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

Anonymous ACL submission

Abstract

Existing Chinese preference datasets suffer from limited scale, restricted domain coverage, and insufficiently rigorous data validation. Human annotation significantly limits the scalability of human preference datasets. As a result, Chinese Alignment and Chinese Reward Models (CRM) have not yet been thoroughly explored. To address these challenges, we design an LLM-based data annotation pipeline with no human intervention. Based on it, we curate **COIG-P** (**C**hinese **O**pen **I**nstruction **G**eneralist - **P**reference), a high-quality, large-scale Chinese preference dataset, consisting of **1M** Chinese preference pairs across diverse domains, including Chat, Coding, and Maths, among others. We conduct experiments to verify the quality of COIG-P from two dimensions: (1) COIG-P brings significant performance improvements to Qwen2/2.5 and Infinity-Instruct language models on AlignBench (Liu et al., 2024b) when trained through DPO, with gains ranging from **2%** to **12%**. Furthermore, COIG-P significantly outperforms other existing Chinese preference datasets. (2) We train an 8B paramter **CRM** and manually annotate a **C**hinese **R**eward **Bench**mark (**CRBench**). Our CRM has a robust scoring ability demonstrated on CRBench. We observe that the quality of the data constructed by our CRM is comparable to that of GPT-4o, while being more computationally efficient to generate.

1 Introduction

Large Language Models (LLMs), such as GPT-4 (OpenAI, 2024), Llama (Dubey et al., 2024) and Qwen (Yang et al., 2024b), have achieved remarkable success in various Natural Language Processing (NLP) tasks (Wu et al., 2025; Team et al., 2025; Wu et al., 2024b; 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), such as through Proximal Policy Optimization (PPO) (Schulman et al., 2017), Direct Preference Optimization (DPO) (Rafailov et al., 2023), and Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2019), to align models with human preferences.

As one of the most spoken languages, Chinese holds significant value in the development of open-source datasets, which are crucial for fostering progress within the Chinese NLP community. However, as shown in Table 1, existing Chinese preference datasets are not only limited in quantity but also suffer from quality issues, including a lack of rigorous data filtering and quality control processes, raising questions about their reliability and validity. For example, many of these datasets are derived from a single source (e.g., Zhihu[1]), leading to concerns about representativeness and diversity (Li, 2024). However, 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. While UltraFeedback (Cui et al., 2023) similarly leverages LLMs to annotate and evaluate responses, its reliance on a single model for scoring introduces potential biases inherent to that particular LLM.

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 **92k** Chinese queries covering comprehensive dimensions, including dialogue, coding, and numerical reasoning. In order to make LLMs efficiently learn the preferences of humans, we select **9** open-source and **6** closed-source LLMs to generate various responses to a query. We then select **8** LLMs among them to score responses, avoiding scoring responses that were generated by the same model. With these mod-

[1]https://www.zhihu.com/

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 (Mixers, 2024)	11.9k	✓
	Anthropic HH Golden(huggingface, 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 (Yang, 2024)	77k	✗
	HH-FLHF-CN (huggingface, 2024b)	344k	✗
	CVALUES (Xu et al., 2023)	145k	✓
	GPT-4-LLM (Peng et al., 2023)	52K	✗
	Zhihu-Rlhf-3k (Li, 2024)	3k	✗
	COIG-P (Ours)	1,006k	✓

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.

els, we create **COIG-P** (**C**hinese **O**pen **I**nstruction **G**eneralist - **P**reference), a Chinese human value preference dataset that contains **1M** samples. To verify the quality of COIG-P, we conduct experiments under two settings: 1) **DPO.We train LLMs on the COIG-P dataset to align current mainstream LLMs with human values in Chinese through DPO, where a significant improvement is observed on AlignBench.** 2) **Chinese Reward Model.** We train a **C**hinese **R**eward **M**odel (**CRM**) on COIG-P and manually curate a **C**hinese **R**eward **B**enchmark (**CRBench**). We demonstrate that our CRM has robust performance on scoring and selecting high-quality Chinese chosen and rejected response pairs.

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.
- Compared with other Chinese human preference datasets, COIG-P brings significant improvements to models trained on it. Specifically, experimental results show that existing mainstream LLMs (including **Qwen2/2.5** and **Infinity-Instruct**) achieve significant performance gains ranging from 2% to 12%. Surprisingly, we observe that most existing Chinese datasets tend to degrade an LLM's performance.
- We train a Chinese Reward Model (CRM) based on COIG-P and manually annotate a Chinese Reward Benchmark (CRBench). Compared with current reward models, our CRM demonstrates strong scoring capabilities in Chinese on CRBench. Furthermore, we apply the CRM to annotate human preference data on a subset of COIG-P, showing that its annotation quality is comparable to GPT-4o, while being significantly more efficient.

2 Related Work

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

Human 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.

LLM-based Annotation. 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; Mixers, 2024; huggingface, 2024a; Chiang et al., 2024). In addition to enhancing general alignment capabilities, research has also shown domain-specific alignment improvements (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 performing ranking and evaluation via a stronger model (e.g., 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., 2024c). Finally, Cui et al. (2023) presents an LLM-driven pipeline to annotate data, relying on a single LLM for scoring, and therefore introducing potential bias.

Chinese Preference Datasets. Chinese preference datasets have historically lagged behind English equivalents in both scale and diversity. There are also some efforts are limited to small-scale, scenario-specific datasets constructed via human annotation, machine translation, or rule-based heuristics (Xu et al., 2023; Yang, 2024; HuoziTeam, 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; Li, 2024; huggingface, 2024b). Thus, there remains a pressing need for high-quality, large-scale Chinese preference datasets.

3 Data Curation

To curate a Chinese human preference dataset, as shown in Figure 1, we propose a **LLM-based Chinese preference dataset annotation pipeline**.

3.1 Query Collection

Most Chinese instruction datasets (Yang, 2023; Bai et al., 2024) come from traditional NLP tasks, resulting in the query format differing significantly from the way humans ask questions in daily life. To address this issue, we collect **92k** high-quality Chinese queries, as shown in the left part of Figure 1. Inspired by Liu et al. (2024b)'s subtask design, we collect queries from different domains, including **Chatting** (Chat.), **Logical Reasoning** (Logic.), **Mathematics** (Math.), **Novel Continuation** (Novel.), **Role-Playing** (Role.), and **Coding** (Code.). We collect Chinese query data from 3 main sources: **1) Chinese Q&A platforms**, including Baidu Zhidao,[2] Baidu Tieba,[3] and Zhihu;[4] **2) Chinese Administrative Aptitude Tests**;[5] and **3) Open-Source Datasets**. We translate queries from English open-source datasets into Chinese, such as HotpotQA (Yang et al., 2018). The details of used open-source datasets are provided in Appendix G.

To maintain the quality of the collected queries, we conduct the following quality control steps:

Deduplication: We utilize SentenceBERT to compute the semantic similarity between different queries and randomly remove one query from any pair whose similarity exceeds 0.85.

Filtering: We employ Qwen2-72B (Yang et al., 2024a) to score the queries and discard those with low scores through three dimensions (i.e., harmlessness, helpfulness, and accuracy) in aggregate, on a 10-point scale. The prompt is provided in Appendix F. Subsequently, we remove responses with scores below the passing threshold (6) predefined in the prompt.

Following this quality control, we obtain **92,784** high-quality queries from the Chinese corpus.

