常识推理使人能够在日常情境中导航、进行逻辑推断，并理解我们环境中的隐含信息。虽然这种能力对于人类来说是自然而然的，但事实证明，在当前的语言模型中复制这种能力是最具挑战性的能力之一。最近，在大型推理模型（LRMs）方面的进展，比如 OpenAI 的 o1 系列，以及 DeepSeek R1 等开源模型，在数学、编码和逻辑推理等各种复杂推理任务中显示出有希望的结果。然而，关于系统分析和理解这些模型的常识推理能力的研究相对较少，特别是在涉及多种语言和文化背景的常识的多语言环境中。

已有多个基准被提出以评估语言模型的常识推理能力。CommonsenseQA（CSQA）(Talmor et al., 2019) 通过源自 ConceptNet 的多项选择题来评估一般常识知识。COPA (Roemmele et al., 2011) 专注于日常事件之间的因果关系，而 SocialIQA (Sap et al., 2019) 评估社交常识理解。最近，像 MMLU (Hendrycks et al., 2021) 和 Big-Bench Hard (Suzgun et al., 2023) 这样的综合基准旨在评估模型在各种常识任务中的泛化能力。然而，这些基准在三个关键领域存在显著局限性。首先，它们往往集中于单一资源丰富的语言，如英语 (Talmor et al., 2019) 或中文 (Sun et al., 2024) 。多语言扩展如 X-COPA (Ponti et al., 2020) 和 X-CSQA (Lin et al., 2021) 主要依赖于现有数据集的翻译，从而限制了它们捕捉文化特定细微差别的能力。其次，尽管最近的努力如 mCSQA (Sakai et al., 2024) 使用生成式多语言语言模型进行更加全面和稳健的数据集创建过程，它们仍然缺乏系统的方法来扩展任务难度，这对于评估 LLMs 快速发展的能力至关重要。最后，当前的基准无法提供对 LLMs 所用推理步骤的细粒度分析和分类，而这将提供对其运作的更深入洞察。

为了解决这些限制，我们引入了多语言和可扩展的基准测试，用于技能型常识推理 (mSCoRe)，这是一项全新的基准测试，专门用于全面评估大型语言模型在多种语言和文化背景下的常识推理能力。具体来说，我们的基准测试提供了三个明显的优势：

1. 全面覆盖：mSCoRe 涵盖了包括英语、德语、法语、中文和日语在内的多种语言的 `general` 常识知识，以及多样的文化 `social` 常识知识。
2. 基于技能的分析：mSCoRe 引入了一种新颖的推理分析方法，通过对每个原子推理步骤进行分类，从而可以更精确地分析模型的推理过程。
3. 可扩展性：mSCoRe 采用了如上下文扩展、选项调整和常识隐含等技术，在逐步增加问题复杂度的同时保持了常识答案语义，从而有效地提高任务难度。

我们的贡献可以总结如下：

- 我们介绍了 mSCoRe，这是一种用于评估多语言通用和文化常识推理的新型可扩展基准，具有细粒度的基于技能的分析。
- 使用 mSCoRe，我们广泛评估了八种最先进的大型语言模型，包括商业和开源模型，涵盖不同的推理条件。
- 我们的分析提供了关于模型规模、训练技术以及推理技能类型如何影响性能的洞见，并建议未来提升 LLMs 常识推理能力的方向。

1 相关工作

大规模推理模型：最近在大型语言模型（LLMs）方面的进展展示了其在各种复杂问题解决任务中的卓越能力。像 OpenAI o1 (Jaech et al., 2024) ，Macro-o1 (Zhao et al., 2024) 和 DeepSeek-R1 (DeepSeek-AI et al., 2025) 等推理增强模型在数学和编码方面表现出色，有效模拟了类似人类的分析思维并增强了多步骤推理 (Glazer et al., 2024; Guo et al., 2024) 。这些模型采用多种方法来增强推理能力。尤其是，链式思维提示 (Wu et al., 2023) 已成为一种强大的技术，鼓励逐步推理，显著提高了在复杂任务上的性能。在此基础上，各种链式-X 方法被提出以进一步增强模型的推理能力 (Yao et al., 2023; Lightman et al., 2024; Besta et al., 2024; Chen et al., 2024) 。最近的技术如测试时缩放和强化学习也为改善 LLMs 的推理能力做出了贡献 (Snell et al., 2024; Kumar et al., 2025; Hou et al., 2025) 。虽然这些方法加强了推理路径的总体结构，但它们一般很少关注各个推理步骤的分类。mSCoRe 提出了一种更细化的方法，其中每一步都是原子的并根据推理技能进行标记，从而促进对模型推理过程的更深入和系统的评估。

常识推理基准：尽管在评估大型语言模型（LLMs）数学和科学推理能力方面取得了显著进展 (Cobbe et al., 2021; Glazer et al., 2024; He et al., 2024; Chow et al., 2025) ，常识推理基准相对较少受到近期关注。早期的数据集如 CommonsenseQA (CSQA) (Talmor et al., 2019) 、COPA (Roemmele et al., 2011) 和 SocialIQA (Sap et al., 2019) 主要针对英语常识知识，分别关注于一般事实知识、因果关系和社会互动。最新的综合基准如 MMLU (Hendrycks et al., 2021) 和 Big-Bench Hard (Suzgun et al., 2023) 评估 LLMs 在不同常识推理任务中的泛化能力。多语言扩展如 X-CSQA (Lin et al., 2021) 和 X-COPA (Ponti et al., 2020) 将评估从英语

