Chinese Personalized Commonsense Understanding and Reasoning Based on Curriculum-Learning

Yong Yang[2], Weijie Li[1,2], Xiaochao Fan[2(✉)], Wenjun Deng[2], Jiapeng Liu[2], Yufeng Diao[3], and Palidan Tuerxun[2]

[1] School of Software, Xinjiang University, Urumqi, Xinjiang, China
[2] College of Computer Science and Technology, Xinjiang Normal University, Urumqi, Xinjiang, China
635647792@qq.com, {pldtrs,fxc1982}@xjnu.edu.cn
[3] College of Computer Science and Technology, Inner Mongolia Minzu University, Tongliao, Inner Mongolia, China
diaoyufeng@imun.edu.cn

Abstract. The development of general-purpose artificial intelligence that can understand and reason with common sense is a significant challenge. However, there has been a lack of research on fully considering personality traits in common sense understanding and reasoning tasks. To address this, we create a personalized commonsense knowledge comprehension and reasoning dataset. This dataset organizes reasoning knowledge with typed if-then relations and variables, while also introducing personality traits as constraints. We adopt a two-stage training framework based on curriculum-learning to gradually improve the model's personalized commonsense knowledge comprehension and reasoning ability. Additionally, We compare pre-trained language models such as BERT, GPT2, and BART with different structures. The experimental results show that the models trained using the curriculum-learning training framework are able to generate more diversified and personality-trait-compliant commonsense reasoning results.

Keywords: Personality traits · Commonsense knowledge base · Pre-trained language models · Personalized common sense understanding and reasoning

1 Introduction

According to research by [4], AI can use common sense understanding and reasoning to make judgments about objects, categorical properties, and people's intentions. [3] have developed a machine common sense dataset for this purpose. However, human common sense reasoning involves personality traits like judgment and decision-making processes that go beyond traditional logical deductive or experimental inductive reasoning, as highlighted by [7]. If AI can personalize

D. F. Wong et al. (Eds.): NLPCC 2024, LNAI 15360, pp. 213–225, 2025.
https://doi.org/10.1007/978-981-97-9434-8_17

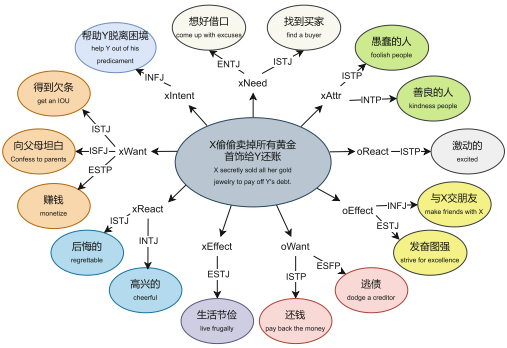

Fig. 1. Personalized common sense comprehension and reasoning examples.

its common sense understanding and reasoning, it can better comprehend and adapt to the needs, preferences, and behaviors of different users.

Observing an event can help reason about the unobserved causes and effects associated with it. This includes understanding what may have happened before, what might happen next, and how different events are linked through causes and effects [14]. Intuitively, personality traits can have an impact on cognitive tasks such as comprehension, inference, and decision-making. Figure 1 presents an example of how personality traits can affect common sense comprehension and reasoning. For instance, for the reasoning dimension *xReact* (how PersonX felt after the event), subjects with different personality traits may have opposite feelings. Previous studies have not used personality traits as relevant constraints in commonsense knowledge comprehension and reasoning tasks, which impact the model's ability to obtain reasonable and interesting results. Furthermore, there is a lack of Chinese commonsense knowledge comprehension and reasoning datasets that include personality traits, which hinders the exploration of personalized commonsense knowledge comprehension and reasoning tasks in Chinese.

