KnowLogic: A Benchmark for Commonsense Reasoning via Knowledge-Driven Data Synthesis

Anonymous ACL submission

Abstract

Current evaluations of commonsense reasoning in LLMs are hindered by the scarcity of natural language corpora with structured annotations for reasoning tasks. To address this, we introduce KnowLogic, a benchmark generated through a knowledge-driven synthetic data strategy. KnowLogic integrates diverse commonsense knowledge, plausible scenarios, and various types of logical reasoning. One of the key advantages of KnowLogic is its adjustable difficulty levels, allowing for flexible control over question complexity. It also includes fine-grained labels for in-depth evaluation of LLMs' reasoning abilities across multiple dimensions. Our benchmark consists of 3,000 bilingual (Chinese and English) questions across various domains, and presents significant challenges for current LLMs, with the highest-performing model achieving only 68.17%. Our analysis highlights common errors, such as misunderstandings of low-frequency commonsense, logical inconsistencies, and overthinking. This approach, along with our benchmark, provides a valuable tool for assessing and enhancing LLMs' commonsense reasoning capabilities and can be applied to a wide range of knowledge domains.

1 Introduction

Reasoning is a higher cognitive function that involves analyzing, inducting, and deducing new information based on existing knowledge. It plays a fundamental role in human intelligence. Evaluating the commonsense reasoning ability of large language models (LLMs) is a crucial area of research in AI. This ability significantly influences LLMs' decision-making capabilities and is vital for advancing towards human-like intelligence in artificial general intelligence (AGI).

The massive natural language corpora on the Internet inherently lack sufficiently dense commonsense knowledge and logical reasoning data, as such information typically exists in implicit forms rather than explicit expressions within natural texts. This inherent deficiency results in the congenital weakness of LLMs' commonsense reasoning capabilities. To effectively evaluate LLMs' commonsense reasoning abilities, it is imperative to employ artificially synthesized reasoning texts embedded with high-density commonsense information. The primary challenge in this endeavor lies in ensuring both the accuracy of commonsense knowledge representation and the reliability of long-range reasoning chains.

Previous commonsense reasoning datasets typically relied on human annotation (Talmor et al., 2019, 2022; Boratko et al., 2020; Geva et al., 2021; Wei et al., 2024), template rules (Weston et al., 2015; Wang and Zhao, 2023; Parmar et al., 2024) or LLMs (Bai et al., 2024; Sakai et al., 2024; Sprague et al., 2024) for data generation. However, the lack of automation capability makes manual annotation challenging for building large-scale datasets, while template rules lead to a lack of diversity for generating varied texts, and LLMs struggle to ensure data quality. Furthermore, these benchmarks lack fine-grained features, which hinders a detailed analysis of model performance, and the data generation process is difficult to precisely control.

To address these issues, we propose a knowledge-driven synthetic data strategy. This involves creating a reliable knowledge base that integrates diverse commonsense knowledge and scenarios, along with logically rigorous reasoning systems capable of controlling the entire inference process to automatically generate accurate test questions and answers. The items in the knowledge base are annotated with fine-grained features, which are carried over to the generated data to support interpretable evaluation. By controlling features such as the length of the reasoning chain and the complexity levels of knowledge, the data can be generated at varying difficulty levels. Ta-

Datasets	Strategy	Accuracy Assurance	Automated Generation	Fine-grained Features	Controllable Difficulty
CommonSenseQA (Talmor et al., 2019) CommonSenseQA 2.0 (Talmor et al., 2022) ProtoQA (Boratko et al., 2020) StrategyQA (Geva et al., 2021) SimpleQA (Wei et al., 2024)	human annotation	✓	×	×	×
bAbI (Weston et al., 2015) TRAM (Wang and Zhao, 2023) LogicBench (Parmar et al., 2024)	generation based on templates	✓	✓	×	×
COIG-CQIA (Bai et al., 2024) mCSQA (Sakai et al., 2024) MuSR (Sprague et al., 2024)	LLM-based data generation	×	✓	×	×
KnowLogic (ours)	knowledge-driven data synthesis	✓	✓	✓	✓

Table 1: Comparison of commonsense reasoning datasets

Domain	Cases of Commonsense Reasoning
Space	David, Jennifer, John and James came to a hot-pot restaurant for a meal. They sat in a four-person booth. Two people sat in each booth, facing each other. David is to the right of Jennifer on the same booth. John is the right neighbour of James. Q: Who is diagonally opposite John? A: David
Time	Jack is a college student, and here are his weekly plans. Jack learns Japanese on Monday and plays badminton on Wednesday. 2 days after learning Japanese, Jack has a group meeting. 1 day after the group meeting, Jack cleans his dormitory room. Q: What will Jack do 4 days after he cleans his dormitory room? A: Learning Japanese.
Social	Alice is Bob's ex-wife, as well as Carol's ex-girlfriend. Dave is Alice's boss, a friend of Bob and also the husband of Eve, who is a classmate of Carol. Q: What is the relationship between Dave's wife and Alice's ex-boyfriend? A: Classmate
Nature	The four enclosures in the zoo keep carp, duck, turkey, and fox. The animal in enclosure No.3 has 4 less legs than the animal in enclosure No.4. The animal in enclosure No.2 can swim. Q: What is kept in enclosure No.1? A: Turkey

Table 2: Four domains of commonsense and reasoning cases

ble 1 compares three kinds of previous datasets with ours, **KnowLogic**[1], generated by knowledge-driven synthetic data strategy.

KnowLogic focuses on four domains of commonsense closely related to everyday human life: space, time, social, and nature. Table 2 shows the cases where these commonsenses are applied in reasoning. The benchmark is bilingual in Chinese and English, with three difficulty levels and diverse knowledge feature labels. The contributions of this paper are as follows:

1. We propose a knowledge-driven data synthesis method for reasoning that ensures accuracy while enabling large-scale automated generation. The transparent and traceable workflow facilitates interpretable evaluations of LLMs.

2. We automatically created a bilingual benchmark consisting of 3,000 commonsense reasoning data points across four domains: space, time, social, and natural knowledge. The dataset includes diverse feature labels and is categorized into three difficulty levels for in-depth LLM evaluation.

3. We conducted evaluations using 14 state-of-the-art reasoning LLMs. The highest-performing model achieved a score of 68.17%, and the average accuracy on the hardest level was below 40%, indicating that KnowLogic is a challenging benchmark.

4. Through case analysis, we identified several significant shortcomings in the commonsense reasoning capabilities of LLMs, including misunderstandings of low-frequency commonsense knowledge, self-contrast in logic, and overthinking, among others.

[1]The dataset will be released once the paper is accepted

2 Strategy of Knowledge-Driven Data Synthesis

2.1 Knowledge Framework for Data Synthesis in Commonsense Reasoning

KnowLogic relies on a knowledge base whose framework involves three core concepts: entities, propositions, and scenarios. Entities are independent perceptible objects such as people, animals, plants, or items. Propositions are used to state the properties of the entities and the relations between them. Scenarios provide the necessary context for commonsense reasoning.

Entities are distinguished by their properties and the corresponding values of those properties. For example, the entity "penguin" has a property called "class" with the value "bird". Based on properties, relations are formed by entities. Some relations are based on comparisons between property values. For instance, penguin and lion both have a property called "Leg-num" with values of 2 and 4 respectively, so there is a comparative relation of "2 less/more legs" between penguin and lion. Appendix A presents more examples of relations.

Propositions also have properties and relations. A proposition specifies its nature or refines its content through properties such as truth value and value precision, as detailed in Appendix B. Between propositions, there are five basic logical relations: equivalence, implication, inclusion, contrariety, and contradiction. For example, the proposition "A is in the first position to the right of B" implies the proposition "A is to the right of B."

In addition to propositions that state the properties and relations of entities, commonsense reasoning also requires an understanding of the specific scenarios in which additional implied information must be inferred to fully grasp the context. For example, the "four-person booth" scenario shown in Figure 1 implies at least the following facts: the involving entities are four people, and each booth has two positions. As such, when describing the spatial relations between entities, we might say "Entity A is to the right of Entity B", but we would not describe the right side of Entity A and Entity C, nor the left side of Entity B and Entity D. This demonstrates how scenarios constrain the way we describe the relations of entities. Appendix K lists the scenarios used by KnowLogic.

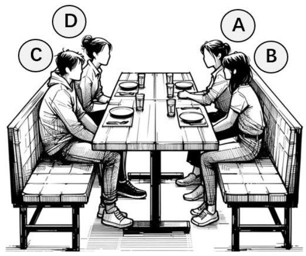


Figure 1: The "four-person booth" scenario

Table 3 presents the composition of the four knowledge bases. Some knowledge, like natural properties of entities, is extracted from existing sources such as HowNet (Dong and Dong, 2003), ConceptNet (Speer et al., 2017) and Wiki. Knowledge that cannot be automatically acquired, such as spatial relations, is written manually. 10 graduate students were hired to verify the correctness of the knowledge. After developing the first version of our knowledge base, we generated a batch of sample test data and hired 15 graduate students to assess and review. Based on the errors identified in the sample data, we traced them back to the knowledge base and made corrections. After three iterations, the accuracy of all 4 knowledge bases reached 100%.

2.2 Data Synthesis Workflow

Data synthesis consists of two stages: (1) preparing the knowledge base and (2) generating the question bank. The first stage involves extracting the necessary knowledge from existing sources, manually creating related templates, classifying the knowledge, and annotating their features. Details about knowledge base construction are shown in Appendix A.

The second stage is fully automated by the Inference Engine, which consists of four stages.

Step 1: Scenario Definition This first step establishes the scenario and enhances its presentation. It involves selecting entities/events from the knowledge base and integrating them into a scenario framework. Crucially, it generates introductory text for context and applies templates for natural language transformation of scenario elements. This combines scenario creation with immediate linguistic refinement for user-friendliness.

Step 2: Inference Data Generation The second step is utilising a Reasoner to generate inference data. The Reasoner generates a fact base by expanding a set of initial facts that describe properties of the entities or events using the relations and logic rules associated with the scenario. Fine-grained

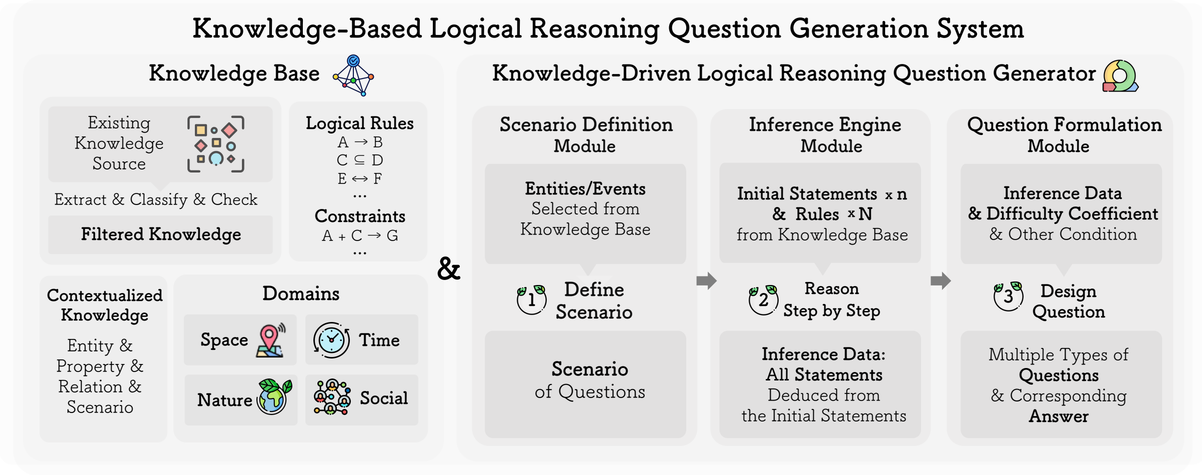


Figure 2: Overall process of data synthesis

Domain	Contextual Attribute	Num of Values	Example Values
Space	Scenarios	4	centrifugal hexagon, three rows two columns, etc.
	Properties	2	human, natural object
	Relations	14	up, down, left, right, east, west, etc.
	Number of Slots	3	4, 5, 6
Time	Scenarios	2	linear scenario, cyclic scenario
	Properties	4	time of occurrence, start time, end time, duration
	Relations	8	earlier, same duration, etc.
	Events	74	get married, play badminton, etc.
Social	Scenarios	2	social relations, kinship relations
	Properties	5	surname, first name, gender, spouse, related people
	Relations	76	father-son, classmates, etc.
Nature	Scenarios	3	farming, zoo ground allocation, items in photos
	Entities	633	123 animals, 147 plants, 363 artifacts
	Properties	18	color, shape, number of legs, etc.
	Relations	11	same color, more legs, etc.

