SCR-Bench：
知识驱动的长链条常识推理数据合成与评价方法

刘群

华为AI语音语义首席科学家

中国多智能体系统大会

2025.06.07 西安

主要贡献者：北京大学 詹卫东 教授 团队

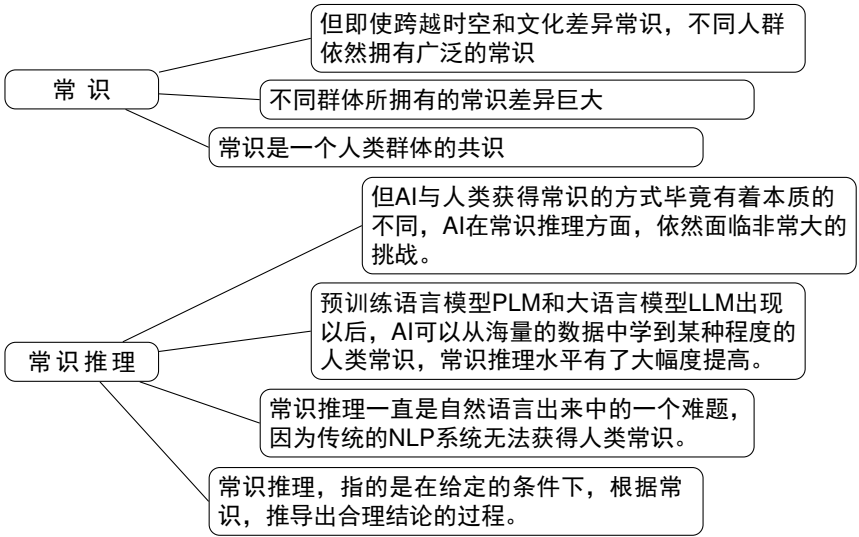
HUAWEI

Content

Content

Content

常识和常识推理



Content

现有的常识推理基准

CommonsenseQA 1.0

问题:
"Where would you most likely find a village located?"
（村庄最可能位于哪里？）

选项:

- (A) In the middle of a lake（湖中心）
- (B) On top of a mountain（山顶）
- (C) **Along a riverbank（河岸旁）** ✅
- (D) Underneath a desert（沙漠下方）
- (E) Floating in the sky（漂浮在空中）

标签：c
推理依据:
人类常识中，村庄通常依赖水源（如河流）生存，因此"河岸旁"符合现实逻辑；其他选项违背地理常识（如湖心无法建村、沙漠下无生存条件）。

Winograd Schema Challege

I. The trophy would not fit in the brown suitcase because it was too **big** (*small*). What was too **big** (*small*)?
Answer 0: the trophy
Answer 1: the suitcase

II. The town councilors refused to give the demonstrators a permit because they **feared** (*advocated*) violence. Who **feared** (*advocated*) violence?
Answer 0: the town councilors
Answer 1: the demonstrators

现有的常识推理基准

RTE (Recognizing Text Entilement)

任务目标判断假设（Hypothesis）是否可从前提（Premise）中推断成立（三分类：蕴含/矛盾/中立）

❌ **样例 2：矛盾关系（Contradiction）**

前提（Premise）：

> “奥巴马是首位获得诺贝尔和平奖的非洲裔美国总统。”

假设（Hypothesis）：

> “奥巴马是唯一获得诺贝尔和平奖的黑人。”

标签（Label）：`Contradiction` ❌

推理依据：

前提仅说明奥巴马是“首位非洲裔美国总统获奖者”，未排除其他黑人获奖者（如南非前总统德克勒克）。假设中的“唯一”添加了前提未支持的限制条件，构成矛盾。

bAbI

📖 **样例 1：单支持事实任务（Task: `qa1`）**

上下文（Context）：

```
1 Mary moved to the bathroom.
2 John went to the hallway.
3 Daniel went back to the hallway.
4 Sandra moved to the garden.
```

