mSCoRe: a Multilingual and Scalable Benchmark for Skill-based Commonsense Reasoning

Nghia Trung Ngo[1], **Franck Dernoncourt**[2] **and Thien Huu Nguyen**[1]
[1] Department of Computer Science, University of Oregon, Eugene, OR, USA
[2] Adobe Research, USA
{nghian@,thien@cs}.uoregon.edu, franck.dernoncourt@adobe.com

Abstract

Recent advancements in reasoning-reinforced Large Language Models (LLMs) have shown remarkable capabilities in complex reasoning tasks. However, the mechanism underlying their utilization of different human reasoning skills remains poorly investigated, especially for multilingual commonsense reasoning that involves everyday knowledge across different languages and cultures. To address this gap, we propose a **M**ultilingual and Scalable Benchmark for **S**kill-based **Co**mmonsense **Re**asoning (**mSCoRe**). Our benchmark incorporates three key components that are designed to systematically evaluate LLM's reasoning capabilities, including: (1) a novel taxonomy of reasoning skills that enables fine-grained analysis of models' reasoning processes, (2) a robust data synthesis pipeline tailored specifically for commonsense reasoning evaluation, and (3) a complexity scaling framework allowing task difficulty to scale dynamically alongside future improvements in LLM abilities. Extensive experiments on eight state-of-the-art LLMs of varying sizes and training approaches demonstrate that **mSCoRe** remains significantly challenging for current models, particularly at higher complexity levels. Our results reveal the limitations of such reasoning-reinforced models when confronted with nuanced multilingual general and cultural commonsense. We further provide detailed analysis on the models' reasoning processes, suggesting future directions for improving multilingual commonsense reasoning capabilities.

Keywords: Commonsense Reasoning, Multilingual, Data Generation, Interpretability

1. Introduction

Commonsense reasoning enables a person to navigate everyday situations, make logical inferences, and understand implicit information in our environment. While this ability comes naturally to humans, it has proven to be one of the most challenging capabilities to replicate in current language models (Davis, 2024). Recent advancements in Large Reasoning Models (LRMs), such as OpenAI's o1 series (Jaech et al., 2024), and open-source models like DeepSeek R1 (DeepSeek-AI et al., 2025), have shown promising results across various complex reasoning tasks, including mathematics, coding, and logical inference (Kazemi et al., 2025). However, relatively little attention has been devoted to systematically analyzing and understanding these models' commonsense reasoning capabilities, especially in multilingual settings which involve common knowledge across diverse languages and cultural contexts (Do et al., 2024).

Several benchmarks have been proposed to assess commonsense reasoning abilities of language models. CommonsenseQA (CSQA) (Talmor et al., 2019) evaluates general commonsense knowledge through multiple-choice questions derived from ConceptNet. COPA (Roemmele et al., 2011) focuses on causal relationships between everyday events, while SocialIQA (Sap et al., 2019) evaluates social commonsense understanding. More recently, comprehensive benchmarks like MMLU (Hendrycks et al., 2021) and Big-Bench Hard (Suzgun et al., 2023) aim to evaluate model's generalization capabilities across diverse commonsense tasks. However, these benchmarks have significant limitations in three key areas. First, they often focus on a single high-resourced language such as English (Talmor et al., 2019) or Chinese (Sun et al., 2024). Multilingual extensions such as X-COPA (Ponti et al., 2020) and X-CSQA (Lin et al., 2021) primarily rely on translation of existing datasets, thus limiting their ability to capture culturally specific nuances. Second, although recent efforts like mCSQA (Sakai et al., 2024) leverage generative multilingual language models for a more comprehensive and robust dataset creation process, they still lack a systematic way to scale task difficulty, which is crucial for assessing the rapid evolving capabilities of LLMs. Finally, current benchmarks are unable to provide fine-grained analysis and classification of the reasoning steps used by LLMs, which would provide deeper insights into their operations.

To address these limitations, we introduce **M**ultilingual and **S**calable Benchmark for **S**kill-based **Co**mmonsense **Re**asoning (**mSCoRe**), a novel benchmark designed explicitly to provide a comprehensive evaluation of LLMs' commonsense reasoning capabilities across multiple languages and cultural contexts. Specifically, our benchmark offers three notable advantages:

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5095–5115
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

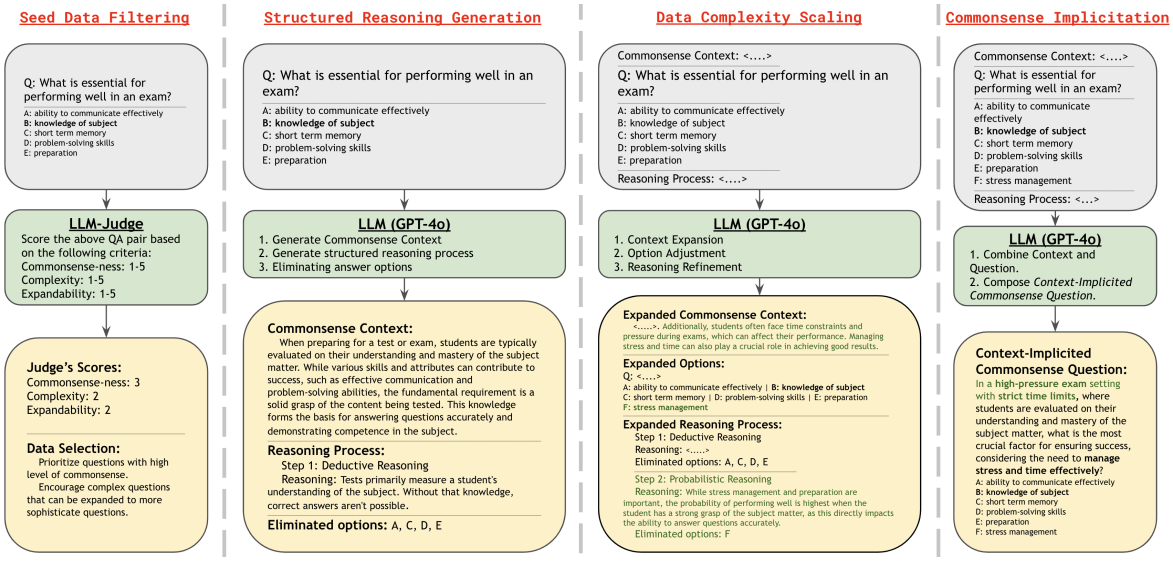

Figure 1: **Data Generation Process.** The four-step data creation pipeline of **mSCoRe**. Each step builds upon the previous one to create progressively more challenging reasoning tasks while maintaining the underlying reasoning skills being evaluated.

1. **Comprehensive Coverage: mSCoRe** covers both `general` commonsense knowledge from across languages including English, German, French, Chinese, and Japanese, and diverse cultural `social` commonsense knowledge.
2. **Skill-based Analysis: mSCoRe** introduces a novel approach to reasoning analysis through the classification of each atomic reasoning step, allowing for more precise analysis of model's reasoning process.
3. **Scalability: mSCoRe** employs techniques such as context expansion, option adjustment, and commonsense implicitation to progressively increase question complexity while preserving commonsense answer semantics, effectively scaling task difficulty.

Our contributions can be summarized as follows:

- We introduce **mSCoRe**, a novel scalable benchmark for evaluating multilingual general and cultural commonsense reasoning with fine-grained skill-based analysis.
- Using **mSCoRe**, we extensively evaluate eights state-of-the-art LLMs, including both commercial and open-source models, across diverse reasoning conditions.
- Our analysis provides insights into how model scale, training techniques, and reasoning skill types impact performance, suggesting future directions for improving commonsense reasoning capability of LLMs.

2. Related works

Large Reasoning Models: Recent advancements in Large Language Models (LLMs) have demonstrated remarkable capabilities in various complex problem-solving tasks. Reasoning-reinforced models like OpenAI o1 (Jaech et al., 2024), Macro-o1 (Zhao et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025) have shown superior performance in mathematics and coding, effectively simulating human-like analytical thinking and enhancing multi-step reasoning (Glazer et al., 2024; Guo et al., 2024). These models employ multiple methods to enhance reasoning capabilities. In particular, Chain-of-thought prompting (Wu et al., 2023) has emerged as a powerful technique that encourages step-by-step reasoning, significantly improving performance on complex tasks. Building upon this foundation, various Chain-of-X approaches have been proposed to further enhance model reasoning capability (Yao et al., 2023; Lightman et al., 2024; Besta et al., 2024; Chen et al., 2024). Recent techniques such as test-time scaling and reinforcement learning have also contributed to improving the reasoning abilities of LLMs (Snell et al., 2024; Kumar et al., 2025; Hou et al., 2025). While these methods enhance the overall structure of the reasoning path, they generally pay little attention to the categorization of each reasoning step. **mSCoRe** proposes a more fine-grained approach in which each step is atomic and labeled according to a reasoning skill, facilitating a deeper and systematic evaluation of model's reasoning process.

Commonsense Reasoning Benchmarks: Despite significant advances in evaluating mathematical and scientific reasoning capabilities of LLMs (Cobbe et al., 2021; Glazer et al., 2024; He et al., 2024; Chow et al., 2025), commonsense reasoning benchmarks have received comparatively less recent attention. Early datasets such as CommonsenseQA (CSQA) (Talmor et al., 2019), COPA (Roemmele et al., 2011), and SocialIQA (Sap et al., 2019) primarily target English-language commonsense knowledge, focusing respectively on general factual knowledge, causal relationships, and so-

Atomic Reasoning Step an indivisible unit of reasoning that predominantly utilizes one reasoning skill. It is a single, coherent thought process that cannot be broken down into smaller steps without losing its meaning. An optimal reasoning path (for multiple-choice QA task) uses a minimum number of atomic steps necessary, ensuring that each step is non-redundant and contributes to narrowing down the possible options by eliminating one or more answer choices.

Figure 2: Atomic Reasoning Step definition.

cial interactions. Recent comprehensive benchmarks like MMLU (Hendrycks et al., 2021) and Big-Bench Hard (Suzgun et al., 2023) evaluate the generalization abilities of LLMs across diverse commonsense reasoning tasks. Multilingual extensions like X-CSQA (Lin et al., 2021) and X-COPA (Ponti et al., 2020) expand the evaluation beyond English by translating existing datasets into multiple languages. More recent approaches such as mCSQA (Sakai et al., 2024) leverage LLM to assist more closely in the data synthesis process. However, these benchmarks are still limited in cultural social commonsense that involves everyday interactions among different cultures. There has been increasing efforts on cultural knowledge bases to develop cultural-aware LLMs. In particular, CulturePark (Li et al., 2024) introduces a novel multi-agent communication framework powered by LLMs to simulate cross-cultural human interactions, whereas CultureBank (Shi et al., 2024) aggregates real-world social interactions from platforms like TikTok and Reddit, structuring annotations around cultural topics. **mSCoRe** builds upon mCSQA and CultureBank to comprehensively cover both *general* and *social* aspects of commonsense reasoning across multiple languages and cultures.

3. Benchmark Creation

3.1. Commonsense Reasoning

Commonsense reasoning involves making inferences about unstated aspects of a scenario using implicit world knowledge – a capability ingrained in human behavior but still challenging for current LLMs. Unlike formal reasoning domains such as mathematics or logic, where rules are explicitly defined and conclusions follow determinate paths, commonsense reasoning requires access to a vast reservoir of implicit knowledge and the ability to apply this knowledge flexibly across diverse situations. Furthermore, there can be multiple reasoning paths that can lead to the correct answer, especially for commonsense questions. However, previous evaluations still primarily focus on answer accuracy (Suzgun et al., 2023; Sakai et al., 2024), providing limited insight into how LLMs construct their reasoning pathways.

To address this limitation, we propose to investigate deeply into model's reasoning process by introducing the concept of "*atomic reasoning steps*" (Fig. 2) as the foundational unit of analysis. Our framework aims to analyze the optimal path utilized by the LLM, defined as the path requiring the minimum number of atomic reasoning steps while maintaining logical coherence. This approach not only enables systematic evaluation of specific reasoning skills, but also provides a clear framework for analyzing how models construct complex reasoning chains. It allows for meaningful comparison of reasoning processes across different models and languages. Finally, it facilitates the scaling of question complexity through the requirement of additional of atomic steps.

