DATATANG (BEIJING) TECHNOLOGY CO.,LTD.

更高质量的数据 更有竞争力的AI

数 据 堂 （ 北 京 ） 科 技 股 份 有 限 公 司

目录

01

企业介绍

数据堂

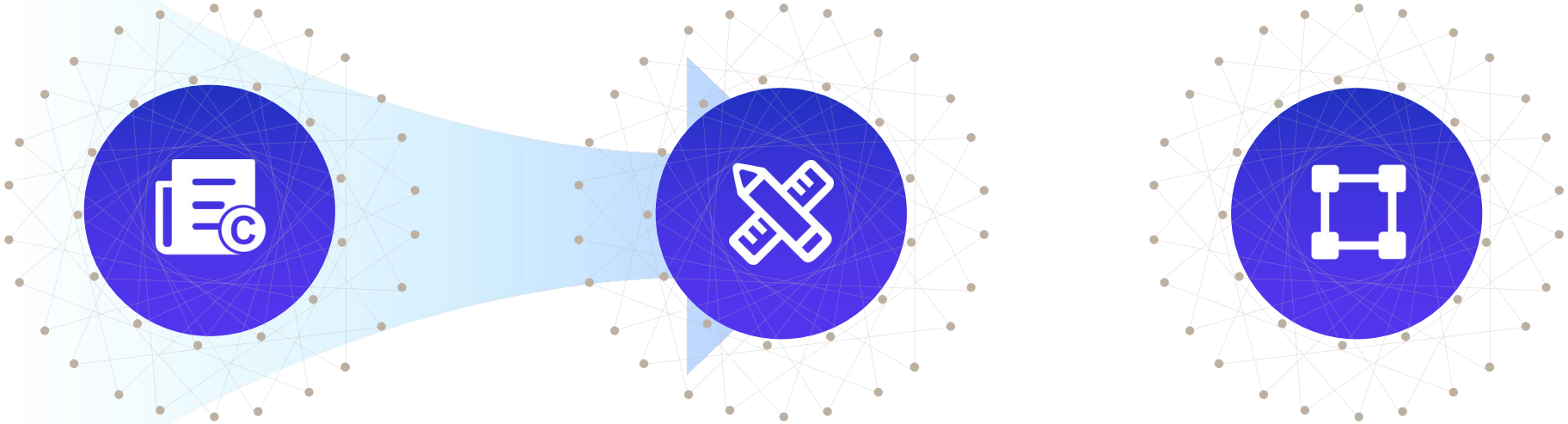
全球领先的人工智能数据服务商

股票代码： 831428

数据堂拥有丰富的训练数据集产品，提供数据定制服务，旗下数加加标注平台通过集成自动化标注工具可以快速降低数据处理成本。

凭借高质量训练数据服务，数据堂已帮助全球上千家企业提升AI模型性能。

我们的产品与服务



版权数据集

即买即用

更高的质量要求

数据定制服务

满足个性化需求

稳定成熟的数据生产线

数据标注平台

多年实施经验积淀

人机协作，成本更低

02

版权数据集

数据堂 Datatang

版权数据集

数据堂现有版权数据涵盖PB级大模型数据，100万+小时语音数据，800TB计算机视觉数据，数据质量经全球AI头部企业考验，值得信赖。

版权

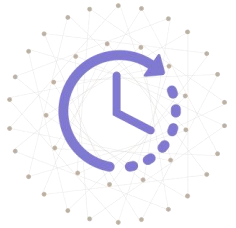
自有版权 清晰可靠

安全

采集授权协议备查

高质量

超越普通数据质量要求



高效

即需即用

分秒交付

快速建立提升AI模型准确率

PB级

大模型数据集

覆盖语言大模型和多模态大模型的各项研发需求

高质量无监督文本数据

- **大规模高质量无标注文本**：覆盖 K12、大学及研究生约60万册教材，1.5亿道中文考试试题，300万册中文图书
- **翻译数据**：目前有30多个中/英和其他语种的翻译数据，超过2亿对大规模平行语料

指令微调（SFT）问答对数据

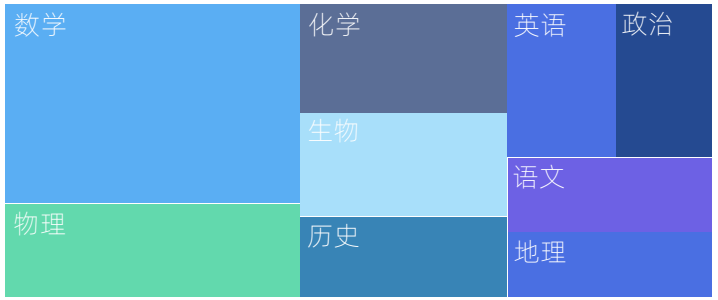
- **基础任务SFT**：20万组中文大模型通用SFT指令微调文本、10万条复杂指令跟随数据
- **垂直任务SFT**：50万条内容安全类文本数据，覆盖31大类别，针对性提升对敏感问题的回答能力

多模态数据

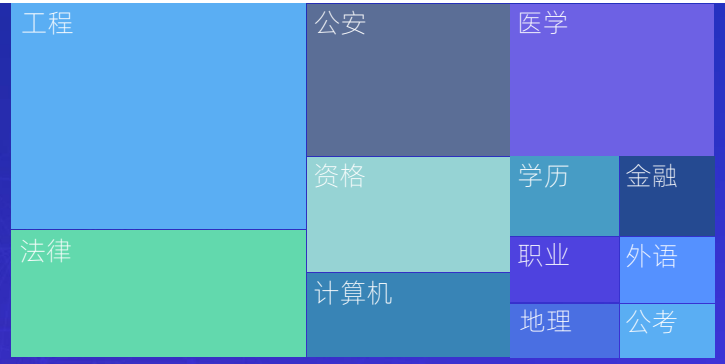
- **图文对**：800TB图片-文本描述数据集，正版图像版权，高清图片，文本描述采用中文、英语进行描述
- **视文对**：1PB视频-文本描述数据集，包括正版视频作品以及描述和标签

1.5亿道中文试题数据集

涵盖K12（2286万道题目）、大学及职业考试（1.35亿道题目），涵盖全学科和全专业，试题中的图片和公式都进行了解析转换。



K12考试
试题数据



大学及职业
考试试题数据


```
{
  "exam_point": "科学记数法—表示较小的数",
  "id": "d0ba8ebf8f8afeab1917a6fea470ea6b",
  "content_type": 2,
  "type": "填空题",
  "grade_band": "初中",
  "difficulty": "一般",
  "grade": "九年级",
  "course": "数学",
  "paper": "2015春•市北区期中",
  "online_test": false,
  "option_split": false,
  "quality": "精品",
  "question_info": {
    "raw_content": {
      "title": "2011年3月11日，日本发生了里氏9.0级大地震，导致当天地球自转时间减少了0.0000016秒，0.0000016用科学记数法表示为
      "option_a": "",
      "option_b": "",
      "option_c": "",
      "option_d": "",
      "option_e": "",
      "answer1": "1.6×10<SUP>−6</SUP>"
    }
  },
  "answer_info": {
    "raw_content": "解：0.0000016=1.6×10<SUP>−6</SUP>，故答案为：1.6×10<SUP>−6</SUP>"
  },
  "solution_info": [
    {
      "solution_info": "绝对值小于1的正数也可以利用科学记数法表示，一般形式为a×10<SUP>−n</SUP>，与较大数的科学记数法不同的是
    }
  ],
  "edition": "人教新版#北师大新版#华师大新版#苏科新版#湘教新版#浙教新版#冀教新版#沪科新版#北京课改新版#沪科版",
  "chapter": "人教新版`七年级上`第1章 有理数`1.5 有理数的乘方`1.5.2 科学记数法#北师大新版`七年级上`第2章 有理数及其运算`2.10
  "knowledge_point": "科学记数法—表示较小的数",
  "children": []
}
```

