Overview of CCL24-Eval Task 8: Evaluation of Commonsense Reasoning and Moral Understanding in Children's Stories

Guohang Yan[1], Feihao Liang[1], Yaxin Guo[1], Hongye Tan[1,2,⋆], Ru Li[1,2], Hu Zhang[1]
[1]School of Computer and Information Technology, Shanxi University,
Taiyuan, Shanxi 030006, China
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing
of Ministry of Education,Shanxi University, Taiyuan, Shanxi 030006,China
{202222407055, 202322408029, 202112407002}@email.sxu.edu.cn
{tanhongye, liru, zhanghu}@sxu.edu.cn

Abstract

This paper provides a comprehensive review of the the CCL24-Eval Task 8: ***C****ommonsense* ***R****easoning and* ***M****oral* ***U****nderstanding in Children's* ***S****tories*(**CRMUS**). This task has designed two sub-tasks, which aim to assess the commonsense reasoning and implicit meaning comprehension capabilities of Large Language Models(LLMs). We heve received registration forms from 33 teams, 15 of which submitted final results that exceeded the baseline score. We present the results of the top 5 teams and our analysis of these results.

1 Introduction

Stories are essential reading material in education, often containing rich knowledge, vivid plots, memorable characters, and profound implicit meanings. They serve as important vehicles for the dissemination of knowledge, cultural inheritance, and value shaping. Story comprehension requires models not only to understand plots based on social, physical, and other common knowledge, but also to analyze character relationships, intentions, and behaviors, and to infer the profound meanings conveyed by the story (Tomasulo et al., 2012; Pelletier and Beatty, 2015; Dorfman and Brewer, 1994). It is suitable for evaluating the cognitive abilities of LLMs.

Therefore, We have constructed a new challenging story comprehension dataset **CRMUS** (***C****ommonsense* ***R****easoning and* ***M****oral* ***U****nderstanding in Children's* ***S****tories*), and designed two sub-tasks based on the cognitive process of human comprehension of stories. Moreover, we organized CCL24-Eval Task 8, **CRMUS**. This evaluation is divided into two tracks: (1) Track 1 allows the use of commercial LLMs through prompt learning; (2) Track 2 allows the use of open-source LLMs through fine-tuning, but the model parameters must not exceed 7 billion. In the end, we received registration forms from 33 teams, of which 15 submitted final results that exceeded the baseline we provided. We found that although LLMs already possess certain text comprehension and reasoning abilities, they still perform poorly in deep semantic comprehension and reasoning tasks that extend beyond the surface meaning of the text, such as commonsense reasoning and implicit meanings comprehension.

The task description is presented in Section 2. The dataset we constructed for this task in Section 3. In Section 4, we provide baselines for two sub-tasks. We discuss the metrics used to rank participant submissions in Section 5. In Section 6, we list participants' information and results from their submissions and provide a more in-depth discussion. We introduce the methods of excellent teams in Section 7. Finally, We conclude the paper in Section 8.

2 Task Description

We designed the following two sub-tasks to evaluate the commonsense reasoning and implicit meaning comprehension abilities of LLMs.

Commonsense Reasoning(CR) Based on a given story and associated commonsense questions, the sub-task requires selecting the correct answer. This sub-task requires the model to reason and answer questions using commonsense knowledge (usually implicit) derived from the story. The questions are in multiple-choice format, each including a question and four options.

Moral Understanding(MU) Based on a given story, select the most appropriate and relevant moral from multiple candidate options that best fits the story plot. This sub-task is a multiple-choice question with four moral options.

There are two tracks set for this evaluation, each containing the two sub-tasks mentioned above. Track 1 allows the use of ChatGPT, GPT-4, ERNIE Bot, and other commercial LLMs through prompt learning; Track 2 allows the use of open-source LLMs such as LLaMA-2 and Qwen-1.5 through fine-tuning, with model parameters not exceeding 7 billion.

