COIG-CQIA: Quality is All You Need for Chinese Instruction Fine-tuning

Yuelin Bai[1*] **Xinrun Du**[2*] **Yiming Liang**[3*] **Yonggang Jin**[2*] **Junting Zhou**[2,4*]
Ziqiang Liu[1] **Feiteng Fang**[5] **Mingshan Chang**[1] **Tianyu Zheng**[2] **Xincheng Zhang**[5]
Nuo Ma[6] **Zekun Wang**[2] **Ruibin Yuan**[2,7] **Haihong Wu**[5] **Hongquan Lin**[5] **Wenhao Huang**[6]
Jiajun Zhang[3] **Chenghua Lin**[2,10] **Jie Fu**[7] **Min Yang**[1] **Shiwen Ni**[1†] **Ge Zhang**[8,9†]
[1]Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]M-A-P [3]Institute of Automation, Chinese Academy of Sciences
[4]Peking University [5]University of Science and Technology of China [6]01.ai [7]HKUST
[8]University of Waterloo [9]Vector Institute [10]University of Manchester

Abstract

Remarkable progress on English instruction tuning has facilitated the efficacy and reliability of large language models (LLMs). However, there remains a noticeable gap in instruction tuning for Chinese, where the complex linguistic features pose significant challenges. Existing datasets, generally distilled from English-centric LLMs, are not well-aligned with Chinese users' interaction patterns. To bridge this gap, we introduce COIG-CQIA, a new Chinese instruction tuning dataset derived from various real-world resources and undergoing rigorous human verification. We conduct extensive experiments on COIG-CQIA, and compare them with strong baseline models and datasets. The experimental results show that models trained on COIG-CQIA achieve highly competitive performance in diverse benchmarks. Additionally, our findings offer several insights for designing effective Chinese instruction-tuning datasets and data-mixing strategies. Our dataset are available at https://huggingface.co/datasets/m-a-p/COIG-CQIA.

1 Introduction

Large Language Models (LLMs) have shown remarkable capabilities as general-purpose assistants. The cornerstone of this advancement is instruction tuning (Zhang et al., 2023b), which significantly enhances the efficacy and safety of models in following human instructions. The core idea is to train models with instruction-output pair data, thus aligning the model's training objective with human intent. This highlights the key role of high-quality instruction tuning datasets in enabling LLMs to function as efficient and reliable assistants. Despite the remarkable progress made in English instruction tuning datasets, datasets for Chinese instruction tuning still remain in the nascent stages. Existing datasets can be roughly categorized into three types: (1) Datasets derived from English instruction datasets (Peng et al., 2023) or NLP datasets (BAAI, 2023; Yang, 2023), (2) Datasets synthesized by LLMs (Guo et al., 2023; Ji et al., 2023; Sun et al., 2023), and (3) Hybrid dataset constructed using different methods (Zhang et al., 2023a). To improve dataset quality, COIG (Zhang et al., 2023a) leveraged multiple methods to construct a human-verified instruction corpus. However, two major challenges still exist in prior Chinese instruction tuning datasets. First, they suffer from insufficient alignment with real-world Chinese users due to the lack of naturally occurring human-generated data. Second, they are still riddled with quality issues due to the high cost of comprehensive human verification. Moreover, it is still under-explored how different data sources impact the downstream Chinese tasks, exacerbating the challenges in constructing Chinese datasets.

To address these challenges, we introduce COIG-CQIA (**C**hinese **O**pen **I**nstruction **G**eneralist - **Q**uality **I**s **A**ll You Need), a new Chinese instruction tuning dataset, distinguished by its incorporation of diverse real-world data resources and rigorous human verification processes. Inspired by LIMA (Zhou et al., 2023), COIG-CQIA focuses on curating a dataset from diverse Chinese internet sources, covering social media and forums, comprehensive encyclopedias, challenging examinations, and existing linguistic corpus. These data undergo a thorough cleaning, restructuring, and careful human verification to secure the quality and diversity. Our aim is to enhance the proficiency of LLMs in following Chinese instructions and executing downstream tasks. We conduct extensive experiments to explore how different data sources impact various downstream tasks and explored the benefits of different data mixing strategies. Additionally, we integrate COIG-CQIA with English data to

[*] Equal contribution.
[†] Corresponding authors.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 8205–8220

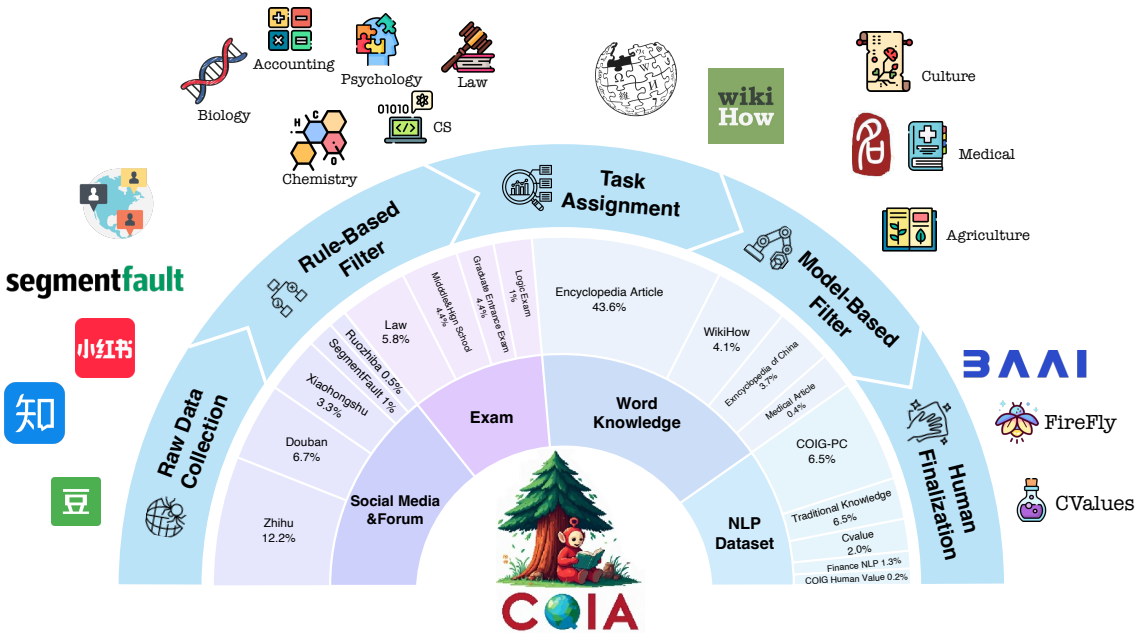

Figure 1: Overview of COIG-CQIA and Statistics of each data source.

investigate the performance of trained models in multilingual scenarios. Further experiments show that COIG-CQIA achieves highly competitive results compared to other Chinese datasets and strong baseline models. Our main contributions are as follows:

- **Resource.** We introduce COIG-CQIA, a high-quality Chinese instruction-tuning dataset built from diverse, real-world sources and verified by humans to ensure quality.
- **Performance.** We demonstrate the effectiveness of COIG-CQIA through extensive experiments, showing its competitiveness against other Chinese datasets and baseline models.
- **Insights.** We systematically investigate the impact of different data sources and mixing strategies on downstream task performance, providing insights into data source influence and training strategy.

