# SCoRE: Benchmarking Long-Chain Reasoning in Commonsense Scenarios

**Weidong Zhan[1], Yue Wang[2], Nan Hu[1], Liming Xiao[1], Jingyuan Ma[2], Yuhang Qin[1],**
**Zheng Li[2], Yixin Yang[2], Sirui Deng[1], Jinkun Ding[1], Qingxiu Dong[2], Wenhan Ma[2], Rui Li[2],**
**Weilin Luo[3], Qun Liu[3], Zhifang Sui[2]***
[1]Center for Chinese Linguistics, Department of Chinese Language and Literature, Peking University
[2]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[3]Huawei Noah's Ark Lab, China
`szf@pku.edu.cn, zwd@pku.edu.cn`

## Abstract

Though large language models (LLMs) have achieved remarkable performance across various commonsense question-answering benchmarks, effective evaluations of their capability to accurately utilize commonsense knowledge in complex reasoning scenarios remain lacking. Existing benchmarks suffer from limitations such as narrow coverage, high construction costs, and overly elementary reasoning trajectories. To address these gaps, we introduce SCoRE (Scenario-based Commonsense Reasoning Evaluation), a benchmark that synthesizes multi-hop questions from scenario schemas of entities, relations, and logical rules to assess long-chain commonsense reasoning. SCoRE contains 100k bilingual multiple-choice questions whose reasoning chains span 1–30 hops and are grouped into various difficulty levels. Each question is accompanied by fine-grained knowledge labels, explicit reasoning chains, and difficulty levels for diagnostic evaluation. Evaluation results on cutting-edge LLMs show that even the best model attains only 69.78% accuracy on SCoRE while models on average attained merely 47.91% accuracy on the hard set, with errors often stemming from rare knowledge, logical inconsistency, and over-interpretation of simple questions. Additionally, SFT experiments indicate SCoRE can also help improve the model performance in both in-domain and out-of-domain evaluations. This proves SCoRE is a scalable, extensible framework, not only for evaluating and diagnosing the long-chain commonsense reasoning abilities of LLMs but also for guiding future advances in model design and training.

**Datasets** — https://github.com/pokerwf/KnowLogic

## 1 Introduction

Reasoning is a high cognitive function that involves analyzing, inducting, and deducing new information based on existing knowledge (Bommasani et al. 2021). It plays a fundamental role in human intelligence. While previous benchmarks have predominantly focused on evaluating LLMs' ability to reason within complex, specialized domains, they have often overlooked a critical aspect of human-like cognition: commonsense reasoning. Evaluating this commonsense reasoning ability in large language models (LLMs) is therefore crucial for the development of AI. This fundamental capability significantly influences LLMs' decision-making in everyday situations and is vital for advancing towards human-like intelligence in artificial general intelligence (AGI).

While commonsense knowledge and reasoning are both indispensable for human-like intelligence, current commonsense reasoning benchmarks mostly exhibit a systemic bias toward knowledge memorization over structured reasoning. 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, Kamigaito, and Watanabe 2024; Sprague et al. 2024) for data generation. Human annotated datasets, such as CommonsenseQA (Talmor et al. 2019), CommonsenseQA 2.0 (Talmor et al. 2022), and ProtoQA (Boratko et al. 2020), guide crowdworkers to generate questions associated with given knowledge. However, the questions are mostly solvable through direct knowledge retrieval or single-step reasoning rather than multi-step reasoning, which appears to be too simple for large reasoning models. Moreover, the high cost of manual annotation makes it difficult to build large-scale datasets. Datasets based on template rules, such as bAbI (Weston et al. 2015) and TRAM (Wang and Zhao 2023), connect commonsense knowledge with explicit logical rules, but their reasoning patterns are strictly confined to predefined templates, exhibiting limited diversity and short reasoning chains. Though some LLM-generated datasets, such as MuSR (Sprague et al. 2024), can combine high-density knowledge with multi-step reasoning, their data quality are not well guaranteed. 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 highlight these issues, we introduce **SCoRE**, a diagnostic benchmark designed to evaluate complex logical reasoning in commonsense scenarios through knowledge-driven synthetic data generation. SCoRE contains 100,000 bilingual (Chinese and English) questions focusing on four domains of commonsense closely related to everyday human life: space, time, social, nature, along with a mix domain for questions with multiple ”domain” labels with adjustable

---

*Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

| **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.<br>Who is **diagonally opposite** John?<br>A. David B. Jennifer C. James D. None of the above |
| 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.<br>What will Jack do **4 days after** he cleans his dormitory room?<br>A. Learning Japanese B. Having a group meeting C. Playing Badminton D. None of the above |
| 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.<br>Q: What is the relationship between Dave's **wife** and Alice's **ex-boyfriend**?<br>A. **Friend** B. **Classmate** C. **Ex-wife** D. **Ex-girlfriend** |
| 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**.<br>What is kept in enclosure **No.1**?<br>A. Carp B. Duck C. Turkey D. Fox |
| Mix | Watermelon, pitaya, green pepper, and tomato are planted on the east, west, north, and south sides of a farmhouse. Two kinds of **fruits** are **on the opposite side**. The plants in the **north and east farms have the same color**. The plant in the **west farm has black seeds**.<br>Q: What is planted in **the east farm**?<br>A. Watermelon B. Pitaya C. Green pepper D. Tomato |

Table 1: Four domains of commonsense and reasoning cases.

reasoning chains up to 30 hops, emphasizing the importance of complex reasoning. Table 1 shows the cases of commonsense reasoning in scenarios of different domains. The questions are generated by a knowledge-driven synthetic data strategy, which involves manually creating a reliable knowledge base that integrates diverse commonsense knowledge and scenarios to ensure data quality and designing an automatic inference engine capable of controlling the entire inference process and recording fine-grained features, to generate questions and answers. 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. This strategy can ensure the scale and accuracy of data with minimal labor costs.

The contributions of this paper are as follows:

1. We propose SCoRE, a bilingual benchmark specifically designed to evaluate complex logical reasoning in commonsense scenarios through multi-step inference chains. It consists of 100,000 commonsense reasoning questions integrating diverse commonsense knowledge, multiple scenarios and long reasoning chains in different difficulty levels.
2. We propose a knowledge-driven data synthesis method combining manual construction of a small-scale knowledge base and automatic question generation to ensure data quality and quantity at a minimal labor expense. This method has a transparent and traceable workflow that facilitates interpretable evaluations of LLMs.
3. We conducted evaluations using 13 state-of-the-art reasoning LLMs and an SFT experiment on a smaller model. The models achieved only 47.91% on the hard set. and the model fine-tuned on SCoRE achieve improvement in both in-domain and out-of-domain evaluations, indicating that SCoRE is not only challenging for the top-performing models but also valuable for future advances in model training.
4. Through case analysis, we identified several significant shortcomings in the commonsense reasoning capabilities of LLMs, including misunderstandings of low-frequency commonsense knowledge, self-contradiction in logic, and overthinking, among others.

## 2 Data Synthesis for SCoRE

The basic methodology for generating our reasoning data revolves constructing knowledge base and ultilizing Inference Engine to generate questions. The core knowledge base is relatively small in scale, and to ensure accuracy, it primarily leverages existing knowledge bases combined with manual construction. The reasoning engine is developed under strict logical rules, which—given the accuracy of the knowledge base—guarantees the precision of the automatically synthesized data.

### 2.1 Scenario-Centered Knowledge Base Construction

SCoRE relies on a knowledge base that includes scenarios, entities, propositions to express properties and relations, and

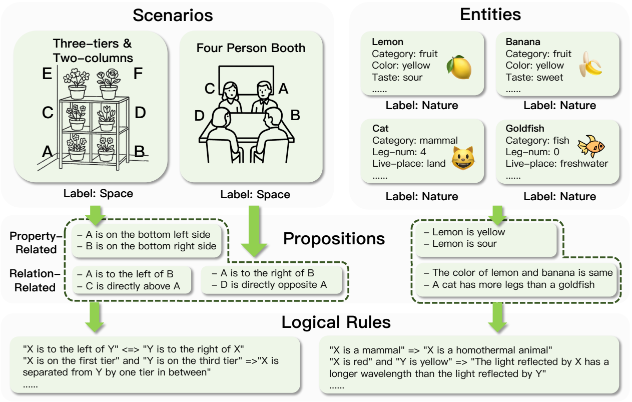


Figure 1: The framework of the knowledge base. Illustrated with scenarios in the space domain and entities in the nature domain.

logical rules for reasoning. The overall framework of the knowledge base is presented in Figure 1.

Scenarios are the core of our knowledge base. We believe that commonsense reasoning requires an understanding of the specific scenarios in which additional implied information must be inferred to fully grasp the context. Each scenario has designated slots for placing entities, and these slots impose selection constraints on entities. For example, in a ”four-person booth” scenario, there should be four entities, each of which should be a person. The scenario also constrains the way we describe the relations of entities; for example, when describing the spatial relations in a ”three-tier and two-column” scenario, we may use left, right, above, and below, but we would not describe the front and back of entities. Scenarios are constructed by manually designing a minimum set of propositions based on the properties and basic relations, and a set of logical rules to infer all relations. Our knowledge base contains 11 scenarios in total. Appendix H list scenario diagrams used in SCoRE.

Entities are candidates to be filled into the scenarios, which also have multiple properties and relations described by propositions and logical rules. Each scenario and entitiy is annotated with a ”domain” label, and each proposition is labelled with its domains, related entities, and relations for fine-grained analysis. The entities and their properties are extracted from existing external resources such as HowNet (Dong and Dong 2003), ConceptNet (Speer, Chin, and Havasi 2017) and Wiki. Based on them, we manually summarize logical rules associated with the properties or relationships of the entities.

The current version of the knowledge base organizes entities, relations between entities, and factual propositions into structured scenarios representing core knowledge across spatial, temporal, natural, and social domains. At this stage, the knowledge base includes 11 scenarios, 707 entities, and 939 rules associated with 29 properties and 109 relations. It is designed to be extensible by adding new scenarios, rules and entities, and will be continuously expanded in future iterations. Details about its construction is shown in Appendix A

## 2.2 Knowledge-Driven Question generation

Once the knowledge base is completed, we can ultilize a knowledge-driven method to automatically generate numerous questions. Our Question generation is fully automated by the Inference Engine, which consists of three steps. The overall process of question generation is shown in Figure 2.

**Step 1: Scenario Definition** This first step establishes the scenario and enhances its presentation. It involves selecting an scenario $S$ with $n$ slots $\{s_1, s_2, ..., s_n\}$ and a set of entities $E = \{e_1, e_2, ..., e_n\}$ that meets the constrains of the scenario from the knowledge base, where entity $e_k$ is filled into slot $s_k$. 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. Once the scenario is defined, a rule set $R = \{r_1, r_2, ..., r_m\}$ that suits this scenario can be determined.

