

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
2024 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、 容诚会计师事务所（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

经公司第三届董事会第四次会议审议，公司拟定 2024 年度利润分配预案如下：

拟以实施权益分派股权登记日登记的总股本扣除回购专户股份后的股本为基数分配利润，向全体股东每 10 股派发现金股利 3.50 元（含税），截至 2025 年 3 月 31 日，公司总股本 60,325,180 股，回购专户股份数 466,117 股，以此计算合计拟分派现金红利 20,950,672.05 元（含税）；根据《上市公司股份回购规则》等有关规定，上市公司以现金为对价，采用要约方式、集中竞价方式回购股份的，视同上市公司现金分红，纳入现金分红的相关比例计算。公司在 2024 年度以集中竞价交易方式累计回购 906,014 股，支付的资金总额为人民币 49,987,510.15 元（不含印花税、交易佣金等交易费用），视同现金分红。综上，公司 2024 年度现金分红总额为 70,938,182.20 元（含税），现金分红比例为 625.77%；2024 年度不进行资本公积转增股本，不送红股。

2024 年度利润分配预案披露后至实施权益分派的股权登记日期间，若公司总股本发生变动，公司拟维持每股现金分红金额不变，相应调整现金分红总额。

上述利润分配方案尚需提交公司 2024 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

1.1 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

1.2 公司存托凭证简况

☐适用 ☒不适用

1.3 联系人和联系方式

	董事会秘书	证券事务代表
姓名	张哲	张哲
联系地址	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
传真	010-62660892	010-62660892
电子信箱	ir@haitianruisheng.com	ir@haitianruisheng.com

2、 报告期公司主要业务简介

2.1 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 1,050 家，覆盖了科技互联网、社交、IoT、智能驾驶、大模型等领域的主流企业，以及政企、教育科研机构。

产品服务矩阵



服务AI产业链



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）是使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。

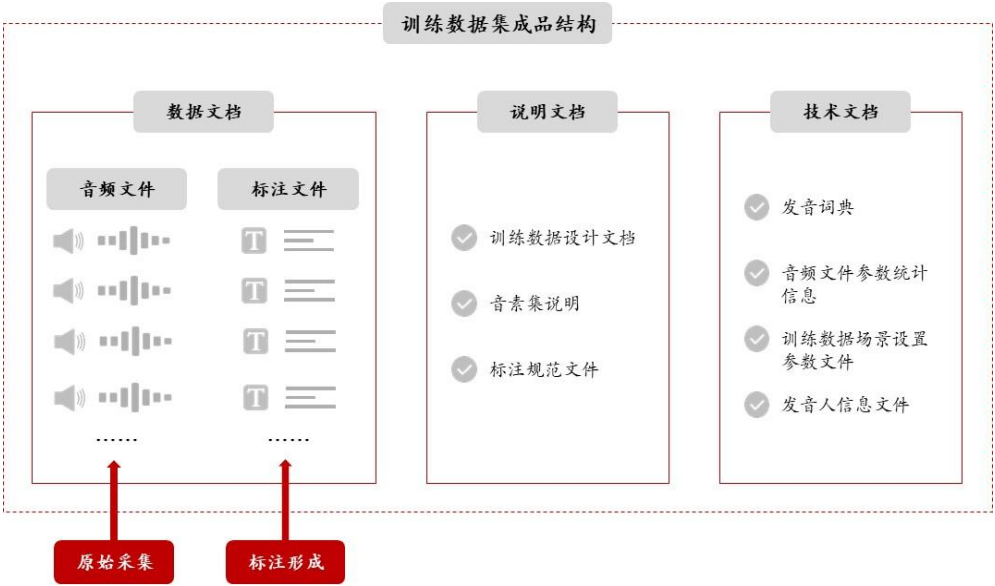
（4）训练数据相关的应用服务

公司开放基于多年行业经验打磨的数据处理工具集及平台，提供包括公有云访问、私有化部署及 SaaS 化服务的多种能力，满足产业链上各类企业对于数据处理工具及平台能力的需求。

公司基于自身生产的训练数据提供算法相关的模型训练服务、模型评测服务及模型应用服务，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

2.2 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：公司基于积累的训练数据和多年行业经验提供数据处理工具集及平台服务、算法相关的模型训练、模型评测及模型应用服务，通常以软件授权或软硬件一体化形式交付平台产品、算法模型拓展、开发成果等，获取让渡资产使用权收入和技术服务等收入。

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

公司开放基于多年行业经验打磨的数据处理工具集及平台，提供包括公有云访问、私有化部署及 SaaS 化服务的多种能力，满足产业链上各类企业对于数据处理工具及平台能力的需求。

公司基于其生产的训练数据提供算法模型相关的训练、评测及应用服务，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应用场景（例如特定行业、特定口音等）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。

