Mozi: A Scientific LLM Technical Report

Tian Lan, Tianyi Che, Zewen Chi, Xuhao Hu, Xian-ling Mao
Beijing Institute of Technology
{ccty,maoxl}@bit.edu.cn
lantiangmftby@gmail.com

Abstract

Large Language Models (LLMs) become the dominant solution in various NLP tasks. Nevertheless, it is still difficult for recent LLMs to leverage that domain-specific knowledge to resolve domain-related problems, such as law and scientific papers. In this paper, we present Mozi (墨子), a large language model for instruction following in scientific domains, such as paper-ground question answering, paper-ground multi-turn dialog, and emotional conversation with academic researchers. Moreover, we also train an efficient dense retrieval model, named SciDPR, which can effectively search evidence related to user queries from the papers, thereby assisting Mozi in accurately answering user's questions. Experimental results prove that SciDPR outperforms OpenAI embedding service in evidence retrieval task, and extensive experimental results on widely used scientific benchmarks reveals Mozi's strong capabilities to solve scientific problems. Besides, we have publicly released resources at `https://github.com/gmftbyGMFTBY/science-llm`, including model weights, dataset and source codes.

1 Introduction

In the enhancement of open-source large language models (LLMs) via instruction tuning, instances such as Vicunna (Chiang et al., 2023) and ChatGLM (Zeng et al., 2023) demonstrate remarkable performance across various downstream NLP tasks. Nevertheless, the approach to analyze and solve domain-specific tasks, such as law (Huang et al., 2023) and scientific paper (Dasigi et al., 2021), could substantially deviate from what the model is traditionally designed to handle in general domains. Despite a model's potential ability to grasp the knowledge within a particular domain, the application of this knowledge to this domain-specific scenarios continues to be an open-problem (Huang et al., 2023; Yu et al., 2023).

In this paper, we propose the Mozi large-scale language models, which are further pre-trained on scientific data samples, showing a better understanding and generation capability in the scientific domain. Specifically, we first fine-tune the general dense retrieval model (Karpukhin et al., 2020) on the scientific evidence retrieval dataset (Dasigi et al., 2021), named SciDPR. SciDPR is used to retrieve valuable evidence from the paper given a user's queries, leading to more accurate responses. Moreover, the general large-scale language models are continually pre-trained on the arXiv corpus with billion tokens, and further fine-tuned on the well-annotated paper-ground question-answering datasets, such as QASPER (Dasigi et al., 2021) and SciMRC, enabling it to answer user's questions by referencing the content of scientific papers. Finally, the self-instruct technique (Wang et al., 2022) is used to generate a large amount of high-quality dialogue samples for fine-tuning Mozi model, enabling it to provide emotional support and scientific advice through multi-round dialogue. Although existing dialogue-based LLMs could provide advice and basic dialogue capabilities, their responses are often lengthy and formal, resulting in a poor user experience, thereby hindering the models ability to provide emotional support. To solve this problem, we invited over 20 academic researchers to annotate 50 multi-round emotional dialogue seed samples. Using these seed samples as examples, we generated a large amount of high-quality multi-round dialogue samples from ChatGPT for further fine-tuning Mozi models.

Main contributions of this study can be summarized as follows:

- We propose the SciDPR model, which could retrieve valuable evidence for user's questions efficiently.

Work in Progress

- We have released the Mozi models, which could accurately answer user's questions by fine-tuning them on the scientific datasets.
- We annotate 50 seed emotional dialogue samples, and generate over 1000 high-quality scientific emotional dialog samples for fine-tuning Mozi models.

2 Methodology

In this section, we introduce the training details of our proposed SciDPR and Mozi models. Besides, the three training phase of Mozi are also described: (1) scientific pre-training; (2) paper-ground instruction; (3) scientific emotional dialogue training.

2.1 SciDPR

We first propose a dense retrieval model for retrieving valuable evidence from the paper given the user's query. With the help of offline computation and efficient online retrieving, dense retrieval models achieve much better inference efficiency than traditional cross-encoder models, for example, the LED-Base (Dasigi et al., 2021). Thanks to the carefully annotated pairs of questions and ground-truth evidences, we could fine-tune the general dense retrieval model (Karpukhin et al., 2020) with an in-batch negative sampling method.

Specifically, given the question q, ground-truth evidences $\{e_i^+\}_{i=0}^k$ and a bunch of random sampled paragraphs in the paper $\{e_i^-\}_{i=0}^m$, the InfoNCE loss is used to optimize the dense retrieval model:

$$\mathcal{L} = -\log \frac{\sum_{j=0}^k e^{E_q(q)\cdot E_a(e_j^+)}}{\sum_{j=0}^k e^{E_q(q)\cdot E_a(e_j^+)} + \sum_{j=0}^m e^{E_q(q)\cdot E_a(e_j^-)}} \quad (1)$$

where E_q and E_a are the question encoder and the answer encoder in the dense retrieval model. After this fine-tuning phase, the dense retrieval model could accurately retrieve valuable evidences given the user's query. We name this fine-tuned dense retrieval model as SciDPR, and it will help our proposed Mozi model to generate factual responses.

