



# OpenING: Open INstruction Generation

## 指令微调数据生成工具

王越<sup>1</sup>, 祁奕诚<sup>1</sup>, 李俊涛<sup>1</sup>, 张民<sup>1</sup>

<sup>1</sup>苏州大学, 计算机科学与技术学院

{lywangnlp,ycqi}@stu.suda.edu.cn,{ljt,minzhang}@suda.edu.cn

### 工具特点

· 未使用闭源模型输出结果:

ChatGPT、Claude、Bard等闭源模型禁止利用其模型输出开发同类模型, 本工具全流程未使用ChatGPT等闭源模型的输出结果, 没有潜在风险

· 支持中英文多种任务指令生成:

通过优化训练数据分布, 本工具既能够生成文本分类, 实体抽取等自然语言理解任务指令, 也能够生成故事、诗歌生成等文本生成任务指令

· 自动过滤高质量指令数据:

根据用户输入生成多个候选指令后, 本工具可以自动过滤出最高质量的指令, 并将最后的指令-输出对自动保存为json格式, 减少用户操作

### 流程介绍

· 基座模型: Baichuan-7B (<https://huggingface.co/baichuan-inc/Baichuan-7B>)

· 训练数据来源(共36511条): Super-NaturalInstructions (<https://github.com/allenai/natural-instructions>)

Dolly (<https://huggingface.co/datasets/databricks/databricks-dolly-15k>)

Firefly (<https://huggingface.co/datasets/YeungNLP/firefly-train-1.1M>)

影响指令微调数据质量的三大因素:

输出回复的可靠性, 输入指令的多样性, 输入指令和输出回复的对齐性

本工具将用户输入的文本作为输出回复, 生成对应的输入指令。因此输出回复的可靠性由用户来控制, 如果从语料库中采样文本输入到本工具中, 那么回复的可靠性就能够得到保证。此外, 本工具先针对一个选中文本生成多个候选指令, 再使用过滤策略来选择最合适的指令, 这也确保了输入指令的多样性, 以及输入指令和输出回复的对齐性。

步骤一: 训练指令微调数据生成模型



步骤二: 生成候选指令微调数据



步骤三: 过滤候选指令微调数据

图一 OpenING指令微调数据生成流程示意图

### 生成样例

OpenING: Open INstruction Generation指令微调数据生成工具

Selected output to generate corresponding instruction

happy

Temperature0.3

Top p0.9

Top k40

Return Sequences Number5

Max tokens256

ClearSubmit

Candidate Instruction 1:

In this task, you're given a context, a sentence, and a character. The sentence describes an action or job of the given character. Also, the context provides more information about the sentence or the character. Your task is to return one of the emotions which are expressed by the Character in the given sentence. For that you can use the Context; however, Context is not provided in all the inputs. Also, in some inputs, there can be zero emotion; for that return 'None'. ### Input: Context: Mary was a new mother. Sentence: She was very happy and excited to be a mother. Character: Mary.

Candidate Instruction 2:

In this task, you're given a context, a sentence, and a character. The sentence describes an action or job of the given character. Also, the context provides more information about the sentence or the character. Your task is to return one of the emotions which are expressed by the Character in the given sentence. For that you can use the Context; however, Context is not provided in all the inputs. Also, in some inputs, there can be zero emotion; for that return 'None'. ### Input: Context: I was happy to see my old friend. Sentence: We hadn't seen each other in years. Character: I (myself)

Candidate Instruction 3:

In this task, you're given a context, a sentence, and a character. The sentence describes an action or job of the given character. Also, the context provides more information about the sentence or the character. Your task is to return one of the emotions which are expressed by the Character in the given sentence. For that you can use the Context; however, Context is not provided in all the inputs. Also, in some inputs, there can be zero emotion; for that return 'None'. ### Input: Context: I had a great day with my friends. Sentence: We went to the park and had a picnic. Character: I (myself)

Candidate Instruction 4:

In this task, you're given a context, a sentence, and a character. The sentence describes an action or job of the given character. Also, the context provides more information about the sentence or the character. Your task is to return one of the emotions which are expressed by the Character in the given sentence. For that you can use the Context; however, Context is not provided in all the inputs. Also, in some inputs, there can be zero emotion; for that return 'None'. ### Input: Context: I had a rough day at work. Sentence: I was really mad at my coworkers. Character: I (myself)

Candidate Instruction 5:

In this task, you're given a context, a sentence, and a character. The sentence describes an action or job of the given character. Also, the context provides more information about the sentence or the character. Your task is to return one of the emotions which are expressed by the Character in the given sentence. For that you can use the Context; however, Context is not provided in all the inputs. Also, in some inputs, there can be zero emotion; for that return 'None'. ### Input: Context: Cameron was a paramedic. He was called to a house to perform CPR on a man. Sentence: The man was talking to Cameron and seemed to be in good spirits. Character: Cameron.

Paired instruction and output in json format

[{"instruction":"In this task, you're given a context, a sentence, and a character. The sentence describes an action or job of the given character. Also, the context provides more information about the sentence or the character. Your task is to return one of the emotions which are expressed by the Character in the given sentence. For that you can use the Context; however, Context is not provided in all the inputs. Also, in some inputs, there can be zero emotion; for that return 'None'.", "output": "happy"}]

OpenING: Open INstruction Generation指令微调数据生成工具

Selected output to generate corresponding instruction

老马让小马把半口袋麦子驮到磨坊去。途中遇到一条小河，小马询问老牛能不能过河。老牛说：“水很浅，刚没小腿，能过去。”就在小马准备过河的时候，从树上跳下的松鼠却阻拦说：河水“深得很哩！昨天，我的一个伙伴就是掉在这条河里淹死的”。小马只好回家问妈妈，妈妈告诉小马：“孩子，光听别人说，自己不动脑筋，不去试试，是不行的。河水是深是浅，你去试一试，就知道了。”小马再次来到河边，小心地到了对岸，他发现原来河水既不像老牛说的那样浅，也不像松鼠说的那样深。

Temperature0.3

Top p0.9

Top k40

Return Sequences Number5

Max tokens256

ClearSubmit

Candidate Instruction 1:

给我讲个小马过河的故事

Candidate Instruction 2:

生成故事，题目：小马过河

Candidate Instruction 3:

生成故事，题目：小马过河

Candidate Instruction 4:

以小马过河为标题，生成一个故事。

Candidate Instruction 5:

生成故事，题目：哲理故事：河水很浅很浅

Paired instruction and output in json format

[{"instruction":"给我讲个小马过河的故事","output":"老马让小马把半口袋麦子驮到磨坊去。途中遇到一条小河，小马询问老牛能不能过河。老牛说：“水很浅，刚没小腿，能过去。”就在小马准备过河的时候，从树上跳下的松鼠却阻拦说：河水“深得很哩！昨天，我的一个伙伴就是掉在这条河里淹死的”。小马只好回家问妈妈，妈妈告诉小马：“孩子，光听别人说，自己不动脑筋，不去试试，是不行的。河水是深是浅，你去试一试，就知道了。”小马再次来到河边，小心地到了对岸，他发现原来河水既不像老牛说的那样浅，也不像松鼠说的那样深。”}]