2 COIG-CQIA CURATION

To ensure data quality and diversity, we curated data from 18 high-quality Chinese Internet sources. We also integrated existing Chinese NLP datasets and examinations to broaden task diversity. Specifically, we categorized all data sources into four types: Social Media & forums, World Knowledge, NLP tasks, and Exam. The statistical information of the data are detailed in the table 1.

2.1 Raw Data Collection

In this initial stage, we aggregated data from diverse sources from the Internet, including social media platforms, encyclopedias, and specialized websites. For instance, we used a web crawler to collect posts from Zhihu, SegmentFault, etc., as well as entries from encyclopedic sources such as the Encyclopedia of China. For the crawled HTML content, we carefully converted them into question-answer pairs or documents. We preserved as much non-text metadata as possible, such as likes, comments, authors, multimedia elements, etc., to facilitate rule-based filtering based on this metadata. Meanwhile, we collected publicly available official exam papers from previous years and used the Mathpix tool [2] to extract the questions and detailed answers from the documents. In Table 1, we mark all the sources that are processed into plain text from crawled or non-text corpora using ■.

2.2 Rule-Based Filter

The rule-based filter is implemented subsequent to the extraction of plain text data. Its primary purpose is to perform preliminary data cleansing, eliminating content that contains harmful or inappropriate information, multimedia elements, or advertisements. Additionally, it removes data that fails to meet specified length criteria or lacks wide human acceptance (e.g., posts with very few likes

[2] https://mathpix.com/

Source	Type	Description	Quantity	Data Processing
Zhihu (A.1)	Forum	Comprehensive Q&A platform	8837	■ ■ ■ ■
Segment Fault (A.1)	Forum	Tech community for IT developers	458	■ ■ ■
Douban (A.1)	Social Media	User-driven platform focused on literature and the arts	3132	■ ■ ■ ■
Xiaohongshu (A.1)	Social Media	Life experiences sharing platform	1508	■ ■ ■ ■
Ruozhiba (A.1)	Forum	Tieba[1] subcommunity interested in logical traps.	240	■ ■ ■ ■
Encyclopedia Article	World Knowledge	Comprehensive encyclopedic knowledge from various website	20020	■ ■ ■ ■
Encyclopedia of China(A.2.1)	World Knowledge	Comprehensive Chinese encyclopedia	1706	■ ■ ■ ■
WikiHow (A.2.1)	World Knowledge	Step-by-step guides and how-tos	1876	■ ■ ■
Medical Ariticle(A.2.2)	World Knowledge	Health-related knowledge	186	■ ■ ■
Middle&High School Exam(A.3)	Exam	Standardized exam for middle&high school students	2000	■ ■ ■
Graduate Entrance Exam(A.3)	Exam	National graduate entrance exam	475	■ ■ ■
Logical Exam(A.3)	Exam	Logistic reasoning exam questions	422	■ ■ ■
Law Exam(A.3)	Exam	Law graduate entrance exam questions	2645	■ ■ ■
COIG PC(A.4)	NLP Dataset	A massive dataset for instruction fine-tuning	3000	■ ■
COIG-Human-Value(A.4)	NLP Dataset	Value-related tasks from COIG	101	■ ■
CValues(A.4)	NLP Dataset	Detoxifying answers written by experts for harmful questions	906	■
Chinese Traditional (A.3)	NLP Dataset	Chinese traditional culture tasks from various datasets	1111	■ ■ ■
Finance NLP Task(A.4)	NLP Dataset	NLP tasks in the financial domain	600	■ ■
Total			**45173**	

Table 1: Overview of different data sources. We list each data source's source type, description, quantity, and data processing stages. The colored squares represent different stages we applied to the data processing: ■ Raw Data Collection; ■ Rule-based filter; ■ Template Curation; ■ Model-Based Filter; ■ Human Finalization.

in forums). This stage is essential while highly efficient, capable of reducing the dataset from millions of entries to hundreds of thousands. Rule-based filtering was applied to nearly all data sources, as marked by green squares ■ in the table 1.

2.3 Task Assignment

This stage is used to convert raw posts and articles into instruction-response formats for instruction fine-tuning. Overall, we designed a variety of instruction or response templates based on the characteristics of the content from different data sources. For example, for encyclopedic entries, the tasks focus mainly on concept explanation, while for metadata-rich sources like Douban, we designed tasks around reviews writing, recommendation, etc. All the details of the data construction process are described in the appendix A. We assign tasks to each data source, but not all of them require template design. Therefore, Table 1 specifically marks those needing Template Curation with orange squares ■.

2.4 Model-Based Filter

While data from social media and forums closely reflects real human interactions and offers great diversity, it's challenging to ensure that all of this data is harmless and accurate. Model-based filtering (marked by pink squares ■) can help eliminate low-quality data that's difficult to remove through rule-based or metadata filtering. This typically includes irrelevant instruction-response pairs, soft advertisements, and potentially harmful content. We used GPT-4 to filter the data sources, as it's widely used in LLMs-as-judge and demonstrates a high correlation with human judgment in assessing data quality. We detail this process for specific sources in the appendix A.

2.5 Human Finalization

To ultimately ensure data quality, we invite human reviewers to re-examine all the data and finalize each data source. Consistent with our model-based filtering criteria, we asked human judges to evaluate the data's usefulness, professionalism, logical coherence, level of detail, objectivity, and harmlessness. Given the flexibility and variability of Chinese language usage, human review allows for filtering out cases that rule-based and model-based approaches struggle to identify. This process also involves making appropriate modifications to instruction-response pairs to ensure accuracy and alignment between responses and the instructions' intent. We detail this process in the appendix A.

3 Data Analysis

3.1 Statistics

We collected a total of 45,173 instances from 18 sources within the Chinese Internet and Community, covering domains ranging from general knowledge and STEM to humanities. Table 1 describe the data statistics for all sources. We demonstrated the distribution in the length of the instructions and responses in Figure 5.

3.2 Semantic Distribution

To visualize and analyze the semantic diversity of our dataset, we employed U-MAP to create a distribution map of all the instructions. Figure 6 in Appendix B illustrates the semantic distribution of COIG-CQIA compared to other datasets. The U-MAP visualization reveals that COIG-CQIA exhibits the most widespread and diverse distribution among all compared datasets.

3.3 Quality

We sample a total of 100 data instances in total from all the sources within the dataset, and then manually evaluate their quality based on four criteria: (1) Is the output correct and an acceptable answer? (2) Does the output meet the instructional requirements and provide a comprehensive and appropriate response to the question? (3) Is the answer complete and sufficiently detailed? (4) Is the answer harmless, avoiding misleading information or the spread of harmful content?

Our human evaluation results in table 10 shows that the data quality has met a very high standard, with human acceptance rates consistently above 95% across four criteria. Regarding the third criteria, we conducted a case study which reveal that human rejections were primarily due to responses not being excessively detailed[3]. Given that our responses are primarily collected from real human interactions on the web, we believe it's acceptable and even natural.

4 EXPERIMENTAL SETUP

In this section, we describe how we use COIG-CQIA to fine-tune models and elaborate our evaluation methods.

