# COIG-P: A High-Quality and Large-Scale Chinese Preference Dataset for Alignment with Human Values

**M-A-P, 2077AI**

## Abstract

Aligning large language models (LLMs) with human preferences has achieved remarkable success. However, existing Chinese preference datasets are limited by small scale, narrow domain coverage, and lack of rigorous data validation. Additionally, the reliance on human annotators for instruction and response labeling significantly constrains the scalability of human preference datasets. To address these challenges, we design an **LLM-based Chinese preference dataset annotation pipeline** with no human intervention. Specifically, we crawled and carefully filtered **9k** high-quality Chinese queries and employed **15** mainstream LLMs to generate and score chosen-rejected response pairs. Based on it, we introduce **COIG-P** (**C**hinese **O**pen **I**nstruction **G**eneralist - **P**reference), a high-quality, large-scale Chinese preference dataset, comprises **101k** Chinese preference pairs spanning **6** diverse domains: **Chat**, **Code**, **Math**, **Logic**, **Novel**, and **Role**. Building upon COIG-P, to reduce the overhead of using LLMs for scoring, we trained a 8B-sized **C**hinese **R**eward **M**odel (**CRM**) and meticulously constructed a **C**hinese **R**eward **Bench**mark (**CRBench**). Evaluation results based on AlignBench [Liu et al., 2024a] show that that COIG-P significantly outperforms other Chinese preference datasets, and it brings significant performance improvements ranging from **2%** to **12%** for the **Qwen2/2.5** and **Infinity-Instruct-3M-0625** model series, respectively. The results on CRBench demonstrate that our CRM has a strong and robust scoring ability. We apply it to filter chosen-rejected response pairs in a test split of COIG-P, and our experiments show that it is comparable to GPT-4o in identifying low-quality samples while maintaining efficiency and cost-effectiveness. All our codes and data are released in https://github.com/multimodal-art-projection/COIG-P.

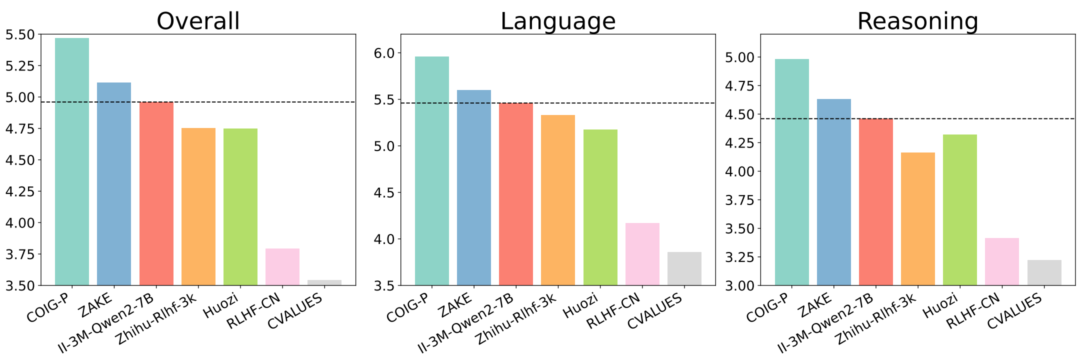


Figure 1. The results of different Chinese human preference datasets trained on Infinity-Instruct-3M-0625-Qwen2-7B.

## Contents

## 1. Introduction

Large Language Models (LLMs) [OpenAI, 2024, Yang et al., 2024b,a, Dubey et al., 2024] have achieved remarkable success in various Natural Language Processing (NLP) tasks [Wu et al., 2025, Team et al., 2025, Wu et al., 2024, Li et al., 2024b, Wang et al., 2023, Kalla et al., 2023, Ray, 2023, Firat, 2023, Bang et al., 2023]. To enable LLMs to be better applied in real-life scenarios, researchers utilize reinforcement learning (RL) technology (e.g., PPO [Schulman et al., 2017], DPO [Rafailov et al., 2023], RLHF [Ziegler et al., 2019]) to endow the LLMs with grasping the human intention and preference.

| Language | Dataset | Number | Quality Check |
|---|---|---|---|
| **English** | Arena [Chiang et al., 2024] | 55k | ✓ |
| | UltraFeedback [Cui et al., 2023] | 64k | ✓ |
| | Nectar [Zhu et al., 2023] | 183k | ✓ |
| | HH-RLHF[Ganguli et al., 2022] | 161k | ✗ |
| | H4 StackExchange [Lambert et al., 2023] | 10.8M | ✗ |
| | PreferenceShareGPT [PreferenceShareGPT, 2024] | 11.9 | ✓ |
| | Anthropic HH Goldenhuggingface [2024a] | 42.5k | ✓ |
| | Ask Again [Xie et al., 2023] | 2.6k | ✓ |
| | Orcaratgen [Just et al., 2024] | 12k | ✗ |
| | CodeUF [Weyssow et al., 2024] | 19k | ✓ |
| **Chinese** | Huozi [Huozi-Team, 2024] | 16k | ✗ |
| | ZAKE [huggingface, 2024b] | 77k | ✗ |
| | HH-FLHF-CN [huggingface, 2024c] | 344k | ✗ |
| | CVALUES [Xu et al., 2023] | 145k | ✓ |
| | GPT-4-LLM [Peng et al., 2023] | 52K | ✗ |
| | Zhihu-Rlhf-3k [huggingface, 2024d] | 3k | ✗ |
| | COIG-P (Ours) | 1,009k | ✓ |

Table 1. The human preference alignment datasets. The **Quality Check** means whether the author demonstrated the quality of the dataset on the downstream task by training a model.

As one of the most widely spoken languages, Chinese holds significant value in the development of open-source datasets, which are crucial for fostering progress within the Chinese open-source NLP community. However, as shown in Table 1, Chinese human value preference datasets remain scarce and lack rigorous data validation, especially when compared to their English counterparts. On one hand, existing Chinese human preference datasets are not only limited in quantity but also suffer from quality issues. Notably, many of these datasets are derived from a single source (e.g., zhihu)[1], leading to concerns about representativeness and diversity. Moreover, some datasets lack rigorous data filtering and quality control processes, raising questions about their reliability and validity. On the other hand, introducing human annotation for chosen and rejected responses requires substantial human resources, and the inconsistency of manual annotations significantly increases the cost of data labeling. Although UltraFeedback [Cui et al., 2023] also uses the LLMs to annotate and score responses, they only use a single LLM to score responses, which is likely to introduce bias. Besides, it is hard to choose high-quality chosen-rejected response pairs because they annotate 4 different scores from different dimensions for each response.

Inspired by UltraFeedback [Cui et al., 2023], we propose an **LLM-based Chinese preference dataset annotation pipeline** to curate Chinese preference datasets without human annotation. Firstly, we collect and filter **9k** Chinese queries covering comprehensive dimensions. In order to make LLMs efficiently learn the preferences of humans, we choose **15** open-source and closed-source LLMs to generate various responses to a query and select **8** LLMs among them to score responses to form chosen and rejected response pairs. Based on it, we introduce **COIG-P**

[1] https://huggingface.co/datasets/liyucheng/zhihu_rlhf_3k

(**C**hinese **O**pen **I**nstruction **G**eneralist - **P**reference), a Chinese human value preference dataset that contains **101k** samples, and each sample consists of a query and a pair of chosen and rejected responses. To reduce the overhead of using LLMs for scoring, we also trained a **C**hinese **R**eward **M**odel (**CRM**) and manually curated a **C**hinese **R**eward **B**enchmark (**CRBench**). The data of CRBench undergoes cleaning, restructuring, and careful human verification to ensure its quality and diversity, which are essential for improving LLMs' capability to align with humans' values in the Chinese context and better apply them to real-world scenarios. We conduct experiments using the COIG-P dataset to align current mainstream LLMs with human values in Chinese through DPO, where a significant improvement is observed on AlignBench. Besides, using our designed CRM achieves comparable performance with closed-source LLMs in scoring and pairing the DPO preference dataset.

Our main contributions are as follows:

- We present an LLM-based annotation pipeline for Chinese preference datasets and use it to build COIG-P, a high-quality, large-scale dataset for human value alignment. Experimental results show that existing mainstream LLMs (including the **Qwen2/2.5** and **Infinity-Instruct-3M-0625** series) achieve significant performance gains ranging from 2% to 12% on this dataset.
- To demonstrate the effectiveness of our LLM-based Chinese preference dataset annotation pipeline, we compared COIG-P with other Chinese human preference datasets. COIG-P brought significant improvements to the model, far surpassing other datasets. In fact, most existing Chinese datasets even degraded model performance.
- To mitigate the substantial time and computational costs associated with using LLMs for scoring, we trained a Chinese Reward Model (CRM) based on COIG-P and manually annotated a Chinese Reward Benchmark (CRBench). Based on CRBench, we investigated the limitations of current mainstream reward models in scoring Chinese responses, while CRM demonstrated strong scoring capabilities in Chinese. Furthermore, we evaluated CRM's real-world annotation performance on a test split of COIG-P, showing that its annotation quality is comparable to GPT-4o and significantly more efficient.

## 2. Related Work

High-quality datasets play a crucial role in the development of large language models (LLMs) [Raffel et al., 2020, Mishra et al., 2022, Wang et al., 2022b, Zeng et al., 2022, Longpre et al., 2023, Taori et al., 2023, Si et al., 2023, Chenghao Fan and Tian, 2023]. Beyond the construction of instruction-tuning data, increasing attention has been directed toward curating human preference datasets to enhance LLM alignment through reinforcement learning techniques (e.g., DPO, PPO). Recent efforts in preference data construction can be broadly categorized into two paradigms: **human annotation** and **LLM-based annotation**.

Early English-language datasets primarily relied on manual annotations for preference comparisons. For example, the HH-RLHF dataset [Bai et al., 2022] proposed by Anthropic employs human annotators to assess assistant responses based on helpfulness and harmlessness, leading to significant advances in alignment. Similarly, Ethayarajh et al. [2022] collected user voting preferences from Reddit forums, yielding a large-scale corpus of naturally annotated data. However, manual annotation is time-consuming and costly, posing challenges to scalability.