3 Datasets

3.1 Dataset Construction

This task uses classic fable stories, manually collected from website[1], as raw materials and meticulously annotates them to construct the **CRMUS** dataset.

Annotation Process We annotates data through the following steps:

- **Preparation** We have developed an annotated outline that includes task definitions and examples. Based on this outline, we invited 10 graduate students with NLP-related knowledge from our team to participate in the annotation process. To enhance efficiency and quality, annotators first independently annotate the same story, then summarize the issues encountered during annotation, and refine the annotation outline accordingly.

- **Initial Annotation** For the commonsense reasoning sub-task, to ensure diversity of problems, at least two annotators are required to propose a minimum of 4 questions for each story and provide corresponding options. The questions should encompass various commonsense types, such as society, biology, time, space, and physics. Additionally, to more effectively highlight the model's limitations, annotators must identify the commonsense types relevant to each problem and provide a detailed explanation for the answers. For the moral understanding sub-task, we use the sentences of story implicit meanings as the correct answers and require annotators to provide three different implicit meanings as incorrect answers. Additionally, we request annotators to annotate two additional questions for each story using LLMs via prompt learning. Specifically, to enhance the diversity of options, annotators are required to create prompt templates, utilize various LLMs to generate multiple implicit meanings based on the story, and then filter and rewrite them to align more closely with the story's existing one. These implicit meanings are used as candidate answers for the remaining two questions.

- **Quality Control** We adopt a cross-checking approach to process the collected data. For the commonsense reasoning task, examiners are required to rate each question on a scale from 0 (unqualified) to 2 (excellent) and make modifications or add additional annotations to some questions as necessary. Finally, non-annotators will conduct secondary verification and remove any non-conforming data. For the moral understanding sub-task, inspectors carefully review each option and modify or re-annotate those that do not meet the requirements.

Finally, we adjusted the distribution of correct answers in the dataset to randomly and evenly spread them across options A, B, C, and D.

3.2 Data Samples

Each example in the development and test sets of the commonsense reasoning sub-task includes the following information: ID, title, story, question, options, answer, and commonsense type. The moral understanding sub-task includes the same information except for commonsense type. Specific examples are detailed in Figure 1:

[1] https://m.thn21.com/Article/chang/3306.html

3.3 Data Statistics

The questions and answers of the commonsense reasoning sub-task are manually annotated, while the moral understanding sub-task uses a combination of automatic generation and manual annotation. The types of common knowledge involved in the commonsense reasoning task include social, biological, temporal, spatial, and physical commonsense. The specific counts of different questions are detailed in Table 1. (Note: Some questions involve multiple types of commonsense)

Commonsense Type	Number
Social	1048
Biological	426
Temporal	308
Spatial	259
Physical	178

Table 1: Number of questions for each commonsense type

The number of question contained in each file is shown in Table 2.

Sub-task	Dev Set	Test Set	Total
Commonsense Reasoning	400	1692	2092
Moral Understanding	252	1056	1308

Table 2: Dataset size of **CRMUS**

4 Baseline

Track 1 utilizes the commercial LLM GLM-3-Turbo from Zhipu AI as the baseline model. Track 2 employs the LLaMA-2 open-source model, chinese-alpaca-2-7b-hf, fine-tuned with a Chinese corpus. For details of the baseline system, refer to the description available at website[1].

5 Evaluation Metrics

The final evaluation **Score** of the participating model is the weighted average of the accuracy of the answers in each sub-task. The specific calculation method is as follows:

$$Score = 0.4 * Acc_1 + 0.6 * Acc_2 \tag{1}$$

Specifically,
Acc_1 = the accuracy of answers for the commonsense reasoning sub-task
Acc_2 = the accuracy for the moral understanding sub-task.

6 Results and Analysis

Table 3 and Table 4 respectively present the top five official rankings of the two tracks, based primarily on the **Score**. Teams in Track 1 and Track 2 surpassed the baseline model scores. It is observed that the overall Score of Track 1 teams surpasses that of Track 2 teams, highlighting the advantage of commercial closed-source LLMs over open-source LLMs with parameters under 7B.

