analyzing Syllogisms as a Logician

We now briefly go through the steps that logicians take for an ordinary categorical syllogism (Copi et al., 2019; Hurley and Watson, 2018).

Translating Categorical Propositions. In practice, rare propositions are in their standard form and we need to make translations. The major benefit of such translation is that the operations and inferences pertinent to standard-form categorical propositions can be directly applied to these statements. Logicians have developed a number of well-tested methods for translating non-standard propositions, although given the richness of ordinary language, these specific rules can not cover all possible cases.

Determining the Mood and Figure. Once a categorical syllogism is written in its standard form, its figure and mood can be determined by comparing it to Tables 2 and 3. The judgment of a syllogism's configuration is then rather straightforward.

Checking Validity. For a given standard-form categorical syllogism, there are at least the following three ways to check its validity: (1) Use the configuration of this syllogism and then compare it against a list of pre-defined valid syllogistic forms; (2) Use the method of Venn diagrams or generalized quantifier theory to perform set operations; or (3) Check to see if the syllogism conforms to certain rules that are developed by logicians.

Handling Non-Standard Cases. When translating into standard-form syllogisms, some specific cases are worth attention, including the treatment of singular propositions, syllogisms with more than three terms, and enthymemes and sorites. We provide the details of these situations in Appendix A.

3 Review of Existing Syllogism Datasets

3.1 Summary of Syllogism Datasets

We categorize all existing syllogism datasets based on their construction methods, i.e., how the text of premises and conclusions are generated. In real practice, although some datasets are originally proposed for predicate (first-order) logic, their construction methods might involve syllogisms, or a portion of or the whole dataset contains only categorical propositions. As these datasets could be formulated as syllogisms, we also list two representative ones for completeness. All syllogism datasets are summarized in Table 4.[3]

Template-based Approach. Datasets falling into this category are normally generated using templates, i.e., four standard propositions in Table 2. The relation triplets are sampled from different sources and then filled into terms positions of these templates to form the complete syllogisms. For example, questions in ProntoQA (Saparov and He, 2023) use ontology generation and contain a series of premises and thus essentially sorites. Eisape et al. (2024) use a list of 30 relation triplets, the terms of which have no obvious semantic associations. The relation triplets in Wu et al. (2023) are sampled from Wikidata and ConceptNet, and the propositions generated from templates are further rephrased by using GPT-3.

Text Generated by Humans. Non-synthetic datasets are normally developed through crowdsourcing efforts. To acquire high-quality inference questions efficiently, these datasets sometimes rely on guidance during the crowdsourcing tasks. SylloFigure (Peng et al., 2021) is built based on the idea of enthymeme reconstruction. Specifically, Peng et al. (2021) select the entailment part of the SNLI (Bowman et al., 2015) dataset and then add the annotations of figures. Avicenna

[2] 15 configurations are "unconditionally valid" and another 9 are "conditionally valid". It is related to existential import in Section 5.1.

[3] Some prior works use syllogism datasets that are not in the format of natural language, such as Dong et al. (2020). We skip the discussions of these studies.

	Data Generation		Annotation				Performance			Meta	
	Method	Source	Term	Mood	Figure	Validity	Task	Model	Acc.	Total	Access
				Syllogisms Datasets							
SylloFigure (Peng et al., 2021)	Entailment part of SNLI		Middle	🟥	1-4	Entail	Figure identification	BERT	92%	8,635	Yes
Avicenna (Aghahadi and Talebpour, 2022)	Crowdsourcing	Books, articles, etc.	Middle	🟥	🟥	valid, invalid	Conclusion generation	GPT-2 trans. learning	32.0%	6,000	Yes
SylloBASE (Wu et al., 2023)	Template w/ GPT-3 rewrite	Wikidata ConceptNet	▲	▲	▲	valid, invalid	Conclusion selection	RoBERTa	72.8%	51,000	No
Logical (Lampinen et al., 2023)	Human authored questions		▲	🟥	🟥	valid belief-consistent	Conclusion validity identification	PaLM 2-L	∼90% (support)	48	No
NeuBAROCO (Ando et al., 2023)	BAROCO (originally designed for human intell. test)		🟥	▲	🟥	entail, contra, neu inference types	Conclusion validity identification	GPT-3.5	51.7% (overall)	375	No
Reasoning (Eisape et al., 2024)	Template	Hand-crafted triples list	△	△	△	valid, invalid	Conclusion selection	PaLM 2	∼75%	1,920	Yes
				First-order Logic Datasets							
FOLIO (Han et al., 2022)	Template w/ crowd -sourcing rewrite	N/A	▲	▲	▲	true, false, unknown	Conclusion truth identification	Logic-LM (GPT-4)	78.1%	1,435	Yes
ProntoQA (Saparov and He, 2023)	Template	Generated ontology	△	🟥	🟥	true, false	Validity of sorites	GPT-3	∼90%	400	Yes