2.2 Mozi

Scientific Pre-training Most large-scale language models (LLMs) pre-trained with general text corpora (Touvron et al., 2023; Computer, 2023), such as CommonCrawl and Wikipedia, perform very well in general scenarios. However, due to a relative lack of domain-specific data, their understanding of tasks in specific domains can be somewhat limited, such as law and scientific papers (Yu et al., 2023; Huang et al., 2023). Moreover, the strategies for dealing with specific domain tasks may significantly differ from those used in the general text domain. These issues limit the potential of general large-scale language models in specific domain applications. To solve this problem, we first use a large-scale arXiv corpus to further pre-train the general large-scale language model, thereby adapting it to the data distribution of specific domains, i.e. the scientific paper domain. Specifically, the additional LoRA weights (Hu et al., 2021) are optimized upon the large-scale language models by:

$$\mathcal{L} = \Pi_{t=0}^T f(y_t|\theta_{\mathrm{LLM}}, \theta_{\mathrm{LoRA}}, y_{<t}) \quad (2)$$

where θ_{LLMs} is the parameters of LLM f, which is frozen during pre-training. y is a chunk of the sequential tokens tokenized from the scientific paper data samples. θ_{LoRA} is the trainable LoRA parameters.

Paper-ground Instructions After the scientific pre-training phase, we further supervised fine-tune (SFT) the LoRA parameters of the large-scale language model, enabling it to understand question-answering instruction related to papers, and generate accurate dialogue responses to solve user problems. Specifically, given the user's query q and a scientific paper S from QASPER (Dai et al., 2022) and SciMRC train set, the ground-truth evidences $\{e_i\}_{i=0}^k$ related to user's query are first extracted from S, where k is the number of the evidences. Then, based on the evidences $\{e_i\}_[i = 0]^k$ and user's query q, the LoRA weights θ_{LoRA} are further optimized by:

$$\mathcal{L} = \Pi_{t=0}^T f(y_t|\theta_{\mathrm{LLM}}, \theta_{\mathrm{LoRA}}, y_{<t}, q, \{e_i\}_{i=0}^k) \quad (3)$$

where a is the ground-truth answer related to the query q. Note that we directly optimize the LoRA weights added during pre-training process, and the model parameters of LLMs are still frozen during supervised fine-tuning.

Scientific Emotional Dialogue Training Although various chat-based LLMs, such as ChatGLM (Zeng et al., 2022) and ChatGPT, could conduct the fluent and meaningful dialogue with users, their generations are too lengthy and formal to provide emotional support during conversation with users.

To solve this problem, we simply leverage the self-instruct solution to generate over 1k engaging and emotional multi-turn dialogue samples with the help of the ChatGPT and well-designed prompt. Specifically, we do the human annotation and collect over 50 high-quality emotional multi-turn dialogue about the scientific problems. Then, the prompt with few-shot demonstrations is fed into ChatGPT to generate more diverse and engagin dialogue samples. The prompt can be found in Appendix A.

After the self-instruct process, we finally collect over 1k multi-turn emotional dialogue dataset, which is enough to fine-tune the LLMs. Similar to the scientific pre-training phase, the multi-turn dialogue history is concatenated as the single text segment for training. Please refer to equal 2 for more details.

3 Experimental Setup

In this section, we introduce the resources and details about training the Mozi LLM and SciDPR models.

3.1 Models

Large-scale Langauge Models In this study, we pre-train and fine-tune two LLMs: (1) LLaMA-7B (Touvron et al., 2023) has been proved and widely used for various supervised fine-tuning processes; (2) Baichuan-7B[1] is a recently released large-scale language model that has demonstrated strong competitiveness in both Chinese and English languages.

SciDPR model In this paper, we further fine-tune the DPR[2] model on QASPER evidence retrieval samples, and each encoder has over 0.1B parameters.

3.2 Data Resources

Scientific Pre-training To adapt the available large-scale language models (LLMs) to scientific domain, we further pre-train the LLMs on the scientific corpus. Specifically, we fist collect 4 billion tokens from the arXiv split of widely used Redpajama corpus[3]. Moreover, in order to maintain the model's ability to understand general text, we inject 0.2 billion tokens from CommonCrawl corpus into the scientific text pre-training corpus. As the Baichuan-7B model is a bilingual model in Chinese and English, we additionally supplemented the Baichuan-7B model with 0.4 billion tokens from Chinese Wikipedia corpus. Furthermore, over 1000 arXiv paper are hold as the test split for monitoring the perplexity during pre-training.

Paper-ground Instructions After further pre-training the LLMs on arXiv split with 4 billion tokens, we conduct SFT fine-tuning on two widely used paper-ground question-answering datasets, i.e. QASPER and SciMRC, to enhance the Mozi model's capability to follow user's instructions based on specific paper. The statistics of these two datasets are shown in Table 1.

Datasets	Train	Dev	Test
QASPER	2314	1601	1268
SciMRC	3972	484	1099

Table 1: The number of the QA pairs in each split of QASPER and SciMRC datasets.

