System Report for CCL24-Eval Task 8: A Two-stage Prompt-Based Strategy for CRMUS Track 1

Mosha Chen
Hangzhou Guangran Digital Technology Co., Ltd, Hangzhou, China
chenmosha@holoflow.cn

Abstract

Large Language Model (LLM) has sparked a new trend in Natural Language Processing, and an increasing number of researchers have recognized the potential of using LLM to unify diverse NLP tasks into a text-generative manner. To explore the potential of LLM for the children's stories domain, CCL2024 has released the **C**ommonsense **R**easoning and **M**oral **U**nderstanding in Children's **S**tories (CRMUS) task. This paper presents a straightforward yet effective two-stage prompt-based strategy for the CRMUS Track 1. In the initial stage, we use the same prompt to obtain responses from GPT-4, ERNIE-4, and Qwen-Max. In the subsequent stage, we implement a voting mechanism based on the results from the first stage. For records with inconsistent outcomes, we query GPT-4 for secondary confirmation to determine the final result. Experimental results indicate that our method achieved an average score of 79.27, securing first place in the closed domain among ten participating teams, thereby demonstrating the effectiveness of our approach.

1 Introduction

With the popularization of ChatGPT, Large Language Model (LLM) has motivated an increasing trend in both industry and academia; an increasing number of researchers are exploring the potential of LLMs (Steven et al., 2023; Chen and Si, 2024; Zhang et al., 2023; Wu et al., 2023; He et al., 2023) across various domains, leading to the proposal of new paradigms for NLP tasks. Following the trend, the **C**ommonsense **R**easoning and **M**oral **U**nderstanding in Children's **S**tories (CRMUS) evaluation task[0] is introduced in CCL2024, which aims to evaluate Chinese pre-trained language models and large language models from multiple perspectives in terms of commonsense reasoning and moral understanding on the children's education domain. In order to investigate different techniques in the field of LLMs, the CRMUS task provides two tracks: prompt-based and fine-tuning of LLM parameters.

The purpose of the prompt-based track in the CRMUS task is to assess LLM's potential in story commonsense reasoning and moral understanding. The types of commonsense involved in the commonsense reasoning task cover a wide range of aspects: temporal commonsense, spatial commonsense, biological commonsense, physical commonsense, and social commonsense. Given the comprehensive reasoning capabilities of LLMs, we hypothesized that a prompt-based approach would effectively fit the CRMUS task. To this end, we propose a straightforward yet effective two-stage prompt engineering pipeline:

- In the first stage, we use a uniform prompt to obtain responses from three advanced commercial LLMs: GPT-4, ERNIE-4, and Qwen-Max.
- In the second stage, we adopted a majority voting strategy for the LLM responses from the first step. For the inconsistent results, we query GPT-4 for a secondary confirmation, with a slightly different prompt from the first step, which narrows down the range of options using only the options returned from the first step. This secondary confirmed choice is chosen as the final submission result.

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

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 311-319, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Our experimental results demonstrate that our method achieved an average score of 79.27, ranking first in the closed domain among ten participating teams. This confirms the effectiveness of our approach. Furthermore, our method leverages the strengths of multiple LLMs and a novel two-stage voting mechanism, offering a robust solution for commonsense reasoning and moral understanding tasks without requiring extensive fine-tuning. This innovation lies in the efficient combination of multiple models' strengths and the strategic use of prompt engineering to enhance performance in the CRMUS task.

2 Related Word

2.1 Prompt Engineering

Recent research has highlighted that the use of prompts can substantially enhance the performance of pre-trained language models in various downstream applications(Brown et al., 2020). This has sparked significant interest in the effective construction of prompts. Manual prompt creation, however, is labor-intensive and not always feasible. To address this, Shin(2020) proposed a discrete word space search algorithm that leverages downstream application training data. While this approach outperforms manual prompt design, it is limited by the weak expressive capability of discrete prompts, resulting in only modest improvements in downstream tasks. To overcome these limitations, some researchers have introduced prompt-tuning methods(Lester et al., 2021) that optimize continuous prompt vectors through gradient backpropagation. These methods have demonstrated considerable performance gains; however, since parameter fine-tuning is not allowed in the prompt-based track, we adopted the approach of manually designing prompts. Unlike these methods, our approach not only involves manually designed prompts but also integrates responses from multiple LLMs through a novel voting mechanism, enhancing the robustness and accuracy of the final results.

2.2 In-context Learning

The release of GPT-3(Radford et al., 2019), OpenAI's former state-of-the-art large language model, has significantly drawn the research community's attention to a novel area: In-Context Learning (ICL). The authors demonstrated that GPT-3, a self-supervised pre-trained model, can effectively perform new tasks without prior specific training, simply by giving a manually designed prompt that includes an optional task description and a few example demonstrations. This groundbreaking capability has spurred extensive research exploring various facets of ICL. The performance of ICL has been demonstrated to be highly sensitive to the selection of demonstration examples(Li et al., 2023). To address this issue, Rubin(2022) proposed methods for learning to retrieve suitable demonstration examples. Li(2023) introduced a series of techniques to enhance demonstration selection performance. Levy(2023) focused on selecting diverse demonstrations to improve in-context compositional generalization. Furthermore, Qin(**?**) developed an iterative approach that selects diverse examples yet closely correlates with the test sample for ICL demonstrations. Our work differs from ICL methods as we do not rely on example demonstrations for task performance. Instead, we focus on leveraging the diverse capabilities of multiple LLMs and a structured voting strategy, which we believe offers a more direct and reliable approach to addressing the CRMUS task.

3 Our Methods

Due to the comprehensive reasoning capabilities of LLM, we believe the prompt-based method could fit the CRMUS task well. We propose a simple yet effective two-stage prompt-based pipeline for the evaluation tasks. The overall processing pipeline, illustrated in Figure 1, applies to the Commonsense Reasoning (CR) subtask. The Moral Understanding (MR) subtask follows the same pipeline except for the prompt.

3.1 First Stage: Query LLMs

In the first stage, we select three of the most advanced commercial LLMs—GPT-4, ERNIE-4, and Qwen-Max—as our testbeds. The same prompt is applied to each LLM.

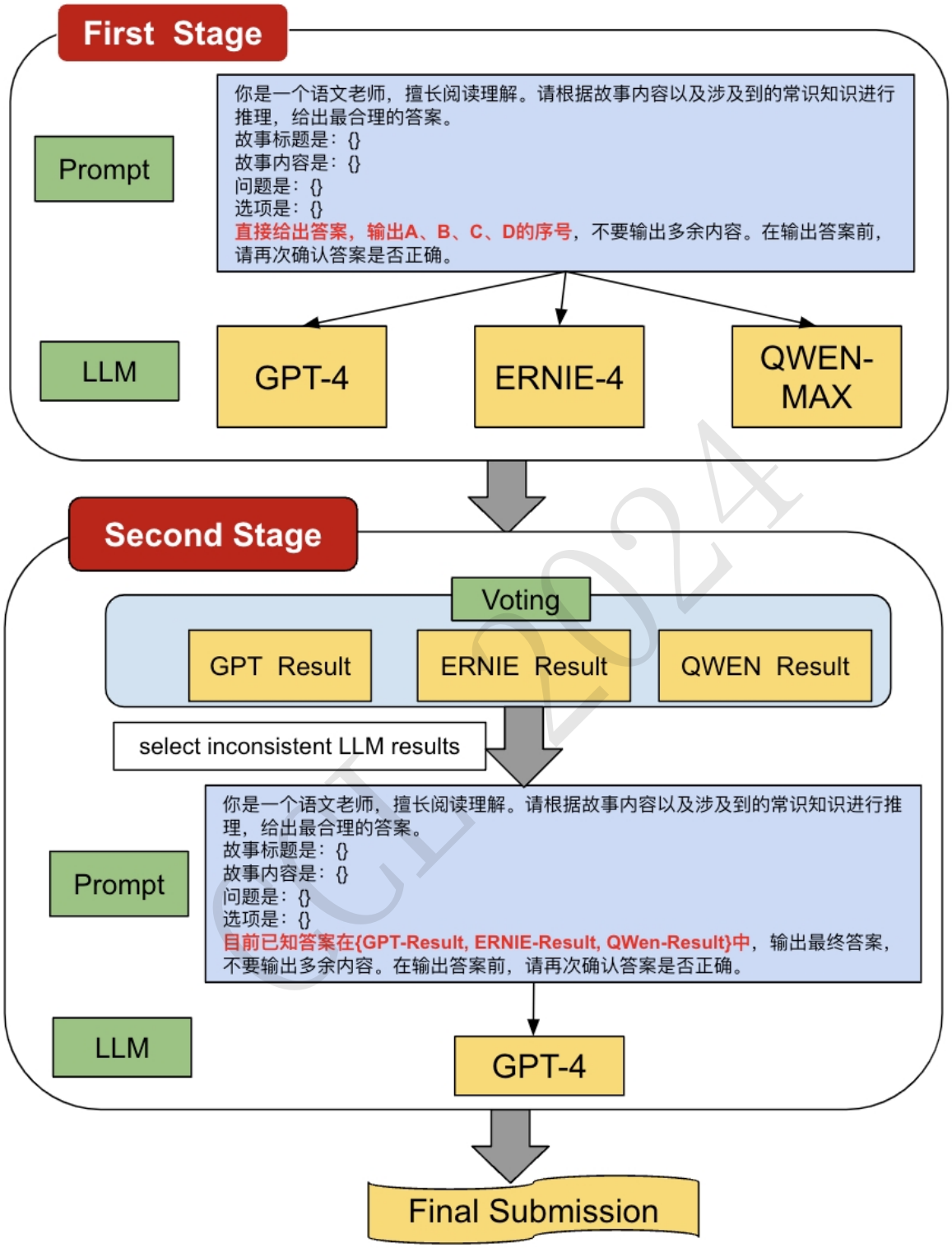

Figure 1: Two-stage prompt-based pipeline for the Common Reasoning (CR) subtask

```
你是一个语文老师，擅长寓言阅读理解。请根据故事内容以及涉及到的常识知识进行推理，给出最合理的答案。
故事标题是：{}
故事内容是：{}
问题是：{}
选项是：{}

直接给出答案，输出A、B、C、D的序号，不要输出多余的内容。在输出答案前，请再次确认答案是否正确。
```

Figure 2: CR subtask prompt for the first stage

```
你是一个语文老师，擅长寓言阅读理解。请基于给定的故事，从4个候选答案中选择最恰当的、最符合故事情节的寓意。
故事标题是：{}
故事内容是：{}
问题是：{}
选项是：{}

直接给出答案，输出A、B、C、D的序号，不要输出多余的内容。在输出答案前，请再次确认答案是否正确。
```

Figure 3: MU subtask prompt for the first stage

3.1.1 Prompt Design

Our designed prompts adhere to conventional prompt elements, including persona (e.g., a language teacher proficient in reading comprehension tasks), instruction (e.g., complete commonsense reasoning or moral understanding tasks for fables), input (e.g., story title, story content, questions, and options), and constraints (e.g., output only the answer option labels). We believe that current LLMs are sufficiently powerful to handle common understanding and reasoning tasks; hence, we did not invest effort in selecting demonstrations that are important in ICL. We have established different prompts for the CR and MU tasks. Please refer to Figures 2 and 3 respectively.

The results returned by LLMs will be further processed to extract the answer label.

3.1.2 Post-Processing

Although we have explicitly asked the LLMs to return the option label only, LLMs always return additional information, such as explanations. Observing the output from LLMs, we have identified a common pattern: in most instances, these models first present an answer label, followed by an explanation. Consequently, we have adopted a simple heuristic rule: traverse the output result from the beginning, and if the current character being traversed corresponds to one of the option labels A˜D, return the matched label as the post-processed result. It is important to note that this rule is not universally applicable. For example,

```
你是一个语文老师，擅长寓言阅读理解。请根据故事内容以及涉及到的常识知识进行推理，给出最合理的答案。
故事标题是：{}
故事内容是：{}
问题是：{}
选项是：{}

目前已知答案在{}几个选项中，给出最终的答案。
输出{}的序号即可，不要输出多余的内容。在输出答案前，请再次确认答案是否正确。
```

Figure 4: CR subtask prompt for the second stage

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 311-319, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

```
你是一个语文老师，擅长寓言阅读理解。请基于给定的故事，从4个候选答案中选择最恰当的、最符合故事情节的寓意。
故事标题是：{}
故事内容是：{}
问题是：{}
选项是：{}

目前已知答案在{}几个选项中，给出最终的答案。
输出{}的序号即可，不要输出多余的内容。在输出答案前，请再次确认答案是否正确。
```

Figure 5: MU subtask prompt for the second stage

models like QWen occasionally follow a different pattern, analyzing each option before providing a final answer. In such cases, our designed heuristic rule fails to apply.

3.2 Second Stage: Vote & Secondary LLM Confirmation

After obtaining results from three LLMs, we conduct a majority vote on the results. For instances where the voting results are inconsistent, we have further designed a second-stage prompting strategy to resolve the discrepancies.

In the development set, we observed that GPT's results significantly outperformed those of ERNIE and QWEN for the CR task. In the MU task, ERNIE's results were slightly better than GPT's, but the advantage was marginal. Therefore, we use GPT's results as the primary basis for voting, categorized into the following four scenarios (assuming the results returned by GPT, ERNIE, and QWEN are denoted as A, B, and C, respectively).

- **Case 1**: If all three models return the same result ($A = B = C$), we adopt this result.
- **Case 2**: If GPT's result matches one other model's result ($A = B$ or $A = C$), we adopt GPT's result.
- **Case 3**: If all three models return different results ($A \neq B \neq C$), we use the second-stage prompting strategy to resolve the discrepancy.
- **Case 4**: If ERNIE and QWEN's results match and differ from GPT's result ($B = C \neq A$), we consider the second-stage prompting strategy to resolve the discrepancy.

We define inconsistencies as cases where the results returned by GPT do not match other LLMs' outcomes, specifically cases 3 and 4. To address these inconsistencies, we have designed a secondary prompt strategy to request GPT once more to obtain the final result. The prompts in the second phase are essentially the same as those in the first phase, with the only difference being the set of answer options. In the first phase, the answer options consist of the initial four complete options from the test set. In the second phase, the prompts include answer options derived from the responses from the three LLMs in the first phase. The second step prompt for the CR & MU tasks is illustrated in Figures 4 and 5.

The outcome generated by GPT in the second step undergoes post-processing using the same method described in Section 3.1.2. The post-processed result is then chosen as the final submission answer.

4 Experiments

4.1 Datasets and Evaluation Metrics

The classic fable used in the CRMUS task was manually collected from web sources. The questions and answers for the CR subtask were manually annotated. A combination of automated construction and manual annotation was employed for the MU subtask. Overall, the annotation quality is high.

Table 1 presents the development and test data statistics.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 311-319, Taiyuan, China, July 25 - 28, 2024.
Volume3: Evaluations

split	subtask type	records number
dev set	CR	400
dev set	MU	252
test set	CR	1,692
test set	MU	1,056

Table 1: Statistics of the dev & test set

	CR Acc	MU Acc
GPT-4	88.00	71.33
ERNIE-4	82.67	72.67
QWen-MAX	82.00	70.67

Table 2: Accuracy for 150 randomly selected dev set records

The evaluation metric for both subtasks is accuracy. The final evaluation score of the competing model is calculated as the weighted average of all evaluation metrics, defined as:

$$\text{Score} = 0.4 \times \text{Acc}_1 + 0.6 \times \text{Acc}_2$$

where Acc_1 represents the accuracy of the CR subtask, and Acc_2 represents the accuracy of the MU subtask.

4.2 Experimental Setup

We use the three most advanced commercial LLMs – GPT-4[1], ERNIE-4[2], and Qwen-Max[3] – as our testbed. We directly utilized the API of the large models, with all model parameters set to the official API default values. Our experimental code is released at https://github.com/Holoflow/CCL2024-CRMUS-Track1.

4.3 Experimental Results

Considering the cost of LLMs, we evaluated[4] the performance of the three LLMs using only 150 randomly selected records from each task in the development set. Specifically, we randomly selected 150 records from the CR subtask and 150 records from the MU subtask. The accuracy of the three LLMs on the dev set is shown in Table 2.

From the dev set, we can draw a preliminary conclusion: GPT's results significantly outperformed those of ERNIE and QWen for the CR task. In the MU task, ERNIE's results were slightly better than GPT's, but the advantage was marginal. This also inspired us to base the voting strategy in the second stage primarily on GPT's results.

Table 3 represents the detailed result for the test set[5]. It is important to note that the **voting** method in the table is not the voting strategy described in Section 3 of this paper, but rather a majority voting strategy. In cases where the results of the three LLMs are inconsistent, the CR subtask uses the results from GPT-4, while the MU task uses the results from ERNIE-4.

It should also be noted that Qwen's experimental results are lower than those of the actual situation. This is because, during the post-processing stage, Qwen's result pattern differs slightly from that of the other two LLMs as explained in 3.1.2.

4.4 Analysis

Based on the experimental results from the dev and test sets, we can draw the following conclusions:

[1]Model snapshot used is gpt-4-turbo-2024-04-09
[2]Model snapshot used is Ernie-4.0-8K-0329
[3]Model snapshot used is qwen-max-0403
[4]We used the official evaluation script to run the results locally
[5]The results are extracted from the public dashboard: http://cuge.baai.ac.cn/#/ccl/2024/crmus

	CR Acc	MU Acc	Weight Score
GPT-4	84.46	68.37	74.80
ERNIE-4	78.30	69.13	72.80
QWen-MAX	76.83	66.57	70.68
Voting	84.99	71.21	76.72
Voting & Secondary Confirmation	86.52	74.43	79.26

Table 3: Results on the test set

- The MU tasks are more challenging than the CR tasks because CR tasks focus on reasoning based on objective facts, whereas MU tasks emphasize the understanding of subjective meanings. In the development set, our analysis of MU tasks revealed the presence of many ambiguous options for human beings. This also explains why the task organizers assigned a higher weight to the MU tasks in the evaluation metrics.
- The distribution of the dev set and the test set is significantly different. The same method shows a noticeable decrease in performance on the test set, with an average drop of 3-4 percentage points. It is speculated that the task organizers included more challenging data in the test set.
- The performance of the three LLMs is consistent across both the dev and test sets.
- A simple majority voting strategy does not significantly improve the performance of the CR subtask, with only a 0.5 percentage point increase compared to the best single model (GPT-4). However, it significantly improves the performance of the MU subtask, with an increase of more than 2 percentage points compared to the best single model (ERNIE-4).
- The secondary prompt enhanced strategy shows a significant improvement over the voting strategy in both the CR and MU subtasks, further validating the effectiveness of our approach.

5 Conclusion

This work proposes a simple yet effective two-stage prompt-based strategy for the CRMUS Track 1. In the first stage, we employ the same prompt to obtain responses from GPT-4, ERNIE-4, and Qwen-Max. We adopt a voting and secondary prompt-based confirmation strategy in the second stage. The experimental results demonstrate that our method achieves an average score of 79.27, ranking first in the closed domain among ten participating teams, thus confirming the effectiveness of our approach. These results further validate the advantages of LLMs in story reasoning and moral understanding, showcasing their potential in addressing complex tasks in the children's education domain. The successful application of our method underscores the robustness and reliability of combining multiple LLMs and utilizing a strategic voting mechanism.

Our approach demonstrates that even with a simple prompt-based strategy, significant improvements can be achieved by leveraging the diverse strengths of different LLMs. The integration of a secondary confirmation step ensures higher accuracy and consistency in the results, which is particularly beneficial for tasks requiring nuanced understanding and reasoning. We believe that LLMs will play an increasingly important role in children's education, offering sophisticated tools for story reasoning and moral comprehension that can adapt to the needs of young learners.

6 Future Work

In the future, we will continue to explore the potential of large models in CRMUS tasks from two main directions:

- Investigating More Effective Prompt Optimization Strategies: We aim to enhance the performance of general LLMs in CRMUS tasks, particularly in the Moral Understanding (MU) sub-tasks. This involves developing and refining prompt engineering techniques to better capture the nuances of

commonsense reasoning and moral judgment. By optimizing prompt design and employing advanced strategies for prompt adaptation, we can improve the LLMs' ability to understand and respond to complex educational scenarios more effectively.

- Developing Specialized Domain Models: We will focus on creating models specifically tailored to children's educational contexts. These specialized domain models will be designed to better adapt to the unique requirements of CRMUS and similar tasks, ensuring that the models can provide more accurate and contextually appropriate responses. By training models on datasets that reflect the specific language, themes, and moral considerations relevant to children's stories, we can enhance the applicability and effectiveness of LLMs in educational settings.

By pursuing these directions, we hope to push the boundaries of what LLMs can achieve in the field of children's education, contributing to the development of intelligent educational tools that support and enhance learning experiences for young learners.

7 Acknowledgements

We greatly thank all anonymous reviewers for their helpful comments. Thanks to Shanxi University and Hefei University of Technology for providing the high-quality CRMUS dataset. We also thank Min Zhou, Ye Yang, Wenyi Jiang, and Zhenjun Wang for helpful discussions.

References

Moore, Steven and Tong, Richard and Singh, Anjali and Liu, Zitao and Hu, Xiangen and Lu, Yu and Liang, Joleen and Cao, Chen and Khosravi, Hassan and Denny, Paul and Brooks, Chris and Stamper, John. 2023. *Empowering Education with LLMs - The Next-Gen Interface and Content Generation*. Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky, pages:32-37.

Chen, Yuetian and Si, Mei. 2024. *Reflections & Resonance: Two-Agent Partnership for Advancing LLM-based Story Annotation*. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, pages:13813-13818.

Zhang, Dell and Petrova, Alina and Trautmann, Dietrich and Schilder, Frank. 2023. *Unleashing the Power of Large Language Models for Legal Applications*. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pages:5257–5258.

Wu, Shijie and Irsoy, Ozan and Lu, Steven and Dabravolski, Vadim and Dredze, Mark and Gehrmann, Sebastian and Kambadur, Prabhanjan and Rosenberg, David and Mann, Gideon. 2023. *BloombergGPT: A large language model for finance*. ArXiv preprint ArXiv: 2303.17564.

Kai He and Rui Mao and Qika Lin and Yucheng Ruan and Xiang Lan and Mengling Feng and Erik Cambria. 2023. *A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics*. ArXiv preprint ArXiv: 2310.05694.

Brown, Tom and Mann, Benjamin and Ryder, Nick and Subbiah, Melanie and Kaplan, Jared D and Dhariwal, Prafulla and Neelakantan, Arvind and Shyam, Pranav and Sastry, Girish and Askell, Amanda and Agarwal, Sandhini and Herbert-Voss, Ariel and Krueger, Gretchen and Henighan, Tom and Child, Rewon and Ramesh, Aditya and Ziegler, Daniel and Wu, Jeffrey and Winter, Clemens and Hesse, Chris and Chen, Mark and Sigler, Eric and Litwin, Mateusz and Gray, Scott and Chess, Benjamin and Clark, Jack and Berner, Christopher and McCandlish, Sam and Radford, Alec and Sutskever, Ilya and Amodei, Dario. 2020. *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems, pages:1877-1901.

Shin, Taylor and Razeghi, Yasaman and Logan IV, Robert L. and Wallace, Eric and Singh, Sameer. 2020. *AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages:4222-4235.

Lester, Brian and Al-Rfou, Rami and Constant, Noah. 2021. *The Power of Scale for Parameter-Efficient Prompt Tuning*. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,pages:3045-3059.

Radford, Alec and Wu, Jeffrey and Child, Rewon and Luan, David and Amodei, Dario and Sutskever, Ilya. 2019. *Language Models are Unsupervised Multitask Learners*. OpenAI Blog 1(8), 9.

Li, Xiaonan and Lv, Kai and Yan, Hang and Lin, Tianyang and Zhu, Wei and Ni, Yuan and Xie, Guotong and Wang, Xiaoling and Qiu, Xipeng. 2023. *Unified Demonstration Retriever for In-Context Learning*. ArXiv preprint ArXiv: 2305.04320.

Rubin, Ohad and Herzig, Jonathan and Berant, Jonathan. 2022. *Learning To Retrieve Prompts for In-Context Learning*. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages:2655-2671.

Levy, Itay and Bogin, Ben and Berant, Jonathan. 2023. *Diverse Demonstrations Improve In-context Compositional Generalization*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages:1401-1422.