Table 3: Commonsense-related attributes involved in the questions

features of each fact are recorded during the generation process to enable in-depth analysis. After the fact base is completed, the Reasoner adds facts to a statement set and verifies them step by step. This process is repeated until the statement set can uniquely determine the slot of each entity or event.

Step 3: Question Design The tird step is utilising a Question Generator to design the question. The Question Generator takes the statement set and the ground-truth arrangement of entities or events in the scenario as input and generates different types of statements to produce different types of questions.

Detailed workflow of data synthesis is shown in Appendix C.

3 The KnowLogic

3.1 Overall Introduction

KnowLogic is a dataset consisting of 3,000 questions spanning four major commonsense domains: space, time, social, and nature, along with a unique mix domain that integrates space and nature. With 600 questions per domain, KnowLogic offers bilingual data in both Chinese and English, except for the social domain, which is only available in Chinese[2]. The dataset's multi-domain cover-

[2]Social relationships in Chinese are more varied, making them more challenging to translate accurately into English. For example, the Chinese terms "爷爷" (paternal grandfather) and "外公" (maternal grandfather) are both typically translated as "grandfather" in English, which can lead to loss of specific cultural meaning. Due to the difference in granularity, translating these terms into English can affect the correctness of the questions.

age ensures a diverse range of commonsense and knowledge-attribute labels, setting it apart from many existing datasets. Moreover, the questions are stratified into three difficulty levels-easy, medium, and hard-allowing for testing LLMs' reasoning capabilities from basic to complex.

All questions are uniformly formatted as four-option multiple-choice questions, with the number of answers ranging from 1 to 4. Our questions can be described along two dimensions: the format of the options and the way the question is posed. For the format of options, we classify them as entity, event, slot and statement. When the options are statements, they are classified as correct and incorrect. For entities, events or slots, we categorize them as precise and vague. Examples of questions are shown in Appendix D.

3.2 Integrating Rich Commonsense into Questions

An independent knowledge base is constructed for each commonsense domain. Each base encompasses entities, properties, relations, and scenarios relevant to that domain. Based on these knowledge bases, we generate multi-domain questions combined with the scenarios. The questions carry a diverse range of commonsense-related attributes derived from the knowledge base, as shown in Table 3, enabling fine-grained evaluation of commonsense knowledge.

3.3 Storing Reasoning Attributes for Questions

Questions in KnowLogic are generated by connecting domain knowledge via logical rules, ensuring a clear and traceable reasoning workflow. Questions that models get wrong can be analyzed and traced back to specific knowledge and reasoning phases. Detailed cases are shown in Appendix J. Futhermore, questions are labeled with reasoning chain lengths, which refers to the minimum number of reasoning steps required to solve the problem. Shorter reasoning chains often indicate simpler questions, while longer chains reveal more intricate tasks that may require deeper reasoning capability.

3.4 Assigning Difficulty Levels to Questions

Calculated based on the commonsense-related and reasoning-related attributes of the questions, KnowLogic includes three difficulty levels: easy, medium and hard. The difficulty calculation formulas for each domain are given in Appendix E. The distribution ratio of difficulty levels is 1:2:3. We believe that current LLMs are in the phase of transitioning from easy to complex reasoning tasks. Therefore, KnowLogic focuses on more medium and hard questions, in order to accelerate the improvement of the model's reasoning capabilities.

4 Experiment

4.1 Experiment Setup

We evaluate a comprehensive set of existing models, including both open-source and closed-source ones. The tested models include `GPT-4o` (Hurst et al., 2024), `O1-Preview`, `O1-Mini`, `O3-Mini`, `GLM-4-Plus`, `GLM-Zero-Preview` (GLM et al., 2024), `DeepSeek-Chat (DeepSeek-V3)` (DeepSeek-AI et al., 2024), `DeepSeek-Reasoner (DeepSeek-R1)` (DeepSeek-AI et al., 2025), `Qwen-Max`, `Qwen2.5-72B-Instruct`, `QwQ-32B-Preview` (Qwen Team, 2024), and `Llama3.3-70B-Instruct` (AI@Meta, 2024). We also try the distilled model by Deepseek-R1, which is `r1-distill-qwen-32b`. (DeepSeek-AI et al., 2025) The experiments are conducted in both Chinese and English, with detailed test procedures provided in Appendix F. The answers are extracted from the model response using a rule based method. Details about the unextracted rates are shown in Appendix G.

4.2 Experiment Result

The complete results are presented in Table 4. Among the models tested, O1-Preview achieves the highest performance. In contrast, DeepSeek-R1, the top-performing open-source model, underperforms relative to the closed-source models. Overall, our benchmark remains a challenging test, effectively highlighting the limitations of current models across various domains. Notably, LLMs specifically trained for inference tend to outperform general LLMs within the same family when tackling such complex questions.

5 Analysis

5.1 General Analysis

Token Count differences Across Difficulty Levels Intuitively, for reasoning models, more challenging problems typically require longer reasoning chains, resulting in increased token counts in model outputs. We conducted experiments on the

Model	Space		Nature		Time		Social	Mix		Avg
	CN	EN	CN	EN	CN	EN		CN	EN	
Closed-Source Models										
o1-preview	**69.83**	**56.67**	**89.83**	**84.33**	79.83	80.33	42.00	**61.17**	**49.50**	**68.17**
o1-mini	66.17	56.17	82.00	75.17	**88.00**	**85.50**	30.33	48.00	40.00	63.48
claude-3-5-sonnet	38.50	37.50	76.00	70.67	60.83	72.00	50.00	36.33	36.67	53.17
glm-zero-preview	41.83	40.00	73.17	71.67	67.67	79.50	33.67	22.83	26.83	50.80
glm-4-plus	31.83	29.33	74.17	64.33	72.17	71.00	33.00	25.50	27.50	47.65
gpt-4o	28.83	30.67	68.17	65.50	66.17	69.83	22.50	23.67	24.17	44.39
qwen-max	28.17	27.50	65.83	63.83	54.00	72.00	38.83	24.17	24.00	44.33
o3-mini	27.17	28.00	57.33	57.67	58.83	62.33	20.83	26.67	30.67	41.06
Open-Source Models										
deepseek-reasoner	58.83	50.33	78.00	73.33	45.33	65.67	**73.50**	46.67	32.83	58.28
qwq-32B-preview	44.67	44.50	76.33	75.00	60.17	77.33	52.33	27.50	30.67	54.28
r1-distill-qwen-32b	44.33	35.17	81.83	60.00	66.17	65.67	46.33	26.50	23.33	49.93
deepseek-chat	32.00	31.83	67.83	63.83	61.33	65.33	40.33	24.83	30.17	46.39
qwen-25-72B	29.50	25.83	68.83	60.67	65.83	77.33	22.33	23.50	23.50	44.04
Llama-3.3-70B-Instruct	24.83	27.33	63.50	59.00	63.50	67.17	41.83	22.33	24.17	43.74

Table 4: Models Performance on Different Domains(Accuracy %). Among them, except for the social domain, all the other domains are English and Chinese topics. mix refers to the mixed problem combining space and nature, and the best models in each domain are already represented by **bold**.

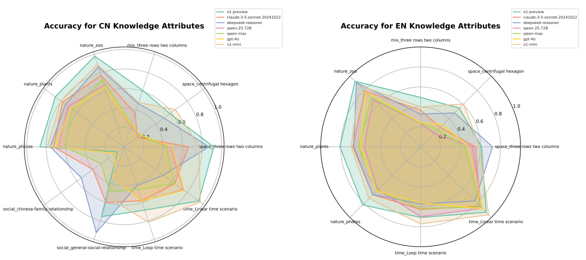


Figure 3: The performance of LLMs on Chinese and English questions with different knowledge attributes. The detailed descriptions of different senarios can be found in Appendix K.

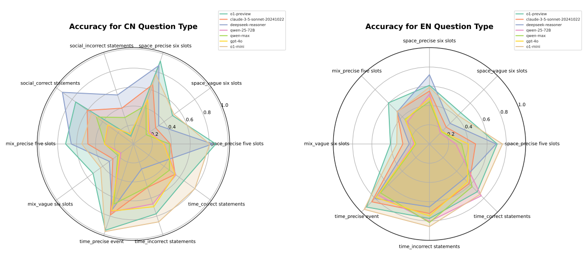


Figure 4: The performance of LLMs on Chinese and English questions with different question types. Here, if the answer to a question involves only a single entity, it is termed "precise." If it involves multiple entities, it is termed "vague".

relationship between token count and question difficulty. The results reveal that reasoning models exhibit a stronger correlation between question difficulty and reasoning length compared to standard models. Details are shown in Appendix I.

Performance Difference Across Domain in Knowledge Attributes We simultaneously analyze the performance of different models across various knowledge domains, and the results are presented in Figure 3. The overall distribution of correct rates follows a similar trend. However, even within the same domain, there are significant performance variations across different scenarios. For instance, models perform better on general social relationships than on family relationships. Similarly, models show weaker performance when addressing spatial scenarios such as the layout of centrifugal hexagon, compared to the arrangement of three rows and two columns.

Performance Difference Across Domain in Question types Additionally, the formulation of the question also plays a crucial role in the model's performance. For example, as shown in Figure 4, the phrasing of a question, such as asking the model to judge whether a statement is correct or incorrect, can influence the accuracy of the response. Similarly, the level of precision in describing spatial relations, such as the difference between vague and precise type of the six-slot scenario in space domain also affects the models to understand and respond correctly.

Performance Difference Across Domains and Levels Figure 5 presents the average accuracy of different domains at three different difficulty levels: easy, medium and hard. The data shows a clear downward trend in accuracy as the difficulty increases, suggesting that the complexity of the task significantly impacts model performance.

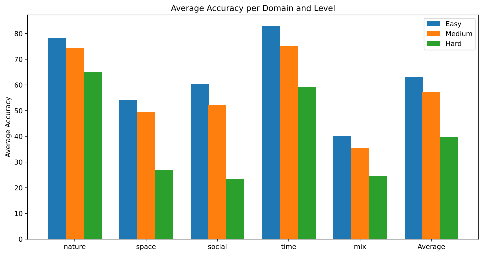


Figure 5: The average performance of LLMs on Chinese and English questions on different levels.

At the easy level, the accuracy is generally high, especially in the "nature" and "time" domains, where the accuracy reaches nearly 80%. This indicates that the models perform well on basic tasks in these areas. However, as the difficulty increases to medium level and hard level, the average accuracy of hard level falls below 40%. This drop reflects the increasing challenge posed by more complex reasoning tasks in these domains. Detailed performances on different domains across difficulty levels are shown in Appendix H.

5.2 Error Types Analysis

Low-Frequency Entity Properties and Similar Social Relationships are Prone to Commonsense Errors The commonsense errors of LLMs can be divided into two categories: errors in entity properties and errors in entity relations. Errors in entity properties mainly occur in low-frequency entities. It is language-dependent as the frequency of an entity can vary in corpora of different languages. For example, the model can correctly identify the properties of the mandarin fish (A freshwater fish primarily distributed in Asia) in Chinese but make commonsense errors in English. Errors in entity relations mainly manifest as the confusion of closely related similar social relationships such as 'classmate' and 'colleague', or misinterpreting asymmetric relationships such as treating "A is B's mentor" and "A is B's apprentice" as identical. Detailed examples are shown in Appendix J.

Models' Internal Capabilities Limitations Lead to Reasoning Errors The reasoning errors of LLMs can be divided into three categories: logical contradiction, imprecision and inference error. Logical contradiction refers to the self-contradiction within different parts of the model's response, for example, confirming A is above B in the preceding analysis, but later stating A is under B. This implies that LLMs have not yet possess logical reasoning capabilities aligned with humans. Imprecision refers to the model being affected by previous output, leading to errors in reasoning. The models may use the closer unrelated information to substitute the right information in the following analysis. The occurrence of this error is strongly related to next token prediction, which is regarded as the cornerstone of LLMs. Inference error refers to making a wrong inference based on a single clue. For example, in the centrifugal hexagonal scenario, the model regards 'the 5th position on the left' as 'the opposite side'. This type of error is related to the model's ability to construct and understand scenario. Detailed examples are shown in Appendix J. model's logical reasoning capabilities

Reasoning-Focused LLMs Tend to Overthink and Provide Additional Special Cases as Conditions Some reasoning LLMs, such as the OpenAI o1 and deepseek-r1 series, may overthink and invoke rare cases to fit the given conditions. For example, o1-preview argues that a birdcage can be decorated with white flowers so it can be the item with white flowers. This implies that the training methods for reasoning-focused models emphasizing logical consistency and depth of reasoning may lead them to miss more straightforward solutions. Detailed examples are shown in Appendix J.