**Step 2: Inference Data Generation** The second step is utilizing a Reasoner to generate inference data. The Reasoner is a rule-based program that takes the relations associated with the scenario $S$ and all properties of the entities in $E$ as input.

It maintains a fact base $F$, initialized by the descriptions of all properties of all entities or events $F_0 = \{f_1, f_2, ..., f_p\}$, such as spatial properties ’strawberry is on the middle floor’, ’lemon is on the top floor’, natural properties ’lemon is sour’, ’strawberry is red’, etc. The program then automatically traverses the each rule in the rule set $S$, matches them with the initial facts to generates new facts. For example, from facts ’strawberry is on the middle floor’ and ’strawberry is red’ and the syllogism rule, it can generate a new fact ’a red item is on the middle floor’. A set of new facts $F_{new} = \{f_1^{'}, f_2^{'}, ...f_q^{'}\}$ is generated through this process. The program then takes the newly added facts in the fact base, along with the original facts, as the new initial facts $F_i = F_{i-1} \cup F_{new}$ and inputs them into the Reasoner again. This process is repeated until $F_i = F_{i-1}$, when no new facts 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 statement set $F_{stat}$ which contains 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 $f$ 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 domains, entities, properties and relations involved in the statements and the number of inference steps. This process is repeated until the answer is ’Yes’. An explicit long reasoning chain can be formed through this repeated stacking of fact base generation and statement selection. The procedure is provided in Algorithm 1.

**Step 3: Question Design** The third step is utilising a Question Generator to design the question. The Question

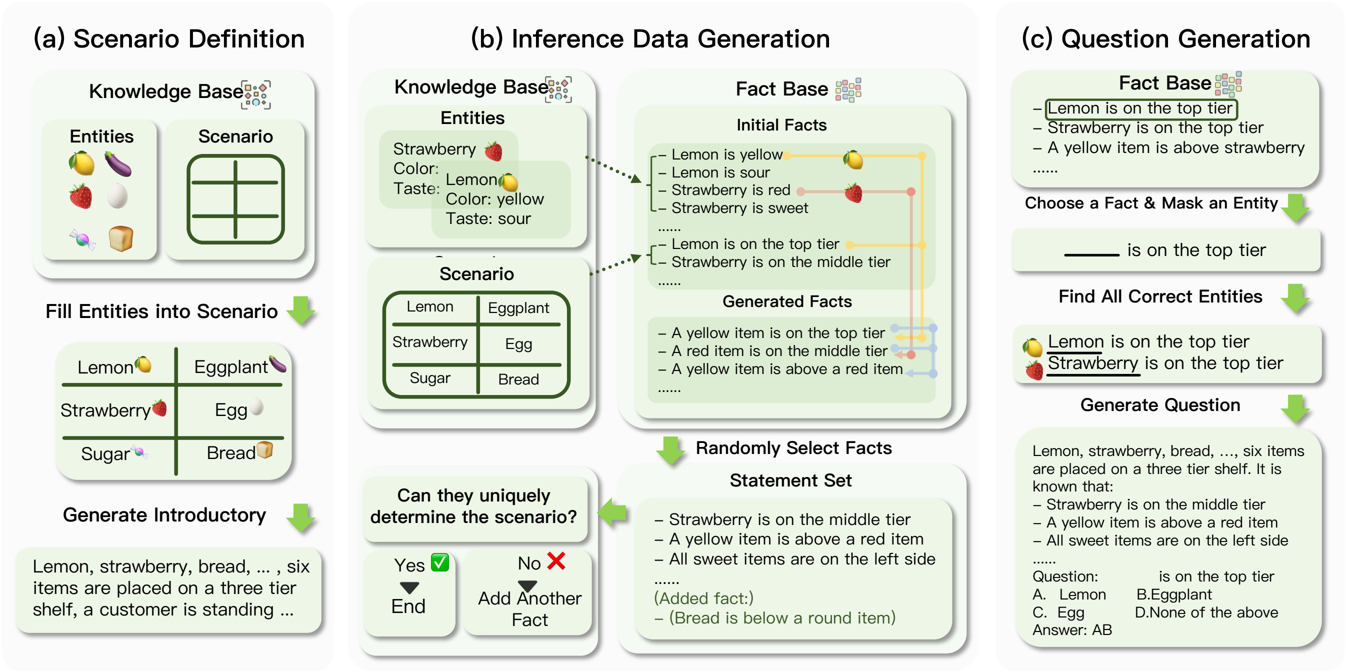

(a) Scenario Definition
Knowledge Base
Entities
Scenario
Fill Entities into Scenario
Lemon
Eggplant
Strawberry
Egg
Sugar
Bread
Generate Introductory
Lemon, strawberry, bread, ... , six items are placed on a three tier shelf, a customer is standing ...
(b) Inference Data Generation
Knowledge Base
Entities
Strawberry
Color:
Taste:
Lemon
Color: yellow
Taste: sour
Scenario
Lemon
Eggplant
Strawberry
Egg
Sugar
Bread
Fact Base
Initial Facts
– Lemon is yellow
– Lemon is sour
– Strawberry is red
– Strawberry is sweet
......
– Lemon is on the top tier
– Strawberry is on the middle tier
......
Generated Facts
– A yellow item is on the top tier
– A red item is on the middle tier
– A yellow item is above a red item
......
Randomly Select Facts
Statement Set
– Strawberry is on the middle tier
– A yellow item is above a red item
– All sweet items are on the left side
......
(Added fact:)
– (Bread is below a round item)
Can they uniquely determine the scenario?
Yes ✅
End
No ❌
Add Another Fact
(c) Question Generation
Fact Base
– Lemon is on the top tier
– Strawberry is on the top tier
– A yellow item is above strawberry
......
Choose a Fact & Mask an Entity
____ is on the top tier
Find All Correct Entities
Lemon is on the top tier
Strawberry is on the top tier
Generate Question
Lemon, strawberry, bread, ..., six items are placed on a three tier shelf. It is known that:
– Strawberry is on the middle tier
– A yellow item is above a red item
– All sweet items are on the left side
......
Question: ____ is on the top tier
A. Lemon B.Eggplant
C. Egg D.None of the above
Answer: AB

Figure 2: Overall process of data synthesis. The pipeline consists of three stages: **(1) Scenario Definition:** selecting entities/events from the knowledge base and constructing natural language descriptions; **(2) Inference Data Generation:** applying the Reasoner to iteratively generate a fact base via rule-based logic and select a minimal statement set that uniquely determines all entity positions; **(3) Question Design:** using the statement set and answer key to generate different question types (e.g., precise, vague, true/false) with appropriate options. This process ensures verifiable reasoning chains, rich logical structures, and natural language fluency.

Algorithm 1: Inference Data Generation

```
Input: Scenario $S$, rule set $R$, entities $E$, initial facts $F_0 = \{f_1, f_2, ..., f_p\}$
Output: Statement set $F_{stat}$
Step 1: Generate Fact Base
Initialize fact base $F \leftarrow F_0$
repeat
    $F_{\text{new}} \leftarrow \emptyset$
    for each rule $r \in R$ do
        Match $r$ with facts in $F$ to get new facts
        Add new facts to $F_{\text{new}}$
    end for
    $F \leftarrow F \cup F_{\text{new}}$
until $F$ does not change
Step 2: Select Statement Set
Initialize statement set $F_{\text{stat}} \leftarrow \emptyset$
repeat
    Randomly select a fact $f$ from $F \setminus F_{\text{stat}}$
    $F_{\text{stat}} \leftarrow F_{\text{stat}} \cup \{f\}$
    Verify if $F_{\text{stat}}$ uniquely determines the slots in $S$
until verification returns Yes
Return $F_{\text{stat}}$
```

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', 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', 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.

Since each step in this process undergoes rigorous reasoning, the correctness of the questions can be guaranteed as long as the knowledge and rules in the knowledge base are accurate. Combining automatic question generation with manually constructing a small-scaled knowledge base, we can ensure the quality and quantity of the questions at a minimal labor expense.

We conducted three rounds of human verification for the quality of questions, and the accuracy reach 100% after this process. Details about human verification are shown in Appendix B.

We illustrate the relationship between the knowledge base and its scalability in generating inference questions. Our data synthesis approach can automatically derive a small set of rules and initial facts into a large fact set, which can be further combined to generate numerous inference questions. The derivation process exhibits rapid combinatorial growth, as inference rules can be recursively applied across compatible fact combinations. Let the rule set $R$ denote the set of inference rules, and the fact set $F_0$ denote the small set of initial facts provided. Under typical conditions, the size of the derived fact set $|F|$ satisfies $|F| = O(|F_0||R|)$, assuming bounded rule composition complexity. However, when new rules are added, the reasoning space expands superlinearly. Specifically, adding $k$ new rules to the knowledge base increases the number of inferable facts to approximately $|F'| = O(|F|^k) = O(|F_0|^{|R|+k})$. As a result, even small extensions to the rule base lead to large-scale amplification of the reasoning space, enabling the automatic generation of millions of high-quality inference items with minimal human effort.

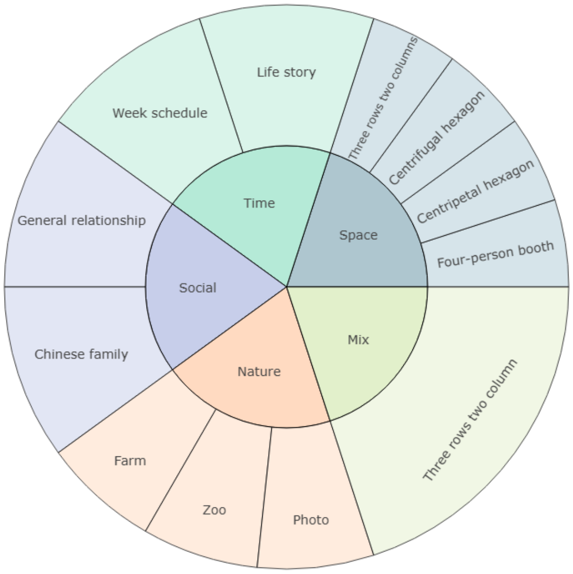


Figure 3: Distribution of questions across domains and scenarios in the SCoRE.

## 3 Data Statistics

The full dataset of SCoRE contains 100,000 questions generated by the above-mentioned method. Considering the computation cost, we intentionally selected 6,000 bilingual questions that can cover all knowledge points in our knowledge base to form the SCoRE test dataset and 5,000 English questions for the SCoRE training dataset.

**Knowledge Coverage** The questions cover four major commonsense domains: space, time, social, and nature, along with a mix domain for questions with multiple "domain" labels. The questions cover all knowledge points in the knowledge base, totaling 11 scenarios, 707 entities, 29 properties and 109 relations. Distribution of questions across domains and scenarios is shown in Figure 3.

