| 百行云游-数据汇总清单 | | | | | | |
|---|---|---|---|---|---|---|
| **序号** | **数据类型** | **数据名称** | **数据量级** | **数据交付方式和格式** | **数据描述** | **样例文件** |
| 1 | 电子图书 | 中文出版图书 | 100万本 | 数据库授权交付<br>PDF EPUB TXT 格式<br>（6：3：1） | 100万册电子图书和期刊等资源，图书类别涉及哲学<br>社会科学 政治 军师 经济 文学 历史 工业<br>综合性图书等各大类，具体见图书数据列表<br>1.正版出版社印刷电子图书，按照国家出版规则经过三审三校，专业知识质量极高；<br>2.数据库授权交付，PDF EPUB TXT<br>格式6：3：1，PDF都是原版图书排版文字格式，可以通过PDF提取工具直接提取内容<br>3.出版社授权，来源合规，交付做正版授权，授权大模型训练场景使用 | https://pan.quark.cn/s/9ef20c19f1f1 |
| 2 | 网文小说 | 中文网络小说 | 10万本 | 数据库授权交付<br>TXT 格式 | 版权说明：正版授权来源合规<br>小说类型：<br>1、女频：萌宝、玄幻言情、种田、马甲、年代、现言脑洞、宫斗宅斗、悬疑脑洞、古言脑洞、医术<br>、快穿、青春甜宠、豪门爽文、悬疑恋爱、职场婚恋、霸总<br>2、男频：科幻、都市日常、都市修真、奇幻仙侠、历史古代、战神赘婿、都市种田、传统玄幻、历史脑洞、悬疑脑洞、都市脑洞、玄幻脑洞、神医、悬疑灵异、抗战谍战、游戏体育<br>3、其他：惊悚、悬疑、职场、官场、古典、外国小说<br>、财经、历史、武侠、军事、魔幻、科幻、社会、乡土 | https://pan.quark.cn/s/2807f18819f0 |
| 2 | 题库 | K12题库数据集 | 2000万 | 数据库授权交付<br>json格式 | 覆盖范围：小学数学、小学语文、小<br>学英语、小学科数学、初中物理、初中化学、初中生物<br>、初中地理初中英语、初中语文、初中历史、初中道德与法治:初中科学、初中信息技术、化<br>学、高中生物、高中地理、高中英语、高中语文、技术<br>、信息技术。<br>题型覆盖：选择题、作图题、填空题、解答题判断题、改错题、问答题、<br>探究题简答题、:书写、其他、实验探究题单选题、多选题、单项选择题、阅读理解、单词拼写、分析说明题<br>、翻译、材料解析题、句型转默写、语言表达、信息匹配、<br>辨析题现代文阅读、双选题、辨析改错题论述题、读图说史题、完形填动多项选择题、短文改错、书面表选词填空、完成句子、<br>补全对达、料题、名句默写、单词拼<br>写、单句语法填空、选词填空、句子单词拼写、综合性学习、名著阅读、不定项选择题、单项选择、看图题、修改病句、古诗阅读、字词书写、单空题、文学类文本阅读等；<br>数据说明：试题属性包含学科、章节、知识点、题型、难度系类来源(所属地区)、能力、分值、解析等十余类;试卷属性包含试卷类型、来源、分值、难度系数、考试时间、命题人等十余类所有学科考点 | https://pan.quark.cn/s/9a2890f785d6 |
| | | 大学和职业考试题库数据集 | 1.6亿 | 数据库授权交付<br>json格式 | **覆盖范围：**大学题库，覆盖大学英语、政治、法学经济学、医学等学科;职业题库，覆盖职业教育所有相关学科，公考、金融类工程类，医学类等。<br>**题型覆盖：**选择题、作图题、填空题题、简答题、书写、其他题、阅读理解、单词拼写换、简析题、诗歌鉴赏、题、现代文阅读、双选题完形填空、多项选择题、补全对话、证明题、连词)写、单句语法填空、选词开放性试题、综合性学习图题、修改病句、古诗阅；<br>**数据说明：**优质资源丰富:近 14<br>年以来的大学职业试题试卷收入高达<br>80%步，模拟试卷收入总量达 20<br>万份以上。数据属性全:试题属性包含学科、章节、知识点、题型、难度系类来源(所属地区)、能力、分值、解析等十余类;试卷属性包含试卷类型、来源、分值、难度系数、考试时间、命题人等十余类所有学科考点覆盖率 100%。 | |
