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

2021 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

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

公司拟以实施权益分派股权登记日登记的总股本为基数分配利润，向全体股东每10股派发现金红利2.50元（含税）。截至2021年12月31日，公司总股本42,800,000股,以此合计拟派发现金红利10,700,000.00元(含税)。本年度现金分红总额占合并报表实现归属于上市公司股东净利润的33.85%；公司本次不进行资本公积转增股本，不送红股。

上述利润分配方案已经公司第二届董事会第七次会议审议通过，尚需提交公司2021年年度股东大会审议。

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	北京市海淀区成府路28号4-801	北京市海淀区成府路28号4-801
电话	010-62660772	010-62660772
电子信箱	ir@speechocean.com	ir@speechocean.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。

算法、算力、数据是人工智能技术发展的三大要素，其中训练数据是算法发展和演进的“燃料”。在当前技术发展进程中，深度学习算法是推动人工智能技术取得突破性发展的关键技术理论，而大量训练数据的训练支撑则是深度学习算法实现的基础。深度学习分为“训练”和“推断”两个环节：训练需要海量数据输入，训练出一个复杂的深度神经网络模型；推断指利用训练好的模型，去“推断”现实场景中的待判断数据，并得出各种结论。训练数据越多、越完整、质量越高，模型推断的结论越可靠。因此，要使算法模型实现从技术理论到应用实践的落地过程，就需要提供大量的训练数据，对算法模型加以训练。通常，从自然数据源简单收集取得的原料数据并不能直接用于深度学习算法的训练，必须经过专业化的采集、加工处理，形成相应的工程化数据集后才能供深度学习算法等算法、模型训练使用。

习近平总书记曾强调：“要构建以数据为关键要素的数字经济。”，《中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见》对加快培育数据要素市场也作出了部署。数据是新的生产要素，是基础性资源和战略性资源。2021 年 3 月，建设人工智能训练数据集、发展全数据产业链已被正式纳入国家十四五规划。训练数据已经成为国家重视、支持和推动的人工智能产业发展所必需的关键产品、关键材料。

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、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 695 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，拍摄自动驾驶场景视频等）、加工（对图像、视频文件进行打点、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。

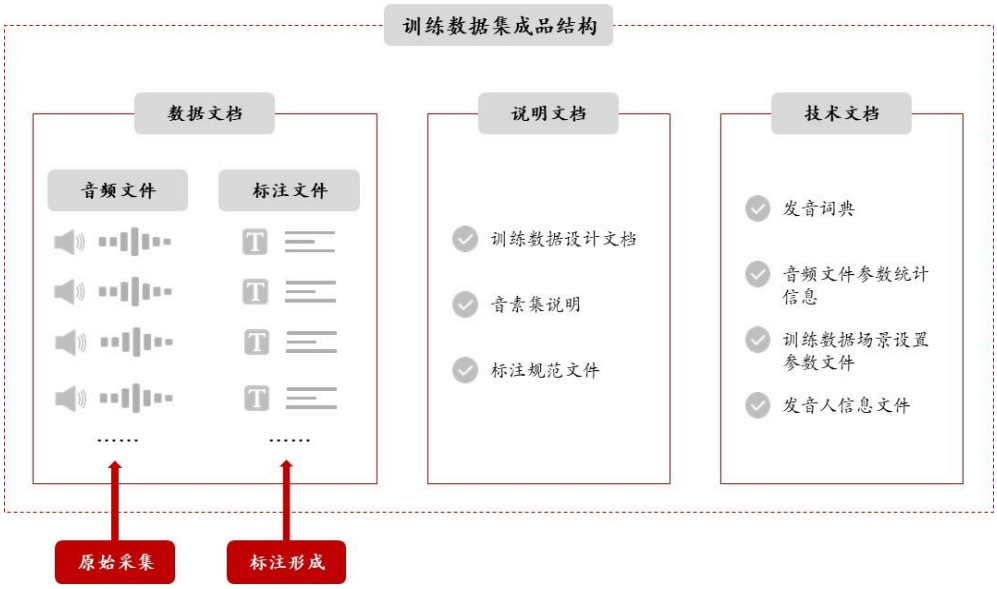
公司通过设计训练数据集结构、采集（收集自然语言文本、对话等数据信息）、加工（对自然语言文本数据进行单词分割、词性标注、语义语法标注、情感属性标注等）、质检（对数据集进行质量检测，如检验文本、词性或者语义的标注结果是否准确等）；或者对客户提供的自然语言文本执行加工、质检工作，最终形成客户所需的自然语言训练数据集。

（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、智能家居、智能客服、机器人、语音导航、智能播报、语音翻译、移动社交、虚拟人、智能驾驶、智慧金融、智慧交通、智慧城市、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，公司仅享有服务费收入，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。

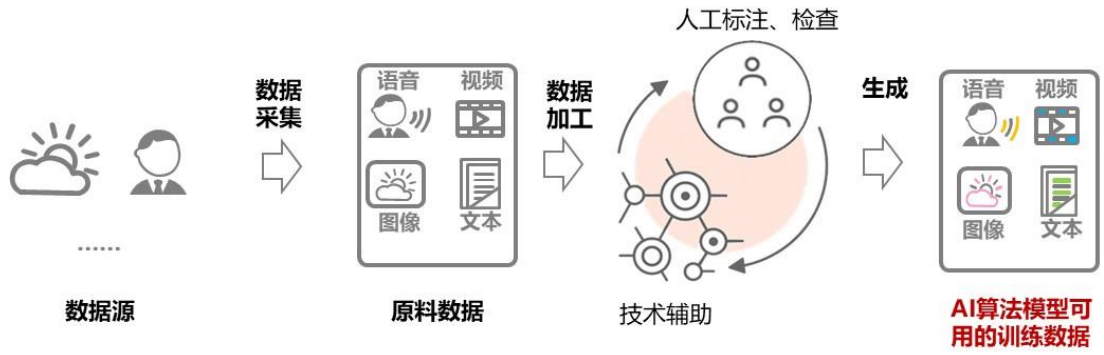
（2）标准化产品：公司开发自有知识产权的训练数据集产品，通过销售训练数据集产品的使用授权许可，获取让渡资产使用权收入。此类训练数据集一经开发完成，可多次销售并获取授权许可收入。

