

大语言模型故事理解能力评价数据集

闫国航^{1,†}, 郭亚鑫^{1,†}, 谭红叶^{1,2,*}, 张虎¹

¹山西大学 计算机与信息技术学院, 山西 太原 030006

²山西大学 计算智能与中文信息处理教育部重点实验室, 山西 太原 030006

{202112407002, 202222407055}@email.sxu.edu.cn

{tanhongye,zhanghu}@sxu.edu.cn

摘要

故事包含大量的社会、物理等常识, 同时蕴含深刻的道理, 是知识传播、文化传承、价值塑造的重要载体。故事理解是NLP中的一项重要任务。近几年, 研究者对大语言模型 (LLMs) 的语言理解能力进行了很多评估与分析, 但由于现有的故事理解数据集大多为答案出现在原文的实体类问题, 因此对LLMs故事理解能力的评价与分析非常有限。为此, 本文构建了一个寓言故事理解数据集CRMUS, 并基于人类故事理解的认知过程: 先进行常识推理, 然后理解故事寓意, 设计了两个任务来评价模型的相应能力。基于CRMUS数据集, 我们对多个代表性的LLMs进行了评估, 发现: LLMs已经可以较好地理解故事中的常识并进行推理, 但在理解故事寓意方面还存在很大提升空间。此外, 我们使用项目反应理论 (IRT) 对数据集进行了质量分析, 表明该数据集是高质量的, 可以有效地评估LLMs。

关键词: 故事理解; 常识推理; 数据集; LLMs; IRT

Benchmarking story comprehension ability of large language model

Guohang Yan^{1,†}, Yaxin Guo^{1,†}, Hongye Tan^{1,2,*}, Hu Zhang¹

¹School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China

²Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China

{202112407002, 202222407055}@email.sxu.edu.cn

{tanhongye,zhanghu}@sxu.edu.cn

Abstract

Stories contain a lot of social, physical and other commonsense, which is an important carrier of knowledge transmission, cultural inheritance and value shaping. Story comprehension is an important task in NLP. In recent years, researchers have witnessed extensive evaluation and analysis of Large Language Models' (LLMs) comprehension abilities. These assessments have been largely confined due to the predominant focus on entity-based questions within existing story comprehension datasets. Consequently, the scope for evaluating LLMs' comprehension of narratives has been severely restricted. Therefore, we propose CRMUS, a new fable story comprehension dataset. Based on the cognitive process of human story comprehension, we design two tasks to evaluate the model's ability in commonsense reasoning and allegorical understanding. Utilizing the

[†] Equal contribution

^{*} Corresponding Author

©2024 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

项目资助: 国家重点研发计划项目(2020AAA0106100)、国家自然科学基金面上项目(62076155)

CSMUS dataset, we evaluated several representative LLMs and found that the LLMs can understand the commonsense of the story and reason effectively, but there is still much room for improvement in understanding the morals of the story. Furthermore, we performed a quality analysis of the dataset using item response theory (IRT), showing that the dataset is of high quality and can effectively evaluate LLMs.

Keywords: Story Comprehension , Commonsense Reasoning , Dataset , LLMs , IRT

1 引言

故事是教育领域中的一种重要阅读材料，通常包含丰富的知识、生动的情节与鲜明的人物形象，同时还蕴含着深刻的道理，是知识传播、文化传承、价值塑造的重要载体。故事理解是自然语言处理（Natural Language Processing, NLP）领域的一项经典任务，在上世纪七十年代就受到了研究者的关注(Charniak, 1972; Winograd, 1972; Stede, 1996; Schubert and Hwang, 2000)。近几年，大语言模型（Large Language Model, LLMs）在NLP领域取得了很大进展，研究者利用SQuAD2.0(Rajpurkar et al., 2018)、RACE(Xu et al., 2017)等机器阅读理解数据集对LLMs的语言理解能力进行了评估与分析(Kocon et al., 2023)，发现其在相应数据集上已经具备类人的语言理解与文本推理能力，然而目前尚未对LLMs的故事理解能力进行系统评估。

相较于一般的机器阅读理解，故事理解不仅需要模型根据社会、物理等常识理解局部情节，还需要分析故事人物、意图和行为之间的关系，推断故事背后传递的深刻道理(Tomasulo et al., 2012; Pelletier and Beatty, 2015; Dorfman and Brewer, 1994)。因此，故事理解也是评估LLMs认知能力的一项重要任务。然而，目前的故事理解数据集大多为答案出现在原文的实体类问题，缺乏足够的挑战性。此外，这些数据集主要关注故事局部的、表层的语义理解，不能有效评估模型的故事深层理解能力。表 1展示了NarrativeQA(Kociský et al., 2018)和FairytaleQA(Xu et al., 2022)数据集中的示例。这两个示例均为故事人物方面的问题，答案“詹姆斯·麦迪逊”、“玛奇”都分别在故事原文中直接出现，且通过简单的词形匹配就能回答。这些问题对LLMs的故事理解能力进行的评价与分析非常有限。

