
Supplementary Materials for Paper 2401: Benchmarking Complex Instruction-Following with Multiple Constraints Composition

Anonymous Author(s)

Affiliation

Address

email

1 Dataset Documentation and Intended Uses

2 1.1 Datasheet

3 We present a datasheet [1] for documentation and responsible usage of COMPLEXBENCH.

4 1.1.1 Motivation

- 5 1. **For what purpose was the dataset created?** It was created as a benchmark for complex
6 instruction-following.
- 7 2. **Who created the dataset (e.g., which team, research group) and on behalf of which entity
8 (e.g., company, institution, organization)?** It was created by the authors of this paper.
- 9 3. **Who funded the creation of the dataset?** Since we choose to submit double-blind, specific
10 information will be disclosed after the paper is published.

11 1.1.2 Composition

- 12 1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people,
13 countries)?** Each instance in the dataset consists of complex instructions with their corresponding
14 annotation of constraint dimensions and composition types, as well as the scoring questions to
15 verify each constraint dimension and composition type.
- 16 2. **How many instances are there in total (of each type, if appropriate)?** The dataset has 1,150
17 instances.
- 18 3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of
19 instances from a larger set?** No.
- 20 4. **Is there a label or target associated with each instance?** No.
- 21 5. **Are relationships between individual instances made explicit?** No.
- 22 6. **Is any information missing from individual instances?** No.
- 23 7. **Are there recommended data splits?** The dataset is recommended to be used for evaluation
24 entirely.
- 25 8. **Are there any errors, sources of noise, or redundancies in the dataset?** Minor noises may
26 be introduced during the manual annotation. We have conducted a strict validation process to
27 alleviate the negative impact of these noises on our dataset and control the data quality.
- 28 9. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g.,
29 websites, tweets, other datasets)?** The dataset is self-contained.

30 10. Does the dataset contain data that might be considered confidential (e.g., data that is
31 protected by legal privilege or by doctor-patient confidentiality, data that includes the
32 content of individuals' non-public communications)? No.

33 11. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threaten-
34 ing, or might otherwise cause anxiety? No.

35 1.1.3 Collection Process

36 1. How was the data associated with each instance acquired? We collect reference instructions
37 from real-world application scenarios and open-source instruction following benchmarks [2, 3, 4]
38 and ask human annotators to create new complex instructions based on the provided reference
39 instructions.

40 2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses
41 or sensors, manual human curation, software programs, software APIs)? The procedure used
42 to collect the dataset was manual human curation.

43 3. Who was involved in the data collection process (e.g., students, crowd workers, contractors),
44 and how were they compensated (e.g., how much were crowd workers paid)? We provide
45 detailed answers in Section 5 of Supplementary Materials.

46 4. Over what timeframe was the data collected? The final version of the dataset was constructed
47 in April 2024.

48 1.1.4 Use

49 1. Has the dataset been used for any tasks already? No.

50 2. Is there a repository that links to any or all papers or systems that use the dataset? No.

51 3. Are there tasks for which the dataset should not be used? No.

52 1.1.5 Distribution

53 1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution,
54 organization) on behalf of which the dataset was created? Yes.

55 2. How will the dataset be distributed (e.g., tarball on website, API, GitHub)? The dataset can
56 be downloaded on GitHub after the paper is published.

57 3. Will the dataset be distributed under a copyright or other intellectual property (IP) license,
58 and/or under applicable terms of use (ToU)? The dataset is distributed under CC BY 4.0. The
59 evaluation code is distributed under the MIT license.

60 4. Have any third parties imposed IP-based or other restrictions on the data associated with
61 the instances? No.

62 5. Do any export controls or other regulatory restrictions apply to the dataset or to individual
63 instances? No.

64 1.1.6 Maintenance

65 1. Who will be supporting/hosting/maintaining the dataset? The authors of this paper.

66 2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
67 Since we choose to submit double-blind, specific information will be disclosed after the paper is
68 published.

69 3. Is there an erratum? No.

70 4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete
71 instances)? Yes.

72 5. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism
73 for them to do so? Yes, they can contact the authors of this paper. Since we choose to submit
74 double-blind, specific information will be disclosed after the paper is published.