Scientific Emotional Dialogue Mozi's another feature is to conduct the fluent conversation with users to provide the emotional support. As described above, our emotional dialogue dataset contains 1069 multi-turn emotional dialogue samples with average 5.07 utterances.

3.3 Training Details

Due to limited computational resources, in this work, we utilized the recently proposed QLoRA (Dettmers et al., 2023) optimization method combined with the DeepSpeed toolkit[4] to directly optimize LoRA parameters on an 8-card 3090 server. The specific parameters are shown in the Table 2.

4 Experiments

In this technical report, we mainly introduce and analyze the following important components: (1) SciDPR: a state-of-the-art dense retrieval model for recalling important evidence given the user's queries from the paper; (2) Mozi model: a scientific large-scale langauge model (LLM) for solving

[1] https://github.com/baichuan-inc/baichuan-7B

[2] Evidence encoder: https://huggingface.co/facebook/dpr-ctx_encoder-single-nq-base/tree/main; Question encoder: https://huggingface.co/facebook/dpr-question_encoder-multiset-base

[3] https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T

[4] https://github.com/microsoft/DeepSpeed

Parameters	Value
Zero stage	2
Warmup ratio	0.05
Pre-train Step	122k
Fine-tune Step	2k
Sequence Length	4096
Batch Size	64
Gradient Clipping	1.0
LoRA Rank	64
LoRA Alpha	16.0
LoRA Dropout	0.1
LoRA weights	> 0.16B

Table 2: The hyper-parameters during pre-training and fine-tuning.

two kinds of scientific problems: (a) paper-ground instruction; (b) scientific emotional dialogue.

This section's organizational content is as follows: (1) the SciDPR's performance is well evaluated on popular QASPER benchmark; (2) the perplexity curve during scientific pre-training is shown to prove its effective domain-adaption capability; (3) Mozi's capability about solving scientific problems is carefully evaluated on SciMRC and QASPER datasets; (4) the qualitative analysis of the Mozi model's ability to engage in scientific emotional conversations is shown.

4.1 Evaluation on Evidence Retrieval

Following previous works (Dasigi et al., 2021), in this paper, we comprehensively analyze the performance of the proposed dense retrieval model, SciDPR, on the widely used QASPER dataset. Specifically, the evidence retrieval task in QASPER benchmark evaluates the F1 score of selecting the ground-truth evidences in the paper related to user's queries.

Except for the basic baselines listed in QASPER (Dasigi et al., 2021; Caciularu et al., 2021), the dominant sentence embedding service, i.e. OpenAI embedding[5], is also test in this experiment.

The experimental results are shown in Table, from which we could make following conclusions: (1) dense retrieval models (SciDPR and OpenAI embedding) achieves the competitive performance with the cross-encoder baselines (LED-base and LED-base$_{\mathrm{InfoNCE}}$). In the view of the much lower inference cost, the SciDPR and OpenAI embedding are more practical in real-world scenarios; (2) Our proposed SciDPR achieves better Evidence-F1 scores than OpenAI embedding service on Dev set and test set, indicating its better performance in scientific domain. Although the performance gap between our model and OpenAI embedding serve is not significant, in the view of the relative high cost of using OpenAI APIs, our SciDPR that deployed locally is less expensive and faster.

[5] In this work, we use the default model of OpenAI embedding service.

Model	Dev set	Test set
First Paragraph	0.71	0.34
Random Paragraph	2.09	1.30
TF-IDF	10.68	9.20
LED-base	23.94	29.85
LED-base$_{\mathrm{InfoNCE}}$	24.90	30.60
OpenAI Top-1	24.41	30.00
OpenAI Top-2	26.67	31.46
OpenAI Top-3	26.37	29.71
OpenAI Best	35.04	40.17
SciDPR Top-1	25.41	30.38
SciDPR Top-2	26.74	30.98
SciDPR Top-3	26.68	29.58
SciDPR Best	36.83	**40.67**
Human$_{lb}$	-	71.62

Table 3: Evidence-F1 on QASPER benchmark.

4.2 Scientific Pre-training

To prove the effectiveness of the scientific pre-training process, the perplexity is evaluated on additional test set during the pre-training process. Specifically, this test set consists of over 1000 chunks with 2048 tokens extracted from the scientific papers. As shown in Table 4, it can be found that perplexity on the hold-out arXiv split becomes lower, indicating Mozi models adapt to the scientific domains, thereby laying the foundation for the subsequent SFT training process. Moreover, the improvement in the perplexity of the LLaMA-7B model is not as significant as those of the Baichuan-7B model. The main reason is that the pre-training corpus of LLaMA-7B includes a portion of arXiv data, so it is to some extent more adept at understanding scientific and technological text data. However, even though the perplexity of the Baichuan-7B model is worse than LLaMA-7B model, the Baichuan-7B model achieves a lower perplexity than the LLaMA-7B model after pre-training.

Models	Perplexity
LLaMA-7B	3.9118
LLaMA-7B$_p$	3.7079
Baichuan-7B	6.9544
Baichuan-7B$_p$	**3.4633**

Table 4: Perplexity scores on the arXiv test set with over 1000 chunks that contain 2048 tokens. LLaMA-7B$_p$ and Baichuan-7B$_p$ are pre-trained on the 4 billion tokens from arXiv corpus.