Table 4: Overview of existing syllogism datasets, along with their construction methods, annotations included, and the documented model performance. △ denotes annotations could be inferred based on the provided dataset construction method, ▲ denotes annotations are generated in the intermediate steps of the dataset construction but are neither released nor inferred, and 🟥 denotes annotations not available or no information.

(Aghahadi and Talebpour, 2022) is a crowdsourcing dataset, and the syllogisms are extracted from various sources, such as books and news articles. Syllogisms in Lampinen et al. (2023) are hand-authored. NeuBAROCO (Ando et al., 2023) originates from BAROCO, which is written in Japanese and is developed to evaluate human syllogistic reasoning abilities. FOLIO (Han et al., 2022) first generates logically valid stories using syllogism templates and then asks human annotators to write logically valid stories in natural language.

Our Newly Collected Test Examples. As shown in Table 4, nearly all datasets with human-generated text lack certain kinds of annotations, thus causing troubles in analyzing them (in Section 3.3). We fill in this missing gap by collecting relevant examples and corresponding exercise questions from standard introduction to logic textbooks (Copi et al., 2019; Kelley, 2013; Baronett, 2018; Hurley and Watson, 2018).

In total, we collect 371 examples of translating statements into standard form, covering all the possible forms of phraseology discussed in Section 3.2; 64 examples for judging the types of standard propositions; and 116 examples for judging the validity of a given syllogism, with complete annotations for the mood and figure. Among these examples, 57 are enthymemes.

3.2 Variations of Categorical Syllogisms

A set of questions that cover all the possible cases could be achieved by varying components of different levels of granularity that we outline in Section 2.1. We consider all possible variations from two angles: syllogisms in standard and non-standard forms. For standard syllogisms, the underlying nature is decided by the combination of mood and figure, which leads to 256 different cases.

For non-standard syllogism, there are variations both on the individual proposition level and the syllogism level. On the proposition level, we consider the different options of quantifiers, terms, and copula: (1) Besides standard quantifiers, the propositions could have non-standard quantifiers (also known as generalized quantifiers), such as “few”, “a few”, “not every”, or “anyone”, and unexpressed quantifiers; (2) Terms could be expressed with only an adjective, a plural noun or a pronoun, and the verbs are in other forms of the verb “to be;” and (3) Certain propositions could be typically translated into categorical propositions. Established categories include singular propositions, conditional statements such as “if ... then,” exclusive propositions that involve words “only,” “none but,” and “none except,” and exceptive propositions in the form of “All except S are P” and “All but S are P”.

On the syllogism level for non-standard syllogism, we vary the following (details in Appendix A): (1) It is possible that the syllogism covers more than three terms; and (2) Besides the normal syllogisms with two premises and one conclusion, there exist situations with more than two premises or missing premises, which we refer to as enthymemes and sorites.

On top of all the options above, instead of putting the propositions in a well-structured format (i.e., explicitly listing them as premises and conclusions), we could mix them all together as ordinary arguments. Some other parts could be varied, such as the order of the two premises. Since the change of the ordering does not change the validity of the conclusion, we skip the discussion of this part.