75 1.2 Usage Method

76 The access method for COMPLEXBENCH and the method for evaluating LLMs using COMPLEXBENCH can
77 be found in the file **README.md** included in our submitted code.

78 2 Author Statement and License

79 COMPLEXBENCH is distributed under CC BY 4.0. The evaluation code of COMPLEXBENCH is distributed
80 under the MIT license. We will bear all responsibility in case of violation of rights, etc.

81 3 Hosting, Licensing, and Maintenance Plan

82 We will use GitHub ¹ as our hosting platform to ensure that COMPLEXBENCH is always accessible.
83 After the paper is published, we will continuously supplement and update COMPLEXBENCH based on
84 the performance of LLMs.

85 4 Detailed Information about NeurIPS Paper Checklist

86 4.1 Experiments Compute Resources

87 In our experiment, the computing resources we mainly use are inference of open-source models and
88 API calls of closed-source models. For the inference of open-source models, we use the vllm [5]
89 framework to generate results for a total of 10 open-source models on 1150 samples of COMPLEXBENCH.
90 For 2 LLMs of the 70B scale, we use 4 A100 GPUs for inference, and for the other 8 LLMs, we
91 use 1 A100 GPU for inference, with each model’s inference time of approximately 6 minutes. For
92 the API calls of closed-source models, our main cost is using the GPT-4-1106 [6] API for automatic
93 evaluation. In total, we need to evaluate the generation results of 15 LLMs. For the result of each
94 LLM, we need to evaluate 5,293 scoring questions separately, with the input token length of a single
95 call of approximately 2,500. In addition, there are some costs associated with the inference of
96 closed-source models generating results on COMPLEXBENCH. The total cost is approximately 2,500\$.

97 4.2 Broader Impacts

98 All data in ComplexBench has been carefully and manually reviewed to ensure that it does not
99 contain any private information or other safety issues, which minimizes the potential negative social
100 impact it may cause. We believe that COMPLEXBENCH will be a useful benchmark in the future for
101 evaluating the complex instruction-following abilities of LLMs, and for promoting further research
102 on the instruction-following of LLMs.

103 4.3 Safeguards

104 We collect reference instructions from real-world application scenarios and open-source instruction-
105 following benchmarks [2, 3, 4] and manually modify the reference instructions to construct COM-
106 PLEXBENCH. All data in COMPLEXBENCH has been carefully manually reviewed to ensure that it does
107 not contain any privacy information or other safety issues.

108 4.4 Licenses for Existing Assets

109 COMPLEXBENCH utilizes existing datasets and models in data construction and experiments. For
110 datasets, we used parts of the following three datasets, all of which have been adequately cited in our
111 paper:

- 112 • **IFeval** [2], which is distributed under the Apache-2.0 license.
- 113 • **FollowBench** [3], which is distributed under the Apache-2.0 license.
- 114 • **InfoBench** [4], which is distributed under the MIT license.

¹<https://github.com/>

115 For models, we used the following five closed-source LLMs for our experiments: GPT-4-1106 [6],
116 Claude-3-Opus [7], GLM-4 [8], ERNIEBot-4 ², GPT-3.5-Turbo-1106 [9], as well as the following
117 open-source models:

- 118 • **Qwen1.5-7B/14B/72B-Chat**^{3,4,5} [10], which are covered by the tongyi-qianwen license⁶.
- 119 • **Llama3-8B/70B-Instruct**^{7,8} [11], which are distributed under the llama3 license⁹.
- 120 • **InternLM2-7B/20B-Chat**^{10,11} [12], which are distributed under the Apache-2.0 license.
- 121 • **Baichuan2-13B-Chat**¹² [13], which is distributed under the Apache-2.0 license.
- 122 • **ChatGLM3-6B**¹³ [14], which is distributed under the Apache-2.0 license.

123 All of these models have been adequately cited in our paper.

124 4.5 New Assets

125 Our paper includes a new dataset COMPLEXBENCH designed to evaluate the complex instruction-
126 following abilities of LLMs. The detailed documentation of this dataset can be referred to in Section
127 1 of Supplementary Materials. CC BY 4.0 is used for COMPLEXBENCH.

128 5 Detailed Information about Human Annotation

129 We recruited 12 college students for data annotation of COMPLEXBENCH, and the total labor cost is
130 approximately 2800\$. We will provide the guidelines for each annotation task as follows.