Evaluation Benchmarks To assess the model's capabilities across various Chinese tasks, we utilize Belle-Eval (Ji et al., 2023) as our open-ended test set. It encompasses 12 different instruction types spanning various domains, making it ideal for evaluating the impact of different data sources on various tasks. We also employed C-Eval (Huang et al., 2024), CMMLU (Li et al., 2023a), and SafetyBench (Zhang et al., 2023c)[4], which are widely used benchmarks for assessing models' knowledge, reasoning, and safety levels in Chinese contexts. To further explore COIG-CQIA's extensibility in non-Chinese scenarios, we evaluated the model on widely-used datasets including BBH (Suzgun et al., 2022), GSM8K (Cobbe et al., 2021), HumanEval (Chen et al., 2021), and TydiQA (Clark et al., 2020).

Baselines To comprehensively evaluate the instruction-following capacity of the models fine-tuned on COIG-CQIA, we compared it with several well-known Chinese instruction-tuning datasets. These include COIG (Zhang et al., 2023a), Firefly (Yang, 2023), Alpaca-ZH (Cui et al., 2023), COIG-PC (BAAI, 2023), and OL-CC (OL-CC, 2023), which were constructed using various methods. We compared subsets of equal size to COIG-CQIA from these datasets. Additionally, we sampled data from WizardCoder (Luo et al., 2023) and MAmmoTH (Yue et al., 2023), with the total size matching that of COIG-CQIA-Sub.[5]

Implementation Details We fine-tuned various models of different architectures and sizes using COIG-CQIA. This included Chinese-centric multilingual models from the Yi series (6B and 34B) (Young et al., 2024) and the Qwen2 series (7B and 72B) (Yang et al., 2024). We merged the 18 data sources of the COIG-CQIA dataset into 12 sources, and manually selected a more balanced subset from these sources, which we refer to as COIG-CQIA-Sub. The merging rules and criteria used for this consolidation are detailed in Appendix D. For detailed statistics and the curation objective of COIG-CQIA-Sub, see Appendix C. To explore the data's performance on non-Chinese-centric models, we also selected the LLaMA2 series (7B, 13B, and 70B) (Touvron et al., 2023) as base models. We set the learning rate to 2e-5, with a batch size of 128 and a maximum sequence length of 4096. The training was conducted for 5 epochs using a cosine scheduler with 5% warmup. For models under 20B parameters, we employ DeepSpeed (Rasley et al., 2020) ZeRO stage 2 optimization, while for models larger than 20B parameters, we use ZeRO stage 3.

5 EXPERIMENTAL RESULTS

5.1 Ablating Instruction Data Sources and Base Models

We finetune the Qwen2-7B (Yang et al., 2024) and LLaMA2-13B (Touvron et al., 2023) models on different data sources from COIG-CQIA to analyze

[3] Such as how AI systems like GPT4 would include redundant information in their responses

[4] Detailed results are provided in Appendix E

[5] See section 5.3.2 for details.

	Dataset	Open-QA	Brain.	CLS.	Gen.	Sum.	Rewrite	Closed-QA	Extract	Math	Code	Average
	Vanilla Models											
	Vanilla Qwen-2-7B	65.5	60.0	46.0	54.3	40.7	53.5	58.7	44.5	46.2	67.1	53.7
	Vanilla LLaMA-2-13B	1.4	3.8	5.0	1.0	6.7	17.5	12.2	13.6	0.0	17.1	6.9
	Qwen2-7B trained on different COIG-CQIA data source											
Qwen-2-7B	Zhihu	65.2	89.6	42.0	91.9	42.7	56.5	36.1	37.3	77.6	80.0	63.7
	Douban	53.8	67.3	15.0	68.1	13.3	34.0	37.8	27.3	81.0	43.6	47.0
	Xiaohongshu	49.3	60.0	12.5	42.9	13.3	12.0	31.7	16.4	71.4	27.1	36.9
	SegmentFault	53.8	68.5	41.5	69.0	33.3	74.5	48.7	42.7	76.2	65.7	58.6
	Ruozhiba	**77.6**	**95.8**	**64.5**	**96.7**	**76.7**	**91.5**	**82.6**	**72.3**	**90.5**	**87.1**	**83.5**
	Exam	51.4	83.8	54.2	75.2	30.7	73.0	72.2	57.3	49.5	71.4	62.9
	Logi QA	52.1	69.2	50.5	78.6	25.3	70.0	53.7	50.0	75.7	65.7	60.2
	Wiki	53.8	80.8	35.0	79.5	25.3	72.5	42.2	30.0	76.2	75.4	59.1
	WikiHow	48.3	28.5	1.0	41.9	20.7	5.0	20.9	12.7	62.4	47.9	30.2
	COIG PC	53.1	95.4	53.0	85.2	47.3	56.5	50.4	60.0	61.9	42.9	62.1
	Chinese Traditional	41.7	73.1	41.0	79.5	28.7	69.5	55.2	41.8	80.0	58.6	58.2
	Human Value	65.5	90.0	60.5	86.7	58.0	85.0	64.8	50.9	78.6	72.9	72.8
	COIG-CQIA-Full	63.8	88.3	55.0	92.9	51.0	59.0	67.8	64.5	66.7	65.7	68.7
	COIG-CQIA-Sub	59.7	86.2	54.0	91.9	54.3	58.5	68.3	70.9	83.3	71.4	70.3
	LLaMA-2-13B trained on different COIG-CQIA data source											
LLaMA-2-13B	Zhihu	23.1	48.5	17.0	47.1	25.3	24.0	26.1	20.9	0.5	25.0	26.5
	Douban	19.0	27.7	9.0	26.7	13.3	25.0	45.7	20.0	11.9	15.7	22.2
	Xiaohongshu	15.9	28.5	0.0	23.8	6.7	25.0	25.2	20.9	1.0	10.0	16.3
	SegmentFault	23.8	23.1	6.0	31.4	23.3	38.0	30.9	20.0	9.5	38.6	24.3
	Ruozhiba	37.6	55.8	**44.5**	51.0	39.3	38.5	**55.2**	34.1	17.6	**47.9**	42.7
	Exam	30.7	60.0	36.0	56.2	26.7	33.0	40.9	37.3	13.8	39.3	38.0
	Logi QA	20.7	23.1	25.0	36.7	23.3	44.0	50.9	**46.4**	15.7	20.0	29.9
	Wiki	25.5	52.3	15.0	50.0	10.0	17.5	31.7	43.6	4.8	40.7	29.1
	WikiHow	26.6	24.2	5.0	34.3	10.0	15.0	21.7	9.1	2.4	28.6	18.6
	COIG PC	22.8	28.8	22.5	22.4	23.3	32.5	40.4	23.6	7.1	12.1	24.2
	Chinese Traditional	17.2	25.8	16.0	51.4	32.0	45.0	45.7	30.0	14.3	7.9	28.7
	Human Value	33.4	61.9	35.0	64.3	25.3	**49.0**	40.0	46.4	4.3	33.6	39.9
	COIG-CQIA-Full	**46.2**	**68.1**	24.0	65.2	25.3	36.5	43.5	39.1	**18.6**	35.0	41.9
	COIG-CQIA-Sub	39.7	64.2	24.5	**68.1**	**40.3**	45.5	50.0	45.5	8.1	44.3	**43.5**

Table 2: The performance of Qwen-2-7B and LLaMA-2-13B trained on various datasets evaluated on BELLE-EVAL using GPT-4o. Brain. refers to Brainstorm; CLS. refers to Classification; Gen. refers to generation; Sum. refers to summerization.