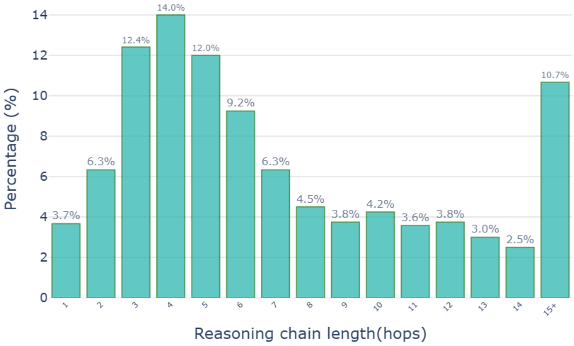


Figure 4: Distribution of reasoning chain length in SCoRE.

**Explicit Reasoning Chains** The reasoning steps are recorded during the data synthesis process. Our dataset includes reasoning questions ranging from 1 to 30 hops, with an average of 7.28 hops, enabling hierarchical evaluation of models' multi-hop reasoning capabilities. Distribution of reasoning chain length are shown in Figure 4

**Fine-grained Labels** Questions in SCoRE are generated by connecting domain knowledge via logical rules, ensuring a clear and traceable reasoning workflow. 7 kinds of labels are recorded during the reasoning process: domains, scenarios, entities, properties and relations used in the reasoning process, along with the reasoning chain length and question type.

**Difficulty Levels** SCoRE includes three difficulty levels: easy (level 1), medium (level 2) and hard (level 3). The difficulty level is calculated as:

$$level = KC + RC + QC$$

Where $KC$ denotes knowledge complexity, $RC$ denotes reasoning complexity, and $QC$ denotes question type complexity. We predefined the difficulty scores for each category of commonsense knowledge, reasoning rules, and question types. The overall difficulty of a question is the sum of the difficulties of all the knowledge, reasoning rules, and question types it involves. The distribution ratio of difficulty levels is 1:2:3.

## 4 Experiments and Analysis

### 4.1 Experimental Setup

Considering the representativeness and evaluation cost, 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),

| Model | Space | | Nature | | Time | | Social | | Mix | | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | CN | EN | CN | EN | CN | EN | CN | EN | CN | EN | |
| **Closed-Source Models** | | | | | | | | | | | |
| o1-preview | **67.17** | **58.83** | **89.83** | **84.33** | 79.67 | 80.67 | 49.67 | 71.67 | **61.17** | **54.83** | **69.78** |
| o1-mini | 62.00 | 56.50 | 82.00 | 75.33 | **88.00** | **85.67** | 33.83 | 56.67 | 48.00 | 45.33 | 63.33 |
| claude-3.5-sonnet | 45.33 | 43.17 | 76.00 | 70.83 | 60.50 | 71.67 | 56.83 | 68.67 | 36.33 | 36.17 | 56.55 |
| glm-zero-preview | 39.50 | 34.50 | 73.17 | 71.83 | 66.83 | 79.17 | 38.67 | 78.33 | 22.83 | 27.33 | 53.22 |
| glm-4-plus | 30.33 | 30.17 | 74.17 | 64.50 | 71.67 | 70.67 | 38.67 | 58.17 | 25.50 | 26.00 | 48.98 |
| gpt-4o | 29.50 | 29.00 | 68.17 | 65.67 | 65.50 | 69.83 | 25.17 | 45.00 | 23.67 | 26.83 | 44.83 |
| qwen-max | 15.83 | 13.17 | 65.83 | 64.00 | 54.17 | 71.67 | 47.67 | 56.33 | 24.83 | 26.00 | 44.50 |
| o3-mini | 19.67 | 22.00 | 57.33 | 57.83 | 58.50 | 62.33 | 24.67 | 70.00 | 26.67 | 25.17 | 42.41 |
| **Open-Source Models** | | | | | | | | | | | |
| deepseek-r1 | 54.67 | 52.17 | 81.33 | 75.33 | 57.67 | 71.83 | **80.83** | **84.00** | 48.67 | 39.17 | 64.65 |
| qwq-32B-preview | 40.50 | 43.50 | 76.33 | 75.00 | 59.67 | 76.83 | 55.67 | 77.50 | 27.50 | 32.67 | 56.52 |
| deepseek-v3 | 36.83 | 35.67 | 67.83 | 64.00 | 60.83 | 65.00 | 45.00 | 82.17 | 24.83 | 33.67 | 50.56 |
| deepseek-r1-distill-qwen-32b | 45.17 | 37.50 | 81.83 | 60.17 | 66.00 | 65.33 | 54.33 | 42.83 | 26.50 | 23.00 | 50.27 |
| qwen-2.5-72B | 30.67 | 26.00 | 68.83 | 60.83 | 66.00 | 77.00 | 29.33 | 51.17 | 22.50 | 23.17 | 45.55 |
| **Human** | | | | | | | | | | | |
| Best | 96.67 | 90.00 | 100.00 | 96.67 | 93.33 | 93.33 | 80.00 | 100.00 | 96.67 | 86.67 | 93.33 |
| Mean | 85.71 | 85.24 | 95.19 | 94.07 | 80.42 | 80.83 | 70.83 | 96.11 | 80.48 | 80.00 | 82.36 |

Table 2: Models Performance on Different Domains (Accuracy %). Best performers in each domain are **bolded**.

`Qwen-Max`, `Qwen2.5-72B-Instruct`, and `QwQ-32B-Preview` (Qwen Team 2024). We also try the distilled model by Deepseek-R1, which is `r1-distill-qwen-32b`. (DeepSeek-AI et al. 2025) We also fine-tune on `Qwen2.5-7B`. The evaluations are conducted in both Chinese and English, and the fine-tuning is conducted in English, with detailed settings provided in Appendix C. The answers are extracted from the model response using a rule-based method. Details about the unextracted rates are shown in Appendix D. For human evaluations, details are provided in Appendix E.

For human evaluation, we recruited 10 undergraduate participants for each domain and randomly sampled 5% of the dataset as test questions, spanning all difficulty levels, scenarios and question types. Each participant was required to complete 60 questions, including 30 in Chinese and 30 in English, in a time limit of 8 hours.

## 4.2 Evaluation Result

The complete results are presented in Table 2. 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.

Compared to human performance, the maximum scores achieved by the models in the space, nature and mixed domains are significantly lower than the mean scores of human participants. The overall performance of the models also falls short of human-level accuracy, indicating a substantial gap in commonsense reasoning capabilities between LLMs and humans. However, in the time and social domain, the best performing model exceeds the mean human performance. We attribute it to the ambiguity of temporal expressions in natural language, which can lead to divergent interpretations and reasoning outcomes among human participants. It should be noted that some human errors appear to stem from carelessness. Given sufficient time and attention, human participants would be capable of achieving significantly higher scores, potentially even perfect accuracy. For a detailed analysis, please refer to Appendix E.

## 4.3 General Analysis

**More Tokens Needed for Harder Questions** 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 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 F.4.

**Downward Accuracy Across Difficulty 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.

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

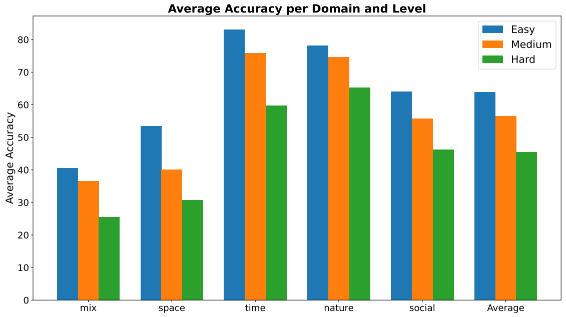


Figure 5: The average performance of LLMs on Chinese and English questions on different levels.

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 is only 47.91%. 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 F.3.

**Cultural and Linguistic Differents Affect Performance** It is generally believed that large language models behave better in English than in non-English languages due to the quality and quantity of training data (Zhang et al. 2023). However, on the contrary, our results show that most models perform better on Chinese questions than on English questions in space, nature and mixed domains. We filter and analyze 100 questions with the greatest performance differences between Chinese and English and find that these differences are primarily due to the following reasons: (1) Some entities are related to geography or culture, which can result in their varying frequency across different language corpora. Although the overall size of English data is larger, the data related to these entities is more abundant in Chinese. The model performs better in understanding commonsense knowledge related to such entities in the Chinese than in English. (2) The same word can have multiple different senses. This semantic ambiguity is more common in English than in Chinese. The same English word may correspond to related or completely unrelated Chinese words, which may bring ambiguity to English questions. The model may confuse different senses during the reasoning process, using different senses in different steps. Detailed examples are shown in Appendix G.

We also find that models tend to exhibit lower performance on rarer scenarios. Changing the way to ask can also affect performance. Detailed analysis is shown in Appendix G.

## 4.4 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. Errors in entity relations mainly manifest as the confusion of closely related, similar social relationships or misinterpreting asymmetric relationships. These types of errors reflect the limitations of LLMs in handling multifaceted commonsense knowledge.

**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. 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. This type of error is related to the model's ability to construct and understand scenario.

**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.

**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. This demonstrates our dataset's ability to expose errors in decision-making strategies, such as premature stopping and inconsistent answer aggregation, which are not well captured by conventional single-choice benchmarks.

Detailed case studies are shown in Appendix G.

## 4.5 Supervised Fine-Tuning on SCoRE

We conducted an SFT experiment to evaluate the effectiveness of SCoRE's data synthesis method. We first select questions from the SCoRE training set that can be answered correctly by DeepSeek-R1. These questions and their corresponding answers were then used to fine-tune a smaller model, Qwen2.5-7B. We evaluate the models before and after SFT on both in-domain (SCoRE) and out-of-domain (CommonSenseQA(Talmor et al. 2022) 2.0 and

MC-TACO(Zhou et al. 2019)) datasets. The result is presented in Table 3.

| Dataset | Before | After | Gain |
|---|---|---|---|
| SCoRE | 25.80 | 64.60 | 38.80↑ |
| - Time | 41.33 | 76.17 | 34.84↑ |
| - Space | 7.00 | 55.50 | 48.50↑ |
| - Nature | 36.33 | 75.00 | 38.67↑ |
| - Social | 34.50 | 83.17 | 48.67↑ |
| - Mix | 9.83 | 33.17 | 23.34↑ |
| CSQA 2.0 | 71.10 | 75.70 | 4.60↑ |
| MC-TACO | 75.52 | 77.07 | 1.55↑ |

Table 3: Performance comparison before and after SFT

**Synthetic data fills in-domain knowledge gaps for smaller models** The SFT model demonstrated remarkable gains across all SCoRE domains, with absolute improvements ranging from 23.3% to 48.67%. Compared to the base model, the SFT model shows a significant reduction in commonsense errors. This suggests that synthetic data can effectively address specific knowledge gaps in smaller models.