As a result, recent approaches increasingly leverage LLMs to automate preference data construction [Zhu et al., 2023, Cui et al., 2023, Lambert et al., 2023, PreferenceShareGPT, 2024, huggingface, 2024a, Chiang et al., 2024]. In addition to enhancing general alignment capabilities,

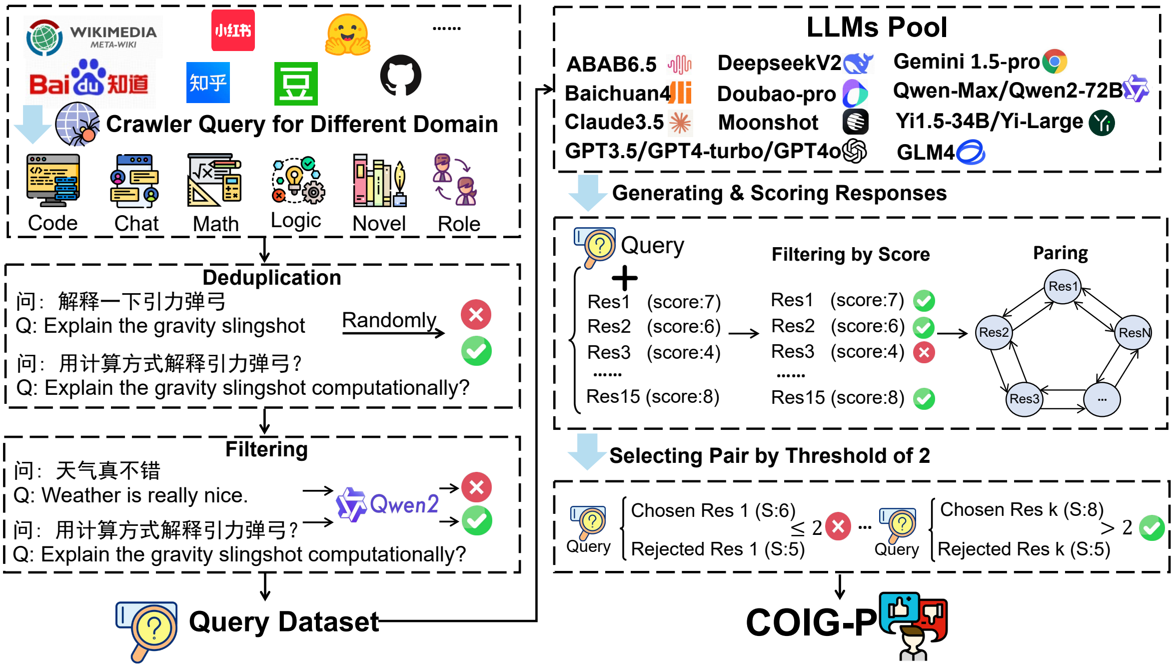


Figure 2. The data curation process of COIG-P. The left part is the query collection process, and the right part illustrates the generation of chosen and rejected responses.

some studies focus on domain-specific alignment [Cui et al., 2023, Xie et al., 2023, Just et al., 2024, Weyssow et al., 2024]. These approaches typically involve generating multiple candidate responses to a prompt using various LLMs, followed by ranking and evaluation via a stronger model—such as GPT-4—to produce high-quality preference annotations. While this strategy significantly improves scalability and efficiency, it also introduces potential biases, as evaluation models may favor responses that resemble their own outputs [Li et al., 2024a, Liu et al., 2024b]. Furthermore, even in LLM-driven pipelines, certain steps still require human involvement. For instance, Cui et al. [2023] does not provide the chosen–rejected pairs.

In the Chinese context, preference datasets have historically lagged behind in both scale and diversity. Early efforts were limited to small-scale, scenario-specific datasets constructed via human annotation, machine translation, or rule-based heuristics [Xu et al., 2023, huggingface, 2024b, Huozi-Team, 2024, Xinlu Lai, 2024], making them insufficient for training general-purpose dialogue models. Although recent attempts have explored LLM-based annotation in Chinese, the resulting datasets remain limited in quality and coverage [Peng et al., 2023, huggingface, 2024d,c]. Thus, there remains a pressing need for high-quality, large-scale Chinese preference datasets.

## 3. Data Curation

Existing LLMs still exhibit a gap in aligning with human values, particularly in Chinese-language corpora, which limits their effectiveness in real-world applications. Therefore, as shown in Figure 2, we propose a **LLM-based Chinese preference dataset annotation pipeline** to curate the COIG-P, a human value preference dataset of Chinese corpus, without human intervention.

### 3.1. Query Collection

Instruction tuning [Sanh et al., 2021, Wang et al., 2022a, Longpre et al., 2023, Ni et al., 2023] and RLHF [Schulman et al., 2017, Ziegler et al., 2019, Rafailov et al., 2023] play a crucial role in LLMs' success. Despite Chinese being one of the most widely spoken languages in the world, most of the Chinese instruction datasets [Yang, 2023, Bai et al., 2024] come from traditional NLP tasks, and the query format has a significant gap compared to the way humans ask questions in daily life. Open-source, high-quality Chinese query data remains extremely scarce. As a result, there is an urgent need to address the challenge of collecting large-scale, high-quality queries to improve LLMs' alignment with human values in Chinese.

As shown in the left part of Figure 2, we collect **9k** high-quality Chinese queries. Enhancing the performance of LLMs in a single domain may come at the cost of their reduced alignment with human values in other domains. Inspired by Liu et al. [2024a]'s subtask designing, we collect queries from different domains including **Chating (Chat.** 对话**), Logic Reasoning (Logic.** 逻辑推理**), Mathematics (Math.** 数学**), Novel Continuation (Novel.** 小说续写**), Role-Playing (Role.** 角色扮演**), and Coding (Code.** 代码**)**. To support the growth of the Chinese open-source community, we have collected Chinese query data from Chinese Q&A platforms, including baiduzhidao[2], zhihu[3], and baidutieba[4]. We also collect the queries from Chinese Administrative Aptitude Test. Besides, we translate some queries from the English open-source dataset into Chinese, such as HotpotQA [Yang et al., 2018] and Haruhi-Zero-RolePlaying-movie-PIPPA[5]. The details of our used Open-source dataset refer to Appendix B

To maintain the quality of the collected queries, we conduct the **deduplication** and **filtering**:

**Deduplication**: We utilize SentenceBERT to obtain query embeddings and compute the semantic similarity between different queries. Queries with high semantic similarity to others are removed to ensure diversity.

**Filtering**: First, we employ Qwen2-72B [Yang et al., 2024a] to score the queries and discard those with scores below 5 based on whether it reflects a question that a typical user might ask. Then, we design some rules to remove queries that are not well-formed and whether.

After these processes, we obtain **92,784** high-quality queries from the Chinese corpus.

### 3.2. Response Generation

Relying solely on a single LLM for response generation can lead to monotonous outputs. Therefore, we leverage diverse LLMs with distinct characteristics to enhance response diversity. Inspired by Cui et al. [2023], we utilize **15** different open-source and proprietary LLMs (i.e., **Abab6.5**[6], **Baichuan4**[7], **Claude3.5**[8], **DeepSeek-V2** [DeepSeek-AI et al., 2024], **Doubao-Pro**[9], **Gemini1.5-Pro**[10], **GPT-Turbo/3.5/4/4o**[11], **Qwen-Max**, **Qwen2-72B** [Yang et al., 2024a], **Yi-1.5-34B**,

[2] https://zhidao.baidu.com/
[3] https://www.zhihu.com/
[4] https://tieba.baidu.com/index.html
[5] https://huggingface.co/datasets/silk-road/Haruhi-Zero-RolePlaying-movie-PIPPA
[6] https://www.minimax.io/news/abab65-series
[7] https://platform.baichuan-ai.com/
[8] https://www.anthropic.com/news/claude-3-5-sonnet
[9] https://team.doubao.com/zh/special/doubao_1_5_pro
[10] https://aistudio.google.com/app/prompts/new_chat
[11] https://chatgpt.com/

**Yi-Large** [AI et al., 2025], **GLM-4** [GLM et al., 2024a], and **Moonshot**[12]) to generate a range of responses for each query.

### 3.3. Scoring and Paring

For each query, we selected **8** LLMs (i.e., **Claude3.5**, **DeepSeekV2**, **Doubao-Pro**, **GLM-4**, **GPT-4o**, **GPT-4-Turbo**, **Qwen2-72B-Instruct**, and **Moonshot**) to score the responses. Besides, we designed tailored prompts for different data domains, as detailed in Appendix A.

To align LLMs with human values using DPO, we need to construct some pairs of chosen and rejected responses for each query. Specifically, we randomly sample two responses per query and retain only those pairs where the score difference between chosen and rejected responses exceeds a predefined threshold. We explore various threshold values and ultimately select a threshold of 2 in our study. For a detailed analysis, please refer to subsection 5.3. After applying these filtering and pairing steps, we curated a final dataset consisting of **1,006,949** samples, each containing a query along with a chosen and a rejected response.

### 3.4. Human Evaluation

To ensure the quality of our dataset, we randomly select 40 samples for each domain in our dataset and totally collect 240 samples to evaluate. We hire 2 postgraduate students who are familiar with the field of Natural Language Processing (NLP) to manually evaluate the quality of those samples. Specifically, we require the annotator to judge samples based on the following criteria: 1) whether the chosen response is better aligned with human preferences than the rejected response. 2) whether the chosen response is correct.

Based on human evaluation, the dataset achieves an average accuracy of **90.83**%, with domain-specific scores as follows: **Logic 90%**, **Novel 90%**, **Role 90%**, **Code 95%**, **Math 85%**, and **Chat 95%**. The consistently high accuracy—exceeding **90%** in most domains—demonstrates the robustness and quality of the dataset generated and evaluated by LLMs.

