



口音英语数据产品介绍

丰强泽



公司基本情况



新三板挂牌公司：831428

9年

成长历程

1.5亿

注册资金

4亿

融资

1000+家

合作伙伴

20+

发明专利

数据堂-专业的AI数据服务提供商

7家

子公司

2个

数据处理中心

200名

员工

45000套

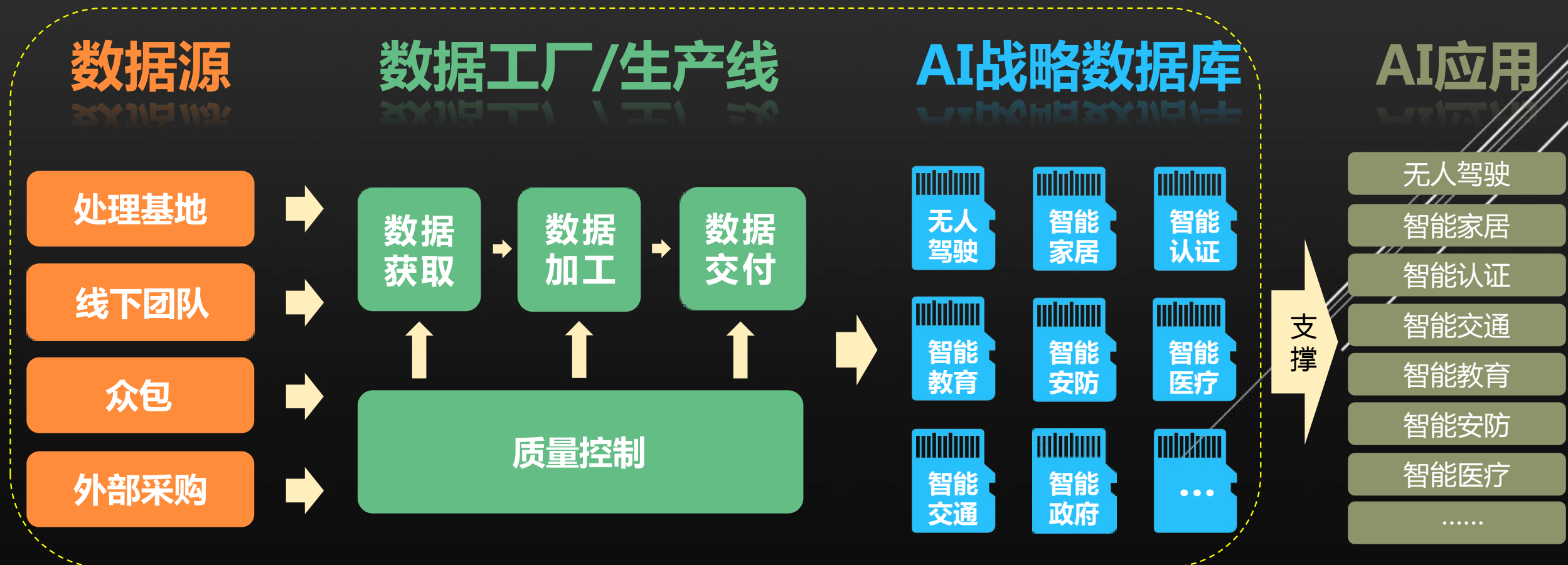
数据量

2500TB

数据规模

公司定位：人工智能数据服务专家

- 致力于打造智能数据工厂，高效采集和处理数据
- 提供类似于石油产业的开采、提炼及成品油服务



我们提供三类服务



数据产品/成品

- 语音识别
- 人脸识别
- 安防场景



定制数据服务

- 语音采集和标注
- 图片采集和标注
- 视频采集和标注



私有化数据工厂部署

- 本地数据处理和标注

语音识别数据概述

原始数据



精细标注

文字：非主流起源于哪里？真正的含义是什么？
性别：男
口音：普通话
噪音：一般噪音

规模与指标

数据规模：十万小时级
语种：重口音、普通话，8大方言、少数民族语言，20多种主流外文（含欧洲语系、亚洲语系、美洲语系）
句准确率：95%~98%
采集设备：50款主流设备（手机、高保真麦克风、麦克风阵列、电话信道）