3.1.1. Reasoning Skills

We develop a structured taxonomy for classifying each reasoning step, enabling systematic evaluation of how LLMs employ human reasoning skills in commonsense tasks. While no clear consensus exists on a comprehensive taxonomy of human reasoning skills, existing categorizations typically serve specific purposes. For example, Bloom's Taxonomy (Oscarini and Bhakti, 2010) provides a hierarchical framework categorizing educational goals into three domains: cognitive (knowledge-based), affective (emotion-based), and psychomotor (action-based). Similarly, Fleishman's taxonomy (Welford, 1986) identifies 52 distinct human abilities across cognitive, perceptual, psychomotor, and physical domains, primarily to facilitate job design, training, and assessment development. Based on the fundamental characteristics of commonsense knowledge identified by (Do et al., 2024) with established reasoning skill categorizations from (Wikipedia contributors, 2025), we propose a taxonomy comprising three major categories:

- **Logical Reasoning** encompasses forms of reasoning that involve structured processes to derive conclusions from given information. This category includes methodologies like **deductive**, **inductive**, and **abductive** reasoning, which are foundational in scientific and analytical disciplines to ensure conclusions are logically sound.
- **Contextual Reasoning** includes skills used to understand relationships, contexts, and dynamics between elements. This category covers various types of reasoning such as **analogical**, **counterfactual**, **probabilistic**, **temporal**, and **spatial**, used to evaluate scenarios, predict outcomes, and solve problems across different contexts.
- **Social and Ethical Reasoning** involves skills focused on understanding social interactions and evaluating ethical principles. This category includes **social** and **moral** reasoning, essential for interpreting behaviors, navigating complex social environments, and making decisions based on ethical considerations.

Detailed descriptions and examples of each rea-

Step 1 - Data Filtering: To limit the cost while maintaining quality and diversity, we sample a small subset from the seed benchmarks. Each sample is scored by a general LLM-judge based on multiple criteria for expansion potential, ensuring that we select instances that will yield meaningful insights when scaled complexity-wise.
Step 2 - Reasoning Generation: Provide a *Commonsense Context* to expand on the given question and a detailed *Reasoning Process* that involves multiple *Reasoning Steps* to arrive at the correct answer. This establishes a gold standard reasoning path for each question.
Step 3 - Complexity Scaling: Modify and expand each question to create more complex variants by expanding its context, modifying the question, adjusting the answer options, and adding additional *Reasoning Steps*. This creates a progression of difficulty levels for each base question.
Step 4 - Commonsense Implicitation: Combine the given *Commonsense Context* with the question to generate a new, concise commonsense question that *implicitly incorporates* the original context. This process aims to evaluate the commonsense reasoning abilities of LLMs by ensuring that the *implicit context* preserves the original reasoning process and maintains the correctness of the answer.

Figure 3: Four steps of data generation process.

Commonsense-ness: Does answering the question rely solely on commonsense knowledge accessible to the general population, or does it require formal reasoning and specialized expertise beyond everyday understanding?
Complexity: How difficult is the question to understand and answer? Does it require minimal reasoning or a complex, multi-step thought process to identify the correct answer?
Expandability: To what extent can the question be expanded or elaborated upon to introduce additional complexity or dimensions?

Figure 4: Three criteria of data filtering.

soning skill are provided in Table 1. While humans employ additional reasoning skills beyond those presented here, our goal is to establish a concise yet comprehensive reasoning taxonomy that maximizes coverage of human reasoning capabilities for commonsense applications, while minimizing the overlap between categories. This reasoning taxonomy will be implemented within our LLM prompts throughout both the data generation and evaluation processes to ensure focus on our considered skills. Each atomic reasoning step will be classified under a single skill from our taxonomy, enabling precise comparison of different reasoning processes.

3.2. mSCoRe

To maintain robust label accuracy, rather than using LLMs to generate a synthetic dataset from scratch, we utilize human-annotated seed datasets and scale up their complexity to create **mSCoRe**. In particular, our benchmark consists of multiple-choice commonsense questions, separated into two subsets based on different seed datasets: (1) **mSCoRe-G** focuses on general commonsense reasoning, building upon multilingual commonsense questions from mCSQA (Sakai et al., 2024) as a seed dataset. This component evaluates understanding of physical causality, temporal relationships, and basic world dynamics across multiple languages. (2) **mSCoRe-S** addresses social commonsense reasoning based on diverse cultural situations from CultureBank (Shi et al., 2024). This component specifically tests understanding of social interactions, cultural norms, and behavioral expectations across different cultural contexts.

The overall data generation process is visualized in Fig. 1, in which each instance in the seed datasets undergoes a four-step process as illustrated in Fig. 3. Through this systematic creation process, **mSCoRe** provides a comprehensive framework for evaluating and analyzing commonsense reasoning capabilities of LLMs.

Context Expansion: Add additional background or situational details to the *Commonsense Context* to increase depth and reasoning requirements to the question.
Option Adjustment: Adjust the existing answer options to align with the new complex question, ensure the correct answer option remains semantically similar to the original. Introduce an additional plausible but incorrect option to increase the complexity of the question that (1) increases the complexity of the question, and (2) requires an additional reasoning step to eliminate.
Reasoning Refinement: Refine the original *Reasoning Process* to fit the new context with an additional reasoning step that eliminates the added incorrect option.

Figure 5: Three sub-steps of Complexity Scaling.

3.2.1. mSCoRe-G: General Commonsense

mCSQA (Multilingual CommonsenseQA) extends the CommonsenseQA dataset (Talmor et al., 2019) to eight languages to evaluate language models' cross-lingual commonsense reasoning capabilities. Building upon ConceptNet, each multiple-choice question-answer (QA) pair in mCSQA mostly revolves around general commonsense knowledge (an example is provided in the first step in Figure 1). To create **mSCoRe-G**, we further process each QA pair through the following 4 steps:
1. Seed Data Filtering: A general LLM-judge evaluates each candidate on three criteria described in Fig. 4: (1) Commonsense-ness, (2) Complexity, and (3) Expandability. The goal is to prioritize questions with a high level of commonsense and complexity, while maintaining flexibility for expansion into more sophisticated questions (full details of the judge model and scoring criteria are provided in our data generation detail appendix).
2. Structured Reasoning Generation: For selected question-answer pairs, we employ LLMs to generate relevant commonsense context that helps identify the correct answer. From the tuple (context, question, options-answer), we then generate a structured reasoning process. Each reasoning step in the process consists of three attributes: (1) **Reasoning Skill** - the specific skill from our reasoning ontology that is predominantly employed in this step, (2) **Reasoning Text** - the model's rationale based on the identified skill, and (3) **Eliminated Options** - the list of options eliminated in this step based on the reasoning.
3. Data Complexity Scaling: From the base (context, question, answer, reasoning process), we implement a procedure to systematically scale up the

Skills	Short Definitions	Examples
Logical Reasoning		
Inductive	Drawing general conclusions from specific observations.	Most technological innovations eventually benefit society.
Deductive	Deriving specific conclusions from general premises.	All communication tools connect people; social media is a communication tool.
Abductive	Forming hypotheses to explain observations.	Rising depression rates suggest social media affects mental health.
Contextual Reasoning		
Analogical	Drawing parallels between similar situations to infer conclusions.	Like town squares facilitated discourse, social media creates digital gathering spaces.
Counterfactual	Considering alternative scenarios and outcomes that did not happen.	Without social media, many social movements would lack momentum.
Probabilistic	Applying principles of probability to make inferences under uncertainty.	Users have a very high chance of encountering misinformation weekly.
Temporal	Understanding sequences and durations of events.	Brief moments scrolling accumulate into hours of lost productivity daily.
Spatial	Visualizing and manipulating objects in space.	Platform designs maximize attention capture through strategic layouts.
Social & Ethical Reasoning		
Social	Understanding social interactions and norms.	Like-based validation systems create unhealthy approval-seeking behaviors.
Moral	Deciding what is right or wrong based on ethical principles.	Prioritizing profit over user wellbeing raises ethical concerns.

Table 1: The ten types of reasoning skills across three categories with short definitions and examples for each skill applied to the question *"Is social media good for society?"*. Detailed descriptions and additional examples are provided in our experimental detail appendix.

difficulty level of each question. The goal is to introduce an additional plausible option at each level that not only increases the complexity of the question, but also requires an additional reasoning step to eliminate. This is achieved through 3 sub-steps, as described in Fig. 5.

4. Commonsense Implicitation: As commonsense knowledge is implicit knowledge about the world that is often unspoken but assumed, this step reduces the context exposed to the LLMs by combining the context and question into a *context-implicit commonsense question*. To answer the modified question, models will have to draw on their internal common knowledge to determine the correct answer, especially when the topic requires more than just logical reasoning.

The whole procedure is illustrated in Fig. 1: After the initial creation of Level 0 instances (Step 1 and 2), Step 3 (Data Complexity Scaling) is iteratively applied to generate questions of increasing complexity. Finally, Step 4 (Commonsense Implicitation) is applied to each generated instance to produce the final context-implicit question for that level. This approach helps mitigate the data leakage (Deng et al., 2024) and shortcut reasoning (Haraguchi et al., 2023) problems, as observed in our experimental results where performance degrades significantly with each level. Furthermore, the scaled complexity forces LLMs to utilize their reasoning capacity more extensively, enabling deeper investigation into their reasoning process. Detailed examples with complete prompts are provided in our prompt detail appendix.

3.2.2. mSCoRe-S: Social Commonsense

There is still a gap in current commonsense benchmarks in terms of social commonsense knowledge and cultural norms (Davis, 2024). To provide a comprehensive evaluation of LLMs' commonsense reasoning capacity, we propose an additional benchmark **mSCoRe-S** that revolves around social situations across diverse cultural contexts. In particular, we utilize CultureBank (Shi et al., 2024) as our seed dataset, which is a knowledge base containing real-world social questions sourced from TikTok and Reddit posts. Each instance in CultureBank is provided with various descriptors containing details about the cultural group, context, behaviors, and an agreement level indicating how widely accepted that behavior is within the community (Fig. 2).

Each seed instance follows the same 4-step process as described in previous section to generate the final *context-implicit commonsense question*. However, minor differences are introduced to adapt to CultureBank data, including:

Seed Data Filtering: In addition to the 3 criteria used for **mSCoRe-G**, we introduce an additional criterion - **multiculture-ness** - for filtering social situations (detailed description provided in our data generation detail appendix). The aim is to select situations that involve the most culturally distinctive elements, allowing us to evaluate models' understanding of diverse cultural contexts and associated commonsense knowledge.

Structured Reasoning Generation: Before generating the context and reasoning process, LLM needs to generate the QA pair first. This acts as the seed QA pair from mCSQA in the previous section and goes through the same procedure.

3.2.3. Dataset Statistic

mSCoRe-G covers 5 languages including English, German, French, Japanese, and Chinese. For each language, we create 200 examples ranging from level 0 (original QA pair) to level 3 (3 steps of expansion). This results in 800 examples per language. For **mSCoRe-S**, we similarly create 200 examples for each source (TikTok and Reddit). In total, **mSCoRe** contains 5,600 instances (4000 for general commonsense and 1600 for social commonsense). Detailed examples at different complexity levels are provided in our prompt detail appendix.

3.3. Human Validations

To ensure the reliability and quality of our automated data generation pipeline, we conducted a multi-faceted human validation study on a randomly sampled subset of 100 instances from the English portion of mSCoRe-G, spanning all complexity levels (L0 to L3). The validation focused on three key aspects of our methodology: the accuracy of skill classification, the effectiveness of complexity

Descriptors	Definitions	Examples
Cultural Topic	Cultural group - topic - scenario	Japanese culture - Gift Giving - Etiquette and Practices
Social Context	Settings the behavior takes place.	During a meeting in Japan, a visiting Western executive wants to express gratitude to their hosts
Actor	Who exhibit the behavior	Visiting executive
Question	The commonsense question regarding the actor's behavior	I'm attending a meeting in Japan and would like to give a gift to my hosts. What should I consider to ensure my gesture is well-received?
Actor Behavior	Behavior of the actor	Offer a gift wrapped in traditional Japanese style as a gesture of appreciation
Recipient	Recipient of the action	Japanese business hosts
Relation	Relation between the actor and the recipient	Business partners
Recipient Behavior	Behavior of the recipient	Receive the gift with both hands and show appreciation

Table 2: An example of a social commonsense question from CultureBank.

scaling, and the integrity of the commonsense implicitation step.