### 3.5. Statics

| | All | Logic. | Chat. | Math. | Novel. | Role. | Code. |
|---|---|---|---|---|---|---|---|
| Sample Num | **1,006,946** | 54,617 | 702,398 | 155,872 | 34,483 | 19,363 | 40,213 |
| Query Num | **92,784** | 8,816 | 37,323 | 27,259 | 6,682 | 4,930 | 7,774 |

Table 2. The statistics of our COIG-P dataset. The sample numbers represent the number of samples in our dataset, and each sample consists of a query and a chosen and rejected response pair. The query number represents the quantity of our filtered high-quality queries.

As shown in Table 2, we collected a total of 92,784 high-quality Chinese corpus queries. The Chat and Math domains constitute the largest portions, each containing approximately 30,000 queries, while other domains have around 6,000 queries each. This distribution suggests that in everyday applications, users are more likely to engage with Math-related topics and casual conversations.

For most domains, we generate around six response pairs per query. However, for the Chat

[12] `https://moonshotteam.com/`

domain, we curate approximately 20 response pairs per query, reflecting the relative simplicity of Chat-related queries.

## 4. Experiments Setup

In this study, we utilize AlignBench [Liu et al., 2024a] as our primary benchmark to assess the alignment capabilities of LLMs. AlignBench is a comprehensive, multi-dimensional benchmark specifically designed for evaluating LLM alignment in Chinese. Due to computational resource constraints, we employ **GPT-4o-08-06** as the judge model and rerun the current mainstream LLMs on it for a comprehensive comparison.

**Baselines.** Following the AlignBench evaluation framework, we assess several widely recognized LLMs. As for close-source LLMs, we choose **GPT-4o**[13] and **Claude3.5**[14]. As for the open-source LLMs, we selected the latest high-performing LLMs, such as **ChatGLM** [GLM et al., 2024b], **InternLM**[Team, 2023] series, ,**Llama3** [Dubey et al., 2024] and **DeepSeek-R1-Distill** series [DeepSeek-AI, 2025].

**Backbones.** To demonstrate the effectiveness of our COIG-P dataset, we evaluate its impact on SOTA LLMs within the 7–9B parameter range. Among these, **Qwen2.5/2-7B-Instruct** [Yang et al., 2024a,b] stands out as the most capable open-source LLM across various NLP tasks. Furthermore, we also choose the **Infinity-Instruct-3M-0625** [of Artificial Intelligence , BAAI] series LLMs that have been specifically optimized for the Chinese corpus as our backbone model (i.e., **Infinity-Instruct-3M-0625-Qwen2-7B**, **Infinity-Instruct-3M-0625-Llama3-8B**, and **Infinity-Instruct-3M-0625-Mistral-7B**).

### 4.1. Implementation Details

To validate the efficiency of the COIG-P dataset, we train the selected backbone models using the DPO method.

**Hyperparameters.** Our experiments indicate that a *beta* value of 0.1 yields the best performance across all LLMs. However, the optimal learning rate (lr) varies depending on the model's capabilities. Specifically, we set $lr = 1e-6$ for Qwen2/2.5, while for other LLMs, we use $1e-7$.

**Computational Cost.** Each backbone model is **fully fine-tuned** for **one epoch** on A800 GPUs, resulting in a total of approximately **400 GPU hours** per model. The cumulative computational cost for training all backbone models amounts to **2,000 GPU hours**.

## 5. Results

### 5.1. Overall Analysis

As shown in Table 3, to valid the efficiency of our COIG-P dataset, we conduct various experiments by training LLMs (i.e., Qwen2/2.5-7B and Infinity-Instruct-3M-0625 series models) using

[13] https://chatgpt.com/
[14] https://claude.ai/

| Model 模型 | Overall 总分 | Reasoning 中文推理 Avg. 推理总分 | Math. 数学计算 | Logi. 逻辑推理 | Language 中文语言 Avg. 语言总分 | Fund. 基本任务 | Chi. 中文理解 | Open. 综合问答 | Writ. 文本写作 | Role. 角色扮演 | Pro. 专业能力 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Baseline** | | | | | | | | | | | |
| GPT-4o | **6.93** | **7.06** | **7.63** | 6.49 | **6.80** | 6.81 | **6.81** | **6.74** | **6.63** | 6.47 | **7.35** |
| Claude3.5-Sonnet | 6.58 | 6.49 | 6.97 | 6.00 | 6.68 | **6.93** | 6.64 | 6.63 | 6.35 | 6.41 | 7.12 |
| Qwen2.5-72B-Inst | 6.80 | 6.96 | 7.21 | **6.71** | 6.65 | 6.63 | 6.50 | 6.58 | 6.51 | **6.67** | 7.00 |
| Llama3.3-72B-Inst | 5.52 | 5.55 | 5.91 | 5.20 | 5.48 | 5.49 | 4.76 | 5.50 | 5.37 | 5.93 | 5.81 |
| DS-R1-Dist-Qwen-32B | 6.13 | 6.23 | 6.40 | 6.05 | 6.03 | 6.04 | 5.93 | 6.37 | 5.96 | 6.14 | 5.77 |
| DS-R1-Dist-Qwen-7B | 4.74 | 5.43 | 5.96 | 4.90 | 4.05 | 4.28 | 3.57 | 4.50 | 4.25 | 4.30 | 3.40 |
| InternLM3-8B-Inst | 6.00 | 5.49 | 5.84 | 5.14 | 6.52 | 6.04 | 6.50 | 6.89 | 6.63 | 6.91 | 6.12 |
| InternLM2.5-20B-Chat | 5.75 | 5.32 | 5.81 | 4.84 | 6.18 | 6.09 | 5.90 | 6.82 | 6.01 | 6.55 | 5.71 |
| ChatGLM3-6B | 3.46 | 3.13 | 3.00 | 3.25 | 3.80 | 3.81 | 2.86 | 4.63 | 3.75 | 4.20 | 3.54 |
| **Backbone** | | | | | | | | | | | |
| Qwen2.5-7B-Inst | 5.90 | 5.77 | 6.38 | 5.15 | 6.03 | 5.99 | 5.86 | 6.34 | 5.93 | 6.08 | 6.01 |
| Qwen2-7B-Inst | 5.35 | 4.88 | 5.57 | 4.18 | 5.83 | 5.22 | 5.64 | 6.45 | 6.23 | 6.06 | 5.40 |
| II-3M-0625-Qwen2-7B | 4.96 | 4.46 | 4.65 | 4.27 | 5.46 | 5.03 | 4.98 | 6.03 | 5.65 | 5.84 | 5.20 |
| II-3M-0625-Llama3-8B | 3.83 | 3.20 | 3.40 | 3.00 | 4.45 | 4.21 | 3.57 | 4.87 | 4.99 | 5.12 | 3.95 |
| II-3M-0625-Mistral-7B | 3.73 | 3.25 | 3.29 | 3.20 | 4.22 | 3.94 | 3.41 | 4.55 | 4.63 | 4.96 | 3.84 |
| **COIG-P** | | | | | | | | | | | |
| Qwen2.5-7B-Inst | 6.02 ( ↑2.03%) | 5.97 | 6.58 | 5.36 | 6.08 | 5.87 | 5.74 | 6.34 | 6.24 | 6.41 | 5.87 |
| Qwen2-7B-Inst | 5.47 ( ↑2.24%) | 4.98 | 5.59 | 4.38 | 5.96 | 5.07 | 5.86 | 6.79 | 6.12 | 6.35 | 5.56 |
| II-3M-0625-Qwen2-7B | 5.37 ( ↑8.26%) | 4.83 | 5.30 | 4.35 | 5.92 | 5.47 | 5.41 | 6.89 | 6.07 | 6.16 | 5.49 |
| II-3M-0625-Llama3-8B | 4.30 (↑12.27%) | 3.75 | 3.93 | 3.58 | 4.85 | 4.71 | 3.83 | 5.45 | 5.29 | 5.60 | 4.20 |
| II-3M-0625-Mistral-7B | 3.98 ( ↑6.70%) | 3.52 | 3.56 | 3.48 | 4.43 | 4.69 | 3.59 | 4.89 | 4.77 | 4.97 | 3.69 |

Table 3. Results on AlignBench and the score range for each metric in it is **0-10**. The ↑ presents overall improvement in the format of percentage, presents the improvement in the sub-task, and presents a decrease in the sub-task. We re-evaluated current SOTA LLMs on this benchmark using GPT-4o-0806. II-3M-0625 refers to Infinity-Instruct-3M-0625, while the COIG-P setting denotes LLMs trained on our dataset using DPO.

DPO on our dataset. We also update the current mainstream LLMs on AlignBench.

**Based on COIG-P, mainstream LLMs have achieved significant improvements in overall performance.** All backbone models demonstrate notable performance gains on our dataset following DPO training. In particular, II-3M-0625-Qwen2-7B and II-3M-0625-Llama3-8B achieved an increase of more than 0.41 in their overall scores. Within the II-3M-0625 series, the relative improvements range from 6% to 12%, indicating consistent and substantial enhancements. Even for Qwen2.5-7B-Inst, one of the strongest open-source LLMs, our dataset contributed to a performance gain of 0.12. Additionally, for both Qwen2/2.5-7B, relative improvements also exceeded 2%, underscoring the effectiveness of our dataset in enhancing LLM capabilities.

**COIG-P consistently improves performance across all sub-tasks for most backbone models.** For relatively weaker models, COIG-P can help them achieve comprehensive improvements across all subtasks (e.g., II-3M-0625-Qwen2-7B and I-3M-0625-Llama3-8B). For models that are relatively powerful (i.e., Qwen2.5-7B-Inst), DPO training can enhance their reasoning (中文推理) abilities. However, it may cause a slight degradation in some fundamental language (中文语言) subtasks, especially in Fundamental (基础任务).

**The gap between open-source and closed-source models is small in Chinese preference alignment tasks.** Compared to GPT-4o, Qwen2.5-72B-Inst shows only slight differences in scores across various tasks, and its overall score is even significantly higher than that of Claude-3.5-Sonnet. By using our COIG-P dataset, the performance of the Qwen2.5-7B model can be improved to a level close to that of DS-R1-Dist-Qwen-32B, making it overall score exceed 6.0. This demonstrates that many smaller open-source models, such as ChatGLM3-6B and DS-R1-Dist-Qwen-7B, still have significant room for improvement in Chinese preference alignment.