131 5.1 Guidelines for Data Annotation

132 This section corresponds to Section 4.1 (Data Annotation and Validation) in our paper, where the
133 guidelines for data annotation are shown in Table 1, and Table 2 presents the English translation ver-
134 sion. In this annotation task, annotators will be provided with reference instructions, the requirements
135 of the minimum number of constraint dimensions in each constraint type, and the minimum number
136 of composition types. Annotators are instructed to construct new complex instructions based on the
137 reference instructions while annotating all the constraint dimensions and composition types within
138 the newly constructed instructions. Then, they are also required to annotate scoring questions for the
139 newly constructed instructions, and the task type of the newly constructed instructions.

140 5.2 Guidelines for Selection Branch Expansion

141 This section corresponds to Section 4.1 (Selection Branch Expansion) in our paper, where the
142 guidelines for selection branch expansion are shown in Table 3, and Table 4 presents the English
143 translation version. In this annotation task, instructions with *Selection* that were annotated in the
144 above task will be provided to the annotators. Annotators are instructed to modify the selection
145 conditions of the original instructions and construct several new instructions to cover all the different
146 selection branches apart from the original instructions. For each new instruction, all the information
147 required by the above annotation task needs to be annotated.

²<https://yiyian.baidu.com/>

³<https://huggingface.co/Qwen/Qwen1.5-7B-Chat>

⁴<https://huggingface.co/Qwen/Qwen1.5-14B-Chat>

⁵<https://huggingface.co/Qwen/Qwen1.5-72B-Chat>

⁶<https://huggingface.co/Qwen/Qwen1.5-72B-Chat/blob/main/LICENSE>

⁷<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

⁸<https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct>

⁹<https://huggingface.co/meta-llama/Meta-Llama-3-8B/blob/main/LICENSE>

¹⁰<https://huggingface.co/internlm/internlm-chat-7b>

¹¹<https://huggingface.co/internlm/internlm-chat-20b>

¹²<https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat>

¹³<https://huggingface.co/THUDM/chatglm3-6b>

下面将给你提供一条参考指令，四个约束类型各自的约束维度数量要求，以及组合方式数量要求，请你依次完成如下标注任务。

1. 基于参考指令构造出一条新的复杂指令，要求该指令中包含的各约束类型的约束维度数量大于或等于要求数量，该指令中包含的各组合方式的数量大于或等于要求数量。你也可以不参考参考指令从头开始编写。
2. 标注新构造的复杂指令的任务类型，在十个类型中选择最接近的一类。
3. 标注新构造的复杂指令中包含的所有约束维度、组合方式。
4. 标注新构造的复杂指令的得分问题。请针对该指令中的每一个约束维度、组合方式，分别设计一个可以用“是/否”回答的得分问题判定其是否得到满足。
5. 标注得分问题之间的依赖关系。我们规定，对于链式组合方式，后续任务的所有得分问题依赖于判断前序任务是否完成的得分问题；对于选择组合方式，判断正确分支执行情况的所有得分问题依赖于判断正确分支是否被选择的得分问题。在标注每个得分问题时，同时需要标注其依赖于哪些得分问题。

在构造复杂指令时，必须遵循如下三个总体原则：

1. 合理性：指令必须没有歧义，有相对明确正确答案。
2. 约束有效性：指令中包含的每一个约束维度，均应该对输出产生实质影响。
3. 难度：指令必须是足够困难的，原则上应该比参考指令更为复杂。

[参考指令]

{reference_instruction}

[任务要求]

词级约束最少数量：{number_of_lexical_constraints}

格式约束最少数量：{number_of_format_constraints}

语义约束最少数量：{number_of_semantic_constraints}

整体约束最少数量：{number_of_utility_constraints}

并列组合最少数量：{number_of_And}

链式组合最少数量：{number_of_Chain}

选择组合最少数量：{number_of_Selection}

请你根据参考指令构造出新的复杂指令：{newly_constructed_instruction}

请选择构造指令的任务分类，从以下十项中选择一项：{task_type_of_newly_constructed_instruction}

- A. 基本能力 B. 中文理解 C. 综合问答 D. 实用文本写作 E. 创意写作 F. 专业文本写作 G. 个性化写作 H. 逻辑推理 I. 角色扮演 J. 专业能力