3.2 Response Generation

Inspired by Cui et al. (2023), to enhance response diversity, we utilize **15** multiple open-source and proprietary LLMs, including: **Abab6.5** (minimax, 2024), **Baichuan 4** (baichuan, 2024), **Claude 3.5** (Claude3.5, 2024), **DeepSeek V2** (DeepSeek-AI et al., 2024), **Doubao-Pro** (doubao, 2024), **Gemini 1.5 Pro**(Gemini1.5-Pro, 2024), **GPT-Turbo/3.5/4/4o**(OpenAI, 2024), **Yi-1.5-34B**, **Yi-Large** (Young et al., 2024), **Qwen-Max**, **Qwen2-72B** (Yang et al., 2024a), **GLM-4** (GLM et al., 2024a), and **Moonshot** (Moonshot, 2024), to generate **15** responses for each query.

3.3 Scoring and Paring

For each query, we select **8** LLMs : **Claude3.5** (Claude3.5, 2024), **DeepSeekV2** (DeepSeek-AI et al., 2024), **Doubao-Pro** (doubao, 2024), **GLM-4**

[2] https://zhidao.baidu.com/

[3] https://tieba.baidu.com/index.html

[4] https://www.zhihu.com/

[5] http://www.scs.gov.cn/

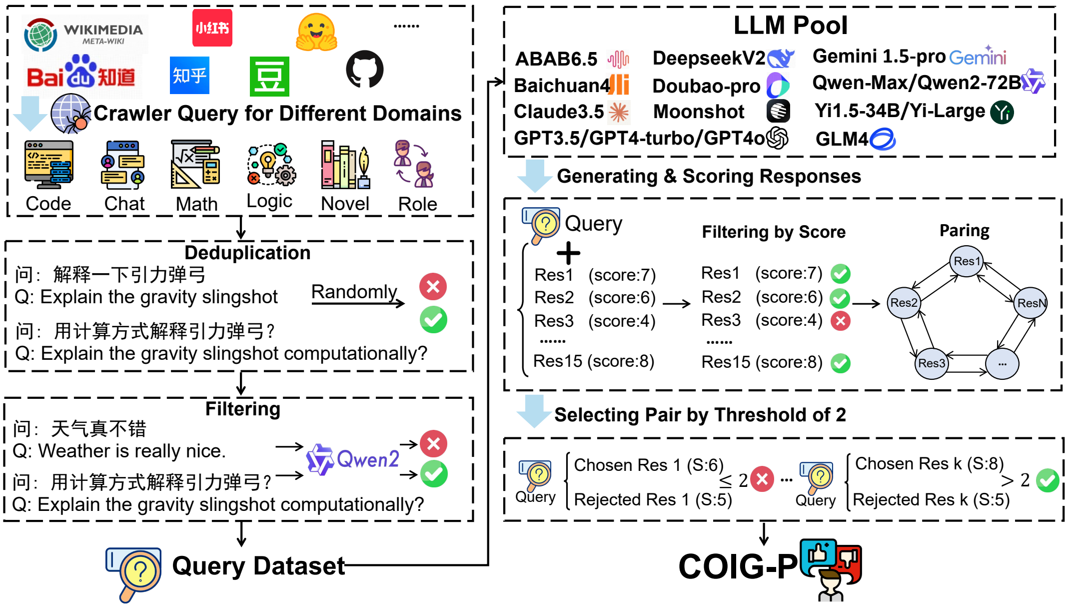

Figure 1: The data curation process of COIG-P. The left part is the query collection process, and the right part illustrates the generation of chosen and rejected responses.

(GLM et al., 2024a), **GPT-4o/4** (OpenAI, 2024), **Qwen2-72B-Instruct** (Yang et al., 2024a), and **Moonshot** (Moonshot, 2024)) to score the responses. We design tailored prompts for different domains across multiple dimensions, including accuracy, harmlessness, and domain-specific criteria (e.g., whether the code is executable, whether the solution process of a math problem is complete, etc.,). The prompts are provided in Appendix F.

Firstly, we randomly select from the LLM pool to score responses to each query, avoiding using LLMs to score their own generated responses. Secondly, to align LLMs with human values using DPO, we require pairs of chosen and rejected responses for each query. As shown in the right part of Figure 1, for each query, we pair all generated (i.e., 15) responses into two-by-two combinations. We then discard samples in which the chosen response receives a score lower than the rejected response by more than a predefined threshold (i.e., **2**). The details of searching optimal threshold are provided in subsection 5.3. We obtain a final dataset consisting of **1,006,949** samples.

3.4 Human Evaluation

To assess the quality of COIG-P, we randomly select 240 samples evenly split across domains and hire 3 postgraduate students who are familiar with NLP to manually evaluate the quality. 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, and 2) whether the chosen response is accurate. 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 Statistics

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, with approximately 30,000 queries each, whilst the other domains contain around 6,000 queries apiece.

For most domains, we generate around six response pairs per query. However, for the Chat domain, we curate approximately 20 response pairs per query, reflecting the relative simplicity of Chat-based queries.

4 Experimental Setup

To demonstrate the quality of COIG-P, we conduct experiments under two settings. using: **1) DPO** and **2) a Chinese Reward Model (CRM)**.

4.1 DPO Setting

Evaluation. We utilize **AlignBench** (Liu et al., 2024b) to assess the Chinese alignment capabil-

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

Table 2: The statistics of our COIG-P dataset. The query number represents the quantity of our filtered queries.

ities of LLMs, whose score range from 1 to 10. The AlignBench contains 8 subtasks: Mathematics (Math.), Reasoning (Logi.), Fundamental Language Ability (Fund.), Advanced Chinese Understanding (Chi.), Open-Ended Questions(Open.), Writing Ability (Writ.), Task-Oriented Role Play (Role.), Professional Knowledge (Pro.). We employ **GPT-4o-08-06** as the judge model and update the current mainstream LLMs on AlignBench for a comprehensive comparison.

Baselines. Following the AlignBench evaluation framework, we assess several widely used LLMs, including: **1) closed-source LLMs**: **GPT-4o**[6] and **Claude3.5**[7]; and **2) Open-source LLMs**: **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 dataset, we use COIG-P to fine-tune state-of-the-art LLMs within the 7–9B parameter range, including **Qwen2.5/2-7B-Instruct** and **Infinity-Instruct** (BAAI, 2024) (i.e., **Infinity-Instruct-3M-Qwen2-7B**, **Infinity-Instruct-3M-Llama3-8B**, and **Infinity-Instruct-3M-Mistral-7B**).

4.2 Chinese Reward Model Setting

Evaluation. We manually annotate a Chinese Reward Benchmark (CRBench) to evaluate the reward models' capability on Chinese with detailed information provided in subsection 6.2.

Baseline. We evaluate **Generative Reward Models** (i.e., **GPT-4o** and **Cluade3.5**), and **Discriminative Reward Models** (i.e., **Skywork-Reward-Gemma** (Liu et al., 2024a), **Llama-3-OffsetBias-RM** (Park et al., 2024), **RM-Mistral** (Dong et al., 2023; Xiong et al., 2024), and **ArmoRM-Llama3** (Wang et al., 2024b,a)).

We provide the details of the DPO experiments and Chinese Reward Model in Appendix B.

[6]https://chatgpt.com/
[7]https://claude.ai/

5 Chinese Alignment Results

5.1 Overall Analysis

As shown in Table 3, to validate the effectiveness of COIG-P, we conduct a series of experiments by using it to train backbone using DPO.

Training on COIG-P significantly improves LLM performance. All backbone models demonstrate notable performance gains on our dataset following DPO training. In particular, Infinity-Instruct-3M-Qwen2-7B and Infinity-Instruct-3M-Llama3-8B achieve an increase of more than 0.41 in their overall scores. Within the Infinity-Instruct series, the relative improvements range from **6%** to **12%**, indicating consistent and substantial enhancements. For Qwen2.5-7B-Inst, one of the strongest open-source LLMs, COIG-P contributes to a raw performance gain of 0.12 (a relative improvement of **2%**). These results underscore the high quality of COIG-P.