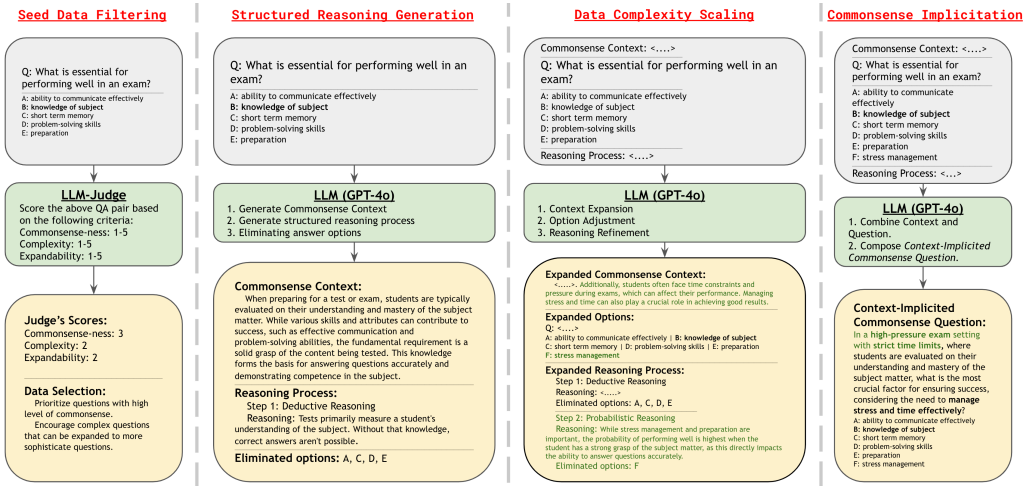

Figure 1: 数据生成过程。mSCoRe 的四步数据创建流程。每一步都建立在前一步的基础上，逐步创建更加具有挑战性的推理任务，同时保持对核心推理能力的评估。

扩展到通过将现有数据集翻译成多种语言。更新的方法如 mCSQA (Sakai et al., 2024) 利用 LLM 更紧密地辅助数据合成过程。然而，这些基准在涉及不同文化的日常互动的文化社会常识方面仍存在局限性。针对文化知识库的发展，以创建文化意识的 LLMs 的努力正在增加。特别是，CulturePark (Li et al., 2024) 引入了一种由 LLMs 驱动的新型多代理沟通框架，以模拟跨文化的人类互动，而 CultureBank (Shi et al., 2024) 则汇聚了来自 TikTok 和 Reddit 等平台的现实社会互动，并围绕文化话题构建注释。mSCoRe 在 mCSQA 和 CultureBank 的基础上，全面覆盖多语言和多文化常识推理的一般和社会方面。

2 基准创建

2.1 常识推理

常识推理涉及使用隐含的世界知识对场景的未明示方面进行推断——这种能力植根于人类行为中，但对当前的大型语言模型来说仍具挑战性。与诸如数学或逻辑等形式推理领域不同，那些领域的规则被明确定义，结论遵循确定的路径，而常识推理需要访问大量隐含知识的储备，并能够灵活地在不同情境中应用这些知识。此外，尤其是对于常识性问题，可能会有多条推理路径可以导致正确答案。然而，以前的评估仍然主要集中在答案的准确性 (Suzgun et al., 2023; Sakai et al., 2024)，因此对于大型语言模型如何构建其推理路径提供的见解有限。

> 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: 原子推理步骤定义。

为了解决这个限制，我们通过引入“原子推理步骤"（图 2）这一概念，深入研究模型的推理过程，作为分析的基础单位。我们的框架旨在分析 LLM 所采用的最优路径，该路径定义为在保持逻辑一致性的前提下所需原子推理步骤最少的路径。这一方法不仅能系统地评估具体推理技能，还能为分析模型如何构建复杂推理链提供一个清晰的框架。它允许跨不同模型和语言对推理过程进行有意义的比较。最后，通过增加原子步骤的要求，它有助于问题复杂性的扩展。

我们开发了一种结构化分类法，用于对每个推理步骤进行分类，从而系统地评估大型语言模型在常识任务中如何运用人类推理技能。虽然对人类推理技能的全面分类没有明确共识，但现有的分类通常用于特定目的。例如，Bloom 分类法提供了一个将教育目标分为三个领域的层次化框架：认知（基于知识）、情感（基于情感）和心理运动（基于行动）。类似地，Fleishman 分类法确定了跨认知、感知、心理运动和生理领域的 52 种不同人类能力，主要用于促进工作设计、培训和评估开发。基于 Do et al. (2024) 识别的常识知识的基本特征和 Wikipedia contributors (2025) 的已建立推理技能分类，我们提出了一种由三个主要类别组成的分类法：

