System Report for CCL23-Eval Task 9: HUST1037 Explore Proper Prompt Strategy for LLM in MRC Task

Xiao Liu, Junfeng Yu, Yibo He, Lujun Zhang, Kaiyichen Wei, Hongbo Sun, Gang Tu*
School of Computer Science and Technology, Huazhong University of Science and Technology
{liuxiaocs,heiheiyoyo,Heyibo,sheli,vichayturen,sunhb}@hust.edu.cn
tugang@hust.edu.cn

Abstract

Our research paper delves into the Adversarial Robustness Evaluation for Chinese Gaokao Reading Comprehension (GCRC_advRobust). While Chinese reading comprehension tasks have gained significant attention in recent years, previous methods have not proven effective for this challenging dataset. We focus on exploring how prompt engineering can impact a model's reading comprehension ability. Through our experiments using ChatGLM, GPT3.5, and GPT4, we discovered a correlation between prompt and LLM reading comprehension ability, and found that prompt engineering improves the performance of each model. Our team submitted the results of our system evaluation, which ranked first in three indexes and total scores.

***Keywords*— LLM, Prompt, Chinese Reading Comprehension**

1 Introduction

Machine Reading Comprehension (MRC), involves machines reading and comprehending human natural language text. Based on this, the machines are expected to answer questions related to the information in the text. This task is often used to evaluate the machine's ability to comprehend natural language, which can help humans quickly identify relevant information from a large amount of text. Additionally, it can help reduce the cost of manual information acquisition. MRC has strong application value in the fields of text Q&A, information extraction, and dialogue systems. In recent years, machine reading comprehension has gained significant attention from both industry and academia and has become one of the research hotspots in the field of natural language processing.

The competition, entitled GCRC_advRobust: Adversarial Robustness Evaluation for Chinese Gaokao Reading Comprehension, adds inference logic perturbation strategies to the regular reading comprehension task to improve the robustness of the machine reading comprehension model. Neural MRC has shown superiority over traditional rule-based and machine-learning-based MRC, and has gradually become the mainstream in the research community. However, compared with the large model, this method has obvious shortcomings. For example, BERT(Devlin et al., 2018) cannot handle too long input, and its input is only 512 tokens, which cannot solve the task at all. Taking into account the setting of the competition questions and the shortcomings of the existing solutions, We chose a suitable large language model that works efficiently with Chinese and adjusted the prompting strategy continuously to enable the model to reason based on the original text. Considering the length of the paragraphs in the reading comprehension task, we first want to find sentences that are highly relevant to the options and questions from the paragraphs and then analyze them, but the effect is not very good. So we tried to feed the entire paragraph into the model, and then analyzed the test results to improve the prompt continuously. Our final approach involved combining the model output with the original text to obtain the correct answer. Our team achieved significant improvements in all indicators, with a 39.2% increase in score, 44.1% in Acc0, 46.87% in Acc1, and 32.4% in Acc2 compared to the official benchmark model.

*Corresponding author: Gang Tu.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 310-319, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2 Related Work

Assessing the scalability of machine reading comprehension models relies heavily on their robustness in practical applications(Jia and Liang, 2017). Although current models have made significant strides in performing well on closed test datasets, their robustness in open, dynamic, real-world environments for reasoning and decision-making remains inadequate(Ren et al., 2022). To evaluate model robustness, previous studies have introduced text noise(Náplava et al., 2021) or rephrased the problem(Tang et al., 2021). However, these methods have limitations in measuring model performance due to their narrow focus on a single attack and relatively low topic difficulty.

Thanks to the fast evolution of big language models and their vast amount of semantic information, along with their impressive reasoning abilities, adopting a prompt modification paradigm can outperform the original models in certain NLP tasks. For instance, in reading comprehension tasks, the model can read the original text and select the correct answer from the options. This task is especially fitting as it involves a lengthy original text, and the options have overlapping content.

In our research, we examined the common methods that are currently used for large language models, which inspired our approach. We use p_θ to denote an LLM with parameters θ, and lowercase letters to represent an input language sequence, e.g. $x = (x_1, \cdots, x_n)$ where each x_i is a token, so that $p_\theta(x) = \prod_{i=1}^{n} p_\theta(x_i|x_1, \cdots, x_{i-1})$. We use uppercase latter to denote an output language sequence.

