Evaluating Logical Reasoning Ability of Large Language Models

Anonymous submission

Abstract

It is observed that large language models exhibit reasoning abilities when they are sufficiently large. A key open question is whether these models reason or simply recite memorized texts encountered during their training. We propose a new dataset comprising of unseen reasoning questions requiring semantic and deductive logical reasoning skills to evaluate the reasoning ability of large language models. The proposed evaluation framework has several desirable properties, including resilience to training data contamination, ease of answer verification, extensibility and automated test case generation.

1 Introduction

It has been shown that large language models (LLMs) exhibit emergent abilities, including the ability to reason, when their number of parameters is sufficiently large (Wei and Tay 2022). Nonetheless, despite their strong performance in certain reasoning tasks, it remains unclear whether LLMs are indeed reasoning when providing a correct answer to a given question, and to what extent they are capable of reasoning. For example, LLMs perform badly on common planning tasks (Valmeekam and Olmo 2022). The failure to transfer the reasoning ability from one task to others casts doubt on whether LLMs can reason. Hence, a question that often crops up is whether LLMs truly reason or merely recite texts they have encountered during their training routine, without the need for reasoning (Bender and Koller 2020).

Devising tests to evaluate the reasoning ability of LLMs is an active research area. Various datasets and benchmarks have been created to test the arithmetic, common sense, symbolic and logical reasoning of LLMs. However, these tests do not help answer the question on whether the LLMs are reasoning or parroting texts. Many of the questions in these tests are taken from existing public examinations and thus susceptible to training data contamination (Narayan and Kapoor 2023). Other tasks are relatively simple, and they cannot truly evaluate the cognitive limit or reasoning ability of LLMs. The key challenge to devising tests is to ensure that the questions have not appeared on the internet before; otherwise, the questions and their solutions could have been part of the corpus of the training texts previously used to train LLMs, and LLMs may answer them correctly without utilizing their reasoning ability.

To fill this gap in evaluating the reasoning ability of LLMs, we devise a novel dataset of deductive, logical reasoning puzzles. Each question in the dataset presents a room assignment problem to LLMs requiring them to find a unique room assignment subject to the given constraints. To ensure that the questions are unseen to LLMs, the constraints are encoded in symbols with novel, proprietary definitions. Verifying a correct room assignment would require LLMs to reason on situation awareness and apply deductive logic. Compared to the existing work on deductive logical reasoning (Madusanka et al. 2023) and spatial understanding (Rozanova et al. 2021) of LLMs, the proposed dataset consists of reasoning tasks about spatial relationships requiring multi-step inference, and the complexity of the tasks is adjustable while the same logical structure is required to solve them. This enables the study of LLM reasoning via the lens of problem complexity. Experiments are run using the dataset to evaluate the logical reasoning ability of GPT-3, GPT-4, Gemini-1.5, Llama-3.1, Claude-2.5 and DeepSeek-V3. Based on the experimental results, the proposed dataset consists of questions of sufficient difficulty to evaluate the cognitive limit of state-of-the-art LLMs. Besides, it is noted that LLMs reason more like algorithms than human intelligence.

2 A New Logical Reasoning Dataset

A new dataset of questions is crafted in a way to minimize the possibility that LLMs may have encountered these questions in its training corpus, with the aim to minimize LLMs' tendency to recite memorized answers. The proposed dataset consists of questions at 3 levels of difficulty. Each test case in the dataset requires LLMs to assign a given number of rooms to a group of persons and their pets subject to certain restrictions or constraints.

Figure 1 depicts a sample question in the dataset. The question is equivalent to finding a total order of a set of objects subject to constraints. At first, LLMs are given a common definition for use in all questions, and this definition explains the meaning of the symbolic expressions of the constraints or rules that a correct room assignment needs to satisfy. In addition, a set of common instructions are given before the questions are asked. This sets the domain knowledge required. The questions are designed to minimize the need to use external domain knowledge or common sense. On one hand, this would minimize the possibility that LLMs need to look up other memorized texts and link the problem to irrelevant texts, causing them to parrot non-sensible answers. On the other hand, minimizing the domain knowledge can help focus on testing the logical reasoning ability of LLMs.