（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，以及极少量硬件销售收入。

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：

1. 数据服务采购：公司在数据采集、加工环节中，向人力资源服务公司等采购的，非核心技术环节的原料数据采集、标注服务。

2. 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。

3. 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，由采购部

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行政部负责。财务部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《供应商管理制度》、《采购管理制度》、《业务采购实施细则》、《岗位服务采购实施细则》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） **销售模式**

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。公司以高品质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、口碑传播、参与学术会议和行业展会、官方网站和自媒体展示等方式建立品牌知名度、与客户建立联系，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

根据国家统计局《战略性新兴产业分类（2018）》，公司所从事的训练数据生产业务属于“新一代信息技术产业—新兴软件和新型信息技术服务—新型信息技术服务—信息处理和存储支持服务—数据加工处理服务”行业，是国家重点支持的“新一代信息技术领域”的战略性新兴产业。公司通过设计训练数据集结构、执行数据采集、加工处理过程，生产用于算法模型开发训练用途的专业数据集，并以软件形式向客户交付，所属行业为软件和信息技术服务业。

根据中国证监会颁布的《上市公司行业分类指引》（2012 年修订），公司所属行业为“软件和信息技术服务业”，行业代码为“I65”。

1.1 行业的发展阶段、基本特点

1） **训练数据作为 AI 算法发展和演进“燃料”的作用继续凸显**

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。在当前人工智能行业发展进程中，有监督的深度学习算法是推动人工智能技术取得突破性发展的关键技术理论，而大量训练数据的支撑则是有监督的深度学习算法实现的基础，训练数据早已成为算法模型发展和演进的“燃料”。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，2012-2016 年期间，人工智能行业不断优化算法增加深度神经网络层级，利用大量的数据集训练提高算法精准性，ImageNet 数据集的超过 1,400 万张训练图片和 1,000 余种分类便在其中起到重要作用。2021

年，全球人工智能和机器学习领域最权威的学者之一吴恩达教授提出二八定律：AI 研究 80%的工作应该放在数据准备上，确保数据质量是最重要的工作；业界如果更多地强调以数据为中心而不是以模型为中心，那么机器学习的发展会更快。

然而，从自然数据源简单收集取得的原料数据并不能直接用于有监督的深度学习算法训练，必须经过专业化的采集、加工，形成相应的工程化训练数据集后才能供深度学习算法等训练使用。目前，应用有监督学习的算法对于训练数据的需求远大于现有的标注效率和投入预算，基础数据服务将持续释放其对于算法模型的基础支撑价值。

2） AI 产业对训练数据服务的需求持续产生、规模继续扩大

AI 产业对训练数据的需求主要来源于成熟算法模型的拓展性需求和新生算法模型的前瞻性需求。在成熟的拓展性需求方面，Mckinsey Global Institute 的研究报告表明：

深度学习模型对训练数据的数据量、多样性和更新速度方面提出较高要求。为充分发挥技术潜能，深度学习模型需要海量且涵盖图像、视频及语音在内等多种类型的训练数据进行模型训练。此外，人工智能技术要求算法模型根据潜在的应用场景变化而持续更新，因此，算法模型所使用的训练数据亦需要定期更新。具体而言，约 1/3 的算法模型每月至少更新一次，约 1/4 的算法模型每日至少更新一次，算法模型持续更新的特点将进一步拓展各领域训练数据的需求空间。

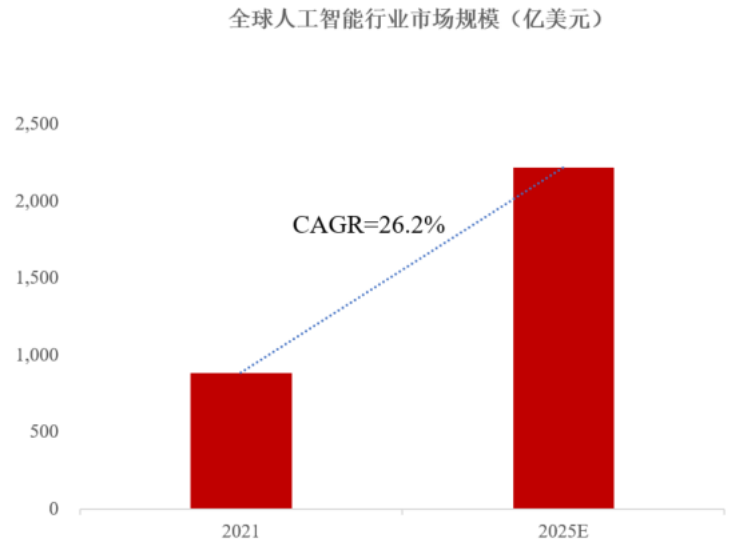
而在新生的前瞻性需求方面，随着人工智能商业化进程的演进，新兴应用场景如智能驾驶、智联网 AIoT、AI PaaS、产业互联网等将展现出巨大的发展潜力，并逐步促进 AI 技术和算法模型的优化和创新。因此，在创新应用场景和新型算法的带动下，具有前瞻性的训练数据产品和高定制化的训练数据服务需求将逐步成为主流。