Prompt is the approach of adding extra information for the model to condition on during its generation of Y(Lester et al., 2021), which has become the prevailing method in the field of NLP. With prompt, we can turn the input x into output Y with LLM: $p_\theta(y|prompt(x)$, where $prompt(x)$ warps input x with extra information for the problem, prompt methods include the following.

Zero-Shot Large language models like GPT-3 are capable of performing certain tasks without any prior training due to their ability to follow instructions and being trained on vast amounts of data. Recent studies have shown that instruction tuning can enhance zero-shot learning(Wei et al., 2022a).

Few-Shot Although large-language models have impressive zero-shot abilities, they struggle with more complex tasks when relying solely on this approach. To address this, few-shot prompting can be used to facilitate in-context learning. By providing demonstrations within the prompt, we can guide the model towards better performance. These demonstrations act as conditioning for subsequent examples where we want the model to generate a response. In a study by (Min et al., 2022), the importance of these demonstrations for the success of in-context learning was explored.

CoT Introduced in (Wei et al., 2022b), Chain-of-Thought (CoT) prompting enables complex reasoning capabilities through intermediate reasoning steps. By using CoT with few-shot prompting, one can achieve even better results on challenging tasks that require reasoning before responding.

In our research, considering that the original paragraph is long, we first construct the Prompt based on Zero-shot, then add examples using the Few-shot prompt model to the extent allowed by the Token, and finally construct CoT prompt samples based on different question types.

3 Task Description

3.1 Data

To improve the robustness of machine reading comprehension models in complex, realistic adversarial environments, construct a subset of adversarial robustness based on GCRC(Tan et al., 2021), the dataset of Gaokao Chinese Reading Comprehension and proposed the task of "GCRC_advRobust: Adversarial Robustness Evaluation for Chinese Gaokao Reading Comprehension". This assessment designs four adversarial attack strategies (keyword perturbation, inference logic perturbation, spatio-temporal attribute perturbation, causality perturbation), focusing on the model's robustness under various adversarial attacks. We can use ChatGPT, ChatGLM and other large models in the open track.

In the following classification, the correct option refers to the option that matches the meaning of the original text; the incorrect option refers to the option that does not match the meaning of the original text; the positive confrontation option is the same as the original option positive or incorrect; the negative confrontation option is the opposite of the original option positive or incorrect. The inference logic per-

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 310-319, Harbin, China, August 3 - 5, 2023.

Option	Text
Original Option(Wrong)	自然资源丰富的湿地，是人类的“衣食父母”， 为人类生存发展提供了所有物资，如食物、饮水、能源等。
Positive Confrontation Option	自然资源丰富的湿地，是人类的“衣食父母”， 为人类生存发展提供了全部物资，如食物、饮水、能源等。
Negative Confrontation Option	自然资源丰富的湿地，是人类的“衣食父母”， 为人类生存发展提供了部分物资，如食物、饮水、能源等。

Table 1: Keyword Scrambling Strategy Example

Option	Text
Original Option(Wrong)	由于19世纪中叶中国文化在与西方文化的抗争中处于 弱势地位，人们才提出“保存国学”“振兴国学”的口号， “国学”一词由此出现。
Positive Confrontation Option	20世纪中叶中国文化在与西方文化的抗争中处于弱势地位， 人们才提出“保存国学”“振兴国学”的口号。
Negative Confrontation Option	19世纪中叶中国文化在与西方文化的抗争中处于弱势地位， 20世纪初，人们才提出“保存国学”“振兴国学”的口号。

Table 2: Spatio-temporal attribute perturbation strategy

turbation strategy mainly attacks the reasoning process of concluding the original text through inductive or deductive reasoning.

1. **Keyword scrambling strategy**: as shown in Table 1, interfere with keywords that affect the semantics of the options by word substitution or rephrasing. In the following example, the positive confrontation option replaces 所有(all) with 全部(all), these two words are very close in meaning in Chinese, and the negative confrontation option replaces 全部(all) with 部分(some) so that the meaning changes.

2. **Spatio-temporal attribute perturbation strategy**: as shown in Table 2, interfere with the spatio-temporal information in the options by changing the temporal or spatial attributes. In the example below, the positive confrontation option uses ”20世纪中叶” (mid-20th century) to replace ”19世纪中叶” (mid-19th century) in the original option, but neither option matches the meaning of the original, so both are incorrect. The time of the corresponding event in the negative confrontation option is correct.