请选择构造指令中包含的所有词级约束，多选，一个选项可以选择多次：{lexical_constraints_in_newly_constructed_instruction}

- A. 输入词匹配 B. 输出关键词

请选择构造指令中包含的所有格式约束，多选，一个选项可以选择多次：{format_constraints_in_newly_constructed_instruction}

- A. Json格式 B. Markdown格式 C. 分点格式 D. 标点格式 E. 输出长度 F. 开头格式 G. 结尾格式 H. 基于模板格式

请选择构造指令中包含的所有语义约束，多选，一个选项可以选择多次：{semantic_constraints_in_newly_constructed_instruction}

- A. 语言风格 B. 角色属性 C. 话题 D. 情感

请选择构造指令中包含的所有整体约束，多选，一个选项可以选择多次：{utility_constraints_in_newly_constructed_instruction}

- A. 目标语言 B. 支持性 C. 连贯性 D. 事实正确性 E. 满足用户需求

请选择构造指令中包含的所有组合方式，多选，一个选项可以选择多次：{composition_types_in_newly_constructed_instruction}

- A. 并列 B. 链式 C. 选择

请标注构造指令的所有评分问题，同时标注出其评测的约束维度/组合方式，每个得分问题参考如下例子编写：

1. 模型生成的文章语言是否为英文？（目标语言）

{scoring_questions_for_newly_constructed_instruction}

请标注评分问题之间的依赖关系，每个得分问题参考如下例子编写（没有依赖关系则不必编写）：

4. 模型回复字数是否在300字左右？（输出长度，依赖于1）

{dependencies_of_scoring_questions}

Table 1: Guidelines for data annotation. The blue part is the information provided to the annotators, and the red part is content that requires the annotators to make annotations.

148 5.3 Guidelines for Overall Preference Annotation

149 This section corresponds to Section 5.1 (Agreement Evaluation) in our main paper, where the
150 guidelines for overall preference annotation and their English translation are shown in Table 5. Given
151 an instruction and two model responses (denoted as A and B), the human annotators are instructed to
152 compare the quality and choose from 3 options, namely A better than B, tie, and B better than A.

153 5.4 Guidelines for Scoring Questions Verification

154 This section corresponds to Section 5.1 (Agreement Evaluation) in our main paper, where the
155 guidelines for scoring questions verification and their English translation are shown in Table 6. Given
156 an instruction and a corresponding model response, as well as a scoring question for the instruction,

157 human annotators are instructed to judge whether the requirements of the scoring question are satisfied
158 by the model response.

159 5.5 Guidelines for Instruction Decomposition

160 This section corresponds to Section 5.2.3 (Decomposition of instructions with composition types) in
161 our main paper, where the guidelines for instruction decomposition and their English translation are
162 shown in Table 7. Given an instruction containing composition types, human annotators are instructed
163 to decompose the instruction based on composition types (e.g., *Chain* into sequential tasks, *Selection*
164 into selection and execution branches, while *And* remains intact) and split the scoring questions of
165 original instructions into corresponding decomposed instructions.

166 References

- 167 [1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach,
168 Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):
169 86–92, 2021.
- 170 [2] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and
171 Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*,
172 2023.
- 173 [3] Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang,
174 Qun Liu, and Wei Wang. Followbench: A multi-level fine-grained constraints following benchmark for
175 large language models. *arXiv preprint arXiv:2310.20410*, 2023.
- 176 [4] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu,
177 Fei Liu, Pengfei Liu, and Dong Yu. Infobench: Evaluating instruction following ability in large language
178 models. *arXiv preprint arXiv:2401.03601*, 2024.
- 179 [5] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.
180 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving
181 with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*,
182 2023.
- 183 [6] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- 184 [7] Anthropic. Introducing the next generation of claude, 2024. URL <https://www.anthropic.com/news/claude-3-family>.
- 186 [8] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi
187 Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*,
188 2022.
- 189 [9] OpenAI. Introducing chatgpt, 2022. URL <https://openai.com/blog/chatgpt>.
- 190 [10] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han,
191 Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- 192 [11] AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- 194 [12] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi
195 Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.
- 196 [13] Baichuan-Inc. Baichuan 2. Online, August 1 2023. URL <https://github.com/baichuan-inc/Baichuan2>.
- 198 [14] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General
199 language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting
200 of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

Below, a reference instruction, the requirements of the minimum number of constraint dimensions in each constraint type, and the minimum number of composition types will be provided. Please complete the following annotation tasks in order.