3） 全球和中国 AI 基础数据服务行业规模持续扩张

i. 全球和中国 AI 产业市场规模

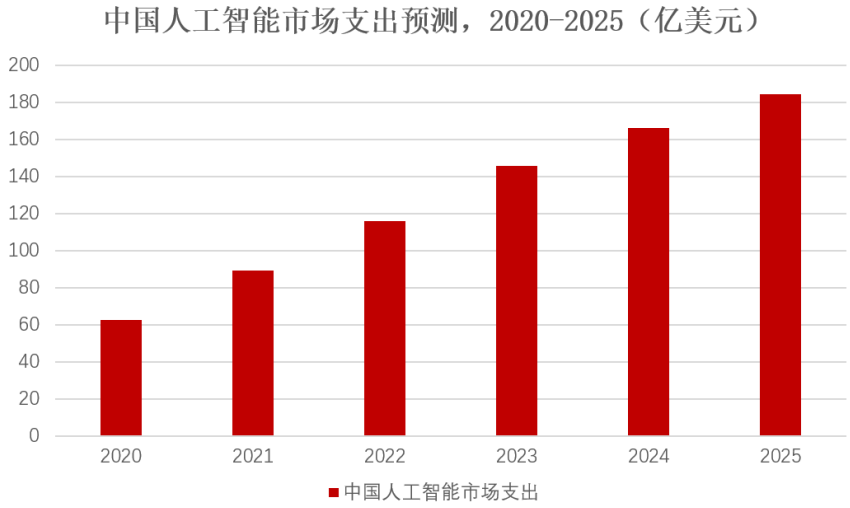
经过多年的发展，人工智能技术已在人机交互、智能家居、智能驾驶、智慧金融、智能安防等多个领域实现技术落地，且应用场景愈来愈丰富，AI 产业已进入全方位商业化的发展阶段。

根据国际数据公司（IDC）的数据，2021 年，全球人工智能市场规模将达到 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）

当前我国人工智能产业加速发展，从基础支撑、核心技术到行业应用的产业链条基本形成，一批创新活跃、特色鲜明的创新企业加速成长，新模式、新业态不断涌现，整体呈现蓬勃发展态势。政策支持、投资引导和巨头布局将推动中国 AI 产业的结构调整，进一步扩大市场规模。根据国际数据公司（IDC）的数据，中国人工智能市场规模预计 2025 年有望达 184.3 亿美元，年复合增长率达到 24.4%。

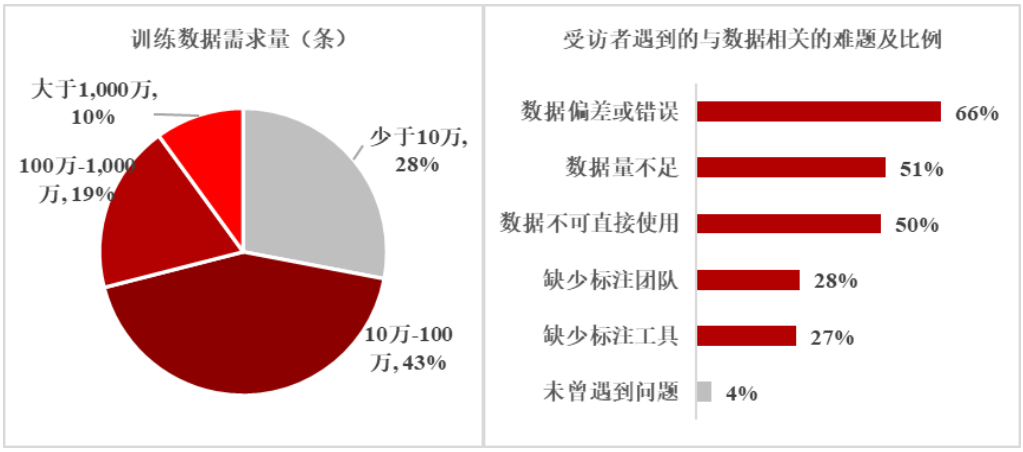


数据来源：国际数据公司（IDC）

ii. 全球和中国 AI 基础数据服务行业发展情况及规模

全球基础数据服务行业处于快速成长期，市场规模具有较大的增长空间。应用场景的创新和机器学习算法的流行直接带动了训练数据需求的大幅增长，这种趋势导致训练数据难以获取和数

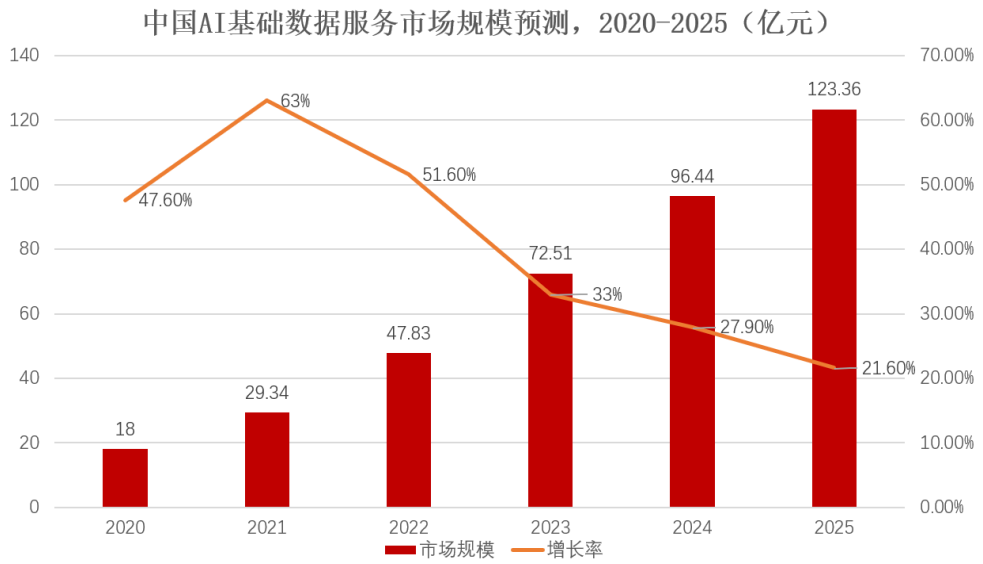
据科学家、数据工程师等人力资源稀缺成为制约 AI 产业发展的两大挑战。根据 Dimensional Research 的全球调研报告，72%的受访者认为至少使用超过 10 万条训练数据进行模型训练，才能保证模型有效性和可靠性，96%的受访者在训练模型的过程中遇到训练数据质量不佳、数量不足、数据标注人员不足等难题。为应对训练数据所带来的多方面挑战，AI 企业开始从第三方购买原料数据收集、训练数据生产和数据专家咨询等服务，调研结果指出，外包服务能够有效加快算法模型落地应用的速度。因此，得益于训练数据需求增长和对外采购意识的形成，全球基础数据服务行业进入快速成长期，市场规模具有较大的增长潜力。