3.3 Coverage of Current Datasets

In Section 3.2, we have enumerated all the possible cases of categorical syllogisms. In this section, we will use this checklist to evaluate the coverage of current syllogism datasets. We mainly consider the following aspects: (1) the forms of phraseology covered, and (2) the mood and figure covered in these syllogism datasets.

3.3.1 Building Tools for Assessing Coverage

Most of the datasets do not have the annotations needing to be assessed (details in Table 4). Making up these missing pieces would require human annotators with linguistic background. Given the huge amount of human effort for such annotations, we take the approach of directly asking LLMs to perform as an annotator for labeling.

To ensure that we can build prompts with reasonable performance, we calibrate them on our newly collected textbook questions (discussed in Section 3.1). We also use the fact about the validity of syllogisms mentioned in Section 2.1 for cross-checking: for a valid inference, if a predicted configuration is not one of the valid syllogism forms, then there is something wrong with this prediction.

Translating the Propositions. When translating statements, besides a deep understanding of the given statement, we need to follow some established rules set by logicians (for example, the treatment of singular propositions discussed in Appendix A). We thus base our prompt design on a 2-step translation process: (1) determine the nature of a proposition by classifying it into categories listed in Table 5), and (2) then perform the translation based on the set rules within that category. To make sure the translated proposition is in the standard form, we also set up a mechanism for a second-round translation. We observe that GPT-4o performs well in identifying the forms of phraseology, while it is easy to incorrectly classify some statements into singular propositions. A manual check for the translated propositions shows that GPT-4 achieves 87.3% accuracy on 371 textbook problems, with 68 propositions translated twice.

		SylloFigure	Avicenna	Reasoning
Proposition	Standard (%)	0.9	0.6	100
	Singular (%)	64.7	27.2	0
	Condition (%)	2.3	9.5	0
	Exclusive (%)	0.1	1.0	0
	Others (%)	32.0	61.7	0
	Total	2,448	1,864	2,560
Configuration	Coverage (%)	>4.3	>2.7	100
	Actual count	>11	>7	256
	Syllo assessed (%)	71.1	60.9	100
Total syllogisms		868	622	2,560

Table 5: Forms of phraseology and configurations of categorical syllogisms covered in datasets.

Judging the Mood and Figure. We can not first translate individual propositions and then simply combine the detected proposition types together to form the mood of the syllogism, due to the issue of having potentially more than three terms (in Appendix A). Thus, we feed the syllogism as a whole and ask GPT-4 to generate the mood and figure simultaneously. The principles and rules discussed above for translating propositions are also incorporated into the prompt. Experimental results on 116 textbook examples reveal an accuracy of 87.9% for mood detection, 48.3% for figure detection, and 44.8% for configuration detection. A further review of mood detection results reveals that this high accuracy is due to the fact that most of the collected textbook examples are standard-form propositions.

3.3.2 Datasets Coverage Observations

We apply our calculating tools developed in Section 3.3.1 to all three categorical syllogism datasets currently released. For the SylloFigure and Avicenna datasets, we conduct analyses only on the test sets, while for the Reasoning dataset, we randomly sample 10 relation triples out of 30 and then generate the complete syllogisms. We use the whole dataset for assessing the proposition forms, since it is rather straightforward. Regarding the underlying configuration of syllogisms: As the Reasoning dataset is generated by using templates, the whole dataset could be accurately assessed (since we have all the annotations such as mood and figure). Using the cross-checking method discussed in Section 3.3.1, we estimate 60.9% of syllogisms could be properly assessed in the Avicenna dataset, while a higher 71.1% for the SylloFigure dataset, as it contains human annotated figures.

Our assessment results are reported in Table 5 (detailed configurations are in Figure 1). In Figure 2 we also provide the distribution of the estimated proposition types in the SylloFigure dataset.

We observe that both the proposition types (A, E, I, O) and forms of phraseology (types in Table 5) are distributed highly unevenly, and datasets normally have different distributions. Regarding the coverage of configurations, we observe that compared to template-based datasets, datasets using human-generated text are normally centered on a few specific moods and figures, i.e., Avicenna covers only over 7 different syllogisms configurations, calculated from 60.9% of the whole dataset.