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: 三类中十种推理技能的简短定义和示例，应用于“社交媒体对社会有益吗?”这一问题。详细描述和更多示例见附录 3.2 。

- 逻辑推理包括使用结构化的过程来从给定的信息中得出结论的推理形式。该类别包括诸如演绎推理、归纳推理和溯因推理等方法，这些方法在科学和分析学科中是基础，以确保结论在逻辑上是合理的。
- 情境推理包括用于理解元素之间关系、情境和动态的技能。该类别涵盖了用于评估情景、预测结果和解决不同情境中问题的各种推理类型，如类比推理、反事实推理、概率推理、时间推理和空间推理。
- 社会和伦理推理涉及专注于理解社会互动和评估伦理原则的技能。该类别包括社会和道德推理，这对于解释行为、在复杂社会环境中游走以及基于伦理考虑进行决策是必不可少的。

表格 1 中提供了每种推理技能的详细描述和示例。虽然人类还采用了超出本文中所提出的额外推理技能，我们的目标是建立一个简洁但全面的推理分类，以最大限度地涵盖人类在常识应用中的推理能力，同时将类别之间的重叠降到最低。这个推理分类将在整个数据生成和评估过程中，在我们的 LLM 提示中实施，以确保关注我们考虑的技能。每个基本的推理步骤将根据我们的分类归入单一技能，从而能够精确比较不同的推理过程。

2.2 mSCoRe

为了保持强大的标签准确性，我们不是从头开始使用 LLMs 生成合成数据集，而是利用人工注释的种子数据集并扩展其复杂性来创建 mSCoRe。具体而言，我们的基准测试由多项选择的常识问题组成，根据不同的种子数据集分为两个子集：(1) mSCoRe-G 侧重于一般常识推理，基于来自 mCSQA (Sakai et al., 2024) 的多语言常识问题作为种子数据集。这个组成部分评估跨多种语言的物理因果关系、时间关系和基本世界动态的理解。(2) mSCoRe-S 针对来自 CultureBank (Shi et al., 2024) 的多样化文化情境的社会常识推理。这个组成部分特别测试在不同文化背景下对社会互动、文化规范和行为期望的理解。

整体数据生成过程在图 1 中可视化，其中种子数据集中的每个实例都经过图 3 所示的四步过程。通过这种系统化的创建过程，mSCoRe 提供了一个综合框架来评估和分析 LLM 的常识推理能力。

2.2.1 mSCoRe-G : 通用常识

mCSQA（多语言 CommonsenseQA）将 CommonsenseQA 数据集 (Talmor et al., 2019) 扩展到八种语言，以评估语言模型的跨语言常识推理能力。基于 ConceptNet，mCSQA 中的每个多选问题-答案（QA）对主要围绕一般常识知识（在图 1 中的第一步提供了一个示例）。为了创建 mSCoRe-G，我们进一步通过以下 4 个步骤处理每个 QA 对：

1. 种子数据过滤：一般的 LLM 裁判根据图 4 中描述的三个标准来评估每个候选：(1) 常识性，(2) 复杂性，和 (3) 可扩展性。目标是优先考虑具有高水平常识和复杂性的问题，同时保持向更复杂问题扩展的灵活性（裁判模型和评分标准的完整细节在附录 4.1 中提供）。

2. 结构化推理生成：对于选定的问题-答案对，我们使用 LLM 生成相关的常识背景，以帮助识别正确答案。从元组（背景，问题，选项-答案）中，我们然后生成一个结构化的推理过程。过程中每一步的推理由三个属性构成：（1）推理技能 - 在此步骤中主要使用的来自我们推理本体的特定技能，（2）推理文本 - 基于已识别技能的模型推理，（3）排除选项 - 在此步骤中基于推理被排除的选项列表。

3. 数据复杂性扩展：基于（上下文、问题、答案、推理过程），我们实施一个程序，以系统地提高每个问题的难度水平。我们的目标是在每个层级引入一个额外的合理选项，不仅增加问题的复杂性，还要求额外的推理步骤来排除。这个目标通过 3 个子步骤实现，如图 5 所示。

4. 常识隐含化：常识知识是关于世界的隐性知识，通常是未言明但被假设的，这一步通过将上下文和问题结合成隐含上下文的常识性问题，减少暴露给 LLMs 的上下文。要回答修改

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: 来自 CultureBank 的一个社会常识问题的例子。

后的问题，模型将必须借助其内部的常识来确定正确答案，特别是在主题需要的不仅仅是逻辑推理时。

整个过程（图 1 ）被重复以创建复杂性逐渐增加的问题。这种方法有助于缓解数据泄漏 (Deng et al., 2024) 和捷径推理 (Haraguchi et al., 2023) 问题，正如我们实验结果中观察到的那样，性能在每个层级显著下降。此外，扩展的复杂性迫使 LLMs 更广泛地利用其推理能力，使得能够更深入地研究它们的推理过程。附录 5 中提供了完整的提示和详细示例。