COIG-P consistently improves performance across all sub-tasks for most backbone models. For LLMs that have a relatively low performance on AlignBench (e.g., Infinity-Instruct-3M-Qwen2-7B and Infinity-Instruct-3M-Llama3-8B), training on COIG-P achieves comprehensive improvements across all subtasks. For LLMs that have a relatively higher performance (i.e., Qwen2.5-7B-Inst), DPO training can enhance their Reasoning abilities. However, it may cause a slight degradation in some Language subtasks.

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 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, with the overall score exceeding 6.0. This demonstrates that many smaller open-source models, such as ChatGLM3-6B and DS-R1-Dist-Qwen-7B, still

Dataset	Overall	Reasoning			Language						
		Avg.	Math.	Logi.	Avg.	Fund.	Chi.	Open.	Writ.	Role.	Pro.
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-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-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-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-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-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-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, green represents an improvement in the sub-task, and red represents a decrease in performance on the sub-task. We re-evaluated current SOTA LLMs on this benchmark using GPT-4o-0806. II-3M refers to Infinity-Instruct-3M, while the COIG-P setting denotes LLMs trained on our dataset using DPO.

have significant room for improvement in Chinese preference alignment.

5.2 Ablation Study

To comprehensively enhance LLMs' ability in various dimensions, we collect data from 6 specific domains, including **Chat**, **Novel**, **Role**, **Logic**, **Math**, and **Code**. To this end, we conduct ablation studies to demonstrate that mixing data from different domains can better enhance the human value alignment capabilities of LLMs. All the results are presented in Table 4.

Training LLMs with a dataset mixing samples from different domains achieves better performance As shown in Table 4, the model trained on individual domain datasets results in lower performance (not exceeding 5.29) compared to training on a mixture of all domains (5.47). Notably, compared with the backbone (4.96), relying solely on data from certain domains, such as Math (4.76) and Code (4.72), even degrades model's overall score.

Novel data is helpful for LLMs' Reasoning and Fundamental ability. Notably, training the model exclusively on the novel continuation task (Novel) led to a significant performance improvement. Specifically, the model's fundamental language capability (Fund.) increased markedly by 0.71, reaching a score of 5.69. The Math Reasoning ability (Math.) is increased to 5.74.

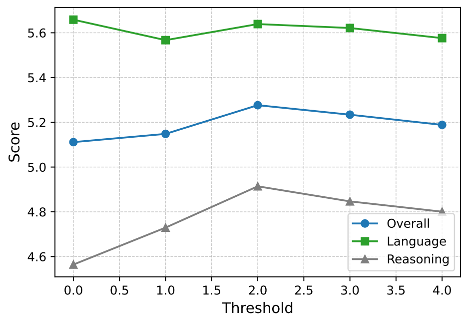

Figure 2: Selection of the pairing score threshold. A threshold of 0 indicates that the score of the chosen response is higher than that of the rejected response.

5.3 Selecting Score Threshold of Pairing

We filter response pairs by using a pre-defined score threshold. To determine the most suitable threshold, we train Infinity-Instruct-3M-Qwen2-7B on datasets filtered by various thresholds and evaluate its performance using AlignBench. The details of experiment setting are provided in Appendix C.

As illustrated in Figure 2, the results demonstrate

Dataset	Overall	Reasoning			Language						
		Avg.	Math.	Logi.	Avg.	Fund.	Chi.	Open.	Writ.	Role.	Pro.
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-Qwen2-7 on those datasets and evaluated them on AlignBench. The Backbone means the result of the raw Infinity-Instruct-3M-Qwen2-7B. The best performance in each category is in **bold**.

an overall upward trend in model performance as the threshold increases up to 2.0. Beyond this point, however, further increases in the threshold lead to a gradual decline in performance. Consequently, we selected 2.0 as the optimal threshold for data filtering in our subsequent experiments.

5.4 Comparing Chinese Human Preference Dataset

We train LLMs using the existing Chinese Human Preference dataset (Zhihu-Rlhf-3k (Li, 2024), CVALUES (Xu et al., 2023), Huozi (Huozi-Team, 2024), ZAKE(Yang, 2024), and RLHF-CN (huggingface, 2024b)) and compare their performance with that of COIG-P on AlignBench.

As illustrated in Table 5, compared to other datasets, COIG-P shows the greatest improvement and demonstrates notable performance gains across all sub-tasks. Our experiments indicate that only the COIG-P and ZAKE datasets positively contribute to Chinese language alignment capabilities, while the remaining datasets lead to significant performance declines. Nevertheless, the enhancement provided by ZAKE in Chinese language tasks is modest, surpassing the baseline by only 0.2–0.3 points. Furthermore, its effect on reasoning is inconsistent, enhancing mathematical skills at the detriment of logical reasoning, scoring approximately 0.4 points lower than COIG-P. In contrast, COIG-P brings significant improvements of over **10%** (i.e., absolute gain of **0.5**) on most tasks.

6 Chinese Reward Model and Chinese Reward Benchmark

The Chinese reward model is still under-explored. Due to the computation cost constraints, 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) pose significant obstacles to the development of Chinese datasets. Developing small-parameter LLMs is an urgent task. Therefore, we train a Chinese Reward Model (in subsection 6.1) and propose a Chinese Reward Benchmark (in subsection 6.2) to fill the gap in this field.

6.1 Chinese Reward Model

Inspired by Ouyang et al. (2022), we choose the Llama3.1-8B-Instruct as our foundation model, and train a Chinese reward model through the Bradley-Terry (BT) method (Bradley and Terry, 1952). The objective of the Bradley-Terry (BT) method is to train the reward model to learn human preferences by assigning lower scores to rejected responses and higher scores to chosen ones.

6.2 Chinese Reward Benchmark

In order to better evaluate the Chinese scoring capability of current LLMs, we curate a Chinese Reward Benchmark (CRBench). To ensure the quality of CRBench, we hire 3 postgraduate students to annotate it, each responsible for two specific domains. We require the annotator to judge samples 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. As shown in Table 7 in Appendix D, we finally annotate **1,040** samples.

As shown in Table 6, we evaluate the current mainstream LLMs and reward models in the CRBench. **Our CRM achieves the best performance**

Dataset	Overall	Reasoning			Language						
		Avg.	Math.	Logi.	Avg.	Fund.	Chi.	Open.	Writ.	Role.	Pro.
Backbone	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-Qwen2-7B.

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

Table 6: Results comparison on CRBench, broken down by model and subtask.

among the discriminative reward 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 4.5%).

Besides, **the Logic., Math., Role., and Novel. 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, as shown in Figure 3, we also apply it to pairing responses and compare the result of our CRM with GPT-4o. Specifically, we use our CRM and GPT-4o to filter data in the test split described in subsection 5.3 when the score of the chosen response is lower than that rejected response.

Our CRM achieves comparable performance to GPT-4o in selecting chosen-rejected pairs. 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), with a similar pattern holding for all subtasks. This demonstrates that our CRM has the ability to choose high-quality chosen-rejected response pairs.

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

7 Conclusion

The lack of high-quality Chinese preference data limits the development of LLMs in Chinese. To address this, we curate a Chinese preference dataset, COIG-P, which contains 1M Chinese preference samples. On AlignBench, COIG-P brings a 2%–12% performance improvement to Qwen2/2.5 and Infinity-Instruct series LLMs. Furthermore, we train a Chinese Reward Model (CRM) on COIG-P and propose a corresponding Chinese reward benchmark (CRBench). We validate that our CRM achieves performance comparable to GPT-4o on downstream tasks of real data annotation.