Since we use LLMs instead of human effort to make up the missing mood and figures, the coverage percentages in Table 5 can only be treated as rough estimates. Nevertheless, our key point is clear: datasets that are from crowdsourcing efforts are skewed to certain linguistic styles and cover only limited configurations of syllogisms. We thus suggest researchers take the actual variations covered by the datasets into account when interpreting experimental results.

4 Evaluating LLMs for Analyzing Syllogisms

4.1 What Do We Know So Far?

Reported Results for Validity Inferences. We observe prior studies mainly make use of the following approaches to evaluate the validity of categorical syllogisms: (1) given two premises, select a correct conclusion from multiple choices (Wu et al., 2023; Eisape et al., 2024), (2) given two premises and a conclusion, identify if the logic inference is valid (Lampinen et al., 2023; Ando et al., 2023), and (3) given two premises or more, generate the conclusion (Aghahadi and Talebpour, 2022; Saparov and He, 2023). In general, most prior works report LLMs have an accuracy of around 75% when evaluating the validity of given syllogisms. We provide more performance evaluation details in Table 4.

Error Analysis. One trend for analyzing the errors that LLMs make is to compare them with human cognition biases. Lampinen et al. (2023) find that like humans, LLMs give out more accurate answers when the semantic content of a task supports the logical inferences. Ando et al. (2023) analyze the models' errors from three aspects: belief biases, conversion errors, and atmosphere effects. Eisape et al. (2024) provide more direct observations that LLMs replicate some human biases discovered in psychology studies, while LLMs could overcome these biases in certain situations.

Dataset	#	GPT-4	GPT-4o
SylloFigure	868	74.3	70.2
Avicenna	622	72.5	53.4
Reasoning	2,560	90.2	95.4

Table 6: Accuracy (%) for checking the validity of categorical syllogisms.

4.2 LLMs' Performance Breakdowns by Syllogisms Configurations

In this section, we reproduce the experimental results of LLMs for judging the logical validity of syllogisms and check to see if prior findings still hold. We will also break down the error rate by the configurations of syllogisms.

4.2.1 Setups

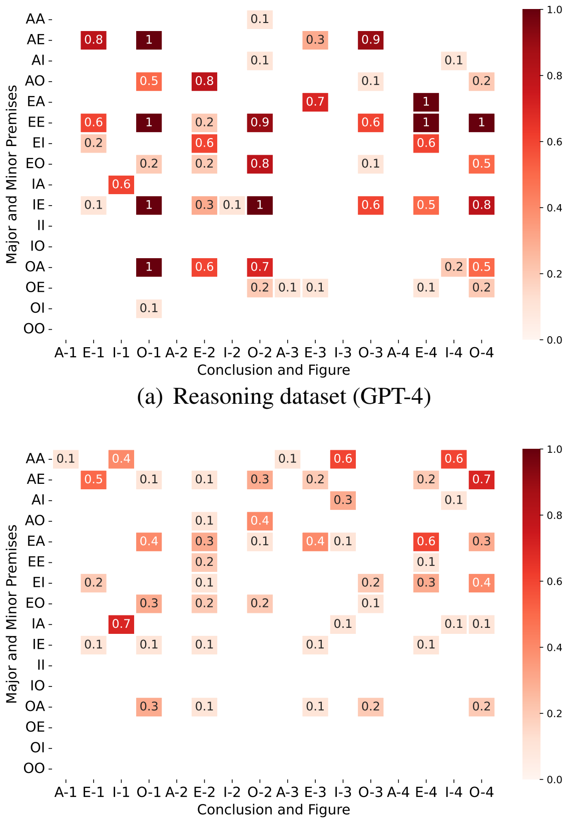
Models and Datasets. We conduct our experiments using OpenAI's GPT models, as they are commonly used large language models with compelling performance on a variety of inference tasks (OpenAI et al., 2024). All our experiments are done using GPT-4 and GPT-4o. We use the same set of datasets that we assess in Section 3.3. The details of these datasets are provided in Section 3.3.2.