**Transfer reasoning capability to out-of-domain datasets** SFT models can learn reasoning patterns, as detailed in Appendix G that can be transferred to out-of-domain datasets. After SFT, the model's average response length increased from 490.46 tokens to 7611.74 tokens, and reasoning patterns such as ”wait” and ”verify,” which were not present before SFT, emerged. The model's accuracy improves 4.60% on CSQA2.0 and 1.55% on MC-TACO, indicating that knowledge and reasoning capabilities acquired from SCoRE synthetic data can transfer to unseen, out-of-domain tasks.

These results demonstrate that SCoRE's data synthesis method can generate high-quality training data that significantly enhances both domain-specific and general reasoning capabilities.

## 5 Related Work

According to the construction method, the previous commonsense 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), MC-TACO (Zhou et al. 2019), ProtoQA (Boratko et al. 2020), StrategyQA (Geva et al. 2021) and SimpleQA (Wei et al. 2024). These datasets utilizes crowdsourcing approach to generate questions related to a given topic, which 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.

**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), SPARTQA (Mirzaee et al. 2021), StepGame (Shi, Zhang, and Lipani 2022), Spatial-EvalLLM (Yamada et al. 2024), TRAM (Wang and Zhao 2023), RoomSpace (Li, Hogg, and Cohn 2024) and LogicBench (Parmar et al. 2024). These datasets design templates and rules for each subtask to gather basic knowledge points. 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. mCSQA (Sakai, Kamigaito, and Watanabe 2024) and MuSR (Sprague et al. 2024). COIG-CQIA (Bai et al. 2024) collects questions from web data, and prompts GPT-4 to generate the responses, while mCSQA (Sakai, Kamigaito, and Watanabe 2024) and MuSR (Sprague et al. 2024) uses LLMs to generate the whole question following pre-defined instructions. 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.

## 6 Conclusion

We introduce SCoRE, a bilingual dataset designed to evaluate complex logical reasoning in commonsense scenarios through multi-hop inference chains generated through a knowledge-driven synthetic data strategy. The dataset contains 100,000 questions across three difficulty levels and commonsense in four domains. Experimental results show that, despite advances in areas like code generation and math problem-solving, LLMs still face challenges in complex commonsense reasoning. Case studies reveal common errors, such as misinterpreting low-frequency commonsense, logical inconsistencies, and overthinking. The proposed data synthesis approach enables automated generation of large-scale, high-accuracy reasoning data with fine-grained labels and explicit reasoning path, which could serve as training data to further improve LLMs' reasoning capabilities.

## Acknowledgements

This research was supported by NSFC project 62476009, the Major Project of the Ministry of Education, China, titled “Research on a Comprehensive Linguistic Knowledge Base for Machine Language Competence Evaluation” (Project No. 22JJD740004) and the Open Project Fund of the State Key Laboratory of Multimedia Information Processing (Project No. SKLMIP-KF-2025-01).

# References

Bai, Y.; Du, X.; Liang, Y.; Jin, Y.; Liu, Z.; Zhou, J.; Zheng, T.; Zhang, X.; Ma, N.; Wang, Z.; et al. 2024. COIG-CQIA: Quality is All You Need for Chinese Instruction Fine-tuning. *arXiv preprint arXiv:2403.18058*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R. B.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N. S.; Chen, A. S.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L. E.; Goel, K.; Goodman, N. D.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M. S.; Krishna, R.; Kuditipudi, R.; and et al. 2021. On the Opportunities and Risks of Foundation Models. *CoRR*, abs/2108.07258.

Boratko, M.; Li, X.; O'Gorman, T.; Das, R.; Le, D.; and Mccallum, A. 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)*, 1122–1136.

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; Li, S. S.; Zhou, S.; Wu, S.; Ye, S.; Yun, T.; Pei, T.; Sun, T.; Wang, T.; Zeng, W.; Zhao, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Xiao, W. L.; An, W.; Liu, X.; Wang, X.; Chen, X.; Nie, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, X. Q.; Jin, X.; Shen, X.; Chen, X.; Sun, X.; Wang, X.; Song, X.; Zhou, X.; Wang, X.; Shan, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhang, Y.; Xu, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Wang, Y.; Yu, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Ou, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Xiong, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Zhu, Y. X.; Xu, Y.; Huang, Y.; Li, Y.; Zheng, Y.; Zhu, Y.; Ma, Y.; Tang, Y.; Zha, Y.; Yan, Y.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Ma, Z.; Yan, Z.; Wu, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Pan, Z.; Huang, Z.; Xu, Z.; Zhang, Z.; and Zhang, Z. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

DeepSeek-AI; Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Guo, D.; Yang, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Zhang, H.; Ding, H.; Xin, H.; Gao, H.; Li, H.; Qu, H.; Cai, J. L.; Liang, J.; Guo, J.; Ni, J.; Li, J.; Wang, J.; Chen, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Song, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhao, L.; Wang, L.; Zhang, L.; Li, M.; Wang, M.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Tian, N.; Huang, P.; Wang, P.; Zhang, P.; Wang, Q.; Zhu, Q.; Chen, Q.; Du, Q.; Chen, R. J.; Jin, R. L.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Xu, R.; Zhang, R.; Chen, R.; Li, S. S.; Lu, S.; Zhou, S.; Chen, S.; Wu, S.; Ye, S.; Ye, S.; Ma, S.; Wang, S.; Zhou, S.; Yu, S.; Zhou, S.; Pan, S.; Wang, T.; Yun, T.; Pei, T.; Sun, T.; Xiao, W. L.; Zeng, W.; Zhao, W.; An, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Li, X. Q.; Jin, X.; Wang, X.; Bi, X.; Liu, X.; Wang, X.; Shen, X.; Chen, X.; Zhang, X.; Chen, X.; Nie, X.; Sun, X.; Wang, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yu, X.; Song, X.; Shan, X.; Zhou, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhu, Y. X.; Zhang, Y.; Xu, Y.; Xu, Y.; Huang, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Li, Y.; Wang, Y.; Yu, Y.; Zheng, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Tang, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Wu, Y.; Ou, Y.; Zhu, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Zha, Y.; Xiong, Y.; Ma, Y.; Yan, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Wu, Z. F.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Huang, Z.; Zhang, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Gou, Z.; Ma, Z.; Yan, Z.; Shao, Z.; Xu, Z.; Wu, Z.; Zhang, Z.; Li, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Gao, Z.; and Pan, Z. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Chang, B.; Sun, X.; and Sui, Z. 2024. A Survey on In-context Learning. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, 1107–1128. Association for Computational Linguistics.

Dong, Z.; and Dong, Q. 2003. HowNet-a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, 820–824. IEEE.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. arXiv:2101.02235.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z.

2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; Madry, A.; Baker-Whitcomb, A.; Beutel, A.; Borzunov, A.; Carney, A.; Chow, A.; Kirillov, A.; Nichol, A.; Paino, A.; Renzin, A.; Passos, A. T.; Kirillov, A.; Christakis, A.; Conneau, A.; Kamali, A.; Jabri, A.; Moyer, A.; Tam, A.; Crookes, A.; Tootoonchian, A.; Kumar, A.; Vallone, A.; Karpathy, A.; Braunstein, A.; Cann, A.; Codispoti, A.; Galu, A.; Kondrich, A.; Tulloch, A.; Mishchenko, A.; Baek, A.; Jiang, A.; Pelisse, A.; Woodford, A.; Gosalia, A.; Dhar, A.; Pantuliano, A.; Nayak, A.; Oliver, A.; Zoph, B.; Ghorbani, B.; Leimberger, B.; Rossen, B.; Sokolowsky, B.; Wang, B.; Zweig, B.; Hoover, B.; Samic, B.; McGrew, B.; Spero, B.; Giertler, B.; Cheng, B.; Lightcap, B.; Walkin, B.; Quinn, B.; Guarraci, B.; Hsu, B.; Kellogg, B.; Eastman, B.; Lugaresi, C.; Wainwright, C. L.; Bassin, C.; Hudson, C.; Chu, C.; Nelson, C.; Li, C.; Shern, C. J.; Conger, C.; Barette, C.; Voss, C.; Ding, C.; Lu, C.; Zhang, C.; Beaumont, C.; Hallacy, C.; Koch, C.; Gibson, C.; Kim, C.; Choi, C.; McLeavey, C.; Hesse, C.; Fischer, C.; Winter, C.; Czarnecki, C.; Jarvis, C.; Wei, C.; Koumouzelis, C.; and Sherburn, D. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.

Li, F.; Hogg, D. C.; and Cohn, A. G. 2024. Reframing Spatial Reasoning Evaluation in Language Models: A Real-World Simulation Benchmark for Qualitative Reasoning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, 6342–6349. ijcai.org.

Mirzaee, R.; Faghihi, H. R.; Ning, Q.; and Kordjamshidi, P. 2021. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tür, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 4582–4598. Association for Computational Linguistics.

Parmar, M.; Patel, N.; Varshney, N.; Nakamura, M.; Luo, M.; Mashetty, S.; Mitra, A.; and Baral, C. 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)*, 13679–13707.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models.

Sakai, Y.; Kamigaito, H.; and Watanabe, T. 2024. mCSQA: Multilingual Commonsense Reasoning Dataset with Unified Creation Strategy by Language Models and Humans. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 14182–14214. Bangkok, Thailand: Association for Computational Linguistics.

Shi, Z.; Zhang, Q.; and Lipani, A. 2022. StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 11321–11329. AAAI Press.

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

Sprague, Z. R.; Ye, X.; Bostrom, K.; Chaudhuri, S.; and Durrett, G. 2024. MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning. In *The Twelfth International Conference on Learning Representations*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 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)*, 4149–4158.

Talmor, A.; Yoran, O.; Le Bras, R.; Bhagavatula, C.; Goldberg, Y.; Choi, Y.; and Berant, J. 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)*.

Wang, Y.; and Zhao, Y. 2023. Tram: Benchmarking temporal reasoning for large language models. *arXiv preprint arXiv:2310.00835*.

Wei, J.; Karina, N.; Chung, H. W.; Jiao, Y. J.; Papay, S.; Glaese, A.; Schulman, J.; and Fedus, W. 2024. Measuring short-form factuality in large language models. arXiv:2411.04368.

Weston, J.; Bordes, A.; Chopra, S.; Rush, A. M.; Van Merriënboer, B.; Joulin, A.; and Mikolov, T. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Yamada, Y.; Bao, Y.; Lampinen, A. K.; Kasai, J.; and Yildirim, I. 2024. Evaluating Spatial Understanding of Large Language Models. *Trans. Mach. Learn. Res.*, 2024.

Zhang, X.; Li, S.; Hauer, B.; Shi, N.; and Kondrak, G. 2023. Don‘t Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7915–7927. Singapore: Association for Computational Linguistics.

Zhou, B.; Khashabi, D.; Ning, Q.; and Roth, D. 2019. “Going on a vacation” takes longer than “Going for a walk”: A Study of Temporal Commonsense Understanding. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3363–3369. Hong Kong, China: Association for Computational Linguistics.

| 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 $X$”;<br>”$X$ 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 $B$”;<br>”$A$ happened in $T$” and ”$B$ happened in $T$” can imply that ”$A$ and $B$ happened at the same time”;<br>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”,<br>red light has a longer wavelength than yellow light,<br>A same stone sinks faster in freshwater than in seawater. |

Table 4: 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.

