数据集名称	领域	数量	任务类型	Prompt	数据提供者	说明	是否开源/研究使用	是否商用	脚本	Done	URL	是否同质
cmrc2018	百科	14,363	问答	问答	Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, Guoping Hu	https://github.com/ymcui/cmrc2018/blob/master/README_CN.md 专家标注的基于维基百科的中文阅读理解数据集，将问题和上下文视为正例	是	否	是	是	https://huggingface.co/datasets/cmrc2018	否
belle_2m	百科	2,000,000	指令微调	无	LianjiaTech/BELLE	belle 的指令微调数据集，使用 self instruct 方法基于 gpt3.5 生成	是	否	是	是	https://huggingface.co/datasets/BelleGroup/train_2M_CN	否
firefily	百科	1,649,399	指令微调	无	YeungNLP	Firefly（流萤） 是一个开源的中文对话式大语言模型，使用指令微调（Instruction Tuning）在中文数据集上进行调优。使用了词表裁剪、ZeRO等技术，有效降低显存消耗和提高训练效率。 在训练中，我们使用了更小的模型参数量，以及更少的计算资源。	未说明	未说明	是	是	https://huggingface.co/datasets/YeungNLP/firefly-train-1.1M	否
alpaca_gpt4	百科	48,818	指令微调	无	Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, Jianfeng Gao	本数据集是参考Alpaca方法基于GPT4得到的self-instruct数据，约5万条。	是	否	是	是	https://huggingface.co/datasets/shibing624/alpaca-zh	否
zhihu_kol	百科	1,006,218	问答	问答	wangrui6	知乎问答	未说明	未说明	是	是	https://huggingface.co/datasets/wangrui6/Zhihu-KOL	否
hc3_chinese	百科	39,781	问答	问答	Hello-SimpleAI	问答数据，包括人工回答和 GPT 回答	是	未说明	是	是	https://huggingface.co/datasets/Hello-SimpleAI/HC3-Chinese	否
amazon_reviews_multi	电商	210,000	问答 文本分类	摘要	亚马逊	亚马逊产品评论数据集	是	否	是	是	https://huggingface.co/datasets/amazon_reviews_multi/viewer/zh/train?row=8	否
mlqa	百科	85,853	问答	问答	patrickvonplaten	一个用于评估跨语言问答性能的基准数据集	是	未说明	是	是	https://huggingface.co/datasets/mlqa/viewer/mlqa-translate-train.zh/train?p=2	否
xlsum	新闻	93,404	摘要	摘要	BUET CSE NLP Group	BBC的专业注释文章摘要对	是	否	是	是	https://huggingface.co/datasets/csebuetnlp/xlsum/viewer/chinese_simplified/train?row=259	否
ocnli	口语	17,726	自然语言推理	推理	Thomas Wolf	自然语言推理数据集	是	否	是	是	https://huggingface.co/datasets/clue/viewer/ocnli	是
BQ	金融	60,000	文本分类	相似	Intelligent Computing Research Center, Harbin Institute of Technology(Shenzhen)	http://icrc.hitsz.edu.cn/info/1037/1162.htm BQ 语料库包含来自网上银行自定义服务日志的 120，000 个问题对。它分为三部分：100，000 对用于训练，10，000 对用于验证，10，000 对用于测试。 数据提供者：哈尔滨工业大学（深圳）智能计算研究中心	是	否	是	是	https://huggingface.co/datasets/shibing624/nli_zh/viewer/BQ	是
lcqmc	口语	149,226	文本分类	相似	Ming Xu	哈工大文本匹配数据集，LCQMC 是哈尔滨工业大学在自然语言处理国际顶会 COLING2018 构建的问题语义匹配数据集，其目标是判断两个问题的语义是否相同	是	否	是	是	https://huggingface.co/datasets/shibing624/nli_zh/viewer/LCQMC/train	是
paws-x	百科	23,576	文本分类	相似	Bhavitvya Malik	PAWS Wiki中的示例	是	是	是	是	https://huggingface.co/datasets/paws-x/viewer/zh/train	是
wiki_atomic_edit	百科	1,213,780	平行语义	相似	abhishek thakur	基于中文维基百科的编辑记录收集的数据集	未说明	未说明	是	是	https://huggingface.co/datasets/wiki_atomic_edits	是
chatmed_consult	医药	549,326	问答	问答	Wei Zhu	真实世界的医学相关的问题，使用 gpt3.5 进行回答	是	否	是	是	https://huggingface.co/datasets/michaelwzhu/ChatMed_Consult_Dataset	否
webqa	百科	42,216	问答	问答	suolyer	百度于2016年开源的数据集，数据来自于百度知道；格式为一个问题多篇意思基本一致的文章，分为人为标注以及浏览器检索；数据整体质量中，因为混合了很多检索而来的文章	是	未说明	是	是	https://huggingface.co/datasets/suolyer/webqa/viewer/suolyer--webqa/train?p=3	否
