公司代码：688787　　公司简称：海天瑞声

北京海天瑞声科技股份有限公司 2023年年度报告摘要

第一节 重要提示

1 本年度报告摘要来自年度报告全文，为全面了解本公司的经营成果、财务状况及未来发展规划，投资者应当到www.sse.com.cn网站仔细阅读年度报告全文。

2 重大风险提示

2023年，公司营业收入较去年同期下滑35.33%，扣非前后归母净利润均为负数。报告期内，受境外部分客户进行阶段性裁员、业务方向及研发节奏周期性调整等影响，部分客户2023年预算释放进度放缓，同时叠加2023年上半年数据出境相关法规落地实施的阶段性影响，公司境外收入同比大幅下滑。境内业务方面，虽然宏观稳经济政策已初见成效，但国内仍面临复杂严峻的内外部环境考验，部分境内客户对集中性研发投入仍持谨慎态度，基础数据服务领域客户预算及需求释放出现阶段性减缓，叠加行业内竞争加剧，综合导致境内收入同比下滑。此外，为配合整体战略发展及业务拓展目标，公司在营销体系建设等方面加大投入，使得销售费用同期较大幅度增长。与此同时，计提坏账金额阶段性增加、持有外币资产增值幅度较同期下降等因素，导致信用减值损失以及财务费用同比增长。以上因素共同导致2023年度归属于母公司所有者的净利润、归属于母公司所有者的扣除非经常性损益的净利润显著下滑并且出现亏损。上述不利因素目前已有所改善，但如果公司收入增长无法覆盖各类投入及期间费用支出，公司业绩存在下滑或亏损的风险。

公司已在本报告中详细描述可能存在的风险，敬请查阅“第三节管理层讨论与分析”之（四）“风险因素”部分，请投资者注意投资风险。

3 本公司董事会、监事会及董事、监事、高级管理人员保证年度报告内容的真实性、准确性、完整性，不存在虚假记载、误导性陈述或重大遗漏，并承担个别和连带的法律责任。

4 公司全体董事出席董事会会议。

5 信永中和会计师事务所（特殊普通合伙）为本公司出具了标准无保留意见的审计报告。

6 公司上市时未盈利且尚未实现盈利

□是　√否

7 董事会决议通过的本报告期利润分配预案或公积金转增股本预案

综合考虑公司目前经营状况以及未来发展需要，为保障公司生产经营的正常运行，增强抵御风险的能力，实现公司持续、稳定、健康发展，更好的维护全体股东的长远利益，公司2023年利润分配预案为：不派发现金红利，不进行公积金转增股本、不送红股。以上利润分配方案已经公司第二届董事会第二十四次会议和第二届监事会第二十三次会议审议通过，尚需公司2023年年度股东大会审议通过。

8 是否存在公司治理特殊安排等重要事项

□适用　√不适用

第二节 公司基本情况

1 公司简介

公司股票简况

√适用　□不适用

公司股票简况				
股票种类	股票上市交易所及板块	股票简称	股票代码	变更前股票简称
人民币普通股（A股）	上海证券交易所科创板	海天瑞声	688787	不适用

公司存托凭证简况

□适用　√不适用

联系人和联系方式

联系人和联系方式	董事会秘书（信息披露境内代表）	证券事务代表
姓名	吕思遥	张哲
办公地址	北京市海淀区知春路68号院1号楼4层401	北京市海淀区知春路68号院1号楼4层401
电话	010-62660772	010-62660772
电子信箱	ir@dataoceanai.com	ir@dataoceanai.com

2 报告期公司主要业务简介

（一）主要业务、主要产品或服务情况

1. 主要业务情况

公司主要从事AI训练数据的研发设计、生产及销售业务。公司通过设计数据集结构、组织数据采集、对取得的原料数据进行加工，最终形成可供AI算法模型训练使用的专业数据集，通过软件形式向客户交付。

自2005年成立以来，公司始终致力于为AI产业链上的各类机构提供算法模型开发训练所需的专业数据集。经过多年发展，公司已成为人工智能基础数据服务领域具有较强国际竞争力的国内头部企业，并实现了标准化产品、定制化服务、相关应用服务全覆盖。公司所提供的训练数据涵盖智能语音（语音识别、语音合成等）、计算机视觉、自然语言等多个核心领域，全面服务于人机交互、智能家居、智能驾驶、智慧金融、智能安防等多种创新应用场景。

公司的产品和服务已获得阿里巴巴、腾讯、百度、科大讯飞、海康威视、字节跳动、微软、亚马逊、三星、中国科学院、清华大学等国内外客户的认可，应用于其研发的个人助手、智能音箱、语音导航、内容生成、搜索服务、短视频、虚拟人、智能驾驶、机器翻译等多种产品相关的算法模型训练过程中。目前公司客户累计数量超过930家，覆盖了科技互联网、社交、IoT、智能驾驶、智慧金融等领域的主流企业，教育科研机构以及部分政企机构。