数据来源：Dimensional Research

从 AI 产业链的发展情况和未来发展趋势来看，中国基础数据服务行业的市场规模将不断扩大。一方面，随着算法模型、技术理论和应用场景的优化和创新，AI 产业对训练数据的拓展性需求和前瞻性需求均快速增长；另一方面，随着行业内对训练数据需求类型的增加以及对服务标准要求的提高，产业链的专业化分工将愈加清晰，专业化的训练数据服务提供商将扮演更加重要的角色。

根据国际数据公司（IDC）2022 年 3 月发布的《IDC Worldwide Artificial Intelligence Spending Guide》预测，2025 年中国人工智能市场规模有望达到 184.3 亿美元（约 1,200 亿人民币），其中，关于基础数据部分，根据 IDC 发布的《2021 年中国人工智能基础数据服务市场研究报告》，预计中国 AI 基础数据服务市场规模近 5 年来的复合年增长率达到 47%，预期 2025 年将突破 120 亿元，达到中国人工智能市场支出总额的约 10%。同时，根据《IDC Worldwide Artificial Intelligence Spending Guide》的预测，2025 年全球人工智能市场规模将达到 2,218.7 亿美元，基础数据服务板块也将是其重要的组成部分之一。



数据来源：国际数据公司（IDC）

4） 以智能驾驶为代表的垂直领域对训练数据需求正在兴起，市场规模可观

当前 AI 技术开始广泛应用于不同产业，展现出可观的商业价值和巨大的发展潜力，为数据服务行业提供巨大的发展红利。产业化应用新产品、新应用、新场景层出不穷，产生了大量新兴垂直领域的基础数据需求，这其中尤以智能驾驶为代表的产业级应用呈现快速增长态势，为数据服务的发展提供了长期向好的基本面。

随着智能化、自动化技术不断成熟，汽车产品正在向智能移动终端快速演进。同时，在智能联网技术的推动下，智能汽车将逐渐接力成为乘用车市场中主要的增长动力。在汽车智能网联化的变革中，汽车电子、软件、算法等价值将因智能驾驶技术而显著提升。先进的通讯、计算机、人工智能等技术不断应用在智能驾驶汽车中，成为愈加重要的生产要素。而在智能驾驶功能实现的过程中，数据扮演着至关重要的角色。

以实现智能驾驶所不可逾越的第一环节--环境感知为例，智能驾驶车辆通过各类传感器如摄像头、毫米波雷达、超声波雷达、激光雷达等获取车辆周边信息，产生图片数据、视频数据、点云图像、电磁波等信息，去除噪点信息后利用不同类型数据形成冗余同时提升感知精度和鲁棒性。对于不同级别智能驾驶汽车和驾驶任务而言，需要的传感器类型、数量和性能也有所区别。这就意味着获取高质量、大规模、多种类、强特征的训练数据是实现高精准度环境感知、进而实现高质量智能驾驶的关键。

根据中金公司研究部预测，仅在高级别自动驾驶领域，随着落地场景的广泛化以及商业化进程的提速，市场规模可达万亿元级别。具体应用场景可分为 2C（乘用车）、2B（商用车）和 2G（政府国企）等。根据中金公司研究部测算，预计我国高速城际物流市场达 3.3 万亿元，自动驾驶出

行服务市场近 1.7 万亿元，矿区无人驾驶市场近 6,700 亿元，无人末端配送市场达 1,700 亿元。目前，智能驾驶数据服务需求处于加速起步阶段，市场规模尚无法准确估量，但随着数据之于 AI 应用技术研发的作用的提升，依托于智能驾驶巨大的市场空间，智能驾驶数据服务领域的市场规模同样具有广阔前景。

除智能驾驶领域外，其他垂直行业（例如智慧金融、工业互联网等）和政企领域也将成为训练数据实现规模化应用的重要方向，是尚未估量的新增市场，且每一个垂直行业内部均有诸多细分，因此市场容量非常可观。

5） 国家政策顶层引领、行业重点支持与规范安全监管协同并进

当前，我国已经开始进入由工业经济迈向数字经济发展的“新阶段”，国家高度重视数字经济，而数据要素是数字经济深化发展的核心引擎。习近平总书记在中共中央政治局就实施国家大数据战略进行第二次集体学习时曾指出：数据是新的生产要素，是基础性资源和战略性资源，也是重要生产力，要构建以数据为关键要素的数字经济。2022 年 1 月 12 日，国务院印发《“十四五”数字经济发展规划》明确提出：数据要素是数字经济深化发展的核心引擎，坚持以数字化发展为导向，充分释放要素价值，激活数据要素潜能。随着数据要素作为国家级战略资源地位不断凸显，一系列国家引领与行业鼓励政策不断推进，数据作为当前最具时代特征的生产要素，成为重点支持领域，为数据资源产业带来了巨大的发展机遇。