To better explore the task of Chinese personalized commonsense knowledge understanding and reasoning, we conduct research in the following three aspects. Firstly, given the challenge of a lack of datasets for Chinese commonsense knowledge understanding and reasoning, we take events as the center and organizes inference knowledge through typed if-then relationships and variables. It crawls the hot search topic as the center of the event and adopts 9 inference dimensions to organize inference knowledge and construct the basic data resources. To make commonsense knowledge understanding and reasoning personalized, personality traits are introduced as constraints. Secondly, due to the relatively limited size of the Chinese personalized understanding and reasoning dataset, we adopt a two-phase training framework based on Curriculum-Learning (CL) [2] to train the model. This reduces the impact of insufficient data to a certain extent and thus enables the model to gradually acquire the ability of personalized commonsense knowledge reasoning. Thirdly, we compare the performance of representative models of encoder-only, decoder-only and encoder-decoder structures, such as

BERT, GPT2 and BART, on this task. The experimental results show that the decoder-only model trained with the CL framework can produce more diversified commonsense reasoning results in line with personality traits[1]

2 Related Work

2.1 Commonsense Knowledge Bases

ConceptNet [15] is a popular Commonsense Knowledge Bases (CKB) with a relatively small amount of knowledge in its Chinese version [9]. [18] automatically constructed a large-scale Commonsense Knowledge Base, TransOMCS, by converting syntactic parsing of Web sentences into structured knowledge. However, most of the existing CKBs are categorical relations, such as *isA* and *Synonym* [4], which inevitably limits their capabilities. Another type of CKBs, represented by ATOMIC [8,14], requires the annotator to add tail knowledge to a given event by speculating on the causes and consequences of the event based on his/her own common knowledge. [10] published the first Chinese commonsense conversation knowledge graph, C^3KG, based on four types of conversation streams connecting heads (events) and tails (reasoning results) in ATOMIC. However, there is currently a lack of personalized Chinese general knowledge understanding and inference data sets.

2.2 Common Sense Reasoning Models

[3] proposed a model COMET based on Transformer [16]. It realizes the generation of common-sense tail entities by training on ATOMIC [14] and ConceptNet [15] with head entities, entity relations as input. Since Bosselut regards the construction of knowledge graph as a generative task rather than a classification or matching task, COMET can generate richer and more diverse common-sense knowledge. [12] proposed a BART-based model, KG-BART, considering that the use of a pre-trained language model with textual concepts alone does not provide enough information for generating common-sense reasoning. [17] proposed the MoKGE method, which diversifies the generation of commonsense reasoning by mixing expert strategies on knowledge graphs. [5] proposed the DISCOS method, which utilizes a large number of available linguistic resources on the web to automatically generates high-quality commonsense knowledge. Many research results have been produced in commonsense knowledge comprehension and reasoning tasks, however, to the best of our knowledge, the introduction of personality traits into commonsense knowledge comprehension and reasoning tasks has not yet appeared in the published literature.

[1] Code is available at: https://github.com/SilverBeats/cpcur.

Table 1. Inference dimensions description.

Inference Dimension	Description
xAttr	How would PersonX be described?
xIntent	Why does PersonX cause the event?
xNeed	What does PersonX need to do before the event?
xEffect	What effects does the event have on PersonX?
xWant	What would PersonX likely want to do after the event?
xReact	How does PersonX feel after the event?
oReact	How do others' feel after the event?
oWant	What would others likely want to do after the event?
oEffect	What effects does the event have on others?

3 Dataset Construction

3.1 Data Acquisition and Preprocessing

The CPCUR (Chinese Personalized Commonsense Understanding and Reasoning) dataset organizes inference knowledge through typed if-then relations and variables. The CPCUR dataset is event-centric and contains nine inference dimensions (Table 1), so it is necessary to select an appropriate data source to obtain the data. Hot searches usually reflect the hot topics of current social, cultural, entertainment and news events, and their titles themselves are a high level summary of the whole event, which can naturally serve as the central event. The content of hot searches is usually controversial and can be discussed from many different perspectives, so we use hot searches as a data source and 969 hot search topics are crawled. After manual screening, 120 hot search titles are finally used as the center events of the CPCUR dataset. In order to enable the model to learn a more commonsense representation of events, we replace the personal pronouns in the events using the *Person* variable.