数据集	样例（英文）	样例（中文）
NarrativeQA	<p>Story snippet: The United States of America, you will recall, was the style by which the rebellious colonies referred to themselves, in the Declaration of Philadelphia. The James Madison who is mentioned as the current President of the United States is now living, in exile, in Switzerland. His alleged predecessor in office...</p> <p>Question: Who is in exile in Switzerland? Answer: James Madison</p>	<p>故事片段: 你会记得，美利坚合众国是反叛的殖民地在《费城宣言》中所称的风格。被提及为美国现任总统的詹姆斯·麦迪逊目前流亡瑞士。他所谓的前任.....</p> <p>问题: 故事中谁在瑞士流亡? 答案: 詹姆斯·麦迪逊</p>
FairytaleQA	<p>Story snippet: it so happened that the great man was walking in his garden with his daughter madge that morning, and was feeling in an especially happy mood, so that when he suddenly looked up and saw a little boy before him, he said, kindly...</p> <p>Question: what was the name of the man's daughter? Answer: madge</p>	<p>故事片段: 碰巧那天早上，这个人正和他的女儿玛奇在他的花园里散步，心情特别愉快，所以当他突然抬头看到面前有一个小男孩时，他和蔼地说.....</p> <p>问题: 那个男人的女儿叫什么名字? 答案: 玛奇</p>

Table 1: NarrativeQA和FairytaleQA数据集中的示例

为此, 本文构建了一个新的具有挑战性的故事理解数据集CRMUS, 并基于人类理解故事的认知过程, 设计了两个任务: 常识问答、寓意理解。其中常识问答涉及到的常识类型有: 社会常识、生物常识、时间常识、空间常识和物理常识。上述两个任务均采用多项选择问答的形式, 每个问题包含四个选项。为了增加混淆性, 我们要求标注者提供的错误选项要与正确答案存在一定的语义关联。此外, 在寓意理解任务中, 除了由标注者提供的人工风格选项之外, 还利用LLMs生成机器风格的选项, 进一步增加选项的混淆性。具体来说, 该数据集包含436篇经典寓言故事和3400个选择题, 用于探测模型常识推理与寓意理解能力。

基于CRMUS数据集, 本文设置了提示学习和微调两个实验, 对多个LLMs进行了评估。具体来说, 提示学习实验中, 通过调用多个商用LLMs的API进行零样本提示; 在微调实验中, 通过微调开源的LLMs并进行提示。实验结果表明, 现有的LLMs可以较好地理解故事情节并根据常识知识进行推理(最佳结果83.45%), 但在深入理解故事寓意方面仍然存在一定困难(最佳结果67.33%)。此外, 我们还使用项目反应理论(Item Response Theory, IRT) (Baker, 1985)对数据集进行了质量分析, 并对上述LLMs的测试结果进行了排名。结果表明, 我们的数据集是高质量的, 可以有效地评估LLMs。

我们的数据集和代码可以在<http://cuge.baai.ac.cn/#/ccl/2024/crmus> 网站公开获取。

2 相关工作

2.1 故事理解数据集

随着深度学习与NLP技术的进步, 机器阅读理解成为一项研究热点。研究者推出了很多的阅读理解数据集, 其中包含一些与故事理解相关的数据集, 具体来说有以下几类。

多项选择问答数据集 该类数据集问题形式均为多项选择题形式。例如, 微软公司推出的MCTest(Richardson et al., 2013)数据集, 该数据集基于虚构的故事创建, 但仅涉及7岁儿童可理解的概念和词。Facebook公司基于儿童故事书构建了CBT数据集(Hill et al., 2016), 但由于采用自动构建方式, 数据集质量难以保证, 存在很大噪声。上述数据集中的问题都比较简单, 仅根据文本表面信息就能回答(Kaushik and Lipton, 2018)。

生成式问答数据集 该类数据集要求模型生成问题答案, 例如, 基于长篇书籍和电影剧本众包构建的NarrativeQA数据集(Kociský et al., 2018)、基于幼儿和青少年童话故事由专家构建的FairytaleQA数据集(Xu et al., 2022)。这些数据集大多为实体类问题, 只关注故事情节和人物关系, 局限于故事文本表层的语义理解, 对模型能力的评估非常有限。

与上述数据集不同的是, 本文提出的数据集重点考察模型对故事背后传递寓意的理解。此外, 数据集还引入故事理解所需要的常识推理任务, 评估模型结合常识知识进行推理的能力。

2.2 常识推理数据集

常识推理是认知智能的重要组成部分(Liu and Singh, 2004; Storcks et al., 2020), 是LLMs面临的主要挑战(Zhou et al., 2021; Bhargava and Ng, 2022)。近年来研究者构建了很多常识推理数据集, 主要类型有两种: (1) 一般常识推理数据集, 如WSC(Levesque et al., 2011)、OpenBookQA(Mihaylov et al., 2018)、CommonsenseQA(Talmor et al., 2019)等。(2) 特定常识推理数据集, 如科学常识数据集ARC(Clark et al., 2018)、时间常识数据集MC-TACO(Zhou et al., 2019)、物理常识数据集PIQA(Bisk et al., 2020)、社会常识数据集Social IQA(Sap et al., 2019)等。但现有的常识推理数据集存在的问题主要有: (1) 使用的上下文较简单, 大多为一两个句子的简单情景(Davis, 2023)。(2) 回答问题所需的常识可从相应的知识库或语料库中检索得到。(3) 涉及的常识类型单一或没有明确指出常识类型。