与此同时，随着数字经济规模的快速扩张，数字技术广泛应用和法律规范启动落地的相互交融也成为数据产业健康发展的必然趋势，建设规范、安全、合规、高质量的数字经济已成为迫切要求，国家陆续出台包括《数据安全法》、《个人信息保护法》、《汽车数据安全管理若干规定（试行）》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。

主要行业政策及法律法规如下：

序号	发布时间	发文机关	主要行业政策及法律法规	相关内容
1	2022 年 1 月	国务院	《“十四五”数字经济发展规划》	明确指出数据要素是数字经济深化发展的核心引擎，坚持以数字化发展为导向，充分释放要素价值，激活数据要素潜能。

2	2021 年 8 月	国家互联网信息办公室等	《汽车数据安全管理若干规定（试行）》	作为汽车数据安全领域出台的第一份有针对性的管理规定，明确了汽车数据处理者的责任和义务，规范汽车数据处理活动，对防范化解汽车数据安全风险、保障汽车数据依法合理有效利用具有重要意义。
3	2021 年 8 月	第十三届全国人大常委会	《个人信息保护法》	进一步细化、完善个人信息保护应遵循的原则和个人信息处理规则，明确个人信息处理活动中的权利义务边界，健全个人信息保护工作体质机制。
4	2021 年 6 月	第十三届全国人大常委会	《中华人民共和国数据安全法》	我国数据的使用和保护进入有法可依的新阶段，国家统筹发展和安全你，坚持以数据开发利用和产业发展促进数据安全，以数据安全保障数据开发利用和产业发展。
5	2021 年 3 月	十三届全国人大四次会议	《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》	《十四五规划》指出要加快数字化发展，建设数字中国，同时打造数字经济新优势，充分发挥海量数据和丰富应用场景优势，促进数字技术与实体经济深度融合，赋能传统产业转型升级，催生新产业新业态新模式，壮大经济发展新引擎。同时指出要加强关键数字技术创新应用：聚焦高端芯片、操作系统、人工智能关键算法、传感器等关键领域；建设重点行业人工智能数据集，发展算法推理训练场景。
6	2020 年 3 月	中共中央、国务院	《中共中央 国务院关于构建更加完善的要素市场化配置体制机制的意见》	首次将“数据”作为市场化要素写入国家顶层设计级别文件，提出要加快培育数据要素市场，发挥数据在市场化配置中的作用。

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、以及算法辅助数据生产能力，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，可用于检查训练数据集

对算法模型的训练效果，进而反哺指导训练数据集的设计；也可应用于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。

2）平台工具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。

2. 公司所处的行业地位分析及其变化情况

1） 深耕行业多年，拥有丰富的技术积累和行业经验，具备较强竞争优势

海天瑞声是我国最早专业从事训练数据产品与服务研发与及销售的主要企业之一，公司凭借多年的研发积累和创新，不仅完成 930 余个自有知识产权的训练数据标准化产品集的建设，在大规模、高质量、可授权使用数据库存量全球企业排名中稳居前列，形成了大量核心技术与知识产权储备成果，并将基础研究、平台工具、训练数据生产等三大领域积累的核心技术持续应用于训练数据生产的各个环节，在数据库架构设计、开发标准、语言学特征、质检评测等多项技术指标方面凸显竞争优势。

多年积累的核心技术成果和综合专业服务能力，使得公司能够更大规模、更有效率、更加精准地生产 AI 训练数据，在提升自身产出效率的同时也有效提高了训练数据对于客户 AI 算法模型的改善、优化效果。公司与 AI 产业链上的各类企业、研究机构持续保持长期的合作伙伴关系，截

止 2021 年底，企业服务客户数量已达到 695 家，产品及服务能力不断得到优质客户的认可，未来公司将继续完善产品服务体系、升级服务质量，不断增强综合数据服务能力竞争优势。

2) 处于中国 AI 基础数据服务行业第一梯队，拥有稳固的行业地位

作为行业的头部阵营企业，海天瑞声在经营情况、市场地位、技术实力、核心竞争力等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：

项目	海天瑞声	Appen	慧听科技	标贝科技
经营情况				
成立年份	2005 年	1996 年	2011 年	2016 年
市场地位概述	我国领先的训练数据产品服务专业提供商，是我国最早从事训练数据产品服务研发销售的企业之一	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	-	-
员工数量	245	超过 1,125	未公开披露	未公开披露
市场占有率	中国 AI 基础数据服务行业第二名，海天瑞声的市场占有率为 12.9%[1]。	未公开披露	未公开披露	未公开披露
客户结构及客户数量				
主要客户/合作伙伴情况	大型科技公司，如阿里巴巴、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	未公开披露	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

项目	海天瑞声	Appen	慧听科技	标贝科技
客户数量	695 家（截至 2021 年 12 月 31 日累计）	未公开披露	数十家	100 余家
技术指标				
技术实力概述	海天瑞声自主开发了一体化数据处理支撑平台，在基础研究、平台工具、训练数据生产三个维度下均积累核心技术，将多项具体核心技术整合为公司特有的核心技术体系。	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。	采用全程质量监控流程，执行完善的标注流程，配合保密管理手段，提供质量上乘的数据服务。	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。
应用领域覆盖	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、音乐	智能语音、计算机视觉、自然语言、音乐
语种/方言覆盖能力	170 余个	超过 235 个	20 余个	10 余个
已取得专利授权	26 项（24 项发明专利、1 项实用新型专利及 1 项外观设计专利）	2 项	1 项	4 项
计算机软件著作权数量	156 项	未公开披露	14 项	30 项
产品/服务及其侧重点				
拥有的成品训练数据集数量	932 个（截至 2021 年 12 月 31 日累计）	291 个	45 个	179 个
其他经营指标				
主要财务指标（2021 年度/2021 年末）				
营业收入	2.06 亿元人民币	4.47 亿美元	未公开披露	未公开披露
综合毛利率	64.01%	39.98%	未公开披露	未公开披露
净利润	3,160.54 万元人民币	2,851.9 万美元	未公开披露	未公开披露
净利率	15.31%	6.38%	未公开披露	未公开披露