Reasoning Skill Validation: Annotators were tasked with independently classifying the primary reasoning skill for each atomic reasoning step. Our evaluation shows a pass@1 accuracy of 78% (the annotator's top choice matched the LLM-generated skill) and a pass@3 accuracy of 97%. This indicates a strong alignment between the LLM's skill categorization and human judgment, confirming that our taxonomy is practically applicable and the generated labels are meaningful.

Complexity Scaling Validation: Annotators evaluated whether each scaling step (from L0 to L3) successfully increased the question's complexity by introducing plausible distractors that required additional, valid reasoning steps to eliminate. The validation yielded agreement rates of 93% for L1, 87% for L2, and 85% for L3. These results confirm that our scaling framework effectively increases task difficulty, while also highlighting the inherent challenge of creating coherent, high-complexity questions automatically.

Commonsense Implicitation Validation: Finally, annotators assessed whether the final, context-implicit questions preserved the original question's semantic intent and correctly required the use of implicit world knowledge to answer. This step achieved an agreement rate of 98%, validating that the transformation successfully tests internalized commonsense knowledge.

4. Experiments

4.1. Experiment Setup

We conduct comprehensive evaluations using a diverse set of state-of-the-art multilingual language models, selected to represent different approaches to model development and training. Our evaluation considers three key dimensions: model availability, parameter scale, and training methodology. The models evaluated in our study include: **GPT-4o** (OpenAI et al., 2024): A general-purpose LLM representing the current state-of-the-art LLM, trained on large-scale multimodal data from diverse sources. OpenAI **o1** (Jaech et al., 2024): A reasoning-reinforced model based on GPT-4o, specifically optimized for complex problem-solving tasks through an additional training phase utilizing data curated for chain-of-thought reasoning. **LLaMA-3.3-70B** and **LLaMA-3.1-8B** (Grattafiori et al., 2024): Two open-sourced LLMs representing different parameter scales, trained on publicly available sources spanning various domains, allowing us to analyze the impact of model size on reasoning capabilities. Distilled DeepSeek-R1 (**R1-70B** and **R1-8B**) (DeepSeek-AI et al., 2025): A reasoning-focused model derived from the LLaMA architecture, distilled using samples generated by the large-scale LRM DeepSeek-R1. **Aya-32B** (Dang et al., 2024): A universal multilingual model trained on data from 200 languages, providing insights into broad multilingual LLM reasoning capabilities. For evaluation, we employ a consistent prompt for all models, providing the proposed reasoning skill taxonomy with step-by-step instructions to generate the desired reasoning process before answering. Further experimental details are in our experimental detail appendix.

4.2. Main Results

Tables 3 and 4 present our main results on **mSCoRe-G** and **mSCoRe-S**, respectively. Overall, we observe a consistent pattern across all models where performance declines as complexity level increases. For **mSCoRe-G**, GPT-4o achieves the highest overall accuracy on general commonsense reasoning across all languages and complexity levels. While this can be an artifact of the benchmark creation process where GPT-4o was used for data generation, LLaMA-3.3-70B results are very close to GPT-4o. Furthermore, the open-source model significantly outperforms others on social commonsense reasoning (over 5% average improvement across all levels and domains).

Multilingual and Cultural Results: Performance is generally similar across languages in **mSCoRe-G**. This may be due to all languages in the seed dataset mCSQA being medium to high resource languages. Future work should explore other seed datasets with more low-resource languages. For social commonsense reasoning in **mSCoRe-S**, most models perform better on Reddit-sourced questions than TikTok-sourced ones. This could be attributed to Reddit containing more content on general "Community and Cultural Exchange", whereas TikTok focuses more on daily life "personal" aspects like "Social Norms and Eti-

General Commonsense	English				German				French				Chinese				Japanese				Average			
	L0	L1	L2	L3	L0	L1	L2	L3	L0	L1	L2	L3	L0	L1	L2	L3	L0	L1	L2	L3	L0	L1	L2	L3
GPT-4o	**80.5**	70.0	**72.5**	**71.5**	**75.0**	**68.5**	**71.0**	67.5	78.0	**74.0**	**70.0**	63.5	**80.5**	**78.5**	**72.5**	**65.5**	**82.0**	**83.5**	**79.5**	**79.5**	**79.2**	**74.9**	**73.1**	**69.5**
o1	**82.5**	**73.5**	**75.0**	**72.0**	**75.0**	67.5	63.0	67.5	**80.5**	**72.5**	**71.5**	61.0	64.5	63.0	56.0	53.0	80.5	80.0	77.0	73.0	76.6	71.3	68.5	65.3
o1-mini	76.5	70.5	65.5	63.5	69.5	66.0	**69.5**	64.5	71.5	64.5	59.5	55.0	71.0	63.0	60.0	51.5	77.5	75.0	68.5	66.5	73.2	67.8	64.6	60.2
LLaMA-3.3-70B	78.5	**75.0**	69.0	70.0	**75.5**	**72.5**	68.0	**73.0**	**78.5**	72.0	67.0	**64.0**	**80.0**	**74.5**	**70.5**	**67.0**	**82.0**	**85.5**	**76.5**	**78.0**	**78.9**	**75.9**	**70.2**	**70.4**
LLaMA-3.1-8B	23.0	22.5	21.5	21.5	73.0	65.5	63.0	61.0	69.5	61.0	54.5	52.0	60.0	52.0	46.0	43.0	17.5	18.5	17.0	17.5	48.6	43.9	40.4	39.0
R1-70B	79.5	70.5	69.5	69.0	73.0	67.0	67.0	**70.0**	76.0	71.5	69.5	**64.5**	75.0	70.5	61.0	65.0	**83.0**	79.5	72.0	73.5	77.3	71.8	67.8	68.4
R1-8B	67.5	62.0	62.0	55.0	67.5	58.0	61.0	55.5	58.0	45.0	44.0	43.5	69.0	62.0	51.5	58.5	61.5	57.0	59.0	53.5	64.7	56.8	55.5	53.2
Aya-32B	77.5	67.0	66.5	66.0	70.5	65.5	66.5	66.0	76.5	69.0	65.0	60.5	78.0	67.0	64.0	60.0	79.5	80.5	70.0	72.5	76.4	69.8	66.4	65.0

Table 3: Accuracy comparison of models on **mSCoRe-G** from complexity level 0 (L0) to 3 (L3). Best results are shown in **bold underline**, second best in **bold**.

Social Commonsense	TikTok				Reddit				Average			
	L0	L1	L2	L3	L0	L1	L2	L3	L0	L1	L2	L3
GPT-4o	**71.0**	**69.0**	62.5	**63.5**	75.0	67.0	**68.5**	**69.5**	73.0	68.0	**65.5**	**66.5**
o1	69.5	**69.0**	**63.5**	61.5	**77.0**	**71.0**	67.5	69.0	**73.3**	**70.0**	**65.5**	65.3
o1-mini	63.5	62.5	53.0	59.5	72.5	62.0	62.0	59.0	68.0	62.3	57.5	59.3
LLaMA-3.3-70B	**80.0**	**75.0**	**73.5**	**73.0**	**83.5**	**76.5**	**80.0**	**76.5**	**81.8**	**75.8**	**76.8**	**74.8**
LLaMA-3.1-8B	30.5	29.5	29.5	29.5	27.0	27.0	27.0	27.0	28.8	28.3	28.3	28.3
R1-70B	68.5	65.5	62.0	62.5	73.5	67.0	67.5	67.5	71.0	66.3	64.8	65.0
R1-8B	64.0	60.0	56.5	52.5	65.0	58.5	60.5	61.0	64.5	59.3	58.5	56.8
Aya-32B	68.0	64.0	61.0	60.5	71.0	57.5	63.0	59.5	69.5	60.8	62.0	60.0

Table 4: Accuracy comparison of models on **mSCoRe-S**.

quette". This suggests that LLMs might still struggle with more personalized problems, as noted in (Davis, 2024). Unexpectedly, despite being trained on 200 diverse languages, the most multilingual model Aya-32B does not perform very well, even in cultural social commonsense benchmark.

Model Scale: We compare models with different parameter counts, from the 8B and 70B parameters open-source models (LLaMA and R1), to the colossal-scale (hundreds of billions of parameters) closed-source LLMs (GPT-4o and o1). Larger models generally perform better across both benchmarks. The performance gap between 70B and 8B versions was substantial in most cases. However, we observe diminishing returns when moving from 70B to colossal-scale LLMs. This finding suggests it takes more than simple parameter scaling to solve commonsense reasoning, especially in understanding social interactions and cultural norms.

Reasoning-reinforced Training: We compare general instruction-tuned models (GPT-4o, LLaMA) with reasoning-reinforced fine-tuned models (o1 and R1). While the state-of-the-art LRM o1 performs best in English, it lags behind other general LLMs like GPT-4o and LLaMA-3.3-70B in other languages. This suggests that reasoning-reinforced training might decrease commonsense reasoning ability, likely due to the highly specialized training data for more complex tasks like coding and math. Interestingly, LLaMA-3.1-8B fails the task in English and Japanese, but R1-8B performs normally, indicating that reasoning-reinforced training helps smaller-scale models understand tasks better.

5. Analysis

	General (English)							Social (Average)						
	L0	L1	L2	L3	L4	L5	L6	L0	L1	L2	L3	L4	L5	L6
GPT-4o	**80.5**	70.0	**72.5**	**71.5**	**72.0**	**70.0**	**68.0**	73.0	68.0	**65.5**	**66.5**	**66.3**	**67.3**	**66.0**
o1	**82.5**	**73.5**	**75.0**	**72.0**	**70.0**	**69.5**	67.0	**73.3**	**70.0**	**65.5**	65.3	62.8	60.8	63.0
LLaMA-3.3-70B	78.5	**75.0**	69.0	70.0	68.5	68.5	**69.0**	**81.8**	**75.8**	**76.8**	**74.8**	**75.5**	**72.5**	**73.3**
LLaMA-3.1-8B	23.0	22.5	21.5	21.5	20.5	20.5	20.5	28.8	28.3	28.3	28.3	28.0	28.0	28.0
Deepseek-70B	79.5	70.5	69.5	69.0	69.0	69.0	65.5	71.0	66.3	64.8	65.0	60.3	62.8	65.8
Deepseek-8B	67.5	62.0	62.0	55.0	57.0	59.5	55.0	64.5	59.3	58.5	56.8	59.3	59.8	61.5
Aya-32B	77.5	67.0	66.5	66.0	58.0	60.0	54.5	69.5	60.8	62.0	60.0	57.8	56.5	54.0

Table 5: Performance comparison from complexity level 0 to level 6.

5.1. Complexity Scaling Results

To further understand model capacity against scaling question complexity, we expand our results for **mSCoRe-G** (English) and **mSCoRe-S** to complexity level 6. As shown in Table. 5, every model accuracy continues to decline to L6. The most significant performance drop occurs between L0 and L2, indicating that even relatively simple complexity scaling introduces substantial challenges for LLMs. At higher difficulty levels (L3 to L6), the rate of degradation slows down considerably. This plateau suggests that our current approach to scaling complexity through additional context and reasoning steps may reach a saturation point. This could indicate that the multiple-choice question-answer format itself imposes certain limitations on how effectively task difficulty can be scaled. Alternative task formulations that require more sophisticated forms of reasoning beyond the current design might be necessary to create more discriminative benchmarks for future, more capable models.

5.2. Skill Type Utilization

To better understand how models employ different reasoning skills across varying complexity levels, Fig. 6 visualizes the distribution of reasoning skills used in both the reference reasoning processes (from our benchmark creation) and the output reasoning processes generated by o1.