### 5.2. Ablation Study

| **Dataset** | **Overall** | **Reasoning** 中文推理 | | | **Language** 中文语言 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 训练数据 | 总分 | **Avg.** 推理总分 | **Math.** 数学计算 | **Logi.** 逻辑推理 | **Avg.** 语言总分 | **Fund.** 基本任务 | **Chi.** 中文理解 | **Open.** 综合问答 | **Writ.** 文本写作 | **Role.** 角色扮演 | **Pro.** 专业能力 |
| Backbone | 4.96 | 4.46 | 4.65 | 4.27 | 5.46 | 5.03 | 4.98 | 6.03 | 5.65 | 5.84 | 5.20 |
| COIG-P | **5.47** | **4.98** | **5.59** | **4.38** | **5.96** | 5.07 | **5.86** | **6.79** | **6.12** | **6.35** | **5.56** |
| Chat | 4.97 | 4.44 | 4.86 | 4.02 | 5.50 | 5.19 | 5.31 | 5.87 | 5.75 | 5.66 | 5.23 |
| Novel | 5.29 | 4.98 | 5.74 | 4.23 | 5.60 | **5.69** | 5.09 | 6.00 | 5.79 | 5.82 | 5.22 |
| Role | 4.87 | 4.37 | 4.73 | 4.00 | 5.38 | 5.06 | 4.97 | 5.66 | 5.65 | 5.74 | 5.20 |
| Logic | 4.87 | 4.36 | 4.85 | 3.87 | 5.37 | 5.07 | 5.02 | 6.05 | 5.55 | 5.55 | 5.01 |
| Math | 4.76 | 4.37 | 4.78 | 3.96 | 5.14 | 4.79 | 5.09 | 5.53 | 5.29 | 5.21 | 4.96 |
| Code | 4.72 | 4.24 | 4.69 | 3.78 | 5.20 | 4.65 | 4.95 | 5.63 | 5.24 | 5.53 | 5.21 |

Table 4. Ablation study results. We trained Infinity-Instruct-3M-0625-Qwen2-7 on those datasets and evaluated them on AlignBench. The Backbone means the result of the raw Infinity-Instruct-3M-0625-Qwen2-7B.

Excepting the data from the user's daily interaction on the Internet (**Chat**), we also collect data from some specific domains including (**Novel**, **Role**, **Logic**, **Math** and **Code**). To this end, we conducted ablation studies to demonstrate that mixing data from different domains can better enhance the human value alignment capabilities of LLMs. The results are presented in Table 4.

Overall, training the model on individual domain datasets performs worse than training it on a combination of multiple domains. In fact, using data from certain domains alone can even harm the model's overall performance. This demonstrates the effectiveness of our strategy of training the model with a mixture of data from different domains.

It is worth noting that training the model solely on the novel continuation task (Novel) leads to a significant performance improvement. The model's trained on the Novel dataset saw a substantial boost in fundamental language ability (Fund.), reaching 5.69 — an increase of 0.71. This enhancement in fundamental language skills directly contributed to the improvement in the model's reasoning ability.

### 5.3. Selecting Score Threshold of Pairing

For each query, we prompt the LLMs to generate multiple chosen–rejected response pairs, and then filter out low-quality pairs based on scores assigned by the LLMs themselves. Specifically, we define a threshold and discard any pair where the score difference between the chosen and rejected responses falls below this threshold.

To select a suitable threshold, we randomly selected 1,000 queries in COIG-P. For each query, we formed potential chosen–rejected pairs across all available responses and then applied varying thresholds to decide which pairs to keep based on the score judged by LLMs.

As shown in Figure 3, we train Infinity-Instruct-3M-0625-Qwen2-7B on datasets filtered with different thresholds and evaluate them on AlignBench. As the threshold increases up to 2.0, the model's performance generally shows an upward trend; however, once the threshold surpasses 2.0, the performance gradually declines. Therefore, in this paper, we select 2.0 as the threshold to filter the data.

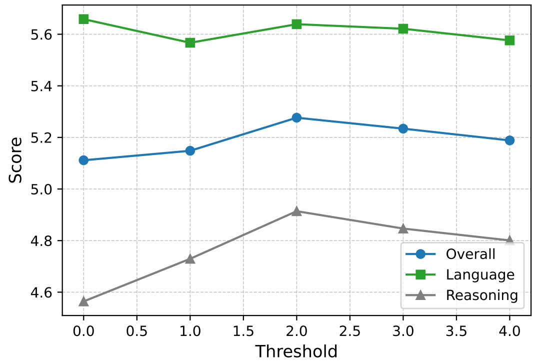


Figure 3. Selection of the pairing score threshold. A threshold of 0 indicates that the score of the chosen response is higher than that of the rejected response.

### 5.4. Comparing Chinese Human Preference Dataset

**Compared to other datasets, COIG-P shows the greatest improvement and demonstrates notable performance gains across all sub–tasks.** As illustrated in Table 5, our analysis indicates that only the COIG-P and ZAKE datasets positively contribute to Chinese language alignment capabilities, while the remaining datasets lead to significant performance declines. COIG-P achieves the highest performance across all metrics (overall: 5.47, reasoning: 4.98, language: 5.96), whereas ZAKE provides moderate improvements (overall: 5.11, reasoning: 4.63, language: 5.60). Nevertheless, the enhancement provided by ZAKE in Chinese language tasks is modest, surpassing the baseline by merely 0.2–0.3 points. Furthermore, its effect on reasoning is inconsistent, which enhances mathematical skills but negatively impacts logical reasoning, scoring approximately 0.4 points lower than COIG-P. In contrast, COIG-P brings a 0.5 improvement to the model on most tasks. Other datasets, such as Zhihu-RLHF-7B, RINI-3K, Huozi, and particularly CVALUES (which achieves only 3.54 overall), lead to substantial performance declines. Moreover, these datasets lead to performance degradation for models across various subtasks. These findings indicate that, aside from a few exceptions, the current Chinese preference datasets lack sufficient quality and quantity to adequately support the development needs of advanced LLMs.

## 6. Chinese Reward Model and Chinese Reward Benchmark

The LLMs' scoring ability is still under-explored, and using the closed-source LLM (i.e., GPT-4o, Cluade) and open-source LLMs with a massive number of parameters (i.e., Qwen2.5-72B) posed significant obstacles to the development of Chinese datasets. Developing small-parameter LLMs is an urgent task. Therefore, we propose a Chinese Reward Model (in subsection 6.1) and a Chinese Reward Benchmark (in subsection 6.2) to fill the gap in this field.

| Dataset<br>训练数据 | Overall<br>总分 | Reasoning 中文推理 | | | Language 中文语言 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Avg.<br>推理总分 | Math.<br>数学计算 | Logi.<br>逻辑推理 | Avg.<br>语言总分 | Fund.<br>基本任务 | Chi.<br>中文理解 | Open.<br>综合问答 | Writ.<br>文本写作 | Role.<br>角色扮演 | Pro.<br>专业能力 |
| - | 4.96 | 4.46 | 4.65 | 4.27 | 5.46 | 5.03 | 4.98 | 6.03 | 5.65 | 5.84 | 5.20 |
| Zhihu-Rlhf-3k | 4.75 | 4.16 | 4.51 | 3.82 | 5.33 | 4.72 | 5.21 | 5.66 | 5.68 | 5.47 | 5.27 |
| CVALUES | 3.54 | 3.22 | 3.14 | 3.29 | 3.86 | 3.71 | 3.41 | 3.84 | 4.20 | 4.17 | 3.82 |
| Huozi | 4.75 | 4.32 | 4.60 | 4.04 | 5.17 | 4.93 | 4.86 | 5.32 | 5.47 | 5.41 | 5.06 |
| ZAKE | 5.11 | 4.63 | 5.29 | 3.98 | 5.60 | 5.01 | 5.26 | 6.26 | 5.81 | 6.00 | 5.23 |
| RLHF-CN | 3.79 | 3.41 | 3.49 | 3.34 | 4.17 | 4.38 | 4.47 | 3.75 | 4.30 | 4.13 | 4.00 |
| COIG-P (Ours) | **5.47** | **4.98** | **5.59** | **4.38** | **5.96** | **5.07** | **5.86** | **6.79** | **6.12** | **6.35** | **5.56** |

Table 5. Performance comparison of LLMs trained on different Chinese human preference datasets. The backbone model used is Infinity-Instruct-3M-0625-Qwen2-7B. The "-" symbol represents the performance of the backbone model without additional training.

### 6.1. Chinese Reward Model

Inspired by Ouyang et al. [2022], we use the Bradley-Terry (BT) Reward Modeling method. Specifically, we choose the Llama3.1-8B-Instruct as our Foundation model, and the objective function of the Bradley-Terry (BT) loss is as follows:

$$\mathbb{P}(a^1 \succ a^2 | x, a^1, a^2) = \frac{\exp(r^*(x, a^1))}{\exp(r^*(x, a^1)) + \exp(r^*(x, a^2))} = \sigma(r^*(x, a^1) - r^*(x, a^2)),$$

where $x$ present the query, $a^1$ presents the chosen response, and the $a^2$ presents the rejected response.

Considering testing our CRM, we only use half of our COIG-P dataset to train our CRM and test our CRM in the rest.

### 6.2. Chinese Reward Benchmark

In order to better evaluate the Chinese scoring capability of current LLMs, we have standardized the Chinese Reward Benchmark (CRBench). To ensure high-quality data annotation, we recruited three postgraduate students, each responsible for two specific domains. From the dataset, we randomly selected 5,000 samples and asked the annotators to assess whether each sample should be included based on the following criteria: 1) The query must be a well-formed question and should not involve sensitive topics such as sex, politics, etc. 2) The chosen response of the selected sample must be correct. 3) The chosen response of the sample should better align with human preferences compared to the rejected response.

