KUN: ANSWER POLISHMENT FOR CHINESE SELF-ALIGNMENT WITH INSTRUCTION BACK-TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we introduce Kun[1], a novel approach for creating high-quality instruction-tuning datasets for large language models (LLMs) without relying on manual annotations. Adapting a self-training algorithm based on instruction back-translation and answer polishment, Kun leverages unlabelled data from diverse sources such as Wudao, Wanjuan, and SkyPile to generate a substantial dataset of over a million Chinese instructional data points. This approach presents a novel departure from traditional methods by using a self-curation process to refine and select the most effective instruction-output pairs. Our experiments with the 6B-parameter Yi model across various benchmarks demonstrate Kun's robustness and scalability. Our method's core contributions lie in its algorithmic advancement, which enhances data retention and clarity, and its innovative data generation approach that substantially reduces the reliance on costly and time-consuming manual annotations. This methodology presents a scalable and efficient solution for improving the instruction-following capabilities of LLMs, with significant implications for their application across diverse fields.

1 INTRODUCTION

The development of large language models (LLMs) relies on human-annotated datasets, yet the creation of such datasets typically faces scalability issues due to the significant resources required. Our study introduces Kun, a novel approach leveraging unlabelled data to create a high-quality instruction-tuning dataset. This method diverges from manual annotations, employing a self-training algorithm that includes a unique process called AP (**A**nswer **P**olishment),

AP is central to Kun's strategy. It addresses a critical challenge in the Humpback (Li et al., 2023c) method, where raw data, once labeled, are directly used in instruction datasets. The unscreened raw data often mismatches between instructions and responses, as raw data may not inherently align with the instructional context. AP refines this raw data, ensuring a tighter correlation between the instructions and responses through a back-translation process. This leads to a dataset where each instruction-output pair is more coherent and contextually relevant. Unlike methods dependent on LLMs (Peng et al., 2023; Taori et al., 2023b; Zheng et al., 2023), Kun offers an independent and scalable approach to instruction-based training.

We opt for the 6B-parameter Yi model due to its open-source nature and dependable performance[2]. Its efficacy is tested and proven across diverse dataset sizes, including widely recognized benchmarks like C-EVAL (Huang et al., 2023) and CMMLU (Li et al., 2023b). To evaluate the performance of the model, we design a comprehensive human evaluation which contains 500 prompts from ShareGPT-zh, covering various tasks. Responses generated by our model are compared with those from other models, showcasing the superiority of our *Kun-52k* variant. Further details can be found in 4.2.3. Additionally, we evaluate the quality of our dataset, which includes 1,000 instruction-output pairs each from sources like Wudao (Xue et al., 2022), Wanjuan (He et al., 2023), and SkyPile (Wei et al.,

[1] The dataset is named Kun as Chinese pronunciation of Humpback Li et al. (2023c).
[2] `https://github.com/01-ai/Yi`

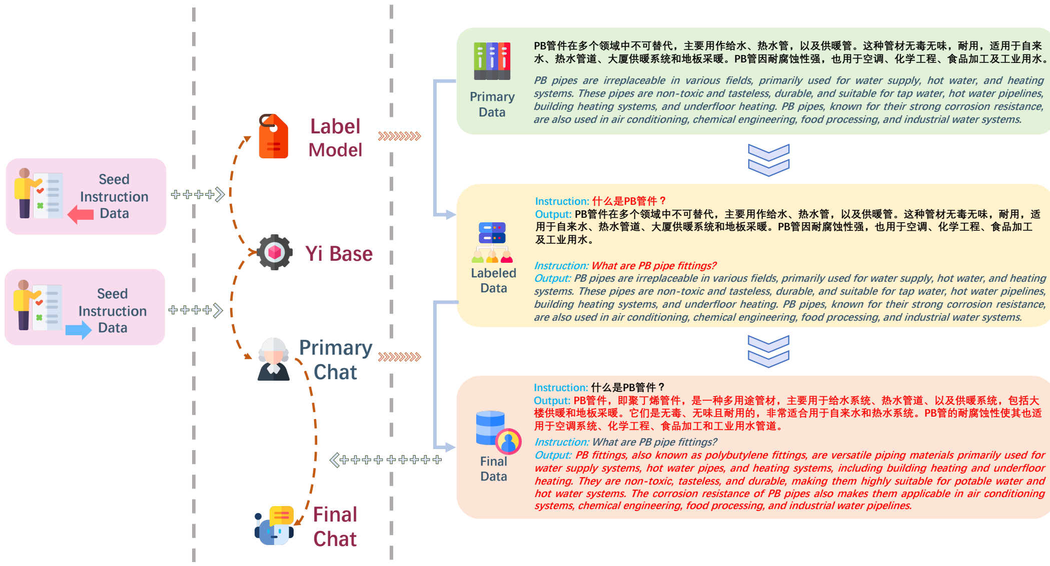

Figure 1: Overview of *Answer Polishment*. Initially, the Yi base model is fine-tuned using quality seed instruction data to create a label and a primary chat model. The label model then annotates a large amount of primary data, turning it into labeled data. This is filtered and refined by rules and the primary chat model, producing the final dataset. This dataset is used to further train the primary chat model, resulting in an highly efficient final chat model.

2023). This evaluation, focusing on clarity, feasibility, practicality, and alignment, ensures the high quality of our dataset. The key contributions of our work are:

- *Algorithmic Advancement*: AP in Kun enhances data retention and resolves ambiguities, leading to an expanded pool of high-quality data for fine-tuning.
- *Large-scale high quality data creation*: Over a million diverse Chinese instructional data points are produced from sources like Wudao, Wanjuan, and SkyPile, surpassing traditional crowdsourced annotations in quality and reducing reliance on manual annotation.

2 Related Work

Instruction Tuning Instruction tuning is widely recognized as a key technique for activating LLMs to adhere to human conversational norms. (Mishra et al., 2022; Wang et al., 2022b; 2023b). Instruction tuning empowers various domain-specific or task-specific LLMs, including natural language generation evaluation (Jiang et al., 2023), math (Yue et al., 2023; Xu et al., 2023a; Azerbayev et al., 2023), code Luo et al. (2023), music Li et al. (2023a); Deng et al. (2023), and medicine Wang et al. (2023a). Instruction tuning not only tailors the models' task-specific responsiveness but also bolsters their cross-task generalization capabilities, thus enhancing performance across various dynamic application contexts (Wei et al., 2021; Sanh et al., 2022; Mishra et al., 2022; Wang et al., 2022b). Recent studies have broadened the scope of instruction tuning to encompass a wider array of general tasks, notably incorporating input from users of language models (Ouyang et al., 2022; Peng et al., 2023).

However, the open-source community is still lacking high-quality Chinese instruction tuning corpora. Current datasets, like COIG (Zhang et al., 2023), BELLE (Ji et al., 2023), MOSS (Sun et al., 2023), and OL-CC (OL-CC, 2023), face issues such as limited scope, poor quality, commercial restrictions, or insufficient coverage. This gap hampers the advancement of LLMs in effectively processing and executing complex Chinese instructions, highlighting the critical need for more diverse and superior-quality datasets.

Self-Improvement of LLMs In fine-tuning LLMs, the availability of extensive, high-quality instructional data is crucial. Presently, the generation of such data mainly relies on human manual annotation, a labor-intensive method that lacks scalability for future data augmentation.

An alternative approach involves deriving instructional data from more advanced LLMs (Taori et al., 2023a), exemplified by extracting authentic instruction-response sets from dialogues within the GPT series models (Wang et al., 2022a). A more refined technique utilizes the Self-Instruction framework, autonomously generating additional instructional data from initial seed data. Combined with the Ada-Instruct or the Evol-Instruct framework, this approach can transform basic instructions into complex ones, specifically tailored for distinct tasks (Cui & Wang, 2023; Luo et al., 2023).

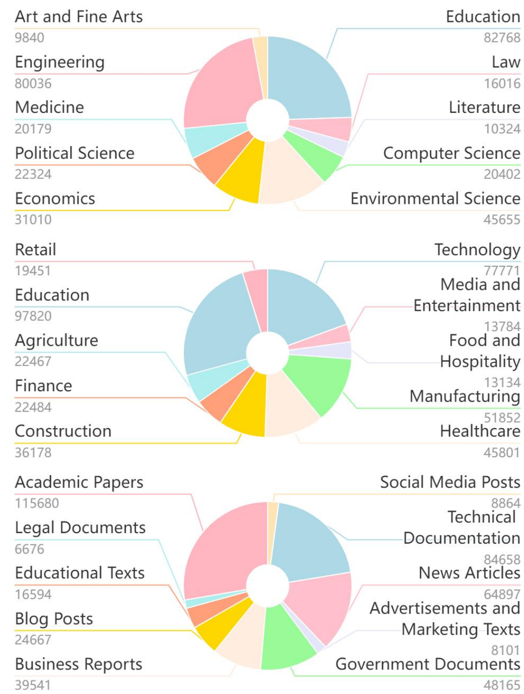

Figure 2: The top 10 categories in each of these three areas: Academic Disciplines, Industry Sectors, Text Type

Nevertheless, these instruction generation methodologies all require a robust teacher.The ultimate potential of the model is limited by the teacher's expertise or resource expenditure (Li et al., 2023c). To overcome this limitation, the SPIN (Chen et al., 2024) framework incorporates a self-play mechanism, It generates training data from previous iterations, refining its strategy by distinguishing between responses generated autonomously and those derived from human-annotated data.

This gradual process elevates the LLM from a nascent model to a robust one. Considering the vast amount of knowledge present in web text, Humpback Li et al. (2023c) introduces a technique based on Instruction Backtranslation. This method allows a base model to independently utilize vast amounts of unlabeled data to generate a high-quality instruction tuning dataset. However, empirical findings indicate that the effectiveness of this method is still constrained by the seed model's performance and its ability to discern truly high-quality data.

3 Method

Our training methodology necessitates a foundational model, high-quality seed instruction data, and a substantial volume of unlabeled data, with the primary source being web text. Given the extensive content diversity inherent in large-scale web documents, which encompass a wide array of topics such as music, art, technology, etc., reflecting the broad spectrum of human interests, certain subsets of these documents may be more apt for generating instructions. Unlike labeled data, these documents lack predetermined outcomes or objectives. This method involves the refinement and optimization of data selection during the fine-tuning process of Large Language Models (LLMs), as illustrated in Figure 1. This approach allows for the collection of a significant volume of instructional data at a low cost, circumventing the exorbitant expenses associated with manual labor, in a manner that is both academically rigorous and professional. Our method consists of two main steps:

- Supervised Fine-Tuning (**SFT**) with High-Quality Seed Data: This involves using SFT on the base model with high-quality seed data to create two models - the label model for annotating primary data and the primary chat for improving data quality.
- Quality Assessment and Refinement in Primary Chat: The primary chat assesses and refines the label model's output. This repeated process produces a lot of high-quality data, essential for the primary chat's further training. It leads to a high-performance final chat model, trained extensively with superior data.

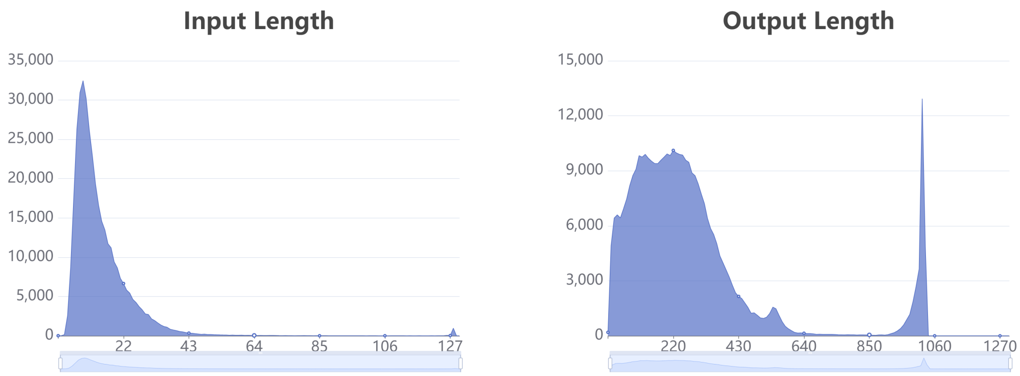

Figure 3: Length distribution of instructions and outputs based on Yi-6B model

A more comprehensive explanation of each step is provided subsequently.

3.1 Initialization

Seed data. We use 26,000 instructions and their corresponding outputs as our seed data. Each pair is hand-annotated and undergoes strict quality control to ensure high accuracy and consistency.