2.3 所处行业情况

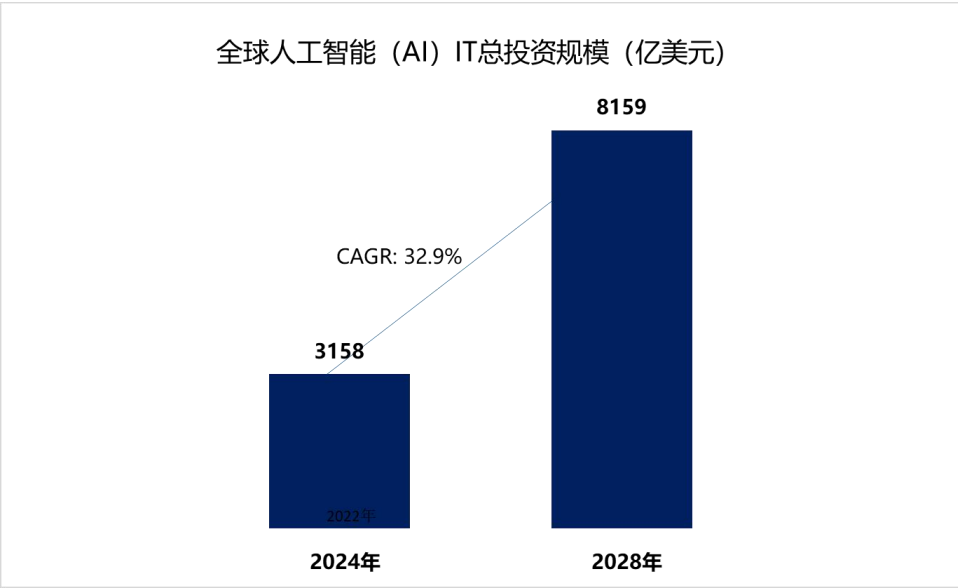
(1). 行业的发展阶段、基本特点、主要技术门槛

1.1 行业的发展阶段、基本特点

(1) 政策、技术、应用协同共振，全球 AI 产业迈入高速发展新阶段

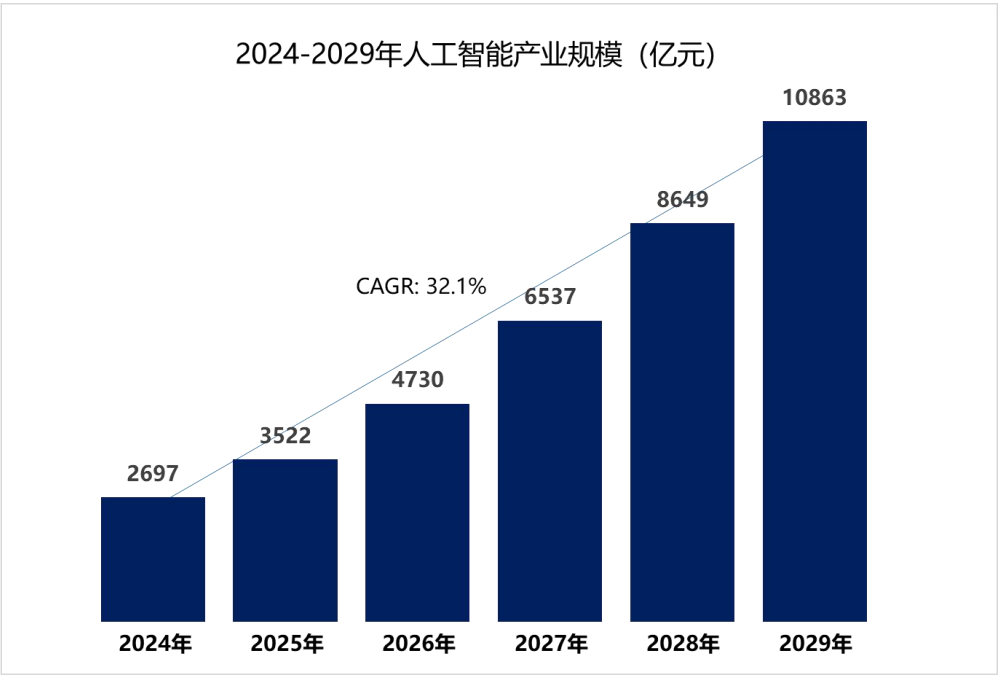
当前，全球人工智能产业正迎来历史性发展拐点，政策、技术与应用的三重共振推动行业进入高速增长通道。中、美、欧三大经济体持续加码 AI 布局：中国通过《新一代人工智能发展规划》和《“十四五”数字经济发展规划》等构建了涵盖技术研发、资源建设和标准制定的全链条政策体系；美国则依托《人工智能行政令》和星际之门计划（Stargate）等，持续巩固其在人工智能基础设施领域的领先优势；欧盟以《人工智能大陆行动计划》为纲领，配套《数字欧洲计划（2025-2027 年工作方案）》专项资金，重点培育 AI 与高性能计算融合创新。技术层面，人工智能也正在经历革命性突破，DeepSeek R1 等开源模型通过性能提升与成本优化的双重突破，显著降低了技术应用门槛；多模态大模型（如 Sora、GPT-4o）也在不断拓展 AI 能力边界，实现全模态互动能力。这种技术平权运动正在加速 AI 向千行百业渗透，大模型已逐步开始向金融、医疗、制造等核心领域进行新一轮的赋能，智能风控、AI 辅助诊断、智能制造等应用场景不断丰富，推动各行业效率提升和模式创新。全球 AI 产业也将在上述政策引导、技术迭代与商业落地的正向循环下，迎来前所未有的发展机遇，加速向规模化、价值化阶段迈进。