问题（Question）：

`Where is Mary?`

答案（Answer）：

`bathroom`

支持句子编号：`1`

推理逻辑：

答案直接源于第 1 句 `Mary moved to the bathroom`，无需跨句子关联。

现有的常识推理基准

TRAM

样例 1：事件精确排序（TRAM 子任务：Temporal Sequences）

问题：
“已知以下事件：

- A：公司发布第一季度财报（3月31日）
- B：公司宣布新产品研发启动（4月15日）
- C：股价单日上涨 8%（4月16日）
  问：事件 C 是否可能由事件 A 直接引起？请解释。”

答案与推理：

- **标签**：否 ❌
- **依据**：
  事件 A（3月31日）与事件 C（4月16日）间隔超过两周，股价波动通常由近期事件驱动（如事件 B 在4月15日）。时间逻辑要求模型识别 **延迟因果的合理性边界**（财报影响通常在1-3天内反映）。

MuSR

样例 1：谋杀悬疑推理（多步因果链整合）

叙事背景（约 1000 词摘要）：

> 富豪 Charles 在书房中毒身亡。前一天，他与妻子 Clara 激烈争吵财产分配问题；女仆 Emma 曾因偷窃被 Charles 威胁解雇；园丁 Thomas 被目击在案发前 1 小时携带农药进入书房。尸检显示毒素为氰化物，而 Thomas 的农药瓶仅含除草剂。警方在 Clara 的手袋中发现未使用的氰化物胶囊。

问题：
谁最可能是凶手？需结合动机、机会、物证三要素逐步推理。

答案与推理链：

1. **动机分析**：
   - Clara 与死者有财产纠纷（直接利益冲突） → **动机强**
   - Emma 虽有怨恨，但解雇威胁已解除（无即时杀人动机） → **动机弱**
   - Thomas 无已知矛盾，且农药无害 → **无动机**
2. **机会验证**：
   - Thomas 进入书房但未携带毒物（农药瓶无害） → **无作案工具**
   - Clara 未被目击进入书房，但手袋藏毒 → **具备隐蔽投毒机会**
3. **物证关联**：
   - 氰化物胶囊与死因匹配，且 Clara 持有同类毒物 → **直接物证链**
4. **结论**：`Clara` 是凶手（需综合三步推理，排除干扰项 Thomas）

常识推理测试基准与模型能力进展

- Before 2018,神经网络语言模型（MLP-LM、LSTM-LM）
  - CommonsenseQA，Winograd Shema Challange，RTE，bAbI
- 2018,预训练语言模型（BERT）
  - CommonsenseQA 2.0，PIQA
- 2023,大语言模型LLM（ChatGPT）
  - 家族排行问题，青蛙跳井问题，弱智吧数据集
- 2025,大推理模型LRM（o1、DeepSeek）
  - MuSR，TRAM，SCR-Bench

声明：常识推理数据集和测试基准非常多，这里只列出少数例子，不是系统性的总结。

Content

研究问题

论文

SCoRE: Benchmarking Long-Chain Reasoning in Commonsense Scenarios

Weidong Zhan[1], **Yue Wang**[2], **Nan Hu**[1], **Liming Xiao**[1], **Jingyuan Ma**[2], **Yuhang Qin**[1], **Zheng Li**[2], **Yixin Yang**[2], **Sirui Deng**[1], **Jinkun Ding**[1], **Qingxiu Dong**[2], **Wenhan Ma**[2], **Rui Li**[2] **Weilin Luo**[3], **Qun Liu**[3], **Zhifang Sui**[2]*

[1]Center for Chinese Linguistics, Department of Chinese Language and Literature, Peking University

[2]School of Computer Science, State Key Laboratory of Multimedia Information Processing, Peking Universi

[3]Huawei Noah's Ark Lab, China

szf@pku.edu.cn, zwd@pku.edu.cn

arXiv:2503.06218v2 [cs.CL] 17 May 2025

拟将本测试基准名称改为SCR-Bench

本文的贡献

1. 我们提出了SCR-Bench，一个双语基准，旨在通过多步推理链评估常识场景下的复杂逻辑推理。它由10万个常识推理问题组成，融合了不同层次的多样常识知识、多场景和长推理链。

