CIF-Bench: A Chinese Instruction-Following Benchmark for Evaluating the Generalizability of Large Language Models

Yizhi Li[2*] Ge Zhang[1,3*] Xingwei Qu[2*] Jiali Li[4] Zhaoqun Li[5] Zekun Wang[6]
Hao Li[2] Ruibin Yuan[7] Yinghao Ma[8] Kai Zhang[9] Wangchunshu Zhou[10] Yiming Liang[11,12]
Lei Zhang[1] Lei Ma[13] Jiajun Zhang[11,12] Zuowen Li[14] Stephen W. Huang[15] Chenghua Lin[2†] Jie Fu[7†]

[1]Stardust.AI m-a-p.ai [2]University of Manchester [3]University of Waterloo [4]National University of Singapore [5]Zhejiang University [6]Beihang University
[7]HKUST [8]Queen Mary University of London [9]Ohio State University [10]AIWaves Inc. [11]Institute of Automation, Chinese Academy of Sciences
[12]School of Artificial Intelligence, Chinese Academy of Sciences [13]Peking University [14]Beijing Foreign Studies University [15]harmony.ai

Abstract

The advancement of large language models (LLMs) has enhanced the ability to generalize across a wide range of unseen natural language processing (NLP) tasks through instruction-following. Yet, their effectiveness often diminishes in less-trained languages like Chinese, exacerbated by biased evaluations from data leakage, casting doubt on their true generalizability to new linguistic territories. In response, we introduce the Chinese Instruction-Following Benchmark (**CIF-Bench**), designed to evaluate the zero-shot generalizability of LLMs to the Chinese language. CIF-Bench comprises 150 tasks and 15,000 input-output pairs, developed by native speakers to test complex reasoning and Chinese cultural nuances across 20 categories. To mitigate data contamination, we release only half of the dataset publicly, with the remainder kept private, and introduce diversified instructions to minimize score variance, totaling 45,000 data instances. Our evaluation of 28 selected LLMs reveals a noticeable performance gap, with the best model scoring only 52.9%, highlighting the limitations of LLMs in less familiar language and task contexts. This work aims to uncover the current limitations of LLMs in handling Chinese tasks, pushing towards the development of more culturally informed and linguistically diverse models with the released data and benchmark[1].

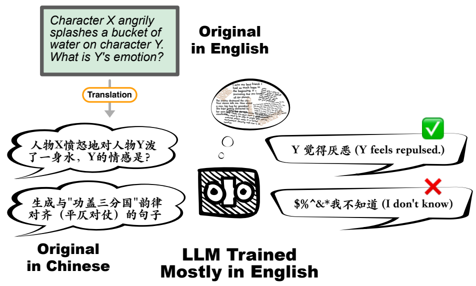

Figure 1: A large language model can tackle English task translated to Chinese, but fail to respond to instruction originally in Chinese.

1 Introduction

The landscape of natural language processing (NLP) has been dramatically reshaped by the emergence of large language models (LLMs), which have demonstrated an ability to generalize across unseen NLP tasks (Lin and He, 2009; Fabbri et al., 2021; Alkaissi and McFarlane, 2023; Wu et al., 2023, 2024), often showcased through the framework of instruction-following (Mishra et al., 2021; Sanh et al., 2021; Wei et al., 2021). Despite these advances, skepticism remains regarding the transferability of this instruction-following capability, particularly in multilingual contexts. The models perform worse when switching to Chinese due to the prevalence of English training data (Huang et al., 2023b; Zhang et al., 2023b), as figured in Fig. 1. This concern is exacerbated by observations that benchmarks designed to assess the capabilities of LLMs may inadvertently suffer from biased evaluations due to data leakage (Sainz et al., 2023), particularly when web-scale datasets are employed to enhance model generalizability (Raffel et al., 2023). Such observations raise a critical question: While the generalizability of LLMs appears intriguing, do these models face significant challenges when evaluated on private and diversified instruction-formatted tasks in less common language contexts?

To answer this question, we introduce the **C**hinese **I**nstruction-**F**ollowing **Bench**mark (**CIF-Bench**), a novel benchmark designed for the zero-shot generalizability evaluation of LLMs, with Chinese serving as an insightful example for multilingual transferred instruction-following tasks. Our benchmark comprises 150 tasks and

*The authors contributed equally to this work.
†Corresponding authors.
[1] https://yizhilll.github.io/CIF-Bench/

Findings of the Association for Computational Linguistics: ACL 2024, pages 12431–12446
August 11-16, 2024

$15,000$ input-output pairs, with the assistance of native speaker annotators, ensuring the inclusion of human-authored tasks that are not only challenging but also naturally expressed. A significant portion (38.7%) of these tasks are designed to test a model's complex natural language inference (NLI) and reasoning capabilities, as well as drawing upon Chinese culture spread across 20 distinct categories. In an effort to mitigate future evaluation biases from data leakage, we decide to publicly release only half of the data instances, reserving the rest as a private dataset to maintain an impartial benchmark. Furthermore, CIF-Bench enhances its robustness by introducing 5 variations of instructions per task, using these to diminish score variance in private split evaluations as discussed in §5. CIF-Bench also pioneers a model-based automatic pipeline designed to tackle the inherent challenges of evaluating open-ended natural language generation outputs (Gehrmann et al., 2021).

By selecting a range of popular LLMs that support Chinese for evaluation, we aim to depict the limits of current instruction-following capabilities in language transfer contexts as the many models follow an English-oriented pre-training paradigm (Huang et al., 2023b). Our findings reveal that even the best-performing model achieves a score of only 52.9% on CIF-Bench, underscoring the gap that exists when LLMs are confronted with tasks in a less-familiar language and unseen data instances. We find that this performance decrement is particularly noticeable in scenarios involving unseen tasks and unseen input-output pairs, contrasting with the models' performance on existing Chinese datasets and translated English-language tasks. Such results suggest that while LLMs exhibit impressive generalizability in a context more aligned with observed data, their effectiveness diminishes when faced with the dual challenges of unacquainted languages and novel tasks.

To summarize our contributions, we:

- Present a new benchmark that addresses a critical gap in existing NLP research by focusing on the generalizability of LLMs to an underrepresented language in terms of training and evaluation resources;
- Construct an instruction-following evaluation dataset with 150 tasks and $45,000$ data samples, and release half of the input-output pairs for future LLM evaluation research;
- Provide an in-depth analysis of 28 LLMs, revealing their limitations in adapting to less familiar languages and task contexts, offering insights into where improvements are needed for instruction-following generalizability.

2 Related Work

Instruction-Following Evaluation. Large-scale pre-trained language models have been found that they can generalize across unseen tasks by fine-tuned on formatted task instructions (Khashabi et al., 2020; Mishra et al., 2021; Wei et al., 2021; Sanh et al., 2021). Early studies attempt to fine-tune and evaluate such a capability in a few-shot manner by providing input-output examples (Ye et al., 2021; Mishra et al., 2021). Following that, another line of research Bach et al. (2022); Wang et al. (2022b) Bai et al. (2024) improves the evaluation reliability from the perspective of scaling the task quantity and providing well-defined corresponding instructions. A more recent concurrent work FollowBench proposes to craft multiple instructions for a single task to evaluate the LLMs, similar to CIF-Bench. A core distinction between CIF-Bench and the FollowBench is that we focus on assessing whether models can stably perform given diversely expressed, but semantically identical instructions, while FollowBench aims to extend the basic instruction with different additional requirements.

Chinese LLM Benchmarks. There have been important efforts, such as CLUE (Xu et al., 2020) and CUGE (Yao et al., 2021), made to evaluate the pre-trained language on extensive tasks in the Chinese context, which consider the traditional taxonomy of natural language understanding and generation. As these benchmarks are restricted in the prediction formats and could not fully measure the cross-task generalization of LLMs in the free-form outputs, more recent studies (Huang et al., 2023b; Li et al., 2023) propose to reformat the tasks into multi-choice question answering, mostly examining the knowledge-base abilities in Chinese. However, such a strict format could impede the models from fully generalizing to more complex reasoning and creative tasks. Thereby, we argue that there is a lag in evaluating LLMs instruction-following capacity in the Chinese language.

3 The Challenging Chinese Instruction-Following Benchmark

The Challenging Chinese Instruction-Following Benchmark unifies the NLP tasks in the prompt-based instruction-following schema (Mishra et al., 2021) and evaluates the LLMs in a zero-shot manner, which is to say that the models are expected to directly provide the correct output given the concatenation of the task instruction and data input texts. Formally, for each data sample in CIF-Bench, the three components we refer to are:

- An instruction that is provided as the introductory information for a specific NLP task, which is an implicit definition of a “mapping function" (i.e., task background context) that must be interpreted by the models before proceeding.
- An span of input text that encompasses the context to define the specific task scenario.
- A reference as the (potentially) standard output in the data instance.

