按照指令任务的类型划分，可以分为传统 NLP 任务指令和通用对话指令两类。表5.1给出了部分开源指令集的汇总信息。

表 5.1 部分开源指令数据集

指令集名称	指令集大小	语言	构建方式	指令类型
Super-Natural Instructions	500 万	多语言	手动构建	NLP 任务指令
Flan2021	44 万	英语	手动构建	NLP 任务指令
pCLUE	120 万	中文	手动构建	NLP 任务指令
OpenAssistant Conversations	16.1 万	多语言	手动构建	通用对话指令
Dolly	1.5 万	英语	手动构建	通用对话指令
LIMA	1000	英语	手动构建	通用对话指令
Self-Instruct	5.2 万	英语	自动生成	通用对话指令
Alpaca_data	5.2 万	英语	自动生成	通用对话指令
BELLE	150 万	中文	自动生成	通用对话指令

传统 NLP 任务指令集：将传统的 NLP 任务使用自然语言指令的格式进行范式统一。

- Super-Natural Instructions [①]是由 Allen Institute for AI (AI2) 发布的一个指令集合。其包含 55 种语言，由 1616 个 NLP 任务、共计 500 万个任务实例组成，涵盖 76 个不同的任务类型（例如文本分类、信息提取、文本重写等）。该数据集的每个任务由“指令”和“任务实例”两部分组成，“指令”部分不仅对每个任务做了详细的描述，还提供了正、反样例以及相应的解释，“任务实例”即为属于该任务的输入-输出实例。
- Flan2021 [②]是一个由 google 发布的英文指令数据集，通过将 62 个广泛使用的 NLP 基准（如 SST-2、SNLI、AG News、MultiRC）转换为输入-输出对的方式构建而成。构建时，先手动编写指令和目标模板，再使用来自数据集的数据实例填充模板。
- pCLUE [③]是由 CLUEbenchmark 发布的，使用 9 个中文 NLP 基准数据集，按指令格式重新构建而成的中文指令集。包含的中文任务包括：单分类 tnews、单分类 iflytek、自然语言推理 ocnli、语义匹配 afqmc、指代消解-cluewsc2020、关键词识别-csl、阅读理解-自由式 c3、阅读理解-抽取式 cmrc2018、阅读理解-成语填空 chid。

通用对话指令集：更广义的自然语言任务，通过模拟人类行为提升大模型的交互性。

- OpenAssistant Conversations [④]是由 LAION 发布的人工生成、人工注释的助手风格的对话语料库，旨在促进将大语言模型与人类偏好对齐。该数据集包含 35 种不同的语言，采用众包的方式构建，由分布在 66497 个对话树中的 161443 条对话数据组成。它提供了丰富且多样

① https://github.com/allenai/natural-instructions
② https://github.com/google-research/FLAN
③ https://github.com/CLUEbenchmark/pCLUE
④ https://github.com/LAION-AI/Open-Assistant