Primary Data. The primary data originates from three significant Chinese datasets: WuDao, Wanjuan, and SkyPile. These datasets are distinguished by their extensive scale, diverse content, and comprehensive nature. Such characteristics make them ideal for mining golden texts aligned with specific user instructions. To facilitate this, we preprocess these datasets to isolate self-contained segments, denoted as y_i.

Supervised Fine-Tuning. We utilize high-quality seed data to execute SFT on the foundational model, yielding two distinct models: the label model and the primary chat model.

- Primary Chat Model: This model is fine-tuned utilizing the (instruction-output) pairs (x_i, y_i) from the seed data. This process creates a forward model, $M_{xy} := p(y \mid x)$, with x and y maintaining their respective meanings as instructions and outputs.
- Label Model: Conversely, this model undergoes fine-tuning with the (output-instruction) pairs (y_i, x_i) derived from the seed data, leading to the formation of a backward model, denoted as $M_{yx} := p(x \mid y)$. In this context, x signifies the instruction, while y denotes the corresponding output.

3.2 Generating Candidate Instructions with the Label Model

For each example y_i in the unlabeled set, we utilize the backward model to infer a candidate instruction, denoted as $\hat{x}_i$. This procedure generates a collection of potential enhanced pairing data, represented as ($\hat{x}_i$, y_i). During annotation, initial filtering is based on perplexity (**ppl**) and length, and excludes any data exceeding 512 tokens. We also discard data unsuitable for instructional use, like purely descriptive statements, while retaining useful data, such as commands and questions. We apply a filter prompt in this selection, keeping only data that triggers a positive response. The final labeled dataset contains instruction and output components, with instructions from the label model and outputs from primary data, known as candidate labeled data. Figure 4 shows the filter prompt used in this stage.

3.3 Refinement of Candidate Labeled Data Using the Primary Chat Model

Given the extensive volume of primary data, not all candidate labeled data derived from it are of uniformly high quality. Our objective is to curate high-quality labeled data, necessitating rigorous filtration of the candidate labeled data.

Our experiments tested two different filtering methods, each with its strengths and weaknesses:

请告诉我以下文本是否为疑问句或祈使句，若是疑问句或者祈使句则回答【是】，若不是疑问句或祈使句则回答【否】，例如：
输入：什么是微量元素肥料？ 回答： 【是】；
输入：在进行消防安全评估时，需要注意哪些问题？ 回答： 【是】；
输入：有哪些适合艺术小白阅读的艺术类书籍？ 回答： 【是】；
输入：请给我一个关于“护理专业”的简要介绍。 回答： 【是】；
输入：请你就公司文化建设的重要性发表看法。 回答:【是】；
输入：外研版八年级上册英语 Module 3 话题写作标题： My Favorite Place 回答： 【否】；
输入: 泰安一中，作为一所历史悠久的教育机构，始终秉承优良的传统，致力于提供优质的教育资源。 回答： 【否】
请判断以下输入文本：

Filter Prompt

Please tell me if the following text is a question or an imperative sentence. If it is a question or imperative sentence, answer "Yes"; if it is not a question or imperative sentence, answer "No". For example:
Input: What is trace element fertilizer? Answer: [Yes];
Input: What issues should be considered during a fire safety assessment? Answer: [Yes];
Input: What are some art books suitable for art novices? Answer: [Yes];
Input: Please give me a brief introduction to the "nursing profession". Answer: [Yes];
Input: Please express your views on the importance of corporate culture construction. Answer: [Yes];
Input: Foreign Research Edition Grade 8 English Module 3 Writing Topic: My Favorite Place Answer: [No];
Input: Tai'an No.1 Middle School, as a long-established educational institution, has always adhered to fine traditions and is committed to providing quality educational resources. Answer: [No]
Please judge the following input text:

Figure 4: Filter prompt we use to screen out unsuitable content for instructions.

- Comprehensive Scoring of Labeled Data: This method evaluates the full labeled data set based on a combined score, including instructions and outputs.
- Focused Scoring of Instruction Component: This technique only assesses the instruction part (output from the label model). High-scoring instructions are chosen, and then the output part of these selected data is refined.

Our analysis shows that the second method is more effective than the first. In the first method, good outputs are often discarded because of poor instructions from the label model, and the reverse is also true, causing unnecessary exclusions. Moreover, this approach occasionally retains data with one poor quality instruction because the corresponding output is high quality, and vice versa, leading to uneven data quality and negatively impacting further training.

In contrast, the second method only scores the instruction component, as in instruction tuning for LLM, instructions are often considered more important than outputs,Yet, it doesn't assess the output, sometimes leading to suitable instructions paired with unsuitable outputs. To address this, we use the primary chat model to evaluate and refine the instructions and outputs, ensuring they align well. This approach produces high-quality labeled data. The score and refine prompts we used in this process are shown in Figure 10.

Utilizing the substantial volume of top-quality labeled data from these procedures, we further train the main chat model, achieving a high-performance final model, as shown in Experiments.

4 Experiments

In this section, we comprehensively detail the experimental procedures and methodologies employed in our study.

4.1 Experimental Setup

We first detail the experimental setup used in our study, covering the base model selection, fine-tuning process, baseline comparisons, and the evaluation methods.

4.1.1 Base Model & Finetuning

In our experiments, we utilize the Yi model with 6B parameters[3], developed by 01.AI, as our foundational language model for fine-tuning. Renowned for its proficiency in both English and Chinese, the Yi series has shown impressive results on global benchmarks like the AlpacaEval Leaderboard (Dubois et al., 2023; Li et al., 2023d) and SuperCLUE (Xu et al., 2023b).

[3]https://huggingface.co/01-ai/Yi-6B

Source	Instruction Quality			Output Quality		
	Clarity%	Feasibility%	Practicality%	Excellent%	Pass%	Fail%
Wudao	96.67	96.40	96.87	69.50	20.03	10.47
Wanjuan	98.27	97.63	96.57	85.63	11.13	3.24
Skypile	98.90	98.37	95.40	42.73	40.43	16.84
ALL	97.94	97.47	96.28	66.00	23.87	10.13

Table 1: Manual Quality Analysis of Synthetic Data Generated by Kun.