根据国际数据公司（IDC）的数据，2024 年全球人工智能(AI)IT 总投资规模为 3,158 亿美元，预计在 2028 年增至 8,159 亿美元，年复合增长率(CAGR)为 32.9%。



数据来源：国际数据公司（IDC）

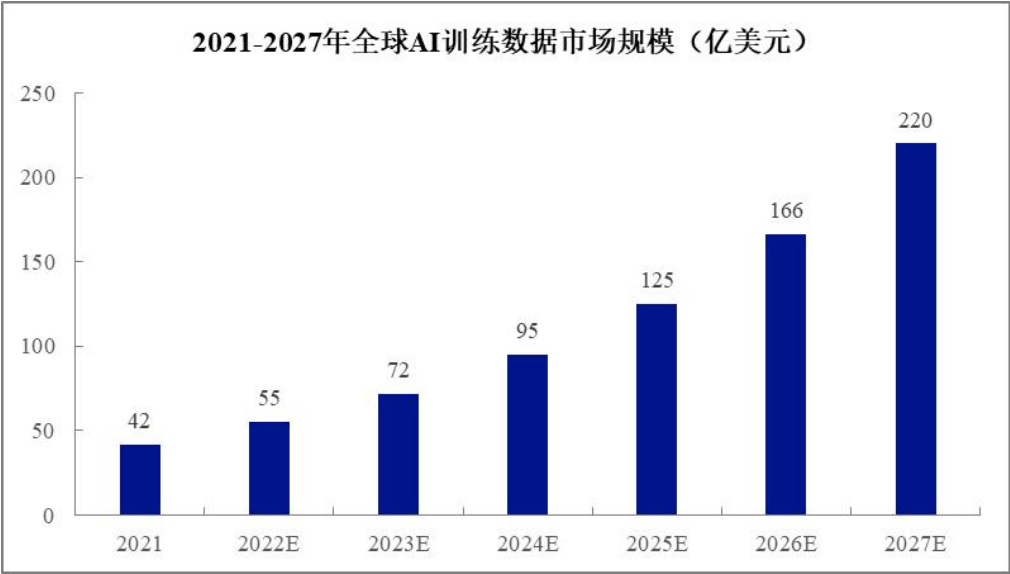
中国作为全球科技大国，也深度受益 AI 技术发展。根据艾瑞咨询的数据，2024 至 2029 年中国 AI 产业将保持 32.1% 的年均复合增长率，在 2029 年突破 1 万亿的市场规模。



(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

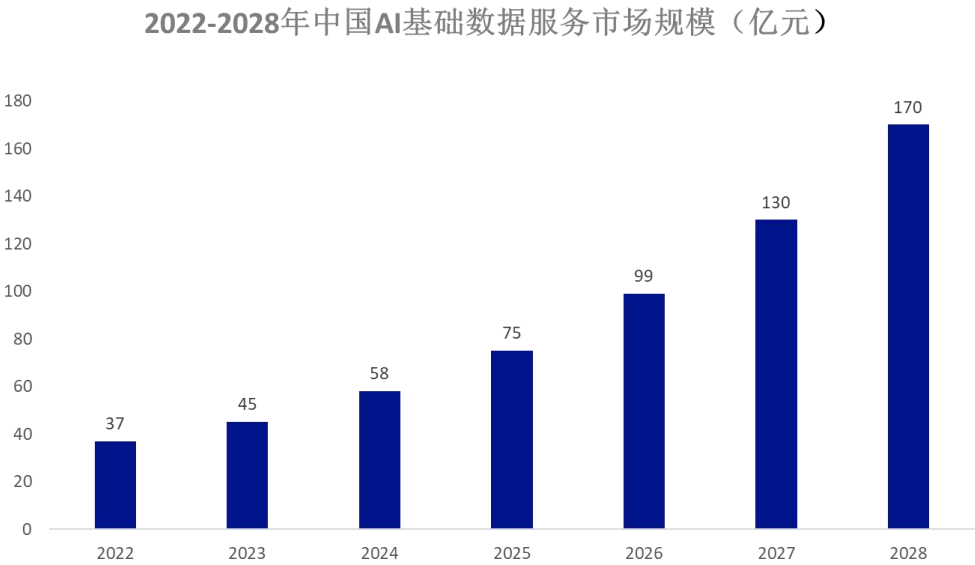
在 AI 产业链中，算法、算力和数据共同构成技术发展的三大核心要素。过去十多年，人工智能产业以算法为核心，通过深度学习算法的不断创新，推动人工智能产业的快速发展。但未来，当更多先进算法趋于开源，算法的差异化壁垒逐渐消除；此外，以 DeepSeek 为代表的算法架构创新和工程优化，带来了训练以及推理成本的大幅下降，这也意味着算力将不再是制约 AI 发展的核心因素。在此背景下，训练数据的重要性被进一步放大，成为未来 AI 发展的关键驱动因素。数据的质量、规模和精准性将直接决定模型能力的上限，高质量的训练数据不仅能显著提升模型的推断可靠性，还能有效减少幻觉和错误信息的产生，成为模型落地效果的关键。因此，高质量数据需求将呈现快速增长趋势。