Model's Tendency in Single-Choice Questions and Logical Contradictions Lead to Answer Aggregation Errors Despite mentioning that our questions are multiple-choice in the prompt, the models still tend to treat them as single-choice questions. For multiple-choice questions, the models sometimes output the answer immediately after finding one correct answer. This may be due to the high frequency of single-choice questions in inference data. Additionally, logical contradictions may also contribute to errors in answer aggregation. There are examples where the model infers A is right but ultimately answers B. Detailed examples are shown in Appendix J.

6 Related Work

According to the construction method, the previous common sense reasoning datasets can be divided into three categories:

Based on Human Annotation These datasets are constructed through manually crafting questions and answer options, with notable examples including CommonSenseQA (Talmor et al., 2019), CommonSenseQA 2.0 (Talmor et al., 2022), ProtoQA (Boratko et al., 2020), StrategyQA (Geva et al., 2021) and SimpleQA (Wei et al., 2024). CommonSenseQA utilizes a crowdsourcing approach to generate questions based on given subgraphs from ConceptNet. CommonSenseQA 2.0 guides workers in a game-like format to pose questions that are likely to cause LLMs to err using provided theme words and relationships, thereby constructing a more challenging dataset for models. ProtoQA gathered 9,762 open-ended questions from the TV show FAMILY-FEUD, collecting 100 answers for each question on a crowdsourcing platform and manually clustering them to assess models' understanding of commonsense in open-ended questions. StrategyQA hired annotators to decompose strategy questions and SimpleQA hired AI trainers to manually create 4326 questions. These datasets emphasize the quality and accuracy of data and covers a wide range of knowledge with diverse question styles. However, the high cost of manual annotation makes it difficult to build large-scale datasets using such methods.

Based on Template Rules These datasets automatically generate questions and answers through predefined task templates and rules, with notable examples including bAbI (Weston et al., 2015), TRAM (Wang and Zhao, 2023) and LogicBench (Parmar et al., 2024). The bAbI dataset comprises 20 subtasks, utilizing 9 action templates and a small number of entity templates to generate event descriptions. The TRAM dataset includes 10 subtasks. It collects historical events and time nodes from Wikipedia and uses rule templates to produce a total of 526,668 questions. LogicBench generates natural language questions by utilizing 18 logic inference rules to gather basic sentences generated by ChatGPT. This approach enables the rapid generation of large-scale data while significantly reducing construction costs. However, predefined rule templates in these datasets were limited in number and simple in structure, resulting in the lack of diversity and realism of the questions, making it difficult to ensure their generalizability.

Based on LLMs These datasets are built by leveraging existing corpora and utilizing LLMs to generate responses or questions, like COIG-CQIA (Bai et al., 2024), mCSQA (Sakai et al., 2024) and MuSR (Sprague et al., 2024). COIG-CQIA collects a vast amount of questions from web data, and prompts GPT-4 to generate the correspondings responses. mCSQA follows the construction procedure of CommonSenseQA but replaces manual annotation with LLMs. MuSR leverages LLMs to generate long-context reasoning questions by constructing a reasoning tree and crafting a story based on it. This approach reduces costs while enriching the diversity of questions and responses. However, the performance of current LLMs on reasoning tasks lags behind that of humans, making them unable to fully replace human input. As a result, the quality of the generated data is difficult to guarantee. Additionally, if a substantial amount of model-generated data is used for training, the performance ceiling of the trained model will be limited by the model that generated the data, making it more susceptible to model collapse during the training process.

7 Conclusion

In this paper, we introduce KnowLogic, a bilingual dataset synthesized from a reliable knowledge base and a program guiding logical reasoning, incorporating extensive commonsense knowledge. The dataset contains 3,000 questions across three difficulty levels and commonsense in four domains. Our experimental results show that, despite advances in areas like code generation and math problem-solving, LLMs still face challenges in commonsense reasoning. Case studies reveal common errors, such as misinterpreting low-frequency commonsense, logical inconsistencies, and overthinking. The proposed knowledge-driven synthesis method can be extended to other domains and used to generate large-scale reasoning datasets, providing valuable training data to enhance LLMs' commonsense reasoning abilities.

Limitations

Scales of Knowledge Base and Dataset Our knowledge base covers multiple domains and types of knowledge, and based on this, we utilize a knowledge-driven data synthesis strategy to create a dataset containing 3,000 questions. However, the scales of the knowledge base and the dataset are still relatively small, which is insufficient to support model fine-tuning. In the future, we plan to expand the knowledge base by adding more entities, properties, relationships, and constructing more scenarios. Our automatic data synthesis method has the potential to generate an infinite number of questions. We plan to enlarge the dataset to support model fine-tuning.

Limited Combinations in Mix Domain Currently, our mix domain involves only the integration of space and nature domains, as the knowledge base has not yet included scenarios for the integration of other domains. In the future, we plan to design more scenario types to enable diverse domain integrations.

Manual Error Type Analysis Currently, the error type analysis is conducted manually. This manual process is time-consuming and labor-intensive, and cannot comprehensively analyze all error samples. Our questions are annotated with diverse and fine-grained features, and the data synthesis process is done step by step, with a clear and traceable reasoning chain. In the future, we will explore automatic error analysis through methods such as error feature localization and reasoning path backtracking.

Ethical Consideration

This work involved human annotation. We have provided appropriate compensation for all annotators. The total cost of annotation for the project is about 20k RMB. For all annotators, we explicitly informed them about the use of the data and required them to ensure that the questions included in KnowLogic do not involve any social bias, ethical issues or privacy concerns during the annotation process.

References

AI@Meta. 2024. Llama 3 model card.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. 2024. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *arXiv preprint arXiv:2403.18058*.

Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew Mccallum. 2020. Protoqa: A question answering dataset for prototypical common-sense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1122–1136.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1107–1128. Association for Computational Linguistics.

Zhendong Dong and Qiang Dong. 2003. Hownet-a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 820–824. IEEE.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Preprint*, arXiv:2101.02235.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214, Bangkok, Thailand. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of ai through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Yuqing Wang and Yun Zhao. 2023. Tram: Benchmarking temporal reasoning for large language models. *arXiv preprint arXiv:2310.00835*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

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

A Knowledge Base Construction

The knowledge bases in four domains are constructed under the same overarching knowledge framework, which includes entities, scenarios, and propositions to express properties and relations. We first extracted knowledge from existing external resources, then manually created proposition templates. These templates, once filled up, will generate factual statements during the data synthesis process. The logical rules among the templates needed by the Reasoner are also connected manually. After three rounds of manual check, we formed our current knowledge bases.

External Resources In time domain, 9 historical events that occurred between 1900 and 2000, such as the birth year of Michael Jackson, were selected from Wikipedia and added to the knowledge base. In social domain, 58 complex family relationships spanning two to three generations are comes from a calculator[3] for computing Chinese social relationships, such as grandfather is equivalent to father's father. In nature domain, 633 entities are from HowNet and then get their properties from HowNet, ConceptNet and the Contemporary Chinese Dictionary. The properties are clustered according to the Attribute Value list in HowNet.

Manual Construction Table 5 presents the number of knowledge entries written by humans in four knowledge bases. Scenarios are shown in Appendix K.

B Properties of statements

Our knowledge bases define certain properties to refine the context of statements, as listed in table 6.

C Technical Details of Data Synthesis

Detail information about the Reasoner and Question Generater is as follows:

Details of the Reasoner The Reasoner takes the relations associated with the scenario and all properties of the entities or events as input. It maintains a fact base, initialized by the descriptions of all properties of all entities or events, such as 'strawberry is on the middle floor', 'Tom was born in 1958', 'Mary is Tom's wife', and 'strawberry is red'. The program then automatically traverses the logic relationships and inference rules in the knowledge base, matches them with the initial facts, generates new facts, and adds them to the fact base. The program then takes the newly added facts in the fact base, along with the original facts, as the new initial facts and inputs them into the Reasoner again. This process is repeated until no new facts

[3] https://github.com/mumuy/relationship/

Domain	Type of Entry	#	Example of Entry
Space	Spatial Property	59	X faces south, X is on the first floor, etc.
	Spatial Relation	244	X is to the left of Y, X is facing away from Y, etc.
	Logical Rule	861	"X is to the left of Y" is equivalent to "Y is to the right of X"; "X is on the first tier" and "Y is on the third tier" can imply that "X is separated from Y by one tier in between".
Time	Temporal Event	39	X get married in T, X played badminton on T, etc.
	Temporal Relation	8	A happened earlier than B, A happened T days after B, etc.
	Logical Rule	51	"A happened before B" is equivalent to "A happended after B"; "A happened in T" and "B happened in T" can imply that "A and B happened at the same time"; In a person's life, getting married happens later than starting elementary school.
Social	Individual Property	42	surname:Li, first name:Dawei, gender:male
	Social Relation	18	X is Y's father, X is a friend of Y, etc.
Nature	Natural Property	16	X is a kind of V, X has V legs, etc.
	Comparative Relation	10	X has a longer wavelength than the light reflected by Y.
	Logical Rule	27	"X is a bird" can imply that "X is a homothermal animal", red light has a longer wavelength than yellow light, A same stone sinks faster in freshwater than in seawater.

Table 5: Knowledge entries written by human in four knowledge bases. The X and Y in the entry are placeholders for entities that can be filled in. The A and B are for events. The T is for the times. The V is for the value of natural properties.

Property	Definition	Statement	Value
truth value	Truth value of a statement, either true or false	Tom isn't the son of Thomas.	false
value precision	Number of values determined by the statement, either precise (single-valued) or vague (multi-valued).	Jack is within the range to the left of Tom.	vague
time	Time that a event in the statement happens	Mary retired in 1955.	1955

Table 6: Examples of the property of statements in the knowledge base

can be generated. Each fact is labeled with the properties or relations involved to enable fine-grained analysis.

With the fact base completed, the Reasoner selects a set of facts that can uniquely determine the slot of each entity or event in the scenario step by step. For each step, the Reasoner randomly selects a fact from the fact base, adds it to the statement set, and verifies whether the statement set can uniquely determine the slot of each entity or event. During this process, the program automatically records the properties or relations involved in the statements and the number of inference steps. This process is repeated until the answer is 'Yes'.

Details of the Question Generator The Question Generator takes the statement set and the ground-truth arrangement of entities or events as input. It first chooses a question type. If the question type is 'Correct Statement' or 'Incorrect Statement', the generator will randomly select four pairs of entities or events to produce statements, either correct or incorrect, as options. When the question type is 'Precise Entity/Event/Slots', the generator generates a proposition that can uniquely determine the slot of the entity or event in the scenario and mask the relevant information. When the question type is 'Vague Slots', the generator generates a proposition that multiple entities can satisfy this proposition and identifies all those entities as potential answers. For options, if the scenario has 4 slots, then all entities/events/slots will be the options. Otherwise, it randomly selects three of them as options A, B, and C, then add 'None of the above' as option D.

D Question examples of different domains and question-types

Table 7 shows the examples of questions across different domains and question-types in our dataset.