Prompts Used. For comparison purposes, we follow the chain-of-thought prompt used in Eisape et al. (2024) and test how LLMs perform logical inferences under a zero-shot learning setting.

4.2.2 Results

We visualize the error rate of GPT-4 and GPT-4o on the complete 256 configurations of syllogisms from the Reasoning dataset in Figures 1(a) and 1(b). The error rate in the SylloFigure and Avicenna datasets are reported in Figure 1(c). We also report the total accuracy of validity judgment in Table 6 for reference purposes.

We observe the following trends. (1) Comparing Figure 1(a) with Figure 1(b), we observe different patterns for the configurations of syllogism that LLMs fail. For example, GPT-4 nearly has no errors when two premises are in AA format, while GPT-4o makes even more than half of the mistakes for AAI-3 and AAI-4. However, GPT-4o performs better than GPT-4 for configurations that GPT-4 has 0% accuracy. (2) For two datasets with human-written text, GPT-4 seems to have more stable performance compared to GPT-4o, i.e., the error rate in Figure 1(c) is roughly the same for AAA-1, AAI-1, and AII-1. (3) We observe that for the same configuration, LLMs generally have a higher error



(a) Reasoning dataset (GPT-4)

(b) Reasoning dataset (GPT-4o)

Figure	SylloFigure				Avicenna			
	Mood	#	GPT-4	GPT-4o	Mood	#	GPT-4	GPT-4o
1	AAA	47	0.21	0.28	AAA	310	0.20	0.42
	AAI	38	0.32	0.42	AAI	12	0.33	0.42
	AII	502	0.21	0.26	AII	25	0.28	0.68
	N/A	56	0.34	0.32	EAE	2	1	0.50
2	EAE	1	0	0	EAE	3	0	0
	N/A	180	0.28	0.36	AEE	3	0.67	0.33
3	AAI	2	1	0.5	AAI	1	0	0
	AII	26	0.54	0.38	AII	4	0	0
	N/A	8	0.38	0.38	IAI	2	0.50	0.50
4	IAI	1	1	0	IAI	14	0.29	0.64
	N/A	7	0.71	0.43	AAI	3	0	0.33
N/A		/			N/A	243	0.35	0.52

(c) SylloFigure and Avicenna datasets

Figure 1: Error rate (↓) of GPT-4 and GPT-4o using zero-shot chain-of-thoughts. (a) and (b): Breakdowns on all 256 configurations of categorical syllogisms in the Reasoning dataset, calculated over 10 different combinations. A white block indicates an error rate of 0 (thus 100% accuracy) in that specific configuration. (c): Breakdowns by configurations in the SylloFigure and Avicenna datasets. We mark the predicted configuration as "N/A" if it does not pass the cross-check discussed in Section 3.3.1.

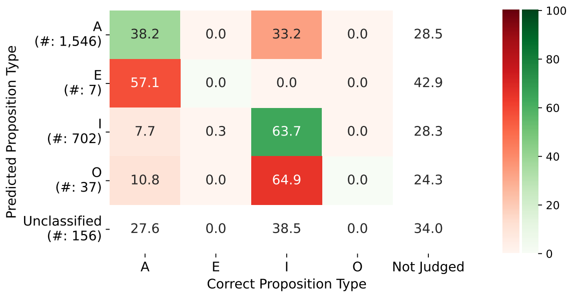


Figure 2: Percentage breakdowns of the correct propositions within each predicted proposition type (by GPT-4). 156 propositions (last row) could not be classified and we can not automatically verify the correctness of predictions without human efforts (last column).

rate in human-generated SylloFigure and Avicenna datasets (Figure 1(c)), compared to the template-based Reasoning dataset (Figures 1(a) and 1(b)). It seems to suggest that translating the syllogisms to the standard form is the bottleneck for LLMs to behave well, as the only difference that the Reasoning dataset has is the expressed way of the premises and conclusions. The underlying ability required to infer remains unchanged: if LLMs can translate ordinary text into the standard format, then it should work well. This observation also aligns with the challenges of the logicians' approach for analyzing syllogisms: as discussed in Section 2.1, the most difficult part is translating the propositions – once the mood and the figure are determined, then checking the validity of the syllogism is trivial.