Table 1: The Statistics of CIF-Bench instruction data. #Instruction and #Input-Output refer to the quantity of examples contained in each task.

Split→	Private	Public
#Task	150	
#Instruction	5	1
#Input-Output	50	50
Total Instances	37,500	7,500

We define a total of 150 curated tasks, constructed according to Chinese linguistic and societal backgrounds, as well as from existing NLP tasks in Chinese and English. To improve the evaluation robustness, we provide a diversified set of 5 instructions with the same semantics for each task. Considering the potential data leakage issue of LLM benchmarks, we split two halves of 100 input-output pairs in each task into *private* and *public* partitions, and only test and release the *public* split which contains one instruction variant. In sum, there are $45,000$ human-annotated [instruction, input, output] instances produced in CIF-Bench, as suggested in Table 1. In addition, we provide detailed instructions for all the tasks in Appendix A.1.

3.1 Data Collection.

Collecting Sources. CIF-Bench is designed for the extensive evaluation of Chinese comprehension and generation capabilities in LLMs, particularly focusing on aspects such as creative generation and linguistic abilities that existing benchmarks, such as C-Eval (Huang et al., 2023b) and C-MMLU (Li et al., 2023), struggle to assess. First, we select 113 diverse existing English NLP tasks, as shown in Table 1 from Super Natural Instructions (**SNI**) (Wang et al., 2022b) and other research work (full list in Appendix A.1). We then describe these task instructions in Chinese and a semantically balanced distributed subset from each original English NLP task as the ***Public*** split of CIF-Bench. We further ask expert native Chinese speakers, who minimally have undergraduate degrees, to annotate 100 samples per task based on the translated task instructions. These samples are further deduplicated according to their semantic embeddings. We finally select 50 samples per task as the ***Private*** split of CIF-Bench, to guarantee each sample's validity and the balanced label distribution of each task.

Table 2: The statistics of existing and newly designed Tasks. The existing tasks and instances include those translated from English as well as original Chinese data.

Task	Instance: Existing	Instance: Newly Annotated
Existing (113)	5,650	5,650
Newly Designed (37)	N/A	3,700
Total	5,650	9,350

Annotation Protocol. To be specific, we set up a robust three-stage pipeline in our annotation process. In ***stage 1***, to ensure high annotation quality, we hire native speakers with college backgrounds to annotate the data samples in the form of triplet <instruction, input, output> in cooperated with the annotation platform Stardust[2]. In ***stage 2***, the data annotation specialists from the platform conduct a second round of checking on the quality of the samples. The specialists first use the GPT-4 as an auxiliary verification, and the samples scored lower than 6 out of 10 would be directly deleted. The specialists then manually check on the rest of the samples and deleted the unqualified ones. Next, annotators from the *stage 1* would continue the annotation until collecting 100 input-output pairs per task. The specialists also check on the distribution of the labels and answers, to avoid similar input-output pairs for the task. In ***stage 3***, four researchers with NLP backgrounds conduct a final check by inspecting randomly sampled 20 data

[2] https://stardust.ai

points from the 150 tasks. If one of the samples does not satisfy the annotation requirements, the task will be returned to the beginning of the annotation pipeline until it passes verification. Such a pipeline of three stages costs approximately $24K.

Detailed Categories. To further improve CIF-Bench's task diversity, we create 37 additional new tasks and state the related Chinese instructions. Specifically, we focus on adding Chinese tasks about **Creative Natural Language Generation**, **Traditional Chinese**, and **Complex Role-Playing Text Games**. We ask the expert native speakers to annotate 200 samples per task based on the translated task instructions. These samples are deduplicated and we further select the ***Public*** and ***Private*** split from it. Each task is further annotated with 4 ***Private*** paraphrased instructions to test whether LLMs understand the Chinese instructions' meanings or overfit to the instructions in the ***Public*** split. Each sample and instruction is manually verified or written by the authors to make sure that CIF-Bench is reliable.

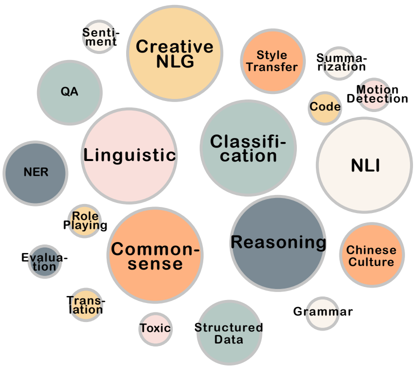

Figure 2: Task Category Distribution in CIF-Bench. The radii have three groups, determined by the number of tasks contained (≤ 10, ≤ 20, and > 20).

system intel

你是一个现代汉语专家。下面你会看到一个判断给定文本修辞手法的问题，对应的标准答案，以及一个学生答案。你需要判断学生答案是否正确，以及给出对应的原因。请以标准JSON形式返回，形如：{"原因": 你判断的原因, "标准答案对或错": 这里只能返回对或错, "对或错": 这里只能返回对或错}。

You are an expert in Modern Chinese. You will find a question to assess the rhetorical methods used in a given text, along with the reference answer and a student's answer. You need to determine whether the student's answer is correct, and provide the corresponding reason. Please return the response in the standard JSON format, as follows: {"reason": your judegment, "correctness of the reference answer": can only be either true or false, "correctness": can only be either true or false}.

user prompt

题目要求是：{instruction}
题目输入：{input}
标准答案：{refernce output}
学生答案：{output}

requirements: {instruction}
input: {input}
reference answer: {refernce output}
student answer: {output}

Figure 3: An Exemplar Prompt for GPT-4 Evaluator for the Task "Chinese Rhetoric Detection".

3.2 Task Category

Whilst diverse tasks are provided in CIF-Bench, it would be difficult to analyze the extensive scores from all of the tasks. By reviewing and summarizing the existing NLP tasks and instruction-following benchmarks, we accordingly categorize the 150 tasks into **20** basic types in a multi-label fashion (i.e., a task can be belong to more than one category). Each category consists of 2 to 36 tasks and the quantity distribution is revealed in Figure 2. Other than the 36 "commonsense" tasks requiring a wide-ranging knowledge base, there are two dominant categories that aim to challenge the logical reasoning abilities of LLMs in CIF-Bench, including 30 "natural language inference (NLI)" and 29 "reasoning" tasks. In particular, there are 18 tasks designed to require knowledge of unique Chinese cultural contexts. We describe the definition of each category and the task numbers in Appendix A.2.

3.3 Task-based Automatic Evaluation

As the CIF-Bench aims to provide a comprehensive evaluation of the LLM instruction-following capability, we argue that the metrics should be designed case by case in task granularity to evaluate the open-ended textual outputs, rather than simply reformatting all tasks into choice questions and using the conditional probability to approximate the models' predictions.

After a thorough review of the task instructions, we categorize the output requirements into the four following types and design corresponding task-level metrics. **Multi-class Classification**: We use **accuracy** as the metric if the task requires the model to predict one label from 2 or more classes in the output. **Multi-label Classification**: We use **F1 score** as the metric if the task requires the model to predict one label from 2 or more classes in the output. **Creative Generation**: Regarding the tasks that have no absolute criteria of the standard answer, we require a model-based evaluator to pro-

Table 3: Overall results in CIF-Bench *Private* split with diversified instructions (1/2). The first column is the average score across *all* the tasks, and the other columns are average scores grouped by task categories. The cells are highlighted with fading colors from maximum to minimum in a column.