1. Construct a new complex instruction based on the reference instruction, ensuring that the number of constraint dimensions in each constraint type within the instruction is greater than or equal to the requirements, and the number of composition types within the instruction is greater than or equal to the requirements. You may also create the new complex instruction from scratch without referencing the provided reference instruction.
2. Annotate the task type of the newly constructed complex instruction, choosing the closest type from the ten options provided.
3. Annotate all the constraint dimensions and composition types within the newly constructed complex instruction.
4. Annotate the scoring questions for the newly constructed complex instruction. Please design a "yes/no" question for each constraint dimension and composition type to verify if it is satisfied.
5. Annotate the dependencies of scoring questions. Specifically, for *Chain*, all the scoring questions of the subsequent task depend on the answers to those of the preceding task. And for *Selection*, all the scoring questions of the selection branch depend on whether the correct selection branch is selected. When annotating each scoring question, please also annotate which scoring questions it depends on (if any).

When constructing complex instructions, you should adhere to the following three general principles:

1. Clarity & Reasonableness: The instruction should be easy to understand, unambiguous, and realistic, with at least one reasonable answer.
2. Validity of Constraints: Every constraint within the instruction should substantially influence the output.
3. Complexity & Difficulty: The instruction should be challenging for most LLMs and be capable of distinguishing the complex instruction-following abilities of different LLMs.

[Reference Instruction]

{reference_instruction}

[Task Requirements]

The minimum number of lexical constraints: {number_of_lexical_constraints}

The minimum number of format constraints: {number_of_format_constraints}

The minimum number of semantic constraints: {number_of_semantic_constraints}

The minimum number of utility constraints: {number_of_utility_constraints}

The minimum number of *And*: {number_of_And}

The minimum number of *Chain*: {number_of_Chain}

The minimum number of *Selection*: {number_of_Selection}

Please construct a new complex instruction based on the reference instruction: {newly_constructed_instruction}

Please choose the task category for the constructed instruction from the following ten options: {task_type_of_newly_constructed_instruction}

A. Fundamental Language Ability B. Advanced Chinese Understanding C. Open-ended Questions D. Practical Writing E. Creative Writing F. Professional Writing G. Custom Writing H. Logical Reasoning I. Task-oriented Role Play J. Professional Knowledge

Please choose all lexical constraints within the constructed instruction, multiple selections are allowed, and an option can be chosen more than once: {lexical_constraints_in_newly_constructed_instruction}

A. Word Matching B. Keywords

Please choose all format constraints within the constructed instruction, multiple selections are allowed, and an option can be chosen more than once: {format_constraints_in_newly_constructed_instruction}

A. Json Format B. Markdown Format C. Bullets Format D. Punctuation E. Length F. Start with G. End with H. Template

Please choose all semantic constraints within the constructed instruction, multiple selections are allowed, and an option can be chosen more than once: {semantic_constraints_in_newly_constructed_instruction}

A. Language Style B. Personalization C. Topic D. Sentiment

Please choose all utility constraints within the constructed instruction, multiple selections are allowed, and an option can be chosen more than once: {utility_constraints_in_newly_constructed_instruction}

A. Target Language B. Supportiveness C. Consistency D. Factuality E. Helpfulness

Please choose all composition types within the constructed instruction, multiple selections are allowed, and an option can be chosen more than once: {composition_types_in_newly_constructed_instruction}

A. *And* B. *Chain* C. *Selection*

Please annotate all scoring questions for the constructed instruction, and indicate the constraint dimensions/composition types they evaluate. Each scoring question should be formatted as follows:

1. Is the language of the article generated by the model in English? (Target language)

{scoring_questions_for_newly_constructed_instruction}

Please annotate the dependencies of scoring questions. Each scoring question should be formatted as follows (no need to write if there is no dependency):

4. Is the number of words in the model's response more than 300? (Length, depends on 1)

{dependencies_of_scoring_questions}

Table 2: Guidelines for data annotation (translated into English). The blue part is the information provided to the annotators, and the red part is content that requires the annotators to make annotations.