## 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[1] 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 (Dong and Dong 2003), ConceptNet[2] (Speer, Chin, and Havasi 2017) and the Contemporary Chinese Dictionary. The properties are clustered according to the Attribute Value list in HowNet.

**Manual Construction** Table 4 presents the number of knowledge entries written by humans in four knowledge bases. Scenarios are shown in Appendix H.

## B Human Verification and Refinement

Human oversight is a quality control measure for the knowledge base. We conducted a total of three rounds of human verification.

In the first round, we directly audited the knowledge base. Eight graduate students were recruited to review all knowledge entries. Approximately **15%** of the entries were revised based on their evaluation.

In the second round, we performed a stratified sampling inspection on the synthesized data. For each domain, 10% of the generated data was randomly sampled. The sampled items spanned all difficulty levels, scenarios, and question types, covering as many distinct knowledge entries as possible. Ten graduate students participated in this round of verification. They were instructed to assess the following aspects:

- Correctness of the question: whether there were contradictions, commonsense violations, or missing conditions;
- Correctness of the answer: whether the answer of the questions are correct;
- Linguistic Quality: whether the questions were expressed clearly and naturally, rated on a 3-point scale.

The results showed that approximately **30%** of the questions had issues in at least one of the above three dimensions. We traced each error back to its originating knowledge entry and made the necessary corrections.

The third round followed the same procedure as the second, focusing on evaluating the correctness of the question and answers. This round revealed that, with the exception of the social domain, where errors were caused by the reasoning program, all other domains achieved **100%** correctness. After fixing the errors of the program, both the dataset and the knowledge base reached complete accuracy.

## C Experiments Setting

All models are tested using their official APIs. For models that allowed temperature adjustment, we set the temperature

---

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

[2] These resources are used under the CC-BY-4.0 license.

to 0.7. The English and Chinese prompts used in the evaluations are shown in Figure 6.

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.

## D 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 5) 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.

| Model Name | Unextracted Rate |
|---|---|
| o1-mini | 0.0007 |
| deepseek-chat | 0.0387 |
| gpt-4o | 0.0341 |
| deepseek-r1-distill-qwen-32b | 0.0244 |
| claude-3-5-sonnet-20241022 | 0.0083 |
| o1-preview | 0.0156 |
| qwq-32B | 0.0463 |
| qwen-25-72B | 0.0074 |
| qwen-max | 0.0305 |
| glm-zero-preview | 0.0059 |
| o3-mini | 0.0126 |
| deepseek-reasoner | 0.0056 |
| glm-4-plus | 0.0017 |
| Overall Average | 0.0176 |

Table 5: Unsuccessful extraction rates of different models

## E Human Evaluation

We recruited a total of 50 undergraduate students from Peking University to participate in the human evaluation, with 10 participants randomly assigned to each commonsense domain. Each domain consisted of 60 test questions. Except for the social relationship domain, which included 60 questions in Chinese, all other domains consisted of 30 questions in Chinese and 30 in English. To help participants stay focused during the evaluation, we divided the 60 questions into four online questionnaires, each containing 15 questions. Participants were required to complete each questionnaire within 2 hours. We released one questionnaire every two days, and the evaluation lasted 8 days. Participants are allowed to use pen and paper during the evaluation. The instruction presented to the participants was similar to the prompts given to LLMs, as presented in figure 6. The entire human evaluation cost approximately $10k.

After human evaluation, we calculated the accuracy of each participant in the 60 questions. For each domain, we computed the average accuracy and standard deviation of the 10 participants. Any score falling below the threshold of “average accuracy minus two standard deviation” was considered an outlier and excluded. In total, we collected 39 valid questionnaires. The number of valid results for each domain is reported in table 6.

| **Space** | **Time** | **Nature** | **Social** | **Mix** | **Total** |
|---|---|---|---|---|---|
| 7 | 8 | 9 | 8 | 7 | 39 |

Table 6: The number of valid human evaluation results per commonsense domain

**Social Domain** As shown in table 7, we found that human participants perform poorly on the questions of the hard level in social domain. The questions in the easy level and medium level cover general social relationships, where interpersonal connections are typically sparser and do not require relational calculations, in contrast to the Chinese family relationships in the hard-level questions. For example, a ’father’s colleague’ in a general social context lacks a single specific term, while a ’father’s father’ in a family context is simply ’grandfather’. This distinction makes family relationship questions inherently more difficult than those about general social relationships. Moreover, hard-level questions involve a larger number of entities, which further increases the difficulty for human participants to construct complex relationship networks (such as family networks).

| Level | Human Test Accuracy |
|---|---|
| Easy | 0.9857 |
| Medium | 0.9625 |
| Hard | 0.4458 |

Table 7: Human accuracy on 3 levels in social domain

**Time Domain** Due to the potential ambiguity of the temporal expression in language, it is difficult for human to understand the temporal relationship of the events. For instance, given the statements ”There is a 1-day gap between Jack reads research papers and he cleans his dormitory room” and ”Jack read research papers on Thursday”, some may think that Jack cleans his dormitory room on Friday, while others may think Saturday. This ambiguity affects the human’s understanding of the temporal relationship between events.

4 types of statements are questioned in time domain (see the ”Imprecision” section in Appendix G). The accuracy of human understanding on each type of statement is shown in Table 8 below. We found that the Arith. questions are the most difficult for human to answer, since they are the most vague statements. Besides, Seq. & Arith. questions are also ambiguious, leading to humans’ misunderstanding.

| Language | Time. | Seq. | Arith. | Seq. & Arith. |
|---|---|---|---|---|
| cn | 1.00 | 0.85 | 0.49 | 0.71 |
| en | 1.00 | 0.60 | 0.62 | 0.73 |

Table 8: Human accuracy on 4 types of statements in time domain. Human has low accuracy on Seq. & Arith. and Arith. statements, which are ambiguous.

# F Detailed Analysis

## F.1 Lower Performance on Rarer Scenarios

We simultaneously analyze the performance of different models across various knowledge domains and scenarios, and the results are presented in Figure 7. 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.

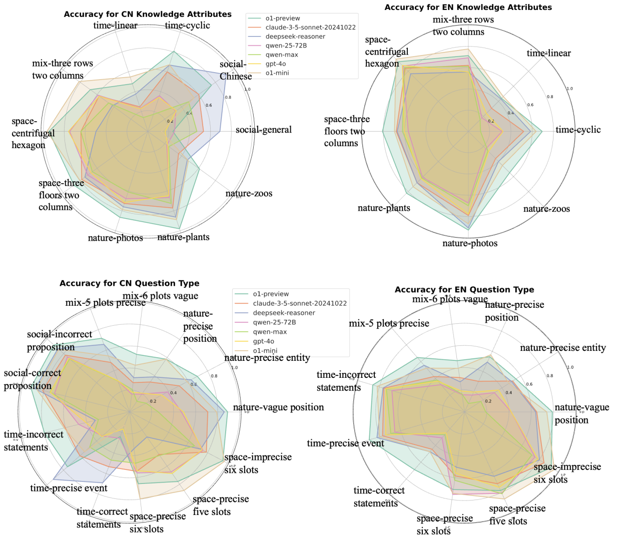


Figure 7: The performance of LLMs on Chinese and English questions with different knowledge attributes and question types. Here, if the answer to a question involves only a single entity, it is termed ”precise.” If it involves multiple entities, it is termed ”vague”.

## F.2 Changing the Way to Ask Affects Performance

Additionally, the formulation of the question also plays a crucial role in the model’s performance. For example, as shown in Figure 7, 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.

## F.3 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 8, 9, 10, 11, and 12, our difficulty classification system effectively differentiates question difficulty across most models, demonstrating its reliability in assessing model capabilities.