Domain	Scenario	Question-type	Example of question
Space	centrifugal hexagon	6 slots - vague	周伯通、郝大通、柯镇恶、赵志敬、刘处玄、王重阳六位道士在终南山重阳宫内盘腿席地打坐，围成一个圆圈，修炼内功，六人的位置恰好形成一个正六边形。六人都面朝外背对圆心而坐。任意相邻两人之间的间距相等，大约为一米。已知： 刘处玄的右边接着就是赵志敬， 郝大通在赵志敬的右边，二者相邻， 从赵志敬的左边数起第二个位置是王重阳， 从柯镇恶的左边数起第五个位置是王重阳， 从王重阳的左边数起第二个位置是周伯通。 问题： 赵志敬与___之间隔着两个位置。 选项：A.刘处玄 B.周伯通 C.柯镇恶 D.郝大通 答案：C David, Jennifer, Elizabeth, Michael, John, James, —these six Taoist priests are seated cross-legged on the ground inside the Chongyang Palace on Zhongnan Mountain, arranged in a circle as they practice internal martial arts. The positions of the six priests form a perfect hexagon. Each priest is facing outward, with their backs toward the center of the circle. The distance between any two adjacent priests is equal, approximately one meter. It is known that: Michael is directly to the right of John; Jennifer is directly to the right of Michael; James occupies the second position to the left of Michael; James occupies the fifth position to the left of Elizabeth; David occupies the second position to the left of James. question: Michael and ___ are separated by two positions. Options: A.John B.David C.Elizabeth D.Jennifer Answer:C
		5 slots - precise	柯镇恶、王处一、郝大通、尹志平、刘处玄、赵志敬六位道士在终南山重阳宫内盘腿席地打坐，围成一个圆圈，修炼内功，六人的位置恰好形成一个正六边形。六人都面朝外背对圆心而坐。任意相邻两人之间的间距相等，大约为一米。已知： 尹志平在赵志敬右边数起第五个位置， 柯镇恶的左边紧接着就是刘处玄， 尹志平在郝大通左边数起第二个位置。 问题： ___在尹志平左边数起第三个位置。 答案：D 选项：A.郝大通 B.刘处玄 C.柯镇恶 D.以上选项都不是 Elizabeth, Robert, Jennifer, Susan, John, Michael, —these six Taoist priests are seated cross-legged on the ground inside the Chongyang Palace on Zhongnan Mountain, arranged in a circle as they practice internal martial arts. The positions of the six priests form a perfect hexagon. Each priest is facing outward, with their backs toward the center of the circle. The distance between any two adjacent priests is equal, approximately one meter. It is known that: Susan occupies the fifth position to the right of Michael; John is directly to the left of Elizabeth; Susan occupies the second position to the left of Jennifer. question: ___ occupies the third position to the left of Susan. Options: A.Jennifer B.John C.Elizabeth D.None of the above Answer:D
	three rows two columns	6 slots - vague	月季、水仙、茉莉、君子兰、天竺葵、郁金香六盆花放置在三层花架上呈列，花架紧靠大厅南墙放置，每层两格，各放一盆花，一在东，一在西。画师站在花架前，面对花架支起画架，为花架中六盆花画素描。在描述各花的方位关系时，约定以画师自身左右方位为参照，即东侧花盆为左，西侧花盆为右。东侧花盘在西侧花盘左边，西侧花盆在东侧花盆右边。已知： 郁金香的正上方是月季的正下方， 月季在天竺葵左上方且二者隔了一层， 天竺葵在一层西侧， 君子兰在天竺葵左上方且二者不隔层， 月季在茉莉左边， 水仙在君子兰右边。

Domain	Scenario	Question-type	Example of question
			问题： ___所在层和天竺葵所在层相邻。 答案：D 选项：A.茉莉　B.月季　C.郁金香　D.以上选项都不是 Monthly Rose, Narcissus, Jasmine, Clivia, Geranium, Tulip, —these six pots of flowers are arranged on a three-tiered flower stand, placed against the south wall of the hall. Each tier is divided into two sections, with one pot placed in each section—one on the east side and one on the west. The artist stands in front of the flower stand, facing it, and sets up an easel to sketch the six pots of flowers. When describing the positional relationships of the flowers, the artist's own left and right are used as a reference, with the eastern pot being on the left and the western pot being on the right. In other words, the eastern pot is to the left of the western pot, and the western pot is to the right of the eastern pot. It is known that: The position directly above Tulip is directly below Monthly Rose; Monthly Rose is located at the upper left side of Geranium and there is a tier between them; Geranium is on the west side of the first floor; Clivia is located in the upper left corner of Geranium and the two are not separated by a layer; Monthly Rose is to the left of Jasmine; Narcissus is somewhere to the right of Clivia. question: The tier where ___ is located above or below the tier where Geranium is located. Options: A.Jasmine B.Monthly Rose C.Tulip D.None of the above Answer:D
Space		5 slots - precise	波斯菊、天竺葵、君子兰、郁金香、茉莉、茶花六盆花放置在三层花架上呈列，花架紧靠大厅南墙放置，每层两格，各放一盆花，一在东，一在西。画师站在花架前，面对花架支起画架，为花架中六盆花画素描。在描述各花的方位关系时，约定以画师自身左右方位为参照，即东侧花盆为左，西侧花盆为右。东侧花盘在西侧花盘左边，西侧花盆在东侧花盆右边。已知： 茉莉在二层左侧， 郁金香在茶花左边且二者同层， 茉莉在郁金香正上方且二者不隔层， 茉莉在波斯菊左下方且二者不隔层， 君子兰的右邻在茉莉的右上方。 问题： ___在茉莉右边且二者同层。 答案：D 选项：A.郁金香　B.君子兰　C.波斯菊　D.以上选项都不是 Cosmos, Geranium, Clivia, Tulip, Jasmine, Camellia, —these six pots of flowers are arranged on a three-tiered flower stand, placed against the south wall of the hall. Each tier is divided into two sections, with one pot placed in each section—one on the east side and one on the west. The artist stands in front of the flower stand, facing it, and sets up an easel to sketch the six pots of flowers. When describing the positional relationships of the flowers, the artist's own left and right are used as a reference, with the eastern pot being on the left and the western pot being on the right. In other words, the eastern pot is to the left of the western pot, and the western pot is to the right of the eastern pot. It is known that: Jasmine is on the left side of the second tier; Tulip is to the left of Camellia and both are on the same level; Jasmine is directly above Tulip and they are not seperated by any tier; Jasmine is located at the lower left of Cosmos and there is no tier between them; The right side neighbor of Clivia is above the upper right side of Jasmine. question: ___ is to the right of Jasmine and both are on the same tier. Options: A.Tulip B.Clivia C.Cosmos D.None of the above Answer:D

Domain	Scenario	Question-type	Example of question
Time	Linear Scenario	Precise Event	小明的女儿正在给朋友讲述父亲的一生: (1)在迈克尔·杰克逊出生之前58年，他出生; (2)他度过一生的时长为99年; (3)在联合国成立的38年之前，他开始上小学; (4)他小学毕业的时间比迈克尔·杰克逊出生早45年; (5)在1913年，他开始上初中; (6)他上初中一共3年; (7)在联合国成立的30年之前，他遇见未来的妻子; (8)他开始上高中的时间比他出生晚16年; (9)他高中毕业的时间比他开始上初中晚6年; (10)在他遇见未来的妻子之后4年，他开始上大学; (11)在他初中毕业之后8年，他大学毕业。 问题：____是在迈克尔·杰克逊出生的42年前。 选项：A.他开始上小学　B.他去世　C.他初中毕业　D.他高中毕业 答案：C Jack's daughter is telling her friends about the story of her father's life: (1)Jack was born 58 years before Michael Jackson was born; (2)Jack lived his life for 99 years; (3)Jack started elementary school 38 years before the United Nations was founded; (4)Jack graduated from elementary school 45 years before Michael Jackson was born; (5)Jack started junior high school in 1913; (6)Jack studied in junior high school for 3 years; (7)Jack met his future wife 30 years before the United Nations was founded; (8)Jack started high school 16 years after he was born; (9)Jack graduated from high school 6 years after he started junior high school; (10)Jack started university 4 years after he met his future wife; (11)Jack graduated from university 8 years after he graduated from junior high school. Question: ____ 42 years before Michael Jackson was born. Options: A.Jack started elementary school B.Jack passed away C.Jack graduated from junior high school D.Jack graduated from high school Answer: C
		Correct Statements	小明的女儿正在给朋友讲述父亲的一生: (1)在1900年，他出生; (2)在联合国成立的54年之后，他去世; (3)他开始上初中的时间比第一届普利策奖颁发早4年; (4)在迈克尔·杰克逊出生的42年之前，他初中毕业; (5)在1915年，他遇见未来的妻子; (6)他开始上高中的时间比他出生晚16年; (7)在他开始上初中之后6年，他高中毕业; (8)他成为父亲的时间比他遇见未来的妻子晚23年; (9)他退休的时间比他出生晚64年。 问题：以下选项中正确的是____ 选项: A. 谷歌创立和联合国成立之间相隔97年。 B. 他成为父亲的时间比谷歌创立早20年。 C. 他开始上初中是在第一届普利策奖颁发之后。 D. 他开始上高中和他出生之间相隔16年。 答案：D Jack's daughter is telling her friends about the story of her father's life: (1)Jack was born in 1900; (2)Jack passed away 54 years after the United Nations was founded; (3)Jack started junior high school 4 years before the first Pulitzer Prizes were announced; (4)Jack graduated from junior high school 42 years before Michael Jackson was born; (5)Jack met his future wife in 1915; (6)Jack started high school 16 years after he was born;

Domain	Scenario	Question-type	Example of question
Time			(7)Jack graduated from high school 6 years after he started junior high school; (8)Jack became a father 23 years after he met his future wife; (9)Jack retired 64 years after he was born. Question: Select the correct statement(s): ____ Options: A.The gap between the time the United Nations was founded and the time Google was founded is 97 years. B.Jack became a father 20 years before Google was founded. C.Jack started junior high school after the first Pulitzer Prizes were announced. D.The gap between the time Jack was born and the time he started high school is 16 years. Answer: D
	Cyclic Scenario	Incorrect Statements	小明是一名大学生，以下是他的每周安排: (1)星期三，他打羽毛球; (2)周三，他开组会; (3)在星期三，他跑步; (4)在他打羽毛球之后1天，他阅读科幻小说; (5)在他开组会的2天之后，他练习吉他; (6)在他开组会之后3天，他看论文。 问题：以下选项中不正确的是____ 选项: A.在他打羽毛球的5天之前，他跑步。 B.在他阅读科幻小说的6天之后，他打羽毛球。 C.在他看论文的3天之前，他阅读科幻小说。 D.在他练习吉他之后2天，他阅读科幻小说。 答案：ACD Jack is a college student, and here are his weekly plans: (1)On Wednesday, Jack plays badminton; (2)Jack has a group meeting on Wednesday; (3)Jack goes jogging on Wednesday; (4)Jack reads sci-fi novels 1 day after he plays badminton; (5)2 days after Jack has a group meeting, he practices the guitar; (6)3 days after Jack has a group meeting, he reads research papers. Question: Select the incorrect statement(s): ____ Options: A.Jack goes jogging 5 days before he plays badminton. B.Jack plays badminton 6 days after he reads sci-fi novels. C.Jack reads sci-fi novels 3 days before he reads research papers. D.2 days after Jack practices the guitar, he reads sci-fi novels. Answer: ACD
Social	General Social Relationship	Correct statements	已知：冯志强是冯秀英的哥哥，也是周强的朋友。王丽是冯志强的妻子，也是周强的同学。郑建国是王丽的前男友，也是李晓静的男朋友。吴晶是郑建国的数位前女友中的一位，也是孙大伟的前妻。周强是吴晶的领导，也是赵伟的下属。李晓静是周强的数位领导中的一位，也是孙大伟的同事。钱静是孙大伟的前女友，也是赵伟的女朋友。 问题：以下选项正确的是___ 选项: A.孙大伟的同事是吴晶的前男友的媳妇 B.孙大伟的前妻是李晓静的男朋友的母亲 C.赵伟的下属是郑建国的某位名字中带有丽的前女友的同学 D.赵伟的下属是郑建国的某位名字中带有丽的前女友的同事 答案：C (Known: Feng Zhiqiang is Feng Xiuying's elder brother and also a friend of Zhou Qiang. Wang Li is the wife of Feng Zhiqiang and also a classmate of Zhou Qiang. Zheng Jianguo is Wang Li's ex boyfriend and also Li Xiaojing's boyfriend. Wu Jing is one of Zheng Jianguo's several ex girlfriends and also the ex-wife of Sun Dawei. Zhou Qiang is the leader of Wu Jing and also a subordinate of Zhao Wei. Li Xiaojing is one of Zhou Qiang's digital leaders and also a colleague of Sun Dawei. Qian Jing is the ex girlfriend of Sun Dawei and also the girlfriend of Zhao Wei Question: Which of the following options is correct Options:

Domain	Scenario	Question-type	Example of question
			A. Sun Dawei's colleague is the wife of Wu Jing's ex boyfriend B. Sun Dawei's ex-wife is the mother of Li Xiaojing's boyfriend C. Zhao Wei's subordinate is a classmate of Zheng Jianguo whose name includes Li's ex girlfriend D. Zhao Wei's subordinate is a colleague of Zheng Jianguo whose name includes Li's ex girlfriend Answer: C)
Social	Chinese Family Relationship	Incorrect statement	已知：赵鹏是赵石的女儿的孙子，也是赵芳的孙子。赵军是赵力的孙女的儿子，也是赵琳的爷爷。赵强是赵秀英的爸爸，也是赵丽的外婆的弟弟。赵秀英是赵晓静的爸爸的侄女，也是赵晓丽的姐姐。赵力是赵晓丽的外孙女。赵玉兰是赵秀兰的弟弟，也是赵盈盈的爸爸的爸爸的妹妹。赵秀兰是赵强的妹妹。赵丽是赵晓静的女儿，也是赵静的外孙女。赵芳是赵盈盈的女儿的孙女。 问题：以下选项不正确的是___ 选项： A.赵石是赵秀兰的侄女的外甥女的孙子。 B.赵玉兰是赵秀英的哥哥的女儿。 C.赵盈盈是赵力的妈妈。 D.赵鹏是赵秀兰的姐姐的女儿的外孙女的孙女的孙子。 答案：B (Known: Zhao Peng is the grandson(son's son) of Zhao Shi's daughter and also the grandson(son's son) of Zhao Fang. Zhao Jun is the son of Zhao Li's granddaughter(son's daughter) and also the grandfather(father's father) of Zhao Lin. Zhao Qiang is the father of Zhao Xiuying and the younger brother of Zhao Li's grandmother(mother's mother). Zhao Xiuying is the niece(brother's daughter) of Zhao Xiaojing's father and also the elder sister of Zhao Xiaoli. Zhao Li(1) is the granddaughter(daugher's daughter) of Zhao Xiaoli. Zhao Yulan is the younger brother of Zhao Xiulan and also the younger sister of Zhao Yingying's father's father. Zhao Xiulan is Zhao Qiang's younger sister. Zhao Li(2) is the daughter of Zhao Xiaojing and also the granddaughter(daughter's daughter) of Zhao Jing. Zhao Fang is the granddaughter(son's daughter) of Zhao Yingying's daughter Question: Which of the following options is incorrect Options: A. Zhao Shi is the grandson(son's son) of Zhao Xiulan's niece(brother's daughter)'s niece(sister's daughter) B. Zhao Yulan is the daughter of Zhao Xiuying's elder brother C. Zhao Yingying is Zhao Li(1)'s mother D. Zhao Peng is the grandson(son's son) of the granddaughter(son's daughter) of Zhao Xiulan's sister's daughter's granddaughter(daughter's daughter) Answer: B)
Nature	Farming	Precise entity	一位勤劳的农夫有四块田，他在四块田中分别种植了南瓜、开心果、瓠子、枇杷四种作物。已知： 1号田中的作物是一种坚果； 2号田中的作物属于蔬菜； 3号田中的作物反射的光比2号田中的作物反射的光波长更长； 3号田中的作物属于蔬菜。 问题：4号田中种的是____。 选项：A.南瓜B.开心果C.瓠子D.枇杷 答案：D A hardworking farmer has four fields. In each field he plants one of four crops: pumpkin, pistachio nut, edible gourd, loquat. It is known that: The plant in field No.1 is a kind of nut; The plant in field No.2 is a vegetable; The light reflected by the plant in field No.3 has a longer wavelength than the light reflected by the plant in field No.2; The plant in field No.3 is a vegetable. Question: ____ is planted in field No.4. Options: A.pumpkin B.pistachio nut C.edible gourd D.loquat Answer: D
	Zoo enclosures	Precise position	动物园里的四个场馆分别养着蝴蝶、章鱼、水牛、白鹭四种动物。已知： 3号场馆中的动物比2号场馆中的动物多2条腿； 3号场馆中的动物属于恒温动物； 4号场馆中的动物比3号场馆中的动物多2条腿。

Domain	Scenario	Question-type	Example of question
Nature			问题：白鹭养在____号场馆中。 选项: A.1 B.2 C.3 D.4 答案：B The four enclosures in the zoo keep four different kinds of animals: butterfly, octopus, buffalo, egret. Now we know that: The animal in enclosure No.3 has 2 more legs than the animal in enclosure No.2; The animal in enclosure No.3 is a homothermal animal; The animal in enclosure No.4 has 2 more legs than the animal in enclosure No.3.
	Items on photos	vague position	一面墙上贴着芒果、苹果汁、杨梅、胡萝卜四种物品的照片。已知: 4号照片上中的物品不属于水果; 3号照片上的物品反射的光比1号照片上的物品反射的光波长更长; 2号照片上的物品的味道是酸的; 4号照片上中的物品不属于蔬菜; 2号照片上的物品属于水果。 问题：水果在____号照片上。 选项: A.1 B.2 C.3 D.4 答案: AB On a wall pasted photos of four different items: mango, apple juice, bayberry, carrot. Now we know that: The item on photo No.4 is not a fruit; The light reflected by the item on photo No.3 has a longer wavelength than the light reflected by the item on photo No.1; The item on photo No.2 tastes sour; The item on photo No.4 is not a vegetable; The item on photo No.2 is a fruit. Question: A fruit is on photo No.____. Options: A.1 B.2 C.3 D.4 Answer: AB
Mix	Three rows two columns	5 slots - precise	香草、铅笔盒、吐司、缝衣针、大葱、花生糖六种商品在三层货架上放置，货架紧靠商店南墙放置，每层两格，各放一种商品，一在东，一在西。顾客站在货架前选购商品。在描述各商品的位置关系时，约定以顾客自身左右方位为参照，即东侧商品为左，西侧商品为右。已知: 属于调味料的蔬菜在花草正下方且二者隔了一层; 花草在顶层; 花草在吐司左边; 甜的加工食品在吐司左下方且二者不隔层; 铅笔盒和属于调味料的蔬菜在同一层。 问题：___在铅笔盒左上方且二者不隔层 选项：A.吐司B.香草C.大葱D.以上选项都不是 答案：D Vanilla, pencil-box, toast, sewing needle, scallion, peanut brittle, six items are placed on a three-tier shelf, which is positioned against the south wall of the store. Each tier has two sections, with one type of item placed in the east section and one in the west section. A customer is standing in front of the shelf. When describing the positional relationships of the items, it is agreed that the customer's own left and right will be used as a reference, with the east section being on the left and the west section being on the right. It is known that: The seasoner and vegetable is located directly below the flower or grass and separated by one tier, The flower or grass is on the top floor, The flower or grass is to the left of the toast, The sweet processed food is located at the lower left of the toast and there is no tier between them, The pencil-box and the seasoner and vegetable are on the same level. ___ is located in the upper left corner of pencil-box and the two are not separated by a layer Options: A.toast B.vanilla C.scallion D. None of the above Answer: D

Domain	Scenario	Question-type	Example of question
Mix	Three rows two columns	6 slots - vague	南瓜、白鲢、红豆、柠檬、购物袋、龙须面六种商品在三层货架上放置，货架紧靠商店南墙放置，每层两格，各放一种商品，一在东，一在西。顾客站在货架前选购商品。在描述各商品的位置关系时，约定以顾客自身左右方位为参照，即东侧商品为左，西侧商品为右。已知： 加工食品在鱼上一层； 工具在顶层； 一层左侧是开黄色花的橙色的物品； 加工食品在黄色的物品正上方； 红色的物品在黄色的物品正下方； 二层西侧是黄色的物品。 问题：___红豆上方 选项：A.白鲢B.龙须面C.柠檬D.以上选项都不是 答案：ABC Pumpkin, silver carp, red bean, lemon, carrier bag, dragon whiskers noodles, six items are placed on a three-tier shelf, which is positioned against the south wall of the store. Each tier has two sections, with one type of item placed in the east section and one in the west section. A customer is standing in front of the shelf. When describing the positional relationships of the items, it is agreed that the customer's own left and right will be used as a reference, with the east section being on the left and the west section being on the right. It is known that: The processed food is one tier above the fish, The tool is on the top floor, The orange item with yellow flower is on the left side of the first tier, The processed food is above the yellow item in a straight line, The red item is directly under the yellow item, The yellow item is on the west side of the second floor. ___ is above the red bean Options: A.silver carp B.dragon whiskers noodles C.lemon D. None of the above Answer: ABC

Table 7: Question examples of different domains and question-types

E Difficulty Calculation Formula of Reasoning Problems in Various Domains

All domain-specific difficulty formulas follow a unified design philosophy: they combine weighted indicators reflecting cognitive load (e.g., entity count, reasoning chain length), knowledge complexity (e.g., property difficulty, scene type), and information completeness (e.g., explicit mentions, coverage). The coefficients balance domain-specific priorities while maintaining comparable difficulty scales across categories.

Domain-Specific Formulas:

Space:

$$level = 0.3nm + 0.5pr + 0.2al$$

- nm: Number of entities
- pr: Precision, $pr = 1$ when the number of entities involved in the question exceeds 2
- al: Whether entity coverage of the text is incomplete, $al = 1$ when entity coverage is incomplete

Time:

$$\begin{aligned} level =& 0.02lc + 0.2cg + 0.25na + 0.05dk \\ &+ 0.5dq + dc \end{aligned}$$

- lc: The length of COT
- cg: Difficulty of the most challenging statement
- na: Number of options
- dk: Difficulty of knowledge
- dc: Difficulty of scenario, $dc = 0.5$ for cyclic scenarios, 0 otherwise
- dq: Difficulty of core question

Social:

$$level = 0.4cl + 0.3nm + 0.3ne$$

- cl: Reasoning chain length
- nm: Whether the entities in the question are explicitly mentioned in the text, $nm = 1$ when entities are not explicitly mentioned
- ne: Relationship edges in network

Nature:

$$level = 0.4kl + 0.3cl + 0.5nm$$

- kl: The sum of property difficulty
- cl: Reasoning chain length
- nm: Whether the entity in the question is mentioned in the text, $nm = 1$ when entities are not mentioned

Mix:

$$level = 0.4level_{\text{nature}} + level_{\text{space}} + \frac{nm_q + nm_a}{2}$$

- $level_{\text{space}}$: Space difficulty, $level_{\text{space}} = 1$ for vague spatial templates
- $level_{\text{nature}}$: Nature difficulty, equal to kl in nature domain
- nm_q/nm_a: Whether all entities in the question/answer are explicitly mentioned in the text, $nm_q/nm_a = 1$ when at least of entity in the question/answer are implicit

F Experiments Setting

All models are tested using their official APIs. For models that allowed temperature adjustment, we set the temperature to 0.7. The English and Chinese prompts used in the evaluations are shown in Figure 6.

G Answer Extractor

In this test, we extract model-generated answers using a rule-based method. Since all questions are multiple-choice and we do not adopt the ICL (Dong et al., 2024) approach. As a result, despite explicitly specifying the answer format in the prompt, some responses could not be extracted. However, the percentage of failed extractions is very low(which is shown in Table 8) and do not impact overall performance. In the future, we may consider using a LLM as the answer extractor to improve accuracy, though this would come at a higher computational cost. We also observe that some models experienced unexplained interruptions when calling the API, which may be related to issues with the website links. Given sufficient resources, we plan to conduct further tests in the future to ensure the stability and reliability of the evaluation results.

Chinese prompt

```
prompt = (
f"{text}\n\n问题：{question}\n\n"
选项：\n{choice_text}\n\n"
"题目均为不定项选择题。多选或漏选均不得分。\n"
"答案选项必须与标准答案完全一致才能得分。\n"
"请逐步思考，并最终将答案选项放在【】中。\n回答："
)
```

English prompt

```
prompt = (
f"{text}\n\nQuestion:
{question}\n\n"
Options:\n{choice_text}\n\n"
"All questions are multiple-choice with one or more correct answers.\n"
"No partial credit will be given for incorrect or incomplete answers.\n"
"Answer choices must exactly match the standard answer to be considered correct.\n"
"Please think step by step and finally place the answer choices in [].\nAnswer:"
)
```

Figure 6: The English and Chinese prompts used in the evaluations.

H Detailed Performance on Different Domains

In this section, we provide a detailed analysis of model performance across different domains and difficulty levels. As illustrated in Figures 7, 8, 9, 10, and 11, our difficulty classification system effectively differentiates question difficulty across most models, demonstrating its reliability in assessing model capabilities.

I Analysis of the Relationship Between Token Count and Question Difficulty

Intuitively, models with reasoning capabilities typically require longer reasoning chains for more challenging questions, resulting in increased token counts in their outputs. Our experimental data corroborates this hypothesis and further reveals that reasoning-specialized models exhibit a stronger correlation between question difficulty and reasoning length compared to general models.

Model Name	Unextracted Rate
o1-mini	0.0011
deepseek-chat	0.0476
gpt-4o	0.0150
deepseek-r1-distill-qwen-32b	0.0322
claude-3-5-sonnet-20241022	0.0144
o1-preview	0.0104
qwq-32B	0.0496
Llama-3.3-70B-Instruct	0.0080
qwen-25-72B	0.0078
qwen-max	0.0111
glm-zero-preview	0.0065
o3-mini	0.0183
deepseek-reasoner	0.0137
glm-4-plus	0.0031
Overall Average	0.0173

Table 8: Unsuccessful extraction rates of different models

We conducted experiments using both deepseek-reasoner and deepseek-chat models across datasets of varying difficulty levels, collecting response metadata including token length and answer correctness. For each dataset and model combination, we generated scatter plots illustrating token length versus question difficulty level, supplemented with line plots demonstrating the relationships between difficulty levels and three key metrics: average token count, average token count for correct responses, and average token count for incorrect responses. The result is shown in Figure 12.