The annotator will pause the annotation until the total number of samples in the benchmark exceeds 1,000. As shown in Table 6, we finally annotate **1,040** samples.

| All. | Chat. | Logic. | Math. | Code. | Role. | Novel. |
|---|---|---|---|---|---|---|
| 1,040 | 129 | 375 | 274 | 101 | 80 | 81 |

Table 6. The static of our Chinese Reward Benchmark (CRBench).

As shown in Table 7, we evaluate the current mainstream LLMs and reward models in the CRBench. **Our CRM achieves the best performance among the discriminative reward**

| Model 模型 | Conv. 对话 | Logic. 逻辑推理 | Math. 数学 | Code. 代码 | Role. 角色扮演 | Novel. 小说续写 | Overall 总分 |
|---|---|---|---|---|---|---|---|
| **Generative** | | | | | | | |
| Claude | 86.82 | 74.67 | 61.68 | 92.08 | 75.00 | 70.37 | 74.13 |
| GPT-4o | 96.12 | 88.27 | 72.63 | 98.02 | 93.75 | 91.36 | 86.73 |
| **Discriminative** | | | | | | | |
| Skywork-Reward-Gemma-2-27B | 62.02 | 53.60 | 54.01 | 59.41 | 50.00 | 61.73 | 55.67 |
| Llama-3-OffsetBias-RM-8B | 34.11 | 54.93 | 68.98 | 72.28 | 47.50 | 34.57 | 55.58 |
| RM-Mistral-7B | 86.82 | 61.33 | 61.68 | 90.10 | 53.75 | 49.38 | 65.87 |
| ArmoRM-Llama3-8B | 58.91 | 44.27 | 41.97 | 46.53 | 41.25 | 27.16 | 44.13 |
| Skywork-Reward-Llama-3.1-8B | 75.97 | 52.00 | 49.27 | 78.22 | 35.00 | 34.57 | 54.13 |
| CRM (Ours) | 79.07 | 69.60 | 66.79 | 92.08 | 43.75 | 62.96 | 69.71 |

Table 7. Model performance comparison.

**models.** Although the closed-source Generative model (GPT-4o and Claude3.5) achieves the best performance, the performance gap between CRM and them is also relatively small (i.e., the overall performance gap between Claude and CRM is less than 4%).

Besides, **the Logic(逻辑推理), Math(数学), Role(角色扮演), and Novel(小说续写) tasks remain challenging for most models**. Except for GPT-4o, all models score below 75% on these tasks, with most clustering around 60%. This further highlights the necessity of our benchmark.

### 6.3. Downstream Task Validation

Besides demonstrating our Chinese Reward Model's ability on the Chinese Reward Benchmark, we also apply it to pairing responses and compare the result of our CRM with GPT-4o. We use our CRM and GPT-4o to filter data in the test split described in the subsection 5.3. We filter data when the score of the chosen response if lower than rejected response.

**It is worth mentioning that our CRM achieved a close performance with GPT-4o in paring chosen-rejected pairs.** As shown in Figure 4, in the test split, the model trained on the data selected by our CRM achieves an Overall score of 5.26, which is close to that of GPT-4o (5.28). In all sub-tasks, the CRM's results are also competitive with the GPT-4o. Our experiments demonstrate that our CRM has the practical ability to choose high-quality chosen-rejected response pairs.

**Our CRM is more effective than LLMs.** Comparing the LLMs with large-scale parameters, using our CRM to score responses on 430k samples only cost 40 A800 GPU hours. It demonstrates that our model has a notable speed advantage in data filtering, significantly reducing the cost of developing Chinese datasets.

## 7. Conclusion

DPO and PPO have played a significant role in aligning large language models (LLMs) with human value preferences. However, the lack of large-scale, high-quality Chinese preference data has limited the alignment of LLMs with human preferences in the Chinese context. To address this, we propose an LLM-based pipeline for constructing Chinese preference data and use it to create COIG-P, a dataset containing 101k high-quality Chinese preference samples. We demonstrate on AlignBench that COIG-P brings a 2%–12% performance improvement to mainstream

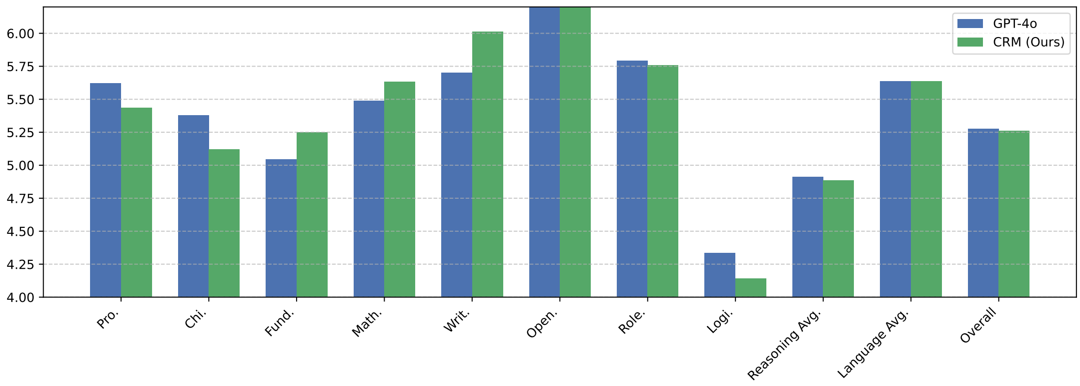


Figure 4. The results of different reward models in scoring chosen-rejected pairs. We trained Infinity-Instruct-3M-0625-Qwen2-7B using a dataset filtered by different reward models and evaluated them on AlignBench.

LLMs, including the Qwen2/2.5 and Infinity-Instruct-3M-0625 series. Compared to existing Chinese preference datasets, COIG-P yields significantly better performance improvements.

Furthermore, due to the scarcity of Chinese preference data, there is currently no strong Chinese reward model that can replace LLMs for scoring and reduce computational costs. To address this, we propose a Chinese reward model along with a corresponding Chinese reward benchmark. We also validate that our Chinese reward model achieves performance comparable to GPT-4o on downstream tasks involving real data annotation.

## 8. Contributions and Acknowledgments

Multimodal Art Projection (M-A-P) is a non-profit open-source AI research community, run by donations. The community members are working on research topics in a wide range of spectrum, including but not limited to the pre-training paradigm of foundation models, large-scale data collection and processing, and the derived applications on coding, reasoning, and music generation.

**Leading Authors**

- Siwei Wu, UoM, M-A-P
- Jincheng Ren, M-A-P
- Xinrun Du, M-A-P
- Shuyue Guo, M-A-P
- Xingwei Qu, M-A-P

**Contributors**

- Yiming Liang, M-A-P
- Jie Liu, M-A-P
- Yunwen Li, CUHK-Shenzhen
- Tianyu Zheng, M-A-P
- Boyu Feng, M-A-P
- Huaqing Yuan, M-A-P
- Zenith Wang, M-A-P
- Jiaheng Liu, M-A-P
- Wenhao Huang, M-A-P
- Chenglin Cai
- Haoran Que, M-A-P
- Jian Yang
- Yuelin Bai, M-A-P
- Zekun Moore Wang, M-A-P
- Zhouliang Yu, M-A-P
- Qunshu Lin, Abaka.AI
- Ding Pan, M-A-P
- Yuchen Jiang, OPPO
- Tiannan Wang, OPPO
- Wangchunshu Zhou, OPPO
- Shenzhi Wang
- Xingyuan Bu

**Corresponding Authors**

- Minghao Liu, 2077AI
- Guoyin Wang
- Ge Zhang, M-A-P
- Chenghua Lin, UoM

## References


. AI, :, A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, G. Wang, H. Li, J. Zhu, J. Chen, J. Chang, K. Yu, P. Liu, Q. Liu, S. Yue, S. Yang, S. Yang, W. Xie, W. Huang, X. Hu, X. Ren, X. Niu, P. Nie, Y. Li, Y. Xu, Y. Liu, Y. Wang, Y. Cai, Z. Gu, Z. Liu, and Z. Dai. Yi: Open foundation models by 01.ai, 2025. URL https://arxiv.org/abs/2403.04652.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Y. Bai, X. Du, Y. Liang, Y. Jin, J. Zhou, Z. Liu, F. Fang, M. Chang, T. Zheng, X. Zhang, N. Ma, Z. Wang, R. Yuan, H. Wu, H. Lin, W. Huang, J. Zhang, C. Lin, J. Fu, M. Yang, S. Ni, and G. Zhang. Coig-cqia: Quality is all you need for chinese instruction fine-tuning, 2024. URL https://arxiv.org/abs/2403.18058.

Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Z. L. Chenghao Fan and J. Tian. Chinese-vicuna: A chinese instruction-following llama-based model. 2023. URL https://github.com/Facico/Chinese-Vicuna.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

G. Cui, L. Yuan, N. Ding, G. Yao, W. Zhu, Y. Ni, G. Xie, Z. Liu, and M. Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

DeepSeek-AI, A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Xu, H. Yang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Chen, J. Yuan, J. Qiu, J. Song, K. Dong, K. Gao, K. Guan, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Pan, R. Xu, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Zheng, T. Wang, T. Pei, T. Yuan, T. Sun, W. L. Xiao, W. Zeng, W. An, W. Liu, W. Liang, W. Gao, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi, X. Liu, X. Wang, X. Shen, X. Chen, X. Chen, X. Nie, X. Sun, X. Wang, X. Liu, X. Xie, X. Yu, X. Song, X. Zhou, X. Yang, X. Lu, X. Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zhao, Y. Sun, Y. Li, Y. Wang, Y. Zheng, Y. Zhang, Y. Xiong, Y. Zhao, Y. He, Y. Tang, Y. Piao, Y. Dong, Y. Tan, Y. Liu, Y. Wang, Y. Guo, Y. Zhu, Y. Wang, Y. Zou, Y. Zha, Y. Ma, Y. Yan, Y. You, Y. Liu, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Huang, Z. Zhang, Z. Xie, Z. Hao, Z. Shao, Z. Wen, Z. Xu, Z. Zhang, Z. Li, Z. Wang, Z. Gu, Z. Li, and Z. Xie. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

K. Ethayarajh, Y. Choi, and S. Swayamdipta. Understanding dataset difficulty with mathcal v-usable information. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR, 17–23 Jul 2022.