2. 我们提出了一种知识驱动的数据合成方法，将人工构建小规模知识库和自动问题生成相结合，以最小的人力成本保证数据质量和数量。该方法具有透明和可追踪的工作流程，便于对LLM进行可解释的评估。

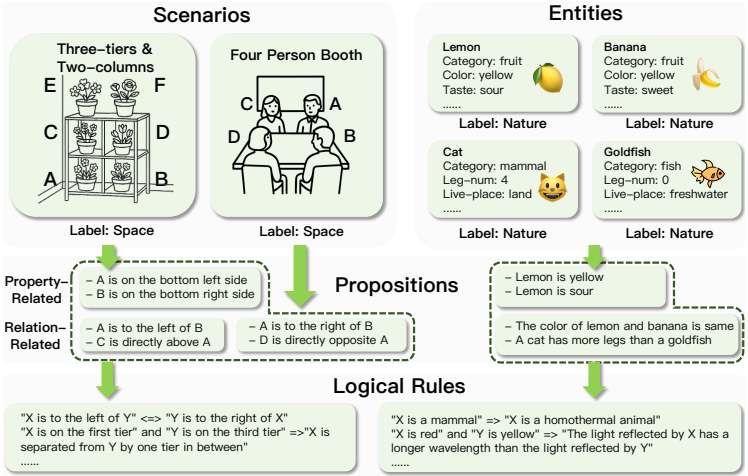
3. 我们使用13个最先进的推理LLM进行了评估。表现最好的模型取得了69.78%的分数，在难题子集上的平均准确率只有47.91%，说明SCR-Bench对于表现最好的模型来说是很有挑战性的。

4. 通过案例分析，我们发现了LLM在常识推理能力方面存在的几个显著缺陷，包括对低频常识知识的误解、逻辑上的自我矛盾和过度思考等。

Content

Content

知识库框架



以空间域和自然域实体为例

知识库框架说明

- 场景是我们知识库的核心：
  - 每个场景都有用于放置实体的指定槽，这些槽对实体施加了选择约束。
  - 场景是通过人工设计一个基于属性和基本关系的最小命题集，以及一组逻辑规则来推断所有关系来构建的。
  - 通过判断这些命题和规则在新的场景中是否适用，可以在构造新的场景时重用这些命题和规则。
- 实体是被填充到场景中的候选者，它也具有由命题和逻辑规则所描述的多种性质和关系：
  - 每个场景和实体都有“领域”标签注释，每个命题都用其领域、相关实体和关系标记，以便进行细粒度分析。
  - 实体及其属性是从现有的外部资源（如HowNet、ConceptNet和维基百科）中提取的。
  - 基于它们，我们手动总结与实体的属性或关系相关联的逻辑规则。
- 当前版本的知识库将实体、实体之间的关系和事实命题组织成结构化的场景，表示跨空间、时间、自然和社会领域的核心知识。知识库是可扩展的。

  目前，知识库包含11个场景、707个实体和与29个属性和109个关系相关联的939条规则。
- 尽管知识库的规模相对较小，但它具有产生大量问题的潜力。

  以“三层两列”场景为例，知识库中有633个实体满足该场景的约束条件。只要选择6个实体填充到此方案中，就会产生$A_{633}^{6} = 6 * 10^{16}$种可能的组合。