下面将为你提供一条包含选择逻辑的指令，请你保持该指令中所有的选择分支不变，仅修改该指令中选择函数的选择条件，构造若干条新指令，覆盖与原指令不同的所有选择分支。例如，对于单层的选择逻辑，选择函数有M个不同的取值，则你应该构造M-1条新指令，改变选择条件以覆盖选择函数与原指令不同的所有取值。

在构造出新指令后，和数据构造任务相同，你需要标注出新构造指令中的任务类型，包含的所有约束维度、组合方式，并编写新构造指令的评分问题和其互相之间的依赖关系。

[原指令]
{instruction}

请你填写根据原指令，修改选择条件构造的第一条新指令： {newly_constructed_instruction_1}

请选择构造指令的任务分类，从以下十项中选择一项： {task_type_of_newly_constructed_instruction_1}

A. 基本能力 B. 中文理解 C. 综合问答 D. 实用文本写作 E. 创意写作 F. 专业文本写作 G. 个性化写作 H. 逻辑推理 I. 角色扮演 J. 专业能力

请选择构造指令中包含的所有词级约束，多选，一个选项可以选择多次： {lexical_constraints_in_newly_constructed_instruction_1}

A. 输入词匹配 B. 输出关键词

请选择构造指令中包含的所有格式约束，多选，一个选项可以选择多次： {format_constraints_in_newly_constructed_instruction_1}

A. Json格式 B. Markdown格式 C. 分点格式 D. 标点格式 E. 输出长度 F. 开头格式 G. 结尾格式 H. 基于模板格式

请选择构造指令中包含的所有语义约束，多选，一个选项可以选择多次： {semantic_constraints_in_newly_constructed_instruction_1}

A. 语言风格 B. 角色属性 C. 话题 D. 情感

请选择构造指令中包含的所有整体约束，多选，一个选项可以选择多次： {utility_constraints_in_newly_constructed_instruction_1}

A. 目标语言 B. 支持性 C. 连贯性 D. 事实正确性 E. 满足用户需求

请选择构造指令中包含的所有组合方式，多选，一个选项可以选择多次： {composition_types_in_newly_constructed_instruction_1}

A. 并列 B. 链式 C. 选择

请标注构造指令的所有评分问题，同时标注出其评测的约束维度/组合方式，每个得分问题参考如下例子编写：

1. 模型生成的文章语言是否为英文？（目标语言）

{scoring_questions_for_newly_constructed_instruction_1}

请标注评分问题之间的依赖关系，每个得分问题参考如下例子编写（没有依赖关系则不必编写）：

4. 模型回复字数是否在300字左右？（输出长度，依赖于1）

{dependencies_of_scoring_questions}

.....

与以上格式相同，请你填写根据原指令，修改选择条件构造的第n条新指令： {newly_constructed_instruction_n}

.....

Table 3: Guidelines for selection branch expansion. The blue part is the information provided to the annotators, and the red part is content that requires the annotators to make annotations.

Below, an instruction containing *Selection* will be provided. Please keep all selection branches unchanged, and only modify the selection condition based on the selection function to construct multiple new instructions, covering all selection branches different from the original instruction. For example, for single-layer *Selection*, if the selection function has M different values, you should construct M-1 new instructions, changing the selection conditions to cover all values different from the original instruction.

After constructing the new instructions, similar to the data annotation task, you need to annotate the task types of the instructions and all constraint dimensions and composition types within the instructions. Furthermore, you should annotate the scoring questions for the newly constructed instructions and their dependencies.

[Original Instruction]
{instruction}

Please annotate the first new instruction by modifying the selection conditions of the original instruction: {newly_constructed_instruction_1}
Please choose the task category for the constructed instruction from the following ten options: {task_type_of_newly_constructed_instruction_1}
A. Fundamental Language Ability B. Advanced Chinese Understanding C. Open-ended Questions D. Practical Writing E. Creative Writing F. Professional Writing G. Custom Writing H. Logical Reasoning I. Task-oriented Role Play J. Professional Knowledge

Please choose all lexical constraints within the constructed instruction, multiple selections are allowed, and an option can be chosen more than once: {lexical_constraints_in_newly_constructed_instruction_1}
A. Word Matching B. Keywords