根据 Cognilytica 数据统计显示，预计 2027 年全球 AI 训练数据市场规模将增长到 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。根据艾瑞咨询的数据，2024 年中国人工智能基础数据服务市场规模为 58 亿元，2028 年规模将达到 170 亿元，年复合增长率为 30.84%。



数据来源：艾瑞咨询

（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

党的二十届三中全会对加快构建促进数字经济发展体制和机制，完善数据要素市场制度规则等作出部署，为中国数据要素市场化改革发展指明了方向。2023 年国家数据局成立以来，我国坚持推进数据要素市场化配置改革“一条主线”，统筹数字中国、数字经济、数字社会“三个建设”的工作思路，推动数据工作更好服务高质量发展。同时，探索完善数据基础制度体系，加快推进

数据基础设施建设，提升数据资源开发和利用水平。国家数据局联合多部门制定《“数据要素×”三年行动计划（2024—2026 年）》《关于促进数据产业高质量发展的指导意见》《国家数据基础设施建设指引》《关于促进数据标注产业高质量发展的实施意见》等一系列数据产业发展指导意见，体现国家高度重视数据产业高质量发展，《关于促进数据产业高质量发展的指导意见》指明了数据技术和产业重点发展方向；《关于促进数据标注产业高质量发展的实施意见》明确提出培育壮大数据标注产业对于提升数据供给质量，推动人工智能创新发展具有重要支撑作用。

全国各地数据要素市场化配置改革也在加速推进。例如，北京市制定 2024 数据工作要点，在探索和实践数据基础制度先行示范区、人工智能数据训练基地的基础上，提出“一区三中心”的建设总体思路，开展“高端数据标注基地”建设，不断完善数据基础制度，激发数据要素活力。2024 年国家公布了 7 个数据标注基地试点城市，数据标注规模再创新高，数据标注总规模达 17282TB，相当于中国国家图书馆数字资源总量的 6 倍左右。其中，成都市发展改革委印发《2024 年成都市数字经济发展工作要点》，成都市大数据集团探索精准安全的公共数据开发运营新模式，搭建“成都市公共数据运营服务平台”，保障不同市场主体平等获取数据要素，实现数据要素配置效益最大化、效率最优化。2024 年长沙市政府印发《长沙市关于推进国家数据标注基地建设的工作方案》，同时配套出台《长沙市关于推进国家数据标注基地建设若干政策（试行）》，从聚焦加快培育标注产业、支持技术创新攻关、加强专业人才培养、促进数据服务赋能等四大方面推出了 14 条产业政策，加大扶持力度。沈阳市推动数据要素市场化配置改革，探索数据交易服务机制，实施《沈阳市公共数据授权运营管理办法（试行）》，开展公共数据授权运营探索实践。其他地方政府也非常关注高质量数据集和数据标注产业的发展。例如，呼和浩特市政府印发《关于构建数据基础制度更好发挥数据要素作用的实施方案》，加快构建数据制度，激活数据要素潜能，更好发挥数据要素作用，增强经济发展新动能。打造内蒙古数据要素产业园，创建数据标注基地。

总体来看，随着“人工智能+”和“数据要素×”等国家战略行动计划的实施，全国各地均在加速推进数据要素市场化配置改革，未来将以高质量数据集建设，夯实人工智能数据基石，推动数字经济高质量发展为目标，围绕数据的汇聚、治理、标注、运营、评测、开发利用以及流通交易全流程，将产生包括数据汇聚、数据治理、数据标注、数据运营、数据评测、数据应用、数据流通、可信数据空间、数据标注基地运营等在内的新业态、新模式，将成为我国数字经济发展的新增长极。