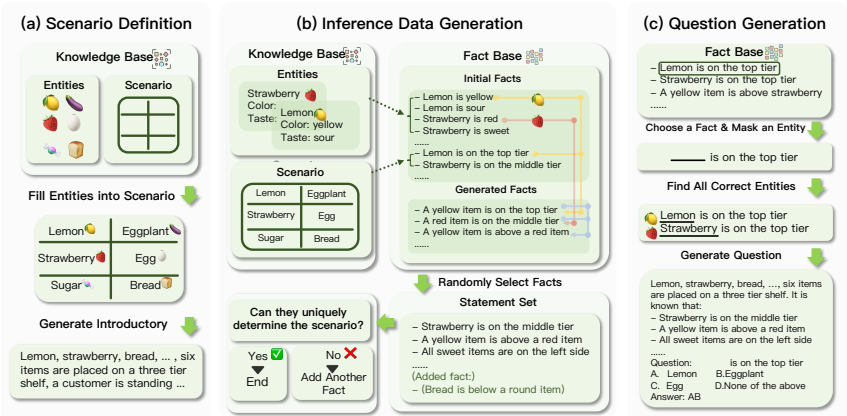
Content

以场景为中心的知识驱动的常识推理数据合成与评价方法

以场景为中心的常识推理知识库构建

知识驱动的常识推理问题合成

数据合成过程



(a) 场景定义：从知识库中选择实体/事件，构建自然语言描述；
(b) 推理数据生成：应用推理机通过基于规则的逻辑迭代地生成事实库，并选择唯一确定所有实体位置的最小语句集；
(c) 问题设计：使用语句集和答案关键，生成不同的问题类型（如精确、模糊、真/假），并带有适当的选项。
这个过程确保了可验证的推理链、丰富的逻辑结构和自然语言的流畅性。

数据合成过程详细介绍

第一步 场景定义：建立场景并生成其文本描述。

它包括从知识库中选择一个场景和少量满足场景约束的实体，并将这些实体填充到场景中。

这个过程将为上下文生成介绍性文本，我们采用预定义模板对场景元素进行自然语言转换。

数据合成过程详细介绍

第二步 推理数据：使用推理机来生成推理数据。

推理机是一个基于规则的程序，它将与场景关联的关系以及实体或事件的所有属性作为输入。

它维护一个事实库，刚开始事实库只包含所有实体和事件属性的描述。

- 程序会自动遍历知识库中的逻辑规则，将其与初始事实进行匹配，生成新的事实，并将其添加到事实库中。
- 程序将事实库中新添加的事实与原始事实一起作为新的初始事实，并将它们再次输入到推理机中。
  - 重复这个过程，直到不能生成新的事实为止。每个事实都用所涉及的属性或关系标记，以实现精确的分析。

事实库完成后，推理机会选择一组事实，这些事实可以唯一地确定场景中每个实体或事件的槽位。这一组事实可以构成一条长思维链，推导出所定义的场景。

- 对于每一个步骤，推理机从事实库中随机选择一个事实，将其添加到语句集，并验证语句集是否可以唯一确定每个实体或事件的槽。
- 在这个过程中，程序会自动记录语句中涉及的域、实体、属性和关系以及推理步骤的数量。
- 重复这一过程，直到答案为“是”。通过这种事实库生成和语句选择的反复叠加，可以形成一条显性的长推理链。

数据合成过程详细介绍

第三步 问题设计：使用问题生成器来设计问题。

问题生成器将语句集和实体或事件的ground-truth排列作为输入。它首先选择一个题型。

- 如果问题类型是“正确陈述”或“不正确陈述”，生成器将随机选择四对实体或事件来生成正确或不正确的陈述作为选项。
- 当题型为“单选”时，生成器生成一个命题，该命题可以唯一确定场景中实体或事件的槽位，并屏蔽相关信息。
- 当问题类型为“多选”时，生成器生成多个实体可以满足此命题的命题，并将所有这些实体识别为潜在答案。对于选项，如果场景有4个潜在答案，则所有实体/事件都将是选项。否则，它会随机选择其中的三个作为选项A、B和C，然后添加“以上都不是”作为选项D。

数据合成过程详细介绍

- 为了数据的质量，我们进行了三轮人工验证，经过这个过程，数据的准确率达到了100%。
- 这种方法自动将一个小的常识知识库转换为一个包含显式、可验证推理链的多步骤推理问题的大型数据集。它提供了四个关键的优点：