50万条中文大模型内容安全类文本数据集

包括40万条覆盖网信办31大类别的敏感指令集，10万条为各类脏话。可针对性提升大模型对敏感问题的识别能力。

首页 | 简 | 繁 | EN | 登录 | 邮箱 | 无障碍

首页 > 政策 > 国务院政策文件库 > 国务院部门文件　　字号：默认 大 超大 | 打印 收藏 留言

标　　题：	生成式人工智能服务管理暂行办法	发文机关：	国家网信办 国家发展改革委 教育部 科技部 工业和信息化部 公安部 广电总局
发文字号：	国家互联网信息办公室 中华人民共和国国家发展和改革委员会 中华人民共和国教育部 中华人民共和国科学技术部 中华人民共和国工业和信息化部 中华人民共和国公安部 国家广播电视总局令第15号	来　　源：	国家网信办网站
主题分类：	科技、教育\科技	公文种类：	命令（令）
成文日期：	2023年07月10日		

国家互联网信息办公室
中华人民共和国国家发展和改革委员会
中华人民共和国教育部
中华人民共和国科学技术部
中华人民共和国工业和信息化部
中华人民共和国公安部
国家广播电视总局
令
第15号

《生成式人工智能服务管理暂行办法》已经2023年5月23日国家互联网信息办公室2023年第12次室务会会议审议通过，并经国家发展和改革委员会、教育部、科学技术部、工业和信息化部、公安部、国家广播电视总局同意，现予公布，自2023年8月15日起施行。

国家互联网信息办公室主任 庄荣文
国家发展和改革委员会主任 郑栅洁
教育部部长 怀进鹏
科学技术部部长 王志刚
工业和信息化部部长 金壮龙
公安部部长 王小洪
国家广播电视总局局长 曹淑敏
2023年7月10日

回到顶部

作为我国首部生成式人工智能服务领域的专门立法，**《办法》鼓励优先采用安全可信的软件、工具、计算和数据资源**，促进生成式人工智能技术健康发展和规范应用。

1PB高质量视频描述数据集

摄影师发布的正版视频作品以及作者撰写的描述和标签列表，其中700万条为英文描述，300万条为中文描述。视频涵盖人物、风景、动植物等多种类别，视频分辨率1080p及以上。

描 述

Little girl writing and drawing. Doing homeworks. Child. Adorable girl.

标 签

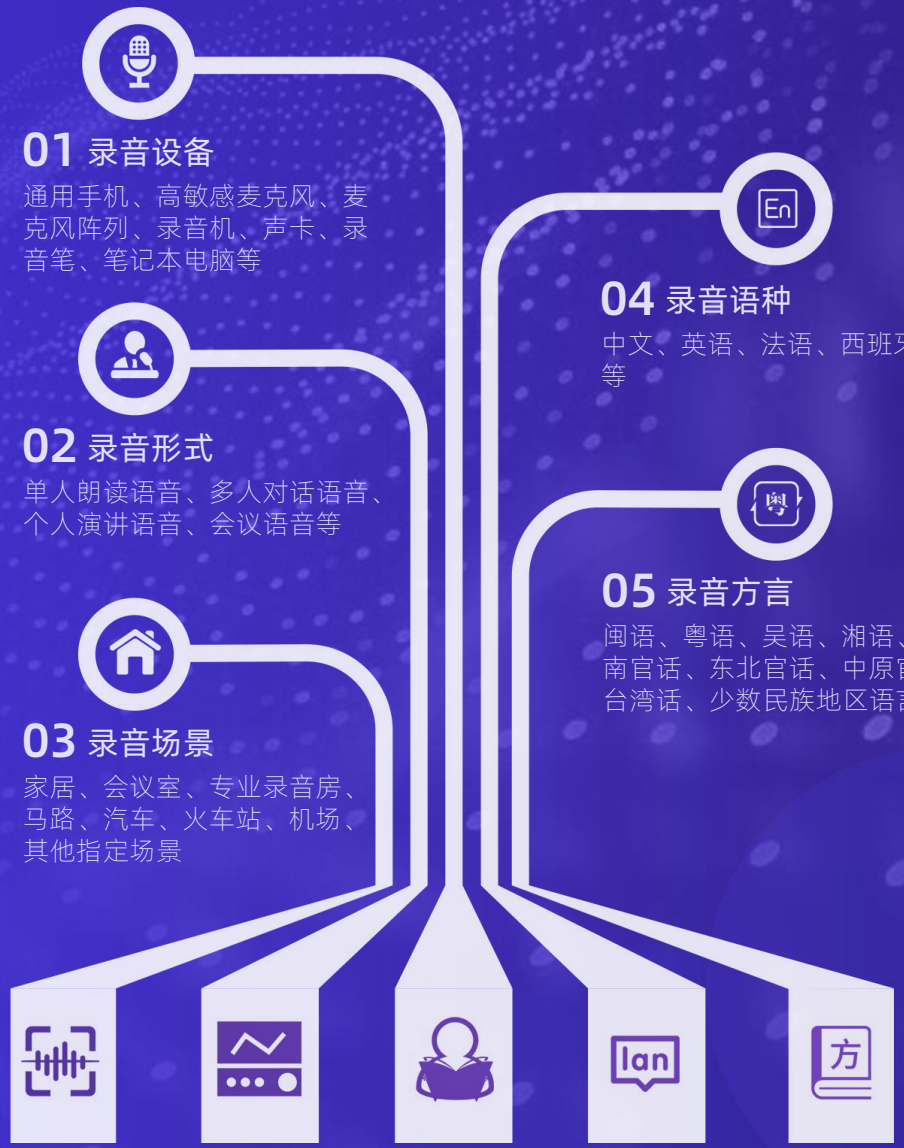
caucasian,child,childhood,color,cute,drawing,education,female,girl,happy,learning,little,beautiful,crayon,kid,paper,school,sitting,young,care,family,home,house,people,portrait,working,writing,colorful,leisure,pencil,angel,book,carefree,concentrated,concentration,day,hair,hand,homework,infant,innocence,interior,laugh,life,lifestyle,minor,new,painting,pen,pure.