Based on the model proposals submitted by participating teams, it was observed that most teams employ the prompt design strategy to prompt LLMs to identify commonsense knowledge within the story,

[1] https://github.com/SXU-YaxinGuo/CRMU

Team Name	Organization	Rank	Score	CR score	CR score
Arabian Nights	South China Normal Univerity Shandong University	1	85.13	82.86	86.65
holoflow	Individual	2	79.27	86.52	74.43
AIAYN	Beijing Jiaotong University	3	77.66	86.05	72.06
XCZL	China Telecom	4	77.39	83.39	73.39
ZZU_NLP	Zhengzhou University	5	75.66	84.46	69.79
Basiline	-	-	61.15	68.79	56.06

Table 3: Track 1 results (Unit: %)

Team Name	Organization	Rank	Score	CR score	MU score
ytkj	Huazhong University of Science and Technology	1	80.82	66.96	90.06
ZZU_NLP	Zhengzhou University	2	74.38	72.87	75.38
Arabian Nights	South China Normal Univerity Shandong University	3	73.85	59.34	83.52
zyy	Shanghai University	4	71.42	70.74	71.88
XJTLU-DKE	Xi'an Jiaotong-liverpool University	5	71.22	70.8	71.5
Basiline	-	-	32.4	31.15	33.24

Table 4: Track 2 results (Unit: %)

perform commonsense reasoning, and select appropriate morals that align with the narrative. Strategies include assigning specific roles to LLMs, establishing ”task completion precautions,” defining ”task completion standards,” and similar approaches.

Certain participating teams performed fine-tuning experiments on open-source LLMs using methods like LoRA(Hu et al., 2021), selecting optimal parameters and fine-tuning modules to enhance LLMs' performance in tasks related to commonsense reasoning and implicit meaning comprehension. Overall, while these teams explored novel and interesting approaches and achieved some results, the innovativeness of these techniques was limited. They focused on activating the capabilities of LLMs for specific tasks without fundamentally enhancing the models' innate ability in commonsense reasoning and deep semantic comprehension.

7 Participant Systems

This evaluation includes two tracks. Track 1 primarily assesses the performance of different commercial models in tasks related to commonsense reasoning and implicit meaning comprehension, alongside evaluating the efficacy of various prompt strategies in enhancing model capabilities. Track 2 focuses on investigating whether open-source LLMs with limited parameter sizes can enhance their commonsense reasoning and implicit meaning comprehension abilities through pretraining and fine-tuning. Presented below are the technical approaches adopted by select outstanding teams across two tracks.

Track 1

In Track 1, ***holoflow*** proposes a straightforward yet effective two-stage prompt engineering.

• Initially, they used identical prompts to obtain responses from three advanced commercial LLMs: GPT-4, ERNIE-4, and Qwen-Max, respectively.

• Subsequently, they implemented a majority voting strategy for the LLM responses obtained in the first step. In cases of inconsistency, they queried GPT-4 for secondary confirmation using a slightly modified prompt compared to the first step, narrowing down the options to those returned initially. The choice confirmed in this secondary phase was selected as the final submission result.

The experimental results demonstrate that their method achieved the final Score of 79.27, placing first in the closed dataset of Track 1 among 10 submitted results, thereby confirming its effectiveness. The results further validate the efficacy of the prompt-based approach in addressing the CRMUS task.

Track 2

ZZU_NLP secured first place in the closed data of Track 2. Their approach primarily involved designing effective prompt templates, fine-tuning LoRA parameters, and utilizing data augmentation techniques.

In the instruction fine-tuning stage, they chose two LLMs mainly in Chinese, namely Qwen1.5-Chat-7B (Bai et al., 2023) and Internlm2-Chat-7B (Cai et al., 2024). Among them, Internlm2-Chat-7B is the main fine-tuning model, and Qwen1.5-Chat-7B is the auxiliary model to verify the optimal LoRA parameters. By testing the combination of different LoRA parameters and fine-tuning modules, it was ultimately determined that two sets of parameters can provide the optimal Acc indicators for CR and MU, respectively.