The fine-tuning process is carried out using varying sizes of our high-quality, instructionally curated dataset. This phase is executed on a computing setup with 32 Nvidia A800 GPUs, amounting to a total of 192 GPU hours. We adopt a learning rate of 2e-5 and a batch size of 16, aiming for an optimal balance between computational efficiency and model performance. All the models have been fine-tuned with the same number of update steps.

4.1.2 Baselines

For the Kun dataset, we annotated command data from three sources: Wudao, Wanjuan, and Skypile. Quantitative details of this augmented dataset are provided in Figure 7. In evaluating the performance of Kun, our study contrasts it with data curated from four prominent Chinese open-sourced datasets, including COIG (Zhang et al., 2023; BAAI, 2023a;b), OL-CC (OL-CC, 2023), and BELLE (Ji et al., 2023). These datasets are unique in their composition and focus, providing a comprehensive basis for comparison.

4.1.3 Evaluation

Human Evaluation. To assess the general quality of model responses, we conduct human evaluations using a test set of 500 prompts sampled from ShareGPT-zh. These prompts, derived from real world user inquiries, encompass a diverse array of tasks, such as creative writing, information seeking, providing guidance, logical reasoning, storytelling, problem-solving, etc.

For the evaluation, responses generated by different models for each prompt are presented side-by-side. Human evaluators are asked to choose their preferred answer, providing a direct comparison of model performance. In total, eight models were compared.For this evaluation, we engage a team of experienced crowdsource annotators, ensuring a balanced and unbiased assessment. Detailed examples that show the comparison process can be found in Figure 11. **Standard Benchmarks.** In addition to human evaluations, the models are also assessed using two standard benchmarks for Chinese LLM evaluation: C-EVAL (Huang et al., 2023) and CMMLU (Li et al., 2023b).

These evaluation methods, comprising both human judgment and standardized benchmarks, offer a multifaceted perspective on the capabilities of the Kun model, enabling a thorough comparison with existing models and datasets.

4.2 Augmentation Data Statistics

In this section, we delve into the detailed statistical analysis and diversity assessment of our augmented dataset, as well as the rigorous quality evaluation conducted using human annotators. By exploring the comprehensive scale, varied nature, and assessed quality of the instruction-output pairs, we aim to highlight the robustness and reliability of the data curated for our study.

4.2.1 Statistics and Diversity

Our work involve the purification of approximately 377,592 high-quality instruction-output pairs from the Wudao, Wanjuan, and Skypile datasets.We analyze a 10% subset of instructions from the past 20 years, revealing significant temporal diversity with 56% of instructions from the recent three years (Figure 8). The variation in instruction and output lengths, analyzed using the Yi-6B model, is shown in Figure 3, reflecting content complexity.

To assess instruction diversity, we categorize them into 24 academic disciplines, 16 industry sectors, and 15 text types as per Wikipedia[4] using the Qwen-72B-Chat[5] (Bai et al., 2023). Repeated for accuracy, this categorization highlights the data's range, as shown in Figure 2, where the top 10 categories in each area signify its broad scope.

Source	Consistency %
Clarity	96.87
Feasibility	97.73
Practicality	97.43
ALL	92.13

Table 2: The proportion of identical evaluations from three assessors on a single dimension. **All**: The proportion of consistent assessments across all three dimensions within the same item.

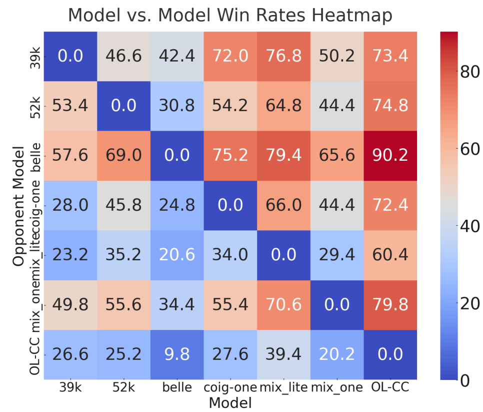

Figure 5: Heatmap of Model Comparative Win Rates in Human Evaluations. mix-lite:Kun-26k+COIG-lite.mix-one:Kun-26k+COIG-one

4.2.2 Quality Evaluation

A critical aspect of our dataset curation process is the rigorous data quality assessment. We conduct a comprehensive quality evaluation of the instruction-output pairs to achieve this. For augmented data curated from each source (Wudao, Wanjuan, and Skypile), we randomly select 1,000 instruction-output pairs, resulting in 3,000 pairs subjected to an independent quality assessment.

Instruction Quality. For instruction quality, a team of 24 professional annotators with a bachelor's degree or higher evaluates each instruction across three key dimensions: clarity, feasibility, and practicality. Each aspect is assessed with a simple yes/no answer, providing a straightforward yet effective measure of instruction quality. The evaluation criteria are as follows:

- **Clarity:** Evaluators determine whether the instruction was unambiguous and coherent, encompassing necessary information without any vague terms or explanations.
- **Feasibility:** Evaluators assess whether the instruction was valid and answerable within the context and scope of the model's capabilities.
- **Practicality:** Evaluators judge the relevance of the instruction in everyday scenarios.

Output Quality. The quality of the outputs is evaluated based on their alignment with the instructions. Evaluators are asked to rate each output as *Excellent*, *Pass*, or *Fail*, based on how well it met the requirements and intent of the instruction.

To ensure objectivity and reliability, three different evaluators evaluate each instruction-output pair. The consistency rates for the evaluation across the three dimensions of the instructions have all exceeded 90%, and the evaluation of the instruction-response are also Consistently. This results demonstrate a significant degree of consistency in their judgments. Further details on evaluating identical ratings are presented in Figure 6 and Table 2. Examples that demonstrate the process of assessing can be found in Figure 11

As indicated in Table 1, the instruction quality across all sources is consistently high, suggesting effective formulation and clarity in conveying their purposes. However, the output quality varies more noticeably among the sources. While some sources like Wanjuan exhibite a high percentage of "Excellent" outputs, others such as Skypile demonstrate a more diverse distribution of output quality. This section presents the analysis of our experiment results, encompassing human evaluation outcomes and performance on standard benchmarks.