Dataset	Instruction-Follow							Knowledge & Reasoning		Average
	Gen.& Sum.	Q&A	CLS.	Rewrite	Extract	Math	Code	C-EVAL	CMMLU	
Qwen2-7B(Vanilla Model)										
	-	-	-	-	-	-	-	83.2	83.9	-
Qwen2-7B trained on baseline datasets										
COIG	68.60	48.05	39.5	59.0	46.8	42.9	28.6	69.5	72.0	62.0
Firefly	80.93	63.45	**64.0**	**91.5**	50.9	51.0	57.9	77.1	78.4	73.8
Alpaca-zh	79.77	56.30	57.5	88.0	63.6	15.2	40.7	77.7	77.3	69.9
COIG-PC	78.40	51.10	44.7	62.5	67.3	12.9	54.3	74.9	75.8	66.7
OL-CC	**85.53**	57.30	56.0	71.5	64.5	24.3	51.1	78.7	**81.2**	72.5
Qwen2-7B trained on COIG-CQIA										
COIG-CQIA-Full	77.40	**65.80**	55.0	59.0	64.5	66.7	65.7	76.9	77.5	73.1
COIG-CQIA-Sub	77.47	64.00	54.0	58.5	**70.9**	**83.3**	**71.4**	**78.9**	79.5	**74.8**

Table 3: Performance comparison with grouped Instruction-Follow tasks and detailed Knowledge & Reasoning. We merged tasks with a high correlation and similar task types. Gen. Sum. includes brainstorming, generation, and summarization, while QA encompasses both open-domain and closed-domain question answering.

the impact of data sources on model capabilities across various domains. Then, we evaluate each model's performance on various types of assistant-style tasks using GPT-4o as LLM-as-Judge evaluator on Belle-Eval (Ji et al., 2023). Evaluation details are provided in Appendix D.

Tabel 2 shows the performance of Qwen2-7B and LLaMA2-13B models fine-tuned on different

Source	I.F.	K&R	Average
Single Domain			
NLP	63.0	76.7	69.9
Exam	**70.6**	80.8	**75.7**
Wiki	59.2	83.1	71.2
Social&Forum	67.3	**83.2**	75.3
Mixed Domain			
NLP+Wiki	57.5	68.9	63.2
NLP+Exam	65.2	80.7	73.0
NLP+Social&Forum	65.6	78.9	72.3
Exam+Wiki	69.0	79.6	74.3
Exam+Social&Forum	**70.4**	**81.1**	**75.8**

Table 4: Comparison of Data Mixing Strategy on Instruction-Follow, Knowledge&Reasoning.

subsets. The table indicates that all fine-tuned models achieved significant improvements across various domains. Notably, the Qwen model trained on the Ruozhiba dataset performed remarkably well, even surpassing high-quality data subsets like COIG-PC and Zhihu. Despite the fact that Ruozhiba is not commonly recognized in the Chinese academic community and often contains humorous or absurd content, we believe these characteristics contributed to its effectiveness. The Ruozhiba dataset has inherent logical structures, includes cognitive and linguistic traps, and features jokes and riddles, as well as artistic and abstract rhetorical techniques. These elements, in turn, challenge the model's multi-hop reasoning capabilities, enhancing its understanding of the Chinese language during fine-tuning and improving its capacity for complex logical reasoning. Human Value ranks second on average across all subsets, which aligns with expectations, as this subset contains a substantial amount of high-quality human-annotated data that aligns well with human values. This data not only improved instruction-following capabilities during fine-tuning but also prevented models from biasing towards specific values, enhancing universality. Moreover, WikiHow scores only 30.2 on Qwen and 18.6 on LLaMA-2-13B, likely due to the lack of diversity in its "how-to" instructions.

We also evaluated different base models with varying parameter sizes fine-tuned on the COIG-CQIA-Sub. The table 6 presents the performance differences across models in instruction-following and knowledge & reasoning tasks. As expected, model size correlates with improved performance across all tasks. The Yi and Qwen2 series show strong results, with Qwen2-72B leading overall. LLaMA-2 series lags behind as it wasn't specifically designed for Chinese language understanding. Additionally, we assess the safety performance of various fine-tuned models on Safety-Bench (Zhang et al., 2023c). The detailed experimental results can be found in Appendix E.

5.2 Comparison with Other Chinese Instruct-tuning Datesets

Table 3 illustrates how models trained on different datasets perform various instruction-following tasks. COIG-CQIA stands out in Q&A, extraction, math, and coding tasks, indicating its strength in knowledge-intensive and reasoning areas. However, for classification, summarization, and rewrite tasks (e.g., translation and text editing), COIG-CQIA underperforms. Our case study attributes this to the limited representation of these tasks in the dataset.[6] To improve performance in these areas, we recommend augmenting with datasets such as Firefly and Alpaca-Zh.

5.3 Exploration of Data Mixture Strategy

5.3.1 Mixture of Different Domain

We categorized COIG-CQIA into four main sources as described earlier: NLP datasets, Exams, World Knowledge, and Social Media & Forums. We evaluated different combinations of these sources across tasks, as shown in Table 4. Our results show that data from Social & Forum and Exam sources most significantly enhance the model's instruction-following ability. World Knowledge and Social & Forum data, meanwhile, contribute to improved knowledge performance, aligning with expectations: Social & Forum and Exam data cover broader, more complex tasks, while NLP and World Knowledge data tend to focus on more constrained, traditional tasks. World Knowledge naturally excels in knowledge-intensive tasks such as C-Eval and CMMLU.

When mixing data sources, we observed that combinations of two domains rarely outperform the stronger individual source. Mixing weaker sources, such as NLP and Wiki, can result in further degradation, especially on instruction-following tasks. However, combining two strong sources tends to maintain high performance, such as mixing Exam and Social & Forum data for instruction-following, or Wiki with Social & Forum for knowledge and reasoning tasks.

[6]Tasks such as translation or text editing constitute only a small proportion of our dataset.