当前的常识基准在社会常识知识和文化规范方面仍存在差距 (Davis, 2024) 。为了全面评估 LLMs 的常识推理能力，我们提出了一个名为 mSCoRe-S 的附加基准，围绕跨多元文化背景的社会情境。特别是，我们利用 CultureBank (Shi et al., 2024) 作为我们的种子数据集，它是一个包含来自 TikTok 和 Reddit 帖子的真实社会问题的知识库。CultureBank 中的每个实例都提供了包含有关文化群体、情境、行为细节的各种描述符，以及指示该行为在社区内被广泛接受程度的同意级别（图 2 ）。

每个种子实例遵循前面章节中描述的相同的四步过程来生成最终的上下文隐含常识问题。然而，为了适应 CultureBank 数据，引入了一些微小的差异，包括：

种子数据筛选：除了用于 mSCoRe-G 的 3 个标准外，我们还引入了一个额外的标准——多文化性——用于筛选社交情境（详细描述见附录 4.1 ）。其目的是选择那些包含最具文化特色元素的情境，允许我们评估模型对多样文化背景及相关常识知识的理解。

结构化推理生成：在生成上下文和推理过程之前，LLM 需要首先生成 QA 对。这作为上一节中 mCSQA 的种子 QA 对，并经历相同的过程。

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: 数据生成过程的四个步骤。

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: 数据过滤的三个标准。

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: 复杂性缩放的三个子步骤。

2.2.2 数据集统计

mSCoRe-G 覆盖了包括英语、德语、法语、日语和中文在内的 5 种语言。对于每种语言，我们创建了 200 个例子，范围从等级 0（原始问答对）到等级 3（3 个扩展步骤）。这导致每种语言有 800 个例子。对于 mSCoRe-S，我们同样为每个来源（TikTok 和 Reddit）创建 200 个例子。总的来说，mSCoRe 包含了 5,600 个实例（4000 个用于一般常识，1600 个用于社交常识）。不同复杂性等级的详细例子在附录 5 中提供。

我们使用一组多样的最先进的多语言模型进行了全面的评估，这些模型被选择用来代表模型开发和训练的不同方法。我们的评估考虑了三个关键维度：模型的可用性、参数规模和训练方法。我们研究中评估的模型包括：

GPT-4o (OpenAI et al., 2024)：一个代表当前最先进的通用大型语言模型，经过来自多种来源的大规模多模态数据训练。

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: 各个模型在 mSCoRe-G 上从复杂度等级 0 (L0) 到 3 (L3) 的准确性比较。

OpenAI o1 (Jaech et al., 2024)：一个基于 GPT-4o 的推理增强模型，特别通过一个额外的训练阶段进行优化，以利用为连锁推理量身定制的数据来解决复杂问题的任务。

LLaMA-3.3-70B 和 LLaMA-3.1-8B (Grattafiori et al., 2024)：两个具有不同参数规模的开源大型语言模型，训练于跨越各种领域的公开可用资源，使我们能够分析模型大小对推理能力的影响。

蒸馏版 DeepSeek-R1（R1-70B 和 R1-8B）(DeepSeek-AI et al., 2025)：一种从 LLaMA 架构派生的、以推理为重点的模型，使用大规模 LRM DeepSeek-R1 生成的样本进行蒸馏。

Aya-32B (Dang et al., 2024)：一个通用的多语言模型，基于来自 200 种语言的数据进行训练，提供了对广泛的多语言大语言模型推理能力的洞察。

为了评估，我们为所有模型采用一致的提示，提供所提出的推理技能分类（章节 2.1 ）并逐步给出指导，以在回答之前生成所需的推理过程。更多的实验细节见附录 **??** 。

2.3 主要结果

表 3 和 4 分别展示了我们关于 mSCoRe-G 和 mSCoRe-S 的主要结果。总体而言，我们观察到在所有模型中，随着复杂性水平的增加，性能都呈现下降的趋势。对于 mSCoRe-G，在所有语言和复杂性水平上，GPT-4o 在普通常识推理方面的整体准确性最高。虽然这可能是因为在基准创建过程中使用了 GPT-4o 进行数据生成，但 LLaMA-3.3-70B 的结果非常接近于 GPT-4o。此外，开源模型在社交常识推理上明显优于其他模型（在所有水平和领域上平均提升超过 5 %）。

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: 各模型在 mSCoRe-S 上的准确性比较。

多语言和文化结果：在 mSCoRe-G 中，不同语言的表现总体相似。这可能是由于种子数据集 mCSQA 中的所有语言都是中等到高资源语言。未来的工作应该探索包含更多低资源语言的其他种子数据集。在 mSCoRe-S 中的社会常识推理方面，大多数模型在来自 Reddit 的问题上的表现优于 TikTok 的问题。这可能归因于 Reddit 包含更多关于“社区和文化交流”的内容，而 TikTok 更关注于“社会规范和礼仪”等日常生活“个人”方面。这表明大型语言模型可能仍然在处理更个性化的问题时遇到困难，正如在 Davis (2024) 中所指出的。出乎意料的是，尽管 Aya-32B 模型经过 200 种不同语言的训练，但它在多语言文化社会常识基准测试中表现并不理想。