## F.4 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.

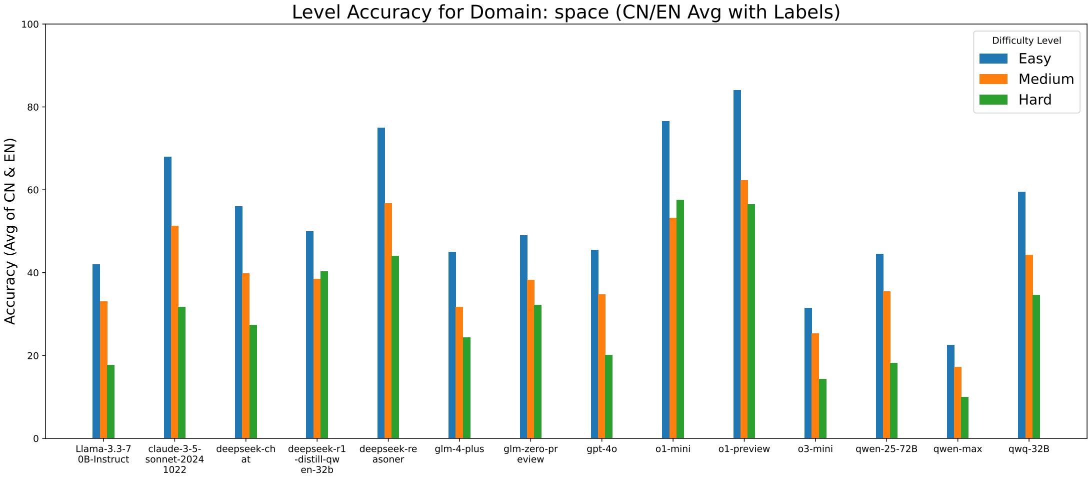


Figure 8: Level Accuracy on Space Domain

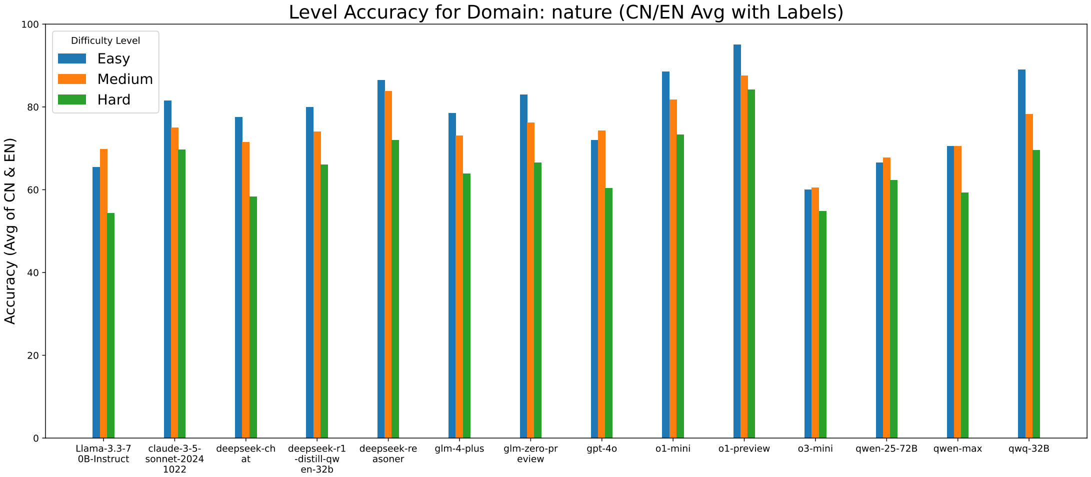


Figure 9: Level Accuracy on Nature Domain

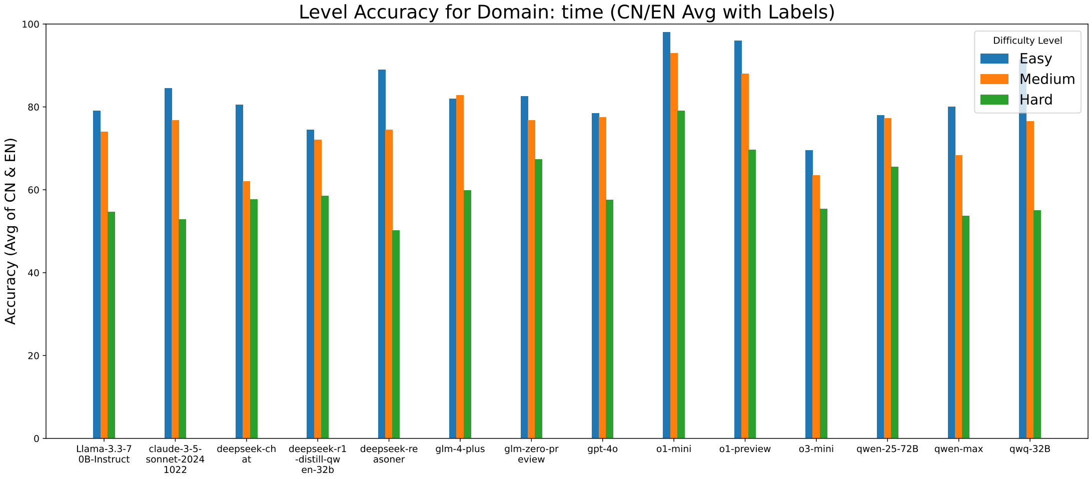


Figure 10: Level Accuracy on Time Domain

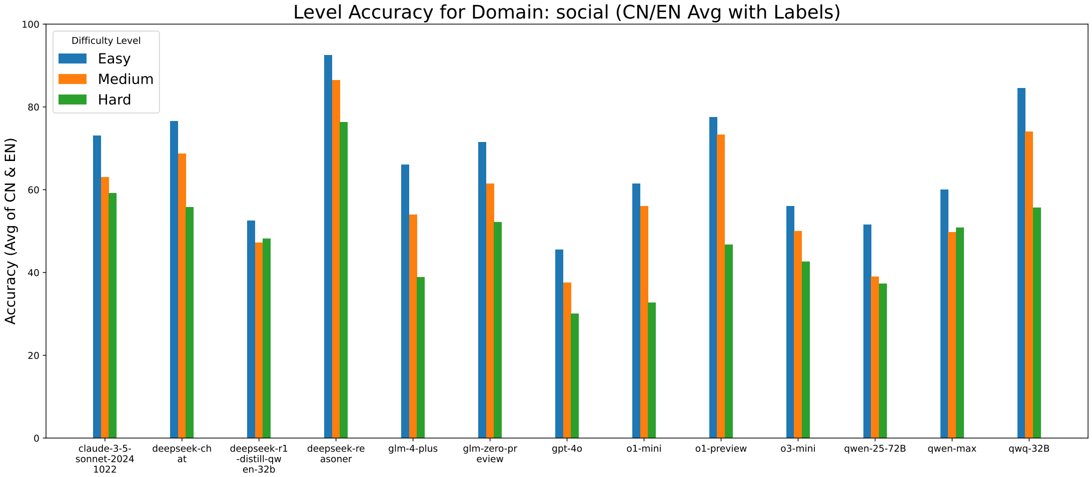


Figure 11: Level Accuracy on Social Domain

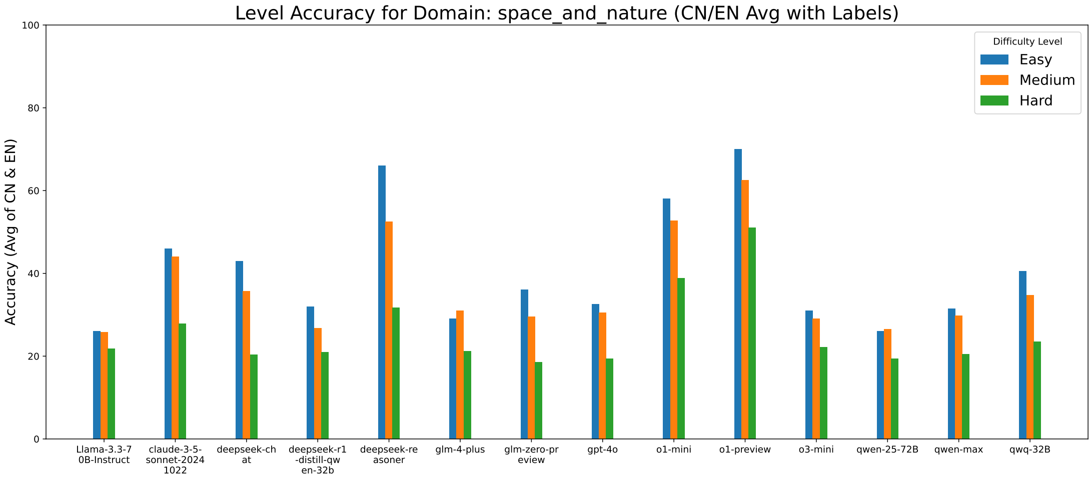


Figure 12: Level Accuracy on Mix Domain

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.

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 13.

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 responses 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.

# G Case Study

## G.1 Commonsense Error

**Entity Property Error** We found that the model still lacks understanding of some low-frequency entity properties. As shown in Figure 14, 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.

**Entity Relation Error** We find that the model often confuses closely related social relationships. As shown in Figure 16, 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 17, 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.”

To further investigate the models' comprehension of social relationships, we compiled statistics on the frequency of errors made by the models on test questions pertaining to different social relations. As shown in Figure 15, Current Large Language Models exhibit a general proclivity for error when processing social relationship terminologies, particularly with affinal (in-law) relationships, generational terms (especially those requiring distinctions between ma-

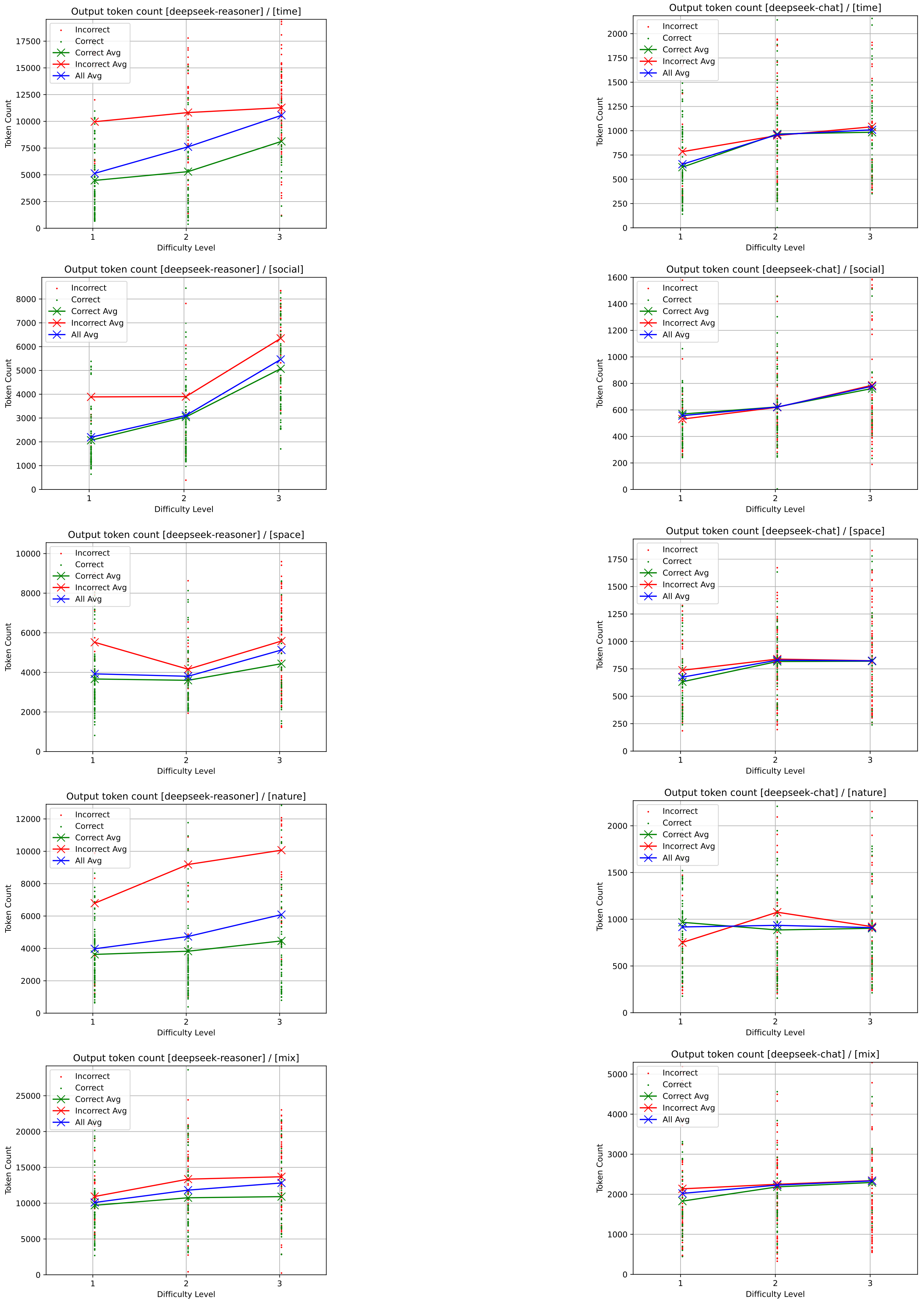


Figure 13: Relationship between response token count and question difficulty level. Left column: Results from deepseek-reasoner; Right column: Results from deepseek-chat. From top to bottom: Time, Social, Space, Nature, and Mix datasets

**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 14: Example of entity property error in the model's response.

ternal/paternal lines, relative age, and gender), and certain everyday yet ambiguously-bounded non-consanguineous relationships (e.g., friends, colleagues). This predisposition towards error likely stems from inadequacies in the models' capacity to learn and infer the conventional naming systems, cultural contexts, and complex network structures inherent in human social organization. While error hotspots vary slightly among different models, possibly reflecting variations in their training datasets and algorithmic emphases, the overarching challenge remains.