| | | 英文题库数据集 | 2000万 | 数据库授权交付<br>json格式 | 覆盖范围：Biology、Business<br>Studies、Chemistry、Computers、Geography、History、Mathematics、Physics、Science；<br>数据说明：总量2000余万，其中500-<br>600万有解析，全部数据来自英语母语国家（英美澳新印） | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | 高质量对话数据 | 中文多轮对话（机器采集） | 2000万轮 | 数据库交付 | 通识内容，Agent采集，L1（机器）加工或者L2（机器+人工）加工 | https://pan.quark.cn/s/91ef44e3c8d5 |
| | | 中文多轮对话（人工采集） | 1000万轮 | 数据库交付 | 主要包含自由对话 客服<br>会议几个领域对话，客服主要是金融/运营商/物流/二手车/电商领域，数据采集是中级人员的人类自然对话，统一采用L3L3（人工-》机器-》人工，两次人工）加工方式 | |
| 4 | 预训练基础数据（最新数据更新到24年5月左右） | 百度百科 | 2000万 | ftp库交付 | 数据量级统计到23年12月底，后续数据也有在持续更新，一般n-2个月 | https://pan.quark.cn/s/132e74a6c2db |
| | | 百度知道 | 2.6亿 | ftp库交付 | 数据量级统计到23年12月底，后续数据也有在持续更新，一般n-2个月 | |
| | | 公众号 | 30亿 | ftp库交付 | 数据量级统计到23年12月底，后续数据也有在持续更新，一般n-1个月<br>文章数据每天更新90万条，每月更新2700万条，包含个人公众号 | |
| | | 知乎 | 3亿 | ftp库交付 | 数据量级统计到23年12月底，后续数据也有在持续更新，一般n-2个月<br>2.7亿回答，2000万提问；700万专栏文章 | |
| | | 小红书 | 2.2亿 | ftp库交付 | 数据量级统计时间范围：2023.3.15-2023.12.30，后续数据也有在持续更新，一般n-2个月 | |
| | | 头条新闻 | 6000万 | ftp库交付 | 数据量级统计时间范围：2023.5-2023.12.30，后续数据也有在持续更新，一般n-2个月 | |
| | | 研究报告 | 240万 | ftp库交付 | 数据量级统计到23年12月底，后续数据也有在持续更新，一般n-2个月 | |
| | | 中文期刊/专利摘要数据 | 千万级别 | ftp库交付 | 数据量级统计到23年12月底，后续数据也有在持续更新，一般n-2个月 | |
| | | 科情头条数据 | 十万级别 | ftp库交付 | 科情头条数据是2019年10月8日至今；<br>信息来源六大类：<br>专业科技媒体25%（MIT科技评论，PNAS.刀、新智元、亿欧网..... ScienceDaily、Nature、Science,AlphaGalileo）<br>科技领先国家科技部门及高校院所40%（美国:美国能源部、美国总统科技政策办公室；英国：英国商业、创新与科技部、英国政府科学办公室；德国：德国联邦经济与技术部；法国：法国科学院；日本：日本文部科学省；韩国：韩国科技部）<br>科技智库16%（巴斯德研究所、<br>兰德公司、斯坦福国际咨询研究所、布鲁金斯学会、哈佛大学国际事务中心）<br>主要金砖国家科技部门及高校院所8%（中国科技部、印度国家科学院、巴西国家科技部）<br>大型企业咨询机构5%（IBM、微软、毕马威、波士顿咨询、德勤，普华永道、埃森哲、麦肯锡）<br>国际组织6%（欧盟、世界经合组织、联合国教科文化组织、世界银行、第三世界科学院组织）" | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 5 | 安全数据 | 价值观数据集上百万字从网络爬取后且经过专家清洗后的中文原始数据，原始数据采集自权威期刊、杂志、法律法规领域权威基础数据库等。