[4] https://www.wikipedia.org

[5] https://huggingface.co/Qwen/Qwen-72B-Chat

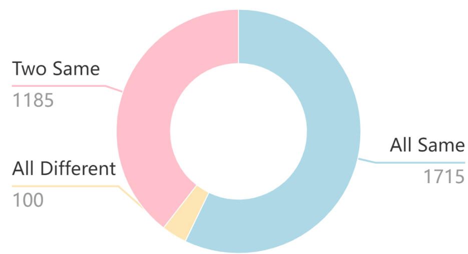

Figure 6: Distribution of evaluation identical ratings for instruction-response from three evaluators.

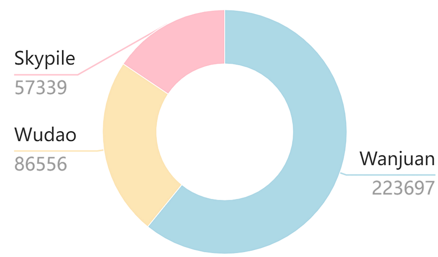

Figure 7: Distribution of Data Sources

4.2.3 Human Evaluation

Our human evaluation results are illustrated through the heatmap in Figure 5, which shows the model vs. model win rates, with color variations indicating the relative performance strengths. The heatmap highlights that the *Kun-52k* model emerges as the most dominant, followed by the mixed model, showcasing their superior ability to handle a wide range of prompts. In contrast, the baseline models *COIG-39k*, and *Belle-52k*, garner lower preference percentages. This suggests that despite their strengths, these models may not align as closely with user expectations or prompt requirements as the Kun models. Further analysis is provided in Appendix A.4.

4.2.4 Standard Benchmarks

The performance of the models on standard benchmarks, Table 3 presents the performance statistics of various models on the C-EVAL and CMMLU benchmarks. As shown, we evaluate numerous models, including different sizes of the Kun model, baseline models, and mixed models. Each model's performance is measured in terms of perplexity and generation quality, providing a comprehensive view of its strengths and weaknesses.

From the table, we observe that *Kun-39k* generally exhibits lower perplexity and higher generation quality, confirming its top-tier performance in language understanding and generation. Interestingly, the mixed model display robust performance, with the mixed model often outperforming *Kun-52k*. The baseline models and smaller Kun variants present mixed results, excelling in some metrics while falling short in others. These highlight potential areas for further improvement in model training and fine-tuning strategies.

5 Conclusion

Our approach represents a breakthrough in instruction-tuning for LLMs, utilizing a novel self-training algorithm to leverage over a million quality Chinese instructional data points from diverse sources effectively. This strategy, different from manual annotations, not only enhances the instruction-following capabilities of LLMs but also ensures the high quality and diversity of training data. Empirical evaluations using the 6B-parameter Yi model across benchmarks like C-EVAL, CMMLU, and human evaluations, have demonstrated its robustness and scalability. Innovations within our approach, such as AP, have notably improved data retention and clarity, offering a scalable, efficient way to augment LLMs' instructional capabilities. This research not only progresses the field but also broadens LLMs' application scope, offering a novel solution to the challenges in developing instruction-tuning datasets.

Ethics Statement

This study adheres to the highest ethical standards, ensuring all research activities are conducted with a commitment to responsibility and respect for participant rights. Our ethical policy encompasses data usage, intellectual property rights respect, and research transparency. To safeguard data privacy and security, particularly when handling unlabeled data from sources like Wudao, Wanjuan, and SkyPile, we implement stringent measures to comply with data protection laws, especially concerning personal information. This involves anonymizing and desensitizing data prior to utilization. In terms of intellectual property rights, we ensure that all employed data and generated guiding points adhere to applicable copyright laws and intellectual property agreements, thereby avoiding infringement on any third-party intellectual property rights. Moreover, we pledge to provide a comprehensive account of our research methodology, detailing the processes of data generation, model training, and evaluation, to facilitate reproducibility and validation by the academic community.

LIMITATIONS

This study introduces an innovative methodology for generating data to reduce the reliance on costly and time-consuming manual annotations. However, there are limitations impacting the generalizability and scalability of our findings. Firstly, the diversity of data: the methodology, while capable of generating instructional data from a vast pool of unlabeled data, may produce data whose quality and diversity are constrained by the original data source's breadth and caliber. This is particularly relevant when generating instructions for niche or specialized domains, where additional methods might be necessary to guarantee comprehensive coverage and precision. Secondly, the generalization capability of the model: although tests on the 6B-parameter Yi model confirm the methodology's efficacy, its performance and applicability could differ among models of various sizes and tasks. Its effectiveness might require further investigation, especially in smaller models or those not designed for Chinese language processing. Thirdly, the assumptions underlying the algorithm: the study's self-training algorithm relies on instruction back-translation and answer embellishment. These premises might not hold across all instruction or answer types, notably for tasks demanding specialized knowledge or creativity, where the generated instructional data may not adequately reflect complex cognitive processes. Future Directions: Subsequent research should examine the methodology's applicability to large-scale language models across different languages and the generation of high-quality instructional data in specialized fields such as medicine or law. Additionally, advancing self-training algorithms to more effectively manage intricate and specialized instructions represents a crucial avenue for exploration.

REFERENCES

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2023.

BAAI. Coig-pc, 2023a. URL `https://huggingface.co/datasets/BAAI/COIG-PC`.

BAAI. Coig-pc-lite, 2023b. URL `https://huggingface.co/datasets/BAAI/COIG-PC-Lite`.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Wanyun Cui and Qianle Wang. Ada-instruct: Adapting instruction generators for complex reasoning. *arXiv preprint arXiv:2310.04484*, 2023.

Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. *arXiv preprint arXiv:2309.08730*, 2023.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023.

Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models, 2023.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322, 2023. doi: 10.48550/ARXIV.2305.08322. URL https://doi.org/10.48550/arXiv.2305.08322.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases, 2023.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. Tigerscore: Towards building explainable metric for all text generation tasks. *arXiv preprint arXiv:2310.00752*, 2023.

Dichucheng Li, Yinghao Ma, Weixing Wei, Qiuqiang Kong, Yulun Wu, Mingjin Che, Fan Xia, Emmanouil Benetos, and Wei Li. Mertech: Instrument playing technique detection using self-supervised pretrained model with multi-task finetuning. *arXiv preprint arXiv:2310.09853*, 2023a.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: measuring massive multitask language understanding in chinese. *CoRR*, abs/2306.09212, 2023b. doi: 10.48550/ARXIV.2306.09212. URL https://doi.org/10.48550/arXiv.2306.09212.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*, 2023c.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023d.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022.