The constraints of the questions are encoded in symbolic expressions (intentionally with non-intuitive symbols), rather than in words, to serve two purposes. First, this arrangement forces LLMs to decode the expressions to understand the constraints, which would require the use of semantic reasoning ability and therefore test this ability of LLMs. Second, since the constraints in the textual form are not directly obtainable from the question itself, LLMs are less ready to recognize that it has seen the question before if, unfortunately, the question happens to be included in the LLMs' training corpus. This indirect presentation of the constraints helps render the question as an unseen question from the perspective of LLMs and force them to apply their reasoning skill to solve it instead of reciting a memorized answer.

Once LLMs are able to decode the symbolic constraint rules, they would need to reason about situation awareness to find the right room assignment through an educated guess. In addition, they also need to apply deductive logical reasoning to validate whether the room assignment they find is a valid assignment that meets all the given constraints.

Definition:
Given X, Y, Z are persons, define the following symbolic relationships:

1. **X: # o o** means that there are 3 rooms, and Person X is assigned to the room marked by #. For this example, X is assigned to the 1st room of the 3 rooms. Similarly, Y: o # o means that Person Y is assigned to the 2nd room.
2. **X: x o o** means that there are 3 rooms, and Person X can't be assigned to the room marked by x. For this example, X cannot be assigned to the 1st room of the 3 rooms. Similarly, Y: o x o means that Person Y can't be assigned to the 2nd room.
3. **X = Y** means Person X's room is next to Person Y's room regardless of the order. That is, X's room can be before after Y's room.
4. **X <> Y** means Person X's room is before Person Y's room. That is, X's room number is smaller than Y's. They need not be adjacent.
5. **X @ Y** means that X and Y are assigned to the same room.
6. **!** is the negation of the subsequent relationship operator. For example, X != Y means that Person X's room is not next to Person Y's room. Similarly, X !@ Y means that X, Y cannot be assigned to the same room.

Instructions:
In the following questions, you are asked to assign rooms to the persons and their pets according to the given restrictions. You should fill in as many rooms as possible, and all persons and their pets should be assigned rooms. No pet can stay alone in any room. Each room can accommodate at most one person and one pet.

Question:
There are 4 rooms, which are assigned to Alice, Bob, Ken and Charlie with a dog, parrot, cat and turtle. You should fill in as many rooms as possible. Here are the restrictions for the room assignment:

1. Ken: o o o #
2. turtle: # o o o
3. cat @ Bob
4. dog: o o o #
5. cat <> Charlie

What are the tenants of the rooms?

Answer:
Room 1: Alice, turtle
Room 2: Bob, cat
Room 3: Charlie, parrot
Room 4: Ken, dog

Figure 1: A Sample Test Case to Test Logical Reasoning Ability of LLMs.

Each test case in the proposed dataset evaluates the semantic reasoning and deductive reasoning of LLMs, as well as their reasoning about situation awareness. Decoding the symbolic rules for room assignment constraints requires LLMs to apply their semantic reasoning ability, whereas, finding and validating a correct room assignment requires them to carry out deductive reasoning. As can be seen, the occurrence of the same or similar combination of words on the internet is less probable, meaning LLMs should not be able to answer a large portion of the questions correctly merely through picking the most plausible word based on the words of the question. In other words, without using some form of reasoning, getting correct answers for these questions is unlikely.

The proposed dataset has several desirable properties, including resilience to training data contamination, ease of response verification and extensibility.

2.1 Resilience to Data Contamination

Through symbolically expressing the room assignment constraints, it is less likely that LLMs can find a piece of memorized text in its training corpus that would resemble a question from the dataset. This would prevent LLMs from parroting memorized texts to get a correct answer. Besides, no explanation texts for the questions are available on the internet. If LLMs give a correct explanation, it is most likely generated by it. Besides, to strengthen the resilience to training data contamination, it is easy to use a different set of symbols to express the room assignment constraints. Even though the current dataset is exposed, the symbolic expressions of its questions could possibly be modified.

2.2 Ease of verification

Each test case has a short answer whose correctness can be easily verified in a mechanistic manner even done manually. This would enable automated verification. Besides, the explanation provided by LLMs could be easily verified based on the room assignment given in the test case.

2.3 Extensibility

The proposed dataset is extensible in the sense that new questions of varying difficulty can be easily added to the dataset. New questions can be easily created by adjusting the number of rooms or selecting the types of room assignment constraints. Currently, once a new dataset or benchmark is published on the internet, LLMs would quickly attain a perfect score (Sawada and Paleka 2023), likely attributed to LLMs' reciting the answers. The proposed dataset in this work is designed to enable the complexity of the questions to scale up with ease to address this issue.

3 Experiment Settings

We use GPT-3 as the baseline model, and compare the performance of GPT-4, Gemini-1.5, Llama-3.1, Claude-3.5 and DeepSeek-V3 on the proposed dataset. An instruction-prompt scheme is adopted for feeding the questions to the LLMs. A zero-shot and a one-shot prompt are used in the experiment. The answer to each question and the respective explanation by LLMs is verified manually for correctness. Based on the answer and the restriction rules in the question, it is easy to identify any reasoning flaw made by LLMs although done manually. Two scores are tallied, one for correct answers and the other for perfectly correct explanation. An explanation is not tallied if any flaw is found in it. For each LLM, three trials are conducted, and memory is cleared between trials. The correct attempts in the three trails are tallied. Similar calculation is performed to obtain the accuracy of LLMs' explanations. To establish a human intelligence baseline for comparison, we randomly sample questions that LLMs perform badly for human participants to attempt to see whether human participants can solve the problems that LLMs find particularly difficult.

4 Results

4.1 LLM Performance Comparison

Depicted in Figure 2 is a comparison of the performance of different LLMs on the proposed dataset. Except DeepSeek-V3, the accuracy of getting a correct answer is observably higher than the accuracy of getting the respective explanation correct for all the other 5 tested LLMs. This gap is higher for an LLM which scores comparatively poorer in terms of answer accuracy. Compared to the other 5 tested LLMs, DeepSeek-V3 behaves obviously differently. It attains a higher explanation accuracy than its answer accuracy. This performance is largely because DeepSeek-V3 reasons correctly in its steps but puts down an

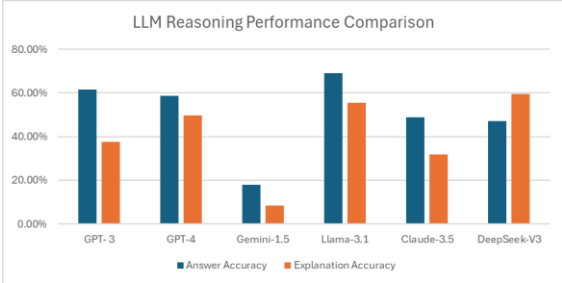


Figure 2: Comparison of the overall performance of LLMs on the proposed logical reasoning dataset.

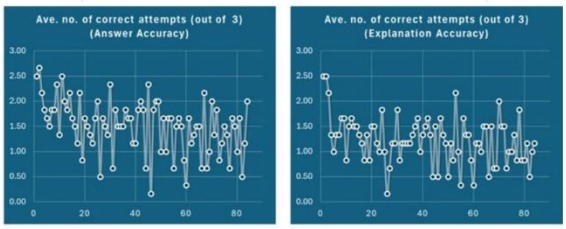


Figure 3: LLM accuracy for different question complexities.

	LLM average	Human average
Low	37.5%	100%
Medium	43.1%	100%
High	13.9%	100%

Table 1: Comparison of answer accuracy of LLMs and human participants.

incomplete answer in the end for some of the questions. This behaviour resembles human intelligence in making careless mistakes although the cause of this behaviour is unknown.

Llama-3.1 and DeepSeek-V3 have the best performance in terms of answer accuracy and explanation accuracy respectively, whereas Gemini-1.5 has the worst performance, performing much poorly compared to the other 5 LLMs. One of the main reasons why Gemini-1.5 performs very poorly, compared to other LLMs, is that it is unable to understand the symbolic rule definition for adjacent room assignment, using the equal sign.

While GPT-4 is frequently highlighted as advanced at reasoning tasks compared to GPT-3, GPT-4's performance in answer accuracy is lower than that of GPT-3, though by just a small fraction. Yet, GPT-4's explanation accuracy is higher than GPT-3's. That could possibly evince GPT-4's higher tendency to reason, compared to GPT-3, on the questions.

We repeated the experiments with a one-shot prompt, and the results are similar (within a deviation of 7% as compared to Figure 2).

4.2 Comparison with Human Performance

Questions that LLMs performed poorly are sampled for 20 human participants to solve. The performance of LLMs and human participants is compared in Table 1. Human participants have no difficulty to solve questions which are challenging for LLMs to solve. This suggests that LLMs' logical reasoning ability, at least on the skill required to solve the proposed dataset, does not match that of human intelligence.

4.3 Salient LLM Reasoning Characteristic

Figure 3 depicts the answer and explanation average accuracy across the tested LLMs. The performance of the tested LLMs do not seem to be dependent on the complexity or difficulty of the questions. In the proposed dataset, the questions are arranged in an order of increasing complexity or difficulty. Later questions are more difficult. Although difference in accuracy can be observed for the three levels of question difficulty, no noticeable trend in performance degradation can be identified for the tested LLMs. We could infer that LLMs use a cognitive structure which is very different from human intelligence to carry out their reasoning. In particular, the observed behaviour suggests that these LLMs reason more like an algorithm rather than human intelligence.

5 Conclusion

We design a novel dataset of questions requiring semantic and deductive logical reasoning to solve. The complexity of the questions are adjustable while requiring the same logical structure to solve them. This dataset exhibits several desirable properties, including resilience to training data contamination, ease of response verification, extensibility and automated test case generation. We use this new dataset to evaluate the reasoning ability different LLMs. While these LLMs exhibit a certain level of reasoning ability, their performance does not seem to match that of human intelligence on the presented dataset. One interesting observation is that LLMs do not reason in the same way as human beings do. While human accuracy would deteriorate with increasing difficulty of the questions, this trend is not observable for all the tested LLMs, except DeepSeek-V3, suggesting that LLMs have a different reasoning structure than human intelligence.

6 Limitations

While this work presents selected questions of the proposed dataset to human participants to compare LLMs' performance with this human baseline, the number of questions only represent a relatively small fraction of the dataset. Besides, the human performance may be skewed due to the sampling method as only questions that LLMs perform poorly are sampled for human participants to solve. While human intelligence may find it easy to solve questions that LLMs find difficult, it remains unclear whether the questions that are solved by LLMs with ease would pose challenge to human intelligence. It is desirable to conduct a larger and more comprehensive experiment involving a larger group of participants to work on the whole dataset with a view to creating a more precise human intelligence baseline covering a wider range of questions.

References

S. Arora and A. Goyal. 2023. A theory for emergence of complex skills in language models. arXiv preprint, abs/2307.15936.

E. Bender and A. Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

P. Bhargava and V. Ng. 2022. Common-sense knowledge reasoning and generation with pretrained language models: a survey. In the Thirty-Sixth AAAI Conference on Artificial Intelligence.

M. J. Cresswell. 1973. Logics and languages (1st ed.), Routledge.

K. Cobbe, V. Kosaraju, et al. 2021. Training verifiers to solve math word problems. arXiv preprint, abs/2110.14168.

M. Geva, D. Khashabi, el al. 2021 Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. Transactions of the Association for Computational Linguistics, 9:346–361.

D. Hendrycks, C. Burns, et al. 2021. Aligning AI with shared human values. In the Proceedings of International Conference on Learning Representations.

J. Huang and K. C.-C. Chang. 2023. Towards reasoning in large language models: A survey. In the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.

Lucja Iwanska. 1993. Logical reasoning in natural language: It is all about knowledge. In Minds and Machines.

T. Kojima, S. S. Gu, et al. 2022. Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems.

A. Lewkowycz, A. Andreassen, et al. 2022. Solving quantitative reasoning problems with language models. In the Proceedings of NeurIPS.

H. Liu, R. Ning, et al. 2023. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4. arXiV preprint, abs/2304.03439.

T. Madusanka, I. Zahid, H. Li, I. Pratt-Hartmann, R. Batista-Navarro 2023. Not all quantifiers are equal: Probing transformer-based language models' understanding of generalized quantifiers. In the Proceedings of EMNLP.

S. Mishra, A. Mitra, et al. 2022. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.

A. Narayan and S. Kapoor. 2023. GPT-4 and professional benchmarks: the wrong answer to the wrong question. https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks.

A. Patel, S. Bhattamishra, and N. Goyal. 2021. Are NLP Models really able to solve simple math word problems? In the Proceedings of NAACL-HLT.

A. Patel, B. Li, et al. 2023. Bidirectional language models are also few-shot learners. In the Proceedings of the International Conference on Learning Representations.

S. Qiao, Y. Ou, et al. 2023. Reasoning with language model prompting: A survey. In the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.

J. W. Rae, S. Borgeaud, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint abs/2112.11446.

J. Rozanova, et al. 2021. Grounding Natural Language Instructions: Can Large Language Models Capture Spatial Information? arXiV preprint 2109.08634.

K. Sakaguchi, R. Le Bras, et al. 2020. Winogrande: An adversarial winograd schema challenge at scale. In the Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8732–8740.

T. Sawada, D. Paleka, et al. 2023. ARB: Advanced reasoning benchmark for large language models. arXiV preprint, abs/v:2307.13692.

P. Shojaee, et al. 2025. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. arXiv:2506.06941.

A. Srivastava, A. Rastogi, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research.

A. Talmor, J. Herzig, el al. 2018. CommonsenseQA: A question answering challenge targeting commonsense knowledge. arXiv preprint abs/1811.00937.

K. Valmeekam, A. Olmo, et al. 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In NeurIPS 2022 - Foundation Models for Decision Making Workshop.

S. Wang, Z. Liu, et al. 2022. From LSAT: The progress and challenges of complex reasoning. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

J. Wei, Y. Tay, et al. 2022. Emergent abilities of large language models. Transactions on Machine Learning Research.

J. Wei, X. Wang, et al. 2022. Chain of thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems.

Appendix A: Limitations and Potential Risks

The limitations and potential risks of the work is given in the paper after **Section 6**.

Appendix B: Information about the Dataset

B.1
This work proposes a new dataset of logical reasoning questions with details given in **Section 2**. The dataset is original and does not have any dependence on existing datasets.

B.2 & B.3
Hence, the work does not violate or need to check on the licensing terms and compatibility with the intended use of any dataset.

B.4
Since the questions in the dataset are synthetic in nature, they do not contain any personal data of real persons.

B.5
For the dataset created in this work per se, it is made freely available on Hugging Face for other researchers to use for evaluating and training LLMs. It is shared based on the MIT license. The details of how to generate the dataset are presented in the paper.

Documentation of the dataset
Tasks: Table Question Answering
Modalities: Text
Formats: JSON
Languages: English
Size: 1K
License: MIT

B.6
Number of examples: 84

Appendix C: Declaration on Computational Experiments

This work does not involve any computational experiment.

Appendix D: Details of Research with Human Participants

D.1
This work has only one experiment involving human participants, described in **Section 4.2**. The human participants were asked to answer 9 selected logical reasoning questions from the dataset, and they were tasked to work on the questions to find a room assignment problem fulfilling the specified constraints in the questions. All the questions are synthetic questions.

The full instruction given to the human participants are given in Figure 1 in **Section 2** and Appendix G. The selected question set is dispatched to the human participant separately through email, and the returned answer sheets (also via email) are anonymous with no participant personal information. No personal data have been collected during the process.

D.2
The participants were recruited in a school where one of the authors currently studies. The invitation was sent via a number of targeted school emails, and the selected question set was sent to responded participants only. No payment was made to the participants for participating in the experiment.

D.3
When returning the answers, the participants were required to provide the consent from them and their parents on the use of their answers to the logical reasoning questions for the publication in a preprint or other publications.

All the participants have been explicitly informed about the purpose and use of the questions and answers provided by them. They all understand that their answers to the selected question set would be used for research and are aware of that the results of the experiment will be published as preprint or other publications. They are aware that only the statistic of the answer accuracy would be published, and no individual answers or personal data will be published.

D.4
No ethics review board approval is required for the data collection protocol.

D.5
The result of the experiment is only the statistics of the accuracy of the answers to a number of synthetic, logical reasoning questions. No characteristics of human are reported in the work. Besides, the demographic information should not be a determining factor for the presented result even though the participants are all Chinese.

Appendix E: Declaration on the Use of AI Assistants

This work does not include any coding, and the dataset created is original. No AI assistants, ChatGPT or other LLMs have been used for the preparation of the paper because the authors prefer to write the paper themselves to get a more concise and succinct version.

Appendix F: Results of the One-shot Experiments

The experiments were repeated with a one-shot prompt setting, and the results are similar to those using a zero-shot prompt.

	Answer accuracy	Explanation accuracy
GPT-3	67.07%	41.67%
GPT-4	63.89%	54.36%
Gemini-1.5	24.61%	13.09%
Llama-3.1	72.62%	59.93%
Claude-3.5	53.18%	36.12%
DeepSeek-V3	51.98%	64.28%

Table F.1: Comparison of performance of LLMs on the proposed dataset with a one-prompt setting.

Appendix G: Prompt for the Experiments

The following is the prompt given to LLMs for all the experiments. The same text in the standard prompt is used as instructions given to human participants for the experiment using selected questions.

Standard prompt:

Please solve the following puzzle with step-by-step explanation.

Definition:
Given X, Y, Z are persons, define the following symbolic relationships:

1. **X: # o o** means that there are 3 rooms, and Person X is assigned to the room marked by #. For this example, X is assigned to the 1st room of the 3 rooms. Similarly, Y: o # o means that Person Y is assigned to the 2nd room.
2. **X: x o o** means that there are 3 rooms, and Person X can't be assigned to the room marked by x. For this example, X cannot be assigned to the 1st room of the 3 rooms. Similarly, Y: o x o means that Person Y can't be assigned to the 2nd room.
3. **X = Y** means Person X's room is next to Person Y's room regardless of the order. That is, X's room can be before after Y's room.
4. **X <> Y** means Person X's room is before Person Y's room. That is, X's room number is smaller than Y's. They need not be adjacent.
5. **X @ Y** means that X and Y are assigned to the same room.
6. **!** is the negation of the subsequent relationship operator. For example, X != Y means that Person X's room is not next to Person Y's room. Similarly, X !@ Y means that X, Y cannot be assigned to the same room.

Instructions:
In the following questions, you are asked to assign rooms to the persons and their pets according to the given restrictions. You should fill in as many rooms as possible, and all persons and their pets should be assigned rooms. No pet can stay alone in any room. Each room can accommodate at most one person and one pet.