100万+小时

语音数据集

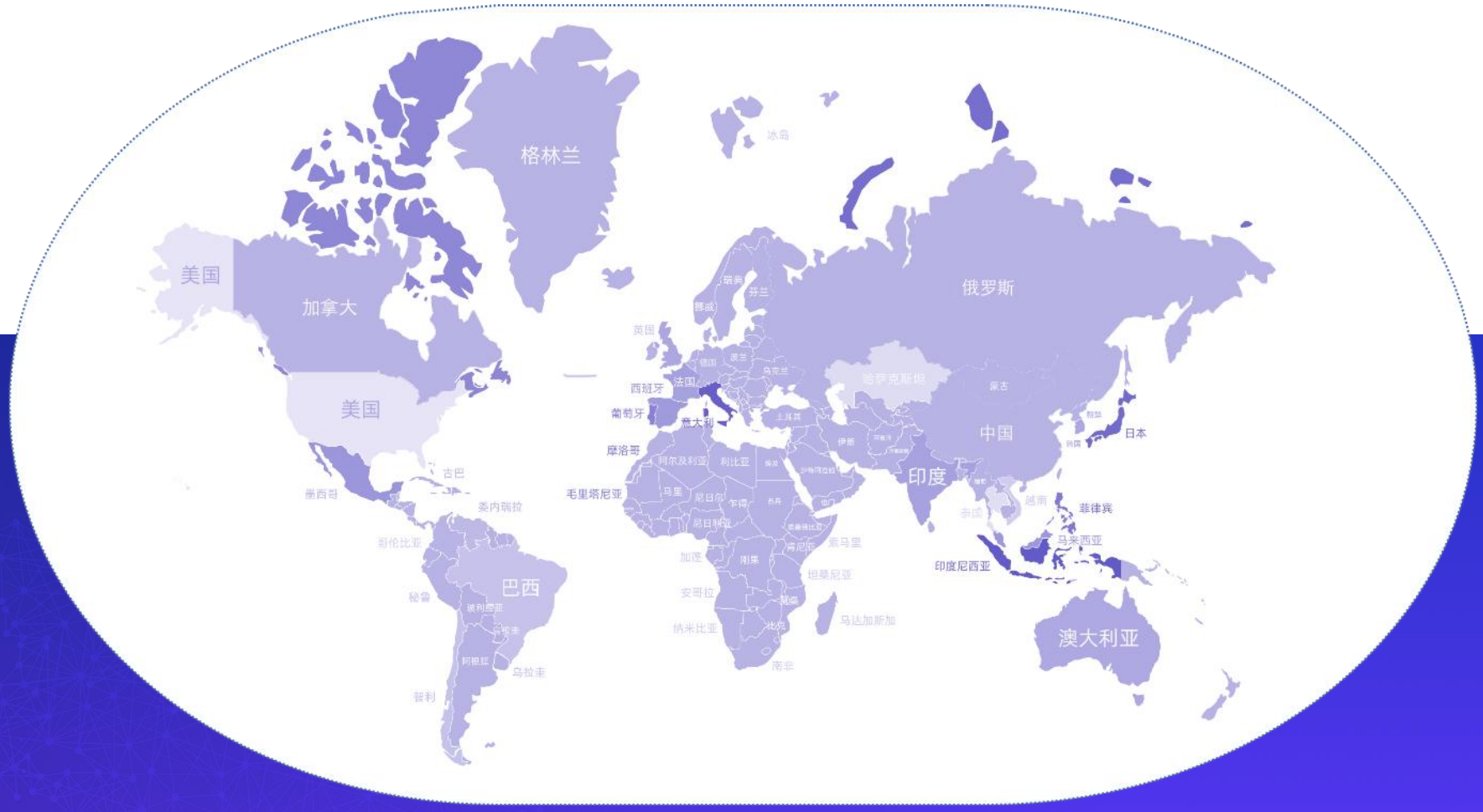
覆盖多种专业录音设备、场景

丰富的语种及录音形式



多语种语音识别数据集

数据堂现有约40万小时多语种语音数据集，这些高质量的外语语音数据集可以为语音识别模型优化提供非常好的帮助。



- 中文普通话
- 美式英语
- 英式英语
- 德语
- 意大利语
- 法语
- 西班牙语
- 欧洲葡萄牙语
- 俄语
- 印地语
- 日语
- 韩语
- 泰语
- 印尼语
- 马来语
- 越南语
- 巴西葡萄牙语

中文方言语音数据集

数据堂拥有20万小时以上中文方言语音数据集，覆盖全国8大方言区域，全部采自本土发音人。

- 闽语
- 粤语
- 吴语
- 湘语
- 西南官话
- 东北官话
- 中原官话
- 少数民族地区
- 台湾

多音色语音合成数据集

数据堂自建专业录音棚，拥有丰富的样音资源，可以完美匹配多个语种环境新闻播报、智能客服、有声阅读、虚拟主播、语音交互、音乐合成等多领域声音特色需求。


女声，意大利语，自然

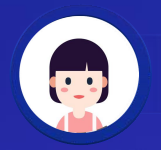

男声，荷兰语，磁性

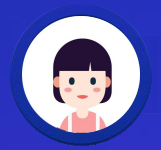

男声，日语，亲切

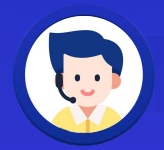

男声，西班牙语，沉稳

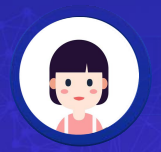

女声，潮汕方言，标准

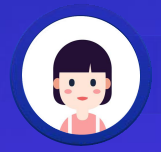

女声，客家方言，标准

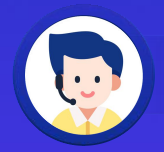

女声，四川方言，标准


女声，中文普通话，愤怒


男声，中文普通话，霸总腔


女声，中文普通话，故事腔


女声，中文普通话，甜美声


男声，中文普通话，客服音

800TB

计算机视觉数据集

数据堂围绕人脸人体、
光学字符识别（OCR）、
交通街景等场景布局并拥有一定量级的数据资源，覆盖约50万人。

01 采集设备
手机、双目摄像头、双目红外摄像头、深度摄像头、工业相机、热成像仪、多角度采集摄像头等

02 采集形式
实景采集、模拟采集、众包采集

03 标注类型
图像分类、目标检测、目标追踪、语义分割、实例分割、全景分割等

04 标注方式
点、线段、矩形框、多边形框、3D立体框

计算机视觉数据集

20万ID、1000万条

人脸图像视频数据

典型人脸数据集

- 2D&3D人脸识别数据
- 2D&3D活体检测数据
- 人脸关键点数据
- 人脸分割数据
- 人脸表情识别数据
- �跨年龄人脸数据

30万ID、100万条

人体图像视频数据

典型人体数据集

- 多场景人体行为识别数据
- 人脸&人头&人体检测数据
- 人体关键点数据
- 人体语义分割数据
- Re-ID数据
- 跨摄像头跟踪数据
- 人体属性数据

2500ID、100万条

手势图像视频数据

典型手势数据集

- 手语手势识别数据
- 娱乐手势识别数据
- 动态手势识别数据

50万张

OCR图像数据

典型OCR数据集

- 常用语言自然场景OCR数据
- 常用语言会议PPT OCR数据
- 常用语言手写体OCR数据
- 票据OCR数据
- 试卷OCR数据

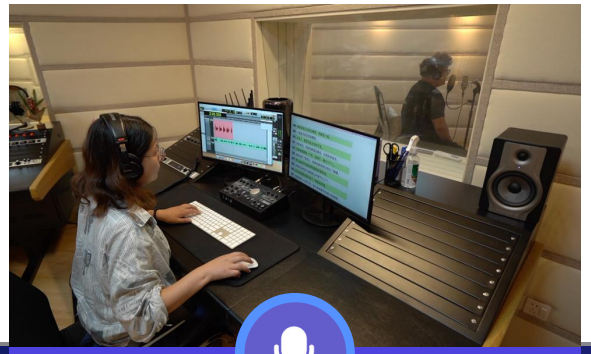

Datatang

03

数据定制服务

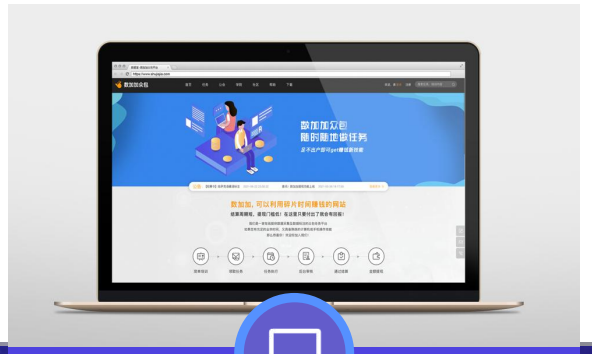
数据堂
Datatang

数据采集定制服务



专业TTS录音棚

布景采集



众包线上采集

数据堂拥有多套专业数据采集设备、工具和环境

项目经理拥有丰富的采集经验及质量管控经验，可以满足客户多种场景与类型的数据采集需求

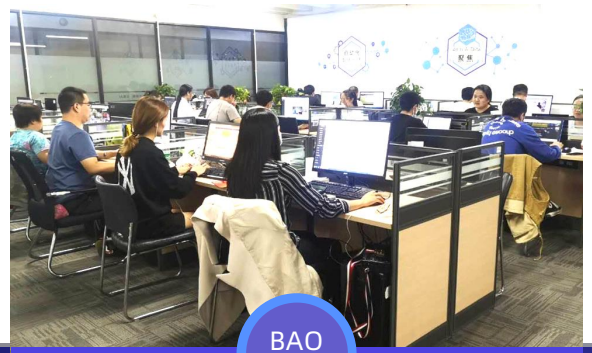
数据标注定制服务

BEIJING

北京数据服务基地

HEFEI

合肥数据服务基地



BAO DING

保定数据服务基地

SEA

东南亚数据服务基地

数据堂在国内自建3个、东南亚1个大型标注基地，现有5000名以上经验丰富的专业标注人员

支持大模型、语音、图像、视频、点云、文本等专业数据标注定制服务


rider
8_car
3_pers
4_car
9_traffi
10_
13_traffic_sign
7_car
11_traffic_
14_traffic_sig
12_traf
5_car