与上述数据集不同的是, 本文数据集提供了完整的故事文本, 设计的问题大多需要对故事整体、深层的语义理解, 且所需的常识并不显式的存在于上下文中, 很难直接从常识知识库中检索得到。此外, 本文数据集涉及故事理解中需要的五类常识, 并明确标注了问题考察的常识类型。

2.3 寓意理解数据集

现实生活中除了采用故事, 人们还可以通过漫画、谚语等载体传递隐含寓意。目前已经有研究者针对漫画、谚语等构建了相应的数据集。如AI2研究院提出的漫画幽默理解任务(Hessel

et al., 2023), 用于评估多模态模型理解和分析漫画隐含幽默意义的的能力; 北卡罗来纳大学提出谚语理解数据集ePiC(Ghosh and Srivastava, 2022), 评估模型谚语意义理解能力。但上述工作没有关注模型理解寓意所需的中间任务。

受上述两个工作的启发, 本文基于寓言故事来探索模型的寓意理解能力。此外, 我们基于人类理解故事的认知过程引入了常识推理来进一步探测模型的故事理解能力。

3 数据集构建

3.1 故事理解所需的常识知识

故事理解涉及到表层语义理解与深层语义理解。其中, 表层语义理解需要结合文本信息和自身具有的常识知识进行推理; 深层语义理解需要进一步结合各种常识知识分析故事中人物、意图和行为之间的关系, 构建故事人物的内部心理状态(Bruner, 1985; Pelletier and Wilde Astington, 2004; Peskin and Astington, 2004), 推断故事背后传递的深刻道理(Tomasulo et al., 2012; Pelletier and Beatty, 2015; Dorfman and Brewer, 1994)。常识作为人类对世界的普遍认知知识, 能够帮助理解故事隐含信息、解释角色行为、预测故事发展, 对故事理解具有重要作用。

本文对故事理解需要的常识进行了分析, 发现涉及到的常识类型主要有以下五种:

- **社会常识** 主要指与故事角色、事件相关的常识, 包括人物性格、事件原因和结果、道德文化、行为准则等。如: 鞠躬表示尊敬、作恶多端的人会被人厌恶等。
- **生物常识** 主要指与故事中的常见生物属性相关的知识, 包括生物体型、捕食关系、行为特征、能力、栖息地等。如: 猴子会爬树、狼是肉食动物等。
- **时间常识** 主要指与故事时间理解相关的常识, 包括事件持续时间、事件顺序、时间点、频率、时间间隔等。
- **空间常识** 主要指对空间的理解和认知, 包括地理位置、方向、距离、物体形状和大小等相关知识。如: 石子体积较小可以装进瓶子里。
- **物理常识** 主要指与故事事实体的物理性质有关的知识, 包括重量、密度、硬度、温度等。如: 石头会沉在水底、盐能溶于水等。

3.2 任务设置

本文提出了常识推理与寓意理解两个任务, 形式均为多项选择, 任务描述如下:

常识推理 该任务要求模型根据故事文本回答相关的常识问题, 每个问题包括4个选项。为了增加混淆性, 我们要求标注者提供的错误选项要与问题主题相关, 且与正确答案在语义或属性等方面存在一定关联。如表 2 示例1中的错误答案“乒乓球”与正确答案“玻璃球”在形状与大小方面具有相似性, 但在质量与密度上不同, 因此, 在该问题场景下具有较强的混淆性。

寓意理解 该任务要求模型从四个选项中选择最符合故事的寓意。我们在该任务中引入了人工风格和机器风格的选项, 其中人工风格的选项由标注者撰写, 机器风格的选项由LLMs生成。类似地, 为了增加错误选项的混淆性, 我们要求人工风格的错误选项要与故事寓意存在一定相关性。我们直觉上认为, 机器风格的选项对问题解答模型更具混淆性。如表 2 示例2所示, 错误选项“强大的力量才能解决困境, 智慧也得依靠力量”与正确选项“聪明机智有时比力量更为有效”都与智慧和力量相关, 因此也具有足够的混淆性。

故事	乌鸦口渴得要命，飞到一只大水罐旁，水罐里没有很多水，他想尽了办法，仍喝不到。于是，他就使出全身力气去推，想把罐推倒，倒出水来，而大水罐却推也推不动。这时，乌鸦想起了他曾经使用的办法，用口叼着石子投到水罐里，随着石子的增多，罐里的水也就逐渐地升高了。最后，乌鸦高兴地喝到了水，解了口渴。
常识推理任务示例	问题：文中乌鸦还可以将什么东西丢到罐子里来喝到水？ 选项：A.石狮子；B.乒乓球；C.树叶；D.玻璃珠
寓意理解任务示例	问题：下列哪个选项最符合故事说明的寓意？ 选项：A.不注重思考和创新，而只依赖过去的经验是不行的。B.聪明机智有时比力量更为有效。C.强大的力量才能解决困境，智慧也得依靠力量。D.有时团队协作能够克服困难。

Table 2: CRMUS数据集任务示例

3.3 数据集构建

数据收集 我们从网站¹下载了436篇《伊索寓言》故事，每个故事包含：题目、故事正文、寓意句²。我们修改了其中的错别字以及不通顺的句子，还对部分冗余的寓意句进行改写，在不改变语义的情况下，使其更为简洁、清晰。