For *general* commonsense, both reference and model-generated reasoning primarily utilize logical reasoning skills, with deductive reasoning being most common. However, the reference distribution shows greater diversification of skills at higher complexity levels, incorporating more contextual reasoning (especially analogical and probabilistic reasoning). In contrast, models like o1 remains heavily dependent on deductive reasoning across all complexity levels. For *social* commonsense, the reference distribution shows more balanced utilization of skills from all three categories, with social and ethical reasoning becomes progressively more important for higher-level questions. While o1 model incorporates some social reasoning skills, it still over-relies on logical reasoning for scenarios where social and contextual reasoning would be more appropriate. Overall, results reveal significant limitations in o1's ability to adapt its reasoning strategy. The rigid reasoning pattern likely explains the

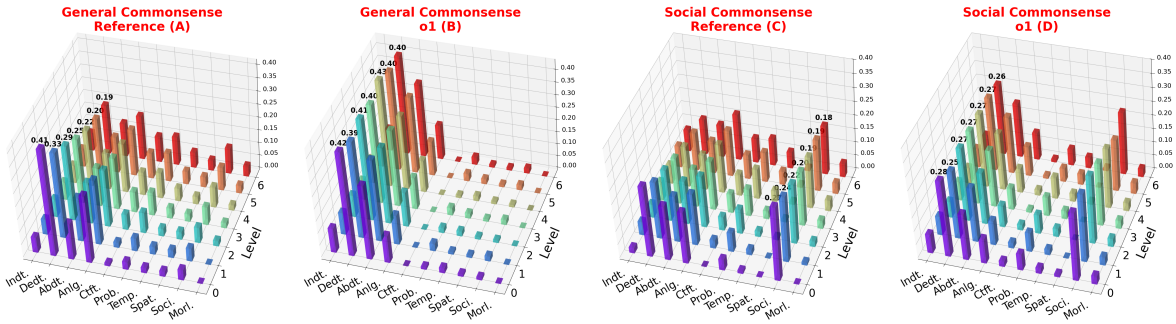

Figure 6: The distribution of reasoning skills of reference reasoning process (**A** and **C**), and o1's reasoning process (**B** and **D**), as question complexity increases from complexity level 0 to 6.

model's performance decrease on higher complexity questions, highlighting the need for more balanced reasoning-reinforced training approaches.

	General				Social			
	L0	L1	L2	L3	L0	L1	L2	L3
o1	76.6	**71.3**	**68.5**	65.3	**73.3**	**70.0**	**65.5**	**65.3**
o1-mini	73.2	67.8	64.6	60.2	68.0	62.3	57.5	**59.3**
cot-o1	75.9	69.3	66.2	61.3	63.3	49.3	44.5	40.3
cot-o1-mini	71.7	65.2	60.2	57.8	60.8	51.5	46.5	45.3
logical-o1	**77.3**	**72.1**	66.3	**65.6**	**72.8**	**64.3**	**59.8**	58.3
logical-o1-mini	73.9	68.3	62.5	62.2	64.8	59.8	59.3	50.5
general-o1	**77.7**	**69.9**	**67.5**	**65.8**	69.3	54.5	51.5	48.3
general-o1-mini	73.3	67.7	61.4	59.6	66.8	61.8	57.0	52.8

Table 6: Results for Different Reasoning Skill Taxonomies.

5.3. Different Reasoning Skill Taxonomies

We investigate how models adapt to different reasoning taxonomies, including: (1) **Chain-of-Thought (CoT)** - Standard chain-of-thought, not requiring skill identification, (2) **Logical** - Only using logical reasoning skills (deductive, inductive and abductive) (3) **General** - Each reasoning step is categorized into one of the three general categories (logical, contextual, and social).

Table 6 shows the average accuracies of o1 and o1-mini for each setting. Interestingly, despite requiring models to distinguish between more skill types, our proposed fine-grained taxonomy yields the best results. As expected from our previous analysis, the **Logical**-only approach performs relatively well on general commonsense tasks but worse on social tasks. The **General** setting also under-performs ours, suggesting that granularity of skill identification benefits commonsense reasoning by encouraging models to consider a broader range of reasoning approaches rather than defaulting to familiar patterns. Finally, **CoT** performs notably worse than all structured skill-based approaches, especially for social commonsense at higher complexity levels. This demonstrates that reasoning without explicit skill categorization may be insufficient for more complex commonsense situations.

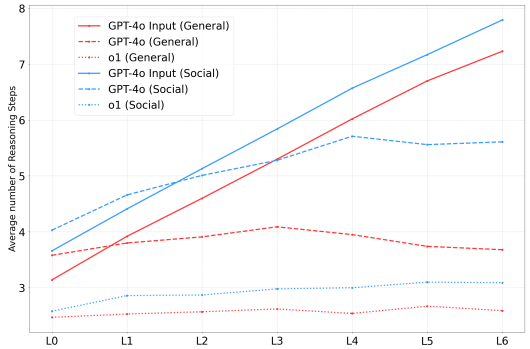

Figure 7: Average number of reasoning steps in the reasoning processes of **mSCoRe**(straight), GPT-4o (barred), and o1 (dotted).

5.4. Reasoning Efficiency

To examine the relationship between reasoning efficiency and task complexity across different models, Fig. 7 visualizes the average number of steps of the reasoning processes of **mSCoRe** and GPT-4o and o1's answers across different complexity levels.

The reference reasoning processes show a clear linear increase in reasoning steps as complexity level increases, with social commonsense reasoning requiring more steps than general commonsense at each level. GPT-4o's reasoning processes show a similar upward trend but with a more gradual slope, whereas o1's reasoning processes maintain a nearly constant step count (around 3 steps) regardless of task complexity. The results indicate higher-level complexities require more steps, and current models are unable to reason longer, unless explicitly forced to (such as in the Complexity Expansion step described in Section 3.2). This is similar to the number of reasoning tokens (more steps is equivalent to more tokens) used in test-time scaling research recently introduced in (Muennighoff et al., 2025). These findings indicate that adapting reasoning depth dynamically based on task demands is likely crucial for sustaining performance as complexity escalates.

6. Conclusion

We introduce **mSCoRe** - Multilingual and Scalable Benchmark for Skill-based Commonsense Reasoning, a novel evaluation framework designed to

address critical gaps in existing benchmarks for commonsense reasoning. By integrating multilingual and diverse cultural coverage, a fine-grained reasoning skill taxonomy, and a dynamic complexity scaling mechanism, **mSCoRe** provides a comprehensive platform for systematically evaluating not only the accuracy but also skill utilization and efficiency of LLMs' commonsense reasoning process. Extensive experiments on eight state-of-the-art LLMs reveal that current models still consistently struggled with higher complexity levels and culturally nuanced social commonsense scenarios. Our analysis highlights several promising directions for improvement, including more robust training methodologies to enhance model's reasoning skill utilization and efficiency. Additionally, **mSCoRe** provides a framework for subsequent benchmarks to scale with the rapid development of LLMs in the future.

7. Bibliographical References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17682–17690. AAAI Press.

Sijia Chen, Baochun Li, and Di Niu. 2024. Boosting of thoughts: Trial-and-error problem solving with large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. 2025. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *The Thirteenth International Conference on Learning Representations*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, and et al. Tom Kocmi. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Ernest Davis. 2024. Benchmarks for automated commonsense reasoning: A survey. *ACM Comput. Surv.*, 56(4):81:1–81:41.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and et al. Aixin Liu. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Benchmark probing: Investigating data leakage in large language models. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning - The Good, the Bad, and the Ugly*.

Quyet V. Do, Junze Li, Tung-Duong Vuong, Zhaowei Wang, Yangqiu Song, and Xiaojuan Ma. 2024. What really is commonsense knowledge? *CoRR*, abs/2411.03964.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in AI. *CoRR*, abs/2411.04872.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and et al. Aurelien Rodriguez. 2024. The llama 3 herd of models.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. Deepseek-coder:

When the large language model meets programming - the rise of code intelligence. *CoRR*, abs/2401.14196.

Daichi Haraguchi, Kiyoaki Shirai, Naoya Inoue, and Natthawut Kertkeidkachorn. 2023. Discovering highly influential shortcut reasoning: An automated template-free approach. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6401–6407. Association for Computational Linguistics.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3828–3850. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. 2025. Advancing language model reasoning through reinforcement learning and inference scaling. *CoRR*, abs/2501.11651.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Kshitij Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. 2025. Big-bench extra hard. *CoRR*, abs/2502.19187.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Salman Khan, and Fahad Shahbaz Khan. 2025. Llm post-training: A deep dive into reasoning large language models.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. Culturepark: Boosting cross-cultural understanding in large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *CoRR*, abs/2501.19393.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and et al. Alex Paino. 2024. Gpt-4o system card.

Sekta Lonir Oscarini and Wati Bhakti. 2010. Bloom's taxonomy: Original and revised.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 4996–5025. Association for Computational Linguistics.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *CoRR*, abs/2408.03314.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking Chinese commonsense reasoning of LLMs: From Chinese-specifics to reasoning-memorization correlations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11205–11228, Bangkok, Thailand. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13003–13051. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

A. T. Welford. 1986. Taxonomies of human performance, edwin a. fleishman and marilyn k. quaintance, academic press, orlando, florida, 1984. no. of pages: xvi + 514. price: $49. *Journal of Organizational Behavior*, 7(2):155–156.

Wikipedia contributors. 2025. Commonsense knowledge (artificial intelligence) — Wikipedia, the free encyclopedia. [Online; accessed 25-March-2025].

Dingjun Wu, Jing Zhang, and Xinmei Huang. 2023. Chain of thought prompting elicits knowledge augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6519–6534. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *CoRR*, abs/2411.14405.

A. Experimental Details

LLMs Details: For closed commercial LLMs (GPT-4o, o1 and o1-mini), we query responses from the models using OpenAI Chat Completions API[1], with temperatures set to 0 for deterministic outputs. Open source models (Deepseek R1-70B[2] and R1-8B[3], LLaMA-3.3-70B[4] and LLaMA-3.1-8B[5], Aya-32B[6]) are run using 2 NVIDIA A100 80GB GPUs. PyTorch 2.1.2[7] and Huggingface-Transformer 4.42.3 [8] are used to implement the models.

Source code with specification of all dependencies, including external libraries: Our data will be released in https://github.com/nghiatrngo/mSCORE.

A.1. Reasoning Skill Details

We provide detail descriptions with abstract and concrete example for each of our reasoning skills in Fig. 8. Abstract examples are generalized representation that uses variables or placeholders to illustrate a pattern or principle of the reasoning skills. In contrast, concrete examples are the application of the corresponding reasoning skills to a specific real-world scenario.

B. Data Generation Details

B.1. LLM-Judge

Judge Model We utilize Flow Judge, a general LLM-as-a-judge model developed by Flow AI[9]. Flow Judge is an open-source 3.8B parameter language model designed for LM-based evaluations, offering high performance and accuracy comparable to much larger models like GPT-4o and Claude 3.5 Sonnet. It is trained on evaluation data across various domains to supports custom evaluation criteria, multiple scoring scales, qualitative feedback, and produces structured evaluation outputs.

Scoring Metrics for Seed Data Filtering step: We provide the full rubrics used for Seed Data Filtering step for **mSCoRe-G** and **mSCoRe-S** in Fig. 9 and 10, correspondingly.

C. Prompt Details

We provide here all of the prompt templates in full version used in our experiments. Fig. 11, 12, and 13 presents the prompt of Structured Reasoning Generation, Data Complexity Scaling and Commonsense Implicitation steps in our data generation process for **mSCoRe-G**. Additionally, Fig. 15 presents the prompt of Structured Reasoning Generation step for **mSCoRe-S** (The other two steps remain the same between the two subsets).

We provide an example of complexity level 0 to 3 for **mSCoRe-G** and **mSCoRe-S** in Fig. 14 and 16, correspondingly.

[1] https://platform.openai.com/docs/guides/text-generation

[2] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B

[3] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

[4] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[5] https://huggingface.co/meta-llama/Llama-3.1-8B

[6] https://huggingface.co/CohereForAI/aya-expanse-32b

[7] https://pytorch.org/get-started/pytorch-2.0/

[8] https://github.com/huggingface/transformers

[9] https://www.flow-ai.com/blog/flow-judge

Inductive Reasoning: Drawing general conclusions from specific observations.
- Description: Inductive reasoning is a method of drawing general conclusions from specific observations. Unlike deductive reasoning, which starts with general premises to reach specific conclusions, inductive reasoning begins with detailed facts and builds up to broader generalizations or theories. This approach is commonly used in scientific research, where repeated experiments and observations lead to the formulation of overarching principles or hypotheses.
- Abstract Example: After witnessing several instances where Event A_1 leads to Event A_2, you infer that Event A_n will similarly result in Event A_2 in future occurrences.
- Concrete Example: After witnessing several instances where the weather forecast predicts rain, you infer that rain will likely continue to fall in the future.

Deductive Reasoning: Deriving specific conclusions from general premises.
- Description: Deductive reasoning involves deriving specific conclusions from general premises. It ensures that if the premises are true and the reasoning is valid, the conclusion must also be true. Deductive logic is fundamental in fields that require rigorous proof, such as mathematics and formal sciences.
- Abstract Example: Given the premise that All X are Y, and knowing that Object x_1 is an X, you deduce that Object x_1 must also be a Y.
- Concrete Example: All birds have feathers. A sparrow is a bird. Therefore, a sparrow has feathers.

Abductive Reasoning: Forming hypotheses to explain observations.
- Description: Abductive reasoning is the process of forming hypotheses to explain observations. It starts with an incomplete set of observations and proceeds to the likeliest possible explanation. Unlike deductive and inductive reasoning, abductive reasoning seeks the simplest and most plausible explanation for a given set of facts, often leading to the generation of new theories or hypotheses.
- Abstract Example: Observing Event B, you hypothesize that Reason 2 is the most plausible explanation among several possible causes.
- Concrete Example: You wake up and see that the street is wet. The most likely explanation is that it rained last night.

Analogical Reasoning: Drawing parallels between similar situations to infer conclusions.
- Description: Analogical reasoning involves drawing parallels between similar situations to infer conclusions. By comparing two objects or systems that share certain characteristics, one can infer that they may share additional, unobserved properties. This form of reasoning is widely used in problem-solving, scientific discovery, and legal reasoning to transfer knowledge from a known domain (source) to an unknown domain (target).
- Abstract Example: Think of Situation C_a, where Component C_a_1 interacts with Component C_a_2 in a specific way. You encounter Situation C_b with Component C_b_1 and Component C_b_2, and infer that they will interact similarly.
- Concrete Example: Just as a gardener waters plants to help them grow, a teacher provides knowledge and guidance to help students develop.

Counterfactual Reasoning: Considering alternative scenarios and outcomes that did not happen.
- Description: Counterfactual reasoning entails considering alternative scenarios and outcomes that did not occur. It involves imagining 'what might have happened' under different circumstances, which is useful for understanding causality, evaluating decisions, and planning future actions.
- Abstract Example: Reflecting on Condition X that did not occur, you imagine that if it had, Outcome Y might have replaced Outcome Z.
- Concrete Example: If you had left the house five minutes earlier, you would have caught the bus on time.

Probabilistic Reasoning: Applying principles of probability to make inferences under uncertainty.
- Description: Probabilistic reasoning involves applying principles of probability to make inferences under uncertainty. It enables individuals to assess the likelihood of different outcomes and make informed decisions based on the probability of various events occurring.
- Abstract Example: Evaluating that Option A has a higher probability (P(A) > P(B)) of success than Option B, you decide to choose Option A.
- Concrete Example: There is a 70% chance of rain tomorrow, so you decide to carry an umbrella when you go out.

Temporal Reasoning: Understanding sequences and durations of events.
- Description: Temporal reasoning is the ability to understand and reason about the sequence and duration of events over time. It involves comprehending time-specific data, such as the order of events, how long events last, and the relationships between different time points.
- Abstract Example: Planning your day, you schedule Event T_1 to occur before Event T_2, ensuring the correct sequence of activities.
- Concrete Example: You observe that the sun will rise in the morning and set in the evening. You infer that the moon will rise and set at the same time.

  * Spatial Reasoning: Visualizing and manipulating objects in space.
- Description: Spatial reasoning entails visualizing and manipulating objects in space. It involves understanding the relationships between different objects, such as their position, orientation, and movement relative to each other. Spatial reasoning is fundamental in fields like engineering, architecture, geography, and various forms of visual arts, enabling individuals to solve problems related to the physical arrangement and movement of objects.
- Abstract Example: While arranging furniture, you visualize Object S_1 and Object S_2 to determine their optimal placement within the room.
- Concrete Example: An architect determining the best location for a window by visualizing the window and the surrounding walls to determine the optimal angle and height.

Social Reasoning: Understanding social interactions and norms.
- Description: Social reasoning involves understanding social interactions and norms. It encompasses the ability to analyze and interpret social situations, recognize appropriate and inappropriate behaviors, and predict others' intentions, emotions, and thoughts. Effective social reasoning is crucial for building successful interpersonal relationships and navigating complex social environments.
- Abstract Example: Noticing that Person A behaves a certain way in Situation S, you adjust your own behavior (Behavior B) to interact effectively.
- Concrete Example: You notice that your friend looks upset after a conversation, so you decide to ask them if they are okay.

Moral Reasoning: Deciding what is right or wrong based on ethical principles.
- Description: Moral reasoning is the process of deciding what is right or wrong based on ethical principles. It involves evaluating actions, intentions, and consequences to make judgments about moral issues. Moral reasoning is central to ethical decision-making and is influenced by various factors, including societal norms, personal values, and philosophical theories.
- Abstract Example: Considering that Action M could harm Person C, you decide it is morally wrong and choose an alternative that respects ethical principles.
- Concrete Example: Seeing someone drop their wallet, you decide to return it instead of keeping the money inside because it is the right thing to do.

Figure 8: Reasoning skill details.

```
### Commonsense-ness
Task: Evaluate the 'Commonsense-ness' of a multiple-choice commonsense question.
Evaluation Criteria: Does answering the question rely solely on commonsense knowledge accessible to the general population, or does it require formal reasoning and specialized expertise beyond everyday understanding?
- Score 1: The question requires formal reasoning and specialized expertise to answer correctly. It demands advanced knowledge in a specific field, technical terminology, or in-depth understanding that goes beyond general life experience. The average person, relying only on commonsense knowledge, would find it challenging or impossible to select the correct answer without additional study or expertise.
- Score 2: The question can be addressed with some commonsense reasoning but may also require moderate specific knowledge or logical deduction. While not entirely dependent on formal expertise, it involves concepts or facts that are not universally known but can be reasoned through by an informed individual. The average person might answer correctly with thoughtful consideration but could also be misled without careful analysis.
- Score 3: The question is answerable using basic commonsense knowledge that is widely shared and understood by the general population. It does not rely on any specialized information or formal reasoning processes. The correct answer should be apparent to most people through everyday experience and general understanding of the world.

### Complexity
Task: Evaluate the 'Hardness/Complexity' of a commonsense question.
Evaluation Criteria: How difficult is the question to understand and answer? Does it require minimal reasoning or a complex, multi-step thought process to identify the correct answer?
- Score 1: The question is very easy to understand, and the correct answer can be quickly identified with a single, straightforward reasoning step. It requires minimal cognitive effort, and most individuals can arrive at the correct answer almost immediately without confusion.
- Score 2: The question is relatively easy to understand, requiring only a couple of straightforward reasoning steps to identify the correct answer. While the question may introduce one or two elements that require brief consideration, the overall context remains clear. Most people can find the correct answer with a small amount of thought.
- Score 3: The question is moderately challenging, necessitating several reasoning steps to accurately comprehend and resolve. It introduces multiple elements or scenarios that require a careful thought process to integrate and analyze. Many individuals will need to pause and deliberately work through the connections or implications before reaching the correct answer.
- Score 4: The question is hard to comprehend and necessitates a complex thought process with multiple reasoning steps. It may involve abstract concepts, less obvious relationships, or misleading information that requires careful analysis. Individuals must invest significant cognitive effort to work through the complexities and identify the correct answer.
- Score 5: The question is very hard to comprehend and requires a long reasoning process with multiple reasoning steps to find the right answer. It demands high-level critical thinking, problem-solving skills, and possibly specialized knowledge. Only with thorough analysis and persistence can individuals navigate the complexity to arrive at the correct answer.

### Expandability
Task: Evaluate the 'Expandability' of a commonsense question.
Evaluation Criteria: To what extent can the question be expanded or elaborated upon to introduce additional complexity or dimensions?
- Score 1: The question cannot be expanded. It is inherently simplistic and covers a very narrow topic or scenario. There is little to no room for introducing additional elements, dimensions, or complexity without altering the fundamental nature of the question. The question stands effectively as a self-contained unit with minimal potential for elaboration.
- Score 2: The question has some potential for expansion. While it currently covers its intended scope adequately, there is moderate room to add a few additional elements or explore related themes that could introduce more complexity. The question can be expanded moderately by incorporating extra conditions, perspectives, or related scenarios, but such additions are not numerous.
- Score 3: The question can be significantly expanded to become a more complex question. It has ample scope for adding new dimensions, scenarios, or layers of reasoning. By introducing additional variables, conditional information, or intricate details, the question can transform into a more challenging problem that requires advanced reasoning and deeper comprehension.
```

Figure 9: Rubrics used for mCSQA Data Filtering Process

```
### Multiculturenes
Task: Evaluate the 'Multicultural-ness' of a commonsense cultural situation.
Evaluation Criteria: Does the situation involve interactions between multiple distinct cultures, reflecting a blend of practices, norms, or etiquette from each?
- Score 1: The situation is primarily rooted in a single culture, without significant influence or interaction from other cultural norms or practices. The interactions and behaviors exhibited are almost exclusively aligned with one cultural tradition, lacking a blend of cultural elements or considerations from another distinct culture.
- Score 2: The situation involves elements from two cultures, showing some level of cross-cultural interaction. While both cultural influences are present, the interaction may largely reflect the dominance of one culture over the other, with limited integration or blending of unique practices, norms, or etiquette from both cultures.
- Score 3: The situation reflects a rich blend of cultural interactions involving more than two distinct cultures. It demonstrates a balanced integration of diverse cultural practices, norms, or etiquette. The interactions and behaviors of the parties involved show a deep understanding and appreciation of multiple cultural perspectives, leading to an enriching multicultural exchange.

### Commonsenseness
Task: Evaluate the 'Commonsense-ness' of a cultural situation.
Evaluation Criteria: To what extent can the situation be understood and addressed using basic commonsense knowledge, without requiring specialized or expert reasoning?
- Score 1: The situation requires formal reasoning and specialized expertise to understand and address appropriately. It involves complex cultural nuances or specific knowledge that goes beyond general commonsense understanding. Responding effectively necessitates familiarity with detailed cultural protocols or insider knowledge.
- Score 2: The situation can be partially addressed using commonsense knowledge, but some elements require a deeper understanding or contextual insights that may not be readily apparent to someone without specific cultural awareness. While general reasoning can guide some actions, certain aspects benefit from additional cultural knowledge or experience.
- Score 3: The situation can be appropriately addressed using basic commonsense reasoning. It involves straightforward cultural interactions that do not demand specialized knowledge. Commonsense understanding of general social norms and human interactions is sufficient to respond suitably and effectively in this context.

### Complexity
Task: Evaluate the 'Complexity' of a cultural situation.
Evaluation Criteria: How intricate is the cultural situation in terms of nuances, number of cultural elements, perspectives, social dynamics, and interactions, requiring varying depths of understanding to navigate appropriately?
- Score 1: The situation is very simple, involving a single cultural aspect with straightforward practices and minimal perspectives or interactions. Understanding and responding require little to no specialized knowledge or awareness of cultural nuances.
- Score 2: The situation has minor complexity, incorporating a couple of cultural elements or perspectives with basic interactions. There are some cultural nuances, but they are easily understood with general awareness. Navigating the situation may require modest cultural sensitivity but is generally manageable.
- Score 3: The situation is moderately complex, involving several cultural elements, multiple perspectives, and noticeable social dynamics. Understanding and responding appropriately require some cultural knowledge and sensitivity to nuances. There is potential for misunderstandings without a moderate level of cultural competence.
- Score 4: The situation is complex, featuring numerous cultural elements, diverse perspectives, intricate social dynamics, and significant interactions. Navigating the situation effectively necessitates considerable cultural competence, an awareness of subtle nuances, and an understanding of how different cultural norms might conflict or interact.
- Score 5: The situation is highly complex, encompassing a multitude of deeply intertwined cultural elements, perspectives, and interactions. It includes profound cultural nuances, ambiguous social cues, and a high potential for misunderstandings. Expert knowledge and significant experience are required to address it appropriately, as the situation may involve conflicting norms and requires advanced cultural navigation skills.

### Expandability
Task: Evaluate the 'Expandability' of a cultural situation.
Evaluation Criteria: Assess the potential for the situation to be expanded by including additional cultural dimensions, participants, interactions, and its adaptability to different contexts.
- Score 1: The situation is tightly defined within a single cultural framework, offering little room for the addition of new cultural dimensions. It does not easily support additional participants or interactions, requiring significant adaptation for expansion. It is context-specific and struggles to adapt to different settings or applications.
- Score 2: The situation allows for the inclusion of some additional cultural dimensions without drastically altering the core context. It can accommodate more participants or interactions with some adjustments to existing dynamics. There is some flexibility for adaptation to similar contexts or applications, albeit with moderate effort needed.
- Score 3: The situation is flexible and open, easily incorporating multiple new cultural dimensions or elements. It naturally supports additional participants and interactions without losing coherence. It is broadly applicable and adaptable across varied contexts and applications, maintaining core effectiveness and relevance.
```

Figure 10: Rubrics used for CultureBank Data Filtering Process

```
### LLM ROLE
You are a language model with advanced commonsense reasoning skills, capable of logical and analytical reasoning, heuristic and intuitive thinking, comparative and
hypothetical analysis, and contextual and specialized understanding.

### TASK DESCRIPTION
Given a multi-choice commonsense questions with the correct option, you task is to provide a "COMMONSENSE CONTEXT" to expand on the given question and a detailed
"REASONING PROCESS" that involves multiple "REASONING STEPs" to arrive at the correct answer.
+ A "COMMONSENSE CONTEXT" to the question refers to the background knowledge or additional details that are generally understood without requiring specialized knowledge,
including factors such as time, place, social norms, cultural influences, and other relevant details that shape the understanding of the topic.
+ Each "REASONING STEP" should be an "ATOMIC REASONING STEP" — an Indivisible Unit of reasoning that predominantly utilizes one reasoning skill. It is a single, coherent
thought process that cannot be broken down into smaller steps without losing its meaning. The "REASONING PROCESS" must be as efficent as possible, only using the minimum
number of steps necessary, ensuring that each step is non-redundant and contributes to narrowing down the possible options by eliminating one or more answer choices.

### STEP-BY-STEP INSTRUCTIONS
Following these Step-by-Step Instructions:
1. Question Comprehension: Read the question carefully along with all the provided answer options.
2. Adding The "COMMONSENSE CONTEXT": Expand on the original question by providing an additional "COMMONSENSE CONTEXT". Ensure that the added context is relevant and
enriches the understanding of the question.
3. Describe your Step-by-Step "REASONING PROCESS" to arrive at the correct answer. Each "ATOMIC REASONING STEP" must following this sequence:
    3.1. Choose a REASONING SKILL below to be used by the REASONING STEP:
        <list of the 10 reasoning skill definitions>
    3.2. Apply the choosen "REASONING SKILL": provide a concise explanation of how the chosen "REASONING SKILL" is applied to eliminate certain answer options or
    reinforce the correct answer option. Ensure the reasoning is clear and cannot be further divided into smaller steps.
    3.3. Eliminate Options: List the options eliminated in this step based on your reasoning.
    3.4. Update Possible Options: Provide the list of remaining possible options after this step.
4. Generate your output in the JSON format with the following structure:
    ```json
    {
      "commonsense_context": "context_text",
      "commonsense_question": "question_text",
      "options": {
          "A": "option_answer_text_A",
          ...
      },
      "correct_answer": ["answer_option", "answer_text"],
      "reasoning_process": {
          "reasoning_step_1": {
              "reasoning_skill": "reasoning_skill_name",
              "reasoning": "reasoning_text",
              "eliminated_options": [list_of_eliminated_options],
              "possible_options": [list_of_remaining_options]
          },
          ...
          "reasoning_step_n": {
              "reasoning_skill": "reasoning_skill_name",
              "reasoning": "reasoning_text",
              "eliminated_options": [list_of_eliminated_options],
              "possible_options": [list_of_remaining_options]
          }
      }
    }
    ```

IN-CONTEXT EXAMPLE:
<.....>

OUTPUT REMINDER
Ensure that your output follows the JSON structure as instructed and demonstrated in the in-context example.

INPUT:
    "question": "What is the best way to experience a live performance?",
    "options": {
        "A": "watch play",
        "B": "go to theatre",
        "C": "open eyes",
        "D": "check showtimes",
        "E": "buy tickets"
    },
    "correct_answer": [
        "B",
        "go to theatre"
    ]
```

Figure 11: Prompt for Structured Reasoning Generation step for **mSCoRe-G** (English).

```
### LLM ROLE
You are a language model with advanced commonsense reasoning skills, capable of logical and analytical reasoning, heuristic and intuitive thinking, comparative and
hypothetical analysis, and contextual and specialized understanding.

### TASK DESCRIPTION
Given a multi-choice commonsense question with its options, your task is to modify and expand it to create a more complex question by expanding its context, modifying
the question, adjusting the answer options, and adding an additional REASONING STEP. Your output should include the expanded context, the modified question, revised
answer options, the correct answer, and a detailed "REASONING PROCESS".

### STEP-BY-STEP INSTRUCTIONS
Following these Step-by-Step Instructions:
1. Question Comprehension: Carefully read the given question and the context, and its answer options.
2. Context Expansion: adding additional backgound or situaltional details to the "COMMONSENSE CONTEXT" to add depth and reasoning requirements to the question.
3. Question Modificatioin:  Utilize the "EXPANDED COMMONSENSE CONTEXT" to craft a more complex question while maintaining its core concept and commonsense.
4. Option Adjustments:
    + Adjust the existing answer options to align with the new complex question
    + Ensure the correct answer option remains semantically similar to the original
    + Introduce an additional plausible but incorrect option to increase the complexity of the question
    + Keep all answer options as concise as the originals
5. Reasoning Refinements: Refine the original "REASONING PROCESS" to fit the new context. The additional "ATOMIC REASONING STEP" must use one of the following "REASONING
SKILLs":
    <list of the 10 reasoning skill definitions>
6. Format the Output using JSON format with the following structure:
    ```json
    {
      "commonsense_context": "context_text",
      "commonsense_question": "question_text",
      "options": {
          "A": "option_answer_text_A",
          ...
      },
      "correct_answer": ["answer_option", "answer_text"],
      "reasoning_process": {
          "reasoning_step_1": {
              "reasoning_skill": "reasoning_skill_name",
              "reasoning": "reasoning_text",
              "eliminated_options": [list_of_eliminated_options],
              "possible_options": [list_of_remaining_options]
          },
          ...
          "reasoning_step_n": {
              "reasoning_skill": "reasoning_skill_name",
              "reasoning": "reasoning_text",
              "eliminated_options": [list_of_eliminated_options],
              "possible_options": [list_of_remaining_options]
          }
      }
    }
    ```

IN-CONTEXT EXAMPLE:
<.....>

OUTPUT REMINDER
Ensure that your output follows the JSON structure as instructed and demonstrated in the in-context example.

INPUT:
<.....>
```

Figure 12: Prompt for Complexity Expansion step for **mSCoRe-G** (English).

```
### LLM ROLE
You are a language model with advanced commonsense reasoning skills, capable of logical and analytical reasoning, heuristic and intuitive thinking, comparative and
hypothetical analysis, and contextual and specialized understanding.


### TASK DESCRIPTION
Your task is to perform "Commonsense Implicitation," which involves combining a given "commonsense_context" with a "question" to generate a new, concise commonsense
question that implicitly incorporates the original context. This process aims to evaluate the commonsense reasoning abilities of LLMs by ensuring that the implicit
context preserves the original reasoning process and maintains the correctness of the answer.

### STEP-BY-STEP INSTRUCTIONS
Following these Step-by-Step Instructions:
1. Analyze the provided "commonsense_context" to understand the underlying assumptions and implicit knowledge required for reasoning
2. Examine the "commonsense_question" and its associated "options" to identify key elements essential for answering the question
3. Rewrite the "commonsense_question" by combining the original context and question to create a more new "commonsense_question" with an "IMPLICITLY IMPLIED COMMONSENSE
CONTEXT". Ensure that the new question remains clear and understandable
4. Verify that the "REASONING PROCESS" remains unchanged in the transformed question, and confirm that the correct answer remains the same as in the original
5. Ensure that all answer options are reasonable, relevant, and maintain their original intent in the context of the rewritten question
6. Retain the structure and content of the "reasoning" section to reflect the logical steps supporting the correct answer. The  "ATOMIC REASONING STEP" must use one of
the following "REASONING SKILLs":
    <list of the 10 reasoning skill definitions>

### IN-CONTEXT EXAMPLE:
<.....>

### OUTPUT REMINDER
Ensure that your output follows the JSON structure as instructed and demonstrated in the in-context example.

### INPUT:
<.....>
```