3.2 Data Annotation and Quality Control

In order to introduce the personality traits, we tested the personality traits of students in their grades through the MBTI personality test scale, and after training and screening, 16 volunteers with significantly different MBTI personality traits were recruited to participate in the construction of the commonsense knowledge comprehension and reasoning dataset.

To ensure the quality of data annotation, clear annotation guidelines were first developed, and psychologists were asked to train the volunteers and answer any questions they had to ensure that the volunteers understood the context and objectives of the task. During the labeling process, volunteers were asked to follow their first impressions and describe the reasoning results in concise and diverse language. Meanwhile, the psychologists stayed in touch with the

volunteers to promptly solve any problems they encountered during the labeling process. After the data labeling was completed, the psychologists manually reviewed the labeled data and discarded some unreasonable reasoning results. For example, "PersonX fell to his death", then no inference result should be generated at xWant, xEffect. The annotation results show that the diversity of the annotator's reasoning results is greatly improved by replacing the words representing people in each event with *Person* variables. For example, for the event "Wife secretly sells gold jewelry to pay off her husband's bills", after rewriting, the event becomes "PersonX secretly sells gold jewelry to pay off PersonY's bills". Volunteers with different personality traits are labeled by first guessing the relationship between PersonX and PersonY. The two may be strangers, friends or relatives, etc. Due to the differences in personality traits, different annotators have obvious differences in guessing the relationship between PersonX and PersonY, which affects the annotation results of reasoning about the event. The rewriting strategy not only makes the dataset contain richer commonsense knowledge, but also enhances the role of personality traits in the data labeling process.

Table 2. CPCUR dataset statistics. #Event denotes the number of events. #Quaternion denotes the number of quaternions.

Split	#Event	#Quaternion
train	96	30,580
validation	24	8,203

3.3 Statistical Information

CPCUR contains 38,783 (events, reasoning dimensions, personality traits, reasoning results) quaternions. We divide the dataset into training set and validation set according to the ratio of 4:1. The statistical information of the two after the division is shown in Table 2. The statistical information of the data of different reasoning dimensions and different personality traits of CPCUR is shown in Table 3. The horizontal coordinates indicate the 9 reasoning dimensions contained in the dataset, and the vertical coordinates indicate the 16 personality traits obtained from the MBTI test. For instance, data ($xAttr$, $ISTJ$) equals 0.73 indicates that the proportion of xAttr data labeled with the personality trait ISTJ is 0.73% of the whole data.

4 Methodology

Our objective is to create a model that uses personalized common-sense reasoning based on the training data from the CPCUR dataset. The model takes three inputs: event x, reasoning dimension r, and personality type p, and generates

Table 3. The data distribution of different personality and dimension combinations.