数据标注 本文按照以下三个步骤标注数据：

- **标注准备** 我们制订了一个包括任务定义和示例的标注大纲。在此基础上，邀请了我们团队具有NLP相关知识的10名研究生参与标注。为了提高标注效率和质量，首先要求不同的标注者对同一故事进行独立标注，然后总结归纳标注过程中存在的问题，并优化标注大纲。
- **初始标注** 对于常识推理任务，为了确保问题的多样性，要求至少两个标注者分别为每个故事至少提出4个问题并撰写对应的选项，问题尽可能覆盖社会、生物、时间、空间、物理五个常识类型。此外，为了更好地揭示模型的局限，还要求标注者标注问题涉及的常识类型和对答案的详细解释。对于寓意理解任务，我们使用故事寓意句作为正确答案，并要求标注者撰写3个不同寓意作为错误答案。此外，我们还额外要求标注者通过提示学习的方式，利用LLMs为每个故事额外标注了两个题目。具体来说，为了提高选项的多样性，要求标注者自行设计提示模板，采用不同的LLMs根据故事生成多个寓意，并由标注者进行筛选和改写，使其更符合故事现有寓意的表述，将其作为另外两个题目的候选答案。
- **质量控制** 我们采取交叉检查的方式处理收集来的数据。对于常识推理任务，我们要求检查者为每个问题打分，0分为不合格，1分为合格，2分为优秀，并对部分题目进行修改或额外标注。最后由非标注者进行二次验证，删除其中的不合格数据。对于寓意理解任务，检查者需要仔细阅读每个选项，对不符合要求的进行修改或重新标注。

最后，我们调整了数据集的正确答案分布，使其随机均匀分布在A, B, C, D四个选项上。

3.4 数据集相关信息

经过统计，我们根据436篇寓言故事，构建了3911个常识推理问题和1308个寓意理解问题。对于常识推理任务，我们尽可能的删去了所有低质量的问题，最终筛选出2092个高质量的问题，平均每个故事约4.8个问题。我们按照一定比例将数据集划分为开发集和测试集（约1:4），具体信息如表3所示。

任务	开发集	测试集	总计
常识推理	400	1692	2092
寓意理解	252	1056	1308

Table 3: CRMUS数据集统计

¹<https://m.thn21.com/Article/chang/3306.html>

²有少量故事不包含寓意，我们为其手动编写了符合故事的寓意。

3.5 数据集规模

常识推理任务中各类型常识统计如图 1 所示，其中 47.23% 问题需要社会常识推理，19.2% 的题目需要生物常识推理。在未来的工作中，希望通过寻找更多涉及这些知识的故事扩充数据集。

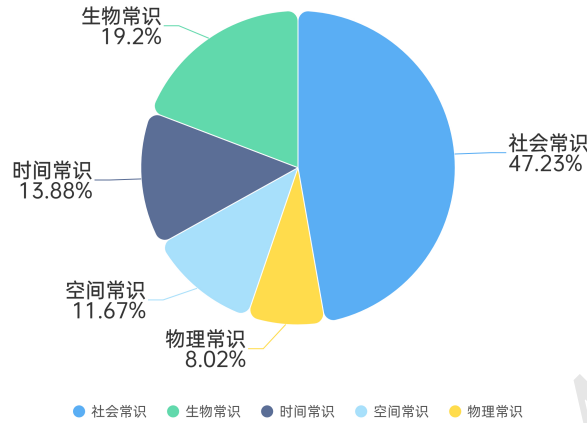


Figure 1: CRMUS 常识推理问题中的常识类型分布占比(%)

4 实验

本文基于 CRMUS 数据集对现有的多个 LLMs 进行了评估。

4.1 基线模型

本文根据新华社研究院推出的《人工智能大模型体验报告3.0》³，选择其中智商指数排名靠前的多个大模型进行测试，包括智谱清言大模型 GLM-3-Turbo⁴、GLM-4⁴、科大讯飞的星火大模型 v3.5⁵、文心一言 ERNIE-3.5-8K-0205、ERNIE-4.0-8K⁶、OpenAI 的 GPT-3.5-turbo (Ouyang et al., 2022)、昆仑万维天工大模型 SkyChat-MegaVerse⁷ 以及阿里通义千问大模型 qwen-turbo⁸。此外，我们还对 Chinese-Alpaca-2 (Cui et al., 2023) 进行了测试。这是一个开源的可微调的 LLMs，它是在 Llama-2 (Touvron et al., 2023) 的基础上使用大规模中文数据增量预训练得到的，其中文基础语义理解和指令理解能力相对于 Llama-2 有较大的提升。

4.2 实验设置

为了进行综合评估，文本设置了提示学习和微调两个实验：

- **提示学习**：调用非开源的超大规模语言模型 API，并进行提示。我们为这些模型提供了统一的任务描述和指令，提示模板和示例如表 4 所示。
- **微调**：本地部署并微调开源大模型。在两个任务上分别测试了 chinese-alpaca-2-7b-hf 和 chinese-alpaca-2-13b-hf 模型。具体来说，我们在开发集上对模型进行 LoRA (Hu et al., 2022) 微调，训练它能够根据指令正确回答 CRMUS 数据集中的问题。

³<https://www.kdocs.cn/l/cr8FBWqL8FB0>

⁴<https://open.bigmodel.cn/>

⁵<https://xinghuo.xfyun.cn/sparkapi>

⁶<https://cloud.baidu.com/product/wenxinworkshop?track=product0>

⁷<https://www.tiangong.cn/>

⁸<https://tongyi.aliyun.com/qianwen/>

提示模板:

根据故事回答下面的单项选择题，只给出答案即可：\n故事{story}: \n问题{question}: \n选项{options}: \n答案:

示例:

根据故事回答下面的单项选择题，只给出答案即可:

故事:有个女主人很勤劳，她雇了几名侍女。夜里每当公鸡一打鸣，她就叫她们起来去干活。侍女们每天日夜劳作，累得精疲力尽，她们十分痛恨那只公鸡，决定要弄死它，她们以为是那公鸡不到天亮叫醒女主人，才使她们受苦受难。然而在她们把公鸡弄死之后，反而比以前更为不幸。那女主人不知道鸡叫的时间，总是在黑夜里更早起把她们叫起来去干活。

问题:故事中的公鸡没死之前，侍女们可能什么时间被叫起来干活?

选项:A.8:00 B.4:00 C.23:00 D.15:00

答案:

Table 4: 本文实验使用的LLMs提示模板和样例

评价指标 本文使用正确率 (Accuracy) 作为两个任务的评价指标。在测试过程中，我们发现即使在提示中强制要求模型只给出答案选项，但模型总会给出形式不一的回答，有时还会附带理由。为了避免输出格式的不同而导致对模型性能的低估，本文对模型的回答采取了一系列过滤措施：如果模型生成的文本包含正确选项，且仅选择了一个候选答案，则将其视为回答正确。

4.3 实验结果与分析

实验结果 表 5 为所有模型在测试集上的表现，其中 $Acc_{cr}(\%)$ 为常识推理任务的正确率， $Acc_{mu}(\%)$ 是寓意理解任务的正确率。具体来说，GLM-4 在常识推理任务中表现最好，正确率达到 83.45%，而在寓意理解任务中（正确率 66.00%）表现略低于 ERNIE-4.0-8K（正确率 67.33%）。这表明目前最先进的 LLMs 已经可以较好的理解故事情节并根据内部的常识知识进行推理，但是在深入理解故事背后的寓意时仍存在困难。值得注意的是，提示学习实验中所有中文大模型的表现均超过 ChatGPT，这可能是由于 ChatGPT 对中文的训练不够充分，一定程度上反应了现有模型的能力更多与语言能力有关，而不是与潜在的理解有关。

此外，我们统计了寓意理解任务中不同风格选项问题的正确率，发现模型回答人工风格选项的问题时正确率（69.47%）高于机器风格（64.23%），这表明模型生成的选项更容易迷惑模型，通过 LLMs 构建数据集来考察 LLMs 是值得研究的。

类型	模型	$Acc_{cr}(\%)$	$Acc_{mu}(\%)$
	random	25.00	25.00
lora	chinese-alpaca-2-7b-hf	31.15	33.02
	chinese-alpaca-2-13b-hf	47.64	44.32
prompt	GLM-3-Turbo	68.79	56.06
	GLM-4	83.45	66.00
	星火大模型V3.5	73.52	65.63
	ERNIE-3.5-8K-0205	74.05	65.15
	ERNIE-4.0-8K	80.38	67.33
	GPT-3.5-turbo	59.93	45.93
	SkyChat-MegaVerse	70.57	53.41
	qwen-turbo	69.15	55.02

Table 5: LLMs在CRMUS两个任务上的测试结果

LLMs存在的问题 我们要求模型再次回答问题并给出理由，发现大模型在理解故事中存在以下问题:

- **多次回答不一致** 对于部分问题，大模型存在多次回答不一致的现象。如GLM-4在两个任务中的多次回答不一致率分别为28.01%和32.86%。

模型	社会常识	生物常识	时间常识	空间常识	物理常识
chinese-alpaca-2-7b-hf	70.42	67.56	69.29	71.62	60.28
chinese-alpaca-2-13b-hf	52.23	52.98	58.92	51.35	46.81
GLM-3-Turbo	27.35	35.71	37.34	33.33	28.37
GLM-4	12.91	17.26	21.16	17.12	26.95
星火大模型V3.5	22.18	29.46	36.10	28.83	27.66
ERNIE-3.5-8K-0205	22.54	25.60	34.44	33.33	21.99
ERNIE-4.0-8K	17.25	19.05	28.22	21.62	18.44
GPT-3.5-turbo	38.50	35.71	53.11	40.99	34.04
SkyChat-MegaVerse	27.70	28.87	34.85	33.33	27.66
qwen-turbo	26.17	30.06	43.98	34.23	29.79

Table 6: LLMs回答不同常识类型问题的错误率(%)

- **不忠实于上下文** 大模型推理时容易与上下文矛盾。如附录中表 9中的示例2所示，在推理小羊的年龄时，忽略了故事中“去年还未出生”这一重要条件。
- **常识能力不均衡** 大模型对于不同类型的常识的推理能力不同。我们统计了不同模型对于不同类型常识的错误率，结果如表 6所示。我们发现LLMs大多在社会常识问题上表现最佳，而在时间常识问题上表现最弱，可能的原因是LLMs的训练语料中包含大量的、显式的人物事件相关信息，而对于时间等隐式信息的直接描述较少。
- **动机理解不全面** 大模型有时不能理解故事中人物行为动机和逻辑。如附录中表 10示例2所示，LLMs没有认识到黄蜂和鹧鸪是在欺骗农夫，并不是真的要去干活。
- **深度思考不充分** 大模型分析故事整体情节的能力较弱，容易根据句子的表层线索走捷径，而不进行更深层次的思考。如附录中表 10示例3所示，LLMs只关注了“时运”，而选择了与“时运”相关的“机会就像流水，稍纵即逝，需要抓住时机。”选项，而故事整体体现的是要感恩真正帮助你的人。