（1） 逻辑可靠性——每个问题都对应于一个形式化的推理过程；
（2） 组合丰富性——规则组合产生多样的常识模式；
（3） 可控的推理链——让我们调整链的长度和难度；
（4） 文本自然性——情景基础产生流利的、上下文丰富的问题陈述。

Content

知识驱动的常识推理基准（数据集）

全数据集包含由上述方法生成的100,000个问题。考虑到测试成本，我们特意选择了6000个能够覆盖我们知识库中所有知识点的问题组成测试数据集。

- 知识密度：测试集由6,000个中英文双语问题组成，涵盖空间、时间、社会和自然四大常识领域，以及带有多个“领域”标签的问题的混合领域，每个领域每个语言600个问题。问题涵盖了知识库中的所有知识点，共计11个场景，707个实体，29个属性和109个关系。
- 显式推理链：问题中的所有事实都由推理引擎生成，并且在数据合成过程中记录所有推理步骤。由于我们的数据合成过程中存在显式且可追踪的推理链，因此这些链的长度是高度可控的。我们的数据集涵盖推理链从2到11跳不等的问题，支持对模型的多跳推理能力进行分层评估。
- 细粒度标签：问题是通过逻辑规则连接领域知识而产生的，保证了推理过程的清晰和可追溯性。在推理过程中记录了7种标签：域、场景、实体、属性和推理过程中使用的关系，以及推理链长度和问题类型。模型出错的问题可以被分析并追溯到具体的知识点或推理步骤。
- 难度级别：根据问题的认知负荷、知识复杂度和信息完备度计算，数据集包括易、中、难三个难度级别。难度等级分布比例为1:2:3。我们认为，当前的LLM正处于从简单推理任务向复杂推理任务过渡的阶段。我们专注于更多中等或难问题，以加快模型推理能力的提升。

数据集样例

周伯通、郝大通、柯镇恶、赵志敬、刘处玄、王重阳六位道士在终南山重阳宫内盘腿席地打坐，围成一个圆圈，修炼内功，六人的位置恰好形成一个正六边形。六人都面朝外背对圆心而坐。任意相邻两人之间的间距相等，大约为一米。已知：
刘处玄的右边接着就是赵志敬，
郝大通在赵志敬的右边，二者相邻，
从赵志敬的左边数起第二个位置是王重阳，
从柯镇恶的左边数起第五个位置是王重阳，
从王重阳的左边数起第二个位置是周伯通。
问题：
赵志敬与___之间隔着两个位置。
选项：A.刘处玄　B.周伯通　C.柯镇恶　D.郝大通
答案：C

数据集样例

月季、水仙、茉莉、君子兰、天竺葵、郁金香六盆花放置在三层花架上呈列，花架紧靠大厅南墙放置，每层两格，各放一盆花，一在东，一在西。画师站在花架前，面对花架支起画架，为花架中六盆花画素描。在描述各花的方位关系时，约定以画师自身左右方位为参照，即东侧花盆为左，西侧花盆为右。东侧花盘在西侧花盘左边，西侧花盆在东侧花盆右边。已知：
郁金香的正上方是月季的正下方，
月季在天竺葵左上方且二者隔了一层，
天竺葵在一层西侧，
君子兰在天竺葵左上方且二者不隔层，
月季在茉莉左边，
水仙在君子兰右边。
问题：
___所在层和天竺葵所在层相邻。
答案：D
选项：A.茉莉　B.月季　C.郁金香　D.以上选项都不是

数据集样例

小明是一名大学生，以下是他的每周安排:
(1)星期三，他打羽毛球;
(2)周三，他开组会;
(3)在星期三，他跑步;
(4)在他打羽毛球之后1天，他阅读科幻小说;
(5)在他开组会的2天之后，他练习吉他;
(6)在他开组会之后3天，他看论文。
问题：以下选项中不正确的是____
选项:
A.在他打羽毛球的5天之前，他跑步。
B.在他阅读科幻小说的6天之后，他打羽毛球。
C.在他看论文的3天之前，他阅读科幻小说。
D.在他练习吉他之后2天，他阅读科幻小说。
答案：ACD