3. **Cause-and-effect perturbation strategy**: as shown in Table 3, interfering with the cause-and-effect relationship in an option by changing or removing the causal link. In the example below, the positive confrontation option replaces the conjunction in the original option. Still, there is no causal relationship before or after the sentence, so both are incorrect options, while the negative confrontation option is correct.

4. **Reasoning logic perturbation strategy**: as shown in Table 4, interfere with the logical reasoning

Option	Text
Original Option(Wrong)	中国之所以选择和平共处五项原则， 是为了在务实的基础上让外界消除误解。
Positive Confrontation Option	因为中国选择了和平共处五项原则， 所以在务实的基础上让外界消除误解。
Negative Confrontation Option	中国选择和平共处五项原则， 并积极在务实的基础上让外界消除误解。

Table 3: Cause-and-effect perturbation strategy

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 310-319, Harbin, China, August 3 - 5, 2023.

Option	Text
Original Option(Right)	气味分子在属于G蛋白的嗅觉受体的作用下 从化学信号转变成为电信号。
Positive Confrontation Option	与属于G蛋白的嗅觉受体结合后，在它的作用下， 气味分子从化学信号转变成为电信号。
Negative Confrontation Option	气味分子与嗅觉受体结合后， 气味分子便自行从化学信号转变成为电信号。

Table 4: Reasoning logic perturbation strategy

DataSet	Validate	Test
Questions/Options	336/1344	288/1152
Keyword Scrambling Strategy	504	418
Spatio-temporal Attribute Perturbation Strategy	619	543
Cause-and-effect Perturbation Strategy	192	172
Reasoning Logic Perturbation Strategy	29	19

Table 5: Dataset

process of the options by rewriting the premises or conclusion. In the example below, both the positive-opposition option and the original option are correct in the original text. Ill, the negative-opposition option needs to include the prerequisite for G protein action and is therefore incorrect.

3.2 Evaluation

This evaluation provides GCRC raw data as a training set, the number of questions is 6994, and GCRC_advRobust is provided as a verification set and a test set. The scale of the GCRC_advRobust dataset is shown in Table 7.

The final score of the participating system is determined by a combination of Acc_0, Acc_1, and Acc_2 metrics, which are calculated as follows:

$Score = 0.2 * Acc_0 + 0.3 * Acc_1 + 0.5 * Acc_2$

Where:

Acc_0 = number of original questions correctly predicted/total number of questions

Acc_1 = number of correct predictions for the original question and any of the confrontation questions/total number of questions

Acc_2 = number of correct predictions for both the original and the two confrontation questions/total number of questions

4 Model Selection

The open track does not limit the models that can be used. Considering that the task needs to process hundreds of data, we selected the LLM that can be called in the form of api, and introduced some models below.

4.1 Available Models

ChatGLM-6B(Du et al., 2022) is an open bilingual language model based on General Language Model (GLM) framework, with 6.2 billion parameters. ChatGLM-6B uses technology similar to ChatGPT, optimized for Chinese QA and dialogue. The model is trained for about 1T tokens of Chinese and English corpus, supplemented by supervised fine-tuning, feedback bootstrap, and reinforcement learning with human feedback. With only about 6.2 billion parameters, the model can generate answers that align with human preference.

GPT-3.5 models can understand and generate natural language or code. The most capable and cost-effective model in the GPT-3.5 family is gpt-3.5-turbo which has been optimized for chat but works well for traditional completions tasks as well.

GPT-4 is a large multimodal model that can solve difficult problems with greater accuracy than any of the previous models of OpenAI, thanks to its broader general knowledge and advanced reasoning capabilities. Like gpt-3.5-turbo, GPT-4 is optimized for chat but works well for traditional completions tasks both using the Chat Completions API.

4.2 Selection Strategy

We applied the same testing strategy on various models and their respective scores are displayed in Table 7 on validate dataset. After thorough evaluation, we have chosen gpt-3.5-turbo as our preferred model for further enhancements. It has demonstrated exceptional performance on the validation dataset and is also convenient to access through its API. We opted not to use gpt-4 due to the unavailability of a stable API and the tendency for the generated content to be excessively long, making it difficult to discern the correct answer from the provided options. The strategy we used is Strategy 3, shown in the appendix.