IRT排名	模型	IRT得分	Acc排名	位次变化
1	GLM-4	4.35	1	-
2	ERNIE-4.0-8K	1.56	2	-
3	ERNIE-3.5-8K-0205	-0.01	3	-
4	星火大模型V3.5	-0.72	4	-
5	SkyChat-MegaVerse	-2.26	5	-
6	GLM-3-Turbo	-2.71	7	↑1
7	qwen-turbo	-2.93	6	↓1
8	GPT-3.5-turbo	-5.84	8	-
9	chinese-alpaca-2-13b-hf	-16.09	9	-
10	chinese-alpaca-2-7b-hf	-17.78	10	-

Table 7: LLMs在常识推理任务上的IRT评估结果与排名

4.4 IRT评估

本文还使用IRT对CRMUS数据集进行了评估。IRT是一种心理测量方法，该理论可以通过统计的方式建立被试者与测试题目之间的联系，来评估被试者的潜在能力(Lalor and Rodriguez, 2023)。最近，有很多研究者将NLP模型视作受试者，利用IRT来补充传统的评估指标，通过项目信息函数 (Item Response Function, IRF) 来评估模型的能力(Lalor et al., 2024)。本文使用三参数模型，如公式1所示：

$$P(X_{ij} = 1 | \theta_j) = c_i + \frac{1 - c_i}{1 + e^{(-\alpha_i(\theta_j - \beta_i))}} \quad (1)$$

其中 $X_{ij} = 1$ 代表模型 $j(j = 1, 2, 3, \dots, J)$ 在题目 $i(i = 1, 2, 3, \dots, I)$ 上的回答是正确的， θ_j 为模

IRT排名	模型	IRT得分	Acc排名	位次变化
1	ERNIE-4.0-8K	8.08	1	-
2	ERNIE-3.5-8K-0205	7.85	4	↑2
3	星火大模型V3.5	7.74	3	-
4	GLM-4	7.71	2	↓2
5	GLM-3-Turbo	6.23	5	-
6	qwen-turbo	6.16	6	-
7	SkyChat-MegaVerse	5.64	7	-
8	GPT-3.5-turbo	3.85	8	-
9	chinese-alpaca-2-13b-hf	1.40	9	-
10	chinese-alpaca-2-7b-hf	-1.53	10	-

Table 8: LLMs在寓意理解任务上的IRT评估结果与排名

型的潜在能力参数, β_i 为题目的难度参数, α_i 表示题目的区分度。其中 α_i 的值越大说明题目对模型的区分程度越高, 而区分度过低 ($\alpha_i < 0$) 表示所有模型都回答正确或都回答错误, 这样的题目对于评估模型没有意义。因此基于题目的区分度可以评估数据集的质量。

我们将文中测试的十个模型作为受试者, 通过IRT拟合工具py-irt(Lalor and Rodríguez, 2022)对其测试结果进行分析, 得出数据集中区分度小于0的题目仅占2.2%, 证明我们的数据集是较高质量的。模型在两个任务上的IRT评估结果与ACC评估结果对比如表 7和表 8所示。结果表明, 在两个任务中各模型的IRT评估结果与ACC几乎一致, 只有两个模型排名发生了变化, 这可能是有的模型答对了更多相对简单的题。在未来的工作中, 我们将通过IRT进一步分析与完善CRMUS数据集。

5 总结

本文提出了一个新的具有挑战性的故事理解数据集CRMUS, 依据人类理解故事的认知过程, 引入常识推理和寓意理解两个任务, 用于评估LLMs的故事理解能力。通过对目前多个主流LLMs进行评估, 我们发现现有的大型语言模型已经能较好地利用常识知识进行推理, 但仍然存在一定的事实性错误, 对于时间常识等隐式知识的理解与推理能力不佳。此外, LLMs对于故事角色意图和主要矛盾的理解能力不足, 难以正确理解故事背后的寓意。本文利用寓言故事评估LLMs更深层次的理解能力, 在故事风格和类型上略显单一。未来我们将使用更多类型和风格的故事进一步探索LLMs的深层语义理解和推理能力。我们希望这项工作可以用来促进LLMs深层理解能力的研究。

参考文献

- Frank B. Baker. 1985. The basics of item response theory.
- Prajwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.
- Jérôme Seymour Bruner. 1985. Actual minds, possible worlds.
- Eugene Charniak. 1972. *Toward a model of children's story comprehension*. Ph.D. thesis, Massachusetts Institute of Technology, USA.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

- Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *CoRR*, abs/2304.08177.
- Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey.
- Mark H. Dorfman and William F. Brewer. 1994. Understanding the points of fables. *Discourse Processes*, 17:105–129.
- Sayan Ghosh and Shashank Srivastava. 2022. epic: Employing proverbs in context as a benchmark for abstract language understanding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3989–4004. Association for Computational Linguistics.
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 688–714. Association for Computational Linguistics.
- Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children’s books with explicit memory representations. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5010–5015. Association for Computational Linguistics.
- Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguistics*, 6:317–328.
- Jan Kocon, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniec, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocon, Bartłomiej Kopytyra, Wiktoria Mieszczewicz-Kowszewicz, Piotr Milkowski, Marcin Oleksy, Maciej Piasecki, Lukasz Radlinski, Konrad Wojtasik, Stanislaw Wozniak, and Przemyslaw Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Inf. Fusion*, 99:101861.
- John P. Lalor and Pedro Rodríguez. 2022. py-irt: A scalable item response theory library for python. *CoRR*, abs/2203.01282.
- John Patrick Lalor and Pedro Rodriguez. 2023. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*, 35(1):5–13.
- John P. Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. 2024. Item response theory for natural language processing. In *Conference of the European Chapter of the Association for Computational Linguistics*.
- Hector J. Levesque, Ernest Davis, and L. Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.
- Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22, 06.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering.

- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Janette Pelletier and Ruth Beatty. 2015. Children’s understanding of aesop’s fables: relations to reading comprehension and theory of mind. *Frontiers in Psychology*, 6.
- Janette Pelletier and Janet Wilde Astington. 2004. Action, consciousness and theory of mind: Children’s ability to coordinate story characters’ actions and thoughts. *Early Education and Development*, 15(1):5–22.
- Joan Peskin and Janet Wilde Astington. 2004. The effects of adding metacognitive language to story texts. *Cognitive Development*, 19(2):253–273.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for squad. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics.
- Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 193–203. ACL.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4462–4472. Association for Computational Linguistics.
- Lenhart K. Schubert and Chung Hee Hwang, 2000. *Episodic logic meets Little Red Riding Hood: a comprehensive natural representation for language understanding*, page 111–174. MIT Press, Cambridge, MA, USA.
- Manfred Stede. 1996. Scott r. turner, the creative process. A computer model of storytelling and creativity. hillsdale, NJ: lawrence erlbaum, 1994. ISBN 0-8058-1576-7, £49.95, 298 pp. *Nat. Lang. Eng.*, 2(3):277–285.
- Shane Storks, Qiaozhi Gao, and Joyce Y. Chai. 2020. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge.
- Daniel J Tomasulo, James O Pawelski, et al. 2012. Happily ever after: The use of stories to promote positive interventions. *Psychology*, 3(12):1189.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.
- Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

- Yichong Xu, Jingjing Liu, Jianfeng Gao, Yelong Shen, and Xiaodong Liu. 2017. Towards human-level machine reading comprehension: Reasoning and inference with multiple strategies. *CoRR*, abs/1711.04964.
- Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: Fairytaleqa - an authentic dataset for narrative comprehension. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 447–460. Association for Computational Linguistics.
- Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3361–3367. Association for Computational Linguistics.
- Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2021. Evaluating commonsense in pre-trained language models.

附录A.模型具体实验结果

表 9和表 10分别展示了GLM-4模型对两个任务的具体回答。

数据	回答和解释	错误类型
<p>故事: 狼误吞下了一块骨头, 十分难受, 四处奔走, 寻访医生。他遇见了鹭鸶, 谈定酬金请他取出骨头, 鹭鸶把自己的头伸进狼的喉咙里, 叼出了骨头, 便向狼要定好的酬金。狼回答说: “喂, 朋友, 你能从狼嘴里平安无事地收回头来, 难道还不满足, 怎么还要讲报酬?”</p> <p>问题: 狼误吞下骨头的大小最接近以下那个选项?”,</p> <p>选项: A. 一间房屋; B. 一只蚂蚁; C. 一个篮球; D. 一只麻雀</p>	<p>答案: C. 一个篮球</p> <p>解释: 从故事中可以看出, 鹭鸶需要把头伸进狼的喉咙里才能取出那块骨头, 这表明骨头的大小应该是相对较大的, 不可能像蚂蚁那样小, 也不至于像一间房屋那样夸张。而篮球的大小比较适中, 既能够造成狼的不适, 又可以让鹭鸶通过喉咙取出, 因此最接近的选项应该是一个篮球。</p>	对事物的大小等概念认知不全导致事实性错误。
<p>故事: 一只小羊在河边喝水, 狼见到后, 便想找一个名正言顺的借口吃掉他。于是他跑到上游, 恶狠狠地说小羊把河水搅浑了, 使他喝不到清水。小羊回答说, 他仅仅站在河边喝水, 并且又在下游, 根本不可能把上游的水搅浑。狼见此计不成, 又说道: “我父亲去年被你骂过。”小羊说, 那时他还没有出生。狼对他说: “不管你怎样辩解, 反正我不会放过你。”</p> <p>问题: 文中的小羊可能多大了?”,</p> <p>选项: A. 还未出生; B. 5岁; C. 2岁; D. 8个月</p>	<p>答案: C. 2岁</p> <p>解释: 根据故事中小羊的回答: “那时他还没有出生”, 可以推断出小羊至少已经出生了。选项A “还未出生” 显然不对。选项B “5岁” 对于一个羊来说可能过大, 而且故事中并没有提到足够的信息来支持这个年龄。选项D “8个月” 虽然可能, 但通常羊的寿命和成熟期使得 “2岁” 更符合常理, 因此选项C “2岁” 是一个合理且最可能的估计。</p>	忽略了故事中的重要条件。
<p>故事: 鹰站立在岩石上, 想要去捕捉一只兔子。有人一箭射中了他, 那箭扎入他的身上, 带着鹰毛的箭翎却留在鹰的眼前。他望着翎羽说: “我自己的羽毛害死了我, 这种痛苦更难以忍受。”</p> <p>问题: 文中的人最可能从哪个方向射中了鹰?</p> <p>选项: A 鹰的后面; B 鹰的前面; C 鹰的左边; D 鹰的下边;</p>	<p>答案: A. 鹰的后面</p> <p>解释: 根据故事描述, 箭扎入了鹰的身体, 而箭翎即鹰毛留在了鹰的眼前。这意味着箭是从鹰的背后射来的, 因此箭的羽毛端 (箭翎) 会出现在鹰的视线中。如果是从其他方向射来的, 比如前面、左边或下边, 箭翎不太可能出现在鹰的眼前</p>	对方位理解错误, 缺乏相应的常识。