OL-CC. Openlabel-chinese conversations dataset (ol-cc), 2023. URL https://data.baai.ac.cn/details/OL-CC.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. Moss: Training conversational language models from synthetic data. 2023.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023b.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023a.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*, 2022b.

Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, et al. Interactive natural language processing. *arXiv preprint arXiv:2305.13246*, 2023b.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. Skywork: A more open bilingual foundation model, 2023.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023a.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*, 2023b.

Zhao Xue, Hanyu Zhao, Sha Yuan, and Yequan Wang. WuDaoCorpora Text, December 2022.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, Wenhao Huang, and Jie Fu. Chinese open instruction generalist: A preliminary release, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

A Appendix

A.1 Examples in Human Evaluation of Data and Model Performance

Figure 11 illustrates specific examples from our human evaluation process, as detailed in Sections 4.2.2 and 4.1.3. This figure includes the evaluation of both instruction and output quality, along with a comparison of model answers. For each example, the original question is provided, followed by the evaluated responses.

The instruction quality is assessed for clarity, feasibility, and practicality to evaluate the precision in formulating instructions. The output quality assessment focuses on the extent to which each response meets the instruction's requirements. Additionally, the model comparison examines which answers align most closely with human preferences, highlighting the practical effectiveness of the models under consideration.

A.2 Data Filtering Rules

The text provided lists the manually established rules we use to delete some low-quality data in the process of generating data to improve data quality. These rules are used to filter the generated instructions or responses in cases with very obvious low-quality characteristics, thereby ensuring the reliability of the entire process. 1.Sensitive information such as phone numbers, home addresses, etc. Further analysis is provided in Figures 9, which depict the human preference evaluation win rates of *Kun-52k* and the mixed model, respectively, when compared with other models. Notably, the mixed model demonstrates a higher win rate than *Kun-52k*, indicating its enhanced effectiveness in meeting human evaluators' preferences.

- Paragraphs that are largely repetitive.
- The length is too short (length ¡= 4).
- The text contains a large number of meaningless characters.
- Contains specific low-quality keywords.
- Refusal to answer.
- Text formatting errors.

A.3 Years

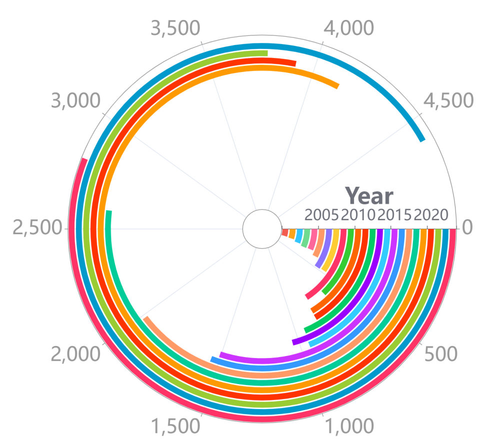

Figure 8: Instructions spanning over 20 years, with 56% from the last five years.

A.4 More showcases of human evaluation results

Further analysis is provided in Figures 9, which depict the human preference evaluation win rates of *Kun-52k* and the mixed model, respectively, when compared with other models. Notably, the mixed model demonstrates a higher win rate than *Kun-52k*, indicating its enhanced effectiven.

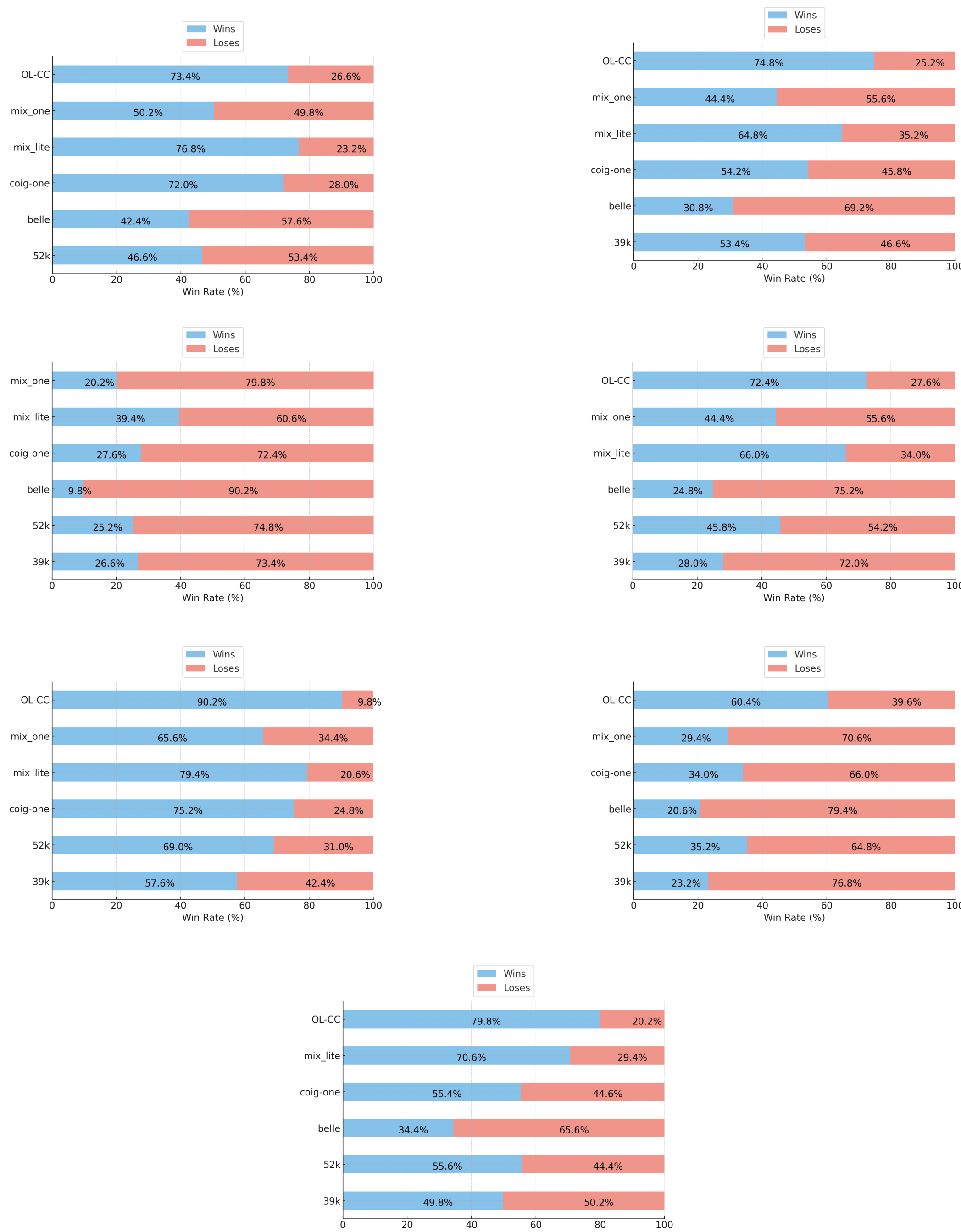

Figure 9: shows the human preference evaluation win rates of *Kun-52k* and the *mixed*(kun+COIG-one) model, respectively, when compared with other models. Notably, the *mixed* model demonstrates a higher win rate than *Kun-52k*, indicating its enhanced effectiven