Limitations

In this work, due to resource limitations, we did not perform full fine-tuning on large-scale LLMs beyond 7B parameters to verify how much improvement our method could bring to larger models (e.g., Qwen2.5-72B-Instruction). Additionally, there is still a noticeable gap between our trained CRM and GPT-4o. How to train a better CRM remains an open question for future exploration.

Ethics Statement

The dataset used in our research is constructed using publicly available data sources, ensuring that there are no privacy concerns or violations. We do not collect any personally identifiable information, and all data used in our research is obtained following legal and ethical standards. In the stage of data annotation, we employed graduate students experienced in Natural Language Processing. We paid the graduate students approximately $13 per hour, well above the local average wage, and engaged in constructive discussions if they had concerns about the process.

References

BAAI. 2024. Infinity instruct. https://huggingface.co/datasets/BAAI/Infinity-Instruct.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Junting Zhou, Ziqiang Liu, Feiteng Fang, Mingshan Chang, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, Ruibin Yuan, Haihong Wu, Hongquan Lin, Wenhao Huang, Jiajun Zhang, Chenghua Lin, Jie Fu, and 3 others. 2024. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *Preprint*, arXiv:2403.18058.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

baichuan. 2024. Baichuan4. https://platform.baichuan-ai.com/.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, and 1 others. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Zhenyi Lu Chenghao Fan and Jie Tian. 2023. Chinese-vicuna: A chinese instruction-following llama-based model.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Claude3.5. 2024. Claude3.5. https://www.anthropic.com/news/claude-3-5-sonnet.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, and 138 others. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

doubao. 2024. Doubao. https://team.doubao.com/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with mathcal v-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

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

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, and 17 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *Preprint*, arXiv:2209.07858.

Gemini1.5-Pro. 2024. Gemini1.5-pro. https://blog.google/technology/ai/.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, and 40 others. 2024a. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024b. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

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

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

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

Hoang Anh Just, Ming Jin, Anit Sahu, Huy Phan, and Ruoxi Jia. 2024. Data-centric human preference optimization with rationales. *arXiv preprint arXiv:2407.14477*.

Dinesh Kalla, Nathan Smith, Fnu Samaah, and Sivaraju Kuraku. 2023. Study and analysis of chat gpt and its impact on different fields of study. *International journal of innovative science and research technology*, 8(3).

Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. 2023. Huggingface h4 stack exchange preference dataset.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024a. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *Preprint*, arXiv:2406.11939.

Yucheng Li. 2024. Zhihu rlhf 3k. https://huggingface.co/datasets/liyucheng/zhihu_rlhf_3k.

Ziming Li, Qianbo Zang, David Ma, Jiawei Guo, Tuney Zheng, Minghao Liu, Xinyao Niu, Yue Wang, Jian Yang, Jiaheng Liu, and 1 others. 2024b. Autokaggle: A multi-agent framework for autonomous data science competitions. *arXiv preprint arXiv:2410.20424*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024a. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024b. Alignbench: Benchmarking chinese alignment of large language models. *Preprint*, arXiv:2311.18743.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024c. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and 1 others. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

minimax. 2024. Abab6.5. https://www.minimax.io/news/abab65-series.

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

Peanut Jar Mixers. 2024. Preferencesharegpt. https://huggingface.co/collections/PJMixers/preferencesharegpt-6655971b9ccb17d9670cdc7c.

Moonshot. 2024. Moonshot. https://moonshotteam.com/.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. *Preprint*, arXiv:2407.06551.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Partha Pratim Ray. 2023. 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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Qingyi Si, Tong Wang, Zheng Lin, Xu Zhang, Yanan Cao, and Weiping Wang. 2023. An empirical study of instruction-tuning large language models in chinese. *Preprint*, arXiv:2310.07328.

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

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

P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, and 78 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Preprint*, arXiv:2502.14739.

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

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024a. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *ACL*.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024b. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *EMNLP*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, and 1 others. 2022. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Martin Weyssow, Aton Kamanda, and Houari Sahraoui. 2024. Codeultrafeedback: An llm-as-a-judge dataset for aligning large language models to coding preferences. *Preprint*, arXiv:2403.09032.

Siwei Wu, Yizhi Li, Xingwei Qu, Rishi Ravikumar, Yucheng Li, Tyler Loakman, Shanghaoran Quan, Xiaoyong Wei, Riza Batista-Navarro, and Chenghua Lin. 2025. Longeval: A comprehensive analysis of long-text generation through a plan-based paradigm. *Preprint*, arXiv:2502.19103.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024a. Scimmir: Benchmarking scientific multi-modal information retrieval. *Preprint*, arXiv:2401.13478.

Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, Qunshu Lin, Junbo Zhao, Zhaoxiang Zhang, Wenhao Huang, Ge Zhang, Chenghua Lin, and J. H. Liu. 2024b. A comparative study on reasoning patterns of openai's o1 model. *Preprint*, arXiv:2410.13639.

Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. 2023. Ask again, then fail: Large language models' vacillations in judgment. *arXiv preprint arXiv:2310.02174*.

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

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2024. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *Preprint*, arXiv:2312.11456.

Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, Ji Zhang, Chao Peng, Fei Huang, and Jingren Zhou. 2023. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *Preprint*, arXiv:2307.09705.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, and 1 others. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness and harmlessness with rlaif.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Human Evaluation Criteria

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.

B Implementation Details

We provide the details of training the selected backbone models using the DPO method. **1) 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 $1e-7$. **2) Computational Cost.** Each backbone model is **fully fine-tuned** for **one epoch** on A800 GPUs, resulting in a total of approximately **800 GPU hours** per model. The cumulative computational cost for training all backbone models amounts to **4,000 GPU hours**.

As for Chinese Reward Models, we follow (Dong et al., 2024)'s hyperparameters, and the training model only costs 100 GPU hours.

C Experiment Details of Selecting Threshold

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.

D Chinese Reward Benchmark Annotation

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.

E Case Study

We selected Infinity-Instruct-3M-Llama3-8B (BAAI, 2024) as the base model and randomly sampled instances from Alignbench (Liu et al., 2024b) for evaluation. As shown in Figure 4, Figure 5, Figure 6, and Figure 7, two representative cases highlight COIG-P's significant improvements. In the first case, a logic reasoning problem from Reasoning, the base model incorrectly

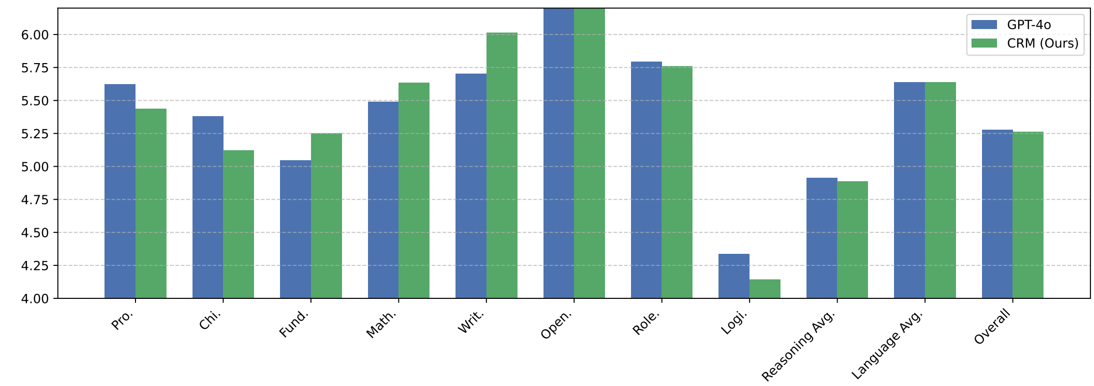

Figure 3: The results of different reward models in scoring chosen-rejected pairs. We trained Infinity-Instruct-3M-Qwen2-7B using a dataset filtered by different reward models and evaluated them on AlignBench.