模型规模：我们比较了不同参数总数的模型，从 8B 和 70B 参数的开源模型（LLaMA 和 R1），到巨型规模（数百亿参数）的闭源 LLMs（GPT-4o 和 o1）。在两个基准测试中，较大的模型通常表现更好。70B 和 8B 版本之间的性能差距在大多数情况下都很明显。然而，我们观察到从 70B 过渡到巨型规模 LLMs 时，收益递减。这一发现表明，解决常识推理问题需要的不仅仅是简单的参数扩展，尤其是在理解社会互动和文化规范时。

推理增强训练：我们比较了通用指令微调模型（GPT-4o，LLaMA）和推理增强微调模型（o1 和 R1）。尽管最先进的 LRM o1 在英语中表现最佳，但在其他语言中，它落后于其他通用 LLM，如 GPT-4o 和 LLaMA-3.3-70B。这表明推理增强的训练可能会降低常识性推理能力，这可能是由于用于更复杂任务（如编码和数学）的高度专业化训练数据造成的。有趣的是，

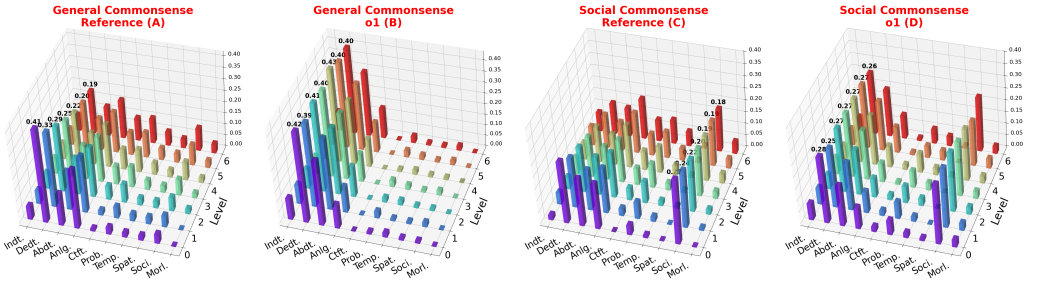

Figure 6: 当问题复杂性从复杂性等级 0 增加到 6 时，参考推理过程（A 和 C）和 o1 的推理过程（B 和 D）的推理技能分布。

LLaMA-3.1-8B 在英语和日语中的任务失败，但 R1-8B 表现正常，这表明推理增强训练有助于更小规模的模型更好地理解任务。

3 分析

为了进一步理解模型在扩展问题复杂度方面的能力，我们将 mSCoRe-G（英语）和 mSCoRe-S 的结果扩展到复杂度等级 6。如表 **??** 所示，每个模型的准确率在 L6 时继续下降。最大幅度的性能下降发生在 L0 和 L2 之间，这表明即使是相对简单的复杂度扩展也会给大语言模型带来重大挑战。在更高的难度级别（L3 到 L6），降解速度显著减慢。这种平稳表明，通过增加上下文和推理步骤来扩展复杂度的方法可能已经达到了一个饱和点。这可能表明，多项选择题-答案格式本身对任务难度的有效扩展施加了某些限制。可能需要替代的任务形式，需要超越当前设计的更复杂的推理形式，以建立未来更强大模型的更具鉴别力的基准。

为了更好地理解模型如何在不同复杂性水平上运用不同的推理技能，图 6 可视化了在参考推理过程（来自我们的基准创建）和由 o1 生成的输出推理过程所使用的推理技能的分布。

对于一般常识，参考和模型生成的推理主要利用逻辑推理技能，其中演绎推理最为常见。然而，在更高复杂度的情况下，参考分布显示出技能的更大多样化，包含更多的情境推理（尤其是类比和概率推理）。相比之下，像 o1 这样的模型在所有复杂度水平上仍然严重依赖于演绎推理。对于社会常识，参考分布显示出三类技能的更加均衡的利用，随着问题层级的提高，社会和伦理推理变得愈发重要。虽然 o1 模型包含了一些社会推理技能，但在需要社会和情境推理的情景中，它仍然过度依赖逻辑推理。总体而言，结果揭示了 o1 在适应其推理策略上的显著局限性。僵化的推理模式可能解释了模型在高复杂度问题上的性能下降，这突显了更均衡的推理强化训练方法的必要性。

3.1 不同的推理技能分类法

我们研究模型如何适应不同的推理分类法，包括：(1) 连锁思维 (CoT) - 标准的连锁思维，不需要技能识别，(2) 逻辑 - 仅使用逻辑推理技能（演绎、归纳和溯因推理），(3) 通用 - 每个推理步骤被归类为三种通用类别之一（逻辑、情境和社会）。

	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 5: 不同推理技能分类的结果。

表 5 显示了 o1 和 o1-mini 在每种设置下的平均准确率。有趣的是，尽管要求模型区分更多的技能类型，我们提出的细粒度分类法却取得了最佳结果。正如我们之前分析所预期的，只使用逻辑的方式在一般常识任务中表现相对较好，但在社交任务中表现较差。通用设置的表现也不如我们的设置，这表明技能识别的细粒度通过鼓励模型考虑更广泛的推理方法而不是依赖熟悉的模式，有助于常识推理。最后，CoT 的表现明显不如所有基于结构化技能的方法，尤其是在较高复杂度级别的社交常识方面。这表明在没有明确的技能分类的情况下进行推理可能不足以应对更复杂的常识情况。