Model	BBH (3-Shot, CoT)	GSM8k (4-shot)	HumanEval (P@10)	TydiQA (GP, 1-shot)	Average
			Official Models		
Yi-6B	45.7	34.5	28.3	47.6	39.0
Yi-6B-Chat	43.0	38.0	31.4	24.8	34.3
Qwen2-7B	59.5	71.5	76.0	62.3	67.3
Qwen2-7B-Instruct	63.7	84.5	87.6	34.5	67.5
LLaMA-2-7B	41.6	14.5	25.2	43.5	31.2
LLaMA-2-7B-Instruct	21.7	8.5	25.2	20.9	19.1
			Models trained on Open-Sourced Data Mixture		
Yi-6B	43.6 (-2.1)	33.0 (-1.5)	41.1 (+12.8)	33.6 (-14.0)	37.8 (-1.2/+3.5)
Qwen2-7B	56.9 (-2.6)	73.2 (+1.7)	79.1 (+3.1)	49.2 (-13.1)	64.6 (-2.7/-2.9)
LLaMA-2-7B	40.1 (-1.5)	27.0 (+12.5)	37.0 (+11.8)	23.1 (-23.4)	31.8 (+0.6/+12.7)
			Models trained on COIG-CQIA		
Yi-6B	43.7 (-2.0)	21.4 (-13.1)	25.9 (-2.4)	50.0 (+2.4)	34.9 (-4.1/+0.6)
Qwen2-7B	60.7 (+1.2)	77.1 (+5.6)	80.3 (+4.3)	63.3 (+1.0)	70.4 (+3.1/+2.9)
LLaMA-2-7B	39.7 (-1.9)	13.2 (-1.3)	24.6 (-0.6)	49.1 (+5.6)	31.7 (+0.5/+12.6)
			Models trained on COIG-CQIA + Open-Sourced Data Mixture		
Yi-6B	44.9 (-0.8)	29.5 (-5.0)	40.8 (+12.5)	37.6 (-10.0)	38.2 (-0.8/+3.9)
Qwen2-7B	60.9 (+1.4)	75.4 (+3.9)	80.9 (+4.9)	64.2 (+1.9)	70.4 (+3.1/+2.9)
LLaMA-2-7B	41.7 (+0.1)	26.5 (+12.0)	38.6 (+13.4)	45.3 (+1.8)	38.0 (+6.8/+18.9)

Table 5: Performance comparison of different models and versions on various non-Chinese tasks. Numbers in parentheses represent differences from the base model. For the averages, the left value represents the difference from the base model, and the right value represents the difference from the chat model.

Model Series	Size	I.F.	K&R	Average
Yi	6B	55.5	74.1	64.8
	34B	62.3	77.6	70.0
LLaMA-2	7B	35.6	33.2	34.4
	13B	43.5	38.0	40.8
	70B	47.7	50.2	48.9
Qwen2	7B	70.3	79.2	74.8
	72B	73.3	89.8	81.6

Table 6: Model performance comparison across different model series and sizes on Instruction-Following and Knowledge & Reasoning tasks.

5.3.2 Mixed with Open-Sourced data

To explore COIG-CQIA's potential on non-Chinese tasks, we extended our evaluation using the open-instruct suite(Wang et al., 2023a) to include four additional non-Chinese tasks: BBH (reasoning), GSM8K (math), HumanEval (code), and TydiQA (multilingualism). The results are shown in the table 5.

Models trained on COIG-CQIA performs on par with official base and chat models on BBH and significantly outperform baselines on TydiQA, highlighting its strength in activating multilingual capabilities. Given that COIG-CQIA is a Chinese-focused dataset with less than 5% of math and code data, it understandably underperforms on English-heavy tasks like GSM8K and HumanEval, particularly when using Yi-6B as the base model. The use of purely Chinese data impacts performance in these tasks. In contrast, experiments with Qwen2-7B and Llama2-7B show COIG-CQIA performing at or above the level of base models, although it still lags behind the more data-engineered chat models. This gap is expected, as these official chat models undergo extensive, costly data engineering.

To address the language mismatch, we experimented with mixing COIG-CQIA with open-source English datasets. We sampled equivalent data from Magicoder (code) and Mammoth (math), labeled as OS Mix, and combined it with CQIA. Our findings show: (1) COIG-CQIA matches or exceeds OS Mix on reasoning tasks and performs significantly better on multilingual tasks. (2) On code and math tasks, COIG-CQIA's performance varies with the base model. With Yi-6B and Llama2-7B, COIG-CQIA lags behind OS Mix, which is specialized for these tasks. Surprisingly, using Qwen2-7B as the base, COIG-CQIA outperforms OS Mix on the same tasks. (3) Combining COIG-CQIA with OS Mix strengthens each dataset's weaknesses, leading to overall performance gains.

5.4 Human Evaluation

We compared Yi-6B(Young et al., 2024) fine-tuned on the COIG-CQIA-Sub with several Chinese open-source chat models. Focusing on real-world

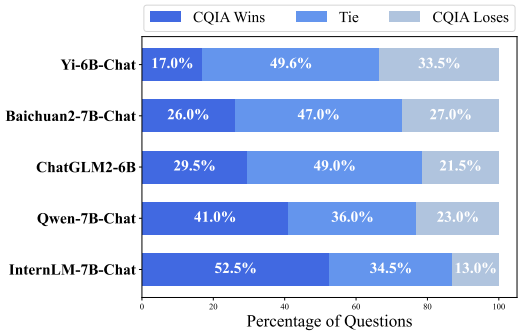

Figure 2: Human evaluation of pair-wise comparison between Yi-6B fine-tuned on COIG-CQIA-Sub and 5 strong baselines.

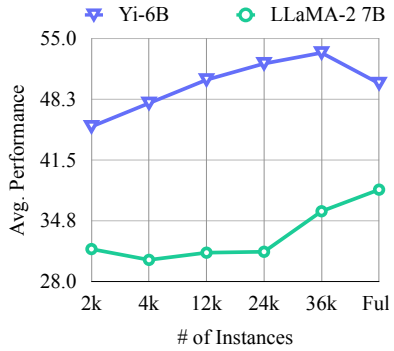

Figure 3: Performance of the Model Trained on Different Data Scales.

questions, we sampled 200 prompts from OL-CC[7] and Zhihu, none of which were part of the training set. We conducted a pairwise comparison to assess how our model performs in real-world scenarios. Figure 2 presents the human evaluation results comparing COIG-CQIA against five baselines: Yi-6B-Chat (Young et al., 2024), Baichuan2-7B-Chat (Baichuan, 2023), ChatGLM2-6B (GLM et al., 2024), Qwen-7B-Chat (Bai et al., 2023), and InternLM-7B-Chat (Cai et al., 2024). The results show that the model trained on COIG-CQIA achieved higher human preference, with over 60% responses being rated as better or tied with the baselines. This demonstrates COIG-CQIA's ability to align more closely with real-world human communication patterns, resulting in higher user preference. See details in Appendix D.

5.5 Data Scaling

The data scaling results in Figure 3 demonstrate the impact of training set size on model performance for Yi-6B and LLaMA-2-7B. Both models exhibit performance improvements as the number of instances increases, underscoring the significance of data quantity in enhancing language model capabilities. Yi-6B shows rapid gains up to 24k instances, after which performance stabilizes with minor fluctuations. This plateau effect may be attributed to the limitations of our dataset scale, where model behavior becomes less predictable at this data size scale. In contrast, LLaMA-2 7B displays a consistent upward trend across the entire range.

[7]https://data.baai.ac.cn/details/OL-CC

6 Related Work

6.1 Instruction-Tuning Dataset

Instruction tuning enhances the conversational and task execution capabilities of large language models (LLMs) by training them to generate responses aligned with input instructions. This approach yields more controllable and predictable models that better align with human intent. Several strategies have been employed to construct instruction-tuning datasets: (1) Manual annotation by human experts (Conover et al., 2023). (2) Repurposing existing NLP datasets (Mishra et al., 2022; Sanh et al., 2022; Chung et al., 2022). (3) Synthesising using LLMs (Honovich et al., 2022; Wang et al., 2023b; Xu et al., 2023a; Ji et al., 2023; Xu et al., 2023b). While efficient, this method may introduce inconsistencies and noise. While numerous English instruction tuning datasets exist, their Chinese counterparts are limited. Some efforts focus on translating English datasets (Peng et al., 2023), while others repurposing existing NLP tasks into instruction formats (BAAI, 2023; Yang, 2023). Notable Chinese datasets include HC3 (Guo et al., 2023), COIG (Zhang et al., 2023a), BELLE (Ji et al., 2023), and MOSS (Sun et al., 2023).