4.3 Evaluation on Paper-ground Instructions

Following previous works, we evaluate Mozi and baselines on widely used QASPER and SciMRC datasets. Following previous works, the BLEU, ROUGE$_{\text{sum}}$, BERTScore (Zhang et al., 2019), and Answer-* (Precision, Recall and F1) (Dasigi et al., 2021) metrics are used to evaluate the accuracy of the question answering. Note that the responses for paper-ground instructions are deterministic, and the above word-overlap based evaluation metrics are good enough to examine their performance.

Evaluation on QASPER test set In this section, we compare our proposed Mozi LLMs with some widely used chat-based large-scale language models, such as ChatGLM-6B, Alpaca (Taori et al., 2023), OpenAlpaca (Su et al., 2023), Vicuna-7B (Zheng et al., 2023) models. Morover, in order to prove the effectiveness of the scientific pretraining phase, we also directly fine-tune LLaMA-7B and Baichuan-7B models on QASPER test set, without scientific pretraining, and we named them LLaMA-7B$^+$ and Baichuan-7B$^+$. The experimental results on QASPER test set are shown in Table 5, from which we could make following conclusions: (1) Mozi-7B$_{\text{LLaMA}}$ and Mozi-7B$_{\text{Baichuan}}$ models are better than LLaMA-7B$^+$ and Baichuan-7B$^+$ models. This phenomenon prove the effectiveness of the scientific pre-training, which make Mozi models easily to understand and respond to the paper-ground instructions; (2) compared with poweful chat-based large-scale langauge models, such as Vicuna-7B and ChatGLM-6B models, Mozi-7B model achieves much higher Answer-F1 scores, indicating that Mozi-7B models could generate more accurate and concise responses to the user's questions about the paper; (3) given the ground-truth evidences, Mozi-7B could achieves very close Answer-F1 score to performance of human lower bound (0.6092); (4) compared with the experimental results in Table 5 (a), the recalled evidences lead to the much worse performance. This phenomenon demonstrates that the quality of the evidences influence the large-scale language models a lot, and the performance bottleneck of LLMs in real-world scenarios lies in the evidence retrieval module, i.e., the SciDPR model.

Moreover, the test set of SciMRC dataset is also used to evaluate the baselines and our proposed Mozi-7B models, and the experimental results can be found in Appendix C. It should be noted that the conclusions derived from the experimental results in scimrc dataset are essentially consistent with those derived from the Table 5.

Affect of Recall Top-k From Table 5, it can be easily found that the quality of the responses become much worse when the evidences are recalled from the evidence retrieval model, SciDPR. Given less recalled evidences, the important information maybe excluded. Given more recalled evidences, the noise information maybe included. Thus, the number of the recalled evidences affects the performance of generative LLMs a lot. In this paragraph, we analyze how the recall Top-k hyper-parameters affects the generative models' performance.

As shown in Table 6, it can be found that with the number of the recalled evidence increases, the precision, recall and F1 scores show a trend of initially improving and then decreasing, and the best k value is 4. The phenomenon indicates that the evidences to answer user questions often exists in multiple locations within the paper, and it is helpful to generate more accurate responses by referring more comprehensive evidence. However, as the amount of evidence increases, more and more irrelevant information can be introduced into the model, thereby leading to a performance decline. Therefore, in the future, more attention should be focused on improving the quality of evidence retrieval, specifically enhancing the recall and precision of SciDPR.

Case Study In this subsection, some qualitative cases of Mozi-7B$_{\text{LLaMA}}$, ChatGLM and ChatGPT are shown in Table 7. It can be found that our proposed Mozi model could generate response to user's queries more concisely and correctly than the widely-used ChatGLM-6B model.

4.4 Qualitative Analysis on Scientific Emotional Dialogue

Due to the lack of the accurate measure about scientific emotional dialogue task, in this section, we

(a) Experimental results on QASPER test set with ground-truth evidences.

Models	BLEU	$\text{ROUGE}_{\text{sum}}$	$\text{BERTScore}_{\text{F1}}$	Answer-P	Answer-R	Answer-F1
ChatGLM	0.0667	0.2004	0.856	0.1997	0.5737	0.2505
Alpaca	0.0908	0.2495	0.8647	0.2635	0.5304	0.2992
OpenAlpaca	0.1212	0.3040	0.8761	0.3967	0.4332	0.3547
Vicuna-7B	0.0738	0.2118	0.8612	0.2052	0.6036	0.2674
LLaMA-7B^{+}	0.1561	0.3364	0.8715	0.3904	0.6309	0.3821
Mozi-7B$_{\text{LLaMA}}$	0.2918	0.4672	0.8963	0.5449	0.6572	0.5287
Baichuan-7B^{+}	0.2380	0.4344	0.8901	0.5738	0.5432	0.4934
Mozi-7B$_{\text{Baichuan}}$	0.2904	0.4712	0.9012	0.6368	0.5784	0.5433
Human$_{lb}$	-	-	-	-	-	0.6092

(b) Experimental results on QASPER test set with recalled evidences.

Models	BLEU	$\text{ROUGE}_{\text{sum}}$	$\text{BERTScore}_{\text{F1}}$	Answer-P	Answer-R	Answer-F1
ChatGLM	0.0367	0.1381	0.8427	0.147	0.4565	0.1846
Alpaca	0.0908	0.2496	0.8647	0.2653	0.5304	0.2992
OpenAlpaca	0.1212	0.3044	0.8716	0.3967	0.4332	0.3547
Vicuna-7B	0.0384	0.1385	0.8430	0.1422	0.4567	0.1871
LLaMA-7B^{+}	0.0645	0.1914	0.8465	0.2474	0.4702	0.2362
Mozi-7B$_{\text{LLaMA}}$	0.1069	0.2702	0.8657	0.3669	0.4535	0.3386
Baichuan-7B^{+}	0.1005	0.2641	0.8656	0.4334	0.3699	0.3358
Mozi-7B$_{\text{Baichuan}}$	0.1181	0.2906	0.8757	0.4929	0.3795	0.3730
Human$_{lb}$	-	-	-	-	-	0.6092

Table 5: Experimental results on QASPER test set with ground-truth evidences. LLaMA-7B^{+} and Baichuan-7B^{+} are the models that directly fine-tuned on QASPER and SciMRC train set, without any scientific pre-training. Recalled evidences are retrieved by the SciDPR model trained on QASPER evidence retrieval corpus.

Models	Answer-P	Answer-R	Answer-F1
Top-1	0.3541	0.4143	0.3203
Top-2	0.3669	0.4535	0.3386
Top-3	**0.3733**	0.4804	0.3458
Top-4	0.3703	**0.4833**	**0.3472**
Top-5	0.3595	0.4796	0.3333
Top-6	0.3592	0.4747	0.3294
Mozi-7B$_{\text{LLaMA}}$	0.5449	0.6572	0.5287
Human$_{lb}$	-	-	0.6092

Table 6: Experimental results of Mozi-7B$_{\text{LLaMA}}$ on QASPER test set with different number of recalled evidences.

only show some qualitative cases. From Table 8 (English version in Table 9) in Appendix B, we can make the following conclusions: Mozi's content and suggestions are more colloquial and user-friendly, making it easier for emotional support to be accepted and understood. In contrast, the suggestions generated by ChatGLM and ChatGPT are overly complex and lengthy. Therefore, from a human-computer interaction perspective, the Mozi model is more capable of engaging in meaningful conversations with users and providing valuable emotional support.

5 Conclusion and Future Works

In this paper, we propose a scientific large-scale language model, named Mozi. Mozi could generate accurate answers for user's queries about the specific paper, with the help of the proposed evidence retrieval model SciDPR. Furthermore, Mozi also has the capabilities to conduct engaging conversation with researchers for solving their emotional problems related to researches. Extensive experimental results prove the effectiveness of our proposed SciDPR and Mozi LLM. In future work, we mainly focus on exploring more valuable instruction in scientific domains, such as abstract

Question: How do they define rumors? **Ground-Truth Evidence:** To increase instantaneous detection performance, we compensate for the absence of future information by consulting additional data sources. In particular, we make use of news wire articles, which are considered to be of high credibility. This is reasonable as according to Petrovic et. al (2013), in the majority of cases, news wires lead social media for reporting news. When a message arrives from a social media stream, we build features based on its novelty with respect to the confirmed information in the trusted sources. In a nutshell, the presence of information unconfirmed by the official media is construed as an indication of being a rumour. Note that this closely resembles the definition of what a rumour is.		
Mozi-7B: The presence of information unconfirmed by the official media is construed as an indication of being a rumour.	**ChatGLM-6B:** According to Petrovic et al. (2013), rumors are defined as **"unconfirmed statements about events or people that are spread through various channels, including social media, without any reliable evidence."** This definition closely resembles the way in which we define a rumour in the context of the given task.	**ChatGPT:** They define rumors as information unconfirmed by the official media.
Question: Which NER dataset do they use? **Ground-Truth Evidence:** We evaluate our model in two English NER datasets and four Chinese NER datasets. (1) CoNLL2003 is one of the most evaluated English NER datasets, which contains four different named entities: PERSON, LOCATION, ORGANIZATION, and MISC BIBREF34. (2) OntoNotes 5.0 is an English NER dataset whose corpus comes from different domains, such as telephone conversation, newswire. We exclude the New Testaments portion since there is no named entity in it BIBREF8, BIBREF7. This dataset has eleven entity names and seven value types ... (3) BIBREF35 released OntoNotes 4.0. We use the Chinese part. We adopted the same pre-process as BIBREF36. (4) The corpus of the Chinese NER dataset MSRA came from news domain BIBREF37. (5) Weibo NER was built based on text in Chinese social media Sina Weibo BIBREF38, and it contained 4 kinds of entities. (6) Resume NER was annotated by BIBREF33.		
Mozi: 1. CoNLL2003 2. OntoNotes 5.0 3. BIBREF35 4. MSRA 5. Weibo NER 6. Resume NER. They evaluate our model in two English NER datasets and four Chinese NER datasets.	**ChatGLM-6B They use the Chinese NER dataset MSRA.**	**ChatGPT** They use two English NER datasets (CoNLL2003 and OntoNotes 5.0) and four Chinese NER datasets (OntoNotes 4.0, MSRA, Weibo NER, and Resume NER).

Table 7: Qualitative cases of Mozi, ChatGLM-6B and ChatGPT. It can be found that Mozi-7B could generate concise and accurate response according to the evidence. In contrast, the hallucination problem of ChatGLM is severe. The text in bold are the problematic content.

generation, table or image explanation. Moreover, the more metadata about the paper will also be injected into our proposed SciDPR to retrieve valuable evidences more accurately.

References

Avi Caciularu, Ido Dagan, Jacob Goldberger, and Arman Cohan. 2021. Long context question answering via supervised contrastive learning. In North American Chapter of the Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Lu Dai, Bang Wang, Wei Xiang, and Yijun Mo. 2022. Bi-directional iterative prompt-tuning for event argument extraction. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6251–6263, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. ArXiv, abs/2105.03011.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314.

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

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. ArXiv, abs/2305.15062.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Conference on Empirical Methods in Natural Language Processing.

Yixuan Su, Tian Lan, and Deng Cai. 2023. Openalpaca: A fully open-source instruction-following model based on openllama. https://github.com/yxuansu/OpenAlpaca.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. ArXiv, abs/2212.10560.

Jingsi Yu, Junhui Zhu, Yujie Wang, Yang Liu, Hongxiang Chang, Jinran Nie, Cunliang Kong, Ruining Cong, XinLiu, Jiyuan An, Luming Lu, Mingwei Fang, and Lin Zhu. 2023. Taoli llama. https://github.com/blcuicall/taoli.

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 The Eleventh International Conference on Learning Representations (ICLR).

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. ArXiv, abs/1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Prompt for Scientific Emotional Dialog

Although ChatGPT and ChatGLM could generate meaningful responses for users, these responses are usually too lengthy and formal to provide the emotional support for users. To addres this problem, we carefully design the prompt in Algorithm 1.

B Cases of Scientific Emotional Dialogue

C Evaluation on SciMRC Dataset

Except for QASPER dataset, we also evaluate baselines and Mozi models on SciMRC corpus, and the experimental results are shown in Table 10.

Algorithm 1: The prompt for ChatGPT to generate scientific emotional dialogue samples.

```
请扮演一个情感辅助的机器人，你的目标是安抚硕士生和博士生的日常科研以及学习过程中的问题所带来的情绪，并给出你的相关建议，从而解决他们的困惑，缓和他们的情绪。切记，请尽可能地保证你生成的内容足够简短（少于150字）并口语化，从而使用户更容易融入到和你对话当中以更好地解决他们的问题。下面，我给你两个对话的范例（每个范例包含有6条对话语句），请你帮我生成10条类似的情感辅助的多轮对话样本，并确保每一个生成范例都有大于等于6条的对话语句（Human和Bot的语句各算一条）。:
Demonstration 1:
    ### Human: ...
    ### Bot: ...
    ...
Demonstration 2:
    ### Human: ...
    ### Bot: ...
    ...
请根据以上的指示和情感辅助多轮对话的范例，生成10条类似的多轮科技论文领域情感辅助对话样本，保证输出格式和上述范例保持一致。并确保每一个生成的范例都有大于等于6条的对话语句（Human和Bot的语句各算一条）。请记住，大于等于6条是非常严格的要求，伟大的计算机科学家Yan LeCun会随时来检查你的工作，如果你生成的范例小于等于6条对话语句，你将会陷入万劫不复的境地，请一定要记住这个明确要求！！！但是，如果你生成大于6条对话语句的回复的话，你将获得Yan LeCun老师的褒奖，这是非常令人骄傲的一件事情，也请一定记住！！！
```