In the process of conducting commonsense reasoning on the development set, they found that the model performed poorly in terms of temporal, spatial, and physical knowledge, and speculated that this may be due to the small amount of data for several commonsense types. Therefore, they used data augmentation methods to address the issue of uneven distribution of different commonsense types in the CRMUS dataset. They created over 200 commonsense reasoning data using ChatGPT, and then manually reviewed and screened 137 high-quality data. The data was then expanded to the development set for fine-tuning, resulting in an improvement in the accuracy.

8 Conclusion and Future Work

This paper presents an overview of the CCL24-Eval Task 8, i.e., *Commonsense **R**easoning and **M**oral **U**nderstanding in Children's **S**tories*(**CRMUS**). This evaluation is conducted using our meticulously annotated CRMUS dataset. These tasks are designed to assess LLMs' ability to understand and reason about commonsense knowledge in stories, as well as their capacity to capture the deep semantics and implicit meanings within the stories. We received a total of 33 completed registration forms, of which 15 teams submitted the final results that exceeded the baseline we provided. Additionally, we offer a comprehensive analysis and summary of the methodologies employed by the participants, which will inform and guide future research in the field of natural language processing.

Finally, we believe that this evaluation remains challenging for LLMs, primarily due to the models' insufficient semantic understanding and reasoning abilities. In the future, we will continue to explore and enhance the CRMUS dataset, aiming to further improve its scale and quality. We also aim to explore additional forms of commonsense and reasoning Q&A, as well as moral examination methods, to better evaluate the commonsense reasoning and deep semantic understanding abilities of LLMs.

9 Acknowledgements

We would like to acknowledge the contributions of the other members of Hongye Tan's team and thank them for annotating the CRMUS corpus. We appreciate the support from Scientific and Technological Innovation 2030 - ”New Generation Artificial Intelligence”(2020AAA0106100) and the National Natural Science Foundation of China(62076155). Lastly, we extend our appreciation to the CCL Evaluation Committee for their support.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