## G.2 Reasoning Error

**Logical Contradiction** We notice that the model's response may sometimes be self-contrast. As shown in Figure 18, 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 con-

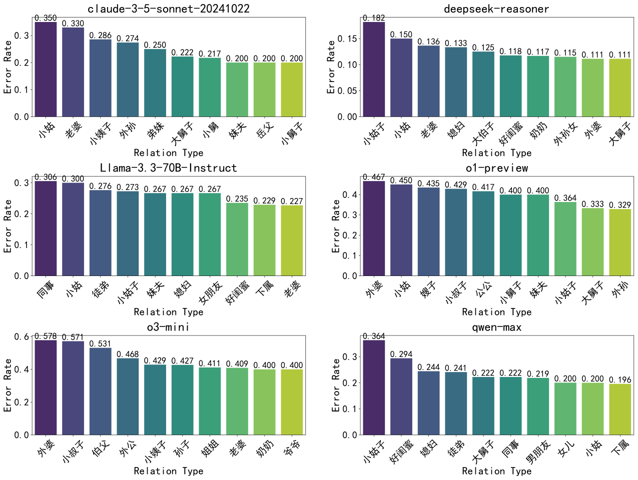


Figure 15: A Statistical Overview of Top 10 High-Frequency Error Terms in Large Language Model Performance on Social Relationship Comprehension

**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 16: Example of confusing different relationship in the model's response.

**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 17: 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 18: Example of logical contradiction in the model's response.

trasts 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.

**Imprecision** In the reasoning process, models may be affected by irrelevant information contained in previous outputs, leading to imprecision in final results. As shown in Figure 19, when calculating the time that "Jack goes jogging", the model uses the existing condition "2 days after learning Japanese". However, it is misled by closer statements that are unrelated, since "Wednesday" in context is misused in the process. This imprecision results in an error in reasoning.

In time domain, 4 types of statements are questioned:

1. Time of Events (Time.), such as "Jack met his future wife in 1920".
2. Sequence of Events (Seq.), such as "Jack met his future wife after he graduated from high school".
3. Arithmetic of Time (Arith.), such as "There is a 2-year gap between Jack graduated from high school and he met his future wife".
4. Sequence of Events and Arithmetic of Time (Seq. & Arith.), such as "Jack met his future wife 2 years after he graduated from high school".

Since arithmetic is more difficult than sequence comparison in reasoning, it is easier to be misled by irrelevant information. Thus, accuracy of the model in Arith. questions is lower than that in Seq. questions, and models has the lowest accuracy in Seq. & Arith. questions (shown in the Table 9).

According to the text of questions, this imprecision could be cause by (1) a wrong understanding of the entity or (2) a wrong understanding of the property. Since the occurrence of this error is related to next token prediction, which is regarded as the cornerstone of LLMs, similiar examples can also be found in other fields.

**Inference Error** As shown in figure 20, 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 under-

| Models | Time. | | Seq. | | Arith. | | Seq. & Arith | |
|---|---|---|---|---|---|---|---|---|
| | cn | en | cn | en | cn | en | cn | en |
| claude-3-5-sonnet | 0.67 | 0.74 | 0.65 | 0.82 | 0.58 | 0.70 | 0.46 | 0.58 |
| deepseek-reasoner | 0.45 | 0.57 | 0.46 | 0.76 | 0.35 | 0.60 | 0.31 | 0.54 |
| gpt-4o | 0.68 | 0.62 | 0.73 | 0.85 | 0.62 | 0.69 | 0.57 | 0.59 |
| o1-mini | 0.88 | 0.75 | 0.96 | 0.95 | 0.84 | 0.79 | 0.81 | 0.79 |
| o1-preview | 0.62 | 0.74 | 0.93 | 0.91 | 0.71 | 0.70 | 0.70 | 0.75 |
| qwen-2.5-72B | 0.65 | 0.84 | 0.60 | 0.83 | 0.54 | 0.74 | 0.43 | 0.65 |
| qwen-max | 0.51 | 0.70 | 0.60 | 0.83 | 0.54 | 0.74 | 0.43 | 0.65 |

Table 9: Accuracy of models on different time reasoning types. Since Arith. is harder than Seq., the accuracy of Arith. is lower than that of Seq.

**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 learns Japanese on Monday, not Wednesday.)
...[text omitted]...
Now, the question asks: "5 days before ____, Jack plays badminton."
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 19: Example of imprecision in the model's response.

stand 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.

### G.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 21, 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. In contrast, general-purpose models like 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 staightforward solutions.

**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 20: Example of inference error in the model's response.

**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 21: Example of overthinking in the model's response.

**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 22: Example of answer aggregation error in the model's response.

## G.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 22, 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 contribute to answer aggregation errors. The model may analyse A is right but answer B, as shown in Figure 18.

## G.5 Analysis on Different Behaviour in Chinese and English

It is generally believed that large language models behave better in English than in non-English languages due to the quality and quantity of training data (Zhang et al. 2023). However, on the contrary, our results show that most models perform better on Chinese questions than on English questions in space, nature and mixed domains. We filtered and analyzed 100 questions with the greatest performance differences between Chinese and English, and found that these differences are primarily due to the following reasons:

**Difference in the frequency of entities** Some entities are related to geography or culture, which can result in their varying frequency across different language corpora. For example, the entity ”mandarin fish” is a kind of freshwater fish primarily found in East Asia and North Asia. As a result, it appears more frequently in Chinese corpora. Although the overall size of English data is larger, the data related to the mandarin fish is more abundant in Chinese. The model performs better in understanding commonsense knowledge related to such entities in the Chinese than in English. As shown in Figure 23, the model can correctly identify the mandarin fish as a freshwater fish in Chinese but mistakenly classified it as a saltwater fish in English, leading to subsequent reasoning errors.

Similar cases occurs with other entities that are common in Chinese but not so in English, such as ”panax notoginseng” (usually used as traditional Chinese medicine) or ”winter bamboo shoots” (usually eaten by Asian people), etc.

**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-CN**

A. 猫
B. 桂鱼
C. 王八
D. 海豚
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.

**Options-EN**

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

**Correct Answer**

B

**Model Name**

o1-preview

**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])

**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] ×

Figure 23: Example of entity property error in the model’s response.

**Difference in semantic ambiguity** A same word can have multiple different senses. This semantic ambiguity is more common than in English than in Chinese. A same English word may correspond to related or completely unrelated Chinese words, which may bring ambiguity to English questions. The model may confuse different senses during the reasoning process, using different senses in different steps. For example, the English word ”straw” can correspond to ”吸管”(a tube used to suck liquids into the mouth) and ”稻草”(thin, hollow stem of grains after harvest) in Chinese. As shown in Figure 24, in the Chinese question, the word ”吸管” has no ambiguity and refers to ”a tube used to suck liquids into the mouth”, which is a tool. In the English question, the word ”straw” is ambiguous. When analyzing the item on photo 4, the model excludes the straw for the item on photo 4 is not a tool, which seems that it uses the sense ”a tube used to suck liquids into the mouth”. However, when identifying the plant, it uses the sense ”thin, hollow stem of grains after harvest” and catagorized the straw as a plant.

**Question-CN**

一面墙上贴着桃子、啤酒、吸管(straw, a tube used to suck liquids into the mouth)、板栗四种物品的照片。已知:
2号照片上的物品的可食用部分是它的种子;
4号照片上中的物品不属于工具;
1号照片上的物品的可食用部分是它的果实;
2号照片上的物品不属于饮品;
3号照片上的物品不属于坚果;
4号照片上中的物品不属于水果;
3号照片上的物品不属于水果;
4号照片上中的物品不属于坚果。
植物在____号照片上。

**Options-CN**

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

**Question-EN**

On a wall pasted photos of four different items: peach, malt liquor, straw, chestnut. Now we know that:
The edible part of the item on photo No.2 is its seed;
The item on photo No.4 is not a tool;
The edible part of the item on photo No.1 is its fruit;
The item on photo No.2 is not a drink;
The item on photo No.3 is not a nut;
The item on photo No.4 is not a fruit;
The item on photo No.3 is not a fruit;
The item on photo No.4 is not a nut.
A plant is on photo No.____.

**Options-EN**

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

**Correct Answer**

AB

**Model Name**

o1-mini

**Model Answer-CN**

... [text omitted]...
照片1：可食用部分是果实，符合桃子的特性。
(Photo 1: edible part is its fruit, peach fits this description.)
照片2：可食用部分是种子，并且不属于饮品，符合板栗的特性。
(The edible part is its seed, chestnut fits this description.)
照片3：不属于坚果和水果，符合吸管的特性。
(Not a nut or a fruit, straw fits this description.)
照片4：不属于工具、水果和坚果，符合啤酒的特性。
(Not a tool, fruit, or nut, malt liquor fits this description.)
关于“植物在____号照片上”，桃子和板栗都是植物的部分（桃树和板栗树的果实和种子），因此植物对应的是照片1和照片2。
(Regarding ”A plant is on photo No.____,” both peach and chestnut are parts of plants (the fruits and seeds of peach trees and chestnut trees), so the plant corresponds to photo 1 and photo 2.)
答案是：【A, B】✓
(Answer:[A, B])

**Model Answer-EN**

... [text omitted]...
- Photo 2: The edible part is its seed. Among the options, chestnut fits this description.
- Photo 1: The edible part is its fruit. Peach is the fruit here.
- Photo 4: Not a tool, fruit, or nut, so it must be malt liquor.
- Photo 3: Not a nut or a fruit, so it must be straw.

Identifying the Plant:
- Straw is derived directly from plants, specifically from cereal grains.

Therefore, the plant is on Photo No.3.
Answer: [C] ×

Figure 24: Example of entity property error in the model’s response.

Similar cases occurs on other ambiguous English entity names that can refer to different entities, such as ”seal” (marine mammal, stamp affixed to a document, or fastener used to securely close something), ”fish” (marine animal or meat of fish), etc.

### G.6 Analysis on Difference Before and After SFT

**Improved commonsense ability after SFT** SFT significantly enhances a model's commonsense reasoning by explicitly calibrating its latent knowledge representations and refining task-specific inference pathways. As shown in Table 25, the base model (Qwen-7B) failed because it lacks the knowledge that bayberry is a fruit. SFT addresses this by exposing the model to human-curated reasoning traces that structurally align entities with this knowledge.