经提取后形成大模型训练所需三元组（问题-优秀答案-不良答案） | 58109条 | 正反向问答对数据库交付 | **产品优势：**<br>1.经网信办专家经过论证及分析后形成的90+个细分维度，如专家以发展的角度论证了科技政治学细分类别的准确性，另外政治体系也从中国政治体系、政府与公共关系、中国特色、国际的政治体系等角度出发将其划分出60多个细分类别，符合社会主义核心价值观要求<br>2.包含了近1万条经过心理、政治、法律领域专家论证后的人工标注高质量调优数据<br>**产品描述：**<br>1.数据内容：中文价值观类数据<br>2.数据规模：58109条（持续增加中，其中1万条为高质量调优数据）<br>3.标注内容：涵盖心理健康21个细分维度、政治敏感60个细分维度、法律法规14个细分维度<br>4.存储格式：以json格式存储<br>5.语言：中文<br>6.数据类别：权威期刊、杂志等<br>**适用场景：**<br>基于构建的三元组数据，训练大模型，提升模型在政治敏感类的回答能力（现在很多大模型敏感问题都不回答）；心理健康类问题回答的更符合中国国情；在法律法规类的问题回答的更准确。 | https://pan.quark.cn/s/4a12bbd7e023 |
| | | 时事热点数据集上百万字从网络爬取近期实时热点新闻数据，原始数据采集自权威期刊、杂志等。经提取后形成大模型训练所需三元组（问题-优秀答案-不良答案） | 20413条 | 正反向问答对数据库交付 | 产品优势：<br>跟进国际时事热点相关的新闻报道，定期更新并增加数据量<br>产品描述：<br>1.数据内容：中文实时热点类数据<br>2.数据规模：20413条（持续增加中）<br>3.标注内容：涵盖近期国际热点新闻<br>4.存储格式：以json格式存储<br>5.语言：中文<br>6.数据类别：权威期刊、杂志、新闻等<br>适用场景：<br>基于构建的三元组数据，训练大模型，提升模型在时事热点类事件的回答更符合中国国情 | |
| | | 社会主义价值观合规负面原始数据集 | 5万条 | 数据库交付 | 数据采集于海外真实数据，经过清洗人工分类和打标签，具体数据采集加工方式和样例见样例级别 | |
| 6 | 多模态数据 | 视频数据-短视频-实拍摄影 | 20万条 | 数据库交付 | 数据分类和量级：20万实拍视频，20万模板视频，每个视频时长15秒-3分钟；视频内容质量极高，包含人物 动物 风景 美食 科技<br>古风等多种丰富场景，1600种分类，每种100个左右<br>数据质量：所有数据带配套标题、多维度标签、描述<br>数据来源：来源合规，正式版权授权交付 | https://pan.quark.cn/s/c691e9e3be85 |
| | | 视频数据-短视频-UGC | 200万条 | 数据库交付 | 数据分类和量级：UGC视频100万条，每条视频平均40秒；社会新闻类型100万条，每条视频平均3分钟。<br>数据质量：提供视频链接，省市区行政码，视频标题，视频时长等信息。而且有对应的文字信息说明，相应的标签都有，这数据对训练文生视频模型，有天然优势。另外数据每天可持续生产。 | https://pan.quark.cn/s/c691e9e3be85 |
| | | 视频数据-长视频-纪录片 | 1万小时 | 数据库交付 | 数据分类和量级：包含历史 美食 科技 文学等多种类型记录片，每个纪录片10-50集不等，单级时长25-60分钟，数据内容素材丰富，质量极高。<br>数据质量：所有数据带配套标题、描述、字幕<br>数据来源：来源合规，正式版权授权交付 | 链接: https://pan.baidu.com/s/1WEPFZ9op-oULvUlcT09a1Q<br>提取码: bxyy |