为了研究不同模型中推理效率与任务复杂性之间的关系，图 7 展示了在不同复杂性水平下，mSCoRe 和 GPT-4o 以及 o1 的答案的推理过程的平均步骤数。

参考推理过程显示，随着复杂性水平的增加，推理步骤呈现出明显的线性增长，其中社会常识推理在每个水平上比一般常识需要更多的步骤。GPT-4o 的推理过程也显示出类似的上升趋势，但坡度更为平缓，而 o1 的推理过程保持几乎恒定的步骤数（大约 3 步），无论任务复杂度如何。结果表明，高级别复杂性需要更多步骤，目前的模型无法进行更长时间的推理，

除非明确强制要求（例如在 2.2 中描述的复杂性扩展步骤）。这类似于最近在 Muennighoff et al. (2025) 中引入的测试时间扩展研究中使用的推理标记数量（更多步骤相当于更多标记）。这些发现表明，基于任务需求动态调整推理深度可能是维持性能随着复杂性升级而不下降的关键。我们介绍了 mSCoRe，一种基于技能的常识推理的多语言和可扩展基准。这是一种新颖的评估框架，旨在解决现有常识推理基准中的关键缺口。通过整合多语言和多样化的文化覆盖、精细的推理技能分类以及动态复杂性扩展机制，mSCoRe 提供了一个全面的平台，不仅系统地评估 LLMs 常识推理过程的准确性，还评估技能的利用和效率。在八个最先进的 LLM 上进行的大量实验表明，当前模型在处理更高复杂性水平和具有文化细微差别的社会常识场景时仍然存在一致的困难。我们的分析强调了一些有前途的改进方向，包括采用更强大的训练方法来增强模型的推理技能利用率和效率。此外，mSCoRe 为后续基准提供了一个框架，以应对未来 LLMs 的快速发展。

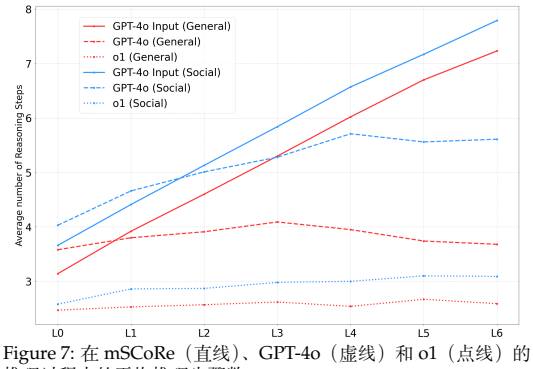

Figure 7: 在 mSCoRe（直线）、GPT-4o（虚线）和 o1（点线）的推理过程中的平均推理步骤数。

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *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*, pp. 17682–17690. AAAI Press, 2024. doi: 10.1609/AAAI.V38I16.29720. URL https://doi.org/10.1609/aaai.v38i16.29720.

Sijia Chen, Baochun Li, and Di Niu. 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, 2024. URL https://openreview.net/forum?id=qBL04XXex6.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Q6a9W6kzv5.

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. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/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. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

Ernest Davis. Benchmarks for automated commonsense reasoning: A survey. *ACM Comput. Surv.*, 56(4):81:1–81:41, 2024. doi: 10.1145/3615355. URL https://doi.org/10.1145/3615355.

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. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 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*, 2024. URL https://openreview.net/forum?id=a34bgvner1.

Quyet V. Do, Junze Li, Tung-Duong Vuong, Zhaowei Wang, Yangqiu Song, and Xiaojuan Ma. What really is commonsense knowledge? *CoRR*, abs/2411.03964, 2024. doi: 10.48550/ARXIV.2411.03964. URL https://doi.org/10.48550/arXiv.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. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in AI. *CoRR*, abs/2411.04872, 2024. doi: 10.48550/ARXIV.2411.04872. URL https://doi.org/10.48550/arXiv.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. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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. Deepseek-coder: When the large language model meets programming - the rise of code intelligence. *CoRR*, abs/2401.14196, 2024. doi: 10.48550/ARXIV.2401.14196. URL https://doi.org/10.48550/arXiv.2401.14196.

Daichi Haraguchi, Kiyoaki Shirai, Naoya Inoue, and Natthawut Kertkeidkachorn. Discovering highly influential shortcut reasoning: An automated template-free approach. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 6401–6407. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.424. URL https://doi.org/10.18653/v1/2023.findings-emnlp.424.

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. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.211. URL https://doi.org/10.18653/v1/2024.acl-long.211.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. Advancing language model reasoning through reinforcement learning and inference scaling. *CoRR*, abs/2501.11651, 2025. doi: 10.48550/ARXIV.2501.11651. URL https://doi.org/10.48550/arXiv.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. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

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. Llm post-training: A deep dive into reasoning large language models, 2025.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. Culturepark: Boosting cross-cultural understanding in large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/77f089cd16dbc36ddd1caeb18446fbdd-Abstract-Conference.html.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 1274–1287, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.102. URL https://aclanthology.org/2021.acl-long.102/.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *CoRR*, abs/2501.19393, 2025. doi: 10.48550/ARXIV.2501.19393. URL https://doi.org/10.48550/arXiv.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 Mdry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and et al. Alex Paino. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL https://aclanthology.org/2020.emnlp-main.185/.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 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, 2011. URL http://www.aaai.org/ocs/index.php/SSS/SSS11/paper/view/2418.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14182–14214, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.844. URL https://aclanthology.org/2024.findings-acl.844/.