Model Name	Overall	Chinese Culture	Classification	Code	Commonsense	Creative NLG	Evaluation	Grammar	Linguistic	Motion Detection	NER
Baichuan2-13B-Chat	0.529	0.520	0.674	0.333	0.641	0.497	0.686	0.542	0.528	0.578	0.563
Qwen-72B-Chat	0.519	0.486	0.630	0.296	0.634	0.508	0.634	0.458	0.520	0.494	0.550
Yi-34B-Chat	0.512	0.483	0.606	0.347	0.623	0.497	0.598	0.480	0.490	0.575	0.525
Qwen-14B-Chat	0.500	0.481	0.582	0.307	0.614	0.494	0.645	0.428	0.475	0.496	0.513
Deepseek-LLM-67B-Chat	0.471	0.467	0.571	0.259	0.577	0.486	0.549	0.442	0.476	0.475	0.509
Baichuan-13B-Chat	0.450	0.408	0.491	0.286	0.552	0.439	0.670	0.417	0.422	0.482	0.486
Chatglm3-6B	0.436	0.381	0.439	0.330	0.541	0.452	0.577	0.310	0.358	0.436	0.453
Yi-6B-Chat	0.417	0.402	0.454	0.313	0.523	0.425	0.506	0.383	0.383	0.487	0.396
Baichuan2-7B-Chat	0.412	0.437	0.647	0.160	0.520	0.402	0.580	0.511	0.444	0.455	0.407
Chatglm2-6B	0.352	0.278	0.469	0.346	0.403	0.424	0.535	0.274	0.397	0.406	0.240
Chatglm-6B-Sft	0.349	0.265	0.454	0.365	0.385	0.462	0.554	0.296	0.379	0.427	0.232
Chinese-Llama2-Linly-13B	0.344	0.250	0.462	0.311	0.399	0.429	0.557	0.273	0.358	0.385	0.268
GPT-3.5-Turbo-Sft	0.343	0.269	0.427	0.298	0.389	0.395	0.575	0.325	0.365	0.389	0.226
Chinese-Alpaca-2-13B	0.341	0.242	0.421	0.356	0.382	0.442	0.602	0.256	0.363	0.430	0.210
Chinese-Alpaca-13B	0.334	0.250	0.399	0.348	0.364	0.435	0.616	0.275	0.349	0.421	0.223
Chinese-Alpaca-7B	0.334	0.216	0.412	0.378	0.381	0.425	0.576	0.265	0.359	0.393	0.243
Chinese-Llama2-Linly-7B	0.333	0.218	0.451	0.330	0.396	0.427	0.583	0.248	0.350	0.410	0.231
Tigerbot-13B-Chat	0.331	0.205	0.397	0.309	0.385	0.420	0.614	0.310	0.379	0.341	0.276
Telechat-7B	0.329	0.267	0.338	0.321	0.420	0.404	0.420	0.272	0.265	0.327	0.320
Ziya-Llama-13B	0.329	0.196	0.402	0.324	0.341	0.428	0.616	0.312	0.349	0.400	0.228
Chinese-Alpaca-33B	0.326	0.234	0.370	0.372	0.364	0.429	0.614	0.246	0.318	0.377	0.221
Tigerbot-7B-Chat	0.325	0.218	0.395	0.306	0.370	0.413	0.631	0.294	0.370	0.368	0.215
Chinese-Alpaca-2-7B	0.323	0.215	0.374	0.335	0.366	0.415	0.546	0.257	0.326	0.395	0.215
Aquilachat-7B	0.309	0.162	0.234	0.291	0.320	0.437	0.344	0.135	0.266	0.309	0.287
Moss-Moon-003-Sft	0.302	0.214	0.405	0.274	0.347	0.380	0.448	0.305	0.341	0.378	0.232
Qwen-7B-Chat	0.301	0.211	0.410	0.289	0.349	0.391	0.531	0.219	0.387	0.404	0.208
Belle-13B-Sft	0.264	0.198	0.307	0.285	0.316	0.349	0.409	0.237	0.305	0.222	0.177
CPM-Bee-10B	0.244	0.234	0.377	0.024	0.278	0.311	0.255	0.302	0.278	0.327	0.148

vide information regarding a given output, including **creativity**, **fluency**, the level of **instruction-following**, and the **confidence** of the evaluator. **Semantic Similarity**: For the remaining tasks that can be evaluated by the semantic similarity between the golden reference and model output, we use a pre-trained language All scores used in CIF-Bench either naturally range from 0 to 1, or are normalized to the same range.

One core dilemma in evaluating the open-ended instruction-following capabilities of LLMs is that model predictions are hard to verify even with reference answers. For instance, it is intractable to handcraft regex rules to extract the predictions from LLMs for the extensive number of tasks, since the answers could be expressed in various formats, or drowned in redundant contexts like reasoning progress. Inspired by G-Eval (Liu et al., 2023), we leverage OpenAI's GPT-4[3] as a relatively reliable evaluator for multi-class classification, multi-label classification, and creative generation tasks, to overcome such issues. The GPT-4 evaluator is prompted to assess the outputs according to the given task instruction and the input-output reference, as shown by the example in Figure 3 and the full list of evaluation prompts in Appendix A.1. the remaining tasks that can be evaluated by the semantic similarity between the golden reference and model output, we use a lightweight multilingual encoder, BLEURT (Sellam et al., 2020), to measure the relevance between the reference and LLM output.

Given a set of task instructions I, we denote the performance score of model m on task t as:

$$S_t^m = \frac{1}{|D_t|} \sum_{d \in D_t} \frac{1}{|I|} \sum_{i \in I} s_t^m(i, d)$$

, where D_t refers to the set of data samples for task t. In the case of the *public* split, the instruction set I is reduced to one single element. In we take the average of task-level scores $\overline{S^m}$ as the indicator of overall performance for a model m.

4 Experiments

Baselines. We compare the performance of existing LLMs that have been trained on Chinese corpora. We select ChatGPT, for which we use `gpt-3.5-turbo-instruct`,[4] which we believe corresponds to instructGPT text-davinci-002. Then we select a series of open-source LLMs, including `ChatGLM` (Zeng et al., 2023), `AquilaChat-7B`.[5] `Baichuan` (Baichuan, 2023), `Deepseek-Llm-67B-Chat` (DeepSeek-AI, 2024), `Qwen` (Bai et al., 2023), `Yi`,[6] `tigerbot-7b-chat` (Chen et al., 2023), `TeleChat` (Wang et al., 2024), `CPM-Bee-10B`,[7], and `Moss-Moon` (Sun et al.,

[3]https://openai.com/gpt-4

[4]https://openai.com/

[5]https://github.com/FlagAI-Open/FlagAI/

[6]https://github.com/OrionStarAI/OrionStar-Yi-34B-Chat/tree/main

[7]https://github.com/OpenBMB/CPM-Bee

Table 4: Overall results in CIF-Bench *Private* split with diversified Instructions (2/2). The first column is the average score across *all* the tasks, and the rest columns are average scores grouped by task categories. The cells are highlighted with fading colors from maximum to minimum in a column.

Model Name	Overall	NLI	QA	Reasoning	Role Playing	Sentiment	Structured Data	Style Transfer	Summarization	Toxic	Translation
Baichuan2-13B-Chat	0.529	0.632	0.569	0.515	0.752	0.624	0.459	0.462	0.332	0.441	0.273
Qwen-72B-Chat	0.519	0.626	0.565	0.528	0.762	0.613	0.496	0.459	0.282	0.608	0.271
Yi-34B-Chat	0.512	0.619	0.554	0.494	0.757	0.580	0.472	0.439	0.346	0.514	0.259
Qwen-14B-Chat	0.500	0.616	0.548	0.507	0.764	0.583	0.469	0.453	0.283	0.575	0.262
Deepseek-LLM-67B-Chat	0.471	0.566	0.496	0.439	0.711	0.546	0.409	0.436	0.262	0.570	0.235
Baichuan-13B-Chat	0.450	0.565	0.505	0.377	0.704	0.552	0.387	0.402	0.350	0.431	0.304
Chatglm3-6B	0.436	0.544	0.503	0.414	0.762	0.560	0.446	0.402	0.321	0.391	0.270
Yi-6B-Chat	0.417	0.523	0.457	0.369	0.754	0.482	0.401	0.380	0.310	0.455	0.227
Baichuan2-7B-Chat	0.412	0.489	0.395	0.406	0.670	0.517	0.342	0.298	0.101	0.463	0.138
Chatglm2-6B	0.352	0.397	0.352	0.326	0.714	0.438	0.298	0.313	0.320	0.461	0.190
Chatglm-6B-Sft	0.349	0.380	0.321	0.292	0.718	0.415	0.296	0.333	0.351	0.441	0.190
Chinese-Llama2-Linly-13B	0.344	0.390	0.330	0.313	0.653	0.433	0.279	0.332	0.292	0.457	0.181
GPT-3.5-Turbo-Sft	0.343	0.382	0.394	0.345	0.710	0.433	0.324	0.266	0.290	0.397	0.225
Chinese-Alpaca-2-13B	0.341	0.376	0.334	0.317	0.714	0.459	0.299	0.316	0.308	0.452	0.200
Chinese-Alpaca-13B	0.334	0.370	0.309	0.319	0.724	0.426	0.285	0.307	0.298	0.445	0.181
Chinese-Alpaca-7B	0.334	0.383	0.326	0.295	0.710	0.409	0.301	0.327	0.325	0.405	0.186
Chinese-Llama2-Linly-7B	0.333	0.367	0.345	0.276	0.698	0.433	0.259	0.315	0.310	0.469	0.168
Tigerbot-13B-Chat	0.331	0.363	0.329	0.301	0.694	0.419	0.280	0.310	0.283	0.393	0.186
Telechat-7B	0.329	0.388	0.355	0.244	0.672	0.344	0.334	0.335	0.299	0.364	0.184
Ziya-Llama-13B	0.329	0.351	0.279	0.313	0.721	0.468	0.311	0.291	0.278	0.431	0.175
Chinese-Alpaca-33B	0.326	0.368	0.300	0.314	0.713	0.428	0.288	0.303	0.295	0.401	0.199
Tigerbot-7B-Chat	0.325	0.355	0.313	0.292	0.713	0.415	0.283	0.315	0.290	0.389	0.171
Chinese-Alpaca-2-7B	0.323	0.375	0.318	0.289	0.698	0.417	0.285	0.303	0.312	0.439	0.193
Aquilachat-7B	0.309	0.337	0.342	0.236	0.609	0.255	0.249	0.400	0.527	0.430	0.306
Moss-Moon-003-Sft	0.302	0.317	0.321	0.267	0.694	0.375	0.251	0.259	0.288	0.424	0.152
Qwen-7B-Chat	0.301	0.325	0.297	0.278	0.681	0.419	0.266	0.251	0.248	0.371	0.157
Belle-13B-Sft	0.264	0.317	0.284	0.242	0.631	0.299	0.244	0.222	0.234	0.296	0.133
CPM-Bee-10B	0.244	0.286	0.224	0.147	0.603	0.277	0.117	0.263	0.220	0.352	0.125

2023), which have been trained from scratch on a large volume of data in both English and Chinese. We additionally select other instruction-following LLMs, such as `Ziya-LLaMA-13B` (Wang et al., 2022a), `Chinese-Alpaca` (Cui et al., 2023), `Linly-Chinese-LLaMA2` (Zhao et al., 2023), and `BELLE` (BELLEGroup, 2023), which are trained with Supervised Fine-Tuning (SFT) on Chinese data, including web texts, books, and code, and then trained via alignment techniques.

Settings. For inference, we use four Nvidia A100 GPUs with 80GB of VRAM. To optimize GPU resource usage, we directly employed the vLLM framework (Kwon et al., 2023) for LLM inference on CIF-Bench where applicable. This setup enables each model to complete all tasks within approximately 6 to 12 hours. For models not supported by the vLLM, we adhere to the configurations specified in official repositories, resulting in an inference duration ranging from 12 to 48 hours. During the evaluation, we use two Nvidia 2080-Ti 12GB GPUs to conduct the BLEURT semantic similarity calculations, and use the `gpt-4-turbo-preview` version of GPT-4 API as the open-ended evaluator for the rest of tasks.

5 Results Analysis

Broadly speaking, we aim to investigate the performance capabilities of current representative Chinese LLMs in a diverse set of NLP tasks to ascertain how well the annotated data with human

Table 5: Comparison between English-translated and newly annotated Chinese tasks in the *Public* split.

Model	SNI Task	New Task
Qwen-72B-Chat	0.588	0.573
Qwen-14B-Chat	0.573	0.535
Deepseek-LLM-67B-Chat	0.529	0.504
gpt-3.5-public-turbo	0.523	0.500
Yi-34B-Chat	0.509	0.514

Table 6: Comparison of the CIF-Bench overall scores in the *Public* split and other leaderboards. The cells are highlighted with fading colors from maximum to minimum for the applicable numbers in a column. * indicates that the performance of pre-trained base LLMs is used to approximate the evaluation of the corresponding unavailable chat models.

Model Name	CIF *Public*	Open LLM	OpenCompass
Qwen-72B-Chat	0.589	*73.60	51.90
Qwen-14B-Chat	0.564	*65.86	45.00
Deepseek-LLM-67B-Chat	0.526	71.79	42.70
gpt-3.5-Public-SFT	0.522	-	46.80
Yi-34B-Chat	0.516	65.32	47.10
Baichuan2-13B-Chat	0.512	-	32.10
Tigerbot-13B-Chat	0.494	*53.42	-
Chinese-Alpaca-2-13B	0.492	57.41	-
Chinese-Alpaca-33B	0.484	55.33	-
Ziya-Llama-13B	0.479	29.96	-
Chinese-Llama2-Linly-13B	0.479	-	-
Tigerbot-7B-Chat	0.478	*47.93	-
ChatGLM3-6B	0.472	-	35.20
Chinese-Alpaca-13B	0.471	-	-
ChatGLM2-6B	0.464	-	-
Chinese-Alpaca-7B	0.452	48.85	-
Chinese-Alpaca-2-7B	0.448	-	-
Chinese-Llama2-Linly-7B	0.443	45.44	-
Qwen-7B-Chat	0.442	*59.19	37.10
ChatGLM-6B	0.440	-	-
Baichuan-13B-Chat	0.426	-	-
Yi-6B-Chat	0.420	*54.08	31.90
CPM-Bee-10B	0.415	-	-
Moss-Moon-003-SFT	0.399	-	-
Belle-SFT-Public	0.397	-	-
Telechat-7B	0.350	-	-
Aquilachat-7B	0.350	-	-
Baichuan2-7B-Chat	0.339	51.42	29.40

Table 7: Overall performance differences in CIF-Bench from *Public* to *Private* splits with single instructions.

Model Name	Score Difference↑	Model Name	Score Difference↑
Aquilachat-7B	-0.050↓	Chinese-Llama2-Linly-7B	-0.122↓
Baichuan-13B-Chat	0.020↑	CPM-Bee-10B	-0.178↓
Baichuan2-13B-Chat	0.006↑	Deepseek-LLM-67B-Chat	-0.060↓
Baichuan2-7B-Chat	0.071↑	gpt-3.5-Public-SFT	-0.187↓
Belle-SFT-Public	-0.145↓	Moss-Moon-003-SFT	-0.110↓
ChatGLM-6B	-0.112↓	Qwen-14B-Chat	-0.068↓
ChatGLM2-6B	-0.124↓	Qwen-72B-Chat	-0.068↓
ChatGLM3-6B	-0.038↓	Qwen-7B-Chat	-0.145↓
Chinese-Alpaca-13B	-0.148↓	Telechat-7B	-0.029↓
Chinese-Alpaca-2-13B	-0.171↓	Tigerbot-13B-Chat	-0.180↓
Chinese-Alpaca-2-7B	-0.138↓	Tigerbot-7B-Chat	-0.163↓
Chinese-Alpaca-33B	-0.170↓	Yi-34B-Chat	-0.014↓
Chinese-Alpaca-7B	-0.125↓	Yi-6B-Chat	-0.008↓
Chinese-Llama2-Linly-13B	-0.147↓	Ziya-Llama-13B	-0.167↓

Table 8: The performance shift caused by unseen data instances and unseen tasks. Note that in the column "Existing" task, only the newly annotated and existing input-output data instances are compared while the task instruction remains the same. In the "Existing→New" setting, both data instances and tasks are changed.

Model ↓ Task→	Existing	Existing→New	Model↓	Existing	Existing→New
Aquilachat-7B	-0.047↓	-0.034↓	Chinese-Llama2-Linly-7B	-0.134↓	-0.047↓
Baichuan-13B-Chat	0.023↑	0.027↑	CPM-Bee-10B	-0.176↓	0.046↑
Baichuan2-13B-Chat	-0.003↓	0.008↑	Deepseek-LLM-67B-Chat	-0.076↓	-0.029↓
Baichuan2-7B-Chat	0.072↑	0.077↑	gpt-3.5-Public-SFT	-0.202↓	-0.029↓
Belle-SFT-Public	-0.167↓	-0.054↓	Moss-Moon-003-SFT	-0.124↓	-0.028↓
ChatGLM-6B	-0.120↓	-0.033↓	Qwen-14B-Chat	-0.088↓	-0.038↓
ChatGLM2-6B	-0.131↓	-0.005↓	Qwen-72B-Chat	-0.082↓	-0.021↓
ChatGLM3-6B	-0.060↓	-0.052↓	Qwen-7B-Chat	-0.157↓	0.005↑
Chinese-Alpaca-13B	-0.164↓	-0.081↓	Telechat-7B	-0.050↓	-0.045↓
Chinese-Alpaca-2-13B	-0.179↓	-0.067↓	Tigerbot-13B-Chat	-0.187↓	-0.017↓
Chinese-Alpaca-2-7B	-0.152↓	-0.051↓	Tigerbot-7B-Chat	-0.162↓	-0.004↓
Chinese-Alpaca-33B	-0.187↓	-0.072↓	Yi-34B-Chat	-0.025↓	-0.002↓
Chinese-Alpaca-7B	-0.147↓	-0.072↓	Yi-6B-Chat	-0.022↓	-0.012↓
Chinese-Llama2-Linly-13B	-0.153↓	-0.031↓	Ziya-Llama-13B	-0.181↓	-0.045↓

standards with the provided instruction-following benchmark. Specifically, we ask: *(i)* Is our benchmark challenging enough? What kind of tasks are difficult? *(ii)* Is it true that LLMs perform worse when language is transferred? *(iii)* Do we measure the instruction-following capability well, by avoiding data contamination? *(iv)* Do the diverse instructions help?