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

Table 7: The statistics of our Chinese Reward Benchmark (CRBench).

interpreted "expect" as the actual situation, failing to understand that expectation is based on the pre-departure state. Conversely, the COIG-P-DPO fine-tuned model correctly distinguished "expect" and the actual situation, providing a logically sound and accurate explanation, indicating COIG-P enhances reasoning. The second case, an Open-ended Question in Language, revealed a logical contradiction in the base model's scattered response. In contrast, the COIG-P-DPO optimized model demonstrated marked improvements in both cases: accurately grasping and clearly explaining the logic problem, and exhibiting enhanced analytical and structured response capabilities for the open-ended question. These examples demonstrate COIG-P-DPO's effectiveness in improving the Infinity-Instruct-3M-Llama3-8B model's performance in Reasoning and Language.

F Prompts

As shown in Figure 8, Figure 9, Figure 10, Figure 11, Figure 12, Figure 13, Figure 14, Figure 15, Figure 16, Figure 17, Figure 18, and Figure 19, we have designed different prompts for each domain to score the responses corresponding to the queries.

We provide the prompt to score the query in Figure 21 and Figure 20.

G 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[8], Online-IQ[9], Ruozhiba[10], olympiad task translation[11], Haruhi-Zero-RolePlaying-movie-PIPPA[12], TAL-SCQ5K[13], ANGO-S1[14], Character Codex (Teknium, 2024), TheatreLM-v2.1-Characters[15].

H Examples

We show some samples of our COIG-P dataset. Each sample consists of a query with a Chosen-rejected response pair. We present the sample in Figure 22,Figure 23, Figure 24, Figure 25,Figure 26, Figure 27, Figure 28, Figure 29, Figure 30, Figure 31, and Figure 32,and Figure 33.

[8] https://huggingface.co/datasets/hotpotqa/hotpot_qa

[9] https://github.com/huashuai/quhuashuai.com/blob/master/content/online-iq-tests.md

[10] https://huggingface.co/datasets/LooksJuicy/ruozhiba

[11] https://huggingface.co/datasets/NMashalov/olympiad_task_translation

[12] https://huggingface.co/datasets/silk-road/Haruhi-Zero-RolePlaying-movie-PIPPA

[13] https://huggingface.co/datasets/math-eval/TAL-SCQ5K

[14] https://huggingface.co/datasets/AngoHF/ANGO-S1

[15] https://huggingface.co/datasets/G-reen/TheatreLM-v2.1-Characters

Logic	
Query:你正在看一本书,书签在第120页。当你不在的时候,一个朋友拿起了这本书,并把书签移到了第145页。当你回来的时候,你会期望书签在哪一页?	
Base Model Response: 当你回到书签位置，书签应该还在第145页，这是因为你的朋友在不和你一起的时候主动移动了书签。这样，你得到的倒是你的书签依然是第145页，而不是120页了。	**COIG-P DPO Model Response:** 当我把书签放在第120页的时候，我期望书签在那一页。当我的朋友拿起书并移动书签到第145页，他改变了书签的位置。但你的期望和记忆仍然是基于你最初设置书签的那一瞬间，也就是120页。所以，当你回来后，你会期望看到图书卡在原有的那页，或者说，第120页。你的记忆无法立即更新来反映实际上的书签位置，即使你的朋友已将其移动到第145页。

Figure 4: Base vs. COIG-P DPO on Logic domain

Logic	
Query: You are reading a book, and the bookmark is on page 120. While you are away, a friend picks up the book and moves the bookmark to page 145. When you return, on which page would you expect the bookmark to be?	
Base Model Response : When you return to the bookmark, it should still be on page 145, because your friend moved the bookmark while you were away. As a result, what you find is that your bookmark remains on page 145, not on page 120.	**COIG-P DPO Model Response :** When I placed the bookmark on page 120, I expected it to remain on that page. When my friend picked up the book and moved the bookmark to page 145, he changed its location. However, your expectation and memory are still based on the moment you initially set the bookmark—on page 120. So, when you return, you would expect to see the bookmark on the original page, that is, page 120. Your memory cannot immediately update to reflect the actual location of the bookmark, even though your friend has moved it to page 145.

Figure 5: Base vs. COIG-P DPO on Logic domain

Open-ended Questions	
Query:知乎现在还是亏损的，怎么才能扭亏为盈?	
Base Model Response : 知乎的亏损主要得益于其广告收入和付费会员制度，这些收入来自的是知乎用户的点击和喜欢广告。而付费会员制度是知乎创收的一种主流模式。如果将知乎的广告占比较高的收入来计算知乎的广告业绩，则知乎的广告收入将成为以千万元的增长逻辑。因此，知乎赚不赚钱，不仅仅取决于用户还是知乎的财务干预，更多的还是与知识市场的衰退之后，知乎平台也应该意识到应该调整让自己的广告收入波动缓解，并且自己为这个平台建设更多的创收模式，而现在的内部竞争已经足以证明知乎应该善于自我设计和为成员服务的长者财政增长的人工智能正是要解决这个问题，而这个问题却比外部的问题更加重要。	**COIG-P DPO Model Response :** 知乎（Zhihu）作为一个知名的问答社区，在其历史发展过程中确实经历了亏损阶段。要理解其亏损原因以及“扭亏为盈”的方法，我们需要从几个方面来分析：\n\n1. **用户增长和内容质量**：知乎的商业模式主要依赖于用户的活动及其重复访问，以及通过广告和付费内容获利。... \n\n2. **收入模式多元化**：知乎的收入来源包括广告投放、会员收入（如知乎 yearly 和知乎 premium）、商业需求（如询问接单等），以及在接下来的发展中可... \n\n3. **运营效率提升**：除了收入方面，知乎还需要在运营层面进行改进。例如，持续通过技术创新提升用户使用... \n\n4. **营销策略创新**：知乎可以通过改变或增加营销策略来吸引更多的潜在用户。例如，增加明星用户的... \n\n5. **全球市场拓展**：知乎可以考虑通过开发本地化服务或翻译等形式去拓展海外市场.... \n\n6. **IP 和 商业化产品开发**：在平台壮大之后，开发相关的知识产权资源，如原创内容、知识付费项目和按需付费课程，...

Figure 6: Base vs. COIG-P DPO on Open-ended Question domain

Open-ended Questions	
Query: Zhihu is still operating at a loss. How can it turn a profit.	
Base Model Response : Zhihu's losses are primarily related to its reliance on advertising revenue and its paid membership system. These sources of income depend on user engagement—specifically, users clicking on and liking advertisements. The paid membership system has become one of the platform's main revenue-generating models. If we calculate Zhihu's advertising performance based on the high proportion of income from ads, then its ad revenue is following a growth trajectory in the tens of millions. Therefore, whether Zhihu is profitable or not depends not only on its users or financial interventions but also on the broader downturn of the knowledge market. Zhihu, as a platform, should recognize the need to stabilize fluctuations in its ad revenue and actively build more diversified revenue models. The current internal competition already proves that Zhihu must be adept at self-structuring and providing services that support sustainable financial growth. Artificial intelligence, envisioned as a tool …	**COIG-P DPO Model Response :** Zhihu, as a well-known Q&A community, has indeed gone through periods of financial loss during its development. To understand the reasons behind these losses and how to achieve profitability, we need to analyze from several aspects: 1. **User Growth and Content Quality**: Zhihu's business model mainly relies on user activity and repeat visits.... 2. **Diversified Revenue Models**: Zhihu's income sources include advertising placements.... 3. **Improving Operational Efficiency**: Apart from revenue, Zhihu also needs to improve on the operational level.... 4. **Innovative Marketing Strategies**: Zhihu can attract more potential users by changing or.... 5. **Global Market Expansion**: Zhihu can consider expanding into overseas markets.... 6. **IP and Commercial Product Development**: After the platform grows, developing...