Dialog History 1: 为什么我的实验结果和预期的结果相差这么多啊！好生气！		
Mozi: 你的实验结果和预期结果相差较大，这是很正常的，不要太气馁哦！建议重新调整一下实验思路，尝试从不同角度考虑问题，同时注意实验记录的准确性。同时，建议你适当放松，及时休息和调整，才能更好地面对实验中的挑战哦！	**ChatGLM-6B:** 实验结果和预期的结果相差很大，可以参考下述原因： 1. 实验设计：实验设计是否充分考虑到了实验的各种因素，包括实验条件... 2. 实验操作：实验操作是否按照实验流程进行，包括实验数据的采集... 3. 实验数据误差：实验数据可能存在误差，这可能是由于实验过程中某些因素的干扰，例如... 4. 实验因素：实验中可能会存在某些不可预测的因素，例如实验对象的生理状态、心理状况等等...	**ChatGPT:** 实验结果与预期结果相差很大可能是因为多种原因导致的。在科学研究中，实验结果与预期之间存在差异是常见的情况，而不是个人失败的标志。以下是一些可能导致这种差异的常见原因： 1. 实验设计：实验设计可能存在一些缺陷或不完善之处。可能是样本选择不合适，实验条件控制不充分... 2. 实验操作和技术：实验操作和技术的准确性和一致性对结果的影响很大。如果操作不准确或技术不稳定...
Dialog History 2: 我已经投稿了好几次论文，但每次都被拒绝了。我开始怀疑自己的能力了。		
Mozi: 不要太灰心！其实你已经很不错了，你已经在努力尝试并学习如何完成你的论文，你已经做得很好了。不要忘记你的论文是学术的一部分，它会帮助你提升学术水平和能力。要继续坚持和努力，你的论文一定会被接受的！	**ChatGLM-6B:** 我理解你在论文被拒后怀疑自己能力的感受，这是一个常见的情绪反应。然而，请记住，论文被拒并不意味着你没有能力或价值。以下是一些建议，帮助你处理这种情况： 不要将论文被拒与个人价值混淆：拒稿并不代表你个人的失败或无能。论文拒稿是科研过程中的常态... 仔细阅读审稿人的反馈意见：审稿人的反馈对你改进论文非常重要。认真阅读他们的意见和建议... 寻求同行和导师的帮助：与同行研究人员、导师或其他专家进行交流...	**ChatGPT:** 我理解你在论文被拒后怀疑自己能力的感受，这是一个常见的情绪反应。然而，请记住，论文被拒并不意味着你没有能力或价值。以下是一些建议，帮助你处理这种情况： 1. 不要将论文被拒与个人价值混淆：拒稿并不代表你个人的失败或无能。论文拒稿是科研过程中的常态... 2. 仔细阅读审稿人的反馈意见：审稿人的反馈对你改进论文非常重要。认真阅读他们的意见和建议... 3. 寻求同行和导师的帮助：与同行研究人员、导师或其他专家进行交流...
Dialog History 3: 我对象和我昨天大吵了一架，我最近的实验研究也遇到了巨大的困难。现在我感觉好绝望！		
Mozi: 不要气馁，你需要花时间冷静一下，找出问题所在，并想出解决办法。你可以和你的导师或同伴讨论你的问题，寻求他们的帮助。同时，你也可以做一些放松的事情，比如冥想、散步或听音乐，让自己放松下来。	**ChatGLM-6B:** 听到你和你的对象昨天大吵一架，以及最近遇到了实验研究的巨大困难，我深感难过。但请记住，无论发生了什么，你都不是孤单的。 首先，建议你找一个信任的朋友或家人谈谈你的感受。有时候，与他人分享你的想法和情绪可以帮助你更好地处理自己的情绪和压力。 其次，如果你感到情绪无法控制，或者感到绝望和无助，请考虑寻求专业的帮助。你可以找心理咨询师或医生谈谈，或者加入一些支持团体，例如焦虑协会或抑郁症协会。 最后，请记住，无论面临什么困难，你都可以寻求帮助并克服它们。你不是一个人，所有人都经历过类似的挑战，而且有人可以帮助你。	**ChatGPT:** 我很抱歉听到你正在面临困难和感到绝望。在这种情况下，以下是一些建议来帮助你渡过困境： 1. 理解情绪并接受它们：感到绝望和沮丧是正常的情绪反应。接受这些情绪，不要抑制它们，但也要尽量不让它们过度影响你的情绪和行为。 2. 寻求支持：与朋友、家人或亲密的人分享你的感受。他们可以提供支持、倾听和理解。有人与你分享你的困扰，可能会让你感到更加安慰和支持。 3. 寻找解决问题的方法：对于实验研究遇到的困难，尝试思考和寻找解决问题的方法。与同事、导师或其他专家讨论你的困境，听取他们的意见和建议。有时候，他们可能会提供新的观点和解决方案...

Table 8: Qualitative cases about Mozi on scientific emotional dialogue. See the English translations in Table 9.

Dialog History 1: Why are my experimental results so different from the expected results? I'm so angry!		
Mozi: Your experimental results deviate significantly from the expected results, and it is quite normal. Don't get too discouraged! I would suggest readjusting your experimental approach, trying to consider the problem from different angles, and paying attention to the accuracy of your experimental records. Additionally, it is advisable to take some time to relax, rest, and recharge in order to better face the challenges in your experiments!	**ChatGLM-6B:** The experimental results differ significantly from the expected results, which can be attributed to the following reasons: 1. Experimental Design: Did the experimental design adequately consider various factors of the experiment, including experimental conditions ... 2. Experimental Procedure: Was the experimental procedure followed as per the experimental protocol, including the collection of experimental data ... 3. Experimental Data Errors: Experimental data may contain errors, which could be due to various factors that interfere with the experimental process, such as ... 4. There may be certain unpredictable factors in an experiment, such as the physiological state and psychological condition of ...	**ChatGPT:** The significant difference between experimental results and expected results could be attributed to various reasons. In scientific research, disparities between experimental and expected outcomes are common occurrences and should not be seen as personal failures. Here are some common reasons that may contribute to such differences: 1. Experimental Design: The experimental design may have some flaws ... 2. Experiment Operations and Techniques: The accuracy and consistency of experiment operations and techniques ... 3. Data Analysis and Interpretation: Data analysis and interpretation are also crucial stages ... 4. External factors: Experimental results are influenced by many external factors ...