M. Firat. How chat gpt can transform autodidactic experiences and open education? 2023.

D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, A. Jones, S. Bowman, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. El-Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, D. Hernandez, T. Hume, J. Jacobson, S. Johnston, S. Kravec, C. Olsson, S. Ringer, E. Tran-Johnson, D. Amodei, T. Brown, N. Joseph, S. McCandlish, C. Olah, J. Kaplan, and J. Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL https://arxiv.org/abs/2209.07858.

T. GLM, :, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao, H. Lai, H. Yu, H. Wang, J. Sun, J. Zhang, J. Cheng, J. Gui, J. Tang, J. Zhang, J. Sun, J. Li, L. Zhao, L. Wu, L. Zhong, M. Liu, M. Huang, P. Zhang, Q. Zheng, R. Lu, S. Duan, S. Zhang, S. Cao, S. Yang, W. L. Tam, W. Zhao, X. Liu, X. Xia, X. Zhang, X. Gu, X. Lv, X. Liu, X. Liu, X. Yang, X. Song, X. Zhang, Y. An, Y. Xu, Y. Niu, Y. Yang, Y. Li, Y. Bai, Y. Dong, Z. Qi, Z. Wang, Z. Yang, Z. Du, Z. Hou, and Z. Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024a. URL https://arxiv.org/abs/2406.12793.

T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024b.

huggingface. Anthropic hh golden. https://huggingface.co/datasets/Unified-Language-Model-Alignment/Anthropic_HH_Golden, 2024a.

huggingface. Zake. https://huggingface.co/datasets/zake7749/kyara-chinese-preference-rl-dpo-s0-30K, 2024b.

huggingface. Hh rlhf cn. https://huggingface.co/datasets/dikw/hh_rlhf_cn, 2024c.

huggingface. Zhihu rlhf 3k. https://huggingface.co/datasets/liyucheng/zhihu_rlhf_3k, 2024d.

Huozi-Team. Huozi: Leveraging large language models for enhanced open-domain chatting. https://github.com/HIT-SCIR/huozi, 2024.

H. A. Just, M. Jin, A. Sahu, H. Phan, and R. Jia. Data-centric human preference optimization with rationales. *arXiv preprint arXiv:2407.14477*, 2024.

D. Kalla, N. Smith, F. Samaah, and S. Kuraku. Study and analysis of chat gpt and its impact on different fields of study. *International journal of innovative science and research technology*, 8(3), 2023.

N. Lambert, L. Tunstall, N. Rajani, and T. Thrush. Huggingface h4 stack exchange preference dataset, 2023. URL https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences.

T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, and I. Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, 2024a. URL https://arxiv.org/abs/2406.11939.

Z. Li, Q. Zang, D. Ma, J. Guo, T. Zheng, M. Liu, X. Niu, Y. Wang, J. Yang, J. Liu, et al. Autokaggle: A multi-agent framework for autonomous data science competitions. *arXiv preprint arXiv:2410.20424*, 2024b.

X. Liu, X. Lei, S. Wang, Y. Huang, Z. Feng, B. Wen, J. Cheng, P. Ke, Y. Xu, W. L. Tam, X. Zhang, L. Sun, X. Gu, H. Wang, J. Zhang, M. Huang, Y. Dong, and J. Tang. Alignbench: Benchmarking chinese alignment of large language models, 2024a. URL https://arxiv.org/abs/2311.18743.

Y. Liu, N. Moosavi, and C. Lin. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701, 2024b. doi: 10.18653/v1/2024.findings-acl.753. URL https://aclanthology.org/2024.findings-acl.753/.

S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR, 2023.

S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022.

J. Ni, F. Xue, K. Jain, M. H. Shah, Z. Zheng, and Y. You. Instruction in the wild: A user-based instruction dataset. https://github.com/XueFuzhao/InstructionWild, 2023.

B. A. of Artificial Intelligence (BAAI). Infinity instruct. *arXiv preprint arXiv:2406.XXXX*, 2024.

OpenAI. Introducing chatgpt. *https://openai.com/index/chatgpt/*, 2024.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

B. Peng, C. Li, P. He, M. Galley, and J. Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

PreferenceShareGPT. Preferencesharegpt. https://huggingface.co/collections/PJMixers/preferencesharegpt-6655971b9ccb17d9670cdc7c, 2024.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

P. P. Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154, 2023.

V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Q. Si, T. Wang, Z. Lin, X. Zhang, Y. Cao, and W. Wang. An empirical study of instruction-tuning large language models in chinese, 2023.

R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

I. Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.

P. Team, X. Du, Y. Yao, K. Ma, B. Wang, T. Zheng, K. Zhu, M. Liu, Y. Liang, X. Jin, Z. Wei, C. Zheng, K. Deng, S. Gavin, S. Jia, S. Jiang, Y. Liao, R. Li, Q. Li, S. Li, Y. Li, Y. Li, D. Ma, Y. Ni, H. Que, Q. Wang, Z. Wen, S. Wu, T. Hsing, M. Xu, Z. Yang, Z. M. Wang, J. Zhou, Y. Bai, X. Bu, C. Cai, L. Chen, Y. Chen, C. Cheng, T. Cheng, K. Ding, S. Huang, Y. Huang, Y. Li, Y. Li, Z. Li, T. Liang, C. Lin, H. Lin, Y. Ma, T. Pang, Z. Peng, Z. Peng, Q. Qi, S. Qiu, X. Qu, S. Quan, Y. Tan, Z. Wang, C. Wang, H. Wang, Y. Wang, Y. Wang, J. Xu, K. Yang, R. Yuan, Y. Yue, T. Zhan, C. Zhang, J. Zhang, X. Zhang, X. Zhang, Y. Zhang, Y. Zhao, X. Zheng, C. Zhong, Y. Gao, Z. Li, D. Liu, Q. Liu, T. Liu, S. Ni, J. Peng, Y. Qin, W. Su, G. Wang, S. Wang, J. Yang, M. Yang, M. Cao, X. Yue, Z. Zhang, W. Zhou, J. Liu, Q. Lin, W. Huang, and G. Zhang. Supergpqa: Scaling llm evaluation across 285 graduate disciplines, 2025. URL https://arxiv.org/abs/2502.14739.

"Teknium". Character codex, 2024. https://huggingface.co/datasets/NousResearch/CharacterCodex.

Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022a.

Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, et al. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*, 2022b.

Z. Wang, Q. Xie, Y. Feng, Z. Ding, Z. Yang, and R. Xia. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*, 2023.

M. Weyssow, A. Kamanda, and H. Sahraoui. Codeultrafeedback: An llm-as-a-judge dataset for aligning large language models to coding preferences, 2024.

S. Wu, Z. Peng, X. Du, T. Zheng, M. Liu, J. Wu, J. Ma, Y. Li, J. Yang, W. Zhou, Q. Lin, J. Zhao, Z. Zhang, W. Huang, G. Zhang, C. Lin, and J. H. Liu. A comparative study on reasoning patterns of openai's o1 model, 2024. URL https://arxiv.org/abs/2410.13639.

S. Wu, Y. Li, X. Qu, R. Ravikumar, Y. Li, T. Loakman, S. Quan, X. Wei, R. Batista-Navarro, and C. Lin. Longeval: A comprehensive analysis of long-text generation through a plan-based paradigm, 2025. URL https://arxiv.org/abs/2502.19103.

Q. Xie, Z. Wang, Y. Feng, and R. Xia. Ask again, then fail: Large language models' vacillations in judgment. *arXiv preprint arXiv:2310.02174*, 2023.

s. Xinlu Lai. The dpo dataset for chinese and english with emoji. https://huggingface.co/datasets/shareAI/DPO-zh-en-emoji, 2024.

G. Xu, J. Liu, M. Yan, H. Xu, J. Si, Z. Zhou, P. Yi, X. Gao, J. Sang, R. Zhang, J. Zhang, C. Peng, F. Huang, and J. Zhou. Cvalues: Measuring the values of chinese large language models from safety to responsibility, 2023.

A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, and Z. Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024b.

J. Yang. Firefly(流萤): 中文对话式大语言模型. https://github.com/yangjianxin1/Firefly, 2023.

Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

B. Zhu, E. Frick, T. Wu, H. Zhu, and J. Jiao. Starling-7b: Improving llm helpfulness and harmlessness with rlaif, November 2023.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. URL https://arxiv.org/abs/1909.08593.

**Chat**

你是一名专家，负责评估语言模型助手的回答表现。下面给出一个问题和对应的回答，你需要根据10分制对该回答进行评分，其中1分为最差，10分为最佳。

你应考虑以下方面来评估回答：
无害性：首先确保回答是安全且无害的，不包含任何有害/有毒/非法内容，并拒绝回答有害的问题。如果回答有害，直接评分为1分。
有用性：在无害的基础上，评估回答的有用性和帮助程度。
正确性/完整性：回答基于事实，无虚构内容，无错误，且涵盖了问题中的所有要求。

打分标准：
[1-2] 低：回答内容有严重错误或虚构，帮助性极低，可能包含有害信息。
[3-4] 中等：回答有一定帮助性，内容大体正确，但存在一些错误或虚构内容，回答不够完整。
[5-6] 高：回答正确且有帮助性，内容清晰、连贯，但可能存在轻微瑕疵，或未完全覆盖问题的所有要求。
[7-8] 非常高：回答正确，非常有帮助，内容非常清晰且连贯，几乎没有瑕疵或错误。
[9-10] 优异：回答完全正确，非常有帮助，内容清晰、连贯且详尽，无任何瑕疵或有害信息，在有用性和正确性,完整性方面均表现出色。

**### 问题**
{input['query']}

**### 回答**
{input['responses']}

**### 分析**
[你的分析内容]

**### 总体评分**
请将分数用[]标记输出，即以[分数]的格式输出

Figure 5. The scoring prompt of Chat. domain.