Please choose all format constraints within the constructed instruction, multiple selections are allowed, and an option can be chosen more than once: {format_constraints_in_newly_constructed_instruction_1}
A. Json Format B. Markdown Format C. Bullets Format D. Punctuation E. Length F. Start with G. End with H. Template

Please choose all semantic constraints within the constructed instruction, multiple selections are allowed, and an option can be chosen more than once: {semantic_constraints_in_newly_constructed_instruction_1}
A. Language Style B. Personalization C. Topic D. Sentiment

Please choose all utility constraints within the constructed instruction, multiple selections are allowed, and an option can be chosen more than once: {utility_constraints_in_newly_constructed_instruction_1}
A. Target Language B. Supportiveness C. Consistency D. Factuality E. Helpfulness

Please choose all composition types within the constructed instruction, multiple selections are allowed, and an option can be chosen more than once: {composition_types_in_newly_constructed_instruction_1}
A. *And* B. *Chain* C. *Selection*

Please annotate all scoring questions for the constructed instruction, and indicate the constraint dimensions/composition types they evaluate. Each scoring question should be formatted as follows:
1. Is the language of the article generated by the model in English? (Target language)
{scoring_questions_for_newly_constructed_instruction_1}

Please annotate the dependencies of scoring questions. Each scoring question should be formatted as follows (no need to write if there is no dependency):
4. Is the number of words in the model's response more than 300? (Length, depends on 1)
{dependencies_of_scoring_questions}

.....

In the same format as above, please annotate the nth new instruction constructed by modifying the selection conditions of the original instruction: {newly_constructed_instruction_n}
.....

Table 4: Guidelines for selection branch expansion (translated into English). The blue part is the information provided to the annotators, and the red part is content that requires the annotators to make annotations.

下面将为你提供一条指令和对应的两个模型回复a、b，请你判断模型回复a、b哪个更好地遵循了指令的要求，质量更高，给出win, lose, tie的标注（win表示回复a更好）。

任务细则

1. 如果两个回复均质量很低，如完全没有理解指令要求，答非所问等情况，可以标注为tie，不需要进行过于细致地区分，但并非仅在此情况下才可以标注tie。
2. 不应该过于关注回复长度，遵循指令要求更为重要。

[指令]

{instruction}

[回复 A]

{response_a}

[回复 B]

{response_b}

你的选择是: {option}

A. win

B. tie

C. lose

简单叙述选择的理由: {explanation}

Below, an instruction and two corresponding model responses, a and b, will be provided. Please judge which model response better follows the instruction's requirements and is of higher quality, and choose 'win', 'lose', or 'tie' ('win' indicates response a is better).

Task Details

1. If both responses are of low quality, such as completely misunderstanding the instruction's requirements or being irrelevant, you can choose 'tie' and there is no need for detailed distinction, but 'tie' is not limited to this situation only.
2. Do not overly focus on the length of the responses. Following the requirements of instruction is more important.

[Instruction]

{instruction}

[Response A]

{response_a}

[Response B]

{response_b}

Your choice is: {option}

A. win

B. tie

C. lose

Briefly state the reason for your choice: {explanation}

Table 5: Guidelines for overall preference annotation and their English translation. The blue part is the information provided to the annotators, and the red part is content that requires the annotators to make annotations.

下面将给你提供一条指令，对应的一个模型回复，以及该指令的一个得分问题，你的任务是判断模型回复是否满足得分问题要求，标注“是”或者“否”。

任务细则

1. “是”的定义必须是完全充分地完成了得分点要求，任何存在错误、不明确、无法判断地回答都应该判定为“否”。不存在“基本上正确”，“部分条件下正确”的说法，这些情况均标注为“否”。如果模型回复中不存在得分问题所评判的对象，也应该标注为“否”。
2. 只需要考虑该得分点是否完全被模型回复满足，而不需要考虑整个输入指令是否被模型输出满足。

{in-context examples}

[指令]

{instruction}

[模型回复]

{model_response}

[得分问题]

{scoring_question}

你的选择是: {option}

A. 是

B. 否

Below, an instruction, a corresponding model response, and a scoring question for the instruction will be provided. Your task is to verify whether the model response meets the requirements of the scoring question, and choose "Yes" or "No."