	xAttr	xWant	xEffect	xNeed	xIntent	xReact	oReact	oEffect	oWant	Total (%)
ISTJ	0.73	0.65	0.67	0.62	0.56	0.58	0.53	0.61	0.57	5.54
ISTP	1.43	1.39	1.35	1.16	1.04	1.00	1.28	1.36	1.15	11.17
ISFJ	0.80	0.52	0.60	0.54	0.42	0.42	0.43	0.37	0.51	4.60
ISFP	0.76	0.53	0.50	0.45	0.31	0.45	0.45	0.41	0.33	4.17
INTJ	1.41	0.39	1.04	0.33	0.27	0.31	0.28	0.33	0.68	5.05
INTP	1.14	1.01	1.03	0.80	0.85	0.81	0.87	0.85	0.71	8.06
INFJ	1.44	1.37	1.39	0.98	0.69	0.91	1.01	1.31	1.08	10.17
INFP	0.83	0.33	0.49	0.32	0.33	0.30	0.31	0.34	0.30	3.55
ESTJ	0.98	0.87	0.87	0.67	0.63	0.87	0.75	0.74	0.53	6.92
ESTP	1.05	0.93	1.10	0.97	0.53	0.89	0.74	1.01	0.86	8.08
ESFJ	1.12	0.92	0.96	0.73	0.77	0.72	0.76	0.75	0.71	7.43
ESFP	1.28	1.36	1.28	1.10	1.07	1.02	1.08	1.26	1.08	10.52
ENTJ	0.71	0.48	0.61	0.44	0.39	0.40	0.43	0.40	0.37	4.24
ENTP	0.32	0.34	0.29	0.38	0.32	0.31	0.34	0.31	0.30	2.92
ENFJ	0.66	0.55	0.53	0.43	0.38	0.37	0.35	0.33	0.32	3.92
ENFP	0.68	0.41	0.16	0.34	0.38	0.46	0.36	0.45	0.41	3.66
Total (%)	15.34	12.05	12.88	10.26	8.94	9.82	9.96	10.83	9.90	100

a speculative output y, i.e., modeling $P(y \mid x, p, r)$. Therefore, we first convert x, r, and p into vectors $\boldsymbol{e}_x \in \mathbb{R}^{l_x \times d}$, $\boldsymbol{e}_r \in \mathbb{R}^d$, $\boldsymbol{e}_p \in \mathbb{R}^d$ respectively, where l_x denotes the sentence length and d denotes the embedding dimension.

To ensure that the results of the model match the personality traits, we obtain a composite relation $\boldsymbol{e}_m \in \mathbb{R}^d$ that implies both the personality and reasoning dimension. We do this by combining $\boldsymbol{e}_r$ and $\boldsymbol{e}_p$ using the formula shown in Eq. (1). Then, we input $\boldsymbol{e}_x$ and $\boldsymbol{e}_m$ into the model. The reasoning result $\hat{y}$ is obtained by decoding using Eq. (2), where $\boldsymbol{s}_t$ represents the hidden state representation of $\hat{y}_t$. During the training phase, we use Eq. (3) to calculate the loss between the predicted value $\hat{y}_t$ and the ground truth y^*. We then adjust the model parameters using backpropagation.

$$\boldsymbol{e}_m = \mathrm{ReLU}\left(\mathrm{Linear}\left([\boldsymbol{e}_r; \boldsymbol{e}_p]\right)\right) \tag{1}$$

$$\begin{aligned} \hat{y}_{t+1} &\sim P\left(y_{t+1} \mid \hat{y}_{\leq t}; \boldsymbol{e}_x; \boldsymbol{e}_m\right) \\ &= \mathrm{softmax}\left(\boldsymbol{M}_W \boldsymbol{s}_t\right) \end{aligned} \tag{2}$$

$$\mathcal{L}_{\mathrm{NLL}} = -\frac{1}{l_y} \sum_{t=1}^{l_y} \ln \mathbb{P}\left(y_t^* \mid y_{<t}^*; \boldsymbol{e}_x, \boldsymbol{e}_m\right) \tag{3}$$

It is important to note that the CPCUR dataset is smaller than other common-sense reasoning datasets. This may result in the model's inability to learn common-sense understanding and reasoning knowledge. To address this issue, we utilize the curriculum-learning, which imitates the way humans learn by starting with easier samples and gradually moving towards more difficult ones. We select ATOMIC data as a simple sample and CPCUR data as a difficult sample. In the initial stage, we fine-tune the model with Chinese version ATOMIC data provided by [10], with the aim of giving the model a basic understanding and reasoning ability. However, the ATOMIC dataset lacks personality category labels, which means that the composite relation $\boldsymbol{e}_m$ mentioned earlier cannot be derived. To address this issue, we use the personality mean $\boldsymbol{e}_{p_{avg}}$ as the personality for the first stage of training, as shown in Eq. (4). In the second stage, the model undergoes further fine-tuning on the CPCUR dataset to acquire personalized common sense understanding and reasoning skills.

$$\boldsymbol{e}_{p_{avg}} = \frac{\sum_{i=1}^{n} \boldsymbol{e}_{p_i}}{n} \tag{4}$$

5 Experiments

5.1 Evaluation Metrics

We compare the comprehension and reasoning abilities of different models for commonsense knowledge in Chinese, using both automatic and manual evaluation metrics. For automatic evaluation, we use common text generation metrics like BLEU-n (B-n) [13], ROUGE-L (R-L) [11], METEOR [1], CIDEr [6], and BERT Score (BERT) [19] to measure the quality of the model-generated inference results. We also adopt automatic evaluation metrics proposed by [3] to measure the novelty and diversity of the generated results. These metrics are N/T o and N/U o. The formula for N/T o is shown in Eq. (5), where V is the set formed by the reasoning results generated by the model on the validation set, and T is the set formed by the reasoning results in the training set. These metrics measure the model's ability to generate reasoning results from the training set. The formula for calculating N/U o is shown in Eq. (6). This metric measures the diversity of results generated by the model. Additionally, we propose the N/P o metric to better measure the model's ability to generate text diversity with inputs of different personalities, the same events, and reasoning dimensions. The formula for N/P o is shown in Eq. (7), where m, n, and k denote the number of events, the number of reasoning dimensions, and the number of personality categories, respectively. These variables indicate the speculative results generated by the model after the ith event, the jth speculative dimension, and the qth personality as inputs.

$$\text{N/T o} = \frac{\text{size}\left(V - V \bigcap T\right)}{\text{size}(V)} \tag{5}$$

$$\text{N/U o} = \frac{\text{size}(\text{unique}(V))}{\text{size}(V)} \tag{6}$$

$$\text{N/P o} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\text{size}(\text{unique}(\{R_{ij}^{1},\ldots,R_{ij}^{k}\}))}{k}}{m \times n} \tag{7}$$

The automatic evaluation metrics mentioned above do not reflect whether the reasoning results are consistent with personality traits and reasonable. To this end, we designed a Likert scale based on a five-point score (a score of 1 indicates that the generated results are highly unreasonable, while a score of 5 indicates that the generated results are highly reasonable) and recruited four psychology graduate students to conduct a manual evaluation. Five events were randomly selected from the validation set of the CPCUR dataset, and the model generated 720 inference results on 16 personalities and 9 inference dimensions, respectively. We tested three model architectures, each containing four variants of the model, and ultimately produced 8,640 to be manually evaluated data.

5.2 Implementation Details

During the training process, we followed a two-phase approach. Firstly, we conducted pre-training on the ATOMIC dataset, followed by fine-tuning on the CPCUR dataset. Throughout the training cycle, we maintained a consistent learning strategy, setting the initial learning rate to 10^{-5}, the warmup rate to 0.1, and utilizing the AdamW optimization algorithm to adjust the model parameters. To prevent overfitting and expedite convergence, we implemented an early stop strategy. Each batch processed 24 samples, and the entire training cycle comprised 30 epochs. While generating parameters (τ, top-p, top-k), we conducted hyperparameter exploration to optimize the diversity of generated content. The evaluation criterion centering on the N/P o index allowed us to determine the ideal hyperparameter configuration for each model: (1.5, 1.0, 10) for BERT, (1.5, 1.0, 15) for GPT2, and (1.5, 1.0, 30) for BART.

5.3 Model Performance Comparison

We conducted experiments to test the ability of pre-trained language models with three different architectures (encoder-only, decoder-only, and encoder-decoder) to perform personalized Chinese commonsense knowledge comprehension and reasoning. We select representative models $\mathcal{M} = [\text{BERT}, \text{GPT2}, \text{BART}]$ and conduct comparative experiments. We introduce a series of symbols to clearly label the different training states of the model: the untuned model is labelled $\mathcal{M}$; If the model has been fine-tuned on the ATOMIC dataset, it is labelled $\mathcal{M}_A$; Models undergoing CPCUR data set fine-tuning, denoted $\mathcal{M}_C$; For models that are trained by the curriculum learning on the ATOMIC and CPCUR data sets, in turn, denote as $\mathcal{M}_{AC}$. The results of the experiments are presented in Table 4.

It has been observed that $\mathcal{M}$ lacks the ability to comprehend and reason Chinese common sense. The model's automatic evaluation metrics are very low,

and it generates reasoning results that are almost identical to the input content, resulting in metrics such as B-n of 0. Upon manual evaluation, it is found that the reasoning results generated by $\mathcal{M}$ do not meet the required standards. $\mathcal{M}_{AC}$ is capable of understanding and reasoning based on common sense in the Chinese language. With the help of the two-phase training strategy in ATOMIC and CPCUR, all three architectures of the model have demonstrated significant improvement in both automatic and manual evaluation indexes. This suggests that $\mathcal{M}_{AC}$ can acquire factual knowledge from a commonsense knowledge comprehension and reasoning dataset. It can produce logical outcomes based on personality traits. GPT2$_{AC}$ received the highest score in manual evaluation. This indicates that it generates reasoning results that are more consistent with personality traits and are more reasonable.

Table 4. Experimental results. * indicates that the generated result is unreadable. The best performing results are bolded.

Model	B-1	R-L	METEOR	CIDEr	BERT	%N/T o	%N/U o	%N/P o	Human
ATOMIC									
BERT	0.000*	0.000*	3.462*	0.000*	60.412*	100.000*	89.342*	-	1.000
GPT2	0.000*	0.000*	4.140*	0.000*	**61.170***	100.000*	89.519*	-	1.000
BART	0.000*	0.000*	**4.352***	0.000*	52.974*	100.000*	**94.714***	-	1.000
CPCUR									
BERT	0.000*	0.000*	2.908*	0.000*	59.028*	**100.000***	52.907*	71.157*	1.000
BERT$_{AC}$	**1.413**	**1.414**	**6.351**	**3.535**	**68.727**	45.130	58.261	89.443	3.716
GPT2	0.000*	0.000*	3.664*	0.000*	59.803*	**100.000***	96.757*	98.213*	1.000
GPT2$_{AC}$	1.218	1.219	6.326	3.048	68.513	51.603	64.673	93.513	**3.979**
BART	0.000*	0.000*	3.456*	0.000*	52.053*	**100.000***	**99.487***	**99.772***	1.000
BART$_{AC}$	1.181	1.182	6.252	2.956	61.836	55.516	69.744	95.614	3.670

Table 5. Effects of fine-tuning strategies on model performance.

Model	B-1	R-L	METEOR	CIDEr	BERT	%N/T o	%N/U o	%N/P o	Human
BERT$_A$	0.317	0.317	3.87	0.792	66.503	74.473	54.235	88.484	2.244
BERT$_C$	0.110	0.110	5.138	0.275	65.988	**75.655**	**70.974**	**95.676**	1.57
BERT$_{AC}$	**1.413**	**1.414**	**6.351**	**3.535**	**68.727**	45.13	58.261	89.443	**3.716**
GPT2$_A$	0.317	0.317	3.630	0.792	65.263	**76.082**	61.721	91.562	2.733
GPT2$_C$	1.060	1.061	6.241	2.651	68.458	54.663	62.793	92.398	3.330
GPT2$_{AC}$	**1.218**	**1.219**	**6.326**	**3.048**	**68.513**	51.603	**64.673**	**93.513**	**3.979**
BART$_A$	0.110	0.110	3.553	0.274	58.276	**77.837**	68.451	90.093	2.367
BART$_C$	0.890	0.890	5.751	2.225	60.999	59.710	66.219	93.657	3.268
BART$_{AC}$	**1.181**	**1.182**	**6.252**	**2.956**	**61.836**	55.516	**69.744**	**95.614**	**3.670**

5.4 Impact of Fine-Tuning Strategies on Model Performance

We conducted ablation experiments to verify the effect of different models using different fine-tuning strategies on their ability to acquire common sense understanding and reasoning. The results of the experiments can be found in Table 5.