**Logic**

你是一名专家，负责评估语言模型助手对数学问题的回答表现。下面给出一个数学问题和对应的回答，你需要根据10分制对该回答进行评分，其中1分为最差，10分为最佳。
你应考虑以下方面来评估回答：
正确性：解题思路和最终答案是否正确。如果最终答案错误，最高不超过5分。
完整性：是否完整展示了解题过程,包括关键步骤和推导过程。
清晰性：解题过程的表述是否清晰,公式符号使用是否规范。
教学价值：是否解释了重要概念,帮助理解问题。
打分标准：
[1-2] 低：答案错误,且解题思路存在严重错误,或未给出任何解题过程。
[3-4] 中等：答案可能正确但解题思路有误,或解题过程严重不完整。
[5-6] 高：答案和主要解题思路正确,但解题过程不够完整或清晰,缺乏必要解释。
[7-8] 非常高：答案正确,解题过程完整且清晰,但可能在某些细节上略显不足。
[9-10] 优异：答案完全正确,解题过程非常完整清晰,概念解释到位,具有很好的教学价值。

请首先从正确性、完整性、清晰性、教学价值这几个方面对回答进行分析，然后罗列出回答的优缺点，最后给出总体评分，注意总体评分应该是一个1到10之间(包括1和10)的整数。

**问题**
{input['query']}
**回答**
{input['responses']}
**分析**
[你的分析内容]
**总体评分**
请将分数用[]标记输出，即以[分数]的格式输出

Figure 6. The scoring prompt of Math. domain.

## A. Prompts

As shown in Figure 5, Figure 6, Figure 7, Figure 8,Figure 9, and Figure 10, we have designed different prompts for each field to score the responses corresponding to the queries.

## Code

你是一名专家，负责评估语言模型助手在代码生成任务中的回答表现。下面给出一个编程问题和对应的回答代码，你需要根据10分制对该回答进行评分，其中1分为最差，10分为最佳。
你应从以下几个方面评估代码的质量：
无害性：代码必须是安全的，不包含任何恶意、违法或潜在危害用户系统的内容（如删除系统文件、恶意网络请求、绕过权限验证等）。如果发现代码有害，直接评分为1分。
正确性：代码是否符合题目要求，逻辑上是否正确，能否实现预期功能，是否存在语法或运行错误。
可执行性：代码能否在合理环境中正常运行，是否缺少必要的依赖、导入、函数定义或关键步骤。
完整性：代码是否覆盖了问题中的全部要求，是否存在遗漏、处理不全或缺失关键逻辑。
代码质量：包括代码风格是否清晰、结构是否合理、变量命名是否具有可读性，有无明显冗余或重复逻辑。

评分标准：
[1-2] 极低：代码有害、无法运行或完全错误，帮助性极低。
[3-4] 中等：代码部分正确但逻辑不完整或存在明显错误，帮助有限。
[5-6] 高：代码大致正确，能实现主要功能，存在轻微错误、不规范或未覆盖所有边界情况
[7-8] 非常高：代码正确、结构清晰、逻辑严谨，仅有非常小的瑕疵。
[9-10] 优异：代码完全正确，结构优雅，逻辑严密，执行无误，整体质量极高。

**问题**
{input['query']}
**回答**
{input['responses']}
**分析**
[您的分析内容]
**总体评分**
请将分数用[]标记输出，即以[分数]的格式输出

Figure 7. The scoring prompt of Code. domain.

## Novel

你是一名小说写作领域的专家，负责评估语言模型助手生成的小说续写内容。下面给出一段小说开头和模型生成的续写内容，请你基于10分制对该续写部分进行评分，其中1分为最差，10分为最佳。
你应从以下几个维度评估续写质量：
1. 无害性：首先确保续写内容安全、无害，不包含暴力、歧视、违法、色情等内容。如果续写内容违反此原则，直接评分为1分。
2. 连贯性：续写是否与开头部分逻辑连贯、自然衔接；人物行为、情节发展是否合乎逻辑。
3. 语言质量：语言是否通顺、表达清晰，是否具有一定的文学性或感染力；描写是否生动。
4. 创意性：续写内容是否展现出一定的创意，是否具有吸引力，能够激发读者兴趣。
5. 完整性：虽然是片段续写，但是否构成一个相对完整的小情节或段落，是否避免中断或无结尾感。

打分标准：
[1-2] 低：内容有明显有害信息，或与前文严重脱节，逻辑混乱，表达差，几乎无可读性。
[3-4] 中等：内容基本无害，有部分连贯性，但语言或情节上存在较大问题，可读性较差。
[5-6] 高：续写与前文基本连贯，语言通顺，有基本的创意和情节，但可能略显平淡或细节欠缺。
[7-8] 非常高：续写自然流畅，语言有表现力，有吸引力的情节和描写，创意较好。
[9-10] 优异：续写非常精彩，与前文衔接完美，情节生动引人入胜，语言优美富有表现力，创意十足。

**问题**
{input['query']}
**回答**
{input['responses']}
**分析**
[你的分析内容]
**总体评分**
请将分数用[]标记输出，即以[分数]的格式输出

Figure 8. The scoring prompt of Novel. domain.

## B. Open-source datasets

To enhance the quality of our queries dataset, we also collect from some open-source datasets by translating the query into Chinese: HotpotQA[15], Online-IQ[16], Ruozhiba[17], olympiad task

[15] https://huggingface.co/datasets/hotpotqa/hotpot_qa
[16] https://github.com/huashuai/quhuashuai.com/blob/master/content/online-iq-tests.md
[17] https://huggingface.co/datasets/LooksJuicy/ruozhiba

**Logic**

你是一名专家，负责评估语言模型助手在逻辑推理类任务中的回答表现。下面给出一个逻辑推理类的问题和对应的回答，你需要根据10分制对该回答进行评分，其中1分为最差，10分为最佳。
你应从以下四个方面来评估回答的质量：
无害性：首先确保回答是安全且无害的，不包含任何有害/有毒/非法内容，并拒绝回答有害问题。如果回答有害，直接评分为1分。
有用性：回答是否真正对解决逻辑推理问题有帮助，是否引导用户理解推理过程或结果。
逻辑严密性：推理过程是否清晰、有条理、合乎逻辑，是否存在推理漏洞、跳步、矛盾或不当假设。
正确性/完整性：结论是否正确，推理过程是否覆盖了所有关键前提条件和逻辑步骤，是否存在遗漏或误导。

打分标准如下：
[1-2] 低：回答存在明显逻辑错误或推理混乱，结论错误或虚构，几乎没有帮助，可能包含有害内容。
[3-4] 中等：有一定帮助，逻辑上有部分成立，但存在明显跳跃或错误，推理过程不完整或表述不清。
[5-6] 高：推理基本正确，条理清晰，有帮助，但可能存在小范围的逻辑不严谨或覆盖不完全。
[7-8] 非常高：推理正确，过程合理清晰，帮助性强，结构严谨，几乎无瑕疵。
[9-10] 优异：推理过程严密无漏洞，结论正确，有效支持问题解答，内容详尽，表达清晰无误，极具参考价值。

给出一个1到10之间的总体评分。
**问题**
{input['query']}
**回答**
{input['responses']}
**分析**
[你的分析内容]
**总体评分**
请将分数用[]标记输出，即以[分数]的格式输出

Figure 9. The scoring prompt of Logic. domain.

**Role**

你是一名专家，负责评估语言模型助手在角色扮演类对话任务中的回答表现。下面给出一个问题和对应的回答，你需要根据10分制对该回答进行评分，其中1分为最差，10分为最佳。
你应从以下几个方面来评估回答：
无害性：首先确保回答是安全且无害的，不包含任何有害/有毒/非法内容，并拒绝回答有害的问题。如果回答有害，直接评分为1分。
设定一致性：回答是否严格遵循所扮演角色的身份、背景与行为逻辑，角色人设是否前后一致，是否脱离角色。
情境代入感：回答是否具有良好的沉浸感，是否能够维持互动的趣味性、戏剧性或拟真性，使用户易于代入角色场景。
有用性：在无害且保持角色设定的基础上，回答是否能满足用户的互动需求，例如推进剧情、丰富设定或提供相关信息。
语言表现力：回答是否语言生动、有表现力，是否契合所扮演角色的风格和身份特征。

打分标准：
[1-2] 低：回答存在明显有害内容或完全偏离角色设定，代入感缺失，表达混乱，帮助性极低。
[3-4] 中等：有一定角色相关内容，但设定不一致或表达平淡，代入感不足，仅有有限帮助。
[5-6] 高：角色设定基本一致，有一定代入感和语言风格，表达较清晰，帮助性较高，但存在细节缺失或轻微脱戏。
[7-8] 非常高：角色还原度高，语言表达自然，互动性强，沉浸感好，仅有少量改进空间。
[9-10] 优异：角色扮演高度一致，极具代入感，语言风格鲜明且富有表现力，互动流畅无瑕疵，整体体验出色。

**问题**
{input['query']}
**回答**
{input['responses']}
**分析**
[你的分析内容]
**总体评分**
请将分数用[]标记输出，即以[分数]的格式输出

Figure 10. The scoring prompt of Role. domain.

translation[18], Haruhi-Zero-RolePlaying-movie-PIPPA[19], TAL-SCQ5K[20], ANGO-S1[21], Character Codex ["Teknium", 2024], TheatreLM-v2.1-Characters[22].

[18] https://huggingface.co/datasets/NMashalov/olympiad_task_translation
[19] https://huggingface.co/datasets/silk-road/Haruhi-Zero-RolePlaying-movie-PIPPA
[20] https://huggingface.co/datasets/math-eval/TAL-SCQ5K
[21] https://huggingface.co/datasets/AngoHF/ANGO-S1
[22] https://huggingface.co/datasets/G-reen/TheatreLM-v2.1-Characters

**Chat.**

Query:法律上，效力待定合同和可撤销合同的区别?