典型客户



应用场景

手机助手



车载智能交互



智能客服



本次竞赛数据

- 数据堂提供8个国家说英语：每个国家20小时训练集、2小时测试集

数据	数据量
俄罗斯人说英语	20小时训练集、2小时测试集
韩国人说英语	20小时训练集、2小时测试集
美国人说英语	20小时训练集、2小时测试集
葡萄牙人说英语	20小时训练集、2小时测试集
日本人说英语	20小时训练集、2小时测试集
英国人说英语	20小时训练集、2小时测试集
中国人说英语	20小时训练集、2小时测试集
印度人说英语	20小时训练集、2小时测试集

口音英语数据产品列表

■ 各国口音英语9000小时

语种	数据量	数据
美式英语	1216小时	201小时北美英语手机采集语音数据 215小时美式英语手机采集语音数据_朗读 800小时美式英语手机采集语音数据_交互场景
英式英语	1030小时	199小时英式英语手机采集语音数据_朗读 831小时英式英语手机采集语音数据
印度英语	1012小时	1012小时印度英语手机采集语音数据
中国英语	1102小时	593小时中国人说英语手机采集语音数据 509小时中国人说英语手机采集语音数据
西班牙英语	183小时	183小时西班牙人说英语手机采集语音数据
法国英语	520小时	520小时法国人说英语手机采集语音数据
德国英语	535小时	535小时德国人说英语手机采集语音数据
意大利英语	200小时	200小时意大利人说英语手机采集语音数据
葡萄牙英语	200小时	200小时葡萄牙人说英语手机采集语音数据
拉丁英语	117小时	117小时拉美人说英语手机采集语音数据

口音英语数据产品列表

■ 各国口音英语9000小时

语种	数据量	数据
日本英语	500小时	500小时日本人说英语手机采集语音数据
韩国英语	300小时	300小时韩国人说英语手机采集语音数据
俄罗斯英语	500小时	500小时俄罗斯人说英语手机采集语音数据
巴西英语	200小时	200小时巴西人说英语手机采集语音数据
澳大利亚英语	200小时	200小时澳大利亚人说英语手机采集语音数据
加拿大英语	200小时	200小时加拿大人说英语手机采集语音数据
新加坡英语	200小时	200小时新加坡人说英语手机采集语音数据
马来英语	200小时	200小时马来人说英语手机采集语音数据

关键数据指标

- 录音文本：

- 语料类型：

- 通用类：领域不限的句子，来源广泛，包括日常口语、新闻等多种内容
 - 交互类：涉及到音乐、天气、出行、生活等不同领域的问句
 - 家居命令类：涉及到对智能家居设备的控制命令
 - 车载命令类：涉及对车载设备的控制命令
 - 数字类：与数字相关的文本，如日期、货币、时间等

- 做了发音平衡

- 重复使用次数：小于3次


关键数据指标

- 音频格式：16kHz，16bit，未压缩wav，单声道
- 录音环境：相对安静的室内，无回声
- 录音设备：苹果手机或安卓手机
- 录音人：本国入，非英语母语国家选取英文水平较高的人群
 - 性别：男女覆盖均匀
 - 年龄：年轻人为主，兼顾其他年龄人群
 - 地域：覆盖本国各大主要地区
 - 数据量：每人录制450句左右，平均有效时长20-30分钟
- 准确率：句准确率95%以上


关键数据指标

■ 数据堂根据《中华人民共和国民法典》、《网络安全法》、《儿童个人信息网络保护规定》、GDPR、CCPA、《（日本）个人信息保护法》、《（韩国）个人信息保护法》等主要法律政策规定，与数据权利人签署授权书、公示《隐私政策》，以此采集、使用数据时保护其个人隐私或信息。