dureader_robust	百科	65,937	机器阅读理解 问答	问答	百度	DuReader robust旨在利用真实应用中的数据样本来衡量阅读理解模型的鲁棒性，评测模型的过敏感性、过稳定性以及泛化能力，是首个中文阅读理解鲁棒性数据集。	是	是	是	是	https://huggingface.co/datasets/PaddlePaddle/dureader_robust/viewer/plain_text/train?row=96	否
csl	学术	395,927	语料	摘要	Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Weijie Liu, Weiquan Mao and Hui Zhang	提供首个中文科学文献数据集（CSL），包含 396,209 篇中文核心期刊论文元信息 （标题、摘要、关键词、学科、门类）。CSL 数据集可以作为预训练语料，也可以构建许多NLP任务，例如文本摘要（标题预测）、 关键词生成和文本分类等。	是	是	是	是	https://huggingface.co/datasets/neuclir/csl	否
miracl-corpus	百科	4,934,368	语料	摘要	MIRACL	The corpus for each language is prepared from a Wikipedia dump, where we keep only the plain text and discard images, tables, etc. Each article is segmented into multiple passages using WikiExtractor based on natural discourse units (e.g., \n\n in the wiki markup). Each of these passages comprises a "document" or unit of retrieval. We preserve the Wikipedia article title of each passage.	是	是	是	是	https://huggingface.co/datasets/miracl/miracl-corpus	否
lawzhidao	法律	36,368	问答	问答	和鲸社区-Ustinian	百度知道清洗后的法律问答	是	是	否	是	https://www.heywhale.com/mw/dataset/5e953ca8e7ec38002d02fca7/content	否
CINLID	成语	34,746	平行语义	相似	高长宽	中文成语语义推理数据集（Chinese Idioms Natural Language Inference Dataset）收集了106832条由人工撰写的成语对（含少量歇后语、俗语等短文本），通过人工标注的方式进行平衡分类，标签为entailment、contradiction和neutral，支持自然语言推理（NLI）的任务。	是	否	否	是	https://www.luge.ai/#/luge/dataDetail?id=39	是
DuSQL	SQL	25,003	NL2SQL	SQL	百度	DuSQL是一个面向实际应用的数据集，包含200个数据库，覆盖了164个领域，问题覆盖了匹配、计算、推理等实际应用中常见形式。该数据集更贴近真实应用场景，要求模型领域无关、问题无关，且具备计算推理等能力。	是	否	否	是	https://www.luge.ai/#/luge/dataDetail?id=13	否

Zhuiyi-NL2SQL	SQL	45,918	NL2SQL	SQL	追一科技 刘云峰	NL2SQL是一个多领域的简单数据集，其主要包含匹配类型问题。该数据集主要验证模型的泛化能力，其要求模型具有较强的领域泛化能力、问题泛化能力。	是	否	否	是	https://www.luge.ai/#/luge/dataDetail?id=12	否
Cspider	SQL	7,785	NL2SQL	SQL	西湖大学 张岳	CSpider是一个多语言数据集，其问题以中文表达，数据库以英文存储，这种双语模式在实际应用中也非常常见，尤其是数据库引擎对中文支持不好的情况下。该数据集要求模型领域无关、问题无关，且能够实现多语言匹配。	是	否	否	是	https://www.luge.ai/#/luge/dataDetail?id=11	否
news2016zh	新闻	2,507,549	语料	摘要	Bright Xu	包含了250万篇新闻。新闻来源涵盖了6.3万个媒体，含标题、关键词、描述、正文。	是	是	否	是	https://github.com/brightmart/nlp_chinese_corpus	否
baike2018qa	百科	1,470,142	问答	问答	Bright Xu	含有150万个预先过滤过的、高质量问题和答案，每个问题属于一个类别。总共有492个类别，其中频率达到或超过10次的类别有434个。	是	是	否	是	https://github.com/brightmart/nlp_chinese_corpus	否
webtext2019zh	百科	4,258,310	问答	问答	Bright Xu	含有410万个预先过滤过的、高质量问题和回复。每个问题属于一个【话题】，总共有2.8万个各式话题，话题包罗万象。	是	是	否	是	https://github.com/brightmart/nlp_chinese_corpus	否
SimCLUE	百科	775,593	平行语义	相似	数据集合，请在 simCLUE 中查看	整合了中文领域绝大多数可用的开源的语义相似度和自然语言推理的数据集，并重新做了数据拆分和整理。	是	否	否	是	https://github.com/CLUEbenchmark/SimCLUE	是
Chinese-SQuAD	新闻	76,449	机器阅读理解	问答	junzeng-pluto	中文机器阅读理解数据集，通过机器翻译加人工校正的方式从原始Squad转换而来	是	否	否	是	https://github.com/pluto-junzeng/ChineseSquad	否