Content

Content

实验结果

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

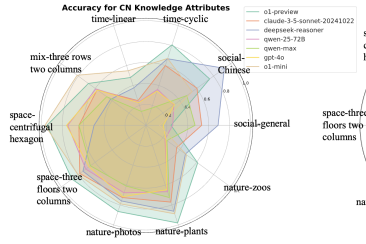
对于人类评估，我们为每个领域招募了10名本科生参与者，并随机抽样数据集的5%作为测试问题，覆盖了所有不同的层次、场景和问题类型。每位参与者需要在8小时的时间内完成60道题，其中30道中文题和30道英文题。

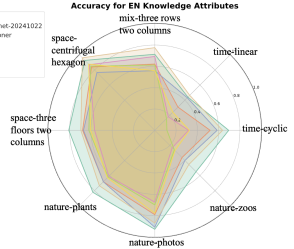
Content

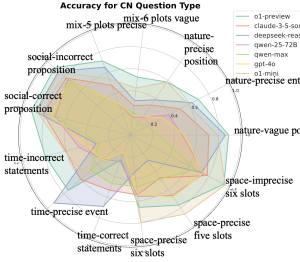
一般分析

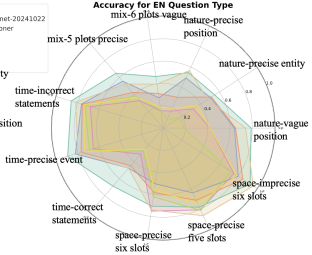
- 更难的问题需要更多token来回答
- 越是少见的场景，回答准确率越低
- 改变提问的方式会影响回复准确率
- 越难的问题，回复准确率越低
- 文化和语言差异也会影响回复准确率

不同知识类型和问题类型在中英文上的差异

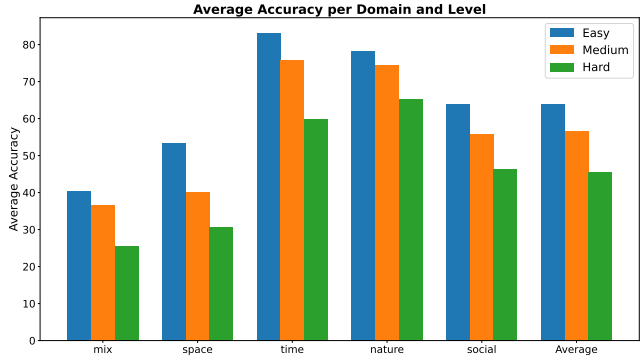








不同知识类型和难度上的差异



错误分析

- 低频实体属性和相似的社会关系容易产生常识错误；
- 模型的内部能力限制导致推理错误；
- 以推理为中心的LLM倾向于过度思考并提供额外的特殊案例作为条件；
- 模型会倾向于把多选题当成单选题来回答，只输出找到的第一个选项；
- 存在模型推理过程与给出的答案不一致的情况。

Content

总结

- 在本文中，我们介绍了SCR-Bench，这是一个双语数据集，旨在通过知识驱动的合成数据策略生成的多跳推理链来评估常识场景下的复杂逻辑推理。
- 该数据集包含10万个跨三个难度级别的问题，以及四个领域的常识问题。
- 我们的实验结果表明，尽管在代码生成和数学问题解决等领域取得了进展，但LLM在常识推理方面仍然面临挑战。
- 案例分析揭示了常见的错误，如曲解低频常识、逻辑不一致和过度思考。
- 本文所提出的数据合成方法能够自动生成具有QA对的大规模、高精度推理数据，这些数据可以作为训练数据，以进一步提高LLM的推理能力。

Thank you!

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI