

FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models

Yuxin Jiang^{1,2*}, Yufei Wang³, Xingshan Zeng³, Wanjun Zhong³, Liangyou Li³,
Fei Mi³, Lifeng Shang³, Xin Jiang³, Qun Liu³, Wei Wang^{1,2}

The Hong Kong University of Science and Technology (Guangzhou)¹,
The Hong Kong University of Science and Technology², Huawei Noah's Ark Lab³
yjjiangcm@connect.ust.hk, wang.yufei1@huawei.com, weiwcs@ust.hk

Abstract

The ability to follow instructions is crucial for Large Language Models (LLMs) to handle various real-world applications. Existing benchmarks primarily focus on evaluating pure response quality, rather than assessing whether the response follows constraints stated in the instruction. To fill this research gap, in this paper, we propose FollowBench, a **Multi-level Fine-grained Constraints Following Benchmark** for LLMs. FollowBench comprehensively includes five different types (i.e., Content, Situation, Style, Format, and Example) of fine-grained constraints. To enable a precise constraint following estimation on diverse difficulties, we introduce a *Multi-level* mechanism that incrementally adds a single constraint to the initial instruction at each increased level. To assess whether LLMs' outputs have satisfied every individual constraint, we propose to prompt strong LLMs with constraint-evolution paths to handle challenging open-ended instructions. By evaluating 13 closed-source and open-source popular LLMs on FollowBench, we highlight the weaknesses of LLMs in instruction following and point towards potential avenues for future work. The data and code are publicly available at <https://github.com/YJiangcm/FollowBench>.

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2022) pre-trained on web-scale corpora have showcased proficiency in generating fluent and realistic text. Yet, human instructions in real-life cases require the model to generate text that not only possesses a high degree of naturalness but adheres to specific constraints (Yang et al., 2023). For instance, the model may be required to recommend ten books that are specifically written in Chinese (Figure 1), or it might be expected to generate responses that have a certain tone.

*Work done during the internship at Huawei Noah's Ark Lab.

L5 + Example	I am interested in Tang Dynasty. In Shakespeare's tone, recommend me ten relevant Chinese books. Use bullet point in your answer. <i>Please response based on the examples: ...</i>	✗	✗
L4 + Format	I am interested in Tang Dynasty. In Shakespeare's tone, recommend me ten relevant Chinese books. <i>Use bullet point in your answer.</i>	✓	✗
L3 + Style	I am interested in Tang Dynasty. <i>In Shakespeare's tone,</i> recommend me ten relevant Chinese books.	✓	✓
L2 + Situation	<i>I am interested in Tang Dynasty.</i> Recommend me ten relevant Chinese books.	✓	✓
L1 + Content	Recommend me ten <i>Chinese</i> books.	✓	✓
	Recommend me ten books.		

Figure 1: FollowBench covers five *fine-grained* constraint categories and is constructed based on the *Multi-level* mechanism, which increasingly adds a single constraint to straightforward instructions. On the right, the model that can follow instructions with more constraints is deemed to possess better instruction-following ability.

The dominant paradigm for assessing if a model can follow instructions involves using human annotators or strongly aligned LLMs to judge its response quality, in terms of helpfulness, relevance, accuracy, depth, creativity, and level of detail (Wang et al., 2023a; Li et al., 2023; Zheng et al., 2023; Xu et al., 2023). However, prior work still has two limitations. Firstly, they ignore the **fine-grained constraints** inside instructions, which are essential and objective standards for evaluating the instruction-following capability. While several benchmarks have rigorously explored individual constraint types, including semantic restrictions (Chen et al., 2022) and complex formatting (Tang et al., 2023), there exists a lack of comprehensive analysis across the diverse spectrum of constraint categories. Secondly, few benchmarks consider the varying difficulty of instructions, which is controlled by the number of imposed constraints. This makes it challenging to precisely assess the degree to which LLMs can follow instructions. Towards this end, our research question is: *how can we systemically and precisely eval-*

uate the instruction-following capability of LLMs?

In this paper, we construct FollowBench, a **Multi-level Fine-grained Constraints Following Benchmark**. FollowBench comprehensively includes five different types of constraints from real-world scenarios, namely Content (i.e., explicit restrictions on the response content), Situation (i.e., specific situation/background information added to the question), Style (i.e., response style requirements), Format (i.e., response format requirements), and Example (i.e., example pattern recognition and following). To precisely estimate the difficulty degree to which LLMs can follow instructions, as shown in Figure 1, we propose a novel *Multi-level* mechanism that incrementally adds a single constraint to straightforward instructions at each increased level. The multi-level mechanism enables us to pinpoint the difficulty level at which LLMs fail to follow instructions, thereby estimating the upper limit of instruction-following capability in LLMs more precisely. Overall, FollowBench consists of 820 meticulously curated instructions from over 50 NLP tasks, including both closed- and open-ended questions. For evaluation purposes, we propose a hybrid evaluation method comprising rule-based and model-based solutions. Given LLMs’ outputs, both solutions judge whether the outputs satisfy each of the constraints in the instructions. The rule-based solutions focus on closed-ended instructions while the model-based solutions are applied to opened-ended instructions. For model-based solutions, instead of merely using current instructions and responses as input, we additionally provide the evolution process of the instructions in the input prompts to LLM judges to better understand each individual constraint. Both the data construction and the evaluation undergo human verification.

In our experiments, we propose three metrics to assess the instruction-following ability of 13 prominent closed-source and open-source LLMs on FollowBench. Our principal observations are: (1) the performance of all tested models declines substantially with an increase in difficulty level (the number of constraints in an instruction); (2) although closed-source models such as GPT-4 and GPT-3.5 **only** consecutively satisfy around three constraints on average, they still markedly surpass all open-source models; (3) certain specific constraint categories, such as Situation and Example, prove to be more challenging for LLMs than

others; (4) beyond capabilities such as knowledge and reasoning, instruction following can offer an additional lens for comprehensively assessing the proficiency of LLMs.

2 Related Work

2.1 Instruction-Following Language Models

Prior research has found that LLMs fine-tuned with annotated “instructional” data, which is composed of language instructional commands and their desired outcomes, can be effective at following general language instructions (Weller et al., 2020; Sanh et al., 2021; Mishra et al., 2022; Jiang et al., 2024). To enhance the understanding of LLMs regarding the intricate and varied intentions of users in real-world scenarios, works like ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) implement instruction tuning across a wide range of human-crafted instructions and task categories. Recent studies (Zheng et al., 2023; Xu et al., 2023; Jiang et al., 2023) have pivoted towards automatically generating high-quality data to enhance the instruction-following capability of LLMs, addressing the challenges posed by labor-intensive human annotation.

2.2 Evaluation for Instruction Following

There are several research efforts in evaluating LLMs’ following capability towards particular tasks. Tang et al. (2023) focuses on evaluating LLMs’ generation capability towards complex structured tabular data in text, HTML, and LaTeX. They first collect tables from existing NLP benchmarks and websites, then construct guiding instructions based on these data. Chen et al. (2022) evaluates whether LLMs can follow particular knowledge-intensive generation instructions. They first provide a list of examples (e.g., a list of sports stars in the UK), followed by a constraint that is contradicted by the examples (e.g., not mentioning any athletes). These benchmarks can only demonstrate particular types of instruction-following capability of LLMs. In contrast, FollowBench comprehensively includes instructions with five different types of fine-grained constraints in multi-level difficulty and FollowBench should provide a well-rounded and precise estimation of instruction-following capability for existing LLMs. For more details on LLM evaluation, we refer to recent surveys (Chang et al., 2023; Wang et al., 2023b).



















Constraint	Task	Avg Len	#Data	Evaluation
Content	Data-to-Text Generation	84	25	
	Document-Level Event Argument Extraction	696	25	
	Document-Level Named Entity Recognition	376	25	
	Text Generation with Language Constraints	88	25	
	Open-ended Question Answering	56	25	
Situation	Suggestion Generation	69	40	
	Role-playing	111	15	
	Complex Situation Reasoning	102	55	
Style	Open-ended Question Answering	64	150	
Format	Text-to-Table Generation	171	30	
	Open-ended Question Answering	74	120	
Example	40 diverse NLP tasks	739	200	
Mixed	Text Editing	96	25	
	Summarization	254	25	
	Machine Translation	91	25	
	Story Generation	34	10	

Table 1: An overview of FollowBench. “Avg Len” is the average word number of instructions.  refers to rule-based evaluation, while  refers to model-based evaluation.

3 FollowBench

As shown in Table 1, FollowBench encompasses five distinct *fine-grained* constraint categories: Content, Situation, Style, Format, and Example. Each category consists of instructions from various NLP tasks. Different from previous benchmarks, we introduce a *Multi-level* mechanism that incrementally adds constraints to an initial instruction (see examples in Figure 2), producing a set of instructions ranging from 1 to 5 constraints. In the following part of this paper, we use “level n ” to denote an instruction containing n constraints. It is worth noticing that the way of adding constraints is meticulously designed for each task within its respective constraint category. The multi-level mechanism enables us to pinpoint the difficulty level at which LLMs fail to follow instructions, thereby estimating the upper bound of instruction-following capability in LLMs more precisely.

To encapsulate, we will introduce the data construction process of FollowBench, including *fine-grained* constraints and the *Multi-level* mechanism, in §3.1. In §3.2, we propose an evaluation protocol with three metrics that seamlessly integrate with the multi-level mechanism.

3.1 Data Construction

Content Constraints Content constraints refer to *explicit* impositions of specific conditions that shape the depth or scope of the response content. An example is shown in Figure 2, which sets spe-

cific criteria for the retrieved object. Ensuring that LLMs adhere to content constraints has become a critical challenge in Controlled Text Generation (Zhang et al., 2022), as it demands models to understand specific guidelines and adapt responses to prescribed conditions (Chen et al., 2022). To this end, we first collect data from the following tasks: (1) Complex Information Extraction aims at retrieving specific information about specific objects from the given text; (2) Text Generation with Language Constraints requires to generate fluent on-topic content while respecting a specified constraint; (3) Open-ended Question Answering comes from real scenarios (e.g., open-source platforms) to prevent the risk of data leakage. Subsequently, we construct multi-level instructions by adding one content constraint to the collected instructions each time. The manners of introducing additional constraints depend on different tasks (see details in Appendix A.1). For Complex Information Extraction, we gradually narrow down the scope of the information to be extracted. For Text Generation with Language Constraints, we incorporate additional restrictions from WordNet (Miller, 1992) and Wikidata (Vrandečić and Krötzsch, 2014). For Open-ended Question Answering, we utilize advanced LLMs like GPT-4 to generate a new instruction with one more constraint based on the given instruction. While the output from the LLMs serves primarily as a reference, we handpick the most relevant and challenging synthesized instructions to ensure data quality.

CONTENT	INITIAL	Recommend 5 films to me.
	LEVEL 1	Recommend me 5 Chinese films.
	LEVEL 2	Recommend me 5 Chinese films released before 1990 .
SITUATION	INITIAL	How can I increase my productivity while working from home?
	LEVEL 1	Since the pandemic began, I've been working remotely. How can I increase my productivity while working from home?
	LEVEL 2	I have a small child at home. Since the pandemic began, I've been working remotely. How can I increase my productivity while working from home?
STYLE	INITIAL	How did US states get their names?
	LEVEL 1	How did US states get their names? Please respond in the writing style of Shakespeare.
	LEVEL 2	How did US states get their names? Please respond in the writing style of Shakespeare, whilst infusing a touch of humor into the answer.
FORMAT	INITIAL	Why can I see the moon during the day?
	LEVEL 1	Why can I see the moon during the day? Answer in a table format with columns "Reason" and "Explanation".
	LEVEL 2	Why can I see the moon during the day? Answer in a table format with columns "Reason" and "Explanation". Each explanation should not exceed 20 words in length.
EXAMPLE	LEVEL 1	question_template_1.format(example_1) + answer_template_1.format(example_1) question_template_1.format(example_2) + answer_template_1.format(example_2) ⋮ question_template_1.format(query)
	LEVEL 2	question_template_1.format(example_1) + answer_template_1.format(example_1) question_template_2.format(example_2) + answer_template_2.format(example_2) ⋮ question_template_1.format(query)

Figure 2: FollowBench covers five *fine-grained* categories of constraints. Within each constraint type, we construct a range of *Multi-level* instructions by incrementally adding constraints (highlighted in red). There are five levels in total; however, we only display the first two levels from each category for demonstration purposes.

Situation Constraints Situation Constraints refer to impositions of specific situations or backgrounds that *implicitly* guide the appropriate answer of the response. For instance, it is necessary to illustrate the situation when asking for customized suggestions, as shown in Figure 2. Another example is to customize LLMs to simulate various characters under certain circumstances, namely Role-playing, which provides a more nuanced interaction for users (Shanahan et al., 2023; Wang et al., 2023c). Situation constraints push LLMs beyond mere factual retrieval or surface-level synthesis, demanding a nuanced understanding, a dynamic adaptation, and complicated reasoning to the situation (Yao et al., 2022; Liu et al., 2023). Besides real-life questions, we also consider Complex Situation Reasoning tasks including Math Word Problems, Time/Spatial Reasoning, and Code Generation. These tasks all require interpreting and solving problems within a given situation, thus matching the definition of situation constraints. We first collect initial instructions from these sources and then manually curate multi-level instructions by incrementally supplementing situation information inside (see Appendix A.2).

Style Constraints Style Constraints control the stylistic variations of output to accomplish specific

stylistic goals, such as tone, sentiment, formality, and empathy (Tsai et al., 2021), as illustrated in Figure 2. The challenges of style constraints for LLMs are the intricate understanding and adaptation of language nuances, ensuring contextually appropriate and stylistically consistent outputs (Smith et al., 2020; Cheng and Li, 2022). Drawing from Open-ended Question Answering datasets and on-line platforms, we collect initial instructions and then leverage LLMs’ in-context learning capability to craft instructions with multi-level style constraints. The prompt template can be viewed in Figure 8. Human experts subsequently review and refine the outputs produced by LLMs.

Format Constraints Format Constraints refer to stipulations governing the structural, linguistic, or output presentation of generated content. An example is shown in Figure 2, which sets limits on word length and requires the format of the response to be a table. Format constraints necessitate a deep, nuanced understanding of language and structure, allowing them to flexibly adapt outputs according to diverse and often intricate specifications (Zhao et al., 2023). Recent work has pointed out that even the most superior LLMs may struggle with tasks that require generating complex, structured outputs such as tables, JSON, HTML, or LaTeX (Tang

et al., 2023). To include a variety of format constraints, we first collect instructions from broader domains, encompassing Text-to-Table Generation and Open-ended Question Answering, then we utilize powerful LLMs to sequentially add format constraints ranging from length and hierarchy to specialized linguistic features and output mediums. See Figure 9 for the prompt template. Finally, we ask human experts to carefully check and refine the synthesized instructions.

Example Constraints LLMs have demonstrated stunning few-shot learning ability (Brown et al., 2020), which enables them to adapt quickly to a new query by recognizing patterns from just a few examples provided in the prompt. However, the robustness of few-shot learning, which means whether LLMs can still follow correct patterns after introducing “noise” examples, has not been explored. Thus, we propose a novel constraint category named Example Constraints to evaluate the example pattern recognition and following capability of LLMs. We automatically craft instructions with multi-level example constraints based on Prompt-Source (Bach et al., 2022), where instructions at level n have $n - 1$ noise examples in the input. The details are illustrated in Appendix A.3.

Mixed Constraints For the above five constraint categories, we construct multi-level instructions by adding the same type of constraint sequentially. Nevertheless, real-world scenarios often require more than one type of constraint to be enforced in a singular instruction. Therefore, we define Mixed Constraints as the composition of varied constraint categories. For instance, in the Text Editing task, we may want to add some content as well as adjust the output format. Besides, we also consider several tasks that are naturally suitable for constructing mixed constraints, including Summarization, Machine Translation, and Story Generation (see Appendix A.4). Instructions with multi-level mixed constraints are produced by specifying the format of generating answers (Format Constraints), requiring the generated text to include or not include certain keywords (Content Constraints), etc.

Data Quality Control To ensure the data quality of FollowBench, we implement a dual-layer verification system for each instruction. Two annotators independently evaluate: (1) the appropriateness of the instruction for its designated constraint category, and (2) the validity of the added

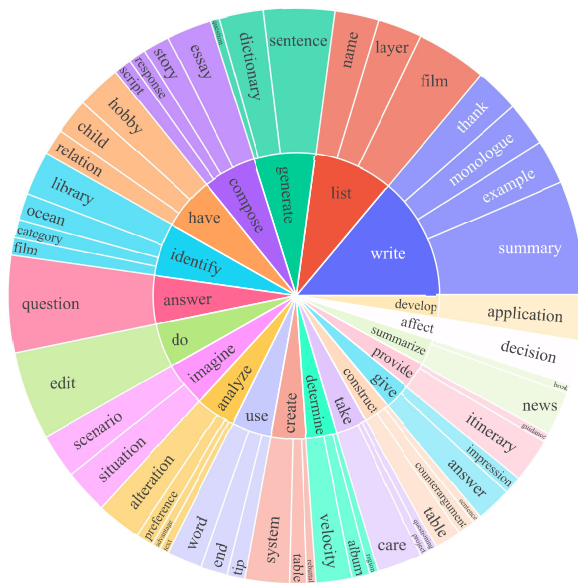


Figure 3: Verb-noun structure of FollowBench Instructions.

constraint within the instruction. In instances of divergent evaluations, a third annotator intervenes for a detailed review to ensure consensus.

We analyze the comprehensiveness and diversity of in FollowBench, which includes 820 instructions in total. To maintain data diversity, we strive to ensure that the ROUGE-L score between any two initial instructions is below 0.7. Figure 3 shows the verb-noun structure of FollowBench instructions, where the top 20 verbs (inner circle) and their top 4 direct noun objects (outer circle) are depicted.

3.2 Evaluation Protocol

Given that nearly half of instructions in FollowBench are open-ended without reference answers, devising a rule-based program to assess the outputs is extremely challenging. To overcome this, inspired by (Gilardi et al., 2023; Huang et al., 2023), we propose to develop a model-based approach by using strong LLMs¹ as judges. Previous works leverage strong LLMs to determine the quality of a response, by prompting them to consider multiple factors such as usefulness, relevance, and level of detail (Li et al., 2023; Zheng et al., 2023). To effectively guide strong LLMs to judge the constraint following capability objectively and faithfully, we propose a *Multi-level-aware* prompt template, as shown in Figure 4. Rather than merely presenting the instruction and

¹We use GPT-4-Preview-1106 in our experiments.

```

Prompt Template for Model-based Evaluation

Given an initial instruction, we add one {constraint_type} constraint per
time and obtain the final instruction with {level_n} additional constraints.

#Initial Instruction#
{initial_instruction}

#Initial Instruction + 1 constraint#
{level_1_instruction}
⋮
#Initial Instruction + {level_n} constraints#
{level_n_instruction}

#Answer of Initial Instruction + {level_n} constraints#
{answer_of_level_n_instruction}

#System#
1) Please identify all {level_n} added constraints.
2) For the {level_n} added constraints, discriminate if the #Answer of
Initial Instruction + {level_n} constraints# satisfies each constraint.
3) In the final line, only output a Python LIST with {level_n} elements
('YES' or 'NO') indicating whether the answer satisfies each constraint.

```

Figure 4: Prompt template for model-based evaluation.

asking LLMs to determine whether all constraints are satisfied, we illustrate the evolution process of the instruction and prompt LLMs to pinpoint the newly added constraint at each level. Exposing the evolution process of the instruction allows for a more granular understanding and identification of individual constraints, enhancing LLMs’ ability to discriminate with precision. The ablation study in §5.1 validates the effectiveness of this strategy.

Moreover, we propose three novel metrics to evaluate the instruction-following ability of LLMs. For an instruction with n constraints (level n), we use the rule-based program or LLM judge (refer to Table 1) to discriminate if the response of a model satisfies each constraint in the instruction. At each level n , given a set of m instructions, we define the Hard Satisfaction Rate (HSR) and Soft Satisfaction Rate (SSR) as follows:

$$\text{HSR} = \frac{1}{m} \sum_{i=1}^m \prod_{j=1}^n s_i^j \quad (1)$$

$$\text{SSR} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n s_i^j \quad (2)$$

where $s_i^j = 1$ if the j -th constraint of i -th instruction is satisfied and $s_i^j = 0$ otherwise. HSR measures the average rate at which all constraints of individual instructions are fully satisfied, while SSR calculates the average satisfaction rate of individual constraints across all instructions.

As described in §3, we construct FollowBench by incrementally adding five constraints to an initial instruction, enabling

us to pinpoint the difficulty level at which LLMs fail to follow instructions. Therefore, we propose a metric called Consistent Satisfaction Levels (CSL) to estimate how many consecutive levels a model can satisfy, beginning from level 1:

$$\text{CSL} = \frac{1}{g} \sum_{i=1}^g \arg \max_l \left(l \times \prod_{n=1}^l S_i^n \right) \quad (3)$$

where g is the group number of initial instructions, $S_i^n = 1$ if all constraints of the i -th instruction at level- n are satisfied and $S_i^n = 0$ otherwise.

4 Experiments

This section first introduces experimental setup in §4.1, and then presents the main experiment results across two key dimensions: difficulty level in §4.2 and constraint category in §4.3.

4.1 Experimental Setup

We evaluate 13 popular LLMs including GPT-4-Preview-1106 (OpenAI, 2023), GPT-3.5-Turbo-1106 (OpenAI, 2022), Qwen-Chat-72B/14B/7B (Bai et al., 2023), LLaMA2-Chat-70B/13B/7B (Touvron et al., 2023), WizardLM-13B-V1.2 (Xu et al., 2023), Vicuna-13B/7B-V1.5 (Zheng et al., 2023), Baichuan2-Chat-7B (Baichuan, 2023), and ChatGLM3-6B (Du et al., 2022). We access GPT-4-Preview-1106 and GPT-3.5-Turbo-1106 via OpenAI API. We access other open-source LLMs from their official repositories. During the inference process, we set the temperature to 0 to ensure deterministic outputs. We set the maximum generation length to 2048. Other parameters use their default values. To facilitate the multilingual evaluation of LLM’s instruction-following ability, we also craft a Chinese version of FollowBench, namely FollowBench-zh, in Appendix D.

4.2 Level-categorized Results

Table 2 provides a comprehensive comparison of various models across five difficulty levels, denoted as L1 to L5. The detailed results for each constraint category are listed in Appendix B. From a bird’s-eye view, we can infer that the performance typically diminishes as we progress from L1 to L5 for almost all models. This trend coincides with the increasing complexity or stringent requirements associated with higher levels. Besides, models with larger architectures generally outperform their smaller counterparts. However, it’s worth noting

Model	HSR (%)						SSR (%)						CSL
	L1	L2	L3	L4	L5	Avg.	L1	L2	L3	L4	L5	Avg.	
GPT-4-Preview-1106	84.7	75.6	70.8	73.9	61.9	73.4	84.7	77.0	75.3	77.0	72.3	77.2	3.3
GPT-3.5-Turbo-1106	80.3	68.0	68.6	61.1	53.2	66.2	80.3	71.2	74.2	69.6	67.1	72.5	2.9
Qwen-Chat-72B	73.8	63.3	54.3	45.2	39.9	55.3	73.8	67.5	63.2	57.6	56.0	63.6	2.4
LLaMA2-Chat-70B	59.9	53.3	46.0	40.2	37.9	47.5	59.9	57.3	55.7	53.3	53.2	55.9	2.1
Qwen-Chat-14B	62.8	56.2	47.7	38.7	30.9	47.3	62.8	61.9	57.7	52.6	51.4	57.3	1.9
WizardLM-13B-V1.2	68.8	64.1	53.1	40.8	35.8	52.5	68.8	65.7	61.8	53.4	53.9	60.7	2.2
LLaMA2-Chat-13B	57.0	56.0	50.4	44.4	38.1	49.2	57.0	60.0	58.0	54.8	52.2	56.4	2.2
Vicuna-13B-V1.5	71.2	60.2	49.6	40.6	34.0	51.1	71.2	64.8	59.9	54.5	53.6	60.8	2.1
Qwen-Chat-7B	55.9	51.7	38.7	33.1	23.3	40.6	55.9	58.2	51.6	48.9	45.9	52.1	1.5
LLaMA2-Chat-7B	58.0	51.3	47.4	39.5	35.3	46.3	58.0	56.5	55.6	52.5	51.4	54.8	1.9
Vicuna-7B-V1.5	60.8	52.0	42.2	33.3	23.9	42.4	60.8	58.6	55.5	48.3	49.0	54.4	1.7
Baichuan2-Chat-7B	58.3	46.1	40.7	30.4	25.5	40.2	58.3	55.4	54.9	49.9	49.3	53.6	1.4
ChatGLM3-6B	60.9	46.6	36.7	27.8	21.4	38.7	60.9	55.3	51.2	47.9	45.0	52.0	1.6

Table 2: Results across five difficulty levels. For each level, we compute the average score of all constraint categories. Proprietary LLMs, open-sourced LLMs (large), open-sourced LLMs (medium), and open-sourced LLMs (small) are distinguished by different colors.

that the scaling law does not apply as effectively to LLaMA2-Chat-70B. It can be observed from Appendix B that while LLaMA-2-Chat-70B does indeed outperform LLaMA-2-Chat-13B in Situation constraints, it shows a relative underperformance in Format and Mixed Constraints categories. More importantly, there’s a marked performance gap between closed-source models (i.e., GPT-4 and GPT-3.5) and open-source models. Regarding CSL, it can be deduced that the instruction-following upper bound for GPT-4 and GPT-3.5 is approximately 3 constraints (level 3) added to an initial instruction. In contrast, open-source models typically have an upper limit of about 2 constraints (level 2). This significant difference underscores the better instruction-following ability of proprietary models, possibly due to superior data quality or optimization strategies such as RLHF (Ouyang et al., 2022). Furthermore, **even the most sophisticated models are limited to following instructions with about three constraints**, suggesting significant potential for further improvement.

4.3 Constraint-categorized Results

As depicted in Figure 5, we assess various models over different constraint categories to succinctly showcase the instruction-following capability of LLMs in a singular dimension. Notably, GPT-4 and GPT-3.5 surpass open-source models in every constraint category, with a pronounced advantage in Content, Situation, Example, and Mixed constraints. Furthermore, most models demonstrated commendable proficiency under the Style constraint. While GPT-4, GPT-3.5, and LLaMA2-

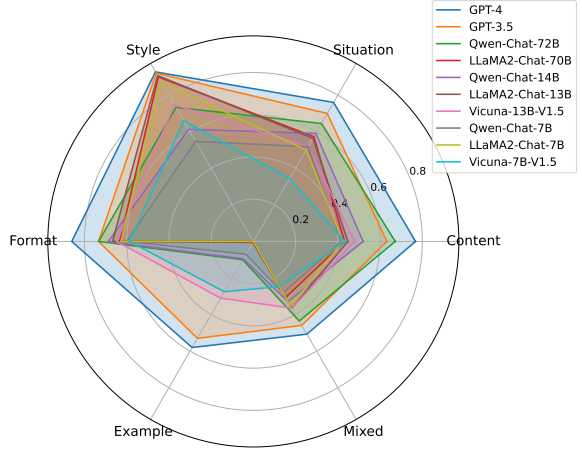


Figure 5: HSR (%) results in diverse constraint categories. For each category, we compute the average score of all difficulty levels.

Chat-70B were the frontrunners, the trend suggests that style adaptation is an area where many models excel, hinting at its utility in real-world applications. However, the Example and Mixed constraints posed a challenge to most models. While GPT-4 led the segment, even its scores were noticeably lower than in other categories. To illustrate, in the “Example” category, we evaluated the instruction-following capabilities of LLMs by introducing “noise examples” with varying natural language templates. The observed performance decline is primarily due to the LLMs’ limited training in processing such noisy inputs within context-based learning scenarios. Typically, LLMs are fine-tuned on clean and uniform datasets, which do not adequately prepare them to sift through and ignore

irrelevant or misleading information. This limitation becomes apparent when faced with the intricacies of real-world data. Our findings underscore the complexity of these constraints and pinpoint an area for potential improvement.

5 Analysis

This section includes: an ablation study confirming our prompt template’s effectiveness for model-based evaluation (§5.1); a comparison of instruction following vs. other LLM’s abilities (§5.2); an examination of failure consistency (§5.3); and an investigation of various decoding strategies (§5.4). In addition, a case study is presented in Appendix C for further analysis.

5.1 Ablation Study of Model-based Evaluation

We randomly sample 100 cases that require LLM evaluation, encompassing five constraints, five distinct levels, and four diverse models to guarantee comprehensive representation. Then we ask three expert-level human labelers to assess whether the model’s response satisfies all the constraints in each case and use the majority voting as the final human annotations. As shown in Table 3, our prompt template (Figure 4) registers an impressive 88% agreement with expert human evaluations, surpassing even the internal agreement among human experts, which stands at 85%. Remarkably, when the evolution process of multi-level constraints is removed from our prompt template, the agreement rate dips by 9%. This underlines the instrumental role played by the detailed portrayal of the instruction’s evolution in enhancing LLM’s precision in discernment. In contrast, we also employ the prompt template from Vicuna (Zheng et al., 2023), a standard prompt for assessing the overall quality of response. This template prompts the LLM to assign a score from 0 to 10 for each response. We consider responses with a score above 5.0 to meet all the constraints of an instruction. This approach achieves 67% agreement with human evaluators. Such a disparity highlights the fundamental difference between assessing the instruction-following ability and the overall response quality.

5.2 Instruction Following vs. Other Abilities

Table 4 presents a comparison of representative LLMs across different abilities, not just instruction following (FollowBench). This includes over-

Prompt	Agreement with Human
Ours	88%
Ours w/o ML	79%
Vicuna-Single	67%

Table 3: Agreement between human and diverse prompt templates. We use ML to denote multi-level.

Model	Following	Overall	Knowledge	Reasoning
GPT-4-Preview-1106	3.3	97.7	86.4	86.7
GPT-3.5-turbo-1106	2.9	86.3	70.0	70.1
LLaMA2-Chat-70B	2.1	92.7	63.0	60.8
WizardLM-13B-V1.2	2.2	89.2	52.7	–
LLaMA2-Chat-13B	2.2	81.1	53.6	40.2
Vicuna-13B-V1.5	2.1	–	55.8	51.5
LLaMA2-Chat-7B	1.9	71.4	45.8	35.6
Vicuna-7B-V1.5	1.7	–	49.8	43.4

Table 4: Model comparison on different abilities.

all response quality (AlpacaEval (Li et al., 2023)), knowledge (MMLU (Hendrycks et al., 2021)), and reasoning (BBH (Suzgun et al., 2022)). We can find that our FollowBench provides an additional perspective for a holistic LLM evaluation. As an illustration, while the performance of WizardLM-13B-V1.2 exceeds that of GPT-3.5 in terms of overall response quality, it notably lags behind in instruction-following ability. Similarly, Vicuna-V1.5 excels over LLaMA2-Chat in the realms of knowledge and reasoning but struggles with instruction-following tasks.

5.3 Does Failure at Lower Level Necessarily Lead to Failure at Higher Level?

For a set of instructions that has five difficulty levels, if a model’s response doesn’t satisfy the constraints at level n , where n ranges from 1 to 4, we define the *failure consistency* as the percentage that the response will also not fulfill the constraints at any subsequent level greater than n . Combining Table 2 and Table 5, it can be seen that models with better instruction-following capability may exhibit lower failure consistency. One possible reason is that the instruction-following ability of more powerful models is less sensitive to the number of constraints in an instruction, thus they are better equipped to adapt and fulfill the requirements even as the constraints increase. This adaptability means that while they may falter at a lower difficulty level, they can still manage to meet the demands of higher difficulty levels, leading to a decrease in failure consistency.

Model	Failure Consistency (%)
GPT-4-Preview-1106	42.2
WizardLM-13B-V1.2	57.3
Vicuna-7B-V1.5	61.8
ChatGLM3-6B	64.0

Table 5: Results on failure consistency.

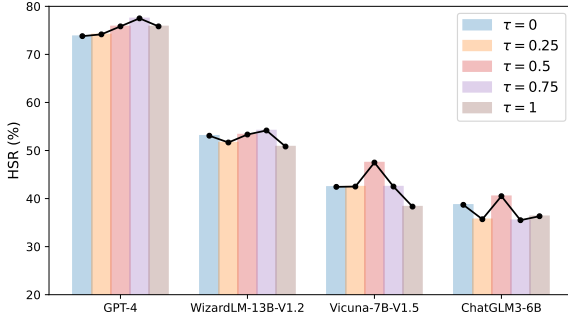


Figure 6: The effect of varying the temperature parameter τ . We use $\tau = 0$ to denote greedy decoding.

5.4 Does Different Decoding Strategies Affect the Instruction-following Ability?

In this section, we systematically investigate the impact of different decoding strategies, represented by the temperature parameter τ , on LLM’s instruction-following ability. The temperature τ is a commonly used parameter that controls the sharpness of the distribution from which we sample the next token:

$$P(w) = \frac{\exp(z_w/\tau)}{\sum_{w' \in V} \exp(z_{w'}/\tau)} \quad (4)$$

where z_w is the logit for word w , V is the vocabulary. Lower values for temperature result in more consistent outputs, while higher values generate more diverse and creative results. As illustrated in Figure 6, the temperature τ has a tangible influence on the instruction-following ability across all four models. The sweet spot seems to be somewhere in the middle where there’s enough variability to capture the nuances and intricacies of complex instructions, yet not so much that the model goes off tangent. This balanced behavior ensures that the model remains within the desired context, producing outputs that align closely with the given instructions while also allowing for a slight creative touch when needed.

6 Conclusion

In this paper, we introduce FollowBench, a Multi-level Fine-grained Constraints Following

Benchmark tailored for gauging the instruction-following capability of LLMs. FollowBench covers five *fine-grained* constraint categories and over 50 NLP tasks, utilizes a novel *Multi-level* mechanism for precisely estimating the upper limit of instruction-following capability. Furthermore, we propose an evaluation protocol with three metrics that seamlessly integrate with the multi-level mechanism. Our extensive tests over 13 popular LLMs reveal a substantial performance advantage for GPT-4 and GPT-3.5 over their counterparts, and there is still significant room for improving the instruction-following ability of current LLMs.

Limitations

While our study contributes valuable insights, it is essential to acknowledge several limitations that warrant consideration.

Firstly, our current investigation is confined to single-round interactions, aiming to offer a controlled environment for evaluation. Future research may extend its scope to multi-round conversations to comprehensively assess the instruction-following proficiency of LLMs in more dynamic and extended dialogues (Kwan et al., 2024).

Secondly, the model-based evaluation framework employed in our experiments, while rigorous, relies on prompt engineering, introducing an inherent imperfection. Despite our meticulous selection of high-performing prompts, the potential for further optimization remains, which may impact the reported evaluation metrics.

Lastly, we refrain from proposing specific solutions to address identified weaknesses of LLMs in instruction following. A plausible avenue for future research involves fine-tuning LLMs using our proposed FollowBench as a benchmark, providing a potential roadmap for enhancing instruction adherence. We defer the exploration of these aspects to subsequent studies, recognizing the need for a comprehensive examination of LLM capabilities across varying interaction complexities.

Ethics Statement

Our paper aims to systemically and precisely evaluate the capability of LLMs to follow natural language instructions. However, it is essential to bear in mind that malicious instructions have the potential to prompt the model to generate harmful or inappropriate outputs. Therefore, ensuring safe and responsible practices when assessing

the instruction-following capability of LLMs is of paramount importance. In FollowBench, each piece of data undergoes a meticulous human review process to identify and eliminate any potentially harmful instructions or offensive content. This rigorous approach underscores our commitment to maintaining a secure and ethical evaluation framework.

Acknowledgments

W. Wang was supported by HKUST(GZ) Grant G0101000028, CCF-HuaweiDBC202302, Guangzhou Municipal Science and Technology Project (No. 2023A03J0003, 2023A03J0013 and 2024A03J0621).

References

arXiv.org submitters. 2023. [arxiv dataset](#).

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. [Prompt-Source: An integrated development environment and repository for natural language prompts](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

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, Sheng-guang 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*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Baichuan. 2023. [Baichuan 2: Open large-scale language models](#). *arXiv preprint arXiv:2309.10305*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the Conference of European Association for Machine Translation (EAMT)*, pages 261–268.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Howard Chen, Huihan Li, Danqi Chen, and Karthik Narasimhan. 2022. Controllable text generation with language constraints. *arXiv preprint arXiv:2212.10466*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebguss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. [Evaluating large language models trained on code](#).

Pengyu Cheng and Ruining Li. 2022. Replacing language model for style transfer. *arXiv preprint arXiv:2211.07343*.

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

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm:

- General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. **Hierarchical neural story generation**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. *Blog post, April*, 1.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. **SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization**. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.
- David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- F Huang, H Kwak, and J An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arxiv*.
- Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. **Lion: Adversarial distillation of proprietary large language models**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3134–3154. Association for Computational Linguistics.
- Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, Qun Liu, and Wei Wang. 2024. **Learning to edit: Aligning llms with knowledge editing**. *CoRR*, abs/2402.11905.
- Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy S Liang. 2019. Spoc: Search-based pseudocode to code. *Advances in Neural Information Processing Systems*, 32.
- Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. **Mt-eval: A multi-turn capabilities evaluation benchmark for large language models**. *CoRR*, abs/2401.16745.
- Sha Li, Heng Ji, and Jiawei Han. 2021. **Document-level event argument extraction by conditional generation**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 894–908. Association for Computational Linguistics.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. **AlpacaEval: An automatic evaluator of instruction-following models**. https://github.com/tatsu-lab/alpaca_eval.
- Pierre Lison and Jörg Tiedemann. 2016. **OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles**. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.
- George A. Miller. 1992. **WordNet: A lexical database for English**. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. **Cross-task generalization via natural language crowdsourcing instructions**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. **A corpus and cloze evaluation for deeper understanding of commonsense stories**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.
- Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. **Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization**. In *Proceedings of the 2018*

- Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. [The E2E dataset: New challenges for end-to-end generation](#). In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- 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*.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role-play with large language models. *arXiv preprint arXiv:2305.16367*.
- Eric Michael Smith, Diana Gonzalez-Rico, Emily Dinan, and Y-Lan Boureau. 2020. Controlling style in generated dialogue. *arXiv preprint arXiv:2009.10855*.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.
- Xiangru Tang, Yiming Zong, Yilun Zhao, Arman Cohan, and Mark Gerstein. 2023. Struc-bench: Are large language models really good at generating complex structured data? *arXiv preprint arXiv:2309.08963*.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Alicia Tsai, Shereen Oraby, Vittorio Perera, Jiun-Yu Kao, Yuheng Du, Anjali Narayan-Chen, Tagyoung Chung, and Dilek Hakkani-Tur. 2021. [Style control for schema-guided natural language generation](#). In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 228–242, Online. Association for Computational Linguistics.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. [Self-instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.
- Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. [Aligning large language models with human: A survey](#). *arXiv preprint arXiv:2307.12966*.
- Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, et al. 2023c. [Interactive natural language processing](#). *arXiv preprint arXiv:2305.13246*.
- Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. 2020. [Learning from task descriptions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, Online. Association for Computational Linguistics.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. [Wizardlm: Empowering large language models to follow complex instructions](#).

Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. 2023. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*.

Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. Large language models are effective table-to-text generators, evaluators, and feedback providers. *arXiv preprint arXiv:2305.14987*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#).

Wanjuan Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. [Agieval: A human-centric benchmark for evaluating foundation models](#). *CoRR*, abs/2304.06364.

A Data Generation Process

Here we outline the sources for our data and provide a detailed description of the data generation process for each constraint category.

A.1 Content Constraints

The data of content constraints is constructed from five tasks as follows:

- **Data-to-Text Generation** We create instructions with 1 to 5 constraints by adapting samples from E2E (Novikova et al., 2017). Different from the original task, we ask the model to extract the flat meaning representations according to the corresponding natural language texts. The number of constraints increases with the number of attributes and the number of restaurants. We use exact match as the evaluation metric.
- **Document-Level Event Argument Extraction** We create instructions by adapting samples from WIKIEVENTS (Li et al., 2021). Given a document, the model is required to extract n events that satisfy a specific event template, where $n \in [1, 5]$ corresponds to the

number of constraints. We use accuracy as the evaluation metric.

- **Document-Level Named Entity Recognition** We derive instructions from samples in the CONLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003). We ask the model to extract a single named entity from a provided document. Notably, as the number of constraints rises, the requirements for the retrieved named entity correspondingly increase. For example, “extract one named entity that is a location” \rightarrow “extract one named entity that is a location in east Asia”. We use accuracy as the evaluation metric.
- **Text Generation with Language Constraints** COGNAC (Chen et al., 2022) is a challenging benchmark wherein models are presented with a topic accompanied by example text and explicit constraints on the text to avoid. We curate data from COGNAC, formulating instructions with 1 to 5 constraints by integrating additional linguistic restrictions from WordNet (Miller, 1992) and Wiki-data (Vrandečić and Krötzsch, 2014).
- **Open-ended Question Answering** We first choose initial instructions from existing datasets including self-instruct evaluation set (Wang et al., 2023a), helpful evaluation released by Anthropic (Bai et al., 2022), Vicuna evaluation (Zheng et al., 2023), and Koala evaluation (Geng et al., 2023), as well as open-source platforms such as Quora², Reddit³, and ShareGPT⁴. Given the challenges associated with iteratively adding constraints to an initial instruction, we prompt GPT-4 with a specific prompt shown in Figure 7 to generate a new instruction with one more constraint based on the given instruction. The above process is repeated five times. Finally, we obtain a set of instructions ranging from 1 to 5 constraints.

A.2 Situation Constraints

The data of situation constraints is constructed from tasks as follows:

- **Suggestion Generation, Role-playing** We collect multi-level instructions that fit within

²<https://www.quora.com>

³<https://www.reddit.com>

⁴<https://sharegpt.com>

the paradigm of situation constraints from Open-ended Question Answering datasets and online platforms. Examples include asking the model to give suggestions under specific circumstances, asking the model to act as a terminal and output based on the given information, etc.

- **Math Word Problems** The initial instructions are collected from GSM8K (Cobbe et al., 2021) and AGIEval (Zhong et al., 2023). We then manually add constraints progressively by enhancing the situation descriptions, ensuring that the core question remains unaltered. We use accuracy as the evaluation metric.
- **Time/Spatial Reasoning** We generate data by refining samples from BIG-Bench Hard (Suzgun et al., 2022). For Time Reasoning, we increase the difficulty level by incorporating additional temporal concepts, such as weeks, months, and years. In the realm of Spatial Reasoning, we opt for a logical deduction task that necessitates deducing the order of a sequence of objects. Here, the number of constraints escalates by augmenting the task with detailed location descriptions for a new object. We use accuracy as the evaluation metric.
- **Code Generation** We sourced initial instructions from HumanEval (Chen et al., 2021) and enhanced the difficulty level by adding complexity to the function descriptions within the instructions. We use pass@1 (Kulal et al., 2019) as the evaluation metric.

A.3 Example Constraints

Specifically, we choose 40 diverse NLP tasks from PromptSource (Bach et al., 2022), where each task has more than 5 question templates. Additionally, we create 29 answer templates (shown in Table 6) that regulate the format of the response. For instructions at difficulty level 1, we utilize the standard 5-shot prompting, where 5 shots are equipped with 1 sampled question template and 1 sampled answer template, and the model is required to respond to a query using the answer template. For instructions at difficulty level n ($1 < n \leq 5$), the 5 shots are randomly paired with n question templates and n corresponding answer templates. Based on the question template of the query, the model is required to recognize the matched question template in the 5 shots and respond using the corresponding

Answer template

```
{question}\n{answer}
{question}\nA: {answer}
{question}\nAnswer: {answer}
{question}\nANSWER: {answer}
{question}\n[Answer]\n{answer}
{question}\n#Answer#\n{answer}
{question}\nThe answer is: {answer}
{question}\n{"answer": "{answer}"}
{question}\n{"Answer": "{answer}"}
{question}\n<body>{answer}</body>
{question}\nResponse: {answer}
{question}\nRESPONSE: {answer}
{question}\n[Response]\n{answer}
{question}\n#Response#\n{answer}
{question}\nThe response is: {answer}
{question}\n{"response": "{answer}"}
{question}\n{"Response": "{answer}"}
{question}\nBot: {answer}
{question}\nBOT: {answer}
{question}\n[Bot]\n{answer}
{question}\n#Bot#\n{answer}
{question}\nThe response of the bot is: {answer}
{question}\n{"bot": "{answer}"}
{question}\n{"Bot": "{answer}"}
{question}\nAI assistant: {answer}
{question}\n[AI assistant]\n{answer}
{question}\n#AI assistant#\n{answer}
{question}\nThe response of the AI assistant is: {answer}
{question}\n{"AI assistant": "{answer}"}
```

Table 6: Answer template of Example Constraints.

answer template. We use accuracy as the evaluation metric.

A.4 Mixed Constraints

In this paper, we consider four below tasks which are naturally suitable for constructing mixed constraints:

- **Text Editing** We start by gathering text from different online sources, like sentences, letters, and emails. Next, we create instructions with multi-level mixed constraints by increasingly adding an editing requirement to the text at each level. For example, “swap the first and last words in the sentence” (Content Constraints), “response using ‘###’ at the beginning” (Format Constraints), etc. We write rule-based programs for individual instructions to assess the satisfaction of internal constraints, employing exact match as the evaluation metric.
- **Summarization** The initial instructions are sampled from CNN/Daily Mail (Nallapati et al., 2016), XSum (Narayan et al., 2018), SAMSum (Gliwa et al., 2019), English Gigaword (Graff et al., 2003), and arXiv (arXiv.org

submitters, 2023). The instructions with multi-level mixed constraints are produced by specifying the format of generating answers (Format Constraints), requiring the generated text to include or not include certain keywords (Content Constraints), etc. We write rule-based programs for individual instructions to assess the satisfaction of internal constraints, employing accuracy as the evaluation metric.

- **Machine Translation** The initial instructions are sampled from OpenSubtitles (Lison and Tiedemann, 2016), TED Talks (Cettolo et al., 2012), and News-Commentary (Tiedemann, 2012). Then we construct instructions from level 1 to level 5 using a similar pipeline as that of Summarization. We write rule-based programs for individual instructions to assess the satisfaction of internal constraints, employing accuracy as the evaluation metric.
- **Story Generation** We collect initial instructions from ROCStories (Mostafazadeh et al., 2016) and WritingPrompts (Fan et al., 2018). Then we add 5 mixed constraints sequentially to the initial instructions based on the ground truth, such as the number of sentences in the generated story (Format Constraints), requiring the generated text to include certain keywords (Content Constraints), specifying the writing style (Style Constraints), etc.

B Detailed Experimental Results

Here we list the experimental results across 5 difficulty levels for each constraint category, including Content Constraints in Table 7, Situation Constraints in Table 8, Style Constraints in Table 9, Format Constraints in Table 10, Example Constraints in Table 11, and Mixed Constraints in Table 12.

C Case Study

Table 13 and Table 14 show the respective responses and evaluation results of GPT-4 and WizardLM-13B-V1.5 when tasked with a level-5 instruction under the category of Content Constraints. It can be observed that GPT-4 meets all five specified constraints, whereas WizardLM-13B-V1.5 fails to fulfill the third constraint, which mandates that the output animals must be able to swim. Besides, these two cases also validate the effectiveness of our model-based evaluation.

D FollowBench-zh

To facilitate the multilingual evaluation of LLM’s instruction-following ability, we have additionally crafted a Chinese version of FollowBench, denoted as FollowBench-zh. This involved employing a data generation process analogous to that utilized in the development of the English version. Overall, FollowBench-zh consists of 790 meticulously curated instructions from over 50 NLP tasks, including both closed- and open-ended questions. The detailed data statistics are listed in Table 15.

Following §3.2 and §4.1, we evaluate 13 popular LLMs on FollowBench-zh. The prompt template for model-based evaluation of FollowBench-zh is shown in Figure 10. It is noticeable that although LLaMA2-Chat-70B/13B/7B, WizardLM-13B-V1.2, and Vicuna-13B/7B-V1.5 are not specifically trained on Chinese corpora, they can still understand and respond in Chinese. Table 16 provides a comprehensive comparison of various models across five difficulty levels, denoted as L1 to L5. Similar to FollowBench, the performance of nearly all models on FollowBench-zh typically diminishes as we progress from L1 to L5. Nevertheless, GPT-3.5 exhibits a notably diminished proficiency in following instructions on FollowBench-zh in comparison to GPT-4, showcasing a more pronounced performance gap than observed on FollowBench. Moreover, models such as Baichuan2-Chat-7B and ChatGLM3-6B, which are pre-trained on a combination of English and Chinese corpora, demonstrate comparable or even better performance compared to their open-source counterparts. This highlights the significance of incorporating diverse linguistic datasets in pre-training to enhance the multilingual instruction-following capability of LLMs. Figure 11 depicts the instruction-following capability of LLMs over different constraint categories, with GPT-4 standing out notably among its counterparts. In a nutshell, there is still a substantial opportunity for enhancing the instruction-following capabilities of existing LLMs.

Prompt Template (Open-ended Question Answering in Content Constraints)

You are an Instruction Rewriting Expert. You need to rewrite #Given Instruction# based on #Rewriting Requirement#, in order to obtain a #Rewritten Instruction#. Basically, #Rewritten Instruction# should adhere to the following guidelines:

1. Your rewriting cannot omit the non-text parts such as the table and code in #Given Instruction#.
2. #Rewritten Instruction# must be reasonable and must be understood and responded by humans.
3. You should try your best not to make the #Rewritten Instruction# become verbose, #Rewritten Instruction# can only add 10 to 20 words into #Given Instruction#.

#Given Instruction#
{given_instruction}

#Rewriting Requirement#
 Please add one proper content constraint to the #Given Instruction#. The content constraints include but are not limited to:

1. Add a Subtask or Another Related Question.
2. Narrow Down the Topic: Instead of a general theme or topic, provide a more specific subset.
3. Set a Higher Standard: Raise the bar for what's considered acceptable or successful.
4. Limit Resources: Restrict the number or type of resources someone can use.
5. Introduce Specific Criteria: Mandate particular components or features that must be included.
6. Specifying Sequence: Dictate the order in which certain steps or actions should be taken.

#Rewritten Instruction#

Figure 7: The prompt template for Open-ended Question Answering in Content Constraints.

Prompt Template (Open-ended Question Answering in Style Constraints)

You are an Instruction Rewriting Expert. You need to rewrite #Given Instruction# based on #Rewriting Requirement#, in order to obtain a #Rewritten Instruction#. Basically, #Rewritten Instruction# should adhere to the following guidelines:

1. Your rewriting cannot omit the non-text parts such as the table and code in #Given Instruction#.
2. #Rewritten Instruction# must be reasonable and must be understood and responded by humans.
3. You should try your best not to make the #Rewritten Instruction# become verbose, #Rewritten Instruction# can only add 10 to 20 words into #Given Instruction#.

#Given Instruction#
{given_instruction}

#Rewriting Requirement#
 Please add one proper style constraint that #Given Instruction# does not have. The style constraints include but are not limited to:

1. Tone and Emotion: Specify the desired emotional tone for the response.
2. Writing Style: Ask the AI to mimic a specific author's writing style.
3. Contradiction: Ask the AI to provide a response that contradicts the previous statement or take a stance opposite to its prior response.
4. Ambiguity: Instruct the AI to create responses with intentional ambiguity or double meanings.
5. Humor or Satire: Request that the response be humorous or satirical, requiring the AI to generate jokes or witty remarks.

#Rewritten Instruction#

Figure 8: The prompt template for Open-ended Question Answering in Style Constraints.

Prompt Template (Open-ended Question Answering in Format Constraints)

You are an Instruction Rewriting Expert. You need to rewrite #Given Instruction# based on #Rewriting Requirement#, in order to obtain a #Rewritten Instruction#. Basically, #Rewritten Instruction# should adhere to the following guidelines:

1. Your rewriting cannot omit the non-text parts such as the table and code in #Given Instruction#.
2. #Rewritten Instruction# must be reasonable and must be understood and responded by humans.
3. You should try your best not to make the #Rewritten Instruction# become verbose, #Rewritten Instruction# can only add 10 to 20 words into #Given Instruction#.

#Given Instruction#
{given_instruction}

#Rewriting Requirement#
 Please add one proper format constraint that #Given Instruction# does not have. The format constraints include but are not limited to:

1. Length: Imposing constraints on the length of individual words, sentences, or paragraphs.
2. Hierarchical Instructions: Providing instructions that have a hierarchical structure, where the AI needs to understand and follow a hierarchy of tasks to construct a response.
3. Special Output Format: Asking the AI to respond by using data format like table, json, HTML, LaTeX, etc.
4. Morphological Constraints: Asking the AI to avoid or use specific morphemes.
5. Multi-lingual Constraints: Asking the AI to respond in multiple languages or switch between languages according to complex patterns.
6. Incorporation of Specific Literary Devices: Requiring the inclusion of specific, and perhaps numerous, literary devices.
7. Following a Specific Grammatical Structure: Requiring the AI to create responses that strictly follow a particular grammatical structure.

#Rewritten Instruction#

Figure 9: The prompt template for Open-ended Question Answering in Format Constraints.

Model	HSR (%)						SSR (%)						CSL
	L1	L2	L3	L4	L5	Avg.	L1	L2	L3	L4	L5	Avg.	
GPT-4-Preview-1106	84.0	76.0	72.0	80.0	72.0	76.8	84.0	78.0	74.7	83.0	80.8	80.1	3.5
GPT-3.5-Turbo-1106	72.0	68.0	72.0	56.0	48.0	63.2	72.0	70.0	76.0	67.0	64.0	69.8	2.7
Qwen-Chat-72B	84.0	72.0	72.0	52.0	56.0	67.2	84.0	76.0	76.0	62.0	66.4	72.9	3.2
LLaMA2-Chat-70B	48.0	44.0	44.0	40.0	40.0	43.2	48.0	48.0	48.0	47.0	47.2	47.6	2.2
Qwen-Chat-14B	56.0	64.0	60.0	44.0	36.0	52.0	56.0	68.0	66.7	55.0	51.2	59.4	2.1
WizardLM-13B-V1.2	68.0	56.0	48.0	44.0	28.0	48.8	68.0	60.0	56.0	51.0	45.6	56.1	2.4
LLaMA2-Chat-13B	48.0	44.0	48.0	48.0	36.0	44.8	48.0	48.0	50.7	50.0	47.2	48.8	2.1
Vicuna-13B-V1.5	60.0	52.0	52.0	44.0	32.0	48.0	60.0	58.0	58.7	53.0	44.8	54.9	2.3
Qwen-Chat-7B	56.0	56.0	36.0	36.0	24.0	41.6	56.0	60.0	48.0	48.0	40.8	50.6	1.7
LLaMA2-Chat-7B	44.0	48.0	44.0	40.0	36.0	42.4	44.0	48.0	46.7	46.0	46.4	46.2	1.8
Vicuna-7B-V1.5	60.0	48.0	52.0	40.0	16.0	43.2	60.0	56.0	61.3	51.0	44.0	54.5	1.9
Baichuan2-Chat-7B	60.0	48.0	40.0	36.0	24.0	41.6	60.0	52.0	45.3	50.0	44.8	50.4	1.7
ChatGLM3-6B	68.0	44.0	44.0	36.0	24.0	43.2	68.0	52.0	49.3	50.0	40.8	52.0	1.9

Table 7: Results of Content Constraints across 5 difficulty levels. Proprietary LLMs , open-sourced LLMs (large) , open-sourced LLMs (medium) , and open-sourced LLMs (small) are distinguished by different colors.

Model	HSR (%)						SSR (%)						CSL
	L1	L2	L3	L4	L5	Avg.	L1	L2	L3	L4	L5	Avg.	
GPT-4-Preview-1106	90.0	90.0	85.0	65.0	50.0	76.0	90.0	90.0	88.3	76.2	69.0	82.7	3.5
GPT-3.5-Turbo-1106	72.7	72.7	72.7	63.6	68.2	70.0	72.7	72.7	75.8	71.6	75.5	73.7	3.2
Qwen-Chat-72B	81.8	81.8	54.6	54.6	50.0	64.6	81.8	86.4	71.2	65.9	60.9	73.2	2.4
LLaMA2-Chat-70B	72.7	68.2	54.6	40.9	50.0	57.3	72.7	70.5	66.7	61.4	68.2	67.9	2.4
Qwen-Chat-14B	72.7	68.2	59.1	45.5	50.0	59.1	72.7	72.7	71.2	56.8	61.8	67.1	2.2
WizardLM-13B-V1.2	65.0	65.0	70.0	35.0	45.0	56.0	65.0	67.5	71.7	50.0	58.0	62.4	1.9
LLaMA2-Chat-13B	63.6	77.3	59.1	45.5	36.4	56.4	63.6	81.8	69.7	58.0	53.6	65.4	2.2
Vicuna-13B-V1.5	68.2	63.6	54.5	31.8	40.9	51.8	68.2	65.9	60.6	47.7	54.5	59.4	1.9
Qwen-Chat-7B	59.1	59.1	54.6	36.4	50.0	51.8	59.1	68.2	65.2	52.3	66.4	62.2	1.6
LLaMA2-Chat-7B	68.2	45.5	54.5	27.3	54.5	50.0	68.2	59.1	63.6	54.5	65.5	62.2	1.8
Vicuna-7B-V1.5	45.5	45.5	31.8	22.7	27.3	34.5	45.5	50.0	43.9	34.1	49.1	44.5	1.4
Baichuan2-Chat-7B	36.4	40.9	40.9	22.7	18.2	31.8	36.4	54.5	54.5	42.0	41.8	45.9	0.9
ChatGLM3-6B	63.6	63.6	40.9	27.3	22.7	43.6	63.6	70.5	56.1	44.3	43.6	55.6	1.8

Table 8: Results of Situation Constraints across 5 difficulty levels. Proprietary LLMs , open-sourced LLMs (large) , open-sourced LLMs (medium) , and open-sourced LLMs (small) are distinguished by different colors.

Model	HSR (%)						SSR (%)						CSL
	L1	L2	L3	L4	L5	Avg.	L1	L2	L3	L4	L5	Avg.	
GPT-4-Preview-1106	96.7	93.3	86.7	96.7	90.0	92.7	96.7	95.0	93.3	98.3	98.0	96.3	4.3
GPT-3.5-Turbo-1106	96.7	93.3	90.0	93.3	86.7	92.0	96.7	96.7	95.6	98.3	97.3	96.9	4.1
Qwen-Chat-72B	90.0	83.3	70.0	66.7	56.7	73.3	90.0	90.0	90.0	89.2	86.0	89.0	3.0
LLaMA2-Chat-70B	96.7	93.3	93.3	83.3	83.3	90.0	96.7	96.7	97.8	95.0	96.0	96.4	4.1
Qwen-Chat-14B	80.0	73.3	46.7	60.0	46.7	61.3	80.0	86.7	75.6	82.5	78.0	80.6	2.2
WizardLM-13B-V1.2	96.7	93.3	80.0	83.3	60.0	82.7	96.7	95.0	91.1	92.5	90.0	93.1	3.6
LLaMA2-Chat-13B	96.7	93.3	90.0	86.7	86.7	90.7	96.7	96.7	95.6	96.7	96.0	96.3	4.1
Vicuna-13B-V1.5	90.0	90.0	60.0	73.3	60.0	74.7	90.0	95.0	83.3	87.5	89.3	89.0	3.1
Qwen-Chat-7B	66.7	73.3	46.7	53.3	33.3	54.7	66.7	86.7	76.7	80.0	64.7	74.9	1.6
LLaMA2-Chat-7B	96.7	93.3	90.0	86.7	70.0	87.3	96.7	96.7	95.6	96.7	93.3	95.8	4.1
Vicuna-7B-V1.5	80.0	80.0	53.3	63.3	53.3	66.0	80.0	88.3	80.0	87.5	82.0	83.6	2.3
Baichuan2-Chat-7B	76.7	83.3	56.7	53.3	50.0	64.0	76.7	90.0	80.0	85.8	87.3	84.0	2.2
ChatGLM3-6B	80.0	60.0	50.0	36.7	33.3	52.0	80.0	76.7	73.3	74.2	74.7	75.8	1.9

Table 9: Results of Style Constraints across 5 difficulty levels. Proprietary LLMs , open-sourced LLMs (large) , open-sourced LLMs (medium) , and open-sourced LLMs (small) are distinguished by different colors.

Model	HSR (%)						SSR (%)						CSL
	L1	L2	L3	L4	L5	Avg.	L1	L2	L3	L4	L5	Avg.	
GPT-4-Preview-1106	90.0	93.3	86.7	93.3	80.0	86.0	90.0	95.0	94.4	98.3	93.3	90.1	4.1
GPT-3.5-Turbo-1106	90.0	76.7	80.0	70.0	50.0	73.3	90.0	85.0	88.9	85.0	82.0	86.2	3.2
Qwen-Chat-72B	86.7	80.0	83.3	60.0	56.7	73.3	86.7	86.7	87.8	83.3	81.3	85.2	3.3
LLaMA2-Chat-70B	83.3	76.7	66.7	53.3	36.7	63.3	83.3	85.0	86.7	78.3	70.0	80.7	2.4
Qwen-Chat-14B	86.7	86.7	80.0	56.7	33.3	68.7	86.7	90.0	86.7	80.8	72.0	83.2	3.1
WizardLM-13B-V1.2	83.3	93.3	73.3	56.7	40.0	69.3	83.3	95.0	86.7	77.5	73.3	83.2	2.9
LLaMA2-Chat-13B	86.7	80.0	70.0	56.7	40.0	66.7	86.7	86.7	86.7	80.0	72.7	82.5	3.1
Vicuna-13B-V1.5	86.7	76.7	76.7	53.3	30.0	64.7	86.7	85.0	86.7	78.3	70.7	81.5	2.6
Qwen-Chat-7B	76.7	76.7	66.7	50.0	30.0	60.0	76.7	83.3	77.8	74.2	70.7	76.5	2.4
LLaMA2-Chat-7B	80.0	80.0	66.7	53.3	33.3	62.7	80.0	88.3	86.7	78.3	68.0	80.3	2.4
Vicuna-7B-V1.5	80.0	76.7	73.3	43.3	20.0	58.7	80.0	86.7	85.6	70.8	67.3	78.1	2.4
Baichuan2-Chat-7B	80.0	56.7	60.0	40.0	36.7	54.7	80.0	73.3	81.1	72.5	72.7	75.9	1.9
ChatGLM3-6B	80.0	60.0	46.7	33.3	23.3	48.7	80.0	71.7	72.2	69.2	68.0	72.2	2.1

Table 10: Results of Format Constraints across 5 difficulty levels. Proprietary LLMs , open-sourced LLMs (large) , open-sourced LLMs (medium) , and open-sourced LLMs (small) are distinguished by different colors.

Model	HSR (%)						SSR (%)						CSL
	L1	L2	L3	L4	L5	Avg.	L1	L2	L3	L4	L5	Avg.	
GPT-4-Preview-1106	87.5	57.5	57.5	45.0	42.5	58.0	87.5	57.5	57.5	45.0	42.5	58.0	2.4
GPT-3.5-Turbo-1106	80.0	50.0	50.0	42.5	42.5	53.0	80.0	50.0	50.0	42.5	42.5	53.0	2.2
Qwen-Chat-72B	30.0	10.0	5.0	2.5	2.5	10.0	30.0	10.0	5.0	2.5	2.5	10.0	0.5
LLaMA2-Chat-70B	0.0	2.5	0.0	0.0	0.0	0.5	0.0	2.5	0.0	0.0	0.0	0.5	0.0
Qwen-Chat-14B	22.5	10.0	5.0	2.5	7.5	9.5	22.5	10.0	5.0	2.5	7.5	9.5	0.4
WizardLM-13B-V1.2	40.0	30.0	27.5	12.5	15.0	25.0	40.0	30.0	27.5	12.5	15.0	25.0	0.9
LLaMA2-Chat-13B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Vicuna-13B-V1.5	57.5	37.5	25.0	17.5	17.5	31.0	57.5	37.5	25.0	17.5	17.5	31.0	1.2
Qwen-Chat-7B	12.5	10.0	5.0	5.0	2.5	7.0	12.5	10.0	5.0	5.0	2.5	7.0	0.2
LLaMA2-Chat-7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Vicuna-7B-V1.5	52.5	32.5	25.0	12.5	15.0	27.5	52.5	32.5	25.0	12.5	15.0	27.5	1.2
Baichuan2-Chat-7B	50.0	30.0	35.0	12.5	12.5	28.0	50.0	30.0	35.0	12.5	12.5	28.0	1.1
ChatGLM3-6B	32.5	22.5	15.0	10.0	7.5	17.5	32.5	22.5	15.0	10.0	7.5	17.5	0.6

Table 11: Results of Example Constraints across 5 difficulty levels. Proprietary LLMs , open-sourced LLMs (large) , open-sourced LLMs (medium) , and open-sourced LLMs (small) are distinguished by different colors.

Model	HSR (%)						SSR (%)						CSL
	L1	L2	L3	L4	L5	Avg.	L1	L2	L3	L4	L5	Avg.	
GPT-4-Preview-1106	60.0	46.7	40.0	66.7	40.0	50.7	60.0	50.0	48.9	66.7	56.0	56.3	1.9
GPT-3.5-Turbo-1106	70.6	47.1	47.1	41.2	23.5	45.9	70.6	52.9	58.8	52.9	41.2	55.3	1.7
Qwen-Chat-72B	70.6	52.9	41.2	35.3	17.7	43.5	70.6	55.9	49.0	42.7	38.8	51.4	1.8
LLaMA2-Chat-70B	58.8	35.3	17.7	23.5	17.7	30.6	58.8	41.2	35.3	38.2	37.7	42.2	1.2
Qwen-Chat-14B	58.8	35.3	35.3	23.5	11.8	32.9	58.8	44.1	41.2	38.2	37.7	44.0	1.4
WizardLM-13B-V1.2	60.0	46.7	20.0	13.3	26.7	33.3	60.0	46.7	37.8	36.7	41.3	44.5	1.4
LLaMA2-Chat-13B	47.1	41.2	35.3	29.4	29.4	36.5	47.1	47.1	45.1	44.1	43.5	45.4	1.5
Vicuna-13B-V1.5	64.7	41.2	29.4	23.5	23.5	36.5	64.7	47.1	45.1	42.6	44.7	48.9	1.5
Qwen-Chat-7B	64.7	35.3	23.5	17.6	0.0	28.2	64.7	41.2	37.3	33.8	30.6	41.5	1.2
LLaMA2-Chat-7B	58.8	41.2	29.4	29.4	17.6	35.3	58.8	47.1	41.2	39.7	35.3	44.4	1.5
Vicuna-7B-V1.5	47.1	29.4	17.6	17.6	11.8	24.7	47.1	38.2	37.2	33.8	36.5	38.6	1.1
Baichuan2-Chat-7B	47.1	17.6	11.8	17.6	11.8	21.2	47.1	32.4	33.3	36.8	36.5	37.2	0.7
ChatGLM3-6B	41.2	29.4	23.5	23.5	17.7	27.1	41.2	38.2	41.2	39.7	35.3	39.1	1.1

Table 12: Results of Mixed Constraints across 5 difficulty levels. Proprietary LLMs , open-sourced LLMs (large) , open-sourced LLMs (medium) , and open-sourced LLMs (small) are distinguished by different colors.

The response and evaluation result of GPT-4-Preview-1106 for a level-5 instruction	
Instruction	<p>Talk about domestic animal: == the norwegian elkhound is one of the northern spitz type breed of dog and is the national dog of norway. == == the irish terrier (irish: brocaire rua) is a dog breed from ireland, one of many breeds of terrier. == == a hearing dog is a type of assistance dog specifically selected and trained to assist people who are deaf or hard of hearing by alerting their handler to important sounds, such as doorbells, smoke alarms, ringing telephones, or alarm clocks. == Could you list three more domestic animals that satisfy ALL of following conditions: 1. Do not talk about dog. 2. Their weights are less than 100kg. 3. Can swim. 4. Have 4 legs. 5. Are hairy.</p>
Response	<p>1. The Maine Coon cat is a large domesticated cat breed that originates from the United States. Known for their size, they typically weigh less than 100kg. They have a thick, water-resistant fur coat, which allows them to swim if necessary.</p> <p>2. The American Mink is a semi-aquatic, fur-bearing mammal that is often domesticated for its fur. It has four legs, can swim, and weighs significantly less than 100kg.</p> <p>3. The Norwegian Forest Cat, also known as the "Wegie", is a breed of domestic cat originating in Northern Europe. This breed is adapted to a very cold climate, with a top coat of glossy, long, water-shedding hairs, and a woolly undercoat for insulation. Despite their large size, they weigh less than 100kg and are known to be good swimmers.</p>
Evaluation	<p>1) The 5 added constraints are: - Do not talk about dog. - Their weights are less than 100kg. - Can swim. - Have 4 legs. - Are hairy.</p> <p>2) For the 5 added constraints, the #Answer of Initial Instruction + 5 constraints# satisfies each constraint as follows: - Do not talk about dog: Yes, the answer talks about Maine Coon cat, American Mink, and Norwegian Forest Cat, not dogs. - Their weights are less than 100kg: Yes, all the animals mentioned weigh less than 100kg. - Can swim: Yes, all the animals mentioned can swim. - Have 4 legs: Yes, all the animals mentioned have 4 legs. - Are hairy: Yes, all the animals mentioned are hairy.</p> <p>3) ['YES', 'YES', 'YES', 'YES', 'YES']</p>

Table 13: The response and evaluation result of GPT-4-Preview-1106 for a level-5 instruction belonging to Content Constraints.

The response and evaluation result of WizardLM-13B-V1.2 for a level-5 instruction	
Instruction	<p>Talk about domestic animal: == the norwegian elkhound is one of the northern spitz type breed of dog and is the national dog of norway. == == the irish terrier (irish: brocaire rua) is a dog breed from ireland, one of many breeds of terrier. == == a hearing dog is a type of assistance dog specifically selected and trained to assist people who are deaf or hard of hearing by alerting their handler to important sounds, such as doorbells, smoke alarms, ringing telephones, or alarm clocks. == Could you list three more domestic animals that satisfy ALL of following conditions: 1. Do not talk about dog. 2. Their weights are less than 100kg. 3. Can swim. 4. Have 4 legs. 5. Are hairy.</p>
Response	<p>1. Cat: Cats are domestic animals that weigh less than 100kg, have four legs, and are hairy. They are also known for their swimming abilities, making them suitable for households with access to water. 2. Rabbit: Rabbits are small, furry animals that weigh less than 100kg and have four legs. They are also excellent swimmers, thanks to their powerful hind legs and water-resistant fur. 3. Hamster: Hamsters are small, furry rodents that weigh less than 100kg and have four legs. They are popular pets due to their gentle nature and low maintenance care.</p>
Evaluation	<p>1) The 5 added constraints are: - Do not talk about dog. - Their weights are less than 100kg. - Can swim. - Have 4 legs. - Are hairy.</p> <p>2) For the 5 added constraints: - Do not talk about dog: The answer does not mention any dogs. It talks about cats, rabbits, and hamsters. So, it satisfies this constraint. - Their weights are less than 100kg: The animals mentioned in the answer (cats, rabbits, and hamsters) all weigh less than 100kg. So, it satisfies this constraint. - Can swim: The answer mentions that cats and rabbits can swim. However, hamsters are not natural swimmers. So, it does not satisfy this constraint. - Have 4 legs: All the animals mentioned in the answer (cats, rabbits, and hamsters) have four legs. So, it satisfies this constraint. - Are hairy: All the animals mentioned in the answer (cats, rabbits, and hamsters) are described as being hairy or furry. So, it satisfies this constraint.</p> <p>3) ['YES', 'YES', 'NO', 'YES', 'YES']</p>

Table 14: The response and evaluation result of WizardLM-13B-V1.2 for a level-5 instruction belonging to Content Constraints.



















Constraint	Task	Avg Len	#Data	Evaluation
Content	Data-to-Text Generation	158	25	
	Document-Level Event Argument Extraction	1,356	15	
	Document-Level Named Entity Recognition	652	25	
	Text Generation with Language Constraints	167	25	
	Open-ended Question Answering	116	25	
Situation	Suggestion Generation	139	40	
	Role-playing	203	15	
	Complex Situation Reasoning	187	55	
Style	Open-ended Question Answering	120	150	
Format	Text-to-Table Generation	305	30	
	Open-ended Question Answering	136	120	
Example	40 diverse NLP tasks	1,556	200	
Mixed	Text Editing	195	20	
	Summarization	481	25	
	Machine Translation	179	10	
	Story Generation	56	10	

Table 15: An overview of FollowBench-zh. “Avg Len” is the average character number of instructions.  refers to rule-based evaluation, while  refers to model-based evaluation.

Model	HSR (%)						SSR (%)						CSL
	L1	L2	L3	L4	L5	Avg.	L1	L2	L3	L4	L5	Avg.	
GPT-4-Preview-1106	86.7	83.9	68.7	67.0	61.1	73.5	86.7	84.8	76.0	74.0	71.8	78.6	3.1
GPT-3.5-Turbo-1106	69.6	65.2	52.8	49.1	39.5	55.2	69.6	70.8	63.8	64.0	59.5	65.5	2.2
Qwen-Chat-72B	66.8	58.9	47.3	42.8	36.5	50.5	66.8	62.3	59.1	59.1	57.9	61.0	2.1
LLaMA2-Chat-70B	52.8	46.4	41.0	30.3	23.5	38.8	52.8	53.1	54.0	51.0	49.1	52.0	1.5
Qwen-Chat-14B	64.0	48.3	42.2	33.6	25.9	42.8	64.0	57.2	56.5	53.7	51.2	56.5	1.6
WizardLM-13B-V1.2	55.9	46.5	37.8	29.4	19.6	37.8	55.9	50.9	51.3	50.2	47.3	51.1	1.6
LLaMA2-Chat-13B	53.3	46.0	36.1	30.6	29.5	39.1	53.3	51.9	50.4	48.3	49.0	50.6	1.6
Vicuna-13B-V1.5	56.4	43.8	36.9	32.4	22.5	38.4	56.4	53.0	52.0	52.2	46.5	52.0	1.5
Qwen-Chat-7B	54.2	45.6	33.2	23.1	18.9	35.0	54.2	52.9	51.2	50.6	46.4	51.0	1.3
LLaMA2-Chat-7B	54.0	44.7	37.6	21.7	21.7	35.9	54.0	51.3	51.0	44.2	44.4	49.0	1.5
Vicuna-7B-V1.5	52.6	37.8	30.0	22.0	13.4	31.2	52.6	48.8	46.6	46.7	40.5	47.1	1.2
ChatGLM3-6B	62.0	45.9	36.6	28.1	17.8	38.1	62.0	53.4	54.3	49.1	45.6	52.9	1.5

Table 16: Results across five difficulty levels of FollowBench-zh. For each level, we compute the average score of all constraint categories. Proprietary LLMs, open-sourced LLMs (large), open-sourced LLMs (medium), and open-sourced LLMs (small) are distinguished by different colors.

```

Prompt Template for Model-based Evaluation (zh)

给定一个初始指令，我们每次添加一个{constraint_type}约束，获得了有{level_n}个附加约束的最终指令。

#初始指令#
{initial_instruction}

#初始指令 + 1个约束#
{level_1_instruction}
:
#初始指令 + {level_n}个约束#
{level_n_instruction}

#初始指令 + {level_n}个约束的答案#
{answer_of_level_n_instruction}

#系统#
1)请识别出添加的{level_n}个约束。
2)对于添加的{level_n}个约束，请判断#初始指令 + {level_n}个约束的答案#是否满足了每个约束。
3)在最后一行，只输出一个包含{level_n}个元素（“YES”或“NO”）的Python列表，表明答案是否满足添加的每一个约束。

```

Figure 10: Prompt template for model-based evaluation of FollowBench-zh.

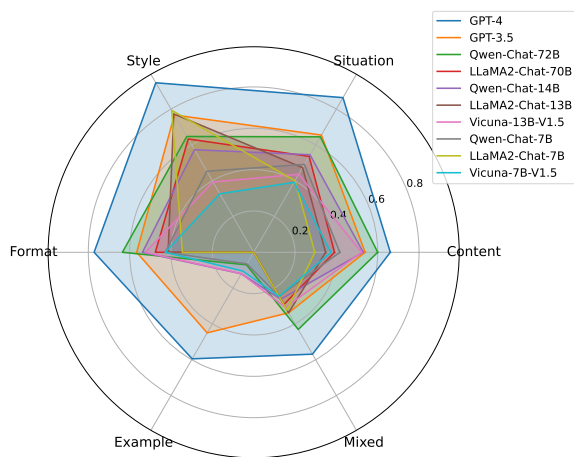


Figure 11: HSR (%) results in diverse constraint categories of FollowBench-zh. For each category, we compute the average score of all difficulty levels.