4.3 Ambiguity of Natural Language

Our observation is that translating into standard propositions is the most challenging part for LLMs and thus causes errors. In this section, we take a closer look at the types of translation errors LLMs make, especially around quantifiers.

We visualize in Figure 2 the percentage of correct propositions within each predicted proposition type in the SylloFigure dataset. We observe that in general, the LLMs have a much higher accuracy in recognizing the "some" quantifier, although it sometimes mixes the particular negative type (O) with the particular affirmative type (I). We also observe LLMs tend to confuse universal affirmative (A) with particular affirmative (I): among 1,546 propositions that are predicted as universal affirmative type (A), 33.2% should be particular affirmative (I). This phenomenon is related to the interpretation of singular propositions (discussed in Appendix A) and is also partially due to the fact that singular propositions represent a huge portion of the SylloFigure dataset (shown in Table 5).

We shall point out that analyzing the sensitivity of quantifiers by LLMs is not entirely new in computation linguistics. One representative work is Cui et al. (2022), where the authors rely on generalized quantifier theory to quantify their contribution to the errors of NLU models. There is a recent work by Madusanka et al. (2023) that tries to investigate how different generalized quantifiers affect LLMs by employing a textual entailment problem. Ando et al. (2023) also suggest the importance of differentiating the problems of interpreting quantifiers and negations from performing logical inferences. In this work, we hold the same standpoints that the comprehension of quantifiers greatly affects the model performance and future models should enhance their abilities to analyze quantifiers. Compared to these prior studies, we present a more complete and comprehensive analysis of quantifiers in a specific syllogism setting.

5 Moving Forward: Future Directions

5.1 Suggestions for Future Datasets

Existential Import. In Section 2.1, we mention that there are 24 valid configurations over all 256 cases, 9 of which rely on the existential import assumption. We notice that nearly all prior works, except Ando et al. (2023), implicitly make such an assumption. We recommend researchers explicitly mention this assumption in their dataset release, as it affects the determination of the validity of syllogisms (Hurley and Watson, 2018).

Complete Annotations. As shown in Table 4, many syllogism datasets lack certain kinds of annotations, thus causing trouble when we try to assess the coverage of language variations in Section 3.3. We notice that in their dataset descriptions, especially datasets that make use of templates, many annotations are actually generated during the dataset construction process (for example, blocks marked with ▲ in Table 4). We suggest researchers consider releasing these annotations from intermediate steps to promote a more accurate assessment of the properties of their datasets.

Ordinary Argument. We observe that all syllogism datasets in Section 3.3 are in a well-structured format, i.e., the premises and conclusions are listed separately. In real life, however, a more realistic situation is that the premises and conclusions are mixed together, with no clear indications or separators. There might even be cases such as enthymemes. Thus, one possible direction is to build datasets that contain ordinary arguments. Building such a dataset will also enable a variety of downstream applications, for example, to evaluate the syllogisms hidden in human forecasts or debates. We note there has been some exploration work in this direction (Jiang and Yang, 2023).

5.2 Enhancing Logical Reasoning Abilities

In prior studies, we observe two lines of research that attempt to enhance the logical reasoning abilities of the LLMs. One line of approach is to rely on external modules. Olausson et al. (2023) make use of an external theorem prover, which symbolically performs deductive inference. Poesia et al. (2023) propose to augment the LLM's reasoning ability by using externally certified reasoning, such as a theorem-proving environment for incremental proof generation. Another line is to directly incorporate the reasoning ability inside the LLMs. Representative work includes Xu et al. (2024), which argues that the reasoning ability should be inherited without using any external blocks. In general, it is unclear which type of approach is better. Specific to our syllogism inference case, if our ultimate goal is to build a trustworthy and reliable system with no tolerance for errors, then enabling some external pure logical solvers would help ensure the accuracy of analyzing syllogisms.