数据定制
服务流程

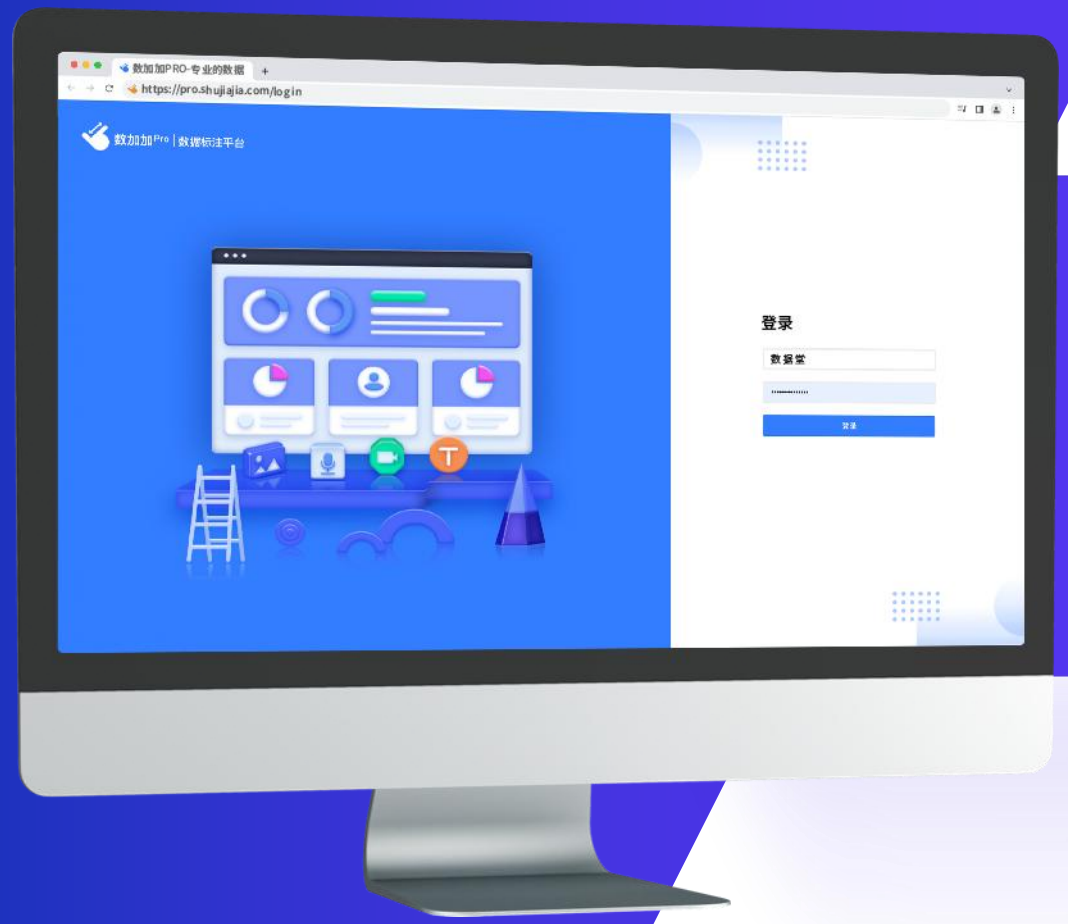

01
需求沟通
接收客户需求
内部评估报价
02
确认合作
客户确认报价
双方签订合同
03
实施规范
细化实施规范
标准可落地
04
资源整合
平台任务部署
协调基地
全球合作伙伴、众包资源
05
立项实施
内部立项
项目启动会
06
样品确认
提交样品给客户
规范磨合至最终确认
07
进入量产
全程质量监控
平台内置自动化质检功能
多轮次质检流程保障
按批次验收
08
交付验收
按客户指定格式交付
结果数据履约删除
09
售后服务
合同约定期内
提交售后服务