（4）训练数据领域的未来发展趋势

随着 DeepSeek、GPT-4o 等成为全球范围内的现象级应用，以及以手机、具身智能为代表

的 AI 终端的快速推出，人工智能迎来了新的发展机遇，其背后的大模型技术也正在带来相关数据需求的变化和增长。

a.多模态大模型正成为人工智能发展的新趋势，多模态数据需求显著上升

当前，大模型正经历从单模态到多模态的范式转变。多模态技术的本质在于跨模态信息融合，即通过协同处理文本、图像、音频、视频等不同形式的的数据，使 AI 具备更接近人类的全维度认知能力。这种演进不仅显著提升了模型在跨模态任务中的鲁棒性，更解锁了诸如视觉问答、跨模态生成、智能语音交互等复杂场景的应用潜力。

以视觉问答为例，系统需同时解析图像中的视觉元素和文本问题的语义信息，并通过模态对齐与知识推理生成准确回答。这一过程的实现，依赖于海量高质量的图文对数据。数据服务商需构建覆盖多样化场景的问答对，通过模拟真实世界的视觉推理逻辑，训练 AI 建立视觉-语言联合表征能力。而在语音合成领域，要实现富有表现力的拟人化发声，仅有时序对齐的文本-语音数据远远不够，还需引入情感标签、语调标记、韵律特征等细粒度标注，让 AI 精确捕捉人类语言中的情感起伏和表达风格。

实践证明，数据质量与多样性是多模态 AI 发展的决定性因素：一方面，优质数据定义了模型的能力上限；另一方面，数据分布的广度直接关系到技术落地的普适性。未来，随着多模态数据生态的持续完善，AI 的感知与认知能力有望实现新的跨越。

b.大模型发展正在从"规模驱动"向"推理驱动"转型，高质量思维链（CoT）数据成为提升大模型复杂推理能力的关键突破口

随着 DeepSeek R1 的推出，思维链技术（Chain of Thought, CoT）迅速成为 AI 领域的研究热点。这项技术通过引导模型模拟人类逐步推理的过程，显著提升了其在逻辑推理、数学计算及复杂问答等任务中的表现。

过往几年，大语言模型（LLM）的发展主要遵循 Scaling Law（规模定律），即通过扩大模型参数量和训练数据规模来提升性能。然而，随着知识积累的边际效益逐渐降低，仅单纯依赖规模的增长已难以实现质的飞跃。于是，让 AI 像人类一样进行深度、有序的“慢思考”，成为突破现有认知能力瓶颈的关键方向。以数学问题为例，传统 LLM 往往直接输出最终答案，而省略中间推理过程，导致错误率居高不下。相比之下，具备 CoT 能力的模型会分步骤解析问题，先理解题意，再梳理已知条件，逐步推导计算，并最终验证结果的合理性。这种显式推理机制不仅大幅提高了模型的准确性，还使其决策过程更加透明，便于开发者理解和优化。

然而，思维链能力高度依赖高质量的 CoT 数据——即包含详细推理步骤的问答对。这类数据

由于专业性强、标注难度高，目前仍属于稀缺资源。具体而言，优质的 CoT 数据需要领域专家进行人工构建，不仅要确保问题本身的复杂性，还需对解题过程中的每个逻辑进行严格验证，这种双重门槛使得数据生产周期往往长达普通标注任务的数倍。当前即便是最前沿的合成数据技术，在生成具有严密因果关系的多步推理时，仍会有相当概率出现逻辑断层，这也使得高阶人工标注目前仍是不可替代的重要数据来源。

c.从通用到垂直，高质量行业数据需求显著提升

以 DeepSeek 为代表的开源大模型，以其高性能、低成本，以及允许全球用户无限制商用等特点，进一步加速了 AI 应用的普及，激发了应用层的创新活力，使 AI 逐渐从通用大模型向更广泛的应用场景和行业领域拓展，AI 不再局限于作为回答通用问题的智能助手，而是能够针对特定行业/场景提供解决方案，或者协助执行各类复杂任务的超级助理（AI Agent）。例如，在医疗领域，AI 可以作为影像科医生辅助诊断；在法律领域，AI 可以担任法律顾问提供专业咨询；同时，AI Agent 也能够独立完成点外卖、制定行程、预订机票等日常任务。

AI 不仅在传统行业与生活场景中展现出独特价值，在智能终端领域，同样掀起了革新的浪潮。例如，AI 在智能手机上的应用，可以帮助用户通过文字快速创作专业级摄影大片；当海量照片让整理相册成为难题时，可以快速通过文字描述定位对应图像，实现即时快速检索。在交互层面，AI 赋能的语音助手也有质的飞跃。如今的语音助手不再是简单的指令执行者，它们能理解各种不同语种、不同场景下，复杂的语义和情感，并用不同的音色情感与用户进行交互。