| | | 视频数据-长视频-影视剧文化节目 | 6000小时 | 数据库交付 | 数据分类和量级：各类图书馆影视剧和文化节目视频，总共9506集。<br>数据质量：所有数据带配套标题、描述、字幕<br>数据来源：来源合规，正式版权授权交付 | https://pan.quark.cn/s/c691e9e3be85 |
| | | 视频数据-采集服务 | 100TB | 数据库交付 | 目前已经采集到的数据约有10万个小时，存量总共有100TB+,主要是目前影视app上的常见影片。都是时长约2个小时左右的长视频。视频资源大多数为1080P，有5%-10%是4k高清资源。<br>未采集完成的数据资源约有1000TB，这里包含了各种综艺，电视剧，电影，短视频等数据，可以按需提供采集服务； | https://pan.quark.cn/s/c691e9e3be85 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 图片数据 | 千万级别 | 支持按需筛选数据<br>数据库交付 | 数据分类和量级：摄影图：62万，创意图：18万，插画：10万，设计模板：22万，免扣：60万<br>数据质量：所有图片数据带配套标题、多维度标签、描述<br>数据来源：来源合规，正式版权授权交付 | https://pan.quark.cn/s/cceb9684e1be |
| | | 3D模型数据 | 百万级别 | 支持按需筛选数据<br>数据库交付 | 数据分类和量级：人物/商品/制造等各风格类别的3D模型数据，数据类型为主流的3D模型风格，支持输出3D图片数据和模型数据；<br>数据质量：所有数据带配套标题、多维度标签、描述<br>数据来源：来源合规，正式版权授权交付 | https://pan.quark.cn/s/86dd8c475750 |
| 7 | 多语种数据 | 多语种中外对应短句 | 百万级别 | 数据库交付<br>txt格式 | 常见主流语种（英语、日语、韩语、土耳其语、俄语、西班牙语、葡萄牙语、泰语、阿拉伯语、泰语、老挝、柬埔寨、越南、缅甸等)的高质量翻译数据(句对+篇章的对齐语料) | https://pan.quark.cn/s/4d0a696504cb |
| 8 | 医药类数据 | 百行云游医疗数据 | 亿条 | ftp库交付 | 具体数据详情见百行云游医疗数据sheet | |
| | | 医疗大模型训练语料 | 百万级别 | 数据库交付 | 数据分类和量级：包含医疗知识库（临床所见65万条；手术及操作10万条；检查检验6万条；药品药物490万条）、知识问答库75万条；图谱三元组145万条；医学文献45万本；<br>数据质量：总量500万条数据，知识库都经过清洗和学校专业老师矫正标注，数据都有合规凭证，正规知识产权授权，数据质量有保证。<br>数据来源：来源合规，正式版权授权交付 | https://pan.quark.cn/s/53fcf7bd97e6 |
| 9 | 法律数据 | 国家法律法规语料库 | 319280条 | ftp库交付 | 数据更新到23年中 | https://pan.quark.cn/s/d9167ab685d6 |
| | | 裁判文书 | 1.3亿 | | 数据说明：司法裁判文书涵盖2000-2023下半年时间范围，2000之前数据很少只有几千份，包含司法中心公布的所有裁判文书<br>交付格式：已深度清洗，统一标准字段内容，数据库交付 | |
| 10 | | 金融证券类数据（证券公告 金融舆情 研报 政策法规） | 千万级别 | 数据库交付 | 所有数据更新到23年底，包含金融舆情信息，国内证券市场公告数据，证券行业监管信息以及证券问答数据，证券行政处罚数据，金融业法律法规数据，企业上市政策以及产业链数据，基本覆盖所有证券机构<br>金融研报2017-2023：170W篇<br>#上市公司反馈问答最早时间2008-02-27<br>#创业板反馈问答最早时间2020-07-08<br>#北交所反馈问答最早时间2020-05-12<br>#科创板反馈问答最早时间2019-04-23<br>#三市公告数据库最早时间1991-12-23 | https://pan.quark.cn/s/8b353b0d4130 |