04

数据标注平台

数据堂 Datatang

数加加Pro | 助您快速搭建自己的数据标注生产线



数据堂多年标注实施经验的结晶

操作体验优化到极致，每一个按钮都经过多次实施考验

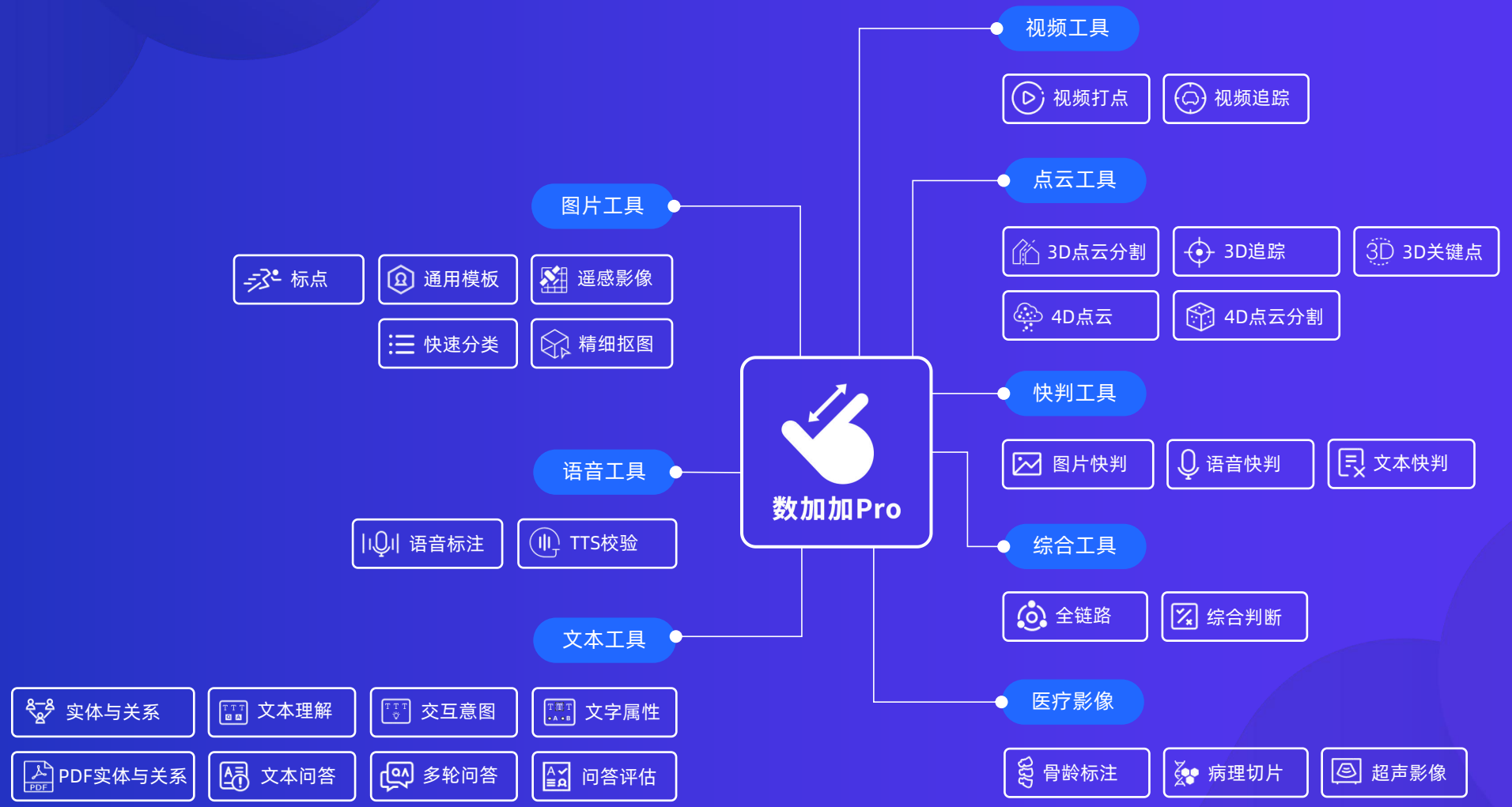
30套成熟的标注模板，全面覆盖大模型、语音、图像视频、3D点云数据标注需求

内置人机交互半自动标注与质检功能

可配置的作业流程，适应不同的作业需求

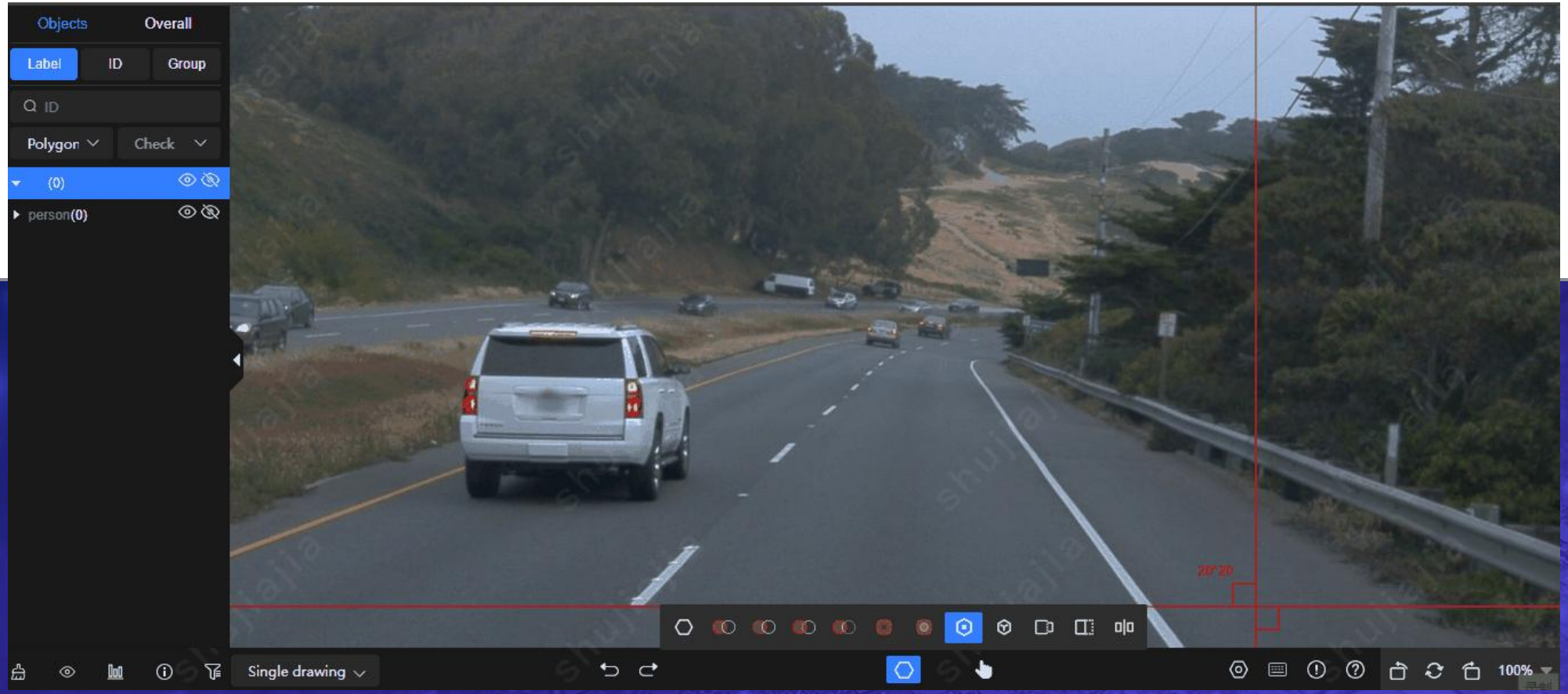
灵活支持独立、私有化等部署方式

数加加Pro | 丰富、高效的数据标注工具组件



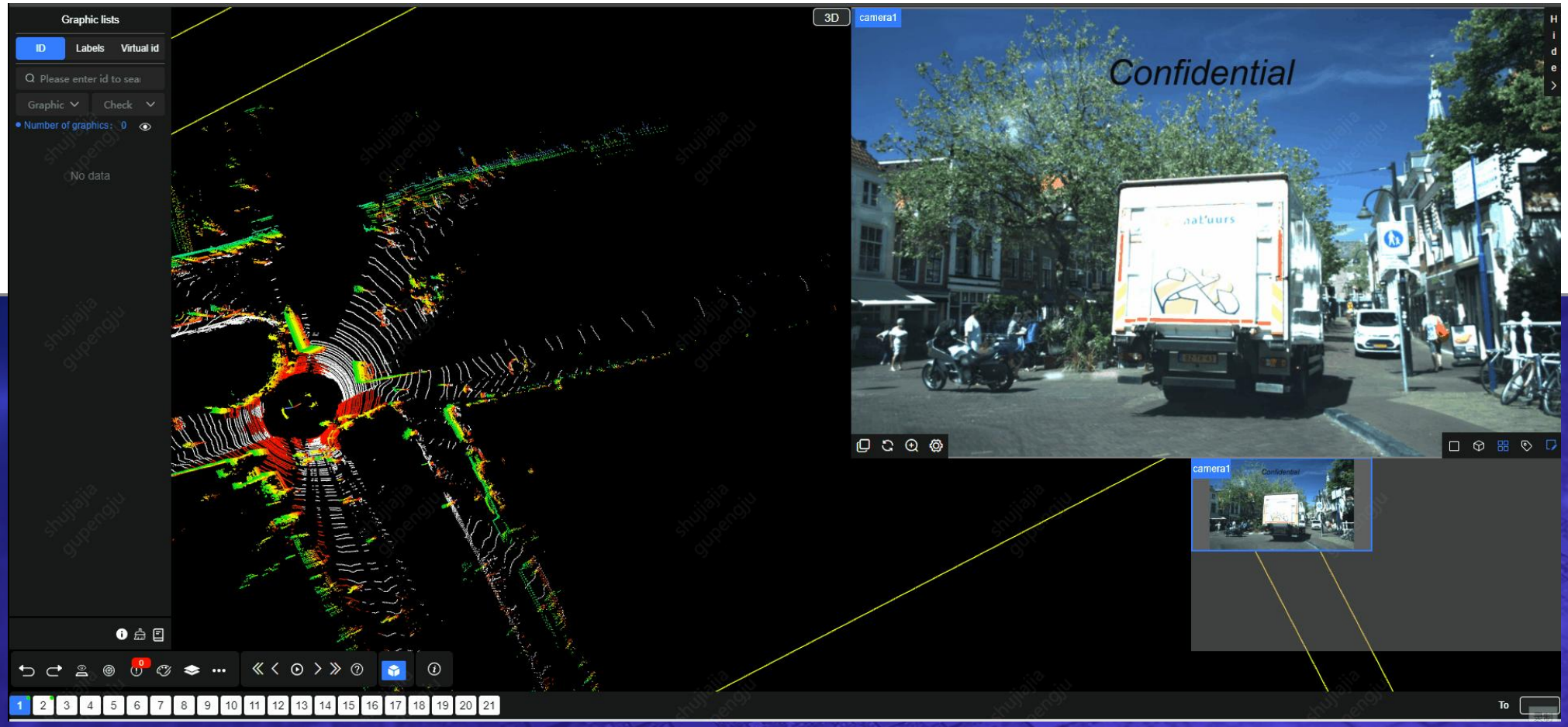
内置预识别引擎

支持人机交互半自动标注，包括语音、图片、点云等多种交互方式预识别引擎，人均标注效率提升30%以上。已成功应用在近5000多个项目的实施过程中。



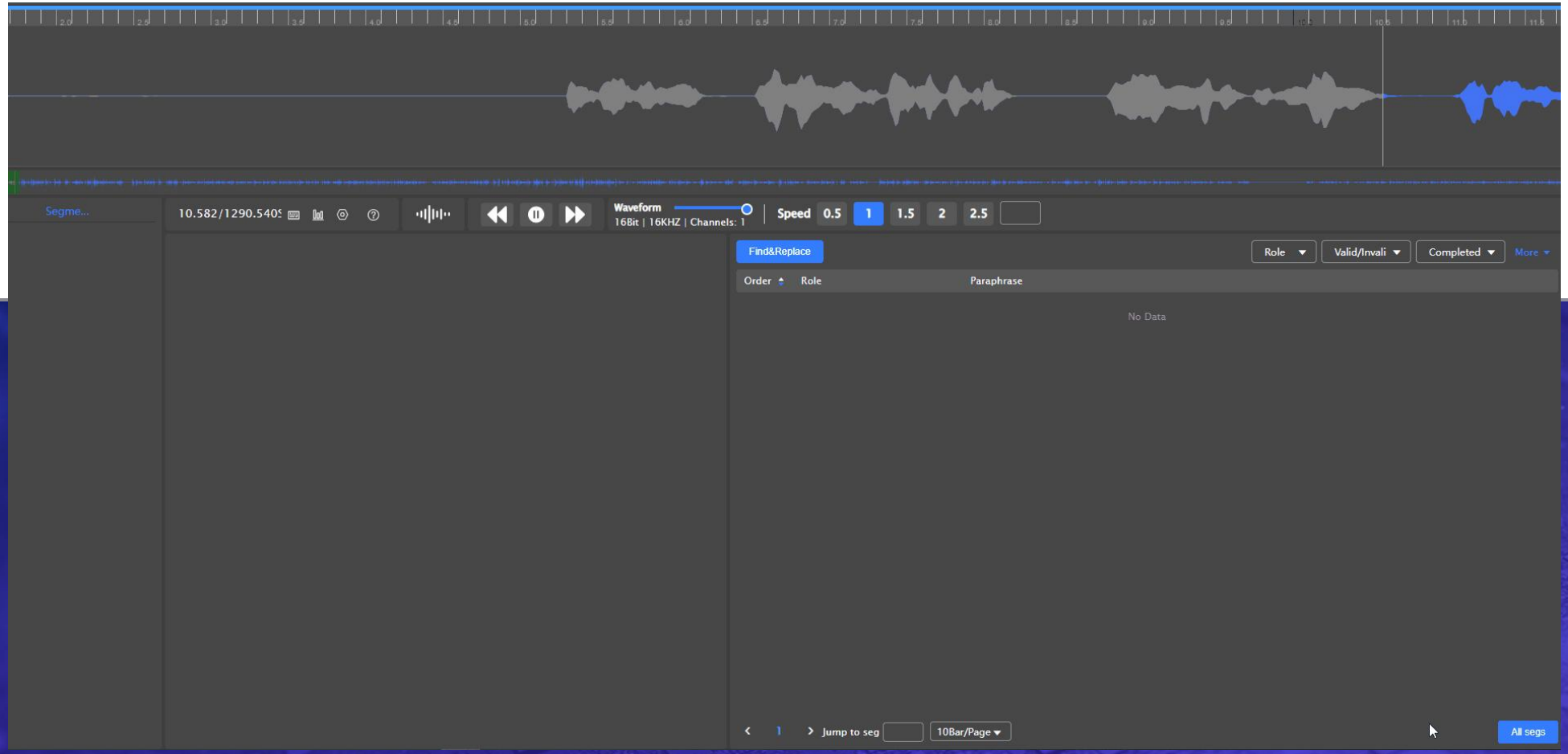
内置预识别引擎

支持人机交互半自动标注，包括语音、图片、点云等多种交互方式预识别引擎，人均标注效率提升30%以上。已成功应用在近5000多个项目的实施过程中。



内置预识别引擎

支持人机交互半自动标注，包括语音、图片、点云等多种交互方式预识别引擎，人均标注效率提升30%以上。已成功应用在近5000多个项目的实施过程中。