**Question**

A hardworking farmer has four fields. In each field he plants one of four crops: persimmon, sesame seed, Osmanthus fragrans, bayberry. It is known that:
The plant in field No.1 is orange;
The plant in field No.2 is not a fruit;
The plant in field No.4 is not a fruit;
The plant in field No.4 is not a flower or grass.
____ is planted in field No.4

**Options**

A. persimmon
B. sesame seed
C. Osmanthus fragrans
D. bayberry

**Correct Answer**

B

**Model Name**

Qwen-7B

**Model Answer**

Let's analyze the given information step by step.
1. **Field No.1**: The plant is orange.
- This means the plant is a fruit (since orange is a fruit). ... [text omitted]...
4. **Field No.4**: The plant is not a fruit.
- Also, it is not a flower or grass.
- Therefore, the plant in Field No.4 must be a type of plant that is neither a fruit nor a flower or grass. The only remaining option from the given crops is bayberry (which is a shrub).
Thus, the plant in Field No.4 is bayberry.
Answer: [D] ×

**Model Name**

Qwen-7B-SFT

**Model Answer**

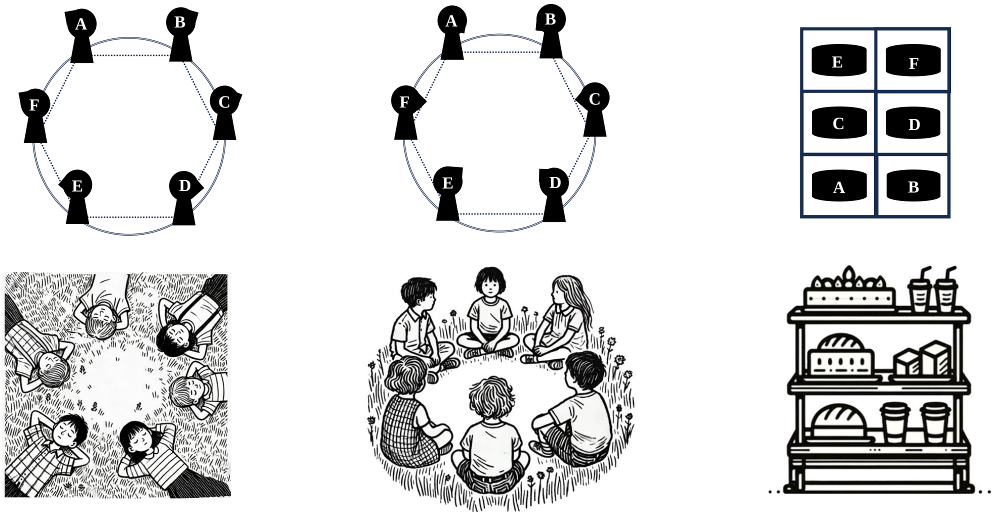
I have a logic puzzle to solve.
... [text omitted]...
First, let's list the clues:
... [text omitted]...
I need to know what categories these plants belong to. Persimmon and bayberry are fruits, I think. Sesame seed is probably a seed, and Osmanthus fragrans is a flower.
... [text omitted]...
Let me confirm:
... [text omitted]...
Field 1 is orange, and we think it's persimmon, but is that necessarily true? Could bayberry be orange? I don't think so. Typically, bayberries are not orange; they are more reddish or purple.
... [text omitted]...
Since field 4 is not a flower or grass, and non-fruit, so it must be sesame seed.
... [text omitted]...
Answer: [B] ✓

Figure 25: Example of improved commonsense ability and reasoning patterns after SFT.

**Improved reasoning pattern after SFT** Smaller, non-reasoning models can also learn advanced reasoning patterns after SFT. As shown in Figure 25, the model simply go through the four fields in order, then make a conclusion. However, after SFT, the model learns a better reasoning pattern to list out all the clues and related commonsense knowledge and reason based on them. During the reasoning process, the SFT model makes self-reflection, where every claim it makes undergoes iterative verification. This reasoning pattern enhances the model's reliability and enables generalization to out-of-domain data.

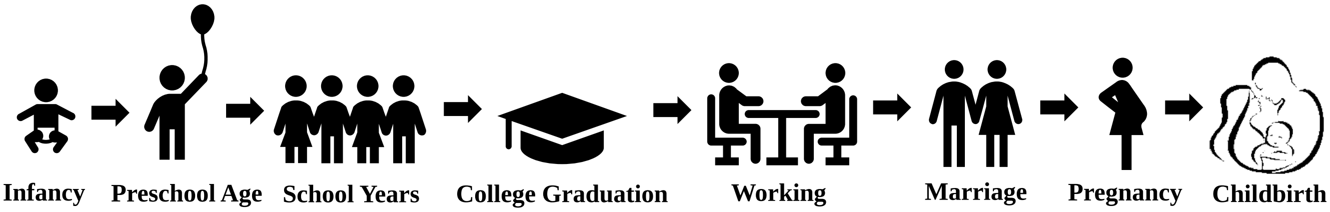
## H Scenario Diagram

The space scenario diagram is shown as Figure 26.
The time scenario diagram is shown as Figure 27.
The social scenario diagram is shown as Figure 28.
The enhanced spatial scenario diagram is shown as Figure 29.

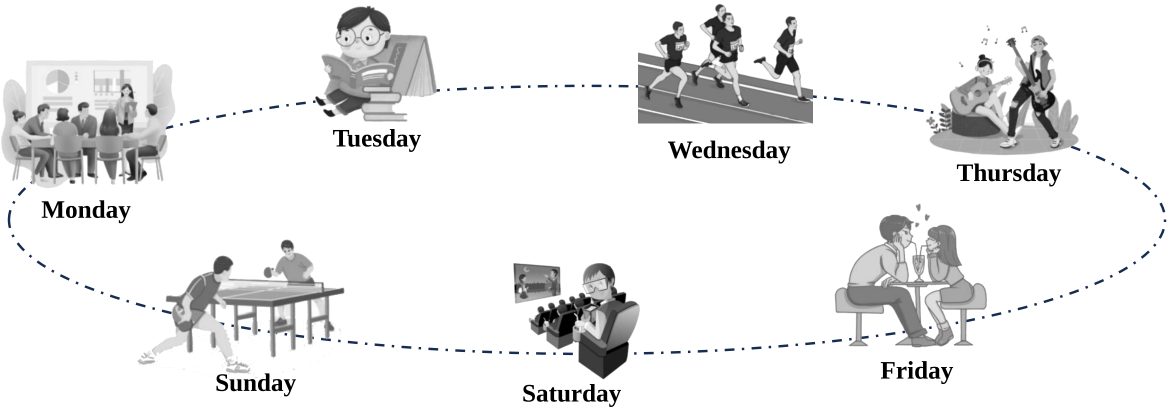


(a) Centrifugal Hexagon Scenario (b) Centripetal Hexagon Scenario (c) Two Rows Three Columns Scenario

Figure 26: Spatial Scenario Diagram

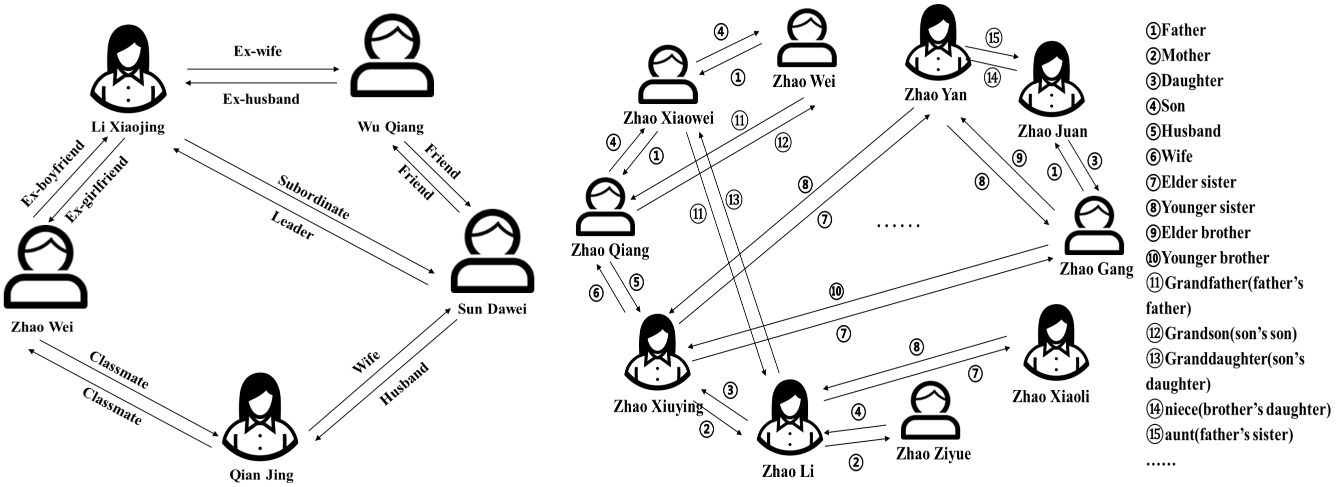


(a) Linear Scenario



(b) Cyclic Scenario

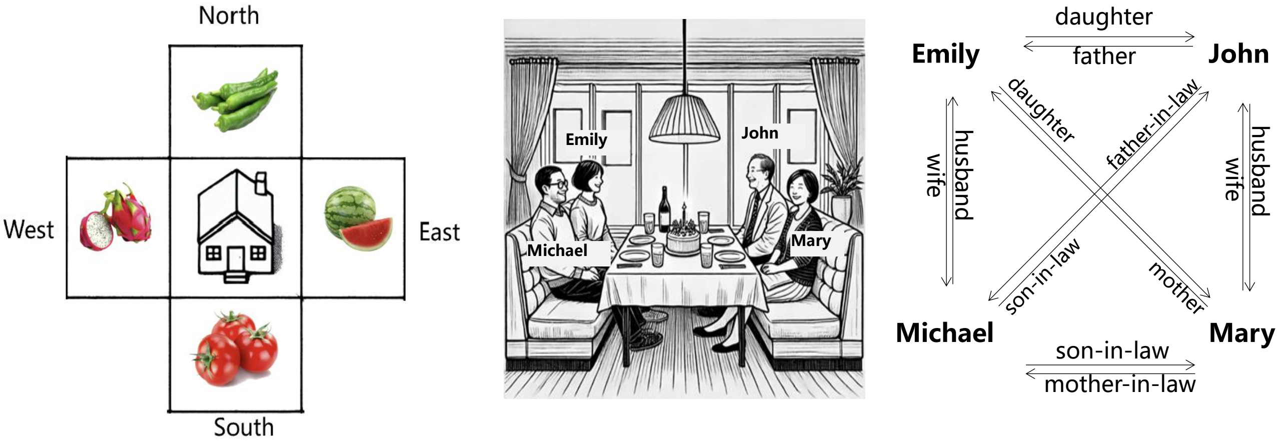
Figure 27: Time Scenario Diagram



(a) General Social Relationship

(b) Chinese Family Relationship

Figure 28: Social Scenario Diagram



(a) Four Plots Farmland Scenario

(b) Family In Four-person Booth Scenario

Figure 29: Enhanced Spatial Scenario Diagram with Entity Properties and Relationships