The graphical analysis yields three principal findings:

(1) In deepseek-reasoner, within the same dataset, higher difficulty levels correspond to increased token counts in model outputs, indicating extended reasoning chains for more complex questions.

(2) The correlation between token count and difficulty level appears less pronounced in deepseek-chat, suggesting that reasoning-optimized models demonstrate superior capability in dynamically adjusting their cognitive processes according to the question complexity.

(3) Notably, in deepseek-reasoner, correct re-

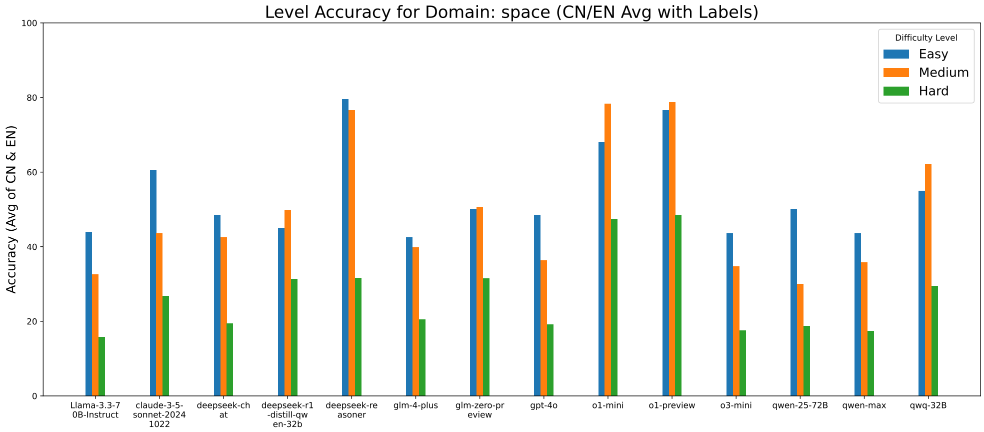


Figure 7: Level Accuracy on Space Domain

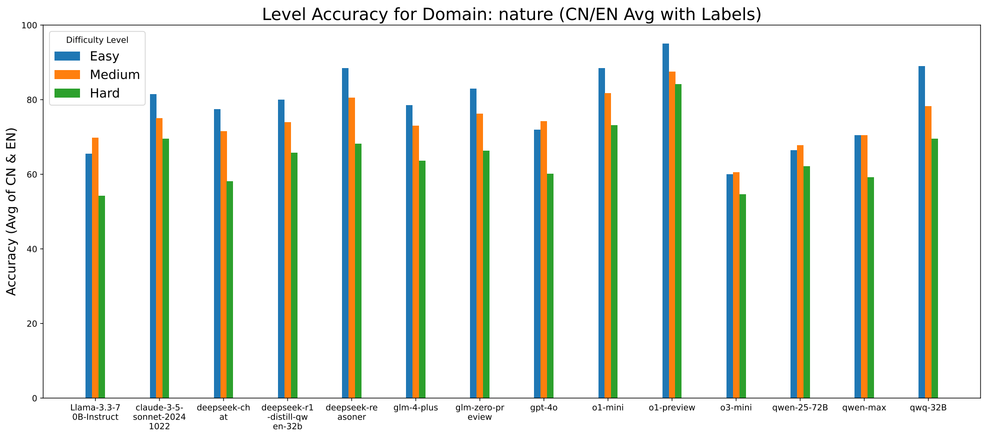


Figure 8: Level Accuracy on Nature Domain

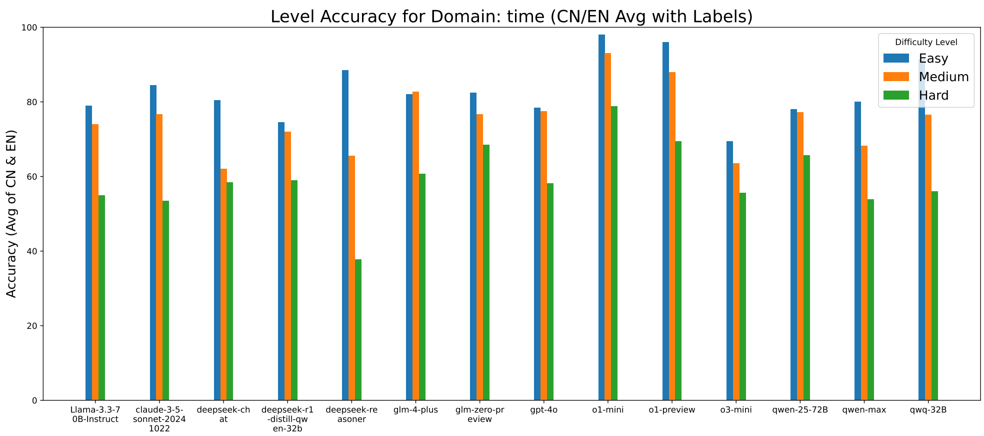


Figure 9: Level Accuracy on Time Domain

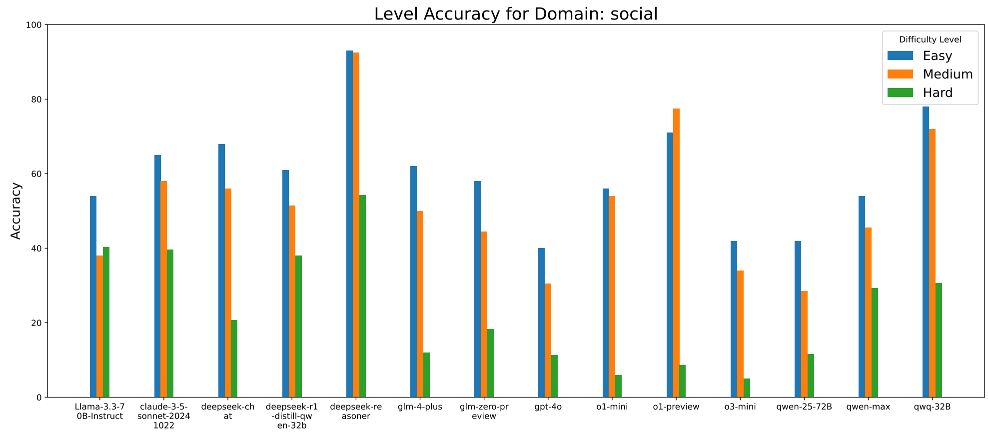


Figure 10: Level Accuracy on Social Domain

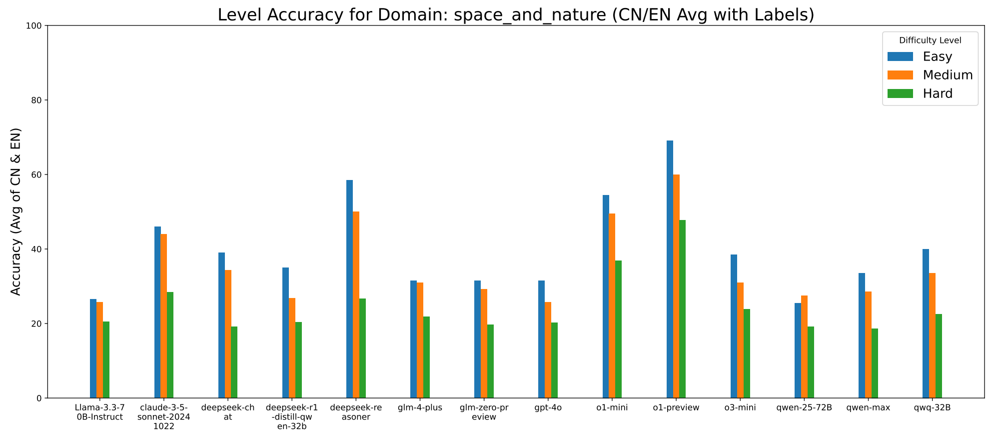


Figure 11: Level Accuracy on Mix Domain

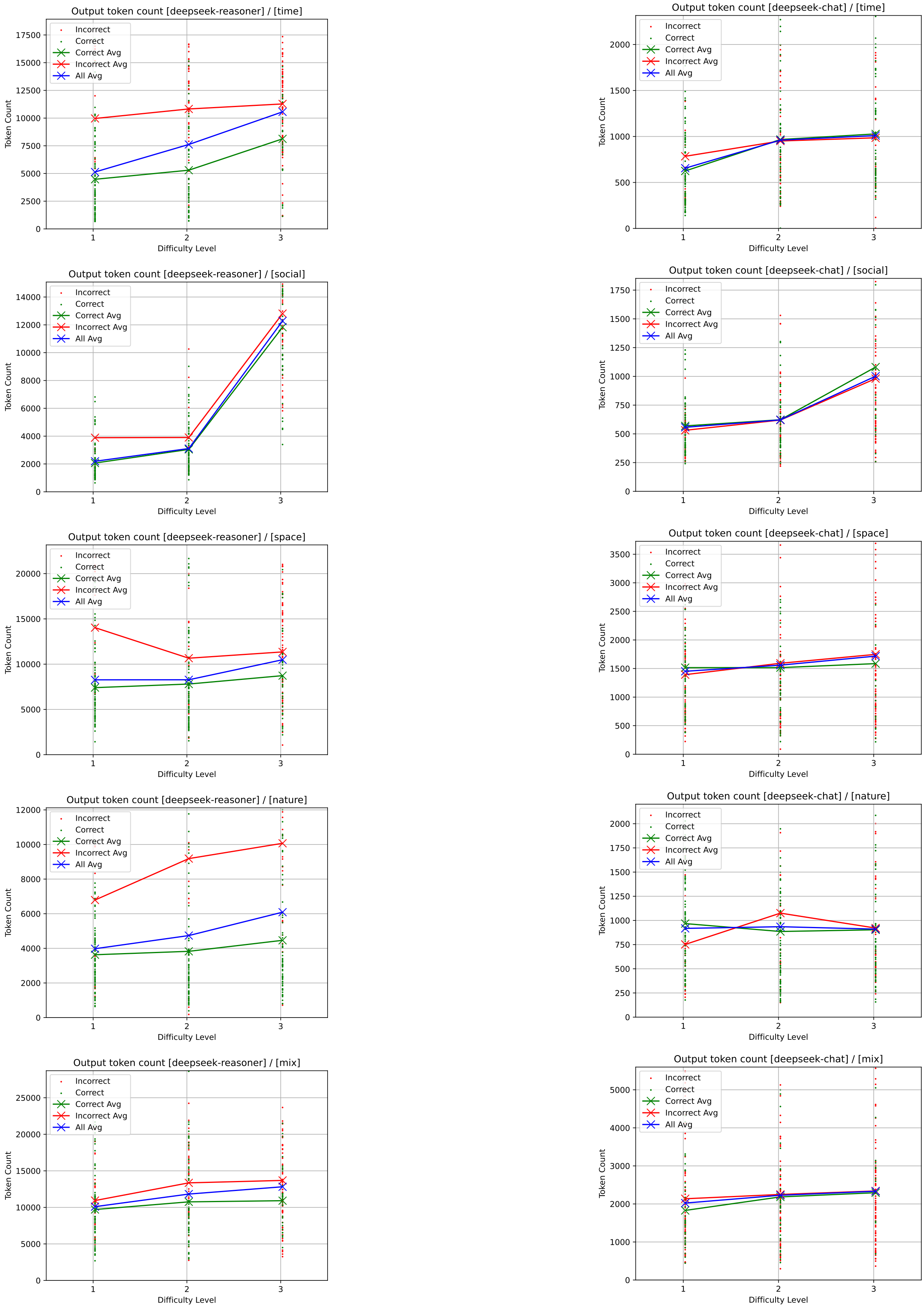


Figure 12: Relationship between response token count and question difficulty level. Left column: Results from deepseek-reasoner; Right column: Results from deepseek-chat. From top to bottom: Time, Social, Space, Nature, and Mix datasets (X-axis: Difficulty level, Y-axis: Token count).

sponses exhibit shorter average token lengths compared to incorrect ones within the same difficulty level. We hypothesize this phenomenon arises from two factors: (a) inherent variability in question difficulty within each difficulty level, and (b) an observable trend where more challenging questions (requiring longer reasoning chains) demonstrate higher error probabilities despite increased processing effort.

J Case Study

J.1 Commonsense Error

J.1.1 Entity Property Error

Question-EN

The four enclosures in the zoo keep four different kinds of animals: cat, mandarin fish, tortoise, snuffer. Now we know that:
The animal in enclosure No.2 and the animal in enclosure No.3 have 8 legs in total;
The animal in enclosure No.3 has a shell;
The animal in enclosure No.1 lives in sea water.
____ is kept in enclosure No.4.