平台内置模板水印、日志审计、登录验证、API授权管理等安全功能，也支持私有化部署，数据不出门，符合安全测评标准。

华为技术认证书
鲲鹏云兼容性认证

北京市科学技术奖
科学技术进步二等奖

数加加标注平台
计算机软件著作权登记证书

数加加标注平台
性能检测报告

CMMI开发模型成熟度3级

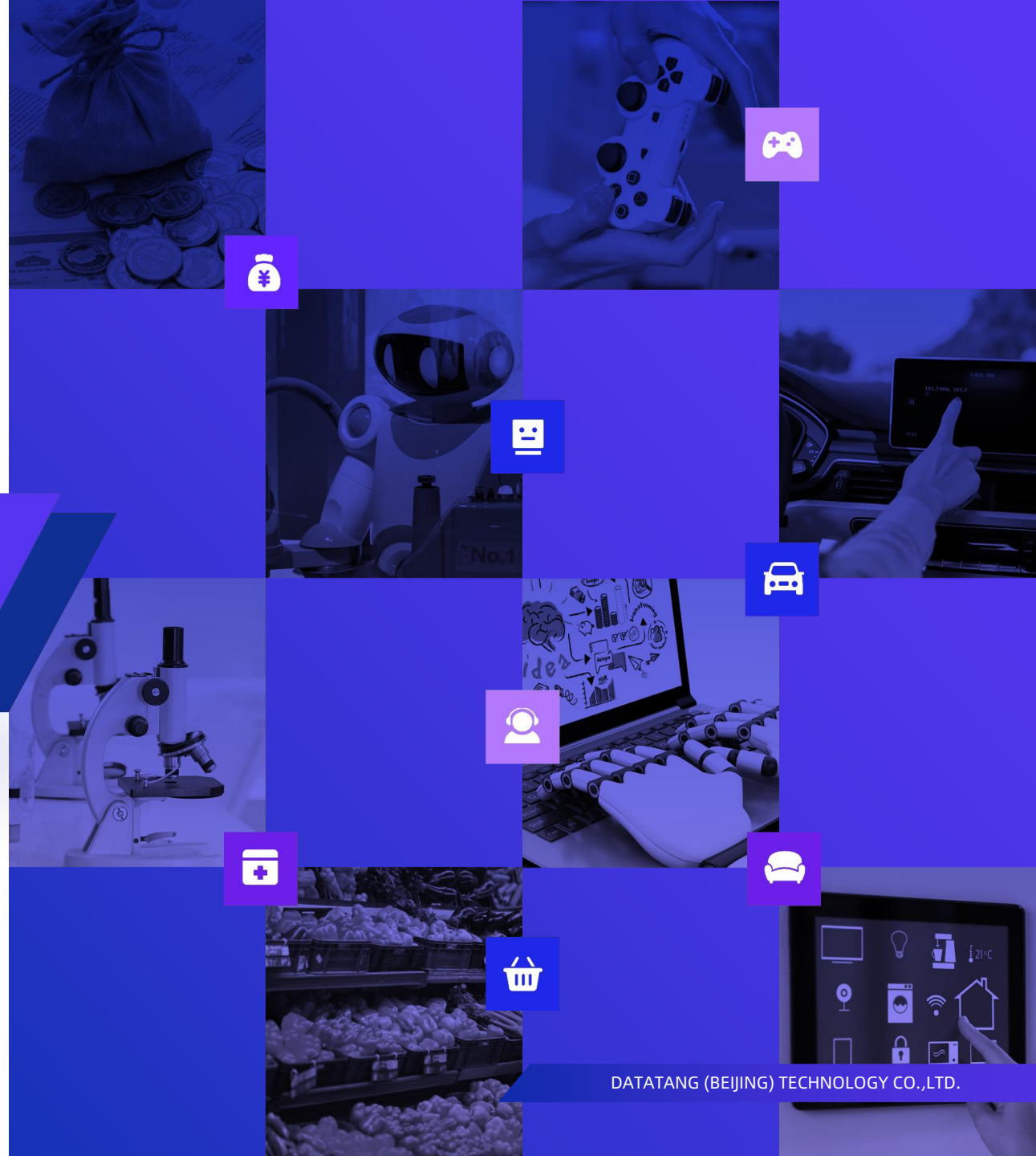
渗透测试报告（中英）

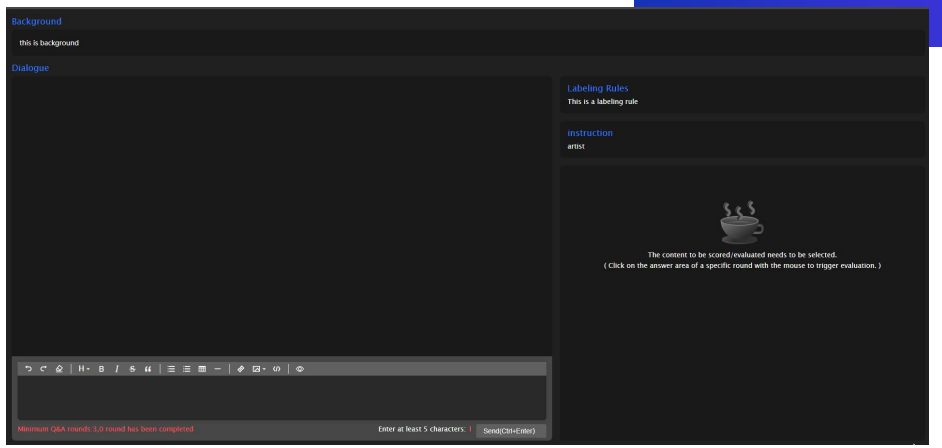
05 场景化数据解决方案

沉淀十余年数据整合处理经验，深刻理解多元业务场景的数据需求

依靠自建数据采集标注平台工具及自动化数据处理能力

数据堂可以提供多场景数据解决方案



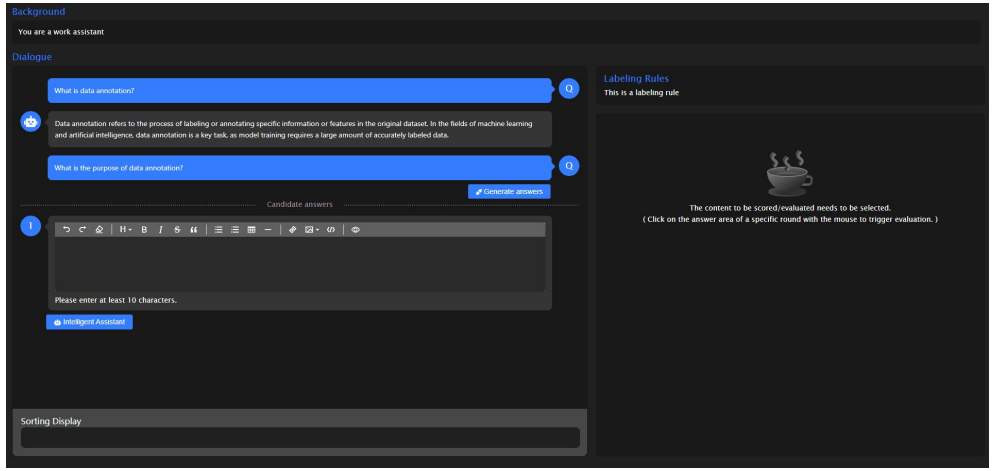


▲ 指令微调数据标注

▲ 多模态数据标注

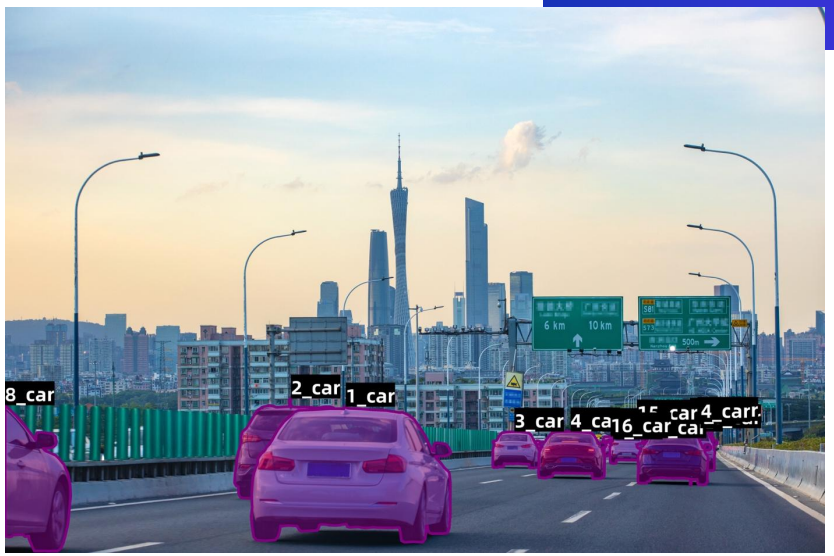
场景化数据解决方案——大模型

提供无监督数据的获取、清洗，也可以为后续监督学习阶段提供定制化数据服务，包括监督微调数据、基于人类反馈的强化学习数据服务



评分排序数据标注 ▶

Bard IFLYTEK Spark Qwen Tencent Hunyuan panguLM Galaxy AI BAAI 智源研究院 GIYUAN LAB Copilot