与以往通用数据处理需求不同，垂向领域的数据处理更加注重专业性。例如，传统的自然语言处理只包括基本的文本处理工作，例如，分词、词性、正则化等标注需求，这些任务无需标注人员具有特定的行业知识；相对地，开发专门领域，例如医疗大模型，则要求标注人员不仅具备扎实的医疗专业知识，还要能够熟练使用医疗术语进行专业场景下的表达，使医疗模型能够理解和生成专业的医疗内容。这种专业化的需求也对数据服务商提出了更高的要求，数据服务商不仅要对行业有深刻的理解，以设计出符合行业需求的数据解决方案，还需要具备更丰富的专家库储备，以提供高质量的数据标注服务以及相关专业咨询服务。

d.具身智能浪潮来袭，数据供给瓶颈亟待突破

具身智能作为通向通用人工智能（AGI）的关键路径与终极载体，正引领着 AI 发展的下一波浪潮。与传统大模型不同，具身智能的核心在于构建能够与物理世界深度交互的智能体。这不仅需要突破性的多模态感知技术（包括视觉、听觉等），更重要的是要具备基于环境感知的自主决策与执行能力。实现这一目标的关键在于海量的物理世界交互数据，目前行业主要依赖四种数据来

源，机器人遥操（通过人工远程操控获取真实场景下的操作数据）、动作捕捉（记录人类在特定环境中的行为模式）、虚拟合成（利用仿真环境生成训练数据）和互联网开源数据集（整合现有的公开机器人数据）等。其中，虚拟合成技术虽解决了初期数据短缺问题，但由于“仿真差距”的存在，其数据质量往往难以满足高精度训练需求。而开源数据集则面临着数据规模有限（通常仅百万级）、采集条件单一等挑战，无法支撑复杂场景下（需要千万级数据）的模型训练需求。当前，在具身智能蓬勃发展的大背景下，高质量的数据供给短缺问题愈发凸显，具身智能的训练数据市场呈现出巨大的供需缺口，需求旺盛且潜力巨大。

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》、《网络数据安全管理条例》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	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	-

公司	海天瑞声	Appen	数据堂	标贝科技
	人工智能训练数据服务企业			
员工数量	238 人	1,130 人	288 人 (截至 2024 年 6 月 30 日)	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、智谱华章、月之暗面、Minimax 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、英伟达、Oracle 等大型科技公司、汽车厂商及政府	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	阿里、腾讯、微软、百度、京东、华为、小米、滴滴、字节跳动、中国移动、中国联通等
客户数量	超过 1,050 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 200 多个国家与 100 多万名众包人员，训练数据涵盖科技、汽车、金融服务、零售、医疗健康、教育、法律和政府等各个领域。	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。	专注以智能语音为代表的 AI 数据服务；基于自研的一站式 AI 数据平台，提供高质量、多语言、跨领域、跨模态的数据采标服务，涵盖语音、视觉、点云、大模型等核心领域，为客户提供垂直领域 AI 数据解决方案。
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	1,716 个	超过 700 个	531 个	188 个
语种/方言覆盖能力	超过 205 个	超过 290 个	100 余个	40 余个
已取得	41 项（39 项发明专利、1 项实用新型专	未公开披露	45	25 项

公司	海天瑞声	Appen	数据堂	标贝科技
专利授权	利及 1 项外观设计专利)			
计算机软件著作权数量	180 项	未公开披露	237 项	64 项
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	ISO27001 信息安全管理体系认证、 ISO27701 隐私信息管理体系认证	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、 CMMI 成熟度 3 级认证证书、武器装备质量管理体系认证证书；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统认证、 ISO27018 公有云中保护个人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、 入选国家数据局国家人工智能数据产业谱图、 2024 年北京企业和信息服务业核心竞争力企业、 福布斯中国-2024 人工智能科技企业 Top50、 《财富》中国科技 50 强、 2024 胡润中国人工智能企业 50 强	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、 北京市科学技术奖项科技进步奖二等奖	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 2024 年 1-12 月/截至 2024 年 12 月 31 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 2024 年 1-12 月/截至 2024 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) DeepSeek 带火 CoT 技术，多领域 CoT 数据需求集中涌现

伴随 DeepSeek R1 的火爆出圈，其背后的思维链（CoT）技术正在成为 AI 领域的新焦点。该技术通过模拟人类“慢思考”认知模式，将复杂问题拆解为逻辑严密的推理链条，使 AI 系统在数学推导、专业决策等场景中准确率大幅提升。