Is CIF-Bench Challenging? To ensure the reliability of our benchmark, the scores in the *private* split with the diversified instructions are referred to as the main results for discussion, as shown in Table 3 and Table 4. Our findings reveal that although large parameter size contributes to performance (`Qwen-72B-Chat`, `Yi-34B-Chat`, and `Deepseek-LLM-67B-Chat`), the effective training methods are still a boost for relatively small models such as `Baichuan2-13B-Chat` and `Qwen-14B-Chat`. Given that the highest score barely reaches 52.9 overall out of 100 and only 4 models exceed 50.0, we conclude that our proposed CIF is a tough benchmark for existing LLMs for question *(i)*.

In addition, we provide finer-grained score aggregation to further analyze the challenging task categories (n.b., most bilingual LLMs perform poorly on tasks in code, summarization, and translation categories). In the code category, the models might misunderstand the semantics expressed in Chinese for the newly defined variable or function. Specifically, models usually perform poorly in a new "programming language" environment that requires the model to understand restricted actions. As for summarization tasks, models could misinterpret the instruction, eg. models sometimes consider the instruction "modify the input into a more friendly expression to non-native speakers" as a Chinese-English translation task and might provide redundant explanations even if not required by the instructions and hence will cause large semantic distances to the golden reference. We point out that Chinese-commented code corpora and parallel translation data of Chinese and other languages are still scarce resources, which might lead to their poor performance on CIF-Bench's code and translation categories. Additionally, we assume that Chinese and English bilingual LLMs, although a major branch of multilingual LLM, do not significantly benefit LLMs' capacity to deal with minorlanguage-related tasks. Part of the tasks in CIF-Bench's summarization category are very challenging, combining counterfactual reasoning and empathy estimation (i.e., task 125 and task 131 referring to Appendix A.1). Thereby, the bilingual LLMs' poor performance on CIF-Bench's summarization category is understandable. Detailed category-based scores on the *public* split are available in Table 13 in Appendix B for further analysis.

Language Transferability. We select the *public* split to investigate LLM language transferability in instruction-following. In the CIF-Bench *public* split, a set of 70 tasks from SNI (Wang et al., 2022b) are used as representative samples of English NLP tasks equipped with directly translated input-output pairs in Chinese. We select the top-5 performing models on the *public* split to show the performance comparison between SNI and our 37 original curated Chinese tasks in Table 5. Although these models maintain instruction-following capability when encountering the translated SNI data, they generally perform worse on tasks newly created in Chinese without a corresponding "copy" in English, which yields an average score decrement of 2.2%.

Data Contamination Does Exist. As mentioned in §3, we evaluate the model performances on the *public* split with half of the input-output pairs in the single instruction setting, with which we can conveniently probe the benchmark data contamination issue of the LLMs.

We first compare the CIF-Bench *public* results with two comprehensive LLM benchmarks, including the Open LLM Leaderboard (Beeching et al., 2023), as well as an English-Chinese leaderboard, OpenCompass (Contributors, 2023). As suggested in Table 6 with rows ranked in the descending order of the overall *public* scores, the results on CIF-Bench are aligned with the other two popular benchmarks, which therefore verifies the reliability of our evaluation pipeline. However, we suspect the highly correlative rankings could be a result of the benchmark data leakage in those "web-scale" pre-training data, since 117 of the constructed tasks and instances in the *public* split are sourced from the internet.

To further confirm such suspicions, we calculate the performance changes of overall scores in the same single instruction setting, but with different input-output pairs from the *public* and *private* splits. Revealed by the differences in Table 7, there is a noticeable performance drop for most (25/28) of the models when a large part of the data translated from public sources is replaced by our original annotations. Consequently, incoming models submitted to the proposed CIF-Bench will restricted to the *private* split for the sake of evaluation reliability.

It is likely that both the leakage of the input-output instances and the tasks themselves contribute to the mentioned evaluation bias. To compare the two factors for the downgraded performances, we analyze the performance shift with the 113 "Existing" tasks translated from English or originally in Chinese and the 37 "New" tasks we crafted from scratch. As revealed in Table 8, the LLMs have impaired performance when given newly curated data instances for a set of seen "Existing" tasks, yielding an average 11.0% score decrease. In contrast, these models on average perform 2.5% worse, with both definitely-unseen tasks and corresponding input-output pairs. We hence conclude that the leakage of the data instances plays a more significant role than the tasks themselves in evaluation biases.

Instruction Diversity for Evaluation Robustness. With the motivation that a model might produce inconsistent output given various instruction,input holding the same semantics, we argue that a diversified instruction set can increase the evaluation robustness by incorporating more corner cases. We separately calculate the task-level score variance in the *private* split for the conditions of using *one* and *five* instructions to verify the improvement. We find that increasing the diversity of the task instructions can bring extra robustness to the evaluation, as the evaluation scores are stabilized to lower variance for all the tested LLMs (see in Table 9).

Human Annotation for Verification. To verify the annotation quality and reliability , we invite 3 annotators with expert-level NLP research backgrounds to assess the model outputs in *public* split with the same task-level instruction. The evaluation dimensions include: "Faithfulness": human experts reflect on the absolute quality of a model's output in a binary (yes/no) form. "Level of preference": a 5-point Likert scale was provided to the experts to assess the relative quality of the model outputs. We randomly sample tasks according to the task category distribution, and pick three models performing differently

Table 9: The difference of variance of task-level scores from single to diverse instruction sets. The variance values are scaled by a factor of 1×10^{-3}.

Model Name	Var. Difference↓	Model Name	Var. Difference↓
Aquilachat-7B	-3.961↓	Chinese-Llama2-Linly-7B	-4.539↓
Baichuan-13B-Chat	-7.049↓	Cpm-Bee-10B	-0.661↓
Baichuan2-13B-Chat	-3.633↓	Deepseek-Llm-67B-Chat	-2.889↓
Baichuan2-7B-Chat	-1.402↓	Gpt-3.5-Turbo-Sft	-6.369↓
Belle-13B-Sft	-0.316↓	Moss-Moon-003-Sft	-6.827↓
Chatglm-6B-Sft	-5.051↓	Qwen-14B-Chat	-0.978↓
Chatglm2-6B	-3.980↓	Qwen-72B-Chat	-1.817↓
Chatglm3-6B	-0.413↓	Qwen-7B-Chat	-3.185↓
Chinese-Alpaca-13B	-8.303↓	Telechat-7B	-6.090↓
Chinese-Alpaca-2-13B	-4.814↓	Tigerbot-13B-Chat	-3.816↓
Chinese-Alpaca-2-7B	-4.494↓	Tigerbot-7B-Chat	-6.004↓
Chinese-Alpaca-33B	-5.000↓	Yi-34B-Chat	-1.942↓
Chinese-Alpaca-7B	-2.961↓	Yi-6B-Chat	-6.397↓
Chinese-Llama2-Linly-13B	-2.961↓	Ziya-Llama-13B	-3.001↓

in general, specifically `Moss-Moon-003-sft` (0.399), `Baichuan-13B-Chat` (0.426), and `Qwen-72B-Chat` (0.589). Considering the diverse and open-ended task, we first measure quality by comparing the pairwise agreement between two annotators, reporting an average agreement of 0.49. Furthermore, we employ Cohen's kappa (Ben-David, 2008) to measure inter-rater reliability, reporting an average of 0.3729 across the 153 questions, implying that the results are substantially reliable. Specifically, the experts scored 0.4966 on the dichotomous form yet 0.2492 on the more varied options, suggesting that completing 153 questions is challenging even for human experts. We further explore the correlation between the model prediction with human evaluation(Spearman's $r = 0.4043$), suggesting that most annotated were indeed truthful and the models can be relied upon to generate output for this task.

6 Conclusion

In summary, CIF-Bench not only exposes the limitations of current LLMs in navigating the complexities of Chinese language instruction-following tasks but also provides a foundational platform for future advancements in LLM generalizability research. Through this work, we aim to facilitate the development of more adaptable, culturally aware, and linguistically diverse language models, capable of truly understanding and interacting with the global tapestry of human language.

Limitations

Recruiting human subjects for annotation limits the reproducibility of human evaluation. In addition, we recognize that there might be more suitable baseline models, whilst in this study only a few of the most advanced models were used. Finally, despite annotation and discrimination by human experts, there may still be offensive content in the data due to both human education and environmental factors. It is worth noting, however, that identifying offensive language is not the purpose of this work.

Ethics Statement

The dataset presented was annotated by a third-party professional annotation company. During the annotation process, we considered the following aspects to ensure the protection of the annotators. (1) Consent: To ensure that our participants agreed to the annotation task, we asked them to read the task guidelines and instructions before starting the work. If they felt uncomfortable, they could withdraw from the task at any time. (2) Confidentiality: The entire annotation process was anonymous and we did not know any information about the participants in the task. (3) Assurance: all data were obtained from open-source datasets or resources.

Acknowledgements

Yizhi Li and Xingwei Qu are Ph.D. students funded by the Department of Computer Science, University of Manchester, UK. Yinghao Ma is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported by UK Research and Innovation [grant number EP/S022694/1]. The research was supported by Theme-based Research Scheme (T45-205/21-N) from Hong Kong RGC, NSFC (No. 62206234) and Generative AI Research and Development Centre from InnoHK.

References

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus*, 15.

Shatha Altammami, Eric Atwell, and Ammar Alsalka. 2020. Constructing a bilingual Hadith corpus using a segmentation tool. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3390–3398, Marseille, France. European Language Resources Association.

Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rachel Bawden, Eric Bilinski, Thomas Lavergne, and Sophie Rosset. 2021. Diabla: A corpus of bilingual spontaneous written dialogues for machine translation. *Language Resources and Evaluation*, 55:635–660.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

BELLEGroup. 2023. Belle: Be everyone's large language model engine. `https://github.com/LianjiaTech/BELLE`.

Arie Ben-David. 2008. Comparison of classification accuracy using cohen's weighted kappa. *Expert Syst. Appl.*, 34(2):825–832.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Ye Chen, Wei Cai, Liangmin Wu, Xiaowei Li, Zhanxuan Xin, and Cong Fu. 2023. Tigerbot: An open multilingual multitask LLM. *CoRR*, abs/2312.08688.

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*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. Textworld: A learning environment for text-based games. *CoRR*, abs/1806.11532.

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

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Content "European Commission, Directorate-General for Communications Networks and Technology.". 2017. "spanish-english website parallel corpus.".

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir

Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023a. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Md Adnanul Islam, Md Saidul Hoque Anik, and ABM Alim Al Islam. 2022. An enhanced rbmt: When rbmt outperforms modern data-driven translators. *IETE Technical Review*, 39(6):1473–1484.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Brenden M Lake and Marco Baroni. 2017. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. arxiv.

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

Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, et al. 2023. Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434*.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Data Market. 2018. shujujishi.com. `http://shujujishi.com/dataset/a037ab86-7727-487b-9a46-2936b0be076b.html`. Accessed 16-02-2024.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Yusuke Oda. 2016. Small parallel enja. `https://github.com/odashi/small_parallel_enja`.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. *arXiv preprint arXiv:2011.03020*.

Jiaxin Pei and David Jurgens. 2021. Measuring sentence-level and aspect-level (un)certainty in science communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Carla Perez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2022. SemEval-2022 task 4: Patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 298–307, Seattle, United States. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012. Morphological processing for english-tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, pages 113–122.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Parth Shah and Vishvajit Bakrola. 2019. Neural machine translation system of indic languages-an attention based approach. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pages 1–5. IEEE.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Yuen-Hsien Tseng, Wun-Syuan Wu, Chia-Yueh Chang, Hsueh-Chih Chen, and Wei-Lun Hsu. 2020. Development and validation of a corpus for machine humor comprehension. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1346–1352.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-2020 task 4: Commonsense validation and explanation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 307–321, Barcelona (online). International Committee for Computational Linguistics.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. 2022a. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Zhongjiang He, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, Yan Wang, Xin Wang, Luwen Pu, Huihan Xu, Ruiyu Fang, Yu Zhao, Jie Zhang, Xiaomeng Huang, Zhilong Lu, Jiaxin Peng, Wenjun Zheng, Shiquan Wang, Bingkai Yang, Xuewei He, Zhuoru Jiang, Qiyi Xie, Yanhan Zhang, Zhongqiu Li, Lingling Shi, Weiwei Fu, Yin Zhang, Zilu Huang, Sishi Xiong, Yuxiang Zhang, Chao Wang, and Shuangyong Song. 2024. Telechat technical report. *CoRR*, abs/2401.03804.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wikipedia. 2024. List of China Mainland Internet Language — Wikipedia, the free encyclopedia. http://zh.wikipedia.org/w/index.php?title=%E4%B8%AD%E5%9B%BD%E5%A4%A7%E9%99%86%E7%BD%91%E7%BB%9C%E7%94%A8%E8%AF%AD%E5%88%97%E8%A1%A8&oldid=81048845.

Yiquan Wu, Yifei Liu, Ziyu Zhao, Weiming Lu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2024. De-biased attention supervision for text classification with causality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19279–19287.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration. *arXiv preprint arXiv:2310.09241*.

Xiangyu Xi, Jianwei Lv, Shuaipeng Liu, Wei Ye, Fan Yang, and Guanglu Wan. 2022. Musied: A benchmark for event detection from multi-source heterogeneous informal texts. *arXiv preprint arXiv:2211.13896*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Liang Xu, Xiaojing Lu, Chenyang Yuan, Xuanwei Zhang, Huilin Xu, Hu Yuan, Guoao Wei, Xiang Pan, Xin Tian, Libo Qin, et al. 2021. Fewclue: A chinese few-shot learning evaluation benchmark. *arXiv preprint arXiv:2107.07498*.

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. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Zhengyan Zhang, Chaojun Xiao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, et al. 2021. Cuge: A chinese language understanding and generation evaluation benchmark. *arXiv preprint arXiv:2112.13610*.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. *arXiv preprint arXiv:2104.08835*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Ge Zhang, Yizhi Li, Yaoyao Wu, Linyuan Zhang, Chenghua Lin, Jiayi Geng, Shi Wang, and Jie Fu. 2023a. Corgi-pm: A chinese corpus for gender bias probing and mitigation. *arXiv preprint arXiv:2301.00395*.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, and Alham Fikri Aji. 2023b. Multilingual large language models are not (yet) code-switchers. *arXiv preprint arXiv:2305.14235*.

Zhe Zhao, Yudong Li, Cheng Hou, Jing Zhao, Rong Tian, Weijie Liu, Yiren Chen, Ningyuan Sun, Haoyan Liu, Weiquan Mao, Han Guo, Weigang Guo, Taiqiang Wu, Tao Zhu, Wenhang Shi, Chen Chen, Shan Huang, Sihong Chen, Liqun Liu, Feifei Li, Xiaoshuai Chen, Xingwu Sun, Zhanhui Kang, Xiaoyong Du, Linlin Shen, and Kimmo Yan. 2023. Tencentpretrain: A scalable and flexible toolkit for pre-training models of different modalities. In *ACL (demo)*, pages 217–225. Association for Computational Linguistics.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Online and Dublin, Ireland. Association for Computational Linguistics.

A Task Details

A.1 Full List of Tasks and Evaluation

We provide a full list of the task names and the source for input-output annotation in this subsection. The comprehensive task descriptions and the corresponding evaluation prompts can be found in the supplementary files.

Table 10: Full task list and source (1/3).

Task ID & Name	**Source**
0 Negotiation Strategy Detection	SNI (Wang et al., 2022b)
1 Grammar Error Correction	SNI (Wang et al., 2022b)
2 Overlap Extraction	SNI (Wang et al., 2022b)
3 Commonsense	SNI (Wang et al., 2022b)
4 Data to Text	SNI (Wang et al., 2022b)
5 Keyword Tagging	SNI (Wang et al., 2022b)
6 Answerability Classification	SNI (Wang et al., 2022b)
7 Dialogue Act Recognition	SNI (Wang et al., 2022b)
8 Cause Effect Classification	SNI (Wang et al., 2022b)
9 Question Rewriting	SNI (Wang et al., 2022b)
10 Textual Entailment	SNI (Wang et al., 2022b)
11 Coreference Resolution	SNI (Wang et al., 2022b)
12 Title Generation	SNI (Wang et al., 2022b)
13 Entity Relation Classification	SNI (Wang et al., 2022b)
14 Punctuation Error Detection	SNI (Wang et al., 2022b)
15 Style Transfer	SNI (Wang et al., 2022b)
16 Sentence Expansion	SNI (Wang et al., 2022b)
17 Poem Generation	SNI (Wang et al., 2022b)
18 Discourse Relation Classification	SNI (Wang et al., 2022b)
19 Mathematics	SNI (Wang et al., 2022b)
20 Text Simplification	SNI (Wang et al., 2022b)
21 Sentence Compression	SNI (Wang et al., 2022b)
22 Spelling Error Detection	SNI (Wang et al., 2022b)
23 Irony Detection	SNI (Wang et al., 2022b)
24 Number Conversion	SNI (Wang et al., 2022b)
25 Word Relation Classification	SNI (Wang et al., 2022b)
26 Paraphrasing	SNI (Wang et al., 2022b)
27 Grammar Error Detection	SNI (Wang et al., 2022b)
28 Text Matching	SNI (Wang et al., 2022b)
29 Fill in The Blank	SNI (Wang et al., 2022b)
30 Speaker Relation Classification	SNI (Wang et al., 2022b)
31 Entity Generation	SNI (Wang et al., 2022b)
32 Summarization	SNI (Wang et al., 2022b)
33 Spam Classification	SNI (Wang et al., 2022b)
34 Stereotype Detection	SNI (Wang et al., 2022b)
35 Dialogue State Tracking	SNI (Wang et al., 2022b)
36 Dialogue State Tracking	SNI (Wang et al., 2022b)
37 Sentence Perturbation	SNI (Wang et al., 2022b)
38 Text Quality Evaluation	SNI (Wang et al., 2022b)
39 Linguistic Probing	SNI (Wang et al., 2022b)
40 Information Extraction	SNI (Wang et al., 2022b)
41 Emotion Prediction	SNI (Wang et al., 2022b)
42 Discourse Connective Identification	SNI (Wang et al., 2022b)
43 Question Generation	SNI (Wang et al., 2022b)
44 Stance Detection	SNI (Wang et al., 2022b)
45 Sentiment Analysis	SNI (Wang et al., 2022b)
46 Story Composition	SNI (Wang et al., 2022b)
47 Program Execution	SNI (Wang et al., 2022b)
48 Gender Classification	SNI (Wang et al., 2022b)
49 Named Entity Recognition	SNI (Wang et al., 2022b)
50 Toxic Language Detection	SNI (Wang et al., 2022b)
51 Question Decomposition	SNI (Wang et al., 2022b)
52 Sentence Ordering	SNI (Wang et al., 2022b)
53 Text to Code	SNI (Wang et al., 2022b)
54 Fact Verification	SNI (Wang et al., 2022b)
55 Speaker Identification	SNI (Wang et al., 2022b)
56 Answer Verification	SNI (Wang et al., 2022b)
57 Wrong Candidate Generation	SNI (Wang et al., 2022b)
58 Dialogue Generation	SNI (Wang et al., 2022b)
59 Text Completion	SNI (Wang et al., 2022b)
60 Pos Tagging	SNI (Wang et al., 2022b)

Table 11: Full task list and source (2/3).

Task ID & Name	Source
61 Explanation	SNI (Wang et al., 2022b)
62 Sentence Composition	SNI (Wang et al., 2022b)
63 Question Understanding	SNI (Wang et al., 2022b)
64 Intent Identification	SNI (Wang et al., 2022b)
65 Word Semantics	SNI (Wang et al., 2022b)
66 Code to Text	SNI (Wang et al., 2022b)
67 Preposition Prediction	SNI (Wang et al., 2022b)
68 Text Categorization	SNI (Wang et al., 2022b)
69 Question Answering	SNI (Wang et al., 2022b)
70 Commonsense Classification	N/A
71 Ancient Chinese Poem Retrieval	N/A
72 Ancient Chinese Translation	N/A
73 Chinese Rhyme Detection	N/A
74 Nationality Detection	N/A
75 Region Detection	N/A
76 Chinese Idiom Explanation	N/A
77 Name Allusion Detection	N/A
78 Chinese Ambiguity Sentence Location	N/A
79 Chinese Winograd Schema Challenge	FewCLUE (Xu et al., 2021)
80 Chinese Modern Abbreviation Explanation	Wikipedia (Wikipedia, 2024)
81 Chinese Epigraph Detection	N/A
82 Chinese Dialect Translation	N/A
83 Chinese Attractions List	N/A
85 Chinese Typo Categorization	N/A
86 Chinese Fiction Characteristic Detection	N/A
87 Chinese Figurative Detection	N/A
88 Chinese Metaphor Explanation	N/A
89 Chinese Medicine Detection	N/A
90 Chinese Pinyin Detection	N/A
91 Chinese Wubi Written	N/A
92 Intimacy Score Prediction	Pei and Jurgens (2020)
93 Sentence Level Uncertainty Judgement	Pei and Jurgens (2021)
94 Chinese Relative Identification	N/A
96 Chinese Heteronomous Language Detection	N/A
97 Code Debug	`https://blog.csdn.net`
98 Code Translate	`https://leetcode.cn`
99 Function Explanation	`https://www.liaoxuefeng.com/`
100 Bias Detoxication	CORGI-PM (Zhang et al., 2023a)
101 MultiLabel Chinese Humor Categorization	Tseng et al. (2020)
102 Legal Term Retrieval	N/A
103 Patronizing Condescending Multilabel	Perez-Almendros et al. (2022)
104 CommonSense Explanation	Wang et al. (2020)
105 Event Type Detection	MUSIED (Xi et al., 2022)
106 Argument Mining	N/A
107 Theory of Mind	Big-Bench Theory of Mind (bench authors, 2023)
108 Game Playing	Big-Bench Language Games (bench authors, 2023)
109 IQ Test	Huang et al. (2023a)
110 Joke Explanation	N/A
111 Role Playing	TRPG `https://bilibili.com` (Côté et al., 2018)
112 Text De-Identification	N/A
113 Outline Generation	N/A
114 Pros Cons Listing	N/A
115 Joke Telling	N/A
116 Affordance	N/A

Table 12: Full task list and source (2/3).

Task ID & Name	Source
117 Material Synthesis	Bara et al. (2021)
118 Tool use	Taskmatrix (Liang et al., 2023)
119 Concept Abstraction	N/A
120 Rhyme Aligned Generation	N/A
121 Advertising	N/A
122 Mind Tree Generation	N/A
123 First Order Logic	FOLIO (Han et al., 2022)
124 Critical Thinking	N/A
125 Empathy Detection	N/A
126 Social Norms Detection	Moral Stories (Emelin et al., 2021)
127 Make Positive	Ziems et al. (2022)
128 Translate to Ancient Chinese	N/A
129 Recipe Generation	Market (2018)
130 Imagination	N/A
131 Compositional Reasoning	Lake and Baroni (2017)
132 Personality Detection	N/A
133 Table Generation	N/A
134 Flowchart Generation	N/A
135 Review Generation	N/A
136 Draw Figure with symbol	N/A
137 CommonsenseQA	CommonsenseQA (Talmor et al., 2018)
138 ReadingComprehensionQA	Rajpurkar et al. (2018)
139 DiscreteOperationQA	DROP(Dua et al., 2019)
140 MultiHopQA	HotpotQA (Yang et al., 2018)
141 CommonsenseNLI	HellaSwag (Zellers et al., 2019)
142 ConversationalQA	CoQA (Reddy et al., 2019)
143 MathQA	GSM8K (Cobbe et al., 2021)
144 English translation	N/A
145 French translation	DiaBLa (Bawden et al., 2021)
146 Arabic translation	Altammami et al. (2020)
147 Japanese translation	Oda (2016)
148 Spanish translation	"European Commission and Technology." (2017)
149 Bengali translation	Islam et al. (2022)
150 Tamil translation	Ramasamy et al. (2012)
151 Gujarati translation	Shah and Bakrola (2019)

A.2 Category Description

We provide the task category description in this subsection.

Chinese Culture (18). Focuses on aspects unique to Chinese history, society, and language, therefore testing the model's understanding of cultural nuances.

Classification (21). Addresses classification tasks, such as determining correctness or whether something belongs to a specific category.

Code (5). Tests the model's proficiency in understanding and generating computer code across various programming languages.

Commonsense (36). Evaluates the model's grasp of general knowledge and everyday reasoning that humans consider obvious.

Creative Natural Language Generation (NLG) (21). Measures the model's ability to produce imaginative and novel text outputs, ranging from stories to creative descriptions.

Evaluation (5). Focuses on assessing other models or systems, therefore testing the ability to judge and provide feedback on performance.

Grammar (10). Assesses the model's understanding of linguistic rules and its ability to apply them correctly in text generation.

Linguistic (24). Involves tasks that test the model's understanding of language structure, including syntax, semantics, and morphology.

Motion Detection (6). Uncommon in LLMs, this refers to tasks related to interpreting descriptions of motion or predicting outcomes based on textual motion descriptions.

Named Entity Recognition (NER) (12). Involves identifying and categorizing key information (e.g., names, places, dates) within the text.

Natural Language Inference (NLI) (30). Tests the model's ability to understand relationships between sentences, such as contradiction, entailment, and neutrality.

Question Answering (QA) (19). Evaluates the model's ability to understand and respond to questions with accurate and relevant answers.

Reasoning (29). Involves tasks that require logical thinking, problem-solving, and deduction to arrive at correct conclusions.

Role Playing (2). Tests the model's ability to adopt personas or roles in conversational contexts, assessing its versatility in generating context-appropriate responses.

Sentiment (8). Evaluates the model's ability to detect and interpret emotional tones in text, such as positive, negative, or neutral sentiments.

Structured Data (16). Involves interpreting and generating responses based on structured information such as tables, charts, and databases.

Style Transfer (20). Tests the model's ability to convert text from one stylistic or tonal form to another while retaining the original content's meaning.

Summarization (9). Assesses the model's ability to condense longer texts into shorter, coherent summaries capturing the essential points.