6 Conclusion

This work tries to address the question of whether current proposed benchmarks can evaluate logical reasoning abilities accurately and thoroughly. We choose categorical syllogism as our main focus, since this logical system has been extensively studied by logicians and has many nice properties, such as a finite number of possible cases, and automated ways of solving it. A categorical syllogism is also arguably the most basic building block for any other more complex reasonings. We draw the inspirations from how logicians analyze categorical syllogisms and construct a list of variations that should be covered by benchmarks. Our results show that there is no single dataset that properly covers all possible situations. We also summarize the current progress made in judging the validity of the categorical syllogisms. Our findings highlight the importance of correctly interpreting different quantifiers. Finally, we provide a discussion of several points that might be worth considering when researchers plan on the future release of categorical syllogism datasets.

Limitations

In this work, we mainly focus on analyzing the existing benchmarks of categorical syllogisms. Among 6 syllogism datasets listed in Table 4, we are only able to assess 3, as others are not publicly released. Also, we use GPT-4 as an annotation tool instead of human annotators to generate the missing annotations, such as mood, figure, and forms of phraseology. Although we have taken steps to control the quality of these annotations (as discussed in Section 3.3.1), it is inevitable that there are errors.

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.

Stan Baronett. 2018. *Logic*. Pearson Education.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv:2303.12712*.

Irving M. Copi, Carl Cohen, and Kenneth D. McMahon. 2019. *Introduction to Logic*. Pearson Education.

Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. 2022. Generalized quantifiers as a source of error in multilingual NLU benchmarks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4875–4893, Seattle, United States.

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

Tiwalayo Eisape, MH Tessler, Ishita Dasgupta, Fei Sha, Sjoerd van Steenkiste, and Tal Linzen. 2024. A systematic comparison of syllogistic reasoning in humans and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. FOLIO: Natural language reasoning with first-order logic. *arXiv:2209.00840*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Patrick J. Hurley and Lori Watson. 2018. *A Concise Introduction to Logic*. Cengage Learning.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, page 417–421, New York, NY, USA.

David Kelley. 2013. *The Art of Reasoning: An Introduction to Logic and Critical Thinking, Fourth Edition*. W. W. Norton, Incorporated.

Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2023. Language models show human-like content effects on reasoning tasks. *arXiv: 2207.07051*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628.

Tharindu Madusanka, Iqra Zahid, Hao Li, Ian Pratt-Hartmann, and Riza Batista-Navarro. 2023. Not all quantifiers are equal: Probing transformer-based language models' understanding of generalised quantifiers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8680–8692, Singapore.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. 2024. GPT-4 technical report. *arXiv: 2303.08774*.

Shiya Peng, Lu Liu, Chang Liu, and Dong Yu. 2021. Exploring reasoning schemes: A dataset for syllogism figure identification. In *Chinese Lexical Semantics*, pages 445–451, Cham.

Gabriel Poesia, Kanishk Gandhi, Eric Zelikman, and Noah D. Goodman. 2023. Certified deductive reasoning with language models. *arXiv: 2306.04031*.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China.

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

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, Toronto, Canada.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, Bangkok, Thailand.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*.

A Handling Special Cases When Analyzing Categorical Syllogisms

Singular Propositions. A singular proposition is defined as making a particular individual or object (for example, a specific person, thing, time, or place) belong to a given class. Although it is arguable about the treatment of these singular propositions, logicians seem to agree that in general, these propositions are generally converted into universal propositions.

Reducing the Number of Terms. A valid syllogism must have exactly three terms. When more than three terms seem to be involved in an argument of apparently syllogistic form, we may need to reduce the number of terms to three, by either eliminating synonyms or eliminating class components (Copi et al., 2019).

Enthymemes and Sorites. In real life, we normally do not make explicit mention of all the premises required to support a given conclusion, especially when the premises are obvious or noncontroversial. A syllogism with an unstated premise is called an enthymeme (Kelley, 2013). Sorites are defined as a chain of categorical syllogisms in which the intermediate conclusions have been left out (Hurley and Watson, 2018). The standard treatment for analyzing sorites is to first make their intermediate conclusions or steps explicit, then test the validity of obtained syllogisms separately.