Figure 13: Prompt for Commonsense Implicitation step for **mSCoRe-G** (English).

```
### ORIGINAL QUESTION FROM mCSQA:
    "question": "What is the best way to experience a live performance?",
    "options": {
        "A": "watch play",
        "B": "go to theatre",
        "C": "open eyes",
        "D": "check showtimes",
        "E": "buy tickets"
    },
    "correct_answer": [
        "B",
        "go to theatre"
    ]

### COMPEXITY LEVEL 0:
  "commonsense_context": "Experiencing a live performance, such as a play, concert, or other theatrical event, typically involves attending a venue where the performance
  is being held. The atmosphere, acoustics, and visual elements of a live performance are best appreciated in person, where the audience can fully engage with the
  performers and the production. This is a social and cultural activity that often involves planning, such as purchasing tickets and checking showtimes, but the core
  experience is being physically present at the venue.",
  "commonsense_question": "What is the best way to experience a live performance?",
  "options": {
      "A": "watch play",
      "B": "go to theatre",
      "C": "open eyes",
      "D": "check showtimes",
      "E": "buy tickets"
  },
  "correct_answer": ["B", "go to theatre"],
  "reasoning_process": {
      "reasoning_step_1": {
          "reasoning_skill": "deductive_reasoning",
          "reasoning": "To experience a live performance, one must be physically present where the performance is taking place. This means attending the venue, such as a
          theatre, where the live event is held.",
          "eliminated_options": ["A", "C"],
          "possible_options": ["B", "D", "E"]
      },
      "reasoning_step_2": {
          "reasoning_skill": "temporal_reasoning",
          "reasoning": "Checking showtimes and buying tickets are preparatory actions that occur before the actual experience of the live performance. They are necessary
          steps but do not constitute the experience itself.",
          "eliminated_options": ["D", "E"],
          "possible_options": ["B"]
      }
  }

### COMPLEXITY LEVEL 1
  "commonsense_context": "Experiencing a live performance, such as a play, concert, or other theatrical event, typically involves attending a venue where the performance
  is being held. The atmosphere, acoustics, and visual elements of a live performance are best appreciated in person, where the audience can fully engage with the
  performers and the production. This is a social and cultural activity that often involves planning, such as purchasing tickets and checking showtimes, but the core
  experience is being physically present at the venue. Additionally, live performances often include unique interactions between the audience and performers, such as
  applause, laughter, or even participation, which enhance the overall experience.",
  "commonsense_question": "In what way can you fully immerse yourself in the unique atmosphere and interactions of a live performance?",
  "options": {
      "A": "watch a recording at home",
      "B": "attend the theatre in person",
      "C": "read reviews online",
      "D": "check showtimes regularly",
      "E": "buy tickets in advance",
      "F": "listen to a podcast about the performance"
  },
  "correct_answer": ["B", "attend the theatre in person"],
  "reasoning_process": {
      "reasoning_step_1": {
          "reasoning_skill": "deductive_reasoning",
          "reasoning": "To fully immerse oneself in the unique atmosphere and interactions of a live performance, one must be physically present at the venue. This
          allows for direct engagement with the performance and the performers.",
          "eliminated_options": ["A", "C"],
          "possible_options": ["B", "D", "E","F"]
      },
      "reasoning_step_2": {
          "reasoning_skill": "temporal_reasoning",
          "reasoning": "Checking showtimes and buying tickets are preparatory actions that occur before attending the performance. They are necessary for planning but do
          not provide the immersive experience itself.",
          "eliminated_options": ["D", "E"],
          "possible_options": ["B", "F"]
      },
      "reasoning_step_3": {
          "reasoning_skill": "social_reasoning",
          "reasoning": "Being present at the theatre allows for social interactions and shared experiences with other audience members and the performers, which are
          integral to the live performance experience.",
          "eliminated_options": ["F"],
          "possible_options": ["B"]
      }
  }

### COMPLEXITY LEVEL 2:
  "commonsense_context": "Experiencing a live performance, such as a play, concert, or other theatrical event, typically involves attending a venue where the performance
  is being held. The atmosphere, acoustics, and visual elements of a live performance are best appreciated in person, where the audience can fully engage with the
  performers and the production. This is a social and cultural activity that often involves planning, such as purchasing tickets and checking showtimes, but the core
  experience is being physically present at the venue. Additionally, live performances often include unique interactions between the audience and performers, such as
  applause, laughter, or even participation, which enhance the overall experience. Furthermore, attending a live performance can create lasting memories and a sense of
  community among attendees, as they share the emotional highs and lows of the performance together.",
  "commonsense_question": "How can you create lasting memories and fully immerse yourself in the unique atmosphere and interactions of a live performance, while also
  contributing to the communal experience?",
  "options": {
      "A": "watch a recording at home",
      "B": "attend the theatre in person",
      "C": "read reviews online",
      "D": "check showtimes regularly",
      "E": "buy tickets in advance",
      "F": "listen to a podcast about the performance",
      "G": "participate in a post-show discussion"
  },
```

```
  "correct_answer": ["B", "attend the theatre in person"],
  "reasoning_process": {
      "reasoning_step_1": {
          "reasoning_skill": "deductive_reasoning",
          "reasoning": "To create lasting memories and fully immerse oneself in the unique atmosphere and interactions of a live performance, one must be physically
          present at the venue. This allows for direct engagement with the performance and the performers.",
          "eliminated_options": ["A", "C", "F"],
          "possible_options": ["B", "D", "E", "G"]
      },
      "reasoning_step_2": {
          "reasoning_skill": "temporal_reasoning",
          "reasoning": "Checking showtimes and buying tickets are preparatory actions that occur before attending the performance. They are necessary for planning but do
          not provide the immersive experience itself.",
          "eliminated_options": ["D", "E"],
          "possible_options": ["B", "G"]
      },
      "reasoning_step_3": {
          "reasoning_skill": "social_reasoning",
          "reasoning": "Being present at the theatre allows for social interactions and shared experiences with other audience members and the performers, which are
          integral to the live performance experience. While participating in a post-show discussion can enhance the communal experience, it does not replace the
          immersive experience of attending the performance itself.",
          "eliminated_options": ["G"],
          "possible_options": ["B"]
      }
  }

### COMPEXITY LEVEL 3:
  "commonsense_context": "Experiencing a live performance, such as a play, concert, or other theatrical event, typically involves attending a venue where the performance
  is being held. The atmosphere, acoustics, and visual elements of a live performance are best appreciated in person, where the audience can fully engage with the
  performers and the production. This is a social and cultural activity that often involves planning, such as purchasing tickets and checking showtimes, but the core
  experience is being physically present at the venue. Additionally, live performances often include unique interactions between the audience and performers, such as
  applause, laughter, or even participation, which enhance the overall experience. Furthermore, attending a live performance can create lasting memories and a sense of
  community among attendees, as they share the emotional highs and lows of the performance together. In recent times, some performances have also incorporated digital
  elements, allowing for a hybrid experience where audiences can engage both in-person and online, adding a new dimension to the traditional live performance.",
  "commonsense_question": "In the context of a modern live performance that incorporates both in-person and digital elements, how can you create lasting memories and
  fully immerse yourself in the unique atmosphere and interactions, while also contributing to the communal experience?",
  "options": {
      "A": "watch a recording at home",
      "B": "attend the theatre in person",
      "C": "read reviews online",
      "D": "check showtimes regularly",
      "E": "buy tickets in advance",
      "F": "listen to a podcast about the performance",
      "G": "participate in a post-show discussion",
      "H": "engage with digital elements during the performance"
  },
  "correct_answer": ["B", "attend the theatre in person"],
  "reasoning_process": {
      "reasoning_step_1": {
          "reasoning_skill": "deductive_reasoning",
          "reasoning": "To create lasting memories and fully immerse oneself in the unique atmosphere and interactions of a live performance, one must be physically
          present at the venue. This allows for direct engagement with the performance and the performers.",
          "eliminated_options": ["A", "C", "F"],
          "possible_options": ["B", "D", "E", "G", "H"]
      },
      "reasoning_step_2": {
          "reasoning_skill": "temporal_reasoning",
          "reasoning": "Checking showtimes and buying tickets are preparatory actions that occur before attending the performance. They are necessary for planning but do
          not provide the immersive experience itself.",
          "eliminated_options": ["D", "E"],
          "possible_options": ["B", "G", "H"]
      },
      "reasoning_step_3": {
          "reasoning_skill": "social_reasoning",
          "reasoning": "Being present at the theatre allows for social interactions and shared experiences with other audience members and the performers, which are
          integral to the live performance experience. While participating in a post-show discussion can enhance the communal experience, it does not replace the
          immersive experience of attending the performance itself.",
          "eliminated_options": ["G"],
          "possible_options": ["B", "H"]
      },
      "reasoning_step_4": {
          "reasoning_skill": "analogical_reasoning",
          "reasoning": "Engaging with digital elements during the performance can enhance the experience but is analogous to supplementary activities. The core immersive
          experience is still best achieved by being physically present.",
          "eliminated_options": ["H"],
          "possible_options": ["B"]
      }
  }
```


Figure 14: An example from **mSCoRe-G** for complexity level 0 to 3 (English).

```
### LLM ROLE
You are a language model with advanced commonsense reasoning skills, capable of logical and analytical reasoning, heuristic and intuitive thinking, comparative and
hypothetical analysis, and contextual and specialized understanding.


### TASK DESCRIPTION
Your task is to create a multiple-choice commonsense question based on a given cultural situation in the following format:
{
  "cultural_topic": "culture group - topic - scenario",
  "social_context": "settings the behavior takes place",
  "actor": "who exhibit the behavior",
  "question": "the commonsense question regarding the actor's behavior",
  "actor_behavior": "behavior of the actor - which are highly agreed upon (the correct answer option)",
  "recipient": "recipient of the action",
  "relation": "relation between the actor and the recipient",
  "recipient_behavior": "behavior of the recipient",
}
The question should implicitly incorporate the cultural context, challenging the AI's ability to utilize commonsense reasoning to arrive at the correct answer. The goal
is to test and enhance the AI's understanding of cultural norms and behaviors in a specific setting.
Provide the detailed "REASONING PROCESS" the arrive at the correct anwser option that involves multiple "REASONING STEPs" to arrive at the correct answer. Each
"REASONING STEP" should be an "ATOMIC REASONING STEP" - an Indivisible Unit of reasoning that predominantly utilizes one reasoning skill. It is a single, coherent
thought process that cannot be broken down into smaller steps without losing its meaning. The "REASONING PROCESS" must be as efficent as possible, only using the minimum
number of steps necessary, ensuring that each step is non-redundant and contributes to narrowing down the possible options by eliminating one or more answer choices.


### STEP-BY-STEP INSTRUCTIONS
Following these Step-by-Step Instructions:
1. Analyze the Provided Cultural Situation: Review the details of the cultural group, context, actor behaviors, and other descriptions to understand the key elements of
the situation.
2. Adding The "COMMONSENSE CONTEXT": Based on the context given in the input, A "COMMONSENSE CONTEXT" to the question refers to the background knowledge or additional
details that are generally understood without requiring specialized knowledge, including factors such as time, place, social norms, cultural influences, and other
relevant details that shape the understanding of the topic.
3. Create the "Commonsense Question": Combine the cultural context and the persona's inquiry to formulate a concise question. Ensure the question IMPLICITLY incorporates
the original context without explicitly stating it. Create the correct answer option based on the "actor_behavior"
4. Provide Other Answer Options: Create 5 multiple-choice options (including the correct answer from the previous step). Two of which should be plausible options. The
other two should be distractors that are relevant and reasonable but incorrect based on the cultural context.
5. Describe your Step-by-Step "REASONING PROCESS" to arrive at the correct answer. Each "ATOMIC REASONING STEP" must following this sequence:
    5.1. Choose a "REASONING SKILL" below to be used by the "REASONING STEP":
        <list of the 10 reasoning skill definitions>
    5.2. Apply the choosen "REASONING SKILL": provide a concise explanation of how the chosen "REASONING SKILL" is applied to eliminate certain answer options or
    reinforce the correct answer option. Ensure the reasoning is clear and cannot be further divided into smaller steps.
    5.3. Eliminate Options: List the options eliminated in this step based on your reasoning.
    5.4. Update Possible Options: Provide the list of remaining possible options after this step.
6. Generate your output in the JSON format with the following structure:

    ```json
    {
      "commonsense_context": "context_text",
      "commonsense_question": "question_text",
      "options": {
          "A": "option_answer_text_A",
          ...
      },
      "correct_answer": ["answer_option", "answer_text"],
      "reasoning_process": {
          "reasoning_step_1": {
              "reasoning_skill": "reasoning_skill_name",
              "reasoning": "reasoning_text",
              "eliminated_options": [list_of_eliminated_options],
              "possible_options": [list_of_remaining_options]
          },
          ...
          "reasoning_step_n": {
              "reasoning_skill": "reasoning_skill_name",
              "reasoning": "reasoning_text",
              "eliminated_options": [list_of_eliminated_options],
              "possible_options": [list_of_remaining_options]
          }
      }
    }
    ```

IN-CONTEXT EXAMPLE:
<.....>

OUTPUT REMINDER
Ensure that your output follows the JSON structure as instructed and demonstrated in the in-context example.

INPUT:
    "cultural_topic": "American culture - Dress Codes - Travel Advising",
    "social_context": "In public settings within American culture, it is common for people to dress casually, often opting for comfortable clothing such as sweatpants
    while still adhering to dress codes. This relaxed approach to attire is widely regarded as the norm by a significant portion of the sampled population. It reflects a
    preference for comfort and practicality in daily dress, showcasing a relaxed and informal attitude towards clothing choices in various public settings.",
    "actor": "people - A business professional from a formal corporate background, planning a first-time trip to the United States for a business conference, eager to
    blend in and avoid any potential faux pas",
    "question": "I'm gearing up for a big conference in the States and I'm a bit worried about what to wear. I come from a formal work environment and I don't want to
    stand out in a negative way. Can you give me some tips on what kind of attire would be appropriate for a business setting over there? Should I be concerned about
    anything specific?",
    "actor_behavior": "dress casually, often in comfortable clothing, with a preference for sweatpants and following dress codes",
    "recipient": "None",
    "relation": "None",
    "recipient_behavior": "None"
```

Figure 15: Structured Reasoning Generation Prompt for **mSCoRe-S**.