6.2 Data Mixture Strategies for SFT

Recent research emphasizes the importance of data quality in instruction tuning. LIMA (Zhou et al., 2023) demonstrates strong performance using only 1,000 high-quality instruction-output pairs. AlpaGasus (Chen et al., 2023) and Humpback (Li et al., 2023b) employ advanced filtering techniques to enhance dataset quality and training efficiency. Studies also explore the impact of mixing different instruction-tuning datasets. Song et al. (2023) investigate various combination approaches, while the Tulu series (Konchakov et al., 2023; Ivison et al., 2023) demonstrates that increasing instruction diversity can improve overall performance. Notably, no single dataset or combination consistently

outperforms others across all metrics, highlighting the complexity of optimizing instruction-tuning data mixtures.

7 Conclusion

This paper presents COIG-CQIA, a high-quality Chinese instruction fine-tuning dataset designed to enhance the performance of large language models in various real-world applications. The dataset is carefully compiled from diverse online sources within the Chinese internet and undergoes a rigorous curation process, including meticulous cleaning, restructuring, and manual review, to ensure its quality, diversity, and relevance. Through extensive experiments, we demonstrate that COIG-CQIA serves as a strong and competitive resource for Chinese instruction tuning, achieving robust performance across multiple evaluation benchmarks. Compared to existing datasets, our dataset exhibits greater linguistic diversity, improved instruction-following capabilities, and stronger generalization across tasks. Additionally, we conduct an in-depth analysis of the impact of data sources and mixing strategies, offering valuable insights into optimizing training data for Chinese NLP applications.

8 Limitation

We acknowledge several limitations in our study. While COIG-CQIA is comprehensive, the inclusion of subjective elements may lead to varying interpretations, potentially impacting data construction. Additionally, our focus on Chinese language data covers only a fraction of human knowledge. The evaluation metrics may not fully capture the models' sophisticated understanding and reasoning abilities. These limitations underscore the need for ongoing refinement and expansion of our dataset. In future work, we aim to collect and aggregate more diverse Chinese instruction-tuning data to improve the models' capability and reliability.

9 Ethics Statement

In developing COIG-CQIA, we strictly adhere to ethical guidelines and legal regulations, ensuring fairness, transparency, inclusivity and respect for all stakeholders. We stress the importance of safeguarding privacy and intellectual property rights, underscoring our commitment to responsible and lawful data management. We have taken steps to anonymize any personal data to protect privacy and have made every effort to minimize harmful or biased content. However, we recognize that biases can inadvertently arise and some information may be potentially offensive. We are committed to continuous monitoring and improvement to mitigate such biases. Furthermore, we encourage users of our dataset to employ it responsibly and to consider the ethical implications of their work, particularly in applications that may impact individuals or communities.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62376262), GuangDong Basic and Applied Basic Research Foundation (2023A1515110718 and 2024A1515012003), China Postdoctoral Science Foundation (2024M753398), Postdoctoral Fellowship Program of CPSF (GZC20232873).

References

BAAI. 2023. Coig-pc.

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

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2023. Alpagasus: Training a better alpaca with fewer data. *Preprint*, arXiv:2307.08701.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world' s first truly open instruction-tuned llm. *Company Blog of Databricks*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *Preprint*, arXiv:2212.09689.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *Preprint*, arXiv:2311.10702.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. 2023. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742*.

R. A. Konchakov, A. S. Makarov, G. V. Afonin, J. C. Qiao, M. G. Vasin, N. P. Kobelev, and V. A. Khonik. 2023. Critical behavior of the fluctuation heat capacity near the glass transition of metallic glasses. *Preprint*, arXiv:2306.00475.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

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

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

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. *Preprint*, arXiv:2104.08773.

OL-CC. 2023. Openlabel-chinese conversations dataset (ol-cc).

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

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

Chiyu Song, Zhanchao Zhou, Jianhao Yan, Yuejiao Fei, Zhenzhong Lan, and Yue Zhang. 2023. Dynamics of instruction tuning: Each ability of large language models has its own growth pace. *Preprint*, arXiv:2310.19651.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, et al. 2023. Moss: Training conversational language models from synthetic data. *arXiv preprint arXiv:2307.15020*, 7:3.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

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

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023a. How far can camels go? exploring the state of instruction tuning on open resources. *Preprint*, arXiv:2306.04751.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

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

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023b. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *Preprint*, arXiv:2304.01196.

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. 2023c. 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, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

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

Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, Wenhao Huang, and Jie Fu. 2023a. Chinese open instruction generalist: A preliminary release. *Preprint*, arXiv:2304.07987.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023c. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

A Details of COIG-CQIA Curation

We provide the data processing code for data construction at https://github.com/paralym/COIG-CQIA, including templates, filtering rules, prompts, and sampling code.

A.1 Social Media & Forums

We curated data from five prominent Chinese social media platforms and forums, each offering unique content characteristics.

Zhihu is a comprehensive Q&A platform where users can ask and answer questions on various topics, making it an extensive repository of knowledge and insights. However, the absence of a review mechanism for answers on Zhihu leads to a large volume of content that falls short of our quality standards. To address this issue, we implemented a multi-step filtering process. Initially, we selected answers that had garnered more than 50 upvotes, which reduced our original dataset from 10 million entries to 2 million. We then applied a rule-based method to filter out content containing sensitive or potentially harmful keywords, further narrowing the dataset to 100K. Subsequently, we leveraged GPT-4 to evaluate the remaining responses on a scale of 1-10, retaining only those that scored above 8, which yield approximately 8K high-quality answers. In the final step, human annotators carefully reviewed and selected the top 5.6K entries, ensuring that only the highest quality, most informative content was included in our final dataset.

SegmentFault is IT-focused Q&A community which is similar to Stack Overflow in its scope and purpose. To ensure the relevance and currency of the dataset, we concentrated on content posted after 2018, acknowledging that earlier posts might be outdated due to evolving programming languages and software versions. Our selection process prioritized "accepted" answers that had received a minimum of 5 upvotes, indicating community validation of their quality and usefulness. To further refine our dataset, we conducted a comprehensive manual review of all instruction-response pairs.

Douban is a social platform focused on literature and arts, where users share content related to books, movies, TV series, music, and more. We sampled data from books, movies, and TV series, collecting metadata such as ratings, actor/crew details, and long reviews. Based on this rich dataset, we created three main tasks: synopsis generation, review generation, and recommendations. For each task, we designed a variety of prompt templates, combining them with metadata to construct comprehensive instructions. In the case of synopsis and review generation, we utilized movie or TV series names in conjunction with these templates, using Douban user-generated content as responses. We then applied quality filtering to elimminate short or irrelevant answers and remove personal information. To improve real-world applicability, we refined some instructions to include more implicit intents, aligning responses more closely with the content.