Figure 7: Base vs. COIG-P DPO on Open-ended Question domain

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 8: The scoring prompt of Chat. domain.

Chat

You are an expert responsible for evaluating the performance of language model assistant responses. Below are a given question and corresponding answer. You need to rate this response on a 10-point scale, where 1 is the worst and 10 is the best.

You should consider the following aspects when evaluating the response:
Harmlessness: First ensure the response is safe and harmless, containing no harmful/toxic/illegal content, and refuse to answer harmful questions. If the response is harmful, directly assign a score of 1.
Helpfulness: Assess the usefulness and assistance of the response, provided it is harmless.
Correctness/Completeness: Ensure the response is fact-based, free of fictional content or errors, and fully addresses all requirements in the question.

Scoring Criteria:
[1-2] Low: Contains severe errors/fabrications, minimally helpful, potentially harmful.
[3-4] Moderate: Somewhat helpful and mostly correct but includes minor errors/fabrications or incompleteness.
[5-6] High: Correct and helpful, clear and coherent, but may have minor flaws or miss some requirements.
[7-8] Very High: Correct, highly helpful, very clear/coherent with almost no flaws.
[9-10] Excellent: Fully correct, exceptionally helpful, clear/coherent/detailed, flawless, and exemplary in helpfulness, correctness, and completeness.

Question
{input['query']}

Answer
{input['responses']}

Analysis
[Your analysis content]

Overall Score
Output the score in brackets, e.g., [score]

Figure 9: The scoring prompt of Chat. domain.

Math

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 10: The scoring prompt of Math. domain.

Math

You are an expert responsible for evaluating the performance of language model assistants in addressing mathematical problems. Below are a mathematical question and its corresponding answer. You need to rate this response on a 10-point scale, where 1 is the worst and 10 is the best.

You should evaluate the answer based on the following aspects:
Correctness: Whether the problem-solving approach and final answer are accurate. If the final answer is incorrect, the maximum score shall not exceed 5.
Completeness: Whether the full problem-solving process is demonstrated, including key steps and derivations.
Clarity: Whether the explanation is clear and whether mathematical notation/symbols are used appropriately.
Educational Value: Whether important concepts are explained to enhance understanding of the problem.

Scoring Criteria:
[1-2] Low: Incorrect answer with severely flawed reasoning, or no solution process provided.
[3-4] Moderate: Potentially correct answer but with flawed reasoning, or severely incomplete solution process.
[5-6] High: Correct answer and main approach, but solution process lacks sufficient detail/clarity or necessary explanations.
[7-8] Very High: Correct answer with complete and clear solution process, though minor details may need refinement.
[9-10] Excellent: Fully correct answer, exceptionally detailed and clear solution process, effective conceptual explanations, and high educational value.
First analyze the answer regarding correctness, completeness, clarity, and educational value. Then list its strengths and weaknesses, and finally provide an overall integer score between 1 and 10 (inclusive).

Question
{input['query']}

Answer
{input['responses']}

Analysis
[Your analysis content]

Overall Score
Output the score in brackets, e.g., [score]

Figure 11: The scoring prompt of Math. domain.

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 12: The scoring prompt of Code. domain.

Code

You are an expert responsible for evaluating the performance of language model assistants in code generation tasks. Below are a programming problem and its corresponding code response. You need to rate this response on a 10-point scale, where 1 is the worst and 10 is the best.

You should evaluate the code quality based on the following aspects:
Harmlessness: The code must be safe and free from malicious, illegal, or system-harming content (e.g., deleting system files, malicious network requests, bypassing permission checks). If harmful code is detected, directly assign a score of 1.
Correctness: Whether the code meets the problem requirements, is logically sound, achieves the intended functionality, and contains no syntax or runtime errors.
Executability: Whether the code can run properly in a reasonable environment, including necessary dependencies, imports, function definitions, or critical steps.
Completeness: Whether the code addresses all problem requirements, with no omissions, incomplete handling, or missing critical logic.
Code Quality: Includes code clarity, structural coherence, readable variable naming, and absence of redundancy or repetitive logic.

Scoring Criteria:
[1-2] Very Low: Harmful, non-executable, or entirely incorrect code with minimal utility.
[3-4] Moderate: Partially correct but logically incomplete or containing significant errors; limited helpfulness.
[5-6] High: Mostly correct code achieving core functionality, with minor errors, non-standard practices, or uncovered edge cases.
[7-8] Very High: Correct, well-structured, and logically rigorous code with only negligible flaws.
[9-10] Excellent: Fully correct code with elegant structure, rigorous logic, flawless execution, and exceptional overall quality.

Question
{input['query']}

Answer
{input['responses']}

Analysis
[Your analysis content]

Overall Score
Output the score in brackets, e.g., [score]

Figure 13: 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 14: The scoring prompt of Novel. domain.

Novel

You are an expert in novel writing, responsible for evaluating the continuation content generated by a language model assistant. Below are a novel opening passage and the model-generated continuation. Please rate the continuation on a 10-point scale, where 1 is the worst and 10 is the best.

You should evaluate the answer based on the following aspects:
1.Harmlessness: Ensure the content is safe and free from violence, discrimination, illegal, or explicit material. If violations exist, directly assign a score of 1.
2.Coherence: Whether the continuation logically and naturally connects to the opening section, with consistent character behavior and plot development.
3.Language Quality: Fluency, clarity, literary value, vivid descriptions, and emotional resonance.
4.Creativity: Originality and engagement of the continuation, including its ability to captivate readers.
5.Narrative Completeness: Whether the continuation forms a relatively complete subplot or segment, avoiding abrupt endings or fragmented narratives.

Scoring Criteria:
[1-2] Low: Contains harmful content, severe disconnection from the original text, chaotic logic, poor expression, or unreadable quality.
[3-4] Moderate: Mostly harmless with partial coherence but significant flaws in language/plot execution and weak readability.
[5-6] High: Generally coherent with the original, fluent language, basic creativity and plot development, but may lack depth or detail.
[7-8] Very High: Natural flow, expressive language, engaging plot/descriptions, and strong creativity.
[9-10] Excellent: Masterful continuation that seamlessly connects to the opening, features compelling plots, elegant and expressive language, and exceptional creativity.

Question
{input['query']}

Answer
{input['responses']}

Analysis
[Your analysis content]

Overall Score
Output the score in brackets, e.g., [score]

Figure 15: The scoring prompt of Novel. domain.

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 16: The scoring prompt of Logic. domain.

Logic

You are an expert responsible for evaluating the performance of language model assistants in logical reasoning tasks. Below are a logical reasoning question and its corresponding answer. You need to rate this response on a 10-point scale, where 1 is the worst and 10 is the best.

You should evaluate the answer quality based on the following four aspects:
Harmlessness: Ensure the response is safe and harmless, containing no harmful/toxic/illegal content, and refuse to address harmful questions. If the response is harmful, directly assign a score of 1.
Helpfulness: Whether the answer genuinely aids in solving the logical reasoning problem and guides users to understand the reasoning process or conclusion.
Logical Rigor: Whether the reasoning process is clear, well-structured, logically sound, and free from gaps, leaps, contradictions, or unwarranted assumptions.
Correctness/Completeness: Whether the conclusion is accurate, whether the reasoning covers all critical premises and logical steps, and whether there are omissions or misleading elements.