数据来源及说明：

1、Appen、慧听科技、标贝科技数据：截至 2021 年 12 月，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（http://cpquery.sipo.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。

2、海天瑞声数据：除特别标注外，均为截至 2021 年 12 月 31 日数据。

与 Appen 相比，海天瑞声规模较小，营收及净利润规模、员工数量等均低于 Appen，在语种/方言覆盖能力方面也具备一定劣势；双方在训练数据的产品领域覆盖能力相当；公司在标准化训练数据集产品储备情况以及我国客户的覆盖程度方面具备一定优势。

与慧听科技、标贝科技相比，根据国际数据公司（IDC）数据，公司市场份额在中国 AI 基础数据服务行业排名第二，较为领先。公司在训练数据覆盖的语种/方言覆盖能力、标准化训练数据集产品储备数量、计算机软件著作权数量方面均高于慧听科技、标贝科技（限于公开信息可查询范围），优势明显；在产品领域覆盖方面，除智能语音、计算机视觉、自然语言外，标贝可覆盖音乐类训练数据；公司未单独列示音乐类训练数据，但语音合成类业务可提供歌曲合成类服务，也已覆盖音乐类训练数据，公司在产品领域覆盖方面比较完善。

在市场地位方面，Appen 是较早从事训练数据开发的训练数据提供商，经营历史较长，规模、体量等相比海天瑞声均具备优势；而公司是我国最早从事训练数据研发、生产、销售的企业之一，在我国市场具备领先地位。多年来，公司深耕训练数据服务领域，伴随了众多国内客户在人工智能领域特别是智能语音领域的开拓、成长，为其持续提供了全球语种语音训练数据的高质量的本土服务，降低了对国外同类训练数据的依赖。

在主要财务指标方面，慧听科技、标贝科技未公开披露其财务数据信息。根据国际数据公司（IDC）数据，公司市场份额在中国 AI 基础数据服务行业排名第二，较为领先；公司在营业收入、净利润规模方面相比 Appen 存在一定劣势，但在盈利能力方面，即综合毛利率、净利率均优于 Appen。

在专利储备方面，通过公开信息渠道可获悉的 Appen、慧听科技、标贝科技的专利储备数量较少，公司在专利技术储备方面具备明显优势。在计算机软件著作权方面，慧听科技及其子公司共拥有计算机软件著作权 14 项，标贝科技及其子公司共拥有计算机软件著作权 30 项；海天瑞声及其子公司共拥有计算机软件著作权 156 项，远高于公开信息可查询的慧听科技、标贝科技的计算机软件著作权数量，具备一定优势。

在语种/方言覆盖能力方面，根据 Appen、慧听科技、标贝科技官方网站的信息，公司的产品和服务可以覆盖超过 170 个语种/方言，覆盖的语种/方言数量少于 Appen，但公司在自有知识产权训练数据产品数量上具备一定优势。与慧听科技、标贝科技相比，公司在产品和服务覆盖的语种/方言个数、自有知识产权训练数据产品数量方面均高于公开信息可查询的慧听科技、标贝科技的相关数量，具有明显的优势。

3） 持续荣获多项资质荣誉，具有基础数据服务行业影响力

近年来，公司的核心技术创新性、先进性不断得到行业、主管部门的高度认可，公司连续多年被评为国家高新技术企业、国家规划布局内重点软件企业，并荣膺工信部国家专精特新“小巨人”企业，2021 年获批“北京市企业技术中心”，并以优异表现获评工信部“新一代人工智能产业创新重点任务揭榜优胜单位”，充分体现了海天瑞声在人工智能基础数据服务领域作为行业标杆和龙头企业的领先作用。公司提供高质量、高效率、高水准的训练数据产品及服务，作为人工智能产业发展所必须的关键要素，补齐攻克了人工智能产业基础发展瓶颈的关键一环，满足了训练数据资源国产化自主创新需求，以专业性和创新性获得了行业与主管机构高度认可，并在推动行业发展上形成了积极的标杆示范效应。

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） 大规模、高精度、高复杂度需求逐渐常态化，训练数据生产智能化成为方向

近年来，随着 AI 应用场景日益拓宽、融合、多元化，客户在数据规模、难度、复杂度方面的要求逐渐抬升，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、更多音色类型、更多维的人像采集、更长尾的物体影像采集等维度，这不仅要求数据服务商具备更丰富的数据采、标经验，并在数据生产过程中不断引入算法、提升平台能力，以持续迭代的智能化人机协作模式来不断提高数据处理质量和效率、降低成本，驱动行业向训练数据生产智能化的方向演进。

2） 语言/语种多元化数据需求不断攀升

随着国家“一带一路”战略的进一步深入推进，我国企业出海布局增多；同时，国外主流的训练数据需求企业大多早已实现全球布局，并呈现不断扩充、细化区域拓展策略的趋势。在此背景下，市场对多语言训练数据的需求迎来新一轮增长，除中、英、法、德、意、西、日、韩等常见语种外，预计客户群体还将在诸如东南亚、一带一路沿线国家地区的罕见小语种（尤其是亚洲小语种）方向产生新的增量需求，只有那些在多语言/语种基础研究方面持续投入、拥有一定积累