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

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério de Paula, and Diyi Yang. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA,*

November 12-16, 2024, pp. 4996–5025. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-emnlp.288.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *CoRR*, abs/2408.03314, 2024. doi: 10.48550/ARXIV.2408.03314. URL https://doi.org/10.48550/arXiv.2408.03314.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. Benchmarking Chinese commonsense reasoning of LLMs: From Chinese-specifics to reasoning-memorization correlations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11205–11228, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.604. URL https://aclanthology.org/2024.acl-long.604/.

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. Challenging big-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.824. URL https://doi.org/10.18653/v1/2023.findings-acl.824.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4149–4158. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1421. URL https://doi.org/10.18653/v1/n19-1421.

Wikipedia contributors. Commonsense knowledge (artificial intelligence) — Wikipedia, the free encyclopedia, 2025. URL https://en.wikipedia.org/w/index.php? [Online; accessed 25-March-2025].

Dingjun Wu, Jing Zhang, and Xinmei Huang. Chain of thought prompting elicits knowl-

edge augmentation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 6519–6534. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.408. URL https://doi.org/10.18653/v1/2023.findings-acl.408.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-o1: Towards open reasoning models for open-ended solutions. *CoRR*, abs/2411.14405, 2024. doi: 10.48550/ARXIV.2411.14405. URL https://doi.org/10.48550/arXiv.2411.14405.

对于封闭的商业大型语言模型（GPT-4o, o1 和 o1-mini），我们使用 OpenAI Chat Completions API [1] 获取模型的响应，温度设为 0 以获取确定性的输出。开源模型（Deepseek R1-70B [2] 和 R1-8B [3]，LLaMA-3.3-70B [4] 和 LLaMA-3.1-8B [5]，Aya-32B [6]）使用 2 张 NVIDIA A100 80GB GPU 运行。使用 PyTorch 2.1.2 [7] 和 Huggingface-Transformer 4.42.3 [8] 来实现这些模型。

我们将在论文被接受后发布我们的数据和源代码。

[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

3.2 推理技能细节

我们在图 8 中为每种推理技能提供了详细描述，包括抽象和具体示例。抽象示例是泛化的表示，使用变量或占位符来说明推理技能的模式或原则。相比之下，具体示例是将相应推理技能应用于特定现实场景。

4 数据生成细节

4.1 LLM-裁判

判定模型 我们使用了由 Flow AI [9] 开发的通用大语言模型裁判模型 Flow Judge。Flow Judge 是一个开源的 3.8B 参数语言模型，专为基于语言模型的评估而设计，提供了与更大模型如 GPT-4o 和 Claude 3.5 Sonnet 相当的高性能和高准确度。它在各个领域的评估数据上进行训练，以支持自定义评估标准、多种评分尺度、定性反馈，并生成结构化的评估输出。

我们在图 9 和 10 中分别提供用于 mSCoRe-G 和 mSCoRe-S 的种子数据过滤步骤的完整规则。

5 提示详情

我们在此提供用于我们实验的完整版本的所有提示模板。图 11 、12 和 13 展示了 mSCoRe-G 的数据生成过程中结构化推理生成、数据复杂性缩放和常识隐喻步骤的提示。此外，图 15 展示了 mSCoRe-S 的结构化推理生成步骤的提示（其他两个步骤在两个子集中保持不变）。

我们在图 14 和 16 中分别提供了 mSCoRe-G 和 mSCoRe-S 的复杂度等级 0 到 3 的例子。

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

```
{
"inductive_reasoning": {
    "short_description": "Drawing general conclusions from specific observations.",
    "long_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": {
    "short_description": "Deriving specific conclusions from general premises.",
    "long_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": "Given All birds have feathers. A sparrow is a bird. Therefore, a sparrow has feathers"
},
"abductive_reasoning": {
    "short_description": "Forming hypotheses to explain observations.",
    "long_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": {
    "short_description": "Drawing parallels between similar situations to infer conclusions.",
    "long_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).
    Analogical reasoning is also used in everyday life to make inferences about the similarities between objects or
    situations.",
    "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 Component C_b_1 and
    Component C_b_2 will interact similarly in Situation C_a.",
    "concrete_example": "Just as a gardener waters plants to help them grow, a teacher provides knowledge and guidance
    to help students develop."
},
"counterfactual_reasoning": {
    "short_description": "Considering alternative scenarios and outcomes that did not happen.",
    "long_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. Counterfactual reasoning is often used
    in fields such as philosophy, psychology, and business to explore the potential consequences of different choices
    or actions.",
    "abstract_example": "Reflecting on Condition X that did not occur, you imagine that if it had, Outcome Y might have
    replaced Outcome Z.",
    "concerte_example": "If you had left the house five minutes earlier, you would have caught the bus on time."
},
"probabilistic_reasoning": {
    "short_description": "Applying principles of probability to make inferences under uncertainty.",
    "long_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. This type of reasoning is crucial in fields like statistics,
    risk assessment, and artificial intelligence.",
    "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
},
"temporal_reasoning": {
    "short_description": "Understanding sequences and durations of events.",
    "long_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. Temporal reasoning is essential in areas like scheduling,
    planning, and understanding narratives.",
    "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": {
    "short_description": "Visualizing and manipulating objects in space.",
    "long_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 object.",
    "abstract_example": "While arranging furniture, you visualize Object S_1 and Object S_2 to determine their optimal
    placement within the room.",
    "concrete_example": "A 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": {
    "short_description": "Understanding social interactions and norms.",
    "long_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": {
    "short_description": "Deciding what is right or wrong based on ethical principles.",
    "long_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: 推理技能详情。