Question-CN

动物园里的四个场馆分别养着猫、桂鱼、王八、海豚四种动物。
已知：
2号场馆中的动物与3号场馆中的动物加起来有8条腿；
3号场馆中的动物有壳；
1号场馆中的动物生活在海水中。
4号场馆中养的是____。

Options-EN

A. cat
B. mandarin fish
C. tortoise
D. snuffer

Options-CN

A. 猫
B. 桂鱼
C. 王八
D. 海豚

Correct Answer

B

Model Name

o1-preview

Model Answer-EN

Let's analyze the given information step by step.
Animals and Their Characteristics:
... [text omitted]...
Mandarin Fish - Legs: 0 - Shell: No - Habitat: Lives in seawater
... [text omitted]...
Answer: [D]

Model Answer-CN

根据题目提供的条件，我们一步一步推理：
(According to the information in the question, we analyze step by step)
分析每个动物的特征：
(Analyze the characteristics of animals)
... [text omitted]...
桂鱼：没有腿，没有壳，生活在淡水。
(Mandarin fish: no legs, no shell, lives in freshwater.)
... [text omitted]...
答案：【B】
(Answer:[B])

Figure 13: Example of entity property error in the model's response.

We found that the model still lacks understanding of some low-frequency entity properties. As shown in Figure 13, in the English question, the model mistakenly classified the mandarin fish (a freshwater fish) as a saltwater fish, leading to subsequent reasoning errors.

It is worth noting that such entity property errors are language-dependent. When the same question was asked in Chinese, the model correctly identified the mandarin fish as a freshwater fish and made the correct inference. We speculate that this is related to the frequency with which the entity appears in different language corpora. The mandarin fish is primarily found in East Asia and North Asia, where it is more frequently mentioned in Chinese corpora, while it is rarer in regions where English is the primary language. The lower occurrence frequency in English-language corpora likely leads to the model's misunderstanding of the mandarin fish's properties in an English context.

J.1.2 Entity Relation Error

Question

已知：李晓静是吴强的前妻，也是赵伟的前女友。孙大伟是吴强的挚友，也是钱静的老公。孙大伟是李晓静的领导。赵伟是钱静的同级生。以下选项正确的是___
(Given: Li Xiaojing is Wu Qiang's ex-wife and also Zhao Wei's ex-girlfriend. Sun Dawei is Wu Qiang's close friend and also Qian Jing's husband. Sun Dawei is Li Xiaojing's supervisor. Zhao Wei is Qian Jing's fellow student. The correct option among the following is ___.)

Options

A. 李晓静的前男友是孙大伟的老婆的同事
(A. Li Xiaojing's ex-boyfriend is the colleague of Sun Dawei's wife.)
B. 钱静的老公是赵伟的前女友的徒弟
(B. Qian Jing's husband is the apprentice of Zhao Wei's ex-girlfriend.)
C. 赵伟的前女友是孙大伟的挚友的前妻
(C. Zhao Wei's ex-girlfriend is the ex-wife of Sun Dawei's close friend.)
D. 赵伟的同级生是李晓静的领导的女朋友
(D. Zhao Wei's fellow student is the girlfriend of Li Xiaojing's supervisor.)

Correct Answer

C

Model Name

GPT-4o

Model Answer

我们来逐一分析每个选项：
(Let us analyze each option one by one:)
... [text omitted]...
A: 李晓静的前男友是孙大伟的老婆的同事。
(A. Li Xiaojing's ex-boyfriend is the colleague of Sun Dawei's wife.)
- 李晓静的前男友是赵伟。
- 孙大伟的老婆是钱静。
- 钱静是赵伟的同级生，因此也是赵伟的同事。
- 所以，李晓静的前男友（赵伟）是孙大伟的老婆（钱静）的同事。
- 该选项正确。
(- Li Xiaojing's ex-boyfriend is Zhao Wei.
- Sun Dawei's wife is Qian Jing.
- Qian Jing is Zhao Wei's classmate,Therefore, he is also Zhao Wei's colleague.
- Thus, Li Xiaojing's ex-boyfriend (Zhao Wei) is a colleague of Sun Dawei's wife (Qian Jing).
- This option is correct.)
... [text omitted]...
所以，正确的答案是【A, C】。
(So, the correct answers are 【A, C】.)

Figure 14: Example of confusing different relationship in the model's response.

We find that the model often confuses closely related social relationships. As shown in Figure

Question

已知：吴强是李晓静的同级生，也是赵伟的挚友。孙大伟是吴强的数位挚友中的一位，也是钱静的前男友。孙大伟是李晓静的师父。赵伟是钱静的领导。以下选项正确的是___
(Given: Wu Qiang is a fellow student of Li Xiaojing and also a close friend of Zhao Wei. Sun Dawei is one of Wu Qiang's several close friends and the ex-boyfriend of Qian Jing. Sun Dawei is Li Xiaojing's mentor. Zhao Wei is Qian Jing's supervisor. The correct option among the following is ___.)

Options

A. 钱静的领导是孙大伟的挚友的哥哥
(A. Qian Jing's supervisor is the elder brother of Sun Dawei's close friend.)
B. 赵伟的下属是李晓静的师父的前女友
(B. Zhao Wei's subordinate is the ex-girlfriend of Li Xiaojing's mentor.)
C. 钱静的前男友是吴强的同级生的徒弟
(C. Qian Jing's ex-boyfriend is the apprentice of Wu Qiang's fellow student.)
D. 吴强的某位名字中带有伟的挚友是孙大伟的前女友的下属
(D. One of Wu Qiang's close friends whose name contains the character "Wei" is the subordinate of Sun Dawei's ex-girlfriend.)

Correct Answer

B

Model Name

GPT-4o

Model Answer

为了确定正确的答案，我们需要一步一步地进行逻辑推理，理清各个角色及其关系。
(To determine the correct answer, we need to proceed with logical reasoning step by step, clarifying the roles and relationships of each individual.)
... [text omitted]...
C: 钱静的前男友是吴强的同级生的徒弟
(C. Qian Jing's ex-boyfriend is the apprentice of Wu Qiang's fellow student.)
根据信息，钱静的前男友是孙大伟，而孙大伟同时也是李晓静的师父，因此孙大伟是吴强的同级生（李晓静）的师父（徒弟），此项正确。
(According to the information, Qian Jing's ex-boyfriend is Sun Dawei, who is also Li Xiaojing's mentor. Therefore, Sun Dawei is the mentor of Wu Qiang's fellow student (Li Xiaojing)(apprentice), this item is correct.)
... [text omitted]...
因此，正确的选项是【B, C】。
(Therefore, the correct options are 【B, C】.)

Figure 15: Example of misinterpreting asymmetric relationships in the model's response.

Question

Flute, conch, fresh kidney bean, mess tin, duckweed, envelope, six items are placed on a three-tier shelf, which is positioned against the south wall of the store. Each tier has two sections, with one type of item placed in the east section and one in the west section. A customer is standing in front of the shelf. When describing the positional relationships of the items, it is agreed that the customer's own left and right will be used as a reference, with the east section being on the left and the west section being on the right. It is known that:
The envelope is located in the upper left corner of the flower or grass and the two are not separated by a layer,
The flower or grass and the animal with no legs are vertically adjacent to each other,
The musical instrument is on the same level as the vegetable,
The item with shell is on the left side of the bottom tier,
The right side neighbor of the musical instrument is above the upper right side of the envelope.
The flute is to the left of ___

Options

A. mess tin
B. conch
C. evnelope
D. None of the above

Correct Answer

D

Model Name

GPT-4o

Model Answer

... [text omitted]...
5. The right side neighbor of the musical instrument is above the upper right side of the envelope.
... [text omitted]...
Since clue 5 states that the right neighbor of the flute (musical instrument) is above the envelope's upper right, this implies that they must be on different tiers, and the flute is above the envelope.
... [text omitted]...
Given clue 5, if envelope is on the top left, the musical instrument (flute) must be in the middle tier.
... [text omitted]...
Thus answering the question, "Flute is to the left of ___", 、flute is to the left of the fresh kidney bean.
Answer: [A: mess tin]

Figure 16: Example of logical contradiction in the model's response.

14, the model correctly handles most reasoning steps, but it treats "classmate" and "colleague" as the same, confusing these relationships. It also confuses "mentor" and "leader," as well as "ex-husband" and "ex-boyfriend."

Additionally, the model misinterprets asymmetric relationships. As shown in Figure 15, it treats "A is B's mentor" and "A is B's apprentice" as identical, failing to understand the mentor-apprentice relationship. Similar errors occur with "A is B's subordinate" and "B is A's subordinate."

J.2 Reasoning Error

J.2.1 Logical Contradiction

We notice that the model's response may sometimes be self-contrast. As shown in Figure 16, the model infers rightly that the flute is above the envelope based on clue 5. However, in the following analysis, the model assumes that the envelope is on the top tier and the flute is on the middle tier based on the same clue, which contrasts the previous analysis. Additionally, the model also exhibit contradiction when summarizing the answer. According to the model's analysis, the flute is to the left of the fresh kidney bean, but in the end, it still outputs the answer as "the flute is to the left of the mess tin". This implies that LLMs have not yet possess logical reasoning abilities aligned with humans.

J.2.2 Imprecision

In the reasoning process, the model may be affected by the preceding output, leading to errors in reasoning. As shown in Figure 17, when calculating the time of "Jack goes jogging", the information "Wednesday" in step 2, instead of "Monday", is used. Thus, the model makes a mistake in reasoning. According to the text of the questions, this imprecision could be caused by (1) a wrong understanding of the entity or (2) a wrong understanding of the property. Since the occurrence of this error

Question

Jack is a college student, and here are his weekly plans:
(1)On Monday, Jack learns Japanese;
(2)On Wednesday, Jack plays badminton;
(3)Jack goes jogging 2 days after he learns Japanese;
(4)Jack cleans his dormitory room 1 day after he goes jogging;
(5)Jack practices the guitar 1 day after he cleans his dormitory room;
(6)1 day after Jack practices the guitar, he reads research papers.
5 days before ____, Jack plays badminton.

Options

A. Jack learns Japanese
B. Jack practices the guitar
C. Jack reads research papers
D. Jack goes jogging

Correct Answer

A

Model Name

DeepSeek-Chat

Model Answer

Let's break this down step by step:
1. **Monday**: Jack learns Japanese.
2. **Wednesday**: Jack plays badminton.
3. **2 days after learning Japanese (Wednesday + 2 days = Friday)**: Jack goes jogging.(In fact, Jack goes jogging on Wednesday)
... [text omitted]...
Jack plays badminton on **Wednesday**.
5 days before Wednesday is **Friday**.
From the schedule, we know that on Friday, Jack **goes jogging**.
Thus, the correct option is: **D: Jack goes jogging**.
The answer is [D].

Figure 17: Example of imprecision in the model's response.

is strongly related to next token prediction, which is regarded as the cornerstone of LLMs, examples can also be found in other fields.

J.2.3 Inference Error

As shown in figure 18, the question states that six entities are arranged in a hexagonal layout, creating a ring where the n-th position on the left is equivalent to the $(6 - n)$-th position on the right. Considering the angular relationships of the hexagon's vertices, it is the third position on the left or the third on the right that is directly opposite. Such an error indicates that the model did not correctly understand the configuration of entities based on the scenario described in the question, leading to incorrect reasoning about their relative positions. This type of error are related to the model's ability to construct and understand scenario.

J.3 Overthinking

We find that some reasoning-specialized LLMs, such as the o1 and r1 series, sometimes overthink and introduce special cases to fit the given conditions. As shown in Figure 19, the condition "The flower of the item on photo No.3 is white" clearly points to coriander in photo No.3. However, o1 suggests that a birdcage can be decorated with white flowers, making it a possible answer.

Question

张天师(David)、曹国舅(Elizabeth)、张果老(Mary)、何仙姑(Jennifer)、汉钟离(William)、吕洞宾(Robert)六座神像在神坛中围成一个圆圈，每座神像都背对神坛中心。六个神像的位置恰好落在正六边形的六个顶点上。任意相邻两个神像之间的距离相等，大约为一米。已知:
张天师的左边起第五个是吕洞宾;
曹国舅在何仙姑左侧紧邻位置;
从吕洞宾的左边数起第三个位置是张果老;
吕洞宾在汉钟离右边数起第二个位置;
张果老在何仙姑右侧紧邻位置。
张天师紧邻___。
(David, Elizabeth, Mary, Jennifer, William, Robert, — these six statues are arranged in a circle on the altar, each facing away from the center of the altar. The positions of the six statues correspond precisely to the six vertices of a regular hexagon. The distance between any two adjacent statues is equal, approximately one meter. It is known that:
Robert is the fifth in sequence starting to the left of David;
Elizabeth is directly to the left of Jennifer;
Mary occupies the third position to the left of Robert;
Robert occupies the second position to the right of William;
Mary is directly to the right of Jennifer.
David is next to ___.)