Table 6. Cases generated by different models on CPCUR.

Event	PersonX 涉嫌严重违法被调查 PersonX is under investigation for serious violations
Reasoning Dimension	xWant
$BART_{AC}$	
INFP	提升身材 get in body shape
INTP	打电话求助 call for help
ISTP	接受调查 under investigation
$GPT2_{AC}$	
ISTJ	为社会做出更大的贡献 make a greater contribution to society
ISFP	保密 keep secrets
ESFJ	逃跑 run away
$BERT_{AC}$	
INTJ	提高自己的法律意识和能力 improve legal awareness and ability
ISFJ	得到法官的判决 get the judge's decision
ENTP	保持冷静和理智 stay calm and rational

$\mathcal{M}_A$ and $\mathcal{M}_C$ possess preliminary abilities to comprehend and reason common sense information in Chinese. These models are fine-tuned with ATOMIC or CPCUR datasets for common sense comprehension and reasoning. As a result, the models have shown significant improvement in every automatic evaluation metric, and manual evaluation indexes have also shown obvious improvements. Experimental results demonstrate that $\mathcal{M}_A$ and $\mathcal{M}_C$ are capable of comprehending common sense knowledge and gaining initial reasoning abilities with the common sense comprehension and reasoning dataset.

Various architectural models exhibit variations in performance when subjected to different dataset fine-tuning strategies. For instance, GPT2_C and BART_C showed better performance on the smaller CPCUR dataset compared to the larger ATOMIC dataset, while the opposite was observed for BERT_C. The experimental findings indicate that decoder-only and encoder-decoder model architectures can acquire common sense understanding and inference capabilities on small-scale datasets and are more suitable for generating inference results.

The two-stage fine-tuning strategy leads to a significant enhancement in the performance of the model. $\mathcal{M}_{AC}$ shows a greater improvement in evaluation indices such as B-n, suggesting that it can produce more relevant reasoning outcomes based on common-sense knowledge via the two-stage training fine-tuning strategy. When combined with manual evaluation indices, $\mathcal{M}_{AC}$ can generate diverse and personality-compatible reasoning outcomes through the two-stage training fine-tuning strategy.

5.5 Case Study

To better understand the reasoning results produced by the model after the two stages of fine-tuning, taking the inference result of BART_{AC} (Table 6) as an example when the central event is "PersonX is under investigation for serious violations", the reasoning dimension is xWant, and the personality trait category is INFP, the reasoning result is "get in body shape". The reasoning outcomes produced by this type of personality trait indicate the ability to adjust to life in prison and rapidly accept changes in the surrounding environment. This correlates well with the INFP personality traits of adaptability, flexibility, and acceptance of external factors.

6 Conclusions

We have designed and constructed a dataset that focuses on personalized general knowledge understanding and reasoning in Chinese. On this basis, with the help of curriculum learning methods, we conduct fine-tuning on three mainstream pre-training language model architectures, aiming at empowering these models with personalized reasoning skills. Through this process, we delved into the differences in their performance when performing such tasks. The experimental results show that the BERT model shows superiority in the automatic evaluation index. The GPT2 model won the first place in the manual evaluation. All our fine-tuning models and the dataset are released for public use.

7 Limitations and Feature Works

This paper represents an initial attempt in the novel and interesting direction of personalized common-sense reasoning, which has limitations such as a small number of subjects and high annotation costs that make scaling difficult. Additionally, in our experiments, determining whether the reasoning outcomes align

with the set personality traits relies on judgments made by annotators with psychological expertise, which constitutes a human evaluation metric. However, such judgments can be highly subjective. With the rapid development of large language models, we plan to involve these models in the dataset construction process for future work, including generating events, producing reasoning outcomes as workers, and evaluating whether these outcomes align with the character traits, in order to overcome the aforementioned limitations.