```
### 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?",
    "rubric": {
        "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.",
        "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.",
        "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?",
    "rubric": {
        "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.",
        "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.",
        "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.",
        "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.",
        "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?",
    "rubric": {
        "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.",
        "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.",
        "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: 用于 mCSQA 数据过滤过程的 评分标准

```
### Multicultureness
{
    "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?",
    "rubric": {
        "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.",
       "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.",
        "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?",
    "rubric": {
       "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.",
        "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.",
        "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?",
    "rubric": {
        "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.",
        "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.",
        "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.",
        "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.",
        "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.",
    "rubric": {
        "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.",
        "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.",
        "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: 用于 CultureBank 数据筛选过程的评分标准

```
### 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:
        + inductive_reasoning: Drawing general conclusions from specific observations.
        + deductive_reasoning: Deriving specific conclusions from general premises.
        + abductive_reasoning: Forming hypotheses to explain observations.
        + analogical_reasoning: Drawing parallels between similar situations to infer conclusions.
        + counterfactual_reasoning: Considering alternative scenarios and outcomes that did not happen.
        + probabilistic_reasoning: Applying principles of probability to make inferences under uncertainty.
        + temporal_reasoning: Understanding sequences and durations of events.
        + spatial_reasoning: Visualizing and manipulating objects in space.
        + social_reasoning: Understanding social interactions and norms.
        + moral_reasoning: Deciding what is right or wrong based on ethical principles.
    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: 为 mSCoRe-G （英文版）生成结构化推理步骤的提示。

```
### 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":
    + inductive_reasoning: Drawing general conclusions from specific observations.
    + deductive_reasoning: Deriving specific conclusions from general premises.
    + abductive_reasoning: Forming hypotheses to explain observations.
    + analogical_reasoning: Drawing parallels between similar situations to infer conclusions.
    + counterfactual_reasoning: Considering alternative scenarios and outcomes that did not happen.
    + probabilistic_reasoning: Applying principles of probability to make inferences under uncertainty.
    + temporal_reasoning: Understanding sequences and durations of events.
    + spatial_reasoning: Visualizing and manipulating objects in space.
    + social_reasoning: Understanding social interactions and norms.
    + moral_reasoning: Deciding what is right or wrong based on ethical principles.
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: 用于 mSCoRe-G 复杂性扩展步骤的提示（英文）。

```
### 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":
    + inductive_reasoning: Drawing general conclusions from specific observations.
    + deductive_reasoning: Deriving specific conclusions from general premises.
    + abductive_reasoning: Forming hypotheses to explain observations.
    + analogical_reasoning: Drawing parallels between similar situations to infer conclusions.
    + counterfactual_reasoning: Considering alternative scenarios and outcomes that did not happen.
    + probabilistic_reasoning: Applying principles of probability to make inferences under uncertainty.
    + temporal_reasoning: Understanding sequences and durations of events.
    + spatial_reasoning: Visualizing and manipulating objects in space.
    + social_reasoning: Understanding social interactions and norms.
    + moral_reasoning: Deciding what is right or wrong based on ethical principles.

### 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: 用于 mSCoRe-G（英语）的常识含隐步骤提示。

```
### 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: 一个来自 mSCoRe-G 的例子，复杂度等级从 0 到 3（英语）。

```
### 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":
        + inductive_reasoning: Drawing general conclusions from specific observations.
        + deductive_reasoning: Deriving specific conclusions from general premises.
        + abductive_reasoning: Forming hypotheses to explain observations.
        + analogical_reasoning: Drawing parallels between similar situations to infer
        conclusions.
        + counterfactual_reasoning: Considering alternative scenarios and outcomes that did
        not happen.
        + probabilistic_reasoning: Applying principles of probability to make inferences under
        uncertainty.
        + temporal_reasoning: Understanding sequences and durations of events.
        + spatial_reasoning: Visualizing and manipulating objects in space.
        + social_reasoning: Understanding social interactions and norms.
        + moral_reasoning: Deciding what is right or wrong based on ethical principles.
    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: 结构化推理生成提示用于 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: 来自 mSCoRe-S 的一个例子，适用于复杂度等级 0 到 3（英语）。