的数据服务商方能抓住此契机、满足客户需求。

3） 智能驾驶领域引领数据需求拓展至更多垂直场景，对行业提出更高要求

随着 AI 产业落地成为主旋律，数据采、标服务需满足的 AI 应用场景比以往明显更加广泛，加之数据质量决定 AI 系统预测的准确度，行业用户对数据服务商在特定垂直场景下专业性要求正在提高，垂直场景的定制化训练数据需求将成为主流。

在细分行业及场景所具备的专业知识、服务经验以及准入资质将成为衡量一家数据服务商是否具备垂直领域数据服务能力的重要考量因素。当前，以智能驾驶为代表的垂直领域已开始释放大规模训练数据需求，行业客户更加需要得到全栈式的闭环数据解决方案，以满足智能驾驶业务的数据处理量更大、数据处理需求的迭代频次更高、合规要求更高等特点，这就要求数据服务商在专业能力（包括但不限于对于交通场景、车辆传感器等要素的综合理解和实施能力）、综合能力（包括但不限于数据处理平台能力、质量管控能力、需求对接能力、项目响应能力、供应链资源管理能力等）、准入资质等方面同时满足达到较高水准。

4） 数据安全与隐私保护要求快速提升，考验数据服务商合规服务能力

近年来，《数据安全法》、《个人信息保护法》、《汽车数据安全管理若干规定（试行）》等主流法律法规已经快速落地实施，行业可以清晰地感受到国家在这方面的法律环境在快速趋严，数据安全相关法律体系的完善对训练数据产业的健康发展将产生深远的影响，有利于规范行业行为、治理行业乱象，提高行业门槛，为实现行业可持续良性发展创造健康环境。

在此背景下，数据安全、隐私保护将成为行业用户选择数据采标服务时的重要考量因素，甚至一些大型需求方在遴选数据服务商时已将此因素提升至重要级别。因此，数据服务商在此方面须紧跟国家法律法规要求的演变，相应调整、升级现行业务开展方式、数据安全管理体系，及时获取合规资质（包括但不限于信息安全管理体系认证、隐私信息管理体系认证、网络安全等级保护测评等），切实提升自身数据安全与合规能力，确保业务始终在健康、合规的环境下开展，并将自身在这方面的积累转化为竞争优势。

3 公司主要会计数据和财务指标

3.1 近 3 年的主要会计数据和财务指标

单位：元 币种：人民币

	2021年	2020年	本年比上年增减(%)	2019年
总资产	840,663,396.09	477,350,038.99	76.11	404,539,351.88
归属于上市公司股	805,908,403.05	437,956,372.58	84.02	355,951,438.36

东的净资产				
营业收入	206,476,533.04	233,373,953.01	-11.53	237,558,118.15
归属于上市公司股东的净利润	31,605,431.79	82,081,021.91	-61.49	81,586,824.49
归属于上市公司股东的扣除非经常性损益的净利润	21,067,433.20	73,015,355.36	-71.15	76,246,636.59
经营活动产生的现金流量净额	-15,548,319.63	51,176,659.14	-130.38	83,363,303.85
加权平均净资产收益率（%）	5.59	20.68	减少15.09个百分点	39.78
基本每股收益（元／股）	0.89	2.56	-65.23	2.72
稀释每股收益（元／股）	0.89	2.56	-65.23	2.72
研发投入占营业收入的比例（%）	29.31	18.64	增加10.67个百分点	17.55

3.2 报告期分季度的主要会计数据

单位：元　币种：人民币

	第一季度（1-3 月份）	第二季度（4-6 月份）	第三季度（7-9 月份）	第四季度（10-12 月份）
营业收入	44,238,877.46	61,762,722.07	24,774,663.82	75,700,269.69
归属于上市公司股东的净利润	16,341,743.76	21,473,377.86	-12,156,689.84	5,947,000.01
归属于上市公司股东的扣除非经常性损益后的净利润	14,324,978.64	19,365,100.35	-14,490,933.13	1,868,287.34
经营活动产生的现金流量净额	18,456,157.25	-8,507,682.68	-23,550,934.24	-1,945,859.96

季度数据与已披露定期报告数据差异说明

□适用 √不适用

4 股东情况

4.1 普通股股东总数、表决权恢复的优先股股东总数和持有特别表决权股份的股东总数及前 10 名股东情况

单位：股

截至报告期末普通股股东总数(户)	4,825
年度报告披露日前上一月末的普通股股东总数(户)	4,875
截至报告期末表决权恢复的优先股股东总数（户）	0
年度报告披露日前上一月末表决权恢复的优先股股东总数（户）	0
截至报告期末持有特别表决权股份的股东总数（户）	0
年度报告披露日前上一月末持有特别表决权股份的股东总数（户）	0

前十名股东持股情况

股东名称（全称）	报告期内增减	期末持股数量	比例(%)	持有有限售条件股份数量	包含转融通借出股份的限售股份数量	质押、标记或冻结情况		股东性质
						股份状态	数量	
贺琳	0	8,669,725	20.26	8,669,725	8,669,725	无		境内自然人
北京中瑞安投资中心（有限合伙）	0	4,954,128	11.58	4,954,128	4,954,128	无		其他
中移投资控股有限责任公司	0	3,855,000	9.01	3,855,000	3,855,000	无		国有法人
唐涤飞	0	3,577,982	8.36	3,577,982	3,577,982	无		境内自然人
北京清德投资中心（有限合伙）	0	2,545,463	5.95	2,545,463	2,545,463	无		其他
上海丰琬投资合伙企业（有限合伙）	0	1,880,374	4.39	1,880,374	1,880,374	无		其他
北京中瑞立投资中心（有限合伙）	0	1,871,560	4.37	1,871,560	1,871,560	无		其他