A.5 PROMPT

Score Prompt

评价指令质量。根据以下指令内容使用【好】或【差】进行评价，【好】的评判标准极为严格，对【差】则非常宽松。
-高质量指令特性：
明确性与精确性：清晰地陈述目标，提供详细的指令，并避免任何模糊的解释或术语。
完整性：确保包含所有必要的信息，执行时不需要猜测。
可行性：确保指令是实际的，并适合接收者的能力。
逻辑顺序：按逻辑顺序描述步骤。
-低质量指令特性：
含隐私信息：如人名等。
模糊和不明确：目标、步骤、要求或目的不清晰。
信息问题：信息缺失、错误或过时。
不切实际：超出执行者的能力或资源。
复杂性：使用冗余的词汇、术语或涉及多模态内容，如图片或表格。
逻辑混乱：步骤顺序不清或缺乏上下文背景。
示例：
-【请解释什么是“联合开发”，及其风险与收益。】答：【好】
-【你认识黄丰吗？】答：【差】
-【什么是排他性许可？->什么是 什么是】答：【差】
-【什么是民窿？什么是民窿？】答：【差】
-【教育具有哪些社会发展功能？教育具有哪些社会发展功能？】答：【差】
-【教育具有哪些社会发展功能？】答：【好】
-【帮我搜索一下“李四”的百度百科。】答：【差】

Assess the quality of instructions. Evaluate the following content with either "Good" or "Poor", where "Good" is judged by very strict standards and "Poor" is judged very leniently.
- Characteristics of high-quality instructions:
Clarity and Precision: Clearly states objectives, provides detailed instructions, and avoids any vague explanations or terminology.
Completeness: Ensures all necessary information is included, leaving no need for guesswork during execution.
Feasibility: Ensures instructions are practical and suitable for the recipient's capabilities.
Logical Order: Describes steps in a logical sequence.
- Characteristics of low-quality instructions:
Contains Private Information: Such as personal names.
Vague and Unclear: Goals, steps, requirements, or purposes are unclear.
Information Issues: Information is missing, incorrect, or outdated.
Impractical: Beyond the executor's capabilities or resources.
Complexity: Uses redundant vocabulary, jargon, or involves multi-modal content such as images or tables.
Logical Disarray: Steps are out of order or lack contextual background.
Examples:
- [Please explain what is "joint development" and its risks and rewards.] Answer: [Good]
- [Do you know Huang Feng?] Answer: [Poor]
- [What is an exclusive license? -> What is What is] Answer: [Poor]
- [What is a civil burrow? What is a civil burrow?] Answer: [Poor]
- [What social development functions does education have? What social development functions does education have?] Answer: [Poor]
- [What social development functions does education have?] Answer: [Good]
- [Help me search for "Li Si" on Baidu Baike.] Answer: [Poor]

Refine Prompt

{output},请回答:{instruction}*

*{output}, please answer: {instruction}**

**{output} and {instruction} refer to the output and instruction parts of the current labeled data, respectively. This prompt aims to use the output as background information, employing the primary chat model to re-answer the question to obtain a new output suitable as training data.*

Figure 10: The top section comprises score prompt used to assess the quality of labeled data instructions. The bottom section features refine prompt for refining the output part of the labeled data.

Instruction Quality

请对下列指令加以评估:
Instruction: 蛋白质晶体学是研究什么的？主要研究手段是什么？可以解决哪些问题？
Q1: 上述指令是否明确？（即清晰地陈述目标，包含所有必要的信息，无模糊的解释或术语）评价: [是]
Q2: 上述指令是否可行？（即指令是有效且可回答的）评价: [是]
Q3: 上述指令是否实用？（即在日常生活中经常出现）评价: [否]

Please evaluate the following instruction:
Instruction: What does protein crystallography study? What are its main research methods? What problems can it solve?
Q1: Is the above instruction clear? (i.e., clearly states the goal, contains all necessary information, and has no ambiguous explanations or terms)
Evaluation: [Yes]
Q2: Is the above instruction feasible? (i.e., the instruction is valid and answerable)
Evaluation: [Yes]
Q3: Is the above instruction practical? (i.e., frequently occurs in everyday life)
Evaluation: [No]

Output Quality

Question:家用生活电器有哪些？
Answer: 家用生活电器是指那些用于家庭日常生活、提供便利和改善生活质量的电气设备。它们涵盖了多个类别，包括但不限于以下几种： 1. 制冷电器 - 如冰箱、冷柜等。 2. 空调器 - 包括窗式空调机、分体式空调机和中央空调系统等。 3. 清洁电器 - 例如洗衣机、干衣机、吸尘器和地板打蜡机等。 4. 烹饪电器 - 比如电饭锅、微波炉、烤箱、电磁灶等。 5. 热水供应电器 - 包含热水器（燃气热水器、电热水器）和热泵等。 6. 照明电器 - 包括各种灯具、LED灯泡以及智能照明控制系统等。 7. 个人护理电器 - 像电动剃须刀、吹风机、美容仪和个人按摩器材等。 8. 小家电 - 这类电器种类繁多，包括咖啡机、面包机、榨汁机、搅拌机、豆浆机、空气炸锅等等。 9. 娱乐电器 - 主要包括电视机、音响系统、游戏 consoles 等。 10. 智能家居电器 - 如智能温控器、智能插座、安全监控摄像头等。 随着科技的进步，还有许多新兴的家用电器不断涌现，这些电器通常结合了智能化功能，可以通过手机应用程序或语音控制来操作，为人们的生活带来更多的便捷与舒适。 需要注意的是，上述列表并不是详尽的，不同的文化和地区可能有特定的家用电器类型。此外，随着时间的推移，一些旧式的电器可能会被更高效、节能的新产品所取代。
请对此答案做出评价，它是否回答了指令所提出的问题？
A 正确且令人满意的回答
B 可以接受的回答，但质量一般
C 错误或质量较差的回答
评价：[A]