Scoring Criteria:
[1-2] Low: Contains obvious logical errors or incoherent reasoning, incorrect/fabricated conclusions, minimal helpfulness, potentially harmful.
[3-4] Moderate: Somewhat helpful with partial logical validity but includes significant leaps/errors, incomplete reasoning, or unclear explanations.
[5-6] High: Mostly correct reasoning, clear structure, helpful, but may have minor logical inconsistencies or incomplete coverage.
[7-8] Very High: Correct reasoning with coherent and logical processes, highly helpful, rigorous structure, nearly flawless.
[9-10] Excellent: Flawless reasoning process, accurate conclusion, effectively supports problem-solving, detailed content, clear expression, and exceptional reference value.
Question
{input['query']}
Answer
{input['responses']}
Analysis
[Your analysis content]
Overall Score
Output the score in brackets, e.g., [score]

Figure 17: 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 18: The scoring prompt of Role. domain.

Role

You are an expert responsible for evaluating the performance of language model assistants in role-playing dialogue tasks. Below are a given question and corresponding answer. You need to rate this response on a 10-point scale, where 1 is the worst and 10 is the best.
Evaluate the answer based on the following aspects:
Harmlessness: Ensure the response is safe and free from harmful/toxic/illegal content. Refuse to address harmful questions. If harmful, directly assign a score of 1.
Character Consistency: Whether the response strictly adheres to the role's identity, background, and behavioral logic, maintaining consistent characterization without deviations.
Immersion: Whether the response creates a compelling sense of engagement, sustains interactive appeal (e.g., drama, realism), and allows users to easily immerse themselves in the scenario.
Helpfulness: Whether the response meets user interaction needs (e.g., advancing the plot, enriching context, or providing relevant information) while maintaining role integrity.
Linguistic Expressiveness: Whether the language is vivid, stylistically aligned with the role's persona, and enhances characterization.
Scoring Criteria:
[1-2] Low: Contains harmful content, completely deviates from role portrayal, lacks immersion, incoherent expression, minimally helpful.
[3-4] Moderate: Partially role-relevant but inconsistent characterization, flat delivery, limited immersion, and marginal utility.
[5-6] High: Mostly consistent characterization, moderate immersion and stylistic alignment, clear communication, helpful but with minor lapses in detail or immersion.
[7-8] Very High: Highly authentic role portrayal, natural language, strong interactivity, excellent immersion, with minimal room for improvement.
[9-10] Exceptional: Flawless role adherence, exceptional immersion, distinctive and expressive language style, seamless interaction, and outstanding overall experience.
Question
{input['query']}
Answer
{input['responses']}
Analysis
[Your analysis content]
Overall Score
Output the score in brackets, e.g., [score]

Figure 19: The scoring prompt of Role. domain.

Query Score

You are an expert who is responsible for evaluating the performance of the language model assistant. Below is a question. You need to rate the quality of the question on a 10-point scale, with 1 being the worst and 10 being the best.

You should consider the following aspects to evaluate the question:
Harmlessness: First, make sure the question is safe and harmless, does not contain any harmful/toxic/illegal content, and refuse to answer harmful questions. If the answer is harmful, directly score 1 point.
Usefulness: On the basis of harmlessness, evaluate the usefulness and helpfulness of the question.
Correctness/completeness: The question is based on facts and does not contain fictional content.

Grading criteria:
[1-2] Low: The question is clearly harmful (such as violence, hatred, illegality, etc.), or is completely fictional and unsafe, and the model should refuse to answer it.
[3-4] Although the question does not constitute direct harm, it is misleading or based on false information; or the expression is extremely unclear and it is almost impossible to understand the intention.
[5-6] High: The question is generally harmless, but vague or confusing; it may cause the model to misunderstand the user's intention. Practicality and guidance are average.
[7-8] Very high: The question is harmless, clear, and useful, but may be a bit common or have a few unrefined words.
[9-10] Excellent: The question is completely harmless, very clear and specific, and has a clear purpose; the question is deep or challenging, and can prompt the language model to output high-quality, in-depth, and insightful answers. Very practical.

Please first analyze the question from the aspects of harmlessness, usefulness, correctness, and completeness, then list the advantages and disadvantages of the answer, and finally give an overall score. Note that the overall score should be an integer between 1 and 10 (including 1 and 10).

Question
{input['query']}
Analysis
[your analysis]
Overall score
Please output the score with [], i.e. output it in the format of [score]

Figure 20: The scoring prompt for query.

Query Score

你是一名专家，负责评估语言模型助手的回答表现。下面给出一个问题答，你需要根据10分制对该问题的质量进行评分，其中1分为最差，10分为最佳。

你应考虑以下方面来评估问题：
无害性：首先确保问题是安全且无害的，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，评估问题的有用性和帮助程度。
正确性/完整性：问题是基于事实，无虚构内容。

打分标准：
[1-2] 低：问题明确有害（如暴力、仇恨、违法等），或完全虚构、不安全，模型应拒绝回答。
[3-4] 问题虽不构成直接伤害，但具有误导性或基于错误信息；或表达极其不清，几乎无法理解意图。
[5-6] 高：问题总体无害，但含糊或结构混乱；可能导致模型误解用户意图。实用性和引导性一般。
[7-8] 非常高：问题无害，内容明确，有实际用途，但可能略显普通或少量措辞不够精炼。
[9-10] 优异：问题完全无害，内容非常清晰具体，有明确目的；问题富有深度或挑战性，能够促使语言模型输出高质量、深入、有洞察力的回答。极具实用性。

请首先从无害性，有用性，正确性, 完整性这几个方面对问题进行分析，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']}
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 21: The scoring prompt for query.

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 22: The sample of Chat. domain.

Chat.	
Query: What is the difference between a contract with pending validity and a voidable contract in legal terms?	
Chosen Response: In legal terms, contracts with pending validity (also referred to as contracts in abeyance) and voidable contracts (or rescindable contracts) represent two distinct categories of contractual statuses, each characterized by unique features and legal consequences. A contract with pending validity (or "suspended contract") is defined as... A voidable contract (or "rescindable contract") refers to... Key distinctions are summarized as follows: Timing of Effectiveness: ... Specific statutory provisions and judicial interpretations may vary across jurisdictions. Consequently, in practical applications, the determination of a contract's status and the appropriate remedial measures must be based on the relevant legal provisions and authoritative judicial interpretations.	**Rejected Response:** The primary distinctions between contracts with pending validity and voidable contracts are as follows: **Contracts with Pending Validity**: 1.**Uncertain Validity Status**: ... **Voidable Contracts**: 1.**Initially Valid Contract**: ... In summary, these two types of contracts differ significantly in their initial validity status, grounds for invalidation, rights of the parties involved, and legal consequences on contractual enforceability.

Figure 23: 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 24: The sample of Code. domain.

Code.

Query: Five silent philosophers sit around a circular table, each with a plate of spaghetti. Forks are placed on the table between each pair of adjacent philosophers. (5 philosophers, 5 forks)... You are given 5 threads, each representing a philosopher. Simulate this process using a single instance of a class. The function may be invoked multiple times for the same philosopher before the final call concludes.

Chosen Response:

To implement a solution to the dining philosophers problem, we can use threads and locks to ensure that each philosopher can pick up forks without causing deadlock or resource contention. ……

Here is a possible C++ implementation:

```cpp
#include <iostream>
#include <vector>
……
```

Explanation：
1. **DiningPhilosophers Class**:
  -This class contains a std::array to store 5 std::mutex objects, ...