| | | 金融数据-金融研报库 | 170万篇 | 数据库交付 | 覆盖范围：自2017年起170万篇研报；<br>数据说明：涵盖研究机构发布的各类研究报告，包括行业研究、行业点评、行业深度、公司研究、宏观策略、晨会报告,港股报告等研报。数据字段：文件标题、正文、研究机构、研报类别 、行业类别<br>交付格式：提供pdf与经过抽取清洗校验的txt文件； | https://pan.quark.cn/s/8b353b0d4130 |
| | | 金融数据-金融法规库 | 100万条 | 数据库交付 | 覆盖范围：基本全部法律法规数据<br>数据说明：包含金融业法律法规数据:1、各证券监管机构-<br>上交所、深交所、证监会、各地证监局、北交所、股转系统、证券业协会、基金业协会;2、金融相关监管机构-财政部、金融法规库发改委、央行、银保监会、银行间市场协会、外管局;3、国家部委-商务部、交通部<br>、环保部等等。字段标签:标题、正文、附件解析内容、发布机构、业务分类、适用范围、时效性。<br>交付格式：提供pdf与txt解析件 | https://pan.quark.cn/s/8b353b0d4130 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | 金融数据 | 金融数据-<br>三市公告数据库 | 1500万 | 数据库交付 | 覆盖范围：更新到最新数据，三市公告数据库最早时间1991-12-23；<br>证券公告数据说明【1000万】：包含上海证券交易所(主板+科创板)、深圳证券板的市场公告。包括年度报告、招股说明书、股东大会公告、董事会公告、业绩预告、股权激励公告、监管处罚等公告，交易所(主板+创业板)、北京证券交易所、新三。包含字段：文件标题、正文、发布主体、业务分类、公告类型、行业统计、市场类型、地域分布<br>交易所公告数据说明【350万】：包含沪深交易所、银行间市场发债主体相关公告。包含字段：文件标题、正文、债券名称、发行人、市场类型、债券品种、公告类型。<br>基金公告数据说明【110万】：包含包含公募基金的基金公告。数据字段包含文件标题、正文、基金名称、公告类型基金管理人。<br>交付方式：数据库推送 | https://pan.quark.cn/s/8b353b0d4130 |
| | | 金融数据-<br>金融证券反馈问答库 | 10万条 | 数据库交付 | 覆盖范围：<br>#上市公司反馈问答最早时间2008-02-27<br>#创业板反馈问答最早时间2020-07-08<br>#北交所反馈问答最早时间2020-05-12<br>#科创板反馈问答最早时间2019-04-23<br>数据说明：拟上市公司在IPO进程中收到的反馈意见及回复的内容;<br>所有A股上市公司在日常经营过程中收到交易所的问询函件和回复内容;<br>发债企业在债券发行过程中收到交易所的问询函件和回复内容，包含字段和清洗标签【股票代码、公司简称、三级分类反馈意见问题标签(目前涵盖一级标签3个经营/财务/法律、二级标签58个、三级标签546个，示例:财务-利润损益-营业收入、经营-重大事 项-<br>项目情况及实施进展、法律-股权结构-<br>红筹架构)、中介机构名称、标题、发布日期、源文件、解析文件、问答提取文件、源url、公司地区】；<br>交付格式：一问一答拆条配对数据库，并且已深度清洗有多维度标签。<br>备注：也有纯pdf文件和txt文件，问题不拆条数据，数据量一样多 | https://pan.quark.cn/s/8b353b0d4130 |
| 11 | 产业企业工商数据 | 百行云游自研工商数据 | 亿条 | ftp库交付 | 具体数据详情见百行云游自研工商数据sheet | https://pan.quark.cn/s/4dd88a085cd7 |
| | | 产业相关数据 | 覆盖全国大部分产业 | ftp库交付 | 有上中下游的产业分类，穿透到企业名单，和企业的工商，司法，专利，商标，软著，财务，舆情等数据<br>全国招标信息，投融资信息，产业舆情，产业政策，行业标准 | |