Task Details

1. A "Yes" must indicate that the scoring point has been fully and sufficiently satisfied. Any response that contains errors, ambiguity, or cannot be judged should be labeled as "No". There is no such thing as "basically correct" or "correct under certain conditions". These should all be labeled as "No". If the model response does not contain the object to be evaluated by the scoring question, it should also be labeled as "No".
2. Please only consider whether the scoring point has been fully satisfied by the model response, without the need to consider whether the entire instruction has been satisfied by the model response.

{in-context examples}

[Instruction]

{instruction}

[Model Response]

{model_response}

[Scoring Question]

{scoring_question}

Your choice is: {option}

A. Yes

B. No

Table 6: Guidelines for scoring questions verification and their English translation. The blue part is the information provided to the annotators, and the red part is content that requires the annotators to make annotations.

下面将给你提供一条指令和该指令对应的得分问题。请你根据该指令中包含的组合方式，将该指令拆分为多条原子指令，要求每条原子指令不含有除了并列之外的其他组合方式。

请将原指令拆分为多条多轮追问形式的原子指令。对于包含链式组合方式的指令，请按每个子任务分别拆分，如果每个子任务中仍然包含组合方式，需要继续进行拆分；对于包含选择组合方式的指令，拆分后其中一条指令为选择正确分支，另一部分指令为执行正确分支，正确分支中仍包含组合方式的，需要继续进行拆分。如果该指令难以拆分的，可以选择跳过该指令不进行拆分。

在你完成指令拆分后，还需要将原指令的得分问题分配到对应的原子指令中。禁止增添或修改得分问题，仅允许在不改变原意的情况下，为增加流畅性对得分问题进行微量修改。

{in-context examples}

[原指令]

{instruction}

[得分问题]

{scoring_questions}

请填写你拆分原指令得到的第一条指令： {sub_instructions_0}

请填写该指令对应的得分问题，必须选取原指令得分问题中对应的连续片段： {scoring_questions_for_sub_instructions_0}

请填写你拆分原指令得到的第二条指令（没有则不必填写）： {sub_instructions_1}

请填写该指令对应的得分问题，必须选取原指令得分问题中对应的连续片段： {scoring_questions_for_sub_instructions_1}

.....

请填写你拆分原指令得到的第n条指令（没有则不必填写）： {sub_instructions_n}

请填写该指令对应的得分问题，必须选取原指令得分问题中对应的连续片段： {scoring_questions_for_sub_instructions_n}

Below is an instruction and its corresponding scoring questions. Please decompose the instruction into multiple atomic instructions according to the composition types it contains, ensuring that each atomic instruction does not contain any composition types other than *And*.

Please decompose the original instruction into multiple atomic instructions in the form of a multi-turn interactive. For instructions containing *Chain*, decompose them by each task separately. If each task still contains composition types other than *And*, continue to decompose further. For instructions containing *Selection*, one of the decomposed instructions should be the choice of the correct branch, and the other part should be the execution of the correct branch. If the correct branch still contains composition types other than *And*, continue to decompose further. If an instruction is difficult to decompose, you may choose to skip and not decompose it.

After you have completed the decomposition of the instructions, you need to assign the scoring questions of the original instruction to the corresponding atomic instructions. Adding or deleting scoring questions is prohibited. Only minimal modifications to increase fluency without changing the original meaning are allowed.

{in-context examples}

[Original Instruction]

{instruction}

[Scoring Questions]

{scoring_questions}

Please annotate the first atomic instruction obtained by decomposing the original instruction: {sub_instructions_0}

Please annotate the corresponding scoring question for this instruction, making sure to select the continuous segment from scoring questions of the original instruction: {scoring_questions_for_sub_instructions_0}

Please annotate the second atomic instruction obtained by decomposing the original instruction (if any): {sub_instructions_1}

Please annotate the corresponding scoring question for this instruction, making sure to select the continuous segment from scoring questions of the original instruction: {scoring_questions_for_sub_instructions_1}

.....

Please annotate the nth atomic instruction obtained by decomposing the original instruction (if any): {sub_instructions_n}

Please annotate the corresponding scoring question for this instruction, making sure to select the continuous segment from scoring questions of the original instruction: {scoring_questions_for_sub_instructions_n}

Table 7: Guidelines for instruction decomposition and their English translation. The blue part is the information provided to the annotators, and the red part is content that requires the annotators to make annotations.