ORIGINAL INSTANCE FROM CULTUREBANK:
"cultural_topic": "Germans culture - Education and Technology - Travel Advising",
"social_context": "In German schools, both the educational institutions and students actively participate in compulsory swimming education, which includes separate classes for students with limited swimming skills. The goal of this initiative is to teach swimming skills and promote integration, with students, including those with immigrant parents, participating in swimming lessons. It is noteworthy that Muslim students are accommodated by wearing burqinis during these swimming classes, showcasing inclusivity and respect for diverse cultural practices. This swimming education typically takes place in 5th or 6th grade and lasts for 2 years, resulting in students acquiring advanced swimming skills. While not universally embraced, a significant portion of the sampled population considers this practice as a standard part of the educational experience in German schools.",
"actor": "German schools and students - A concerned parent planning a family trip to Germany, looking for a comprehensive and inclusive educational experience for their children",
"question": "I'm planning a family trip to Germany and I want my kids to make the most of their time there. I've heard that schools there offer some unique educational experiences. I'm particularly interested in finding activities that are both fun and educational for my kids, especially ones that can help them learn new skills. Could you recommend some family-friendly programs that might be a good fit for us, keeping in mind that we have a diverse family background?",
"actor_behavior": "provide and attend compulsory swimming education, including separate classes for those with limited swimming skills",
"recipient": "German students, including those with immigrant parents",
"relation": "educational institution and attendees",
"recipient_behavior": "participate in swimming lessons, including wearing burqinis for Muslim students"

COMPEXITY LEVEL 0:
"commonsense_context": "In Germany, swimming education is an integral part of the school curriculum, aimed at teaching essential swimming skills and promoting inclusivity. This program is designed to accommodate students from diverse backgrounds, including those with immigrant parents and Muslim students, who are allowed to wear burqinis. The initiative is generally well-received and considered a standard educational practice, providing students with valuable life skills and fostering integration.",
"commonsense_question": "What unique educational experience in German schools could be particularly beneficial for a diverse family visiting Germany, looking for fun and educational activities for their children?",
"options": {
"A": "Participate in compulsory swimming education that includes accommodations for diverse cultural practices.",
"B": "Enroll in advanced mathematics classes to enhance analytical skills.",
"C": "Join a German language immersion program to improve language proficiency.",
"D": "Attend a local history tour to learn about German culture and heritage.",
"E": "Take part in a cooking class to explore traditional German cuisine."
},
"correct_answer": ["A", "Participate in compulsory swimming education that includes accommodations for diverse cultural practices."],
"reasoning_process": {
"reasoning_step_1": {
"reasoning_skill": "social_reasoning",
"reasoning": "Understanding that the swimming education program in German schools is designed to be inclusive and accommodate diverse cultural backgrounds, making it a unique and suitable experience for a diverse family.",
"eliminated_options": ["B", "C"],
"possible_options": ["A", "D", "E"]
},
"reasoning_step_2": {
"reasoning_skill": "deductive_reasoning",
"reasoning": "Considering the context of the question, which emphasizes fun and educational activities, swimming education stands out as it combines physical activity with skill acquisition, unlike a history tour which is more passive.",
"eliminated_options": ["D", "E"],
"possible_options": ["A"]
}
}

COMPLEXITY LEVEL 1:
"commonsense_context": "In Germany, swimming education is an integral part of the school curriculum, aimed at teaching essential swimming skills and promoting inclusivity. This program is designed to accommodate students from diverse backgrounds, including those with immigrant parents and Muslim students, who are allowed to wear burqinis. The initiative is generally well-received and considered a standard educational practice, providing students with valuable life skills and fostering integration. Additionally, German schools often collaborate with local community centers to offer these swimming lessons, ensuring that students have access to safe and well-maintained facilities. This collaboration also allows for the inclusion of parents in some sessions, promoting family involvement in the educational process.",
"commonsense_question": "For a diverse family visiting Germany, interested in engaging in both educational and community activities, what unique experience offered by German schools could be particularly beneficial?",
"options": {
"A": "Participate in compulsory swimming education that includes accommodations for diverse cultural practices and involves community engagement.",
"B": "Enroll in advanced mathematics classes to enhance analytical skills.",
"C": "Join a German language immersion program to improve language proficiency.",
"D": "Attend a local history tour to learn about German culture and heritage.",
"E": "Take part in a cooking class to explore traditional German cuisine.",
"F": "Engage in a community art project to express cultural diversity."
},
"correct_answer": ["A", "Participate in compulsory swimming education that includes accommodations for diverse cultural practices and involves community engagement."],
"reasoning_process": {
"reasoning_step_1": {
"reasoning_skill": "social_reasoning",
"reasoning": "Understanding that the swimming education program in German schools is designed to be inclusive and accommodate diverse cultural backgrounds, making it a unique and suitable experience for a diverse family. The program's collaboration with community centers further enhances its appeal by involving the family in the local community.",
"eliminated_options": ["B", "C"],
"possible_options": ["A", "D", "E", "F"]
},
"reasoning_step_2": {
"reasoning_skill": "deductive_reasoning",
"reasoning": "Considering the context of the question, which emphasizes educational and community activities, swimming education stands out as it combines physical activity, skill acquisition, and community involvement, unlike a history tour which is more passive.",
"eliminated_options": ["D"],
"possible_options": ["A", "E", "F"]
},
"reasoning_step_3": {
"reasoning_skill": "abductive_reasoning",
"reasoning": "While a cooking class and a community art project can be educational and fun, they do not offer the same level of inclusivity, skill development, and structured community engagement as the swimming program, which is a part of the school curriculum.",
"eliminated_options": ["E", "F"],
"possible_options": ["A"]
}
}

COMPLEXITY LEVEL 2:
"commonsense_context": "In Germany, swimming education is an integral part of the school curriculum, aimed at teaching essential swimming skills and promoting inclusivity. This program is designed to accommodate students from diverse backgrounds, including those with immigrant parents and Muslim students, who are allowed to wear burqinis. The initiative is generally well-received and considered a standard educational practice, providing students with valuable life skills and fostering integration. Additionally, German schools often collaborate with local community centers to offer these swimming lessons, ensuring that students have access to safe and well-maintained facilities. This collaboration also allows for the inclusion of parents in some sessions, promoting family involvement in the educational process. Furthermore, these swimming programs often include cultural exchange activities, where students and their families can share and learn about each other's traditions, enhancing mutual understanding and respect.",

```
"commonsense_question": "For a diverse family visiting Germany, interested in engaging in both educational and community activities, what unique experience offered by
German schools could be particularly beneficial, especially in terms of cultural exchange and inclusivity?",
"options": {
  "A": "Participate in compulsory swimming education that includes accommodations for diverse cultural practices, involves community engagement, and offers cultural
  exchange opportunities.",
  "B": "Enroll in advanced mathematics classes to enhance analytical skills.",
  "C": "Join a German language immersion program to improve language proficiency.",
  "D": "Attend a local history tour to learn about German culture and heritage.",
  "E": "Take part in a cooking class to explore traditional German cuisine.",
  "F": "Engage in a community art project to express cultural diversity.",
  "G": "Participate in a multicultural festival organized by the school."
},
"correct_answer": ["A", "Participate in compulsory swimming education that includes accommodations for diverse cultural practices, involves community engagement, and
offers cultural exchange opportunities."],
"reasoning_process": {
  "reasoning_step_1": {
    "reasoning_skill": "social_reasoning",
    "reasoning": "Understanding that the swimming education program in German schools is designed to be inclusive and accommodate diverse cultural backgrounds, making
    it a unique and suitable experience for a diverse family. The program's collaboration with community centers further enhances its appeal by involving the family in
    the local community.",
    "eliminated_options": ["B", "C"],
    "possible_options": ["A", "D", "E", "F", "G"]
  },
  "reasoning_step_2": {
    "reasoning_skill": "deductive_reasoning",
    "reasoning": "Considering the context of the question, which emphasizes educational and community activities, swimming education stands out as it combines physical
    activity, skill acquisition, and community involvement, unlike a history tour which is more passive.",
    "eliminated_options": ["D"],
    "possible_options": ["A", "E", "F", "G"]
  },
  "reasoning_step_3": {
    "reasoning_skill": "abductive_reasoning",
    "reasoning": "While a cooking class and a community art project can be educational and fun, they do not offer the same level of inclusivity, skill development, and
    structured community engagement as the swimming program, which is a part of the school curriculum.",
    "eliminated_options": ["E", "F"],
    "possible_options": ["A", "G"]
  },
  "reasoning_step_4": {
    "reasoning_skill": "analogical_reasoning",
    "reasoning": "Comparing the swimming program with the multicultural festival, the swimming program offers a more structured and ongoing opportunity for cultural
    exchange and skill development, whereas the festival is a one-time event.",
    "eliminated_options": ["G"],
    "possible_options": ["A"]
  }
}

### COMPLEXITY LEVEL 3:
"commonsense_context": "In Germany, swimming education is an integral part of the school curriculum, aimed at teaching essential swimming skills and promoting
inclusivity. This program is designed to accommodate students from diverse backgrounds, including those with immigrant parents and Muslim students, who are allowed to
wear burqinis. The initiative is generally well-received and considered a standard educational practice, providing students with valuable life skills and fostering
integration. Additionally, German schools often collaborate with local community centers to offer these swimming lessons, ensuring that students have access to safe
and well-maintained facilities. This collaboration also allows for the inclusion of parents in some sessions, promoting family involvement in the educational process.
Furthermore, these swimming programs often include cultural exchange activities, where students and their families can share and learn about each other's traditions,
enhancing mutual understanding and respect. The program also emphasizes water safety, which is a crucial skill for everyone, and includes sessions on the importance of
respecting different cultural practices in shared spaces.",
"commonsense_question": "For a diverse family visiting Germany, interested in engaging in both educational and community activities, what unique experience offered by
German schools could be particularly beneficial, especially in terms of cultural exchange, inclusivity, and learning essential life skills like water safety?",
"options": {
  "A": "Participate in compulsory swimming education that includes accommodations for diverse cultural practices, involves community engagement, and offers cultural
  exchange opportunities.",
  "B": "Enroll in advanced mathematics classes to enhance analytical skills.",
  "C": "Join a German language immersion program to improve language proficiency.",
  "D": "Attend a local history tour to learn about German culture and heritage.",
  "E": "Take part in a cooking class to explore traditional German cuisine.",
  "F": "Engage in a community art project to express cultural diversity.",
  "G": "Participate in a multicultural festival organized by the school.",
  "H": "Join a water safety workshop that includes cultural sensitivity training."
},
"correct_answer": ["A", "Participate in compulsory swimming education that includes accommodations for diverse cultural practices, involves community engagement, and
offers cultural exchange opportunities."],
"reasoning_process": {
  "reasoning_step_1": {
    "reasoning_skill": "social_reasoning",
    "reasoning": "Understanding that the swimming education program in German schools is designed to be inclusive and accommodate diverse cultural backgrounds, making
    it a unique and suitable experience for a diverse family. The program's collaboration with community centers further enhances its appeal by involving the family in
    the local community.",
    "eliminated_options": ["B", "C"],
    "possible_options": ["A", "D", "E", "F", "G", "H"]
  },
  "reasoning_step_2": {
    "reasoning_skill": "deductive_reasoning",
    "reasoning": "Considering the context of the question, which emphasizes educational and community activities, swimming education stands out as it combines physical
    activity, skill acquisition, and community involvement, unlike a history tour which is more passive.",
    "eliminated_options": ["D"],
    "possible_options": ["A", "E", "F", "G", "H"]
  },
  "reasoning_step_3": {
    "reasoning_skill": "abductive_reasoning",
    "reasoning": "While a cooking class and a community art project can be educational and fun, they do not offer the same level of inclusivity, skill development, and
    structured community engagement as the swimming program, which is a part of the school curriculum.",
    "eliminated_options": ["E", "F"],
    "possible_options": ["A", "G", "H"]
  },
  "reasoning_step_4": {
    "reasoning_skill": "analogical_reasoning",
    "reasoning": "Comparing the swimming program with the multicultural festival, the swimming program offers a more structured and ongoing opportunity for cultural
    exchange and skill development, whereas the festival is a one-time event.",
    "eliminated_options": ["G"],
    "possible_options": ["A", "H"]
  },
  "reasoning_step_5": {
    "reasoning_skill": "probabilistic_reasoning",
    "reasoning": "While a water safety workshop with cultural sensitivity training is beneficial, the swimming program is more comprehensive, offering ongoing lessons
    that include water safety as part of a broader curriculum.",
    "eliminated_options": ["H"],
    "possible_options": ["A"]
  }
}
```

Figure 16: An example from **mSCoRe-S** for complexity level 0 to 3 (English).