| | | 企业相关数据 | 覆盖全国大部分企业 | ftp库交付 | 企业信息：招投标信息，行业关联信息，投融资信息，舆情信息，招聘信息，核心高管背景，企业行业标准 | |
| 12 | 标注数据 | 语音标注数据 | 几十万小时 | 数据集交付，支持定制需求<br>源数据+标注标签数据 | 具体数据详情见百行云游医疗数据sheet<br>1、TTS语音合成标注数据，带情感和不带情感的标注数据均有，高品质带有准确文本标注，大概1500人的数据量，每个人1000段录音。还有各类语音数据集；<br>2、多场景语音数据：包含智能家居、中文普通话、成人英语、声纹数据、远场语音数据、回忆数据、智能机器人数据等多种语音数据，总时长10万+小时。 | https://pan.quark.cn/s/efdcd56ce702 |
| | | OCR标注数据 | 百万级别 | 数据集交付，支持定制需求<br>源数据+标注标签数据 | OCR多种场景标注数据，例如中文OCR相关数据集：主要包含票据OCR数据，多种中文场景OCR数据，体检报告OCR数据，互联网图像OCR等多种数据 | |
| | | AI视觉标注数据 | 百万级别 | 数据集交付，支持定制需求<br>源数据+标注标签数据 | AI视觉标注标注数据，例如车辆图像视频数据集：主要包含工程车辆数据，车辆检测数据，车辆属性图像描述数据 | |
| | | 标书 | 10000个模版,68G | 2019-2024 | 包含60个行业各类标书范本 | |
| | | 中文论文数据 | 5000万条 | 更新到2023年 | 中文论文数据，PDF源格式，5000万左右 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 13 | 长文档数据 | 国内专利 | 4000万+ | 1985-2023 | 收录了1885至今的中国专利数据库，包括发明专利和实用新型<br>中国专利： 全部可以提供全文<br>数据，查个数。23年前个人的专利没有，企业有全量。<br>国外专利： 23年6月之后 可以全文数据，2千万，<br>23年6月之前，结构化数据，但是缺少全文字段<br>查数，5千万 | https://pan.quark.cn/s/e3e7bc57177f |
| | | 国外专利 | 0.7亿+ | | | |
| | | 合同模版 | 3G,2500个 | 2020-2024 | 各类合同模版2500个，包括公司合同模版1670个 | |
| | | 国家标准 | 300G | 2024版 | 2024最新省标国标数据，包含建筑、结构、排水、电气、暖通各专业，道桥，水利，交通等全工程类设计施工验收等图集规范 | |
| | | 长文档数据国内出 | 亿篇+，150 | 2010-2024 | 1亿+的电子书，包含网络小说，历史、教材等书籍，有中文和英文的 | |
| | | 其他长文档数据海 | 1亿，100TB | 1980-2024 | 约有1亿篇论文，大约80%的文件是发表在期刊上的研究论文 6%是会议论文集（conference proceedings）中的论文，5%是书籍的章节<br>剩下的就是其他种类的文件了。77%发表于1980年到2020年<br>36%发表于2010年到2020年。所有主要科学出版社的覆盖率都在95%以上。。 | |
| 14 | 代码数据 | AI代码数据集 | 67TB | 数据集 | 数据来源：不限GITHUB，多渠道来源，专门为大模型训练整理<br>包含了超过 30 亿个来自 600<br>余种编程与标记语言的文件，全量数据为<br>67.5TB，超过 6TB 的许可源代码文件，涵盖 358<br>种编程语言。 | https://pan.quark.cn/s/654f1dfc6a86 |
| 15 | 其他数据 | 科研类数据 | 见描述 | 数据集&API交付 | 科研类数据，涉及职业安全，医药健康，环境，政务智慧城市星座<br>科技等各方面数据，具体见科研类数据sheet | |