System Report for CCL24-Eval Task 8: Exploring Faithful and Informative Commonsense Reasoning and Moral Understanding in Children's Stories

Zimu Wang[1,3], Yuqi Wang[1,3], Nijia Han[1], Qi Chen[2], Haiyang Zhang[1], Yushan Pan[1], Qiufeng Wang[1], Wei Wang[1†]

[1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University
[2]School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University
[3]Department of Computer Science, University of Liverpool
{Zimu.Wang19,Yuqi.Wang17,Nijia.Han23}@student.xjtlu.edu.cn
{Qi.Chen02,Haiyang.Zhang,Yushan.Pan}@xjtlu.edu.cn
{Qiufeng.Wang,Wei.Wang03}@xjtlu.edu.cn

Abstract

Commonsense reasoning and moral understanding are crucial tasks in artificial intelligence (AI) and natural language processing (NLP). However, existing research often falls short in terms of faithfulness and informativeness during the reasoning process. We propose a novel framework for performing commonsense reasoning and moral understanding using large language models (LLMs), involving constructing guided prompts by incorporating relevant knowledge for commonsense reasoning and extracting facts from stories for moral understanding. We conduct extensive experiments on the Commonsense Reasoning and Moral Understanding in Children's Stories (CRMUS) dataset with widely recognised LLMs under both zero-shot and fine-tuning settings, demonstrating the effectiveness of our proposed method. Furthermore, we analyse the adaptability of different LLMs in extracting facts for moral understanding performance.

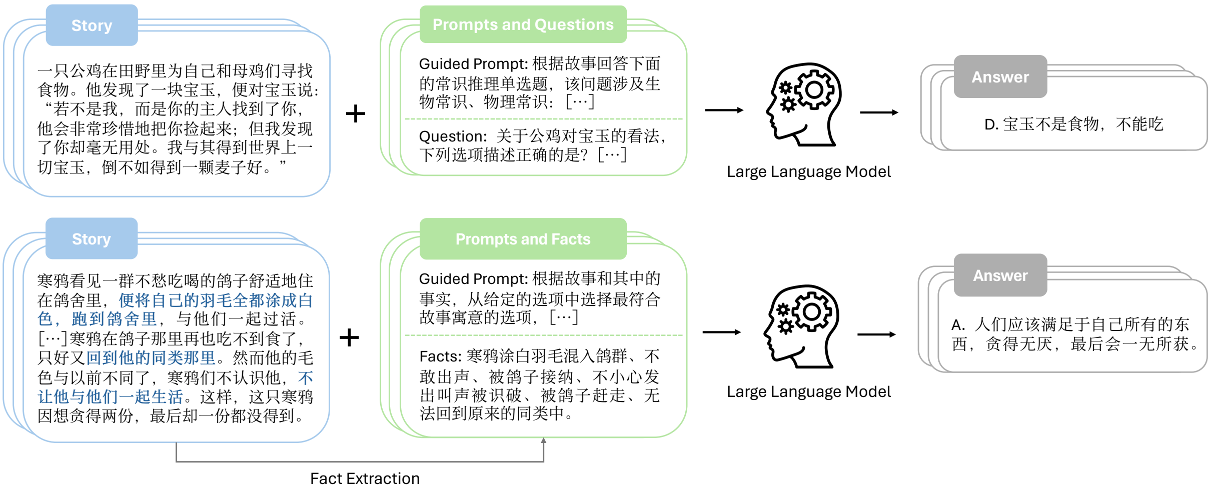
1 Introduction

Proficiency in acquiring reasoning abilities, such as arithmetic, commonsense, and symbolic reasoning, plays an essential role in artificial intelligence (AI) and natural language processing (NLP) (Wang et al., 2023). Unlike arithmetic reasoning, which involves manipulating numbers, and symbolic reasoning, which involves interpreting logic and symbols, commonsense reasoning encompasses counterfactual, abductive, and monotonic reasoning (Ashida and Sugawara, 2022). It is crucial for language understanding and enables humans to navigate daily situations seamlessly (Sap et al., 2020). Applications of commonsense reasoning include text classification (Wang et al., 2019), question answering (Mihaylov and Frank, 2018), and natural language generation (Chen et al., 2019).

Commonsense reasoning is typically framed as a multiple-choice format, where the goal is to determine the plausibility of candidate answers. This approach mirrors how people often consider several plausible choices based on a given situation and their thought processes (Figure 1) (Ashida and Sugawara, 2022). Previous research has focused primarily on utilising pre-trained language models (PLMs) and conducting the reasoning process based on factual time and space information (Talmor et al., 2019), human behaviours (Zhang and Choi, 2021; Emelin et al., 2021), and story texts (Ashida and Sugawara, 2022). Recently, with the development of large language models (LLMs) that have shown remarkable performance in a range of natural language understanding and reasoning tasks (Peng et al., 2023; Na et al., 2024), they have also been leveraged to enhance the reasoning process (Wang and Zhao, 2023; Bian et al., 2024; Krause and Stolzenburg, 2024). Similar to commonsense reasoning, moral understanding is the process of comprehending the moral of the given context from multiple candidates.

Despite the progress made, existing research in commonsense reasoning and moral understanding faces significant challenges, particularly regarding the *unfaithfulness* and *uninformativeness* of the reasoning process. Current LLM-based methods primarily follow the in-context learning (ICL) paradigm

[†]Corresponding author.



一只公鸡在田野里为自己和母鸡们寻找食物。他发现了一块宝玉，便对宝玉说："若不是我，而是你的主人找到了你，他会非常珍惜地把你捡起来；但我发现了你却毫无用处。我与其得到世界上一切宝玉，倒不如得到一颗麦子好。"

Guided Prompt: 根据故事回答下面的常识推理单选题，该问题涉及生物常识、物理常识：[···]

Question: 关于公鸡对宝玉的看法，下列选项描述正确的是？[···]

D. 宝玉不是食物，不能吃

寒鸦看见一群不愁吃喝的鸽子舒适地住在鸽舍里，便将自己的羽毛全都涂成白色，跑到鸽舍里，与他们一起过活。[···]寒鸦在鸽子那里再也吃不到食了，只好又回到他的同类那里。然而他的毛色与以前不同了，寒鸦们不认识他，不让他与他们一起生活。这样，这只寒鸦因想贪得两份，最后却一份都没得到。

Guided Prompt: 根据故事和其中的事实，从给定的选项中选择最符合故事寓意的选项，[···]

Facts: 寒鸦涂白羽毛混入鸽群、不敢出声、被鸽子接纳、不小心发出叫声被识破、被鸽子赶走、无法回到原来的同类中。

A. 人们应该满足于自己所有的东西，贪得无厌，最后会一无所获。

Figure 1: Overall framework of our proposed method to conduct *faithful* and *informative* commonsense reasoning and moral understanding.

(Brown et al., 2020), which conditions the models on a natural language instruction and/or a few demonstrations (Qiao et al., 2023; Wang and Zhao, 2023). However, commonsense reasoning usually involves various types of knowledge applied in different stories and questions, such as temporal, spatial, biological, physical, and social knowledge; and the facts described in the stories, such as their plots, characters, and events, are highly related to the moral that the authors intend to impart. Though effective, the aforementioned information is usually disregarded in previous research.

Motivated by this phenomenon, we design a novel framework to conduct commonsense reasoning and moral understanding, making the reasoning process more *faithful* and *informative*. Unlike the previous work that utilises external knowledge, such as knowledge bases (Mitra et al., 2020; Bian et al., 2021), search engines (Talmor et al., 2021), and the knowledge generated by LLMs (Liu et al., 2022), we construct guided prompts for the two tasks, as shown in Figure 1. For commonsense reasoning, we incorporate the related knowledge concerning the story and the question into the prompt, and for moral understanding, we extract the facts contained in the stories as additional supervision.

We conduct extensive experiments on the Commonsense Reasoning and Moral Understanding in Children's Stories (CRMUS) dataset with the widely recognised LLMs: GLM-3, GLM-4, Moonshot, and Yi-34B for zero-shot prompting and ChatGLM3-6B (Zeng et al., 2023), InternLM2-7B (Cai et al., 2024), Qwen1.5-7B (Bai et al., 2023), and Yi-6B (01.AI et al., 2024) for fine-tuning. Experimental results demonstrate the effectiveness of guided prompts for both commonsense reasoning and moral understanding. Among the models, GLM-4 and InternLM2-7B achieve the best performance in zero-shot and fine-tuning settings, respectively. Furthermore, we conduct additional experiments to analyse the adaptability of extracted facts from different LLMs for moral understanding performance.

The key contributions of this work are summarised as follows:

- We propose a novel framework for commonsense reasoning and moral understanding using LLMs, making the process becomes *faithful* and *informative*.
- We perform extensive experiments on widely recognised LLMs to demonstrate the effectiveness of the proposed method.
- We conduct additional experiments on moral understanding and analyse the adaptability of extracted facts on different LLMs on this task.

2 Background

Commonsense reasoning has received considerable attention over the past decade. Recent research highlights the substantial improvements in this area by incorporating additional knowledge, broadly falling

End-to-End Prompt for Commonsense Reasoning:

根据故事回答下面的单项选择题，只给出答案即可：
[Translation: Please answer the following multiple-choice question based on the story, providing only the answer:]
故事[Story]：{story_text}
问题[Question]：{question}
选项[Options]：{options}
答案[Answer]：

End-to-End Prompt for Moral Understanding:
根据故事从给定选项中选择最符合故事说明的寓意的选项，只给出答案即可： [Translation: Please select the option that best matches the moral from the given choices based on the story, providing only the answer:] 故事[Story]：{story_text} 选项[Options]：{options} 答案[Answer]：

Table 1: End-to-end prompts for the commonsense reasoning and moral understanding tasks, in which {story_text}, {question}, and {options} refer to the story context, the question, and the candidate options, respectively.

into two categories. The first approach involves augmenting the task with external knowledge graphs, such as ConceptNet (Speer et al., 2017) and FreeBase (Bollacker et al., 2008). Noteworthy methods like KAGNet (Lin et al., 2019) and GRF (Ji et al., 2020) operate by reasoning over the links connecting different entities and relationships within the knowledge graphs. However, it is worth mentioning that commonsense knowledge includes a wide range of facts and scenarios that exceed the capacity of a single knowledge graph with a specific schema (Yu et al., 2022).

The second approach focuses on leveraging the internal knowledge of LLMs, which are trained on massive datasets to generate task-specific knowledge. For instance, Zhou et al. (2021) employs self-talk procedures (Shwartz et al., 2020) and inquiry-based discovery learning to generate implicit commonsense before response generation. Similarly, Qin et al. (2020) and Zhao et al. (2023) generate plausible explanations for commonsense reasoning by incorporating future context in decoding algorithms and using posterior regularisation for constraint enforcement. Additionally, Paranjape et al. (2021) prompts GPT2-XL (Radford et al., 2019) for inference using generated contrastive explanations. Furthermore, studies like those discussed by Liu et al. (2022) and Cao and Jiang (2024) emphasise the improvements in commonsense reasoning, even in zero-shot scenarios, through the incorporation of LLM-generated knowledge. Unlike the previous work, we construct guided prompts with knowledge highly related to the stories, which are more faithful to the story contexts and have higher generalisability.

3 System Overview

Following the overall framework illustrated in Figure 1, in this section, we describe the design of the system to conduct *faithful* and *informative* commonsense reasoning and moral understanding in detail.

3.1 Problem Definition

We define our commonsense reasoning and moral understanding tasks as follows. Given a story context $S = \{w_1, w_2, \ldots, w_M\}$ (M is the number of words within the story), a question q, and a list of candidate answers $A = \{a_1, a_2, \ldots, a_N\}$ (N is the number of candidate answers), the aim of the tasks is to select the answer $a^* \in A$ that matches the question q with respect to the story S most. To make the reasoning process *faithful* and *informative*, we add the related knowledge K, a subset of a pre-defined list $\mathcal{K}$, containing the knowledge related to the story and the question (e.g. temporal, spatial, and social knowledge) for commonsense reasoning and the facts that happened in the stories $F = \{f_1, f_2, \ldots, f_P\}$ (P is the number of facts in the story) for moral understanding into the prompt, in which the list of facts F is extracted by an LLM, which could be GLM-4, Moonshot, and Yi-34B.

PROMPT FOR FACT EXTRACTION:
根据故事内容，抽取故事中与寓意有关的事实，只给出答案即可，以顿号分隔： [Translation: Based on the story content, extract the facts relevant to the moral of the story, providing only the answers separated by serial commas:] 故事[Story]：`{story_text}` 答案[Answer]：

Table 2: Fact extraction prompt for the moral understanding task, in which `{story_text}` refers to the story context.

GUIDED PROMPT FOR COMMONSENSE REASONING:
根据故事回答下面的常识推理单项选择题，只给出答案即可，该问题涉及`{reasoning_type}`： [Translation: Answer the following commonsense reasoning multiple-choice question based on the story, providing only the answer. The question involves `{reasoning_type}`:] 故事[Story]：`{story_text}` 问题[Question]：`{question}` 选项[Options]：`{options}` 答案[Answer]：
GUIDED PROMPT FOR MORAL UNDERSTANDING:
根据故事和其中的事实从给定选项中选择最符合故事说明的寓意的选项，只给出答案即可： [Translation: Based on the story and its facts, select the option that best matches the moral of the story from the given choices, providing only the answer:] 故事[Story]：`{story_text}` 事实[Facts]：`{extracted_facts}` 选项[Options]：`{options}` 答案[Answer]：

Table 3: Guided prompts for the commonsense reasoning and moral understanding tasks, in which `{story_text}`, `{question}`, `{reasoning_type}`, `{extracted_facts}`, and `{options}` refer to the story context, the question, the related knowledge, the facts extracted by the LLMs, and the candidate options, respectively.

3.2 End-to-End Prompt Construction

We first construct a prompt to conduct end-to-end commonsense reasoning and moral understanding and consider it as our baseline, as shown in Table 1. It includes an instruction for the target task, a story context, a question, and four candidate options. Because all the questions for moral understanding are the same (i.e. "*Which of the following options best matches the moral of the story?*"), we incorporate the requirement of performing moral understanding into the instruction for that task. The prompt ends with the word "*Answer:*", which asks the language models to answer the question.

3.3 Guided Prompts Construction

We design separate guided prompts for commonsense reasoning and moral understanding with respect to the task characteristics, which are explained as follows:

Commonsense Reasoning Commonsense reasoning usually includes multiple types of knowledge, such as temporal, spatial, biological, physical, and social knowledge, to support the reasoning process. Therefore, we incorporate the relevant knowledge into the prompt, which is obtained from the golden annotations, and we ask the model to consider this information when making accurate predictions, as shown in Table 3.

Moral Understanding For moral understanding, we consider the facts in the stories—such as their plots, characters, and events—as additional supervision, since these elements are often connected to the moral that the author conveys. By analysing the essential facts within the story, LLMs can gain a deeper understanding of the message the author intends to impart. The construction of guided prompts for moral understanding includes two procedures: fact extraction and guided prompt construction.

Model	End-to-End Prompt		Guided Prompt	
	Accuracy (CR)	Accuracy (MU)	Accuracy (CR)	Accuracy (MU)
GLM-3	70.45	56.91	72.10 (↑ 1.65)	56.91 (↑ 0.00)
GLM-4	**83.92**	**65.91**	**84.28** (↑ 0.36)	**66.38** (↑ 0.47)
Moonshot	77.01	65.44	78.66 (↑ 1.65)	65.06 (↓ 0.38)
Yi-34B	74.65	59.38	74.53 (↑ 0.12)	59.09 (↓ 0.29)

Table 4: Performance of LLMs on commonsense reasoning (CR) and moral understanding (MU) under the *zero-shot* setting, in which the best and the second-best results are highlighted in **bold** and underlined, respectively.

Model	End-to-End Prompt		Guided Prompt	
	Accuracy (CR)	Accuracy (MU)	Accuracy (CR)	Accuracy (MU)
ChatGLM3-6B	53.72	61.08	53.49 (↓ 0.23)	62.03 (↑ 0.95)
InternLM2-7B	**70.63**	**71.50**	**70.80** (↑ 0.17)	**70.83** (↓ 0.67)
Qwen1.5-7B	66.96	69.60	66.96 (↑ 0.00)	69.60 (↑ 0.00)
Yi-6B	67.85	65.25	67.02 (↓ 0.83)	63.16 (↓ 2.09)

Table 5: Performance of LLMs on commonsense reasoning (CR) and moral understanding (MU) under the *fine-tuning* setting, in which the best and the second-best results are highlighted in **bold** and underlined, respectively.

Initially, we extract the factual details from the stories by prompting LLMs, the prompt for which is shown in Table 2. Similar to the prompt for reasoning, the prompt for fact extraction includes an instruction and a story, and it ends with the word "*Answer:*" to ask the models to answer the question. Once the facts are extracted, they are incorporated into the end-to-end prompt as additional supervision, and the guided prompt is used to ask the models to conduct the moral understanding process, as depicted in Table 3.

4 Experiments

4.1 Experimental Setup

We conducted experiments on widely recognised LLMs under both zero-shot prompting and fine-tuning settings. Under the zero-shot setting, we experimented on GLM-3[0] (`glm-3-turbo`), GLM-4 (`glm-4`), Moonshot[1] (`moonshot-v1-8k`), and Yi-34B[2] (`yi-34b`), which were accessed by calling their official APIs. During the reasoning process, we set the temperature as 0 to stabilise the output of the models. We also fine-tuned four LLMs, including ChatGLM3-6B (Zeng et al., 2023), InternLM2-6B (Cai et al., 2024), Qwen1.5-7B (Bai et al., 2023), and Yi-6B (01.AI et al., 2024), which were accessed from the Hugging Face[3] repository. During the fine-tuning process, we set the number of epochs as 20, the learning rate as $5e-5$, the batch size as 2, and the gradient accumulation steps as 8, and we adopted LLaMA-Factory (Zheng et al., 2024) for efficient fine-tuning with the LoRA strategy (Hu et al., 2022). All experiments were conducted on a single NVIDIA A10 Tensor Core GPU.

4.2 Experimental Results

Main Results Tables 4 and 5 present the main experimental results of commonsense reasoning and moral understanding under zero-shot and fine-tuning settings, in which we utilised GLM-4 as the model

[0] https://open.bigmodel.cn/

[1] https://platform.moonshot.cn/

[2] https://platform.lingyiwanwu.com/

[3] https://huggingface.co/

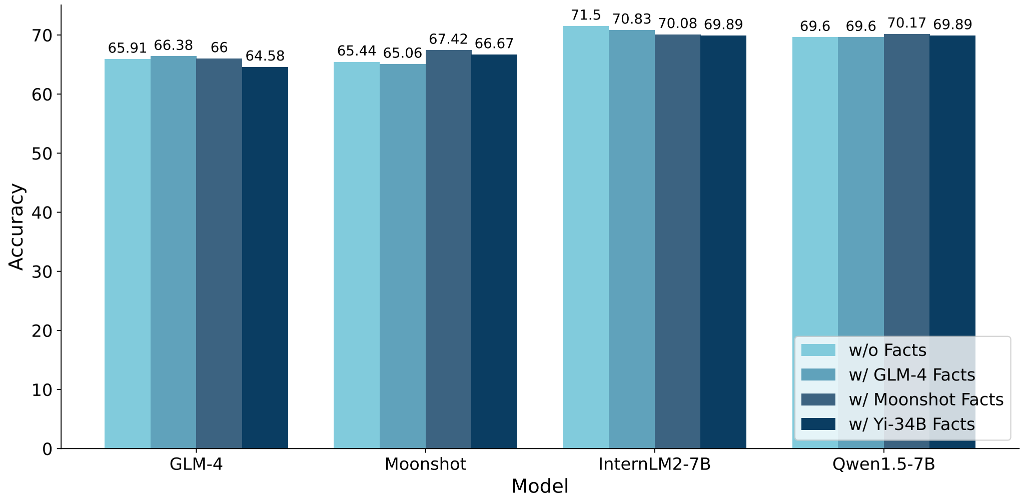


Figure 2: Effects of facts extracted from different LLMs in the moral understanding task.

to extract the facts that occur in the stories. Among all experiments, GLM-4 and IntenLM2-7B achieved the best performance under zero-shot and fine-tuning settings, regardless of the utilisation of the guided prompts. After fine-tuning the smaller-sized LLMs, such as InternLM-7B and Qwen1.5-7B, the models could perform comparable to or even better performance than the larger-scale LLMs, such as GLM-4 and Moonshot, in moral understanding; however, there was still room for improvement in terms of commonsense reasoning.

We also observed the effectiveness of guided prompts in commonsense reasoning and moral understanding. Regarding commonsense reasoning, the use of guided prompts led to performance improvements across nearly all models, indicating that incorporating knowledge successfully enhanced the commonsense reasoning process. However, for moral understanding, guided prompts proved beneficial specifically for GLM-4; thus, further investigations are needed to assess the generalisability of the method in moral understanding.

Effects of the Extracted Facts The main experimental results of guided prompts for moral understanding were remarkable—typically, the facts extracted from the stories should closely align with the moral intended by the authors. It was observed that these facts, extracted by GLM-4, significantly benefited GLM models. This correlation underscores the importance of aligning the models utilised for fact extraction with those used for moral understanding prediction. To further investigate the relationship between fact extraction and moral understanding, we conducted additional experiments using GLM-4, Moonshot, and Yi-34B for fact extraction, and subsequently employing GLM-4, Moonshot, InternLM2-7B, and Qwen1.5-7B for predicting moral understanding.

We presented the experimental results in Figure 2, which substantiated our earlier hypothesis. Generally, using the same LLM for both fact extraction and moral understanding prediction (e.g. GLM-4) contributed significantly to model performance. Interestingly, despite Moonshot initially performed worse than GLM-4 in our main experiments, its performance improved notably when employed it for fact extraction from the stories. This underscores the efficacy of guided prompts in enhancing moral understanding. The impact of extracted facts varied for InternLM2-7B and Qwen1.5-7B, highlighting how different LLMs affect moral understanding performance based on the extracted facts.

5 Conclusion and Future Work

We introduced a novel framework for commonsense reasoning and moral understanding with LLMs, aiming to ensure a *faithful* and *informative* reasoning process. Specifically, we developed guided prompts that integrate relevant knowledge for commonsense reasoning and the facts that happened in the stories extracted by LLMs for moral understanding. We conducted extensive experiments on the CRMUS dataset with widely recognised LLMs under both zero-shot and fine-tuning settings and demonstrated the effectiveness of our proposed method. We further analysed the adaptability of extracted facts of different LLMs on moral understanding. In the future, we will make the guided prompts more diverse, incorporating more useful features to guide the reasoning process. We will also transfer our method on more reasoning tasks to test the generalisability of our proposed method.

Acknowledgements

This research is funded by the Postgraduate Research Scholarship (PGRS) at Xi'an Jiaotong-Liverpool University, contract number FOSA2212008, and partially supported by 2022 Jiangsu Science and Technology Programme (General Programme), contract number BK20221260.

References

01.AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Mana Ashida and Saku Sugawara. 2022. Possible stories: Evaluating situated commonsense reasoning under multiple possible scenarios. In *Proceedings of COLING*, pages 3606–3630, October.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Ning Bian, Xianpei Han, Bo Chen, and Le Sun. 2021. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation. *Proceedings of AAAI*, 35(14):12574–12582, May.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of SIGMOD*, page 1247–1250.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*, volume 33, pages 1877–1901.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report.

Rui Cao and Jing Jiang. 2024. Knowledge generation for zero-shot knowledge-based VQA. In Yvette Graham and Matthew Purver, editors, *Findings of EACL*, pages 533–549, March.

Jiaao Chen, Jianshu Chen, and Zhou Yu. 2019. Incorporating structured commonsense knowledge in story completion. *Proceedings of AAAI*, 33(01):6244–6251, Jul.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of EMNLP*, pages 698–718, November.

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 *Proceedings of ICLR*.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of EMNLP*, pages 725–736, November.

Stefanie Krause and Frieder Stolzenburg. 2024. Commonsense reasoning and explainable artificial intelligence using large language models. In *Artificial Intelligence. ECAI 2023 International Workshops*, pages 302–319.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of EMNLP-IJCNLP*, pages 2829–2839, November.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of ACL*, pages 3154–3169, May.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of ACL*, pages 821–832, July.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2020. How additional knowledge can improve natural language commonsense question answering?

Hongbin Na, Zimu Wang, Mieradilijiang Maimaiti, Tong Chen, Wei Wang, Tao Shen, and Ling Chen. 2024. Rethinking human-like translation strategy: Integrating drift-diffusion model with large language models for machine translation.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of ACL-IJCNLP*, pages 4179–4192, August.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of ACL*, pages 5368–5393, July.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of EMNLP*, pages 794–805, November.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of ACL (Tutorial)*, pages 27–33, July.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of EMNLP*, pages 4615–4629, November.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of AAAI*, 31(1), Feb.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pages 4149–4158, June.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of ai through gamification. In *Proceedings of the NeurIPS (Datasets and Benchmarks)*, volume 1.

Yuqing Wang and Yun Zhao. 2023. Gemini in reasoning: Unveiling commonsense in multimodal large language models.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019. Improving natural language inference using external knowledge in the science questions domain. *Proceedings of AAAI*, 33(01):7208–7215, Jul.

Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, Suparna De, and Amir Hussain. 2023. Fusing external knowledge resources for natural language understanding techniques: A survey. *Information Fusion*, 92:190–204.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of EMNLP*, pages 4364–4377, December.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *Proceedings of ICLR*.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of EMNLP*, pages 7371–7387, November.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. In *Proceedings of ACL*, pages 14883–14896, July.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models.

Pei Zhou, Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Think before you speak: Learning to generate implicit knowledge for response generation by self-talk. In *Proceedings of NLP4ConvAI*, pages 251–253, November.