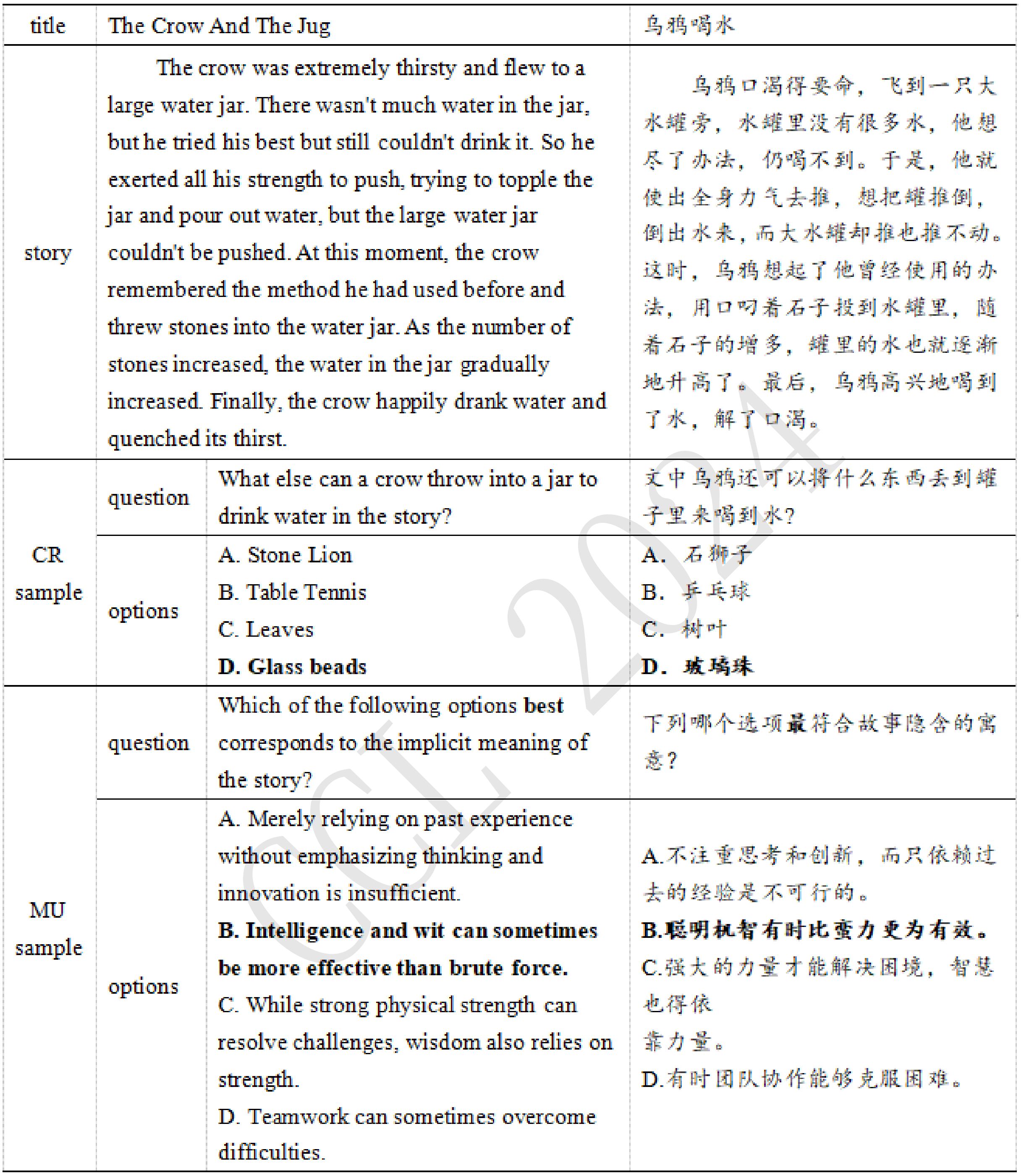
Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Marcy H Dorfman and William F Brewer. 1994. Understanding the points of fables. *Discourse Processes*, 17(1):105–129.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Janette Pelletier and Ruth Beatty. 2015. Children's understanding of aesop's fables: relations to reading comprehension and theory of mind. *Frontiers in Psychology*, 6:146239.

Daniel J Tomasulo, James O Pawelski, et al. 2012. Happily ever after: The use of stories to promote positive interventions. *Psychology*, 3(12):1189.



title	The Crow And The Jug		乌鸦喝水
story	The crow was extremely thirsty and flew to a large water jar. There wasn't much water in the jar, but he tried his best but still couldn't drink it. So he exerted all his strength to push, trying to topple the jar and pour out water, but the large water jar couldn't be pushed. At this moment, the crow remembered the method he had used before and threw stones into the water jar. As the number of stones increased, the water in the jar gradually increased. Finally, the crow happily drank water and quenched its thirst.		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。
CR sample	question	What else can a crow throw into a jar to drink water in the story?	文中乌鸦还可以将什么东西丢到罐子里来喝到水?
	options	A. Stone Lion B. Table Tennis C. Leaves **D. Glass beads**	A. 石狮子 B. 乒乓球 C. 树叶 **D. 玻璃珠**
MU sample	question	Which of the following options best corresponds to the implicit meaning of the story?	下列哪个选项最符合故事隐含的寓意?
	options	A. Merely relying on past experience without emphasizing thinking and innovation is insufficient. **B. Intelligence and wit can sometimes be more effective than brute force.** C. While strong physical strength can resolve challenges, wisdom also relies on strength. D. Teamwork can sometimes overcome difficulties.	A.不注重思考和创新，而只依赖过去的经验是不可行的。 **B.聪明机智有时比蛮力更为有效。** 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。

Figure 1: Samples of **CRMUS**