而构建优质的 CoT 能力体系，需要以结构化思维数据工程为核心基础。通过在训练数据中引入包含中间推理步骤的示例，引导模型学习如何逐步进行推理，并在生成回答时模仿这种推理过程。例如，在数学领域，融入分步证明过程的 CoT 数据，可以使模型掌握从问题解析、公式推导到结论验证的完整认知闭环，使数学问题解决能力大幅提升；在医疗影像诊断领域，借助 CoT 数据训练的 AI 模型可基于影像特征给出逐步推理过程，提高诊断准确性和可解释性。因此，在大模型向应用/垂直领域拓展时，高质量的多领域 CoT 数据需求将快速增加，并成为推动 AI 技术发展的关键因素。

(2) DeepSeek 加速 AI 应用拓展，垂向领域数据需求快速增加，标注难度不断提升

以 DeepSeek 为代表的开源大模型，凭借高性能、低成本和无限制商用等特点，加速了 AI 应用的普及。该技术民主化浪潮推动行业从“暴力堆参数”的通用模型竞赛，转向面向医疗、金融、制造等领域的深度价值挖掘，催生出行业数据处理需求的指数级增长。与通用类数据处理不同，行业数据处理难度更大、更加注重专业性，对数据服务商的综合能力也提出了更高的要求。一方面，数据服务商需具备行业 knowhow，以设计出符合行业需求的数据解决方案；另一方面，随着模型向更专业化和精细化方向发展，丰富的高质量垂类专家资源也至关重要，目前数据服务公司会在高校密集地区建立数据标注基地，以获取更广泛的高质量的数据标注资源。

(3) 具身智能的训练数据市场呈现出巨大的供需缺口，需求旺盛且潜力巨大

具身智能作为实现通用人工智能（AGI）的关键路径与终极载体，正在受到更多的关注。国家层面，工信部等七部门联合印发《关于推动未来产业创新发展的实施意见》，将具身智能纳入未来制造、未来信息等六大战略方向。地方层面，北京、深圳、广州、上海等城市已出台专项行动计划，通过资金以及政策支持，深入推动具身智能产业发展。具身智能需要机器人在复杂的真实世界中实现自主感知、学习和适应，而该能力的构建则依赖大量的高质量的数据。目前，业内主要包含四种数据获取方式，分别是遥操作技术收集的真机数据、3D 环境中模拟的合成数据、动作

捕捉获取的人类动作数据、以及从互联网收集的非结构化视频或图像数据。

然而，具身智能若想达到更佳的训练效果，尤其是在泛化能力提升方面（即让机器人能够在陌生环境下自主识别和完成任务），仍需要海量的（约数十亿级别）来自“真实物理环境”的动态交互数据进行训练，而非简单的合成数据或互联网获取的数据。但目前，由于真实世界数据获取成本高、时间长，数据体量远无法支持通用机器人的发展。因此，目前具身智能的训练数据市场正面临巨大的供需缺口，需求旺盛，未来增长潜力巨大。

（4）数字经济发展催生新型数据服务模式

发展数字经济将成为我国经济“弯道超车”以及挖掘经济内生增长的重要战略举措。国家在数字经济建设方面决心极为坚定，通过《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》、《数字中国建设整体布局规划》等政策文件的密集发布以及组建成立国家数据局等方式，进一步统筹并加速落地数字经济发展战略，而数据要素作为深化数字经济发展的核心引擎，也将迎来新的发展机遇。未来，围绕数据确权、汇聚、处理、利用和流通等环节将会产生巨大的增量市场空间，催生出围绕公共数据以及行业数据开发的新型数据服务需求，以及以基地建设、数据平台开发运营、数据交易为代表的新业态、新模式。

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

	2024年	2023年	本年比上年 增减(%)	2022年
总资产	808,464,516.38	824,507,109.18	-1.95	876,927,792.15
归属于上市公司股东 的净资产	743,282,633.50	782,293,983.51	-4.99	829,522,849.25
营业收入	237,083,030.07	170,010,956.57	39.45	262,887,869.44
归属于上市公司股东 的净利润	11,336,089.30	-30,385,187.56	不适用	29,454,139.23
归属于上市公司股东 的扣除非经常性 损益的净利润	4,973,475.19	-43,470,684.50	不适用	10,149,073.69
经营活动产生的现 金流量净额	28,733,413.17	-31,046,209.61	不适用	30,658,908.30
加权平均净资产收 益率(%)	1.49	-3.77	增加5.26个百分点	3.61
基本每股收益(元 /股)	0.19	-0.50	不适用	0.49
稀释每股收益(元 /股)	0.19	-0.50	不适用	0.49
研发投入占营业收	26.76	34.40	减少7.64个百分点	35.86

入的比例（%）				
---------	--	--	--	--

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	40,645,255.83	51,781,124.52	57,131,120.08	87,525,529.64
归属于上市公司股东的净利润	-634,142.17	1,050,563.43	3,372,306.04	7,547,362.00
归属于上市公司股东的扣除非经常性损益后的净利润	-2,646,313.72	-1,029,535.88	1,737,541.20	6,911,783.59
经营活动产生的现金流量净额	4,985,859.22	10,690,996.98	192,796.74	12,863,760.23