Table 9: GLM-4常识推理任务的错误示例

数据	回答和解释	错误类型
<p>故事: 普罗米修斯创造了人, 又在他们每人脖子上挂了两只口袋, 一只装别人的缺点, 另一只装自己的。他把那只装别人缺点的口袋挂在胸前, 另一只则挂在背后。因此人们总是能够很快地看见别人的缺点, 而自己的却总看不见。</p> <p>问题: 下列哪个选项最符合故事说明的寓意?</p> <p>选项:</p> <p>A.人们应该更全面的认识自己和他人。 B.不要因过于注他人, 而忽略了自己的优点和价值。 C.世上不存在完美的无缺事物, 要带着欣赏的眼光去看待问题。 D.人们要接纳自身不足, 勇于改进和成长。</p>	<p>答案: B.不要因过于注他人, 而忽略了自己的优点和价值。</p> <p>解释: 根据故事内容, 普罗米修斯给人挂了两只口袋, 一只装别人的缺点, 另一只装自己的, 但挂在胸前和背后的位置不同, 使得人们更容易看到别人的缺点而忽视自己的。这说明了人们往往会过度关注别人的不足, 而忽视自身的优点和价值。选项B与这一寓意最为贴切, 强调了不应忽略自己的优点和价值。其他选项虽然也是正面的寓意, 但不是这个故事所要强调的主旨。</p>	<p>对故事中的指代理解错误。</p>
<p>故事: 有一次, 黄蜂与鹧鸪因口渴难忍, 飞到农夫那里求水喝, 他们许诺将报答农夫, 鹧鸪许诺在葡萄园松土, 以便结出累累硕果; 黄蜂许诺守护葡萄园, 用毒刺驱逐偷吃的人。农夫说道: “我有两头牛, 他们从不许诺什么, 但什么活都干, 因此, 我把你们要的水给他们喝, 那不更好吗!”</p> <p>问题: 下列哪个选项最符合故事说明的寓意?</p> <p>选项:</p> <p>A.实干的牛比他们更值得信任。 B.有些人总喜欢抢占他人的劳动成果。 C.作出承诺就要去努力实现, 否则会失去别人的信任。 D.有些人随便许诺却并不打算实干。</p>	<p>答案: A. 实干的牛比他们更值得信任。</p> <p>解释: 故事中, 黄蜂和鹧鸪都许下了承诺, 但是农夫认为他们不如自己的牛, 因为牛虽然没有许下任何承诺, 但是它们实际上做了所有的工作。这个故事强调的是实际行动比空洞的承诺更为重要, 因此, 最符合故事寓意的选项是A, 即实干的牛更值得信任。其他选项虽然与故事内容有关, 但并不是故事想要强调的中心思想。</p>	<p>未能正确分析人物意图。</p>
<p>故事: 有个农夫耕地时, 发现了一块金子, 认为一定是土地女神所赐。于是, 他每日给土地的女神祭奉。命运女神来到他面前, 说: “喂, 朋友, 那块金子是我送给你的发财礼物, 你为什么把它看成是土地女神的恩惠呢? 若时运不同, 这块金子也许会落入别人的手里, 那时候你一定又要怨我命运女神了。”</p> <p>问题: 下列哪个选项最符合故事说明的寓意?</p> <p>选项:</p> <p>A.机会就像流水, 稍纵即逝, 需要抓住时机。 B.在受到帮助后, 我们要搞清楚谁才是真的恩人, 并感激他的帮助。 C.无论是谁给予了帮助, 都应该感恩。 D.成功离不开他人的支持和帮助, 我们应当认清恩人。</p>	<p>答案: A.机会就像流水, 稍纵即逝, 需要抓住时机。</p> <p>解释: 故事中命运女神提醒农夫, 好运和机会不是永久的, 它们可能会随时改变。这个寓言强调了把握时机的重要性。命运女神指出, 金子是她的礼物, 而农夫最初却错误地归功于土地女神, 这说明有时候人们可能不清楚真正帮助他们的是谁, 但更重要的是, 故事传达了机会易逝, 应当及时把握的观点。因此, 选项A“机会就像流水, 稍纵即逝, 需要抓住时机”最符合故事的寓意。其他选项虽然也涉及到感恩和认识恩人的主题, 但它们不是故事的主要教训。</p>	<p>没有理解故事的主要矛盾。</p>

Table 10: GLM-4寓意理解生成的错误示例