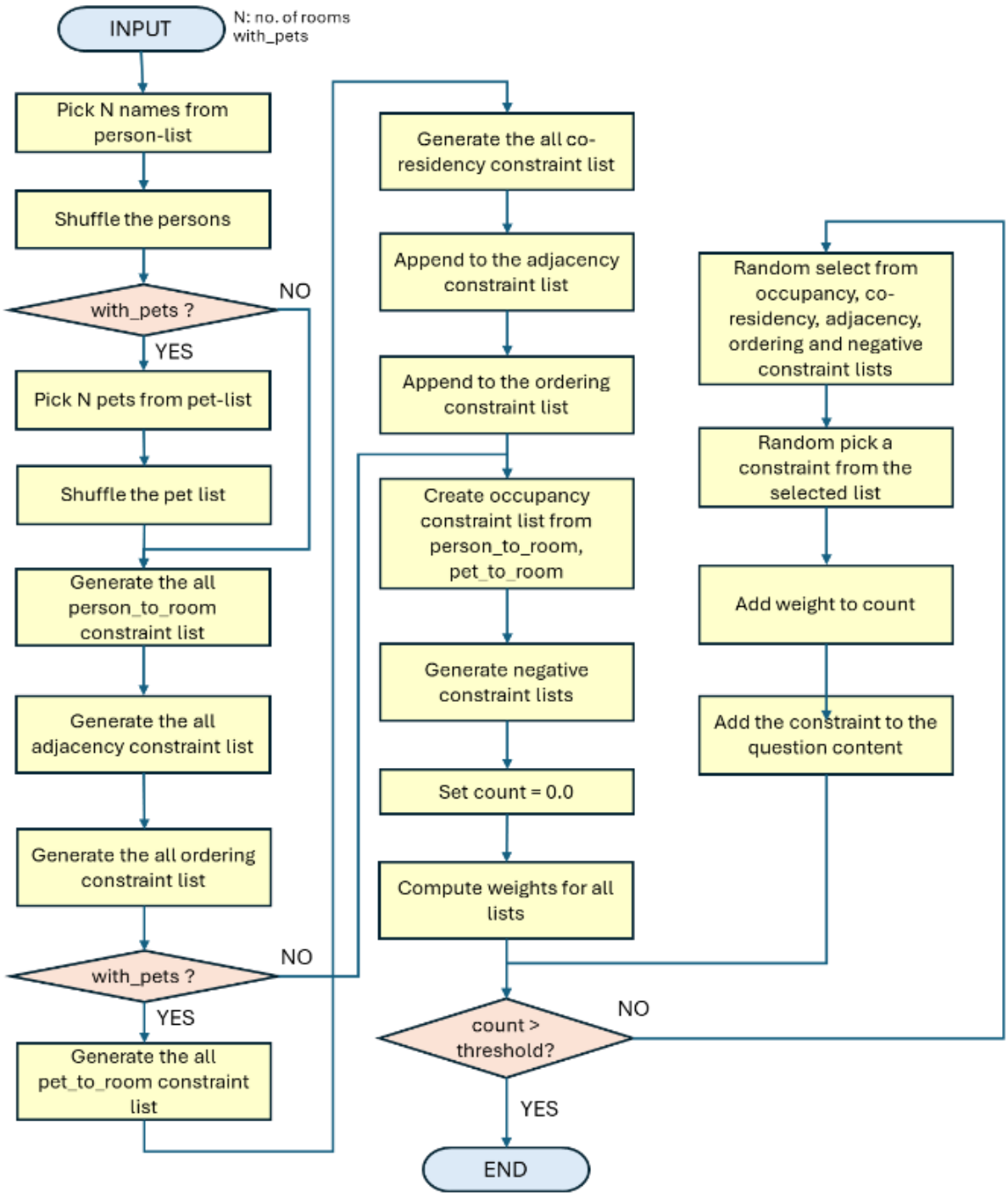
Question:
There are 4 rooms, which are assigned to Alice, Bob, Ken and Charlie with a dog, parrot, cat and turtle. You should fill in as many rooms as possible. Here are the restrictions for the room assignment:

1. Ken: o o o #
2. turtle: # o o o
3. cat @ Bob
4. dog: o o o #
5. cat <> Charlie

What are the tenants of the rooms?

One-shot prompt setting:

For the one-shot prompt setting, LLMs is given a question and answer as an example:

Please refer the following question and answer as an example ……

Appendix H: Automated Test Case Generation