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)					12,577		
年度报告披露日前上一月末的普通股股东总数(户)					12,533		
截至报告期末表决权恢复的优先股股东总数（户）					-		
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）					-		
截至报告期末持有特别表决权股份的股东总数（户）					-		
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）					-		
前十名股东持股情况（不含通过转融通出借股份）							
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	质押、标记或冻 结情况		股东 性质
					股份 状态	数量	
贺琳	0	12,137,615	20.12	-	无	0	境内自然 人
北京中瑞安投资中心（有限合伙）	0	6,935,780	11.50	-	无	0	其他

中移投资控股有限责任公司	0	4,797,881	7.95	-	无	0	国有法人
北京清德投资中心（有限合伙）	0	2,824,448	4.68	-	无	0	其他
宁波丰琬创业投资合伙企业（有限合伙）	0	2,033,324	3.37	-	无	0	其他
唐涤飞	-1,064,029	1,452,125	2.41	-	无	0	境内自然人
全国社保基金四一四组合	100,000	613,694	1.02	-	无	0	其他
苏州博瑞鑫稳管理咨询合伙企业（有限合伙）	578,425	578,425	0.96	-	无	0	其他
银岳（上海）投资管理有限公司—银岳智选私募证券投资基金	280,006	280,006	0.46	-	无	0	其他
银岳（上海）投资管理有限公司—银岳精选私募证券投资基金	见“注 2”	217,006	0.36	-	无	0	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； 2、银岳（上海）投资管理有限公司—银岳智选私募证券投资基金、银岳（上海）投资管理有限公司—银岳精选私募证券投资基金基金管理人均为银岳（上海）投资管理有限公司。 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			不适用				

存托凭证持有人情况

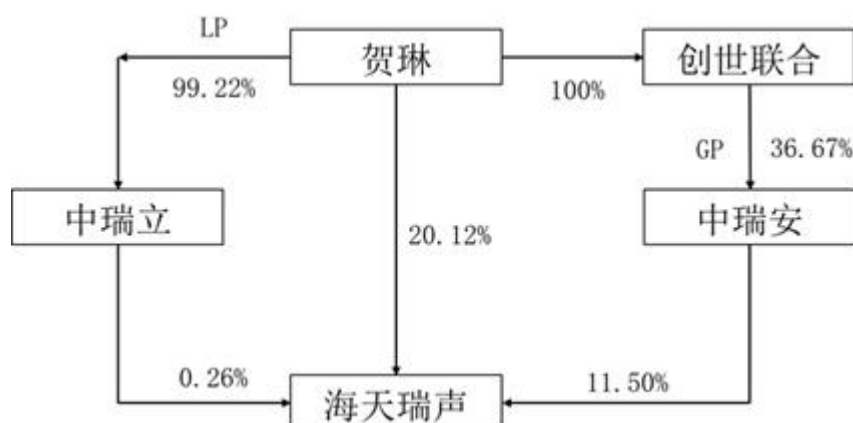
☐适用 ☒不适用

截至报告期末表决权数量前十名股东情况表

☐适用 ☒不适用

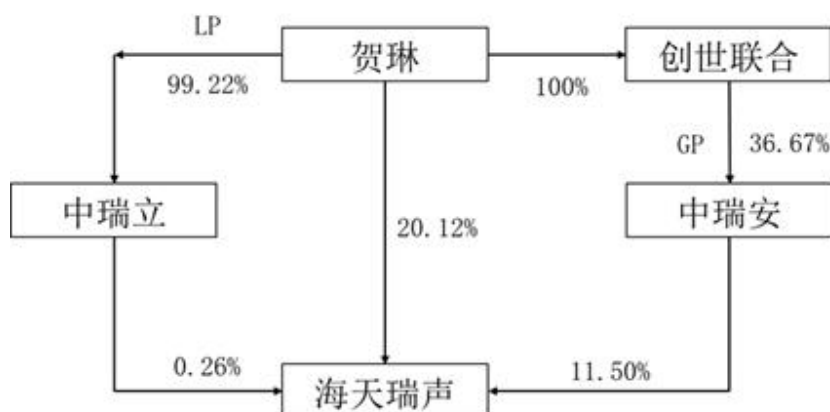
4.2 公司与控股股东之间的产权及控制关系的方框图

☒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，公司实现业务收入 2.37 亿元，较上年同期增加 39.45%；实现归属于母公司所有者净利润 1,133.61 万元，较上年同期增加 4,172.13 万元；扣非后归母净利润 497.35 万元，较上年同期增加 4,844.42 万元。截至报告期末，公司总资产为 8.08 亿元，归属于母公的所有者权益为 7.43 亿元，分别较上年末减少 1.95%和 4.99%。

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