图：公司产品服务矩阵示意

2. 主要产品及服务情况

2.1 主要产品及服务按业务类型分类

公司研发、生产的训练数据覆盖了智能语音、计算机视觉及自然语言处理三大AI核心领域，广泛应用于算法模型的开发、训练、优化、应用场景拓展等环节。此外，公司还提供与训练数据相关的应用服务。

（1）智能语音

人工智能在语音领域的应用技术主要包括语音识别、语音合成等。

语音识别（Automatic Speech Recognition，ASR）是让机器能够“听懂”人类语音的技术，它能使机器自动将语音信号转换为对应的文本信息。

语音合成（Text to Speech，TTS）是让机器能够“说出”人类语音的技术，它使机器能将文字信息转化为流畅的语音“朗读”出来，相当于给机器安上了人工嘴巴。

以日常生活中的情景为例，语音输入法、即时通讯软件运用了语音识别技术将用户输入的语音实时转换为文字，实现了软件“听懂”语音并“听写”出文字的效果；而地图、导航软件则运用语音合成技术，实现了软件“发声说话”的效果，为用户提供即时语音导航。

公司通过设计（设计训练数据集结构、供发音人朗读录制的语料文本或对话场景、发音人分布、录音设备场景等）、采集（定义合适的发音人、选取录音设备及软件、组织发音人朗读录制音频）、加工（对音频文件进行切分、标注各类声音特征，形成带时间戳和特征标签的文本和标注文件等）、质检（对数据集进行质量检测，如音字一致性、标注准确率检查等）等训练数据集生产环节；或者针对客户提供的原料音频文件执行加工、质检工作，最终形成客户所需的智能语音训练数据集。

（2）计算机视觉

计算机视觉（Computer Vision，CV）是使机器具备“看”的功能的技术，它使得智能驾驶、智能家居、手机、安防设备等机器能够代替人眼对目标进行识别、跟踪和测量等。

以日常生活中的情景为例，在汽车的自动驾驶功能中，计算机视觉技术使得汽车能够“看见”并识别行车过程中的各种行人、路况场景，为后续作出相应的反应奠定基础；在机场、车站安检中，计算机视觉技术使得人脸识别设备能够识别被检验人员是否为其出示的身份证件显示的人员。

公司通过设计训练数据集结构、采集（如定义合适的人脸、动作、场景作为采集对象，组织被采集人按照要求拍摄照片、录制视频等）、加工（对图像、视频文件进行打点、拉框、分割标注等）、质检（对数据集进行质量检测，如检验图片、视频文件格式是否正确，检查光照环境、物体种类的数量是否达标，打点标框的准确率是否符合要求等）；或者对客户提供的图像、视频文件执行加工、质检工作，最终形成客户所需的计算机视觉训练数据集。

（3）自然语言处理

自然语言处理（Natural Language Processing，NLP）是以机器能够像人一样理解语言意图的技术。

以日常生活中的情景为例，寄送快递时使用的“智能填写”功能即运用了自然语言处理技术，在输入框中填入整段联系信息，软件应用能够理解语义，并从中识别及提取“收件人”、“联系方式”、“地址信息”等所需信息，完成自动填写；智能客服、聊天机器人等人机交互程序也运用了自然语言处理技术，使得程序、机器能够读懂人类语言的真正意图，并相应做出反应、提供服务等。

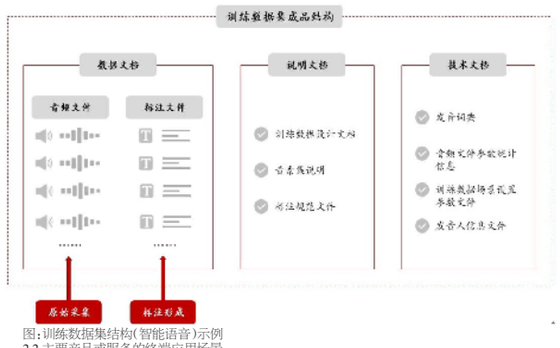
公司通过设计训练数据集结构、采集（收集或编写自然语言文本、对话等数据信息）、加工（对自然语言文本数据进行单词分割、词性标注、语义语法标注、情感属性标注等）、质检（对数据集进行质量检测，如检验文本、词性或者语义的标注结果是否准确等）；或者对客户提供的自然语言文本执行加工、质检工作，最终形成客户所需的自然语言训练数据集。

（4）训练数据相关的应用服务

公司基于自身生产的训练数据提供算法模型相关的训练服务，运用训练数据研发能力助力下游客户完成其算法模型的语言拓展、特定算法模块拓展、垂直应用领域拓展等，为客户定制针对特定应用场景的专属算法模型，提高AI技术应用效果。

前述产品、服务均以公司生产的专业训练数据集为核心或基础。公司通过设计训练数据集结构、组织原料数据采集、对取得的原料数据进行加工，最终形成可供算法模型训练使用的专业数据集。

成品训练数据集主要由数据文档、说明文档、技术文档三部分构成。以智能语音训练数据集为例，成品训练数据集包含原始采集形成的音频文件、与音频文件对应的带有时间戳的标注文件，训练数据集相关的设计文档、训练数据集说明，发音词典，数据集参数信息文件等，图示如下：



图：训练数据集结构（智能语音）示例

2.2 主要产品或服务的终端应用场景

公司提供的高质量、大规模、结构化的训练数据，为算法模型的训练拓展提供了可靠的训练素材，助力AI技术实现实践应用及商业化落地，赋能AI技术与实体经济深度融合。公司提供的训练数据广泛应用于众多主流AI产品及终端应用的训练过程中，覆盖了个人助手、语音输入、内容生成、智能家居、机器人、语音导航、智能客服、智能播报、语音翻译、移动社交、虚拟人、智能驾驶、智慧医疗、智慧教育、智慧交通、智慧城市、智慧金融、机器翻译、智能问答、信息提取、情感分析、OCR识别等多种应用场景。

图：训练数据集服务的算法模型应用场景示意

（二）主要经营模式

1. 盈利模式

与主要产品及服务类型对应，公司的盈利模式主要包括以下三类：

（1）定制服务：公司根据客户需求提供定制训练数据集并收取服务费。在此种模式下、公司享有服务费收入，不享有最终生成的训练数据的知识产权、不可将此类业务生产的训练数据向其他客户重复销售。

（2）标准化产品：公司开发自有知识产权的训练数据集产品，通过销售训练数据集产品的使用授权许可，获取让渡资产使用权收入。此类训练数据集一经开发完成，可多次销售并获取授权许可收入。

（3）训练数据相关的应用服务：公司基于生产的训练数据提供算法模型相关的模型拓展及训练服务，通常以软件授权或软硬件一体化形式交付算法模型拓展、开发成果，获取让渡资产使用权收入和技术服务收入，以及少量硬件销售收入。

2. 生产或服务模式

（1）训练数据集生产模式

公司通过设计训练数据集结构、组织原料数据采集、对取得的原料数据进行加工，最终形成可供算法模型训练使用的专业数据集。

图：训练数据生产过程示意图

公司的训练数据生产过程主要包括四个环节：设计（训练数据集结构设计）、采集（获取原料数据）、加工（数据标注）及质检（各环节数据质量、加工质量检测）。

（2）训练数据相关的应用服务模式

公司基于其生产的训练数据提供算法模型相关训练服务，助力下游客户完成其算法模型的语言拓展、特定算法模块拓展、垂直应用领域拓展等，为客户定制针对特定行业和口音的专属算法模型，提高AI技术应用效果。

以某大型科技公司客户项目为例，客户研发了特定语音识别算法模型，需要根据算法模型的实际场景（如法院庭审场景）开发落地应用。公司承担了部分落地应用拓展相关的开发工作，围绕客户的算法模型和接口开发，最终协助客户算法模型实现多个麦克风收集庭审语音内容并实时转成文字记录入系统的功能。

3. 采购模式

按照采购的内容及主体划分，公司的采购包括：

数据服务采购：公司在数据采集、加工环节中，向人力资源服务等类型的公司等供应商采购的，非核心技术环节的原料数据采集、标注服务。

岗位服务采购：主要针对临时性的、不设长期岗位的业务领域的外包采购，如保洁、临时招聘服务、少量实习生招聘等。

其他采购：（1）训练数据生产所需的资产，主要包括软、硬件设备及其他需求物品采购；（2）日常运营所需的资产及物品，如办公用房、车辆、办公家具、计算机设备等；（3）日常专项服务采购等，主要包括审计服务、会议服务、差旅服务等。

上述原料数据采集、加工环节所涉及的数据服务采购，为公司最主要的采购类别，由集采中心负责；各部门岗位服务采购由人力资源部负责；其余日常运营相关的资产物品采购、专项服务采购等非业务采购由集采中心负责。财务中心负责参与采购供应商的遴选、监督与管理，并对采购费用进行核算及结算。

经过多年的发展，公司已经建设有完善的《海天瑞声采购管理制度》、《海天瑞声项目资源采购管理制度》、《海天瑞声供应商管理制度》、《海天瑞声岗位服务采购管理制度》等内部规范制度，设立有完善的采购流程和体系，并与主要的供应商形成了良好稳定的长期合作关系。

4. 营销模式

公司采用直接对接并服务客户的直销模式进行营销，符合行业通行惯例。公司以高质量的训练数据集及相关服务吸引客户，并在持续服务客户的过程中提升服务价值和客户黏度。公司通过直接拜访潜在客户、参与学术会议和行业展会新产品发布、搭建并持续升级公司官方网站和建立自媒体矩阵等方式提升品牌知名度、开拓新客户，后续再通过商务谈判、招投标等形式获取具体业务机会。

（三）所处行业情况

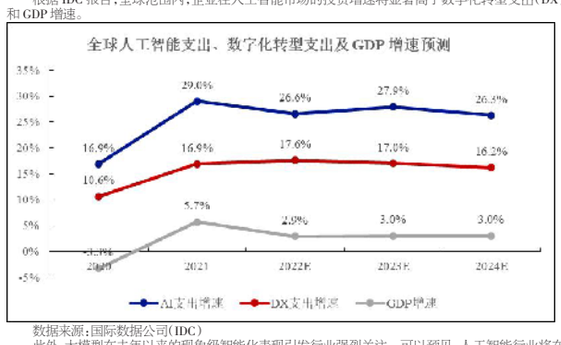
1. 行业的发展阶段、基本特点、主要技术门槛

1.1 行业的发展阶段、基本特点

（1）在数字经济发展以及大模型技术的共同驱动下，全球AI产业进入新一轮加速发展期

当前，新一轮科技革命和产业变革深入发展，数字化转型成为大势所趋，世界主要国家均高度重视发展数字经济，纷纷出台战略规划，重塑数字时代的国际竞争新格局。人工智能作为数字经济发展的底层核心技术之一，正在发挥更加重要的作用。例如，随着数字经济发展的不断深入，数据体量以及复杂度均不断提升，为更好解决产业数字化中数据提取、处理、分析等工作，将会产生更多样化的人工智能需求，人工智能支出也将成为支持企业数字化转型支出的主力因素之一。

根据IDC报告，全球范围内，企业在人工智能市场的投资增速将显著高于数字化转型支出（DX）和GDP增速。



数据来源：国际数据公司（IDC）

此外，大模型在去年以来的现象级智能化表现引发行业强烈关注。可以预见，人工智能行业将在大模型技术的推动下进入新一轮产业高速发展期。

未来，受益于数字经济政策和大模型技术的双重驱动，人工智能将具备更强的产业融合能力，并将深刻影响千行百业的运行规则，以及人们的生活方式，人工智能产业的发展将随之进入快车道。

根据国际数据公司（IDC）的数据，2021年，全球人工智能市场规模为885.7亿美元，预计2025年将达到2,218.7亿美元，年复合增长率达到26.2%。

数据来源：国际数据公司（IDC）

在市场需求拉动和国家政策的支持引导下，当前我国人工智能产业加速发展，已形成基础底层设施、中层技术以及上层应用的完备的产业链生态，一批创新活跃、特色鲜明的创新企业不断涌现，并联合推动中国人工智能产业实现规模增长。根据艾瑞咨询的数据显示，2022年中国人工智能产业规模达1,948亿元，预计2027年市场规模将达到6,122亿元，年复合增长率为25.6%，主要与智算中心建设以及大模型训练等需求拉动的AI芯片市场、无接触服务需求拉动的智能机器人及对话式AI市场等快速增长相关。有望在下游制造、交通、金融医疗等多领域不断渗透，实现大规模落地应用。

数据来源：艾瑞咨询

（2）训练数据作为AI发展和演进“燃料”的作用更加凸显，并逐渐成为大模型竞赛中的重要决定性要素

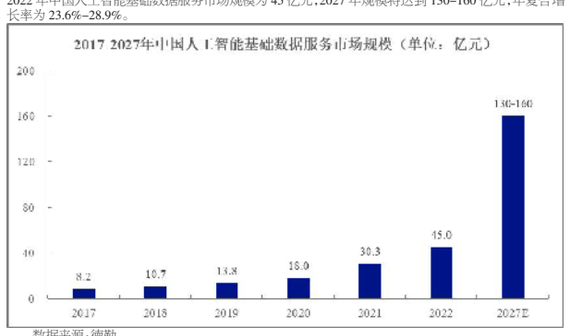
在AI产业链中，算法、算力和数据共同构成技术发展的三大核心要素。算法模型从技术理论到应用实践的落地过程依赖于大量的训练数据。训练数据越多、越完整、质量越高，模型推断的结论越可靠。过去十年，人工智能产业以算法为核心，通过深度学习算法的不断创新，推动人工智能产业的快速发展。但未来，当算法发展趋于开源、算力能力大幅提升及人工智能模型从技术理论应用到更多的垂直场景，想要更快更好提升人工智能能力，数据将发挥更重要的作用。

尤其在大模型时代下的今天，数据正在被视为大模型落地以及竞赛中重要的决定性要素。在大模型领域，过去业界普遍认为模型参数量是模型效果增强的核心要素，模型参数越大，性能表现越好，而如今这一“参数”定律正在打破。Meta开发的新模型Llama（Large Language Model Meta AI）证明，相比于单纯参数量提升，训练数据规模以及多样性的增强，可带来更好的模型效果提升。根据新浪财经报道，Llama-13B虽然在参数规模上相较于GPT-3（175B）小了十几倍，但由于其大幅提升了训练数据规模（Llama-13B训练数据量约为GPT-3的3倍），其表现能力在大部分指标上均超越了后者；与此同时，Llama-65B也是凭借数据规模优势，与谷歌5400亿参数的PaLM-540B在表现上旗鼓相当。

可以看出，数据正在逐渐成为大模型时代下的重要推动力量，并产生快速增长的数据需求。根据Cognilytica数据统计显示，2021年全球AI训练数据市场需求约为42亿美元，并预计到2027年这一需求将增长到220亿美元，2021-2027年复合增长率达32%。

数据来源：Cognilytica

中国作为全球人工智能产业增速最快的国家之一，相关数据需求也在快速增长。根据德勤数据，2022年中国人工智能基础数据服务市场规模为45亿元，2027年规模将达到130-160亿元，年复合增长率为23.6%-28.9%。



数据来源：德勤

（3）数据要素市场蓬勃发展，数据行业迎来更为广阔的发展机遇

近年来，我国数字经济蓬勃发展，数据要素因具有基础性战略资源和关键性生产要素的双重属性，相关市场规模持续增长。尤其在《中共中央、国务院关于构建数据基础制度更好发挥数据要素作用的意见》出台后，我国系统性布局了数据基础制度体系的“四梁八柱”，加速了数据流通交易和数据要素市场发展，进一步推动了公共数据、企业数据、个人数据合规高效流通使用。为更好响应中央号召，北京、上海、广州、深圳、杭州等地数据政策陆续出台，逐步构建了多层次、多元化数据要素市场生态体系。

以北京为例，《关于更好发挥数据要素作用进一步加快发展数字经济的实施意见》（北京市促进通用人工智能创新发展的若干措施）和《关于推进北京市数据专区建设的指导意见》指出，北京市要加快建设“数据基础制度先行先试示范区”（以下简称“先行先试示范区”），“支持北京经济技术开发区等开展数据基础制度先行先试，打造政策高地、可信空间和数据工场”，探索打造数据训练基地，归集高质量基础训练数据集，推动数据要素高水平开放，提升本市人工智能数据标注库规模和质量，并建设针对重大领域、重点区域或特定场景建设专题数据区域，吸纳市场主体和数据、技术、资本等多元要素参与。北京市陆续出台的多项文件旨在打破数据壁垒，推动数据融合利用，加快推动公共数据开放，促进数据要素流通，激发数字市场创新活力，释放和发展数字化生产力，打造多层级数据要素市场，成为具有竞争力和影响力的数字产业集群。按照“政府引导、市场运作、创新引领、安全可控”的原则，“先行先试示范区”有望成为国际领先的数据要素高效流通核心枢纽。

数据要素市场受政策推动，进入高速发展期，未来围绕数据的价值利用以及流通交易，将产生大量新增数据需求，为数据行业开拓了新的增长空间，提供了新的业务拓展机遇。未来，数据要素也将成为数据行业增长的重要推动力量。

（4）训练数据领域的未来发展趋势

a.大模型技术的突破和跃升，将驱动新型数据需求持续增长

随着ChatGPT成为全球范围内的现象级应用，人工智能迎来了新的发展机遇，其背后的大模型技术也将进一步引导人工智能产业变革并带来相关数据需求的变化和增长。

首先，和传统的深度学习模型相比，大模型的数据需求规模将呈指数级增长。传统深度学习技术路线下，训练一个专有小模型大约需要GB级数据，而训练一个大模型通常需要TB级数据。此外，大模型数据丰富程度显著增加，大模型不仅包含海量语言类、知识类信息，还包括各类垂直领域以及多模态数据，通过多样化数据的引入，大幅提升模型的通用能力以及迁移能力，并使其可服务更多的任务类型与场景。同时，数据质量会显著拉开大模型预训练阶段的效果差距。另外，相比于传统模型训练，大模型的数据需求类型也将有所转变，更多模型或将采用类强化学习模式来进行特定领域或特定方向上的优化迭代，以使得机器能够以更加接近于人类期望的方式提供答案输出。对于大模型训练而言，不仅需要持续获取大规模、多样化（多模态、多场景、多垂向）、高质量的数据，更须具备持续迭代的高质量数据清洗和标注策略，以不断提升包括预训练（Pre-training）、模型微调（Fine-tune）及奖励模型（Reward Model）等过程中所需数据（例如指令（instruction）类数据）的质量，确保语言类和常识性知识之外的其他垂直领域的应用场景的能力提升，为大模型精确性、通用性及泛化能力的实现奠定坚实基础。

在以上背景下，一方面，大规模、多样化、高质量数据集重要性凸显，成为模型训练效果的核心支撑之一。另一方面，AI发展所面对的数据前沿性及工程化技术的挑战也更为显著。长期看，只有AI数据处理技术的不断拓新与发展，才能及时适应甚至超前引领大模型技术和应用的发展。

b.多模态数据受AIGC、虚拟人等应用发展驱动，将呈现快速增长趋势

随着AIGC技术发展，AI可在更多维度、更多场景辅助人类进行内容生产以及创作。例如，通过大模型等AIGC技术，人类仅需输入一段简单的文字指令，AI即可按照人类描述生成一幅画、一段语音或一段视频，以此帮助人类完成内容创作。想要实现上述功能，AI除了要具备理解人类文字指令的能力，还需要通过对齐不同独立模态关键特征的方式，建立文字与图、语音、视频等一一映射关系，这背后将依赖大量的多模态数据，AI需要对多模态数据进行学习，以实现跨模态的创作能力。

此外，随着AI虚拟主播、虚拟学生、虚拟员工轮番上岗，数字人概念逐渐走入大众视野，成为人工智能的热门技术赛道。想要让虚拟数字人实现与人类的自然交互，不仅需要发音标准自然、身体动作流畅，其表情、口型与声音也要实现细节的精准匹配，而多模态技术就是打破传统人工智能单一感官局限、让各类AI能力协同使用的重要技术。通过对高质量多模态训练数据集的持续学习，AI可实现图像、视频、音频、语义文本等多维度能力的融合，使得虚拟人在行为上更接近人类。

未来，随着以AIGC、虚拟人为代表的AI技术以及应用的不断发展，多模态数据需求将呈现加速增长趋势，具备多模态数据服务能力，以及多模态数据集储备的企业将获得更多市场机会。

c.人工智能企业全球化布局加速，多语种能力成为企业业务拓展核心支撑

2013年，共建“一带一路”的倡议正式面世，十年来，随着国家“一带一路”战略的深入推进，国内一批具有较强创新能力和过硬技术实力的企业，纷纷踏出国门，积极拓展海外市场，通过不断扩大企业出海战略版图，获得高速发展机会。另一方面，境外头部企业也继续践行“全球化”战略，搭乘全球出海的快车。

随着境内、外企业的全球化扩张成为确定性趋势，多语种能力作为支撑企业顺利出海的核心要素之一，重要意义更加凸显。未来，多语种训练数据将对客户侧在语音助手、智能汽车、智能家居、智能客服、机器人、多语种OCR等各领域产品应用的全球化推广起到积极作用。因此，随着各类客户群体扩张步伐加速，多语种需求也将快速增长，具有强大语言研究能力的数据服务企业将获得更多商业机会。

d.人工智能技术加速向产业渗透融合，催生更多垂向领域数据需求

随着深度学习技术的不断突破，人工智能发展已经进入2.0时代，相关训练需求正逐渐从通用基础能力建设，向更为专业的垂向场景/行业拓展。一方面，以大模型为代表的AI基础技术不断取得重大突破，AI模拟人类认知的能力飞速提升，因此从技术能力维度看，AI已具备与垂直产业融合并规模化应用的前提条件；另一方面，受国家数字经济发展战略推动，产业数字化和智能化将进一步席卷各行各业，智能化技术与传统产业的融合将成为数字经济时代的新发展趋势，并创造出巨大的蓝海空间。

当前AI技术正在加速与各类产业融合，在汽车、金融、医疗、工业等传统行业的渗透率和应用场景不断提升，展现出可观的商业价值和巨大的发展潜力，而数据作为打通算法技术与行业需求的核心桥梁，作用更加凸显，可以说数据能力一定程度上决定了算法模型在对应产业的适用性以及实用性，成为加速AI产业化落地的关键要素。

e.国家法律法规密集落地，对数据安全及合规提出更高要求

近年来，数字经济规模快速扩张，数据作为数字经济时代核心生产要素、重要性更加凸显，但数据不同于传统生产要素，其中可能涉及个人隐私以及国家安全的重要信息，因此，为更好保障数字经济长期稳定的可持续发展，建设规范、安全、合规、高质量的数据安全体系已成为迫切需求。近年，国家陆续出台包括《数据安全法》、《个人信息保护法》等主流法律法规，为解决数据安全问题、净化行业快速发展中的不良乱象提供了切实可行的法律依据。

未来，随着AI技术不断革新，应用行业以及场景不断增加，各行业、各领域数据安全规范逐渐落地将成为趋势，对于以数据生产为主营业务的数据服务企业，数据安全及合规能力将成为数据服务能力的核心评价维度，成熟的安全合规管理体系将成为重要评价标准，能持续跟踪法律环境变化，积极响应监管政策，牢牢把握发展与安全并重的原则的企业将具有更强的市场竞争力。

1.2 行业的主要技术门槛

随着AI技术不断演进、产业应用不断丰富，训练数据的市场需求呈现体量、难度、复杂性、合规性持续上升的趋势，数据服务商须同时具备对人工智能核心算法的理解能力、前瞻性的专业数据集设计能力、丰富的语言覆盖能力及场景采集能力、算法辅助数据生产能力、以及数据合规管理能力，这使得行业的技术门槛持续提升，具体体现为：

（1）在训练数据研发、生产全流程中的算法全面介入

随着AI技术应用落地的规模化效应凸显，客户对于数据规模和处理效率的要求不断提升，数据服务商须在研发、生产流程中全面引入算法以实现高效、合理的人机协作模式，进而实现降本增效的目标。一般而言，在训练数据研发、生产全流程中融入算法技术，可用于训练数据集的设计及训练数据生产的各个环节，例如调度不同类型的标注人员应对不同领域的任务、形成算法自动处理能力以帮助标注人员提升效率、降低对人员的依赖（既有人员数量的降低、也有对人员标注能力要求的降低），并构建训练数据设计、加工相关的核心技术；也可用于检查训练数据集对算法模型的训练效果，进而保障训练数据集质量。

（2）平台工具链功能及适配性要求持续提升

当前，客户侧的数据采集、标注需求范围在逐渐拓宽，数据采集与标注需满足的AI应用场景比以往明显更加广泛、复杂，这就对数据服务商的平台工具能力提出了更高要求，平台上处理过多大规模的数据、这些处理过的数据的多样性和复杂程度如何、算法引擎投票机制如何建立、置信区间如何设置、算法在平台中如何应用、数据流转的工程化程度如何等等这些因素都决定了平台的适配性和能力如何，并最终决定了数据处理的质量、效率、成本。

（3）语音语言学基础研究方面须有深厚积累

伴随语音技术进一步发展落地、并向各行各业和更多垂直场景不断渗透，同时受到中国企业出海需求、国外企业区域拓展需求两方面的支撑，客户在多语种、多音色、音素集、发音规则、发音词典等方面的要求在不断抬升，这意味着只有那些在语音语言学基础研究方面投入更多、拥有深厚积累的数据服务商才能满足客户在这方面的多元化需求。

因此，市场上仅有极少数企业通过长期自主研发的方式能够达到上述核心技术门槛，成为有能力向不同客户群体提供综合、高效、合规的数据产品及服务的供应商。

2. 公司所处的行业地位分析及其变化情况

作为行业的头部阵营企业，海天瑞声在经营情况、技术实力、以及以数据安全为代表的其他综合能力方面都展示出明显优势，并具有较强国际竞争力。近年来公司紧跟AI技术发展趋势，尤其关注在客户资源、技术实力、产品/服务等方面的竞争优势，树立国内领先基础数据服务商的品牌形象，以巩固公司的行业领先地位。与同行业国内外竞争对手的对比情况及优势体现如下：

公司	海天瑞声	Appen	数据堂	标贝科技
基本经营情况				
成立年份	2005年	1996年	2010年	2016年
市场地位概述	是我国最早从事训练数据研发销售的企业之一；国内首家且是目前唯一一家A股上市的人工智能训练数据服务企业	较早从事数据资源开发的数据资源产品服务提供商，经营历史较长，规模、体量较大	新三板挂牌企业，是国内较早从事数据交易、数据采标的服务商之一	-
员工数量	226人	1,037人	292人	未公开披露
主要客户/合作伙伴情况	大型科技公司，如阿里巴巴、Meta、腾讯、百度、字节跳动、微软、三星等；人工智能企业，如科大讯飞、商汤科技、云知声、海康威视等；科研机构，如中国科学院、清华大学、中国科学技术大学等	微软、亚马逊、谷歌等大型科技公司、汽车厂商及政府	包括百度、腾讯、阿里巴巴、奇虎360、联想、科大讯飞等国内互联网和高科技企业，微软、NEC、Canon、Intel、Samsung、Fujitsu等企业及在华研发机构	微软、百度、阿里、腾讯、京东、滴滴、字节跳动、网易、360、三星、小鹏、美的、中科大、中电科、中国银行等
客户数量	超过930家	未公开披露	未公开披露	100余家
技术研发及产品能力				
技术实力概述	海天瑞声拥有自主研发的一体化数据处理平台，所提供的训练数据涵盖智能语音、计算机视觉、自然语言等多个AI核心领域，可服务于个人助手、语音输入、内容生成、机器人、智能驾驶、智慧医疗、智慧教育等22种创新应用场景。	Appen拥有人工智能辅助数据注释平台，在全球170多个国家与100多万名专业承包商合作，训练数据涵盖科技、汽车、金融服务、零售、医疗健康和政府等各个领域。	拥有人工智能数据与生产服务平台，可提供数据定制服务、人工智能数据集产品、人工智能数据处理平台私有化部署服务，数据采集范围遍及全球30多个国家，合作伙伴遍布世界10多个国家。	拥有语音合成模型和算法，通过算法+专业的人工数据处理方式，为客户提供优质的语音合成服务，拥有TOBI标注体系，通过自主研发的TTS评测系统，为客户提供高质量的数据服务。
应用领域	智能语音、计算机视觉、自然语言	智能语音、计算机视觉、自然语言	智能语音、计算机视觉、自然语言	智能语音、计算机视觉、自然语言
拥有的成品训练数据集数量	1,558个	超过600个	291个	190个
语种/方言覆盖能力	超过205个	超过235个	100余个	10余个
已取得专利授权	37项（35项发明专利、1项实用新型专利及1项外观设计专利）	4项	36项（截至2023年12月31日）	20项（截至2023年12月31日）
计算机软件著作权数量	173项	未公开披露	198项（截至2023年12月31日）	30项（截至2023年12月31日）
综合能力				
数据安全能力	乙级测绘资质；ISO27001信息安全管理体系认证、ISO27701隐私信息管理体系认证；信息系统安全等级保护三级；中国信通院数据安全推进计划成员单位；数据知识产权登记	未公开披露	乙级测绘资质；ISO27001信息安全管理体系认证、ISO27701隐私信息管理体系认证；数据知识产权登记	ISO27001信息安全管理体系认证、ISO27701隐私信息管理体系认证、ISO27017云服务信息安全管理体系认证、ISO27018公有云中保护个人身份信息的信息安全管理体系认证、信息系统安全等级保护二级
资质荣誉	国家高新技术企业、国家专精特新“小巨人”企业、“北京市企业技术中心”、工信部“新一代人工智能产业创新重点任务揭榜优胜单位”等多个国家或市级重要奖项、北京数字经济企业100强、第一批入选北京市通用人工智能产业创新伙伴计划	不适用	国家高新技术企业、国家专精特新“小巨人”企业、中国自动化学会CAA科技进步一等奖	国家高新技术企业、中关村高新技术企业、北京市专精特新“小巨人”企业、优秀服务机器人企业奖

注1：数据堂、标贝科技数据：除特别标注外，均为2023年1-6月/截至2023年6月30日数据，前述公司官网及公开披露信息；国家知识产权局中国及多国专利审查信息查询平台（https://www.cnipa.gov.cn/）、中国版权保护中心CPCC微平台等公开信息查询渠道及第三方机构查询信息。

注2：海天瑞声、Appen数据：均为2023年1-12月/截至2023年12月31日数据。

3. 报告期内新技术、新产业、新业态、新模式的发展情况和未来发展趋势

（1）数据需求向海量、高质量、多元化方向演进，智能化水平成为数据服务商核心竞争力

随着AI应用场景日益丰富、以及产品智能化要求的不断提升，客户在数据规模、质量、多元化等方面提出了更高的要求。以智能语音和计算机视觉领域为例，训练数据需求逐渐拓展至更多语种、更复杂场景、更多AI设备、更多音色、更多维的图像采集等维度，数据服务商除了要具备丰富的数据采、标经验，还需要拥有完善的多元化数据处理平台，同时，通过引入算法提升数据处理的质量和效率，降低成本，驱动行业向训练数据生产智能化的方向演进。

（2）全球化发展的大背景下，多语种数据需求不断攀升

随着国家“一带一路”战略的进一步深入推进，我国本土头部企业纷纷走出国门，主动出海；与此同时，国外主流科技企业也在同步加速全球化布局，并呈现不断扩充、细化区域