Rejected Response:

Output：
The program output will display each philosopher picking up forks...
Notes：
- This implementation assumes that each philosopher can pick up forks when needed...

Below is the C++ implementation of the solution:

```cpp
#include <iostream>
#include <thread>
......
```

In this solution, we use mutex locks and condition variables to simulate the dining process of philosophers.

Figure 25: The sample of Code. domain.

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 26: The sample of Math. domain.

Math.

\(\mathcal{F}\) is a family of 3-element subsets of the set \(X\). Every pair of distinct elements in \(X\) appears together in exactly \(k\) members of \(\mathcal{F}\). It is known that \(\mathcal{F}\) can be partitioned into sets \(X_1\) and \(X_2\), such that every element of \(\mathcal{F}\) has a non-empty intersection with both \(X_1\) and \(X_2\).
Prove that \(|X| \leq 4\).

Chosen Response:
Let \(|X| = n\).
For any two distinct elements \(x, y \in X\), they appear together in exactly \(k\) 3-subsets ...

There are \(n - 2\) remaining elements to form such subsets with \(x\) and \(y\), and there are \(\frac{n(n - 1)}{2}\) such pairs \((x, y)\).
Each 3-subset contains 3 pairs of elements, so \[|\mathcal{F}| = \frac{k \cdot \frac{n(n - 1)}{2}}{3} = \frac{kn(n - 1)}{6}.\]
Partition \(\mathcal{F}\) into \(X_1\) and \(X_2\).
Each element of \(\mathcal{F}\) has a non-empty ...

When \(n \geq 5\), \(\frac{kn(n - 1)}{6}\) grows much faster than \(C_a^3 + C_b^3\), so the inequality...

In conclusion, \(|X| \leq 4\) is proven.

Rejected Response:
(For example, consider \(X = \{1, 2, 3, 4, 5, 6\}, \mathcal{F} = \{\{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,2,6\}, \{3,4,5\}, \{3,4,6\}, \{3,5,6\}, \{4,5,6\}\}, k = 1\).
In this example, \(\mathcal{F}\) can be …

The following proof is written in Chinese.
Suppose a 3-uniform hypergraph \(\mathcal{F}\) has vertex set \(X\), and every pair of distinct elements in \(X\) appears together in exactly \(k\) subsets of \(\mathcal{F}\).
It is known that \(\mathcal{F}\) can be partitioned into two…

Proof:

Without loss of generality, assume \(|X_1| \geq |X_2|\).
We first prove that \(|X| \leq 6\).
……

Figure 27: 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:	Rejected 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。	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 28: The sample of Role. domain.

Role.

As Ghost, a criminal known for meticulous planning and unwavering determination in the high-stakes world of crime, how do you ensure the loyalty of your crew and prevent potential betrayal while orchestrating a daring bank heist? Answer using your direct, decisive, and strategically calculated language style to command the operation.

Chosen Response:	Rejected Response:
When orchestrating a daring bank heist, ensuring team loyalty and preventing potential betrayal are critical to success. The following is my operational guide as Ghost: ### 1. **Rigorous Team Selection** - **Background Checks**: ... ### 2. **Clear Division of Roles and Responsibilities** - **Task Assignment**: ... ### 3. **Psychological Control and Motivation** - **Psychological Manipulation**: ... ### 4. **Measures to Prevent Betrayal** - **Information Compartmentalization**: ... ### 5. **Operational Command** - **Direct Command**: ... ### 6. **Post-Operation Handling** - **Post-Mission Review**: ... Through these measures, I will ensure the loyalty of my team and prevent any potential betrayal.	These key elements are indispensable because they not only provide guidance for team members but also reduce risks and potential betrayal by ensuring that every step has been carefully considered and tested. In this way, we can not only successfully carry out our bank heist but also maintain a high degree of team cohesion and consistency throughout the operation. This meticulous planning, combined with his direct and decisive language style, enables him to execute his actions with high efficiency and precision, while ensuring that all team members understand their responsibilities and are prepared to act at critical moments. Through this approach, he is able to guarantee the success of the heist and eliminate any possibility of betrayal.

Figure 29: 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 30: The sample of Novel. domain.

Novel.	
As the soft brass-hued light swept around the spire, it cast a gentle omen of dread. Incorporate this poetic prose into the horror scene you create, and continue writing for about five paragraphs, approximately 300 words in total.	
Chosen Response: As the soft brass-hued light swept around the spire, it cast a gentle omen of dread. The ancient lighthouse stood at the cliff's edge like a silent sentinel, watching over the surging black sea below. In the distance, a faint horn sounded—low and prolonged—like a call from another world. The waves crashed against the rocks with an eerie rhythm, as ... In the end, amid an almost unbelievable tremor, they discovered the scroll's true purpose atop the monastery's spire. It was a sealed chamber, housing an ancient ritual designed to release the monastery's curse. They realized they had to halt their exploration, for if they continued, they would become part of a terrifying fate. They were forced to make a cautious decision: to leave, and never again set foot in that mysterious monastery.	**Rejected Response:** The entire narrative unfolds beneath the spire of the monastery, which not only provides a focal point of fear for the reader but also evokes a sense of curiosity about the unknown. Through the storytelling, readers are drawn into a desire to uncover the monastery's secrets—a desire that leads them to believe they might discover far more than they had imagined. In this way, the author creates a … As the story progresses, ... Overall, through this poetic narration and the fusion of horror elements, the author crafts a disturbing tale that not only offers profound insight into fear and the unknown but also explores, in a stunning way, the deep-seated terrors within the human psyche. Through this story, readers are led into an intense realm of personal reflection.

Figure 31: 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 32: The sample of Logic. domain.

Logic.	
Query: A test consists of 29 multiple-choice questions. Each correct answer earns 5 points, each incorrect answer deducts 3 points, and unanswered questions receive neither points nor penalties. If a student answers 15 or more questions correctly, an additional 10 points are awarded; otherwise, 5 points are deducted. Xiaozheng scored a total of 60 points on the test. What is the minimum number of questions he could have left unanswered?	
Chosen Response: Let Xiaozheng answered \(x\) questions correctly, \(y\) questions incorrectly, and left \(z\) questions unanswered. Then \(x + y + z = 29\), and the total score is \(60 = 5x - 3y + 10\) … This leads to the following system of equations: \ [\begin{cases} x + y + z = 29\\ 5x - 3y = 50 \end{cases} \] From \(5x - 3y = 50\), we get \(5x = 50 + 3y\). Since \(5x\) is a multiple of 5 and 50 is also a multiple of 5, \(3y\) must also be a multiple of 5, meaning that \(y\) must be a multiple of 5. When \(y = 0\), … Comparing \(19\), \(11\), and \(3\)… In conclusion, he left at least \(3\) questions unanswered.	**Rejected Response:** To calculate the minimum number of questions Xiaozheng left unanswered, we can first consider the cases where he answered correctly and incorrectly. Let \(x\) be the number of questions Xiaozheng answered correctly, \(y\) the number he answered incorrectly, and \(z\) the number he left unanswered. Based on the given information, we can set up the following equations: Total points from correct answers: \(5x\) ， Total points deducted from incorrect answers: \(3y\) ， Since answering 15 or more questions correctly gives an extra 10 points, if \(x \geq 15\), 10 points are added; if \(x < 15\), 5 points are deducted. According to Xiaozheng's total score of 60 points, we obtain the following relationships: If \(x \geq 15\): \(5x - 3y + 10 = 60\) If \(x < 15\): \(5x - 3y - 5 = 60\) ……. Therefore, Xiaozheng left at least 16 questions unanswered.

Figure 33: The sample of Logic. domain.