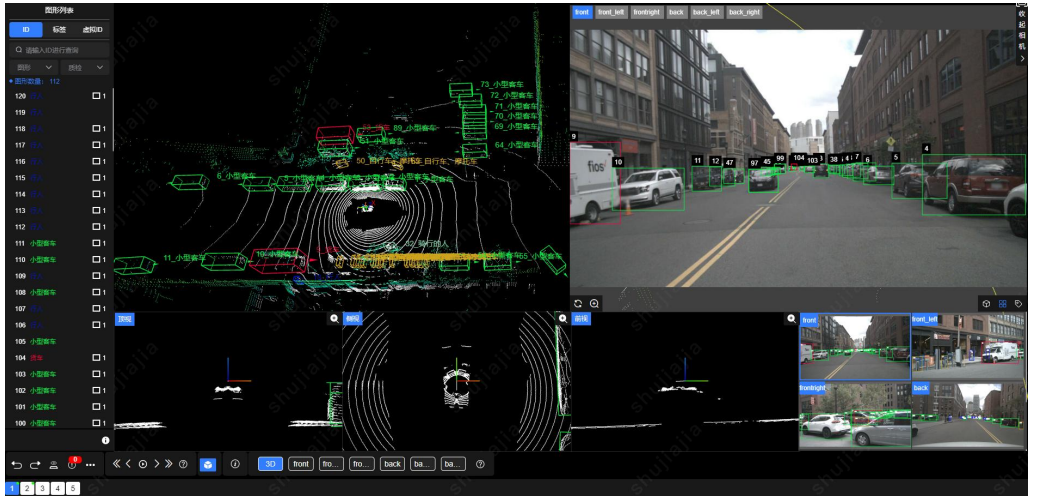


▲ 目标检测与跟踪

▲ 驾乘行为监测

场景化数据解决方案——智能驾驶

从常规目标检测、驾乘行为采标，到专业的点云数据处理数据堂高质量的训练数据可帮助智能驾驶AI模型更加准确，以创造更加安全的驾乘体验



点云标注 ▶

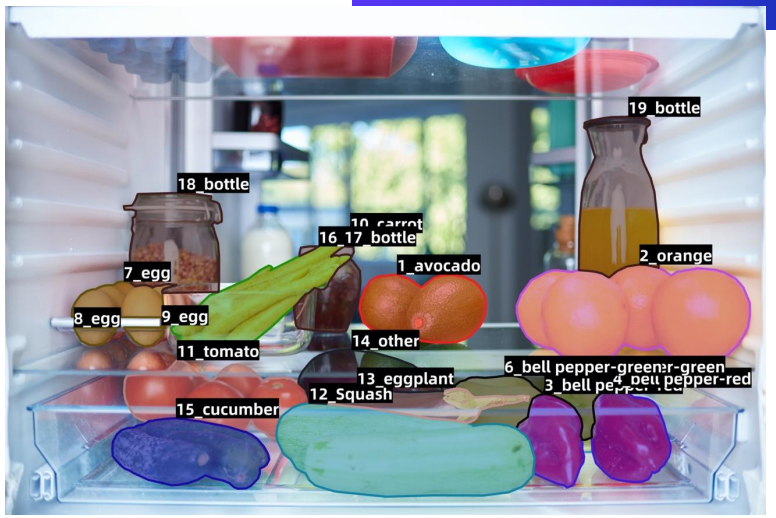
HAOMO.AI 挚途 GEELY 吉利控股集团 BOSCH 上汽集团 SAIC MOTOR 长城汽车 dji 蔚来 NIO 小鹏汽车 理想 HONDA PHANTOM AI 长安汽车 CHANGAN AUTO 中国一汽 FAW GROUP

场景化数据解决方案——智能家居

智能家居数据解决方案可以帮助
家居产品更智能化理解主人需求
时刻关爱家庭成员，人机沟通更顺畅

▼ 语音识别

▼ 异常事件检测



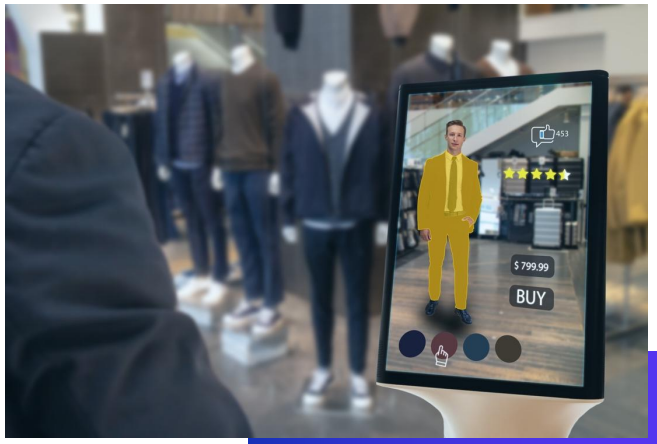
▲ 目标识别与分类

▲ 身份验证

长虹 CHANGHONG　AUCMA 澳柯玛　Segway-Ninebot　云米　KONKA 康佳　TCL　美的 Midea　GREE 格力　PHILIPS　SAMSUNG　HIKVISION 海康威视　POSITEC　LG　SIEMENS