We also explored the impact of the parameters of the gpt-3.5-turbo api on the performance of the model.

Temperature What sampling temperature to use, between 0 and 2. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic.

Top_p An alternative to sampling with temperature, called nucleus sampling, where the model considers the results of the tokens with top_p probability mass. So 0.1 means only the tokens comprising the top 10% probability mass are considered.

Max_tokens The maximum number of tokens to generate in the chat completion.

Presence_penalty Number between -2.0 and 2.0. Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics.

Frequency_penalty Number between -2.0 and 2.0. Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim.

We began by analyzing the API request's system field, which imposes stricter constraints on the model compared to the prompt. Our goal was to transform the model into a reading comprehension tool that could accurately answer questions. Next, we conducted thorough experiments to evaluate the impact of the aforementioned parameters on the model. Ultimately, we identified the optimal parameter settings that yielded the best results.

Parameters	Temperature	Top_p	Max_tokens	Presence_penalty	Frequency_penalty
Value	0.15	0.1	2048	-1.25	-0.75

Table 6: Parameters setting

5 Experiment

5.1 Experiment Setup

During the competition, our team conducted two phases of experiments. The first phase involved testing various large language models for their performance in Chinese reading comprehension. As the official training set provided only one question per passage, we utilized the validation set to evaluate different models, given that the format of the test set was different. In the second phase, we focused on devising an effective hinting strategy for the large language model. We experimented with different answer extraction methods, testing various system prompt words, paragraphs, and options using different splicing techniques to optimize the algorithm.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 310-319, Harbin, China, August 3 - 5, 2023.

5.2 Prompt Strategy

Following phase 1, we acquire a collection of parameter values that exhibit the most exceptional overall performance. These values are subsequently utilized in all subsequent strategy experiments. Table 6 displays the corresponding values.

Model	Score	Acc_0	Acc_1	Acc_2
MacBert	6.91	28.82	3.82	0
chatglm-6b	10.38	27.78	13.19	1.74
vicuna-7b	12.85	29.51	17.36	3.47
gpt-3.5-turbo	31.08	50.00	37.85	19.44

Table 7: Using the same strategy on different models

During our experiments, we utilized different prompts to assess the impact on gpt-3.5-turbo. The outcomes of these varied prompts on the test set are presented in Table 8. We attempted multiple methods to adjust the prompt format in order to produce the desired output option directly from gpt-3.5-turbo, but none proved successful. However, we did notice that gpt-3.5-turbo tends to analyze the question before providing an answer. As a result, we utilized the last option of the regular match response as the answer. Based on our tests, we discovered that gpt-3.5-turbo performed exceptionally well in answering question 2 when tested on the validation dataset. To improve the robustness between questions, we design a strategy. We start by asking question 2 and integrate it into the history record. Then, we move on to question 1, integrate it into the record, and conclude by asking question 3. For further assistance on implementing these strategies, please refer to the appendix.

In our research, we conducted a detailed examination of the model's performance on the validation dataset. When considered individually, the accuracy rates for question one and question two were each approximately 50%, while question three had an accuracy rate of around 30%. Despite the relatively high accuracy rates for each separate question, the model's ability to correctly answer all questions simultaneously was found wanting, leading to a deficiency in the final score. The objective of this task was to test the model's robustness, that is, its capacity to handle multiple questions at once. If the model could correctly answer all three questions in a single attempt, then its score would be significantly boosted. We observed that in human problem-solving processes, if the first question is answered correctly, this provides additional information that greatly increases the likelihood of correctly answering all questions simultaneously. Therefore, we believe that this human problem-solving habit should be considered during the model's training and optimization process. The model should be able to utilize the information provided by correctly answered questions to increase its accuracy when dealing with subsequent questions. This approach may help improve the model's robustness when handling multiple questions simultaneously, ultimately enhancing its overall score.

Model	Strategy	Score	Acc_0	Acc_1	Acc_2
gpt-3.5-turbo	1	33.96	51.39	41.32	22.57
	2	32.47	45.14	38.19	23.96
	3	12.99	28.47	17.36	4.17
	4	31.08	50.00	37.85	19.44
	5	29.86	49.31	35.42	18.75

Table 8: Using different strategies on the same model

5.3 Result

The test dataset for this competition is given in a closed format, with only the original text provided, and each team submits the results file to the online measurement platform. Each team was allowed to submit

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 310-319, Harbin, China, August 3 - 5, 2023.

test set results three times per day. Table 9 shows the results of this competition, **Baseline** is the official baseline, and HUST1037 is our team name.