上海兴富创业投资管理中心（有限合伙）	0	1,323,112	3.09	1,323,112	1,323,112	无		其他
中国互联网投资基金管理有限公司－中国互联网投资基金（有限合伙）	0	1,290,000	3.01	1,290,000	1,290,000	无		其他
华泰证券资管－招商银行－华泰海天瑞声家园 1 号科创板员工持股集合资产管理计划	0	1,070,000	2.50	1,070,000	1,070,000	无		其他
上述股东关联关系或一致行动的说明	上述股东中，1、公司控股股东、实际控制人贺琳持有 100%股权的北京创世联合投资管理有限公司为北京中瑞安投资中心（有限合伙）的普通合伙人、执行事务合伙人，并持有北京中瑞安投资中心（有限合伙）36.67%的出资，贺琳为北京中瑞立投资中心（有限合伙）的有限合伙人，持有北京中瑞立投资中心（有限合伙）5.88%的出资；2、唐涤飞持有 50%股权的北京创慧科瑞投资管理有限公司为北京中瑞立投资中心（有限合伙）的普通合伙人、执行事务合伙人，并持有北京中瑞立投资中心（有限合伙）29.41%的出资，唐涤飞为北京中瑞立投资中心（有限合伙）的委派代表；3、北京清德投资中心（有限合伙）普通合伙人、执行事务合伙人钟山及其配偶志鹏分别持有北京清德投资中心（有限合伙）8.76%、4.31%的出资，志鹏同时持有北京中瑞立投资中心（有限合伙）普通合伙人、执行事务合伙人北京创慧科瑞投资管理有限公司 50%的股权；4、中移投资控股有限责任公司的间接控股股东中国移动通信集团有限公司为中国互联网投资基金（有限合伙）的有限合伙人，持有中国互联网投资基金（有限合伙）9.97%的出资，中国移动通信集团有限公司的全资子公司中移资本控股有限责任公司持有中国互联网投资基金（有限合伙）普通合伙人、执行事务合伙人中国互联网投资基金管理公司（持有中国互联网投资基金（有限合伙）0.33%的出资）16.36%的股权；5、华泰证券资管－招商银行－华泰海天瑞声家园 1 号科创板员工持股集合资产管理计划为公司部分高级管理人员及核心员工参与战略配售所成立的专项资管计划。除此之外，公司未知上述其他股东之间是否存在关联关系或属于一致行动人。							

表决权恢复的优先股股东及持股数量的说明	无

存托凭证持有人情况

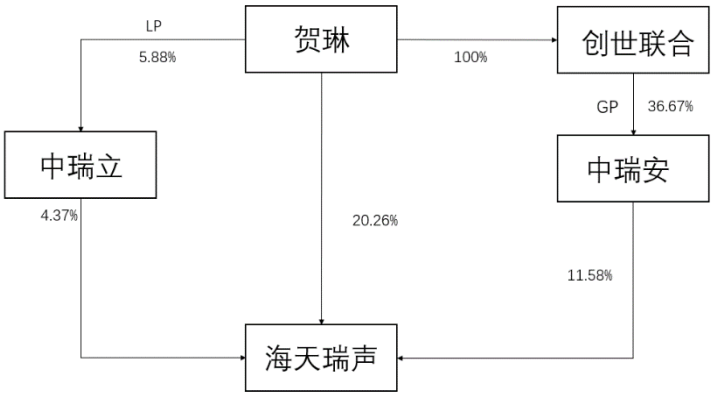
□适用 √不适用

截至报告期末表决权数量前十名股东情况表

□适用 √不适用

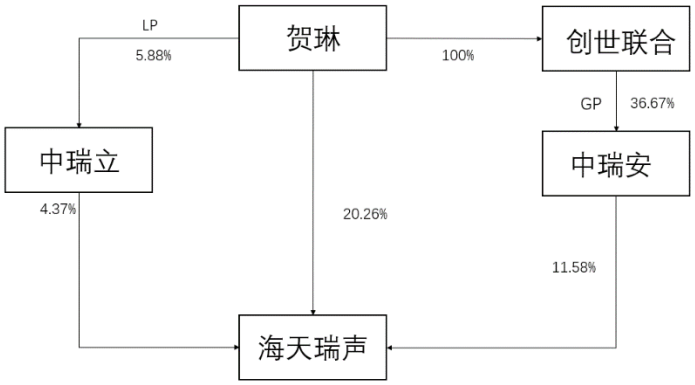
4.2 公司与控股股东之间的产权及控制关系的方框图

√适用 □不适用



4.3 公司与实际控制人之间的产权及控制关系的方框图

√适用 □不适用



4.4 报告期末公司优先股股东总数及前 10 名股东情况

□适用 √不适用

5 公司债券情况

□适用 √不适用

第三节 重要事项

1 公司应当根据重要性原则，披露报告期内公司经营情况的重大变化，以及报告期内发生的对公司经营情况有重大影响和预计未来会有重大影响的事项。

报告期内，公司实现营业收入 20,647.65 万元，实现归属于母公司所有者的净利润 3,160.54 万元，实现归属于母公司所有者的扣除非经常性损益的净利润 2,106.74 万元，分别较上年同期减少 11.53%、61.49%、71.15%。截至报告期末，公司总资产为 84,066.34 万元，归属于母公司的所有者权益为 80,590.84 万元，分别较上年末增加 76.11%和 84.02%。

2 公司年度报告披露后存在退市风险警示或终止上市情形的，应当披露导致退市风险警示或终止上市情形的原因。

□适用 √不适用