《告知及授权书》



Document No.: _____



Document No.: _____

Notification and Authorization⁴¹

Datatang (Beijing) Technology Co., Ltd. (hereinafter referred to as "Datatang") is a global data service and data processing company. We are committed to providing data collection and data annotation services for the research field of artificial intelligence. Datatang knows very well the importance of personal information to you. We will protect your personal information and privacy in accordance with laws and regulations. Datatang hereby elaborates on the relevant rules for the collection and use of your personal information through this "Notification and Authorization", and we will use your personal information strictly in accordance with your authorized scope. Although different countries have different definitions of minors/children in their laws, we regard individuals under the age of 14 as minors/children here. If you are under 14 years old, this "Notification and Authorization" should be read and signed by you and your guardian. Please be sure to read the following in detail. Once you or your guardian sign the paper or electronic version of the this "Notification and Authorization", you or your guardian knows and agrees to the following contents on your behalf:⁴²

- 1. [Collection Contents]** Datatang needs to collect your personal information for the project _____ . Collection method _____ . Collection number _____ . If the collected content is personal biometric information, our project manager will clearly inform you of the relevant rules separately.⁴³
- 2. [Information Application]** The information collected by Datatang includes audio, video, image, text, etc., which is used in the field of artificial intelligence. The purpose of collection is not to know about your personal information, but to conduct artificial intelligence model training in a statistical sense after the personal identification is removed.⁴⁴
- 3. [Organization Structure]** Datatang has a complete management organization structure and a data registration management system to ensure that the use of your information is in compliance with the law. Datatang appoints a data protection officer to express an independent opinion and conduct an independent vote right on your personal information. Datatang will use your personal information within the scope of your authorization. If the use of information exceeds the scope of authorization, Datatang will notify you and get your permission to obtain your authorization again.⁴⁵
- 4. [Acquire the Consideration]** Datatang adopts the method of compensable collection . After the collected information meets the requirements of the project, you will get the corresponding consideration. The collection price varies according to the collection range, and the specific amount subjects to the Datatang paid to you. You or your guardian may refuse to sign this "Notification and Authorization and refuse us to collect your information. If you or your guardian refuse, you will not have the right to obtain information collection consideration. Signing this "Notification and Authorization" by you or your guardian means that you or your guardian approve the business activities of Datatang.
- 5. [Ownership of Rights]** Datatang has the property right of your personal information which it collects, including but not limited to the property right of the information and the derivative right of filing a lawsuit or compensation for the infringement. The relevant rights to the new information formed by the processing of the collected information also belong to Datatang, and you will not be entitled to compensate or claim any rights for this part of information.⁴⁶
- 6. [Use of Information]** Datatang and its affiliates or third parties authorized by Datatang will digitize, modify, research, investigate, change, transfer, copy and create derivative works for your personal information repeatedly, which is used on the needs of artificial intelligence model training and it will not pertaining thereto the

distortion of your personal information and seek your further consent. We will not make up any compensation for the above treatment of your personal information.⁴⁷

7. [Authorized Use] For the purpose of business expansion, Datatang may authorize third parties to use your personal information for free or paid, and limit third parties to use your personal information only in the field of artificial intelligence in accordance with the contract. In no case will Datatang make up for your compensation. *Please note that the scope of third parties is not limited to the territory of your country of nationality. When your personal information needs to be used by a third party outside the country, Datatang will strictly abide by the relevant regulations of your country of nationality regarding data exit to ensure that your personal information is used within the legal scope.*

8. [Information Storage] Datatang will take the shortest necessary time for the collection as its storage principle. However, in general, your personal information will become the basic data for artificial intelligence model training under the processing of Datatang, and this data will be of permanent use for artificial intelligence model training. Therefore, if you do not exercise the right to delete your personal information, Datatang will permanently save your personal information in the server and authorize its use for many times. In principle, Datatang will not collect your personal biometric information and personal identification information at the same time. If they are collected at the same time, the information will be stored separately.⁴⁸

9. [Information Security] In order to ensure your information security, Datatang and its affiliates and the third party which is authorized by Datatang will take all appropriate technical (including but not limited to the dedicated server and data encryption, etc.) and organizational measures to prevent your personal information in the storage and transport links from accessing and using by any unauthorized third party. At present, Datatang became international authority of ISO9001 quality management system and ISO27001 information security management system certified.⁴⁹

10. [Notification of Rights] If you or your guardian want to access, check, correct, and update any personal information or found out that your personal information has been misused or illegally disclosed, please contact our Data Protection Officer at 010-53938660. You can also contact Datatang through official website <https://www.datatang.com> or services@datatang.com. If you or your guardian want to delete your personal information or withdraws consent, please contact 400-650-6137 or project leader. Datatang will delete your personal information after you or your guardian have completed the corresponding rescission procedure with Datatang.⁵⁰

11. [Security Events] In case of data leakage, Datatang will release information on the company's official website and other public channels. We will cooperate with relevant national departments to deal with the data leakage. If the leaked data involves yours, you can communicate with Datatang and get the latest progress of the event and obtain corresponding compensation by legal approaches.⁵¹

Signature of Participant:
Printed Name: _____

Date: _____

Signature of Guardian:⁴¹
Printed Name: _____

Date: _____
Relationship with the Participant:⁴² _____

⁴¹

⁴²

⁴³

⁴⁴

⁴⁵

⁴⁶

⁴⁷

⁴⁸

⁴⁹

⁵⁰

⁵¹

数据堂通过授权书及《隐私政策》保障数据权利人个人隐私或信息知情权、处分权等：

- ✓ 告知数据采集范围、使用目的等信息；
- ✓ 告知行使数据修改、删除等操作方式；
- ✓ 履行法律法规其他应尽的隐私保护义务。

数据示例

数据示例：1,012小时印度英语手机采集语音数据

数据量	2100人，1012小时
性别	男性52%，女性48%
地域	覆盖印度主要大邦
年龄	18~25岁80%，25岁以上20%
录音环境	相对安静室内环境，无回声
录音设备	手机
音频格式	16kHz，16bit，无压缩wav，单声道
录音内容	每人录制450句，覆盖通用类、交互类、家居命令类、车载命令类、数字串。做了发音平衡
标注内容	文本
正确率	句准确率95%

正在做的英语数据

各个国家的儿童英语数据

国家	数据量	数据
中国	500小时	500小时中国儿童说英语手机采集语音数据
美国	51小时	51小时美国儿童麦克风采集语音数据
英国	55小时	55小时英国儿童麦克风采集语音数据

正在做的英语数据

自然对话风格的英语数据

数据	1000小时美式英语自然对话
数据量	2000人，1000小时
性别	女性50%，男性50%
语言	美式英语（美国人录制）
录音环境	真实家居场景；安静环境
录音设备	手机
音频格式	手机：16kHz，16bit，wav，单声道
录音内容	两人对话。给出话题列表，录音人从中挑选多个自己熟悉的话题以确保对话的流畅自然，围绕每个话题展开一段对话并录制
标注内容	文本，起止时间点，说话人标识
正确率	句准确率95%

正在做的英语数据

百万单词级的英文发音词典：所有单词的音标均由人工标注，预计12月底完成

单词	英音	美音
disobeying	[,disə'beɪŋ]	[,disə'beɪŋ]
archbishopric	[,ɑ:rtʃ'biʃəprɪk]	[,ɑ:tʃ'biʃəprɪk]
partook	[pɑ:'tʊk]	[pɑ:r'tʊk]
contriving	[kən'traɪvɪŋ]	[kən'traɪvɪŋ]
dietetics	[,daɪə'tetɪks]	[,daɪə'tetɪks]
drays	[dreɪz]	[dreɪz]
bullfrogs	['bʊlfrɔ:gz]	['bʊlfrɑ:gz]
dilating	[daɪ'lertɪŋ]	[daɪ'lertɪŋ]
highnesses	['haiməsɪz]	['haiməsɪz]
alb	[ælb]	[ælb]
grammarians	[grə'meərɪənz]	[grə'merɪənz]
woodsmen	['wʊdzmən]	['wʊdzmən]
parboiled	['pɑ:bɔɪld]	['pɑ:rbɔɪld]
sandboxes	['sændbɒksɪz]	['sændbɑ:ksɪz]
palpated	[pæl'pertɪd]	[pæl'pertɪd]
counterpoints	['kaʊntəpɔɪnts]	['kaʊntərpoɪnts]
legitimizing	[li'dʒɪtɪməɪzɪŋ]	[li'dʒɪtɪməɪzɪŋ]
reappearances	[,ri:ə'piərənsɪz]	[,ri:ə'piərənsɪz]
terrapins	['terəpɪnz]	['terəpɪnz]

数据堂

<http://www.datatang.com>