Options

A. 张果老(Mary)
B. 何仙姑(Jennifer)
C. 吕洞宾(Robert)
D. 曹国舅(Elizabeth)

Correct Answer

B,C

Model Name

GPT-4o

Model Answer

根据题目提供的信息，我们可以推理出神像的排列顺序:
(Based on the information provided in the question, we can deduce the arrangement order of the statues:)
张天师的左边起第五个是吕洞宾：这意味着吕洞宾在张天师的对面。
(Robert is the fifth in sequence starting to the left of David, This means that Robert is opposite David.)
(In fact, in this scenario, the fifth position on the left is equivalent to the first position on the right, not directly opposite.)
... [text omitted]...
所以，张天师紧邻张果老。答案是：A. 张果老
(so, David is next to Mary. The answer is A. Mary)

Figure 18: Example of inference error in the model's response.

In contrast, general-purpose models like Llama3.3-70B-Instruct and qwen-max do not overthink and correctly answer the question. This implies that the training methods for reasoning-specialized models emphasizing logical consistency and depth of reasoning may lead them to miss more staightforword solutions.

J.4 Answer Aggregation Error

Despite mentioning that our questions are multiple-choice in the prompt, the models still tend to treat them as single-choice questions. As shown in Figure 20, the model outlines the right arrangement but stops checking once it finds one correct answer. This may be due to the high frequency of single-choice questions in inference data, which leads the models to learn a shortcut outputting the answer once they find one answer.

Additionally, logical contradiction may also con-

Question

On a wall pasted photos of four different items: leather jacket, carnation, birdcage, coriander. Now we know that:
The item on photo No.3 is not a clothing;
The item on photo No.2 is not a clothing;
The flower of the item on photo No.3 is white;
The item on photo No.4 is not a tool;
The item on photo No.4 is not a clothing;
The item on photo No.3 is not a flower or grass.
Carnation is on photo No.____.

Options

A. 1
B. 2
C. 3
D. 4

Correct Answer

C

Model-1 Name

o1-preview

Model-1 Answer

Given Constraints:
... [text omitted]...
3. The flower of the item on photo No.3 is white.
... [text omitted]...
First, let's list out the possible categories for each item:
Birdcage: Tool
Coriander: Grass/Herb (has small white flowers)
... [text omitted]...
Perhaps the birdcage is decorated with white flowers or has a design featuring white flowers.
... [text omitted]...
Therefore, the item on photo No.3 must be the birdcage.
... [text omitted]...
Answer: [B]

Model-2 Name

Llama3.3-70B-Instruct

Model-2 Answer

given the items listed (leather jacket, carnation, birdcage, coriander), coriander has small white flowers, and it's not a clothing, tool, or exclusively a flower in the context of these clues
... [text omitted]...
The birdcage, not being clothing, could fit on No.2, given that No.3 has a specific description that doesn't fit the birdcage directly
... [text omitted]...
The answer is [D]

Figure 19: Example of overthinking in the model's response.

tribute to answer aggregation errors. The model may analyse A is right but answer B, as shown in Figure 16.

Question

Rolling luggage, eyebrow pencil, plaice, Chinese herbaceous peony, turkey, port wine, six items are placed on a three-tier shelf, which is positioned against the south wall of the store. Each tier has two sections, with one type of item placed in the east section and one in the west section. A customer is standing in front of the shelf. When describing the positional relationships of the items, it is agreed that the customer's own left and right will be used as a reference, with the east section being on the left and the west section being on the right. It is known that:
The animal with no legs is directly to the eyebrow pencil's left,
The rolling luggage and the eyebrow pencil are separated by one tier,
The bird is on the first floor,
The flower or grass is above the animal with 2 legs in a straight line,
The right side neighbor of the animal with no legs is above the upper right side of the animal with 2 legs,
The drink is somewhere to the right of the flower or grass,
The bird is located at the bottom left of the eyebrow pencil and there is one tier separating the two,
The rolling luggage and the drink are vertically adjacent to each other.
port wine and ___ are vertically adjacent to each other

Options

A. eyebrow pencil
B. turkey
C. rolling luggage
D. None of the above

Correct Answer

AC

Model Name

o1-preview

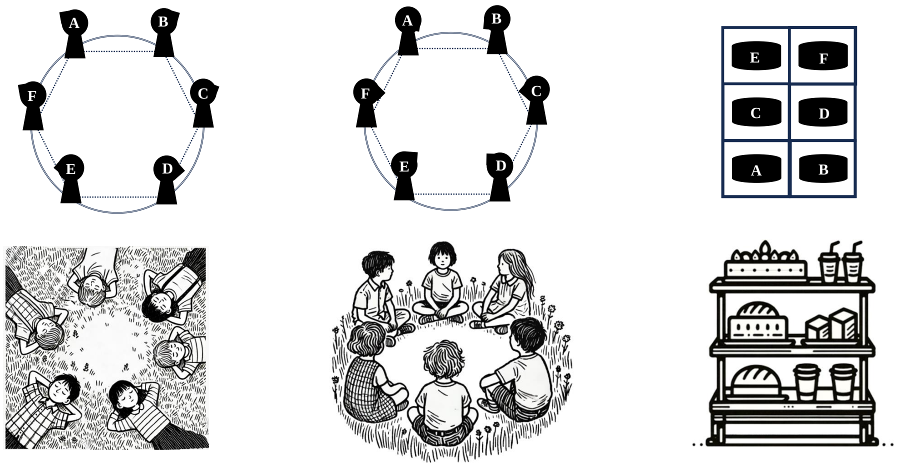
Model Answer

... [text omitted]...
- Top Floor (3rd Tier):
- Left (East): Plaice (animal with no legs) - Right (West): Eyebrow pencil
- Middle Floor (2nd Tier): - Left (East): Chinese herbaceous peony (flower) - Right (West): Port wine (drink)
- Bottom Floor (1st Tier): - Left (East): Turkey (bird, animal with 2 legs)
- Right (West): Rolling luggage
Answering the question: Port wine and rolling luggage are vertically adjacent to each other. So, the correct choice is C.
*Answer:** [C]

Figure 20: Example of answer aggregation error in the model's response.

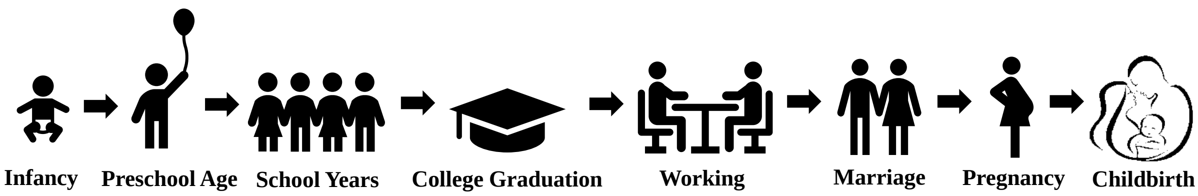
K Scenario Diagram

The space scenario diagram is shown as Figure 21. The time scenario diagram is shown as Figure 22. The social scenario diagram is shown as Figure 23. The enhanced spatial scenario diagram is shown as Figure 24.

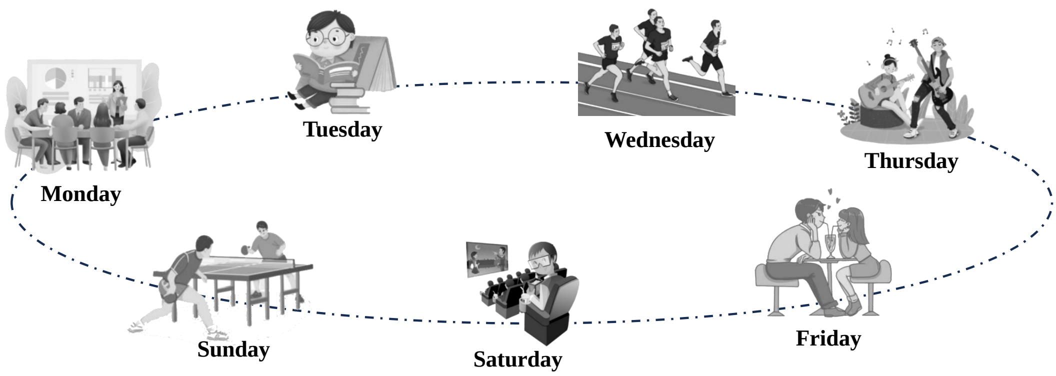


(a) Centrifugal Hexagon Scenario (b) Centripetal Hexagon Scenario (c) Two Rows Three Columns Scenario

Figure 21: Spatial Scenario Diagram

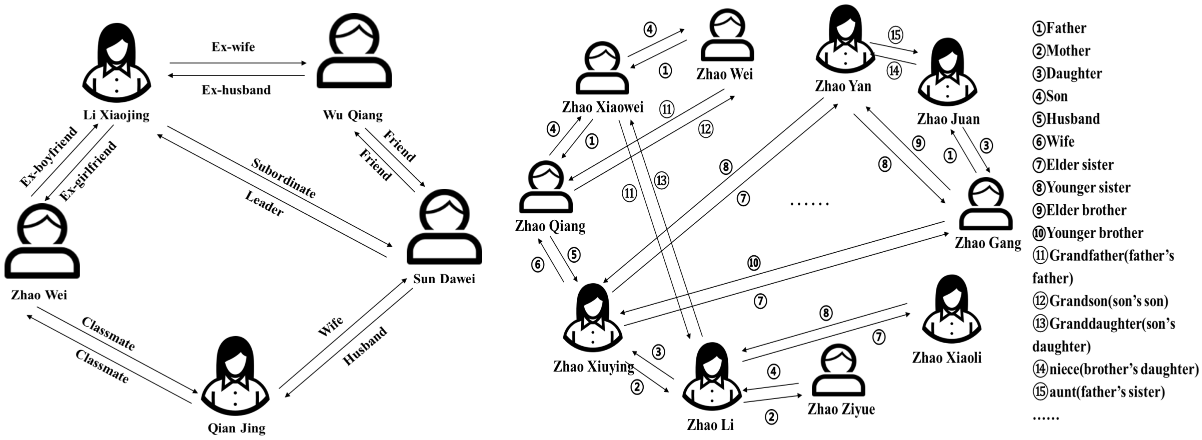


(a) Linear Scenario



(b) Cyclic Scenario

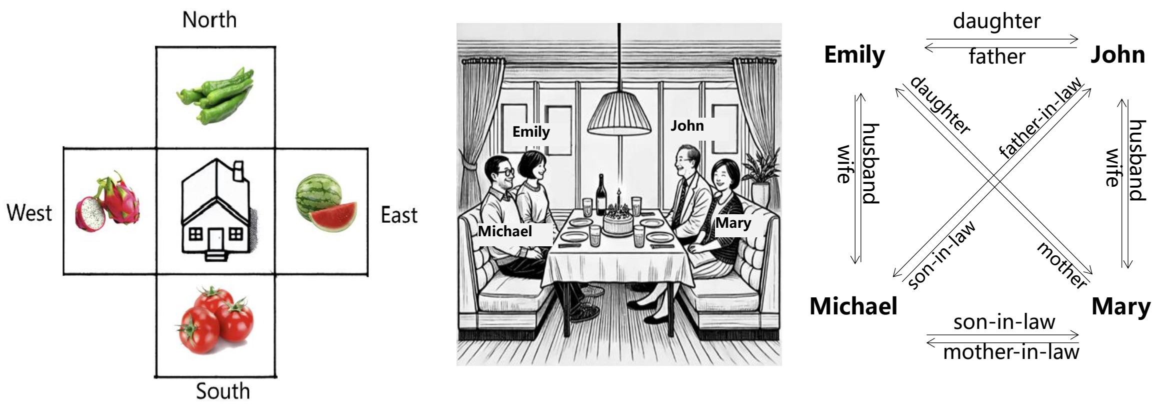
Figure 22: Time Scenario Diagram



(a) General Social Relationship

(b) Chinese Family Relationship

Figure 23: Social Scenario Diagram



(a) Four Plots Farmland Scenario

(b) Family In Four-person Booth Scenario

Figure 24: Enhanced Spatial Scenario Diagram with Entity Properties and Relationships