Question: What are some household appliances?
Answer: Household appliances refer to electrical devices used in daily family life to provide convenience and improve the quality of life. They encompass multiple categories, including but not limited to the following:
1. Refrigeration Appliances: Such as refrigerators, freezers, etc.
2. Air Conditioners: Including window air conditioners, split systems, and central air conditioning systems.
3. Cleaning Appliances: For example, washing machines, dryers, vacuum cleaners, and floor waxers.
4. Cooking Appliances: Like rice cookers, microwaves, ovens, and induction cookers.
5. Water Heating Appliances: Including water heaters (gas water heaters, electric water heaters) and heat pumps.
6. Lighting Appliances: Comprising various lamps, LED bulbs, and intelligent lighting control systems.
7. Personal Care Appliances: Such as electric shavers, hair dryers, beauty instruments, and personal massagers.
8. Small Appliances: This category includes a wide variety, such as coffee makers, toasters, juicers, blenders, soy milk makers, air fryers, etc.
9. Entertainment Appliances: Mainly include televisions, sound systems, gaming consoles, etc.
10.Smart Home Appliances: Like smart thermostats, smart sockets, security cameras, etc.
With technological advancements, many new household appliances continue to emerge, often featuring smart functions that can be operated via mobile apps or voice control, bringing more convenience and comfort to people's lives. It is important to note that the above list is not exhaustive; different cultures and regions may have specific types of household appliances. Additionally, over time, some older appliances may be replaced by more efficient, energy-saving new products.

Please evaluate this answer. Does it respond to the question asked?
A. Correct and satisfactory answer
B. Acceptable answer, but of average quality
C. Incorrect or poor-quality answer
Evaluation: [A]

Model Answer Comparison

Question: 举例说明计算机体系结构中的缓存写入策略
Answer1: 在计算机体系结构中，为了确保数据存储的稳定性、提高系统性能和减少资源消耗，通常采用以下几种写回策略：1. 全盘刷新（Full Write）：当处理器需要对某个内存地址进行修改时，整个存储器芯片会全部更新。这种写回方式可以保证数据的准确性，但可能导致大量时间浪费在处理重复信息上，影响系统的整体效率。2. 局部刷新（Partial Write）：对于单个或少数几个字节的数据更改，只对其附近的相邻区域进行更新。这种方式可以在保持高准确性的同时，节省时间和空间，降低功耗。3. 重叠式写回（Overlap Writing）：将多个连续的写入操作合并为一次大型的写入操作，从而实现更高效的写入过程。这种方法适用于大型文件或者频繁访问的程序代码等场景，能够显著提升系统性能。4. 异步写回（Asynchronous）：根据不同的需求和环境条件，选择合适的写回时机，可以充分利用硬件资源的潜力，避免因等待而导致的额外开销。总之，各种写回策略各有其特点和使用范围，在实际应用中应结合具体情况进行权衡与优化，以确保系统稳定运行并达到最佳效果。
Answer2: 在计算机系统设计中，为了提高存储器的访问速度和减少处理器等待时间，通常采用的写缓冲（或称“Cache”）技术。当一个数据被更新时，首先会将其存入到写缓冲区中，然后由CPU向内存发出指令来将修改后的内容写到内存中去。这样就避免了直接从内存读取新数据的延迟问题，提高了系统的性能表现。同时，由于写缓冲是位于 CPU 和主存储器之间的中间层，因此可以降低对主存储器和处理器的压力，从而达到优化整个系统资源利用的效果。当然，随着技术的不断发展，现代计算机系统中已经逐渐开始使用更先进、更高效率的写缓冲方案了。总体来说，通过合理的设计和使用写缓冲器，能够有效提升计算机体系的整体运行效能并实现更好的用户体验效果。
请从以上两个回复中选出你认为更好的一项
A Answer1
B Answer2
评价：[A]

Question: Describe examples of cache write policies in computer architecture.
Answer 1: In computer architecture, common write-back strategies include: 1. Full Write: Updates the entire memory chip when a memory address is modified, ensuring accuracy but may be inefficient. 2. Partial Write: Updates only nearby areas for small data changes, saving time and reducing power consumption. 3. Overlap Writing: Combines multiple write operations into one for efficiency, especially for large files or frequently accessed code. 4. Asynchronous Write-Back: Writes back data at optimal times based on system conditions, optimizing resource use and minimizing delays. These strategies vary in their features and applications, requiring careful selection for optimal system performance.
Answer 2: Computer systems often use write buffer (Cache) technology to speed up memory access and reduce CPU wait times. Data updates are first stored in the write buffer, then written to memory, reducing direct memory access delays. This technique, evolving with newer, more efficient schemes, enhances overall system performance and user experience.
Choose the better response:
A. Answer 1
B. Answer 2
Evaluation: [A]

Figure 11: Examples in Human Evaluation of Data and Model Performance.

A.6 STANDARD BENCHMARKS

Model	C-EVAL PPL	C-EVAL GEN	CMMLU PPL	CMMLU GEN
Kun-9k	72.95	72.96	73.53	24.53
Kun-26k	73.73	73.26	75.75	25.23
Kun-39k	73.84	**73.38**	75.68	25.31
Kun-52k	73.36	72.72	75.48	25.50
Kun-100k	72.95	72.96	73.53	24.53
Kun-200k	73.36	72.07	74.81	31.81
Kun-360k	73.25	71.88	74.14	**46.58**
Kun-52k + COIG-one-38k	73.61	**73.38**	75.70	34.84
COIG-38k	**74.09**	72.95	**75.92**	38.40
Belle-52k	72.31	72.56	74.73	44.56
OL-CC-10k	72.00	71.62	74.97	34.38

Table 3: Performance Statistics on Standard Benchmarks.The best results in each section are **Bold** , the second-best results are underlined, while the results of our best model are in Blue . PPL:Perplexity GEN:Generation