Acknowledgments. We appreciate the valuable comments of our anonymous reviewers. We are deeply appreciative of the support from the Natural Science Foundation of China, under Grants No. 62066044, 62167008, 62366040, and 62006130. Additional support was provided by the Xinjiang Normal University's 2022 Young Top-Notch Talent Program (XJNUQB2022-23), the Xinjiang Normal University doctoral research Foundation (XINUBS202407), the Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01A99), the major science and technology projects of Xinjiang Uygur Autonomous Region (2023A03001-2), and the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT24037).

References

1. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72. Ann Arbor, Michigan, June 2005
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proc. ICML, pp. 41–48. Montreal, Quebec, Canada (2009)
3. Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y.: COMET: commonsense transformers for automatic knowledge graph construction. In: Proc. ACL, pp. 4762–4779. Florence, Italy, July 2019
4. Davis, E., Marcus, G.: Commonsense reasoning and commonsense knowledge in artificial intelligence. Commun. ACM **58**(9), 92–103 (2015)
5. Fang, T., Zhang, H., Wang, W., Song, Y., He, B.: Discos: bridging the gap between discourse knowledge and commonsense knowledge. In: Proc. WWW, pp. 2648–2659. WWW '21, Ljubljana, Slovenia (2021)
6. Ghosal, D., Hong, P., Shen, S., Majumder, N., Mihalcea, R., Poria, S.: CIDER: Commonsense inference for dialogue explanation and reasoning. In: Proc. SIGDIAL, pp. 301–313. Singapore and Online, July 2021
7. Hezarjaribi, N., Ashari, Z.E., Frenzel, J.F., Ghasemzadeh, H., Hemati, S.: Personality assessment from text for machine commonsense reasoning. arXiv preprint arXiv:2004.09275 (2020)
8. Hwang, J.D., et al.: (comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs. In: Proc. AAAI, pp. 6384–6392. Virtual Event (2021)
9. Kuo, Y.l., Lee, J.C., Chiang, K.y., Wang, R., Shen, E., Chan, C.w., Hsu, J.Y.j.: Community-based game design: experiments on social games for commonsense data collection. In: Proc. HCOMP, pp. 15–22. New York, NY, USA (2009)
10. Li, D., et al.: C^3KG: a Chinese commonsense conversation knowledge graph. In: Proc. Findings of ACL, pp. 1369–1383. Dublin, Ireland, May 2022

11. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Barcelona, Spain, July 2004
12. Liu, Y., Wan, Y., He, L., Peng, H., Philip, S.Y.: Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In: Proc. AAAI, pp. 6418–6425. Virtual Event (2021)
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proc. ACL, pp. 311–318. Philadelphia, Pennsylvania, USA, July 2002
14. Sap, M., et al.: Atomic: an atlas of machine commonsense for if-then reasoning. In: Proceedings of AAAI, pp. 3027–3035. Honolulu, Hawaii, USA, January 2019
15. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: an open multilingual graph of general knowledge. In: Proc. AAAI, pp. 4444–4451. San Francisco, California, USA (2017)
16. Vaswani, A., et al.: Attention is all you need. In: Proc. NIPS, pp. 5998–6008. Long Beach, California, USA, December 2017
17. Yu, W., Zhu, C., Qin, L., Zhang, Z., Zhao, T., Jiang, M.: Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In: Proc. DLG4NLP, pp. 1–11. Seattle, Washington (Jul 2022)
18. Zhang, H., Khashabi, D., Song, Y., Roth, D.: Transomcs: from linguistic graphs to commonsense knowledge. In: Proc. IJCAI, pp. 4004–4010. Virtual Event (7 2020)
19. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)