Toxicity (3). Focuses on identifying and mitigating harmful or stereotypical content in text generation.

Translation (9). Evaluates the model's ability to accurately translate text between languages, testing its linguistic versatility and understanding.

B CIF-Bench Results in Public Split

We provide the category-based results in *public* split here in Table 13.

Table 13: Overall Results in CIF-Bench *Public* Split with Single Instruction. The first column is the average score across *all* the tasks, and the rest columns are average scores grouped by task categories. The cells are highlighted with fading colors from maximum to minimum in a column.

Model Name	**Overall**	**Chinese Culture**	**Classification**	**Code**	**Commonsense**	**Creative NLG**	**Evaluation**	**Grammar**	**Linguistic**	**Motion Detection**	**NER**
Qwen-72B-Chat	0.589	0.512	0.716	0.444	0.706	0.587	0.661	0.424	0.521	0.694	0.515
Qwen-14B-Chat	0.564	0.481	0.678	0.416	0.657	0.567	0.669	0.396	0.485	0.663	0.486
Deepseek-LLM-67B-Chat	0.526	0.477	0.617	0.364	0.609	0.559	0.573	0.374	0.458	0.631	0.493
GPT-3.5-Public-SFT	0.522	0.316	0.611	0.492	0.578	0.538	0.639	0.377	0.447	0.580	0.492
Yi-34B-Chat	0.516	0.452	0.607	0.437	0.624	0.516	0.545	0.254	0.382	0.671	0.398
Baichuan2-13B-Chat	0.512	0.446	0.623	0.403	0.600	0.505	0.582	0.352	0.423	0.633	0.435
Tigerbot-13B-Chat	0.494	0.350	0.558	0.447	0.599	0.528	0.707	0.352	0.447	0.551	0.498
Chinese-Alpaca-2-13B	0.492	0.260	0.572	0.434	0.533	0.562	0.574	0.318	0.417	0.624	0.467
Chinese-Alpaca-33B	0.484	0.274	0.546	0.470	0.527	0.540	0.703	0.332	0.382	0.582	0.464
Ziya-Llama-13B	0.479	0.287	0.550	0.422	0.523	0.551	0.650	0.294	0.384	0.610	0.437
Chinese-Llama2-Linly-13B	0.479	0.286	0.623	0.439	0.549	0.535	0.626	0.286	0.403	0.587	0.468
Tigerbot-7B-Chat	0.478	0.354	0.528	0.440	0.570	0.540	0.708	0.314	0.430	0.528	0.413
ChatGLM3-6B	0.472	0.321	0.488	0.436	0.527	0.503	0.588	0.290	0.328	0.574	0.415
Chinese-Alpaca-13B	0.471	0.264	0.553	0.443	0.495	0.525	0.587	0.334	0.394	0.653	0.457
ChatGLM2-6B	0.464	0.334	0.532	0.436	0.522	0.527	0.651	0.314	0.395	0.536	0.402
Chinese-Alpaca-7B	0.452	0.237	0.536	0.438	0.484	0.502	0.672	0.318	0.389	0.652	0.394
Chinese-Alpaca-2-7B	0.448	0.251	0.472	0.435	0.480	0.532	0.577	0.268	0.348	0.596	0.431
Chinese-Llama2-Linly-7B	0.443	0.264	0.558	0.419	0.497	0.522	0.664	0.236	0.381	0.593	0.381
Qwen-7B-Chat	0.442	0.313	0.549	0.404	0.520	0.515	0.646	0.244	0.411	0.570	0.368
ChatGLM-6B	0.440	0.311	0.499	0.446	0.484	0.548	0.558	0.278	0.382	0.484	0.386
Baichuan-13B-Chat	0.426	0.355	0.416	0.361	0.516	0.416	0.564	0.324	0.374	0.380	0.394
Yi-6B-Chat	0.420	0.320	0.439	0.395	0.489	0.449	0.493	0.230	0.293	0.587	0.341
CPM-Bee-10B	0.415	0.382	0.455	0.284	0.431	0.508	0.300	0.317	0.367	0.494	0.397
Moss-Moon-003-SFT	0.399	0.233	0.465	0.389	0.427	0.482	0.509	0.274	0.369	0.526	0.385
Belle-SFT-Public	0.397	0.196	0.503	0.376	0.426	0.472	0.543	0.269	0.371	0.512	0.356
Telechat-7B	0.350	0.172	0.299	0.438	0.386	0.456	0.400	0.138	0.202	0.412	0.322
Aquilachat-7B	0.350	0.203	0.270	0.357	0.404	0.449	0.394	0.090	0.260	0.348	0.322
Baichuan2-7B-Chat	0.339	0.345	0.595	0.154	0.455	0.327	0.523	0.362	0.354	0.466	0.233

Model Name	**Overall**	**NLI**	**QA**	**Reasoning**	**Role Playing**	**Sentiment**	**Structured Data**	**Style Transfer**	**Summarization**	**Toxic**	**Translation**
Qwen-72B-Chat	0.589	0.695	0.668	0.539	0.752	0.637	0.505	0.587	0.609	0.671	0.466
Qwen-14B-Chat	0.564	0.647	0.609	0.498	0.757	0.638	0.460	0.610	0.629	0.691	0.467
Deepseek-LLM-67B-Chat	0.526	0.588	0.624	0.444	0.694	0.592	0.384	0.576	0.594	0.666	0.439
GPT-3.5-Public-SFT	0.522	0.587	0.565	0.498	0.745	0.583	0.444	0.501	0.620	0.643	0.452
Yi-34B-Chat	0.516	0.631	0.592	0.460	0.761	0.566	0.440	0.551	0.610	0.608	0.408
Baichuan2-13B-Chat	0.512	0.600	0.591	0.474	0.751	0.597	0.434	0.525	0.572	0.494	0.372
Tigerbot-13B-Chat	0.494	0.571	0.569	0.413	0.732	0.560	0.365	0.502	0.607	0.601	0.306
Chinese-Alpaca-2-13B	0.492	0.566	0.545	0.420	0.712	0.595	0.382	0.488	0.641	0.740	0.347
Chinese-Alpaca-33B	0.484	0.550	0.506	0.423	0.732	0.548	0.342	0.494	0.629	0.648	0.334
Ziya-Llama-13B	0.479	0.546	0.499	0.404	0.749	0.582	0.367	0.499	0.629	0.722	0.313
Chinese-Llama2-Linly-13B	0.479	0.563	0.524	0.411	0.676	0.561	0.359	0.482	0.602	0.696	0.313
Tigerbot-7B-Chat	0.478	0.532	0.554	0.393	0.731	0.583	0.351	0.519	0.630	0.614	0.291
ChatGLM3-6B	0.472	0.557	0.526	0.397	0.749	0.612	0.431	0.529	0.620	0.589	0.392
Chinese-Alpaca-13B	0.471	0.524	0.513	0.402	0.726	0.526	0.323	0.486	0.628	0.702	0.336
ChatGLM2-6B	0.464	0.520	0.533	0.407	0.725	0.506	0.363	0.480	0.627	0.661	0.303
Chinese-Alpaca-7B	0.452	0.504	0.501	0.351	0.699	0.543	0.365	0.478	0.623	0.711	0.328
Chinese-Alpaca-2-7B	0.448	0.509	0.493	0.344	0.703	0.510	0.334	0.483	0.637	0.596	0.343
Chinese-Llama2-Linly-7B	0.443	0.496	0.546	0.350	0.713	0.559	0.323	0.495	0.603	0.584	0.293
Qwen-7B-Chat	0.442	0.489	0.514	0.384	0.713	0.563	0.328	0.463	0.576	0.639	0.281
ChatGLM-6B	0.440	0.480	0.483	0.353	0.738	0.460	0.346	0.480	0.633	0.543	0.322
Baichuan-13B-Chat	0.426	0.531	0.584	0.339	0.668	0.478	0.402	0.459	0.559	0.497	0.392
Yi-6B-Chat	0.420	0.496	0.516	0.344	0.742	0.488	0.348	0.498	0.627	0.510	0.285
CPM-Bee-10B	0.415	0.451	0.472	0.304	0.647	0.329	0.284	0.538	0.534	0.486	0.305
Moss-Moon-003-SFT	0.399	0.403	0.457	0.325	0.712	0.450	0.304	0.435	0.594	0.542	0.308
Belle-SFT-Public	0.397	0.450	0.430	0.338	0.645	0.426	0.300	0.398	0.558	0.683	0.224
Telechat-7B	0.350	0.375	0.414	0.261	0.660	0.341	0.320	0.462	0.639	0.494	0.304
Aquilachat-7B	0.350	0.385	0.426	0.274	0.595	0.308	0.267	0.434	0.607	0.409	0.355
Baichuan2-7B-Chat	0.339	0.414	0.349	0.339	0.673	0.429	0.300	0.246	0.097	0.357	0.130