Team	**Score**	Acc_0	Acc_1	Acc_2
Baseline	6.42	22.22	6.6	0
HUST1037	**45.62**	**66.32**	**53.47**	**32.64**
斯灵思	32.47	45.14	38.19	23.96
lostlost	32.08	50.35	39.24	20.49
一二三四	6.04	25	3.47	0
UIRISC	5.45	23.61	2.43	0

Table 9: Official baseline model and top five team metrics

We concluded from the experiment that the prompt should present the task content as clearly as possible and be very concise.

6 Conclusion

Our team used a large language model based on GPT4 for the Adversarial Robustness Evaluation in the Chinese Gaokao Reading Comprehension task. We modified and tested various prompting strategies to enable the model to make logical inferences from the original text. This method utilizes the semantic information and reasoning ability of the large model more effectively compared to the original approach to solving reading comprehension tasks. However, there are still some limitations to the current system. We were unable to try more hinting strategies due to the display limitation of the GPT4 model API. Additionally, the original text in this task was lengthy, and the model input length was restricted, leading to a shorter scalable content. We aim to compress the original text information to enable us to try out more hinting strategies in the future.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Jakub Náplava, Martin Popel, Milan Straka, and Jana Straková. 2021. Understanding model robustness to user-generated noisy texts. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 340–350, Online, November. Association for Computational Linguistics.

Feiliang Ren, Yongkang Liu, Bochao Li, Shilei Liu, Bingchao Wang, Jiaqi Wang, Chunchao Liu, and Qi Ma. 2022. An understanding-oriented robust machine reading comprehension model. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2), dec.

Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. 2021. GCRC: A new challenging MRC dataset from Gaokao Chinese for explainable evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1319–1330, Online, August. Association for Computational Linguistics.

Hongxuan Tang, Hongyu Li, Jing Liu, Yu Hong, Hua Wu, and Haifeng Wang. 2021. DuReader_robust: A Chinese dataset towards evaluating robustness and generalization of machine reading comprehension in real-world applications. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 955–963, Online, August. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*.

A Appendix

A.1 Strategy 1

The gpt-3.5-turbo request format is as follows:

[
{
'role': 'system',
'content': '你现在是一个答题系统，根据输入的段落、问题、选项，回答A、B、C、D其中一个即可。'
},
{
'role': 'user',
'content': ''
}
]

The prompt is as follows:

段落：
The passage field of the test set data
问题：The question field of the test set data
选项:
A: Option A content
B: option B content
C: option C content
D: option D content
根据以上内容选择答案。

A.2 Strategy 2

The gpt-3.5-turbo request format is as follows:

[
{
'role': 'system',
'content': '你现在是一个答题系统，根据输入的段落、问题、选项，回答A、B、C、D其中一个即可。'
},
{
'role': 'user',

'content': ''
}
]
The prompt is as follows:
段落:
The passage field of the test set data
问题1：The question field of the test set data
选项:
A: Option A content
B: option B content
C: option C content
D: option D content
答：answer1
问题2：The question field of the test set data
选项:
A: positive option A content
B: positive option B content
C: positive option C content
D: positive option D content
答：answer2
问题3：The negative_question field of the test set data
选项:
A: negative option A content
B: negative option B content
C: negative option C content
D: negative option D content
答:

A.3 Strategy 3

The gpt-3.5-turbo request format is as follows:

```
[
 {
    'role': 'system',
     'content': 'IMPORTANT: You are a quiz assistant powered by the gpt-3.5-turbo model'
 },
 {
     'role': 'user',
     'content': ''
 }
]
```

The prompt is as follows:
段落:
The passage field of the test set data
问题1：The question field of the test set data
选项:
A: Option A content
B: option B content
C: option C content
D: option D content
问题2：The question field of the test set data
选项:

A: positive option A content
B: positive option B content
C: positive option C content
D: positive option D content
问题3：The negative_question field of the test set data
选项:
A: negative option A content
B: negative option B content
C: negative option C content
D: negative option D content
根据以上内容选择答案。