Xiaohongshu is a popular social media platform in China that serves as a hub for users to share their daily lives, travel experiences, food, and product recommendations. Contents in this platform are renowned on the Chinese internet for their unique expressive style. For our study, we curated a sample of posts ranging from 500 to 2000 characters in length. To maintain focus on the core content, we excluded posts that contained user interactions (such as "@User_Name" mentions) or references to visual media (e.g., "as shown in the picture/video").

Ruozhiba is a sub-forum within Baidu Tieba, China's largest interest-based online community platform. This particular forum is renowned for its linguistic complexity, featuring posts rich in wordplay, including puns, polysemous terms, causal reversals, and homophones. Many of these posts are ingeniously crafted with logical traps that present cognitive challenges even for native speakers. In our study, we focus on the 500 most upvoted threads in this forum. We used the thread titles as potential instructions, carefully filtering out those that were non-instructive (such as mere declarative statements or unanswerable queries) or contained toxic content. For answer curation, human evaluators first identified the traps within the instructions and then prompted GPT-4 to generate responses. This process was repeated until GPT-4 produced correct answers.

A.2 World Knowledge

Introducing world knowledge to LLMs is crucial for enhancing their ability to engage in knowledge-driven interactions. To collect comprehensive data in this broad field of information, we focused on two key areas.

A.2.1 General Encyclopedia

General encyclopedias provide comprehensive coverage of a wide range of topics across various fields. We collected data from three prominent Chinese encyclopedic websites: One Hundred Thousand Whys, wikiHow-zh, and Encyclopedia of China. **One Hundred Thousand Whys** focuses on popular science, featuring articles that ask "why" across diverse topics. We collected data from all 15 categories, using article titles as instructions and content as responses, filtering out responses under 300 characters. **WikiHow-zh**, the Chinese version of WikiHow, covers a wide range of "how-to" articles. We sampled 1.5K entries from all 19 categories, filtered for quality and length, and used titles as instructions and article contents as responses. **Encyclopedia of China** is a comprehensive resource with 500K expert-authored entries. We designed various prompt templates for concept explanation tasks, sampling entries from all 74 categories. Instructions were constructed by combining entry names or subtitles with prompt templates, with corresponding content used as responses.

A.2.2 Domain Specific Knowledge

We collected data from four specific domains: medicine, economic management, electronics, and agriculture. **Medical Domain** data was sourced from three websites: Baobaozhidao, Qianwen Health, and Baikemingyi. The first two feature expert-written Q&A articles, while Baikemingyi offers structured data on diseases and medications. We used article titles as instructions and content as responses, designing various prompt templates for structured data. **Economic Management Domain** data came from MBA Wiki Encyclopedia, a collaborative knowledge platform. We created instructions by combining entry names with designed prompt templates, using entry content as responses. **Electronics Domain** data was collected from the EETrees electronic encyclopedia, following a similar method of combining entry names with prompt templates to create instruction-response pairs. **Agriculture Domain** data was sourced from an agricultural encyclopedia website covering various topics. We constructed instruction-response pairs from article titles and content, applying specific filtering criteria.

A.3 Examinations

To equip the model with robust problem-solving skills and a comprehensive knowledge foundation, we leveraged a diverse range of examination resources in the training process.

The Middle School and College Entrance Examinations data is primarily sourced from the COIG dataset(Zhang et al., 2023a), focusing on China's principal general competency tests. These data cover various humanities subjects and include detailed answer explanations. After filtering and processing, we obtained 1964 (instruction, response) pairs.

Graduate Entrance Examination is one of the most challenging examinations in China, exceeding college entrance exams in difficulty and requiring advanced knowledge application and depth. We have collected a variety of exam papers from recent years across disciplines including mathematics, computer science, chemistry, law, psychology, medicine, etc. Using Mathpix[8] for image-to-text conversion, we extracted questions and answers and converted them into LaTeX format. We eliminate data without analysis and manually verified the accuracy of the questions and answers. We eliminate data without analysis and manually verified the accuracy of the questions and answers. To enhance domain-specific capabilities, we separately curated **Law Exam** questions as an independent data source, ensuring a more focused dataset for legal reasoning tasks.

Logical Reasoning Test data aims to assess critical thinking and problem-solving skills. We collected logic reasoning questions with detailed answer analyses from various online sources.

Chinese Culture Test data investigates the mastery of traditional Chinese culture and history. We compiled multiple-choice questions with answer analyses from online resources.

A.4 NLP Datasets

To further enhance the model's language understanding and generation capabilities, we incorporated several specialized NLP datasets into `COIG-CQIA`. These datasets were carefully selected to cover a wide range of linguistic tasks and cultural contexts.

COIG-PC is a comprehensive collection of Chinese NLP tasks (BAAI, 2023). We initially selected 1,413 Chinese-English tasks from COIG-PC

[8] https://mathpix.com/

and manually refined 250 high-quality tasks covering information extraction, classification, summarization, and more, primarily from traditional NLP datasets. Through temperature sampling, we obtained 3,000 instruction-response pairs, which were further verified by humans to ensure quality. Notably, while this dataset provides valuable task-specific training, its characteristically short outputs required careful integration to avoid compromising the model's overall chat performance across tasks.

COIG Human Value , a subset of the COIG dataset(Zhang et al., 2023a), focuses on instruction fine-tuning data aligned with Chinese cultural values. We manually filtered out data with formatting errors and incorrect answers, retaining only those that provide answer explanations to form instruction-response pairs.

Firefly Chinese Traditional comprises three tasks related to traditional Chinese culture: Classical Chinese Translation, Ancient Poetry Writing, and Idiom Interpretation (Yang, 2023). We filter the responses shorter than 300 characters, and sample 300 instances from each task. Then, we manually filtered out low-quality data such as instruction-response mismatch, response error, and unanswerable instructions.

CValues addresses anti-discrimination and empathy across various dimensions. It includes human-generated prompts and expert-crafted responses aligned with human values. We incorporated all data from CValues(Xu et al., 2023c) to enhance the model's alignment with ethical considerations.

Finance The FinanceNLP task is constructed by selecting additional finance-related tasks from COIG-PC, using the same filtering strategy as COIG-PC.

B Comparison between COIG-CQIA and other Chinese Datasets.

Table 7 and Figure 6 show the comparison of COIG-CQIA and baseline Chinese Datasets.

C COIG-CQIA-Sub

To enrich the data diversity for CQIA, we aimed to expand the coverage of data sources as much as possible. However, this expansion led to an imbalanced distribution of data sources, where certain categories, such as encyclopedia data, constituted a disproportionately large portion of the dataset. To address this, we carefully curated a high-quality subset, COIG-CQIA-Sub, which maintains a more balanced data composition while preserving linguistic diversity, domain coverage, and task relevance.

The selection of COIG-CQIA-Sub from COIG-CQIA-Full was guided by both empirical evaluation and manual curation across multiple dimensions:

Data Diversity: We evaluated linguistic variations, domain-specific expressions, and stylistic differences across sources. For instance, Ruozhiba contributes complex linguistic structures, while Xiaohongshu primarily focuses on copywriting tasks, offering limited diversity in task types.

Difficulty Level: We considered linguistic and reasoning complexity when curating the subset. For example, LogiQA contains highly structured logical reasoning tasks that are valuable for evaluating advanced comprehension capabilities, making its inclusion crucial for certain benchmarks.