场景化数据解决方案——新零售

数据堂在目标检测、语义分割、目标行为识别任务
拥有丰富的数据服务经验
可持续改善新零售场景的客户消费体验



▲ 个性化购物

库存管理 ▶

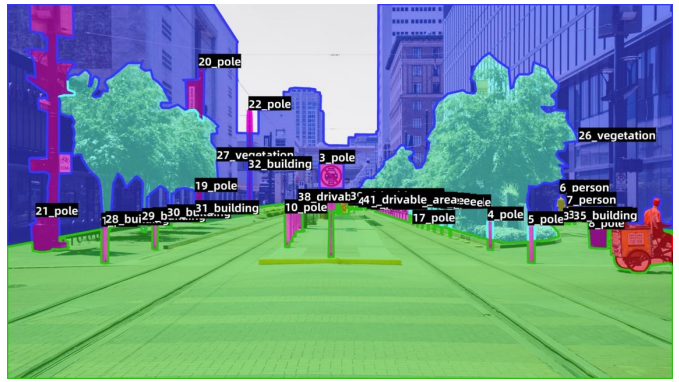
▼ 搜索相关性

场景化数据解决方案——游戏与娱乐

面向游戏与娱乐场景，数据堂建有多套人脸、
人体版权数据集，覆盖50万人
均已获得采集授权，数据安全更有保障

◀ 关键点检测

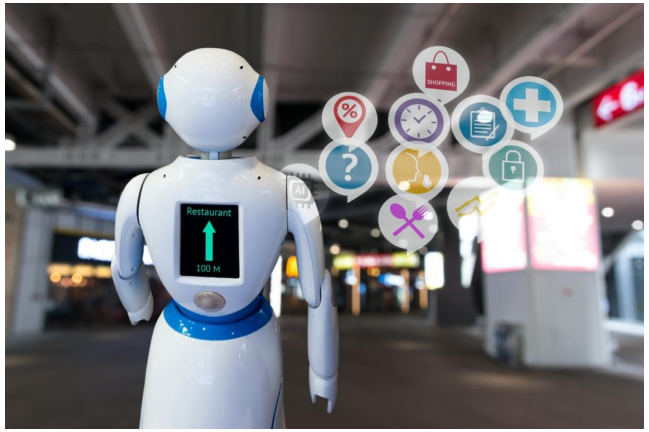
▼ 目标检测与分类



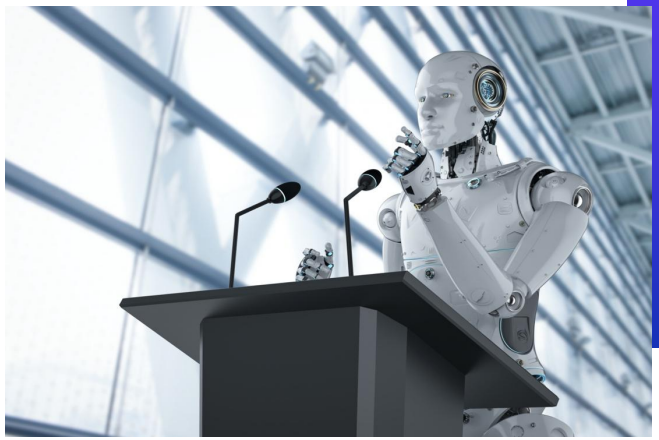
语义分割 ▶

抖音 网易游戏 快手 欢聚时代 YY Inc. 芒果tv oppo Tencent 腾讯 iQIYI 爱奇艺 酷狗音乐 bilibili 小红书 miHoYo

▼ 语音识别



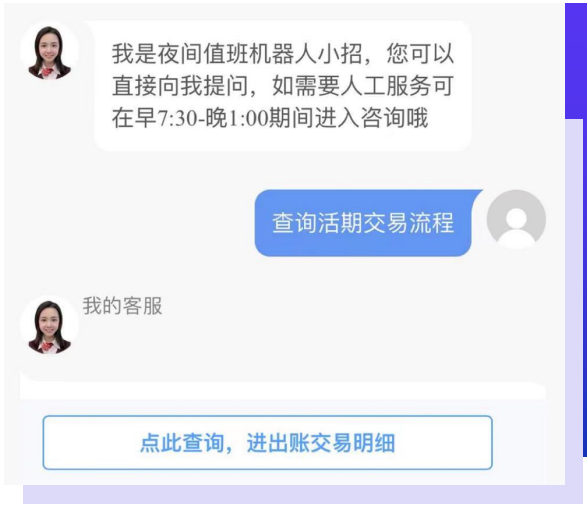
▼ 智能化沟通



场景化数据解决方案——智能客服

数据堂可以帮助客户处理语音、图像视频及文本等多类型数据，针对性提升智能客服AI模型性能

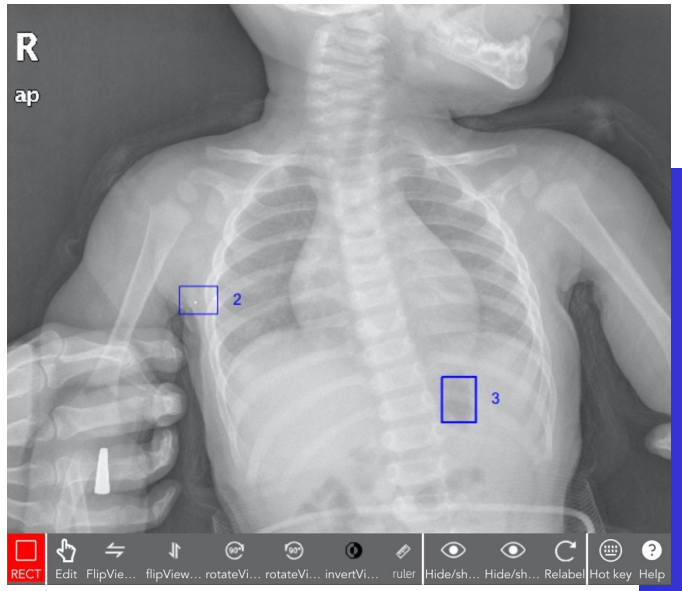
现有10000小时客服语音数据集，即刻应用



◀ 问题分类

场景化数据解决方案——智能医疗

数据堂可以帮助客户收集、标注、整理医疗类型数据
使AI模型可以对患者状况或疾病进行自动分类识别，
协助提高医疗工作效率



◀ 医疗影像标注

▲ 信息提取

◀ 特定病例收集

数据质量管理体系

无论何时，数据堂都将以保障数据质量为客户服务第一目标

我们希望通过更高质量的数据服务，以帮助客户的模型获得更好的表现

智能自检

系统内置智能自检

人机协作

节省更多时间与成本

多轮次质检流程

[采集/标注员自检]

[项目经理初验]

[独立质检部门终验]

三道质量关卡

必须全部过关

ISO9001质量管理认证

通过ISO9001

质量管理认证

不断提高质量追求

数据安全与合规

数据堂在数据集、定制服务、数据标注平台业务均拥有完善的数据安全合规管理方案
全力保障客户权益，可以放心无忧使用我们的数据产品及服务

业务合规

核心业务安全合规方案
不留死角

安全实施

签署授权协议
系统安全
人员安全

ISO体系认证

通过ISO27701
ISO27001认证

相关资质证书

拥有涉外调查许可
测绘资质证书

DATATANG (BEIJING) TECHNOLOGY CO.,LTD.

THANKS

全 球 领 先 的 人 工 智 能 数 据 服 务 商

北京总部 | 美国 | 日本 | 上海 | 深圳 | 郑州 | 南京 | 保定 | 合肥

数据堂 Datatang