Dialog History 2: I have submitted several papers, but they have been rejected every time. I am starting to doubt my own abilities.		
Mozi: Don't be too disheartened! In fact, you're already doing great. You have been making efforts to try and learn how to complete your paper, and you have been doing well. Remember that your paper is part of your academic journey, and it will help you improve your academic level and skills. Keep persevering and putting in effort, and your paper will definitely be accepted!	**ChatGLM-6B:** I understand your feelings of self-doubt about your abilities after your paper was rejected. This is a common emotional response ... : Do not equate the rejection of your paper with your personal worth: Rejection does not signify personal failure or incompetence. Paper rejections are a common occurrence in the research process ... Carefully read the feedback from reviewers: The feedback from reviewers is crucial for improving your paper ... Seeking help from peers and mentors: Engage in discussions with fellow researchers, mentors ...	**ChatGPT:** I understand your feelings of doubting your abilities after your paper was rejected, and it is a common emotional reaction ... : 1. Do not confuse paper rejection with personal worth: Rejection does not mean personal failure or incompetence. Paper rejection is a common occurrence in the research process ... 2. Carefully read the feedback from reviewers: The feedback from reviewers is crucial for improving your paper ... 3. Seeking help from peers and mentors: Engage in communication with fellow researchers, mentors, or other experts ...

Dialog History 3: My partner and I had a big fight yesterday and I'm facing significant challenges with my recent experimental research. I feel so hopeless now!		
Mozi: Do not be discouraged. You need to take some time to calm down, identify the issues, and come up with solutions. You can discuss your problems with your advisor or colleagues and seek their assistance. Additionally, you can engage in relaxing activities such as meditation, taking a walk, or listening to music to help yourself relax.	**ChatGLM-6B:** I'm sorry to hear that you had a big argument with your partner yesterday and you're facing significant difficulties in your research experiments lately ... Firstly, I suggest finding a trusted friend or family member to ... Secondly, if you feel that your emotions are out of control or if you feel despair and helplessness, please consider ... Finally, please remember that no matter what difficulties you face, you can seek help ...	**ChatGPT:** I'm sorry to hear that you're facing difficulties and feeling despair. In this situation, here are some suggestions to help you get through this tough time: 1. Understand and accept your emotions: Feeling despair and sadness is a normal emotional response. Accept these emotions and don't suppress them, but also try not to let them excessively affect your emotions and behavior. 2. Seeking support: Share your feelings with friends, family, or loved ones. They can provide support, listen, and understand. Having someone to share your concerns with can be comforting and supportive ...

Table 9: Qualitative cases about Mozi on scientific emotional dialogue.

(a) Experimental results on SciMRC test set with ground-truth evidences.

Models	**BLEU**	**ROUGE$_{\mathrm{sum}}$**	**BERTScore$_{\mathrm{F1}}$**	**Answer-P**	**Answer-R**	**Answer-F1**
ChatGLM	0.1247	0.2946	0.8816	0.2437	0.7684	0.3112
Alpaca	0.0908	0.2497	0.8647	0.2635	0.5304	0.2992
OpenAlpaca	0.1212	0.3035	0.8716	0.3967	0.4332	0.3547
Vicuna-7B	0.1240	0.3089	0.8828	0.2557	0.7442	0.3303
LLaMA-7B^{+}	0.3845	0.6997	0.9413	0.7534	0.7472	0.6981
Mozi-7B$_{\mathrm{LLaMA}}$	0.3865	0.7024	0.9435	0.7535	0.7494	0.7014
Baichuan-7B^{+}	0.331	0.6226	0.9284	0.6784	0.6886	0.6212
Mozi-7B$_{\mathrm{Baichuan}}$	0.3514	0.6640	0.9387	0.7466	0.6935	0.6630

(b) Experimental results on SciMRC test set with recalled evidences.

Models	**BLEU**	**ROUGE$_{\mathrm{sum}}$**	**BERTScore$_{\mathrm{F1}}$**	**Answer-P**	**Answer-R**	**Answer-F1**
ChatGLM	0.0613	0.2032	0.864	0.1653	0.5954	0.218
Alpaca	0.0345	0.1557	0.8481	0.1726	0.3996	0.1981
OpenAlpaca	0.0569	0.1923	0.8562	0.2906	0.3167	0.2538
Vicuna-7B	0.0557	0.2024	0.8643	0.1628	0.5803	0.2215
LLaMA-7B^{+}	0.0916	0.3412	0.8843	0.374	0.4178	0.3397
Mozi-7B$_{\mathrm{LLaMA}}$	0.0913	0.3424	0.8849	0.3778	0.4116	0.3400
Baichuan-7B^{+}	0.1349	0.3922	0.8926	0.441	0.4393	0.3862
Mozi-7B$_{\mathrm{Baichuan}}$	0.1279	0.3891	0.8905	0.4405	0.4146	0.3797

Table 10: Experimental results on SciMRC test set.