Task Relevance: The alignment of data sources with key NLP applications was a major criterion. Data sources such as COIG PC, which exhibit strong relevance for tasks like question answering, classification, and translation, were prioritized due to their impact on downstream applications.

Rarity: We emphasized retaining rare or long-tail expressions that contribute to the richness of the dataset. For example, Ruozhiba includes unique linguistic constructs and challenging text samples, which add value to the dataset's diversity.

To ensure a well-balanced selection, each source was assigned an importance score based on its impact on downstream performance. These scores were complemented by manual assessment following the criteria outlined above. The subset was designed to maintain a representative but diverse sampling of the full dataset while reducing the over-representation of any single data source.

Our goal with COIG-CQIA-Sub is to provide a high-quality subset that explores the dataset's performance ceiling on key benchmarks. It does not replace COIG-CQIA-Full but complements it, allowing focused experimentation while retaining the full dataset's broad applicability. While COIG-CQIA-Full may not always outperform COIG-CQIA-Sub in specific tasks, it remains a valuable resource for fine-tuning models across diverse domains.

Dataset	# Instances	Source	Human Generated?	Human Verified?
COIG	178k	Existing Dataset&Synthesis	×	×
Firefly	1.1M	Existing Dataset	×	×
Alpaca-zh	51k	Synthesis	×	×
COIG-PC	321M	Existing Dataset	×	×
OL-CC	10k	Human&Synthesis	✓	×
COIG-CQIA	44k	Human	✓	✓

Table 7: Comparison of Different Datasets.

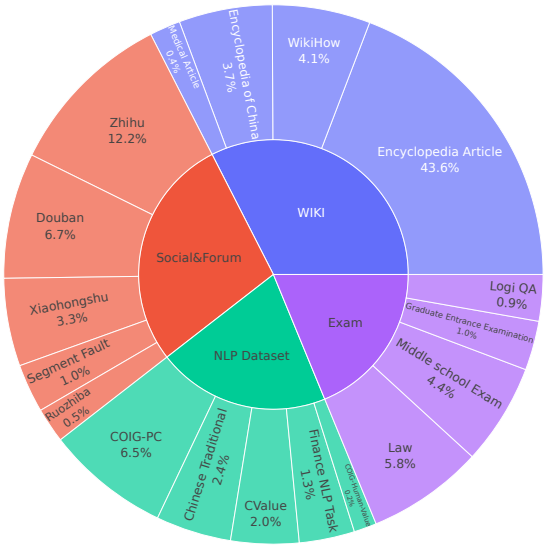

Figure 4: Data distribution of COIG-CQIA.

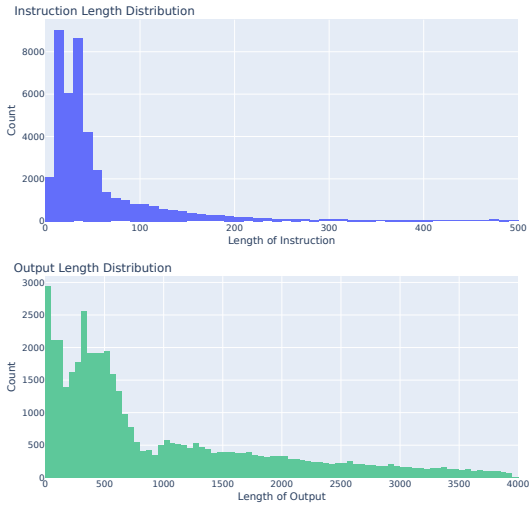

Figure 5: Length distribution of instruction and responses. Note that the instruction is the concatenation of instructions and inputs in COIG-CQIA.

Source	Quantity	Source	Quantity
Zhihu	2733	Douban	300
Xiaohongshu	50	Segment Fault	454
Encyclopedia Article	1350	Encyclopedia of China	200
WikiHow	300	COIG PC	3000
Middle school Exam	200	Graduate Entrance Examination	475
Logi QA	422	CValues	906
COIG-Human-Value	101	Chinese Traditional	1110
Finance NLP Task	500	Ruozhiba	240
Medical Article	186	Law	400
Total		12687	

Table 8: The data composition of CQIA-Sub.

D Experimental Settings of Evaluation

D.1 COIG-CQIA Data Source Merging Strategy

To reduce redundancy in experiments and enhance data organization, we consolidated the 18 original data sources in COIG-CQIA into 12 broader categories. The merging was performed based on task similarity and content overlap as follows:

- **Human Value**: *C-Values* and *COIG-Human-Value* were merged since both pertain to human values and ethical reasoning.
- **Wiki**: *Encyclopedia Article*, *Encyclopedia of China*, and *Medical Article* were grouped into a single "Wiki" category, as all three provide structured, encyclopedia-style knowledge.
- **Exam**: *Middle & High School Exam*, *Graduate Entrance Exam*, and *Law Exam* were merged into a unified "Exam" category, as they all consist of academic and professional examination questions.
- **COIG PC**: Since *COIG PC* comprehensively covers a wide range of existing Chinese NLP datasets, including the *Finance NLP dataset* we collected, we integrated the Finance dataset into COIG PC to avoid redundancy.

D.2 BELLE-Eval

For rapid evaluation, we selected an average of 200 samples from the BELLE-Eval dataset based on task type to serve as our test set. Through our

Figure 6: U-Map Visualization of COIG-CQIA and other Chinese Datasets. COIG-CQIA exhibits the broadest distribution in the semantic space, encompassing the combined semantic distributions of all other datasets.

validation, we found that these 200 samples have a strong correlation with the full BELLE-eval dataset when used for model evaluation.

D.3 Human assessment

In human assessment, for each prompt, we generated one response per model[9], then asked annotators to compare the responses from our model and a baseline, allowing for a "tie" when neither response was better.

Model	SafetyBench
GPT-4-0613	89.2
GPT-3.5-turbo-0613	80.4
Yi-6B	
+Zhihu	75.8
+Douban	76.2
+Xiaohongshu	76.0
+Segmentfault	78.0
+Ruozhiba	81.3
+Exam	77.6
+Logi QA	79.1
+Wiki	75.8
+Wikihow	76.4
+COIG PC	81.2
+Chinese Traditional	76.6
+Human Value	79.1
+COIG-CQIA	**81.7**

Table 9: SafetyBench scores of Yi-6B trained on various data sources.

Criteria	Accept Rate
Is the output correct and an acceptable answer?	98%
Does the output meet the instructional requirements and provide a comprehensive and appropriate response to the question?	96%
Is the answer complete and sufficiently detailed?	95%
Is the answer harmless, avoiding misleading information or the spread of harmful content?	99%

Table 10: Accept Rate by Criteria.

E Safety Evaluation

Table 9 shows that model trained on COIG-CQIA outperforms GPT-3.5-turbo-0613. Models trained on social media and forum data (e.g., Douban, Zhihu, and Xiaohongshu) achieved moderate safety scores, likely due to the diverse and open nature of social media content, which may cause potential harmfulness. Interestingly, models trained on Wiki-style data tended to score lower. We hypothesize that this may be due to the limited diversity of instruction within professional data sources, leading to poor performance on safety which is outside of specialized domains.

[9]Responses were generated using nucleus sampling with p=0.85, k=50, and temperature=0.9.