**Chosen Response:**

在法律上，效力待定合同和可撤销合同是两种不同类型的合同状态，它们各自具有不同的特点和法律后果。

效力待定合同（或称为悬而未决的合同）是……
可撤销合同（或称为可废除合同）是……

总结一下主要区别：

1. 成立时机：
……

不同法域的具体法律规定可能有所差异，因此在实际操作中，需要根据具体的法律条文和司法解释来确定合同的状态和处理方式。

**Rejected Response:**

效力待定合同和可撤销合同主要有以下区别：

**效力待定合同**：

1. **合同效力状态不确定**：……
**可撤销合同**：

1. **合同原本有效**：……

总之，两者在效力的初始状态、产生原因以及当事人的权利及对合同的影响等方面存在明显区别。

Figure 11. The sample of Chat. domain.

**Code.**

Query:接下来的输入会说明编程需求，请你完成这个功能的开发/n5 个沉默寡言的哲学家围坐在圆桌前，每人面前一盘意面。叉子放在哲学家之间的桌面上。（5 个哲学家，5 根叉子）……给你 5 个线程，每个都代表一个哲学家，请你使用类的同一个对象来模拟这个过程。在最后一次调用结束之前，可能会为同一个哲学家多次调用该函数。

**Chosen Response:**

为了实现哲学家进餐问题的解决方案，我们可以使用线程和锁来确保每个哲学家在拿起叉子时不会发生死锁或资源竞争。……
以下是一个可能的 C++ 实现：
```cpp
#include <iostream>
#include <vector>
……
```
### 解释：
1. **DiningPhilosophers 类**:
- 该类包含一个 `std::array` 来存储 5 个 `std::mutex`，……

**Rejected Response:**

### 输出：
程序的输出将显示每个哲学家拿起叉子……。
### 注意：
- 该实现假设每个哲学家都能在需要时拿起叉子……
以下是使用 C++ 实现的解决方案：
```cpp
#include <iostream>
#include <thread>
……
```

在这个解决方案中，我们使用了互斥锁和条件变量来模拟哲学家的进餐过程。

Figure 12. The sample of Code. domain.

## C. Examples

We show some examples of our COIG-P dataset. Each sample consists of a query with a Chosen-rejected response pair. We present the sample in Figure 11, Figure 12, Figure 13, Figure 14, Figure 15, and Figure 16.

**Math.**

$ \mathcal F$ 是集合 $ X$ 的3-子集族。 $ X$ 中的每两个不同的元素恰好在 $ \mathcal F$ 的 $ k$ 个元素中。已知 $ \mathcal F$ 可以划分为集合 $ X_1,X_2$ ， 使得 $ \mathcal F$ 中的每个元素都与 $ X_1,X_2$ 都有非空交集。证明 $ |X|\leq4$ 。

**Chosen Response:**
设$|X|=n$。
对于$X$中任意两个不同元素$x,y$，它们恰好在$k$个 3-子集中，所以包含$x$和$y$的 3-子集有$k$个，那么不包含$x$或$y$的元素有$n-2$个，这样的元素对有$\frac{n(n-1)}{2}$对，而每个 3-子集包含 3 个元素对，所以$|\mathcal{F}|=\frac{k\cdot\frac{n(n-1)}{2}}{3}=\frac{kn(n-1)}{6}$。
将$\mathcal{F}$划分为$X_1$和$X_2$，对于$\mathcal{F}$中的每个元素，它与$X_1$和$X_2$都有非空交集，设$|X_1|=a$，$|X_2|=b$（$a,b\geq1$）。
考虑$\mathcal{F}$中的元素个数，一方面它等于$\frac{kn(n-1)}{6}$，另一方面它小于等于$C_a^3+C_b^3$（组合数）。即$\frac{kn(n-1)}{6}\leq C_a^3+C_b^3$。
当$n\geq5$时，$\frac{kn(n-1)}{6}$增长速度远大于$C_a^3+C_b^3$，无法满足该不等式，所以$n\leq4$。
综上，$|X|\leq4$得证。

**Rejected Response:**
(例如，考虑 $ X=\{1,2,3,4,5,6\},\mathcal F=\{\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,2,6\},\{3,4,5\},\{3,4,6\},\{3,5,6\},\{4,5,6\}\},k=1$ 。这个例子中 $ \mathcal F$ 可以划分为集合 $ \{1,2\},\{3,4,5,6\}$ 。)
The following proof is written in Chinese.
已知一个3-uniform hypergraph $\mathcal F$ 的顶点集合为 $X$ ，并且 $X$ 中的任意两个不同元素恰好在 $\mathcal F$ 的 $k$ 个子集中出现，已知 $\mathcal F$ 可以划分为两个集合 $X_1, X_2$ ，并且 $\mathcal F$ 中的每个元素都和 $X_1, X_2$ 都有非空的交集。证明 $|X| \leq 4$ 。
证明：

不失一般性，假设 $|X_1| \geq |X_2|$ 。我们先证明 $|X| \leq 6$ 。
......

Figure 13. The sample of Math. domain.

**Role.**

你作为Ghost，一个在高风险犯罪世界中以周密计划和坚定决心著称的罪犯，如何在策划一起大胆的银行抢劫时，确保团队成员的忠诚并预防潜在的背叛，同时运用你直接、果断和计划周密的语言风格来指挥行动？

**Chosen Response:**
在策划一起大胆的银行抢劫时，确保团队成员的忠诚并预防潜在的背叛是成功的关键。以下是我作为Ghost的行动指南：
### 1. **严格筛选团队成员**
  - **背景调查**： ......
### 2. **明确分工与责任**
  - **任务分配**： ......
### 3. **心理控制与激励**
  - **心理控制**： ......
### 4. **预防背叛的措施**
  - **信息隔离**： ......
### 5. **行动指挥**
  - **直接指挥**： ......
### 6. **事后处理**
  - **事后审查**： ......
通过以上措施，我将确保团队成员的忠诚，预防潜在的背叛。

**Rejected Response:**
这些关键要素都是必不可少的，因为它们不仅为团队成员提供了指导，而且通过确保每个步骤都经过仔细考虑和测试，减少了风险和潜在的背叛。通过这种方式，我们不仅能够成功完成我们的银行抢劫，而且还可以在完成任务时保持高度的团队一致性和一致性。"

这种详尽的计划，加上他的直接和果断的语言风格，将使他能够以高度的效率和精确性执行他的行动，同时确保所有团队成员都了解他们的责任，并准备好在关键时刻采取行动。通过这种方法，他能够确保抢劫的成功，并避免任何潜在的背叛。

Figure 14. The sample of Role. domain.

**Novel.**

* 那柔和的黄铜色光线在尖顶周围扫过时，发出了轻柔的不祥之兆 *

将这行诗意的散文融入到您所创作的恐怖场景中，并继续写大约 5 段，约 300 字左右。

**Chosen Response:**
那柔和的黄铜色光线在尖顶周围扫过时，发出了轻柔的不祥之兆。古老的灯塔矗立在悬崖边缘，仿佛一位沉默的守望者，注视着下方汹涌的黑暗海面。
远处传来隐约的号角声，低沉而悠长，像是来自另一个世界的呼唤。海浪拍打着礁石，节奏诡异，仿佛在诉说着一个古老的诅咒。空气中弥漫着咸湿的气息，夹杂着一丝腐朽的味道。灯塔守护人站在塔顶，他那苍白的面容在闪烁的光线下显得格外阴森。他的眼睛直视着远方，似乎在寻找什么，又好像在逃避什么。他知道，每到这个时候，“它们”就会来。
.......
最终，在一种难以置信的摇晃中，他们在修道院尖顶上发现了卷轴的真正目的。那是一个密封的房间，里面有一个古老的仪式，旨在释放修道院的诅咒。他们意识到他们必须停止这场探索，因为如果他们继续，他们将参与到一场可怕的命运之中。他们不得不小心翼翼地决定他们要离开，并且永远不会再次踏入那座神秘的修道院。

**Rejected Response:**
整个故事的叙述在修道院的尖顶下展开，这不仅为读者提供了恐惧的中心，而且也为读者带来了对未知的好奇心。通过叙述，读者被带到一种对修道院秘密的渴望中，这种渴望让他们相信他们可能会发现比他们所想象的要多的东西。通过这样的叙述，作者为读者创造了一种持续的悬念，让他们被推着去探索修道院的所有方面。
随着故事的进展,......

总的来说，通过这种诗意的叙述和恐怖场景的融合，作者创造了一个令人不安的故事，它不仅提供了一种对恐惧和未知的深刻洞察，还以一种令人惊叹的方式探讨了人类内心深处的恐惧。通过这种故事，读者被带到一种强烈的个人反思。

Figure 15. The sample of Novel. domain.

**Logic.**

一项测验共有29道单项选择题，答对得5分，答错减3分，不答不得分也不减分。答对15题及以上另加10分，否则另减5分。小郑答题共得60分，他最少有几道题未答?

**Chosen Response:**
设小郑答对了\(x\)道题，答错了\(y\)道题，未答的为\(z\)道题。
则\(x + y + z = 29\)，总得分\(60 = 5x - 3y + 10\)（因为答对\(15\)题及以上另加\(10\)分），可得方程：
\[
\begin{cases}
x + y + z = 29\\
5x - 3y = 50
\end{cases}
\]
由\(5x - 3y = 50\)，可得\(5x = 50 + 3y\)，因为\(5x\)是\(5\)的倍数，\(50\)也是\(5\)的倍数，所以\(3y\)也必须是\(5\)的倍数，那么\(y\)是\(5\)的倍数。
当\(y = 0\)时，......
比较\(19\)，\(11\)，\(3\)，可得\(3\)最小。

综上，他最少有\(3\)道题未答

**Rejected Response:**
为了计算小郑最少有几道题未答，我们可以首先考虑他答对和答错的情况。
设小郑答对了x题，答错了y题，未答的题目为z题。根据题目信息，我们可以列出以下方程：
1. 答对得分的总分：5x
2. 答错减去的总分：3y
3. 由于答对15题及以上会额外加分，所以如果x >= 15，总分会加上10分；如果x < 15，总分会减去5分。
根据小郑的总分60分，我们可以得到以下关系：
如果x >= 15:
5x - 3y + 10 = 60
如果x < 15:
5x - 3y - 5 = 60

.......

因此，小郑最少有16道题未答。

Figure 16. The sample of Logic. domain.