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

2025 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、 本集团董事会及董事、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）为本集团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九次会议审议，公司拟定2025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1.00元（含税），截至2026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5,985,906.30元（含税），现金分红比例为42.40%；2025年度不进行资本公积转增股本，不送红股。

2025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2025年年度股东会审议通过。

母公司存在未弥补亏损

□适用 √不适用

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

2.1.1 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、中国移动、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 1,200 家，覆盖了科技互联网、社交、IoT、具身智能、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。

图：公司产品服务矩阵示意

2.1.2 主要产品及服务情况

2.1.2.1 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。

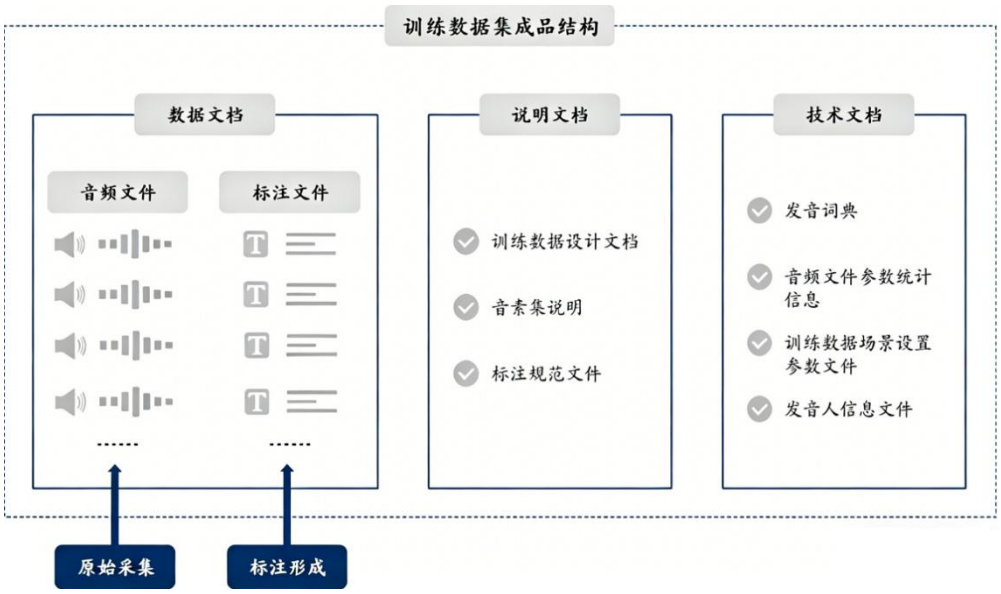
（4）训练数据相关的应用服务

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

2.1.2.2 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

2.2.1 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。

2.2.2 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。

2.2.3 采购模式

公司实行集中采购与分散采购相结合的采购管理模式，建立了规范的采购管理制度与供应商管理体系。

按照采购对象区分，公司的采购主要包括业务项目采购和公司常规采购，按照采购品类进一步划分为数据服务采购、岗位服务采购和其他品类采购。

数据服务采购：指语音识别采集、语音识别转写、语音合成、自然语言处理、图像视频处理、

智能驾驶、具身智能领域等公司主要业务的原材料数据以及原材料数据加工服务采购，主要包括非核心技术环节的原料数据采集、标注服务等。

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。

其他品类采购：（1）常规货物类采购，涵盖日常运营中所需的办公家具、计算机、服务器、办公用品等有形物资；（2）无形资产类采购，包括数据生产、研发活动所需专用平台、专用软件等；（3）日常运营服务类采购，包括云服务、审计服务、差旅服务等。

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，形成权责清晰、流程规范、监督有效的采购管理体系，并与主要的供应商形成了良好稳定的长期合作关系，为公司持续健康发展提供坚实可靠的供应链保障。

2.2.4 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

2.3.1 行业的发展阶段、基本特点、主要技术门槛

2.3.1.1 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处于历史性拐点，政策、技术与应用的三重共振正推动行业进入高速增长通道。

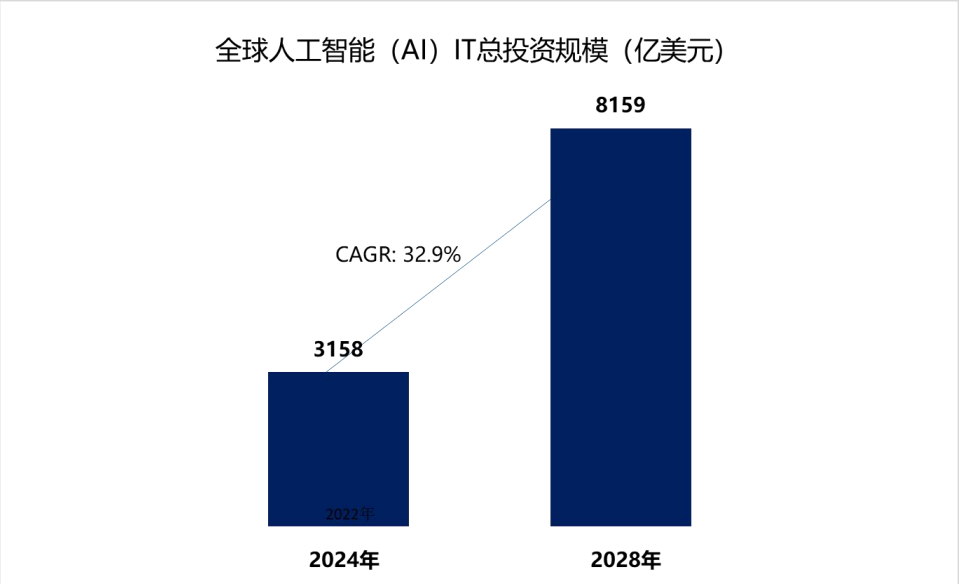
政策层面，主要经济体竞相加码。中国国务院于 2025 年 8 月印发《关于深入实施“人工智能+”行动的意见》，明确提出到 2027 年新一代智能终端、智能体等应用普及率超 70%，至 2035 年全面迈入智能经济与智能社会时代，标志着我国 AI 产业进入“规模化提升”阶段。美国在“星际之门”计划（5000 亿美元基础设施投资）基础上，相继推出“美国 AI 行动计划”与“创世纪计划”，持续扩大领先优势。欧盟于 2025 年 4 月发布《人工智能大陆行动计划》，聚焦算力、数据、

应用、人才与法规五大领域，计划在 2021-2027 年间投资超 100 亿欧元建设 AI 工厂。

技术层面，革命性突破持续涌现。以 DeepSeek R1 为代表的开源模型将 API 调用成本降低 90-95%，大幅降低应用门槛。同时，多模态大模型（如 Google DeepMind 的 Genie 3、OpenAI GPT-5、阿里 Qwen3-VL 等）不断拓展能力边界，实现从语言理解、视觉识别到 3D 世界生成的全模态交互。2025 年被视为“智能体元年”，AI Agent 凭借自主任务规划、动态决策与闭环执行能力，实现从“被动响应指令”向“主动解决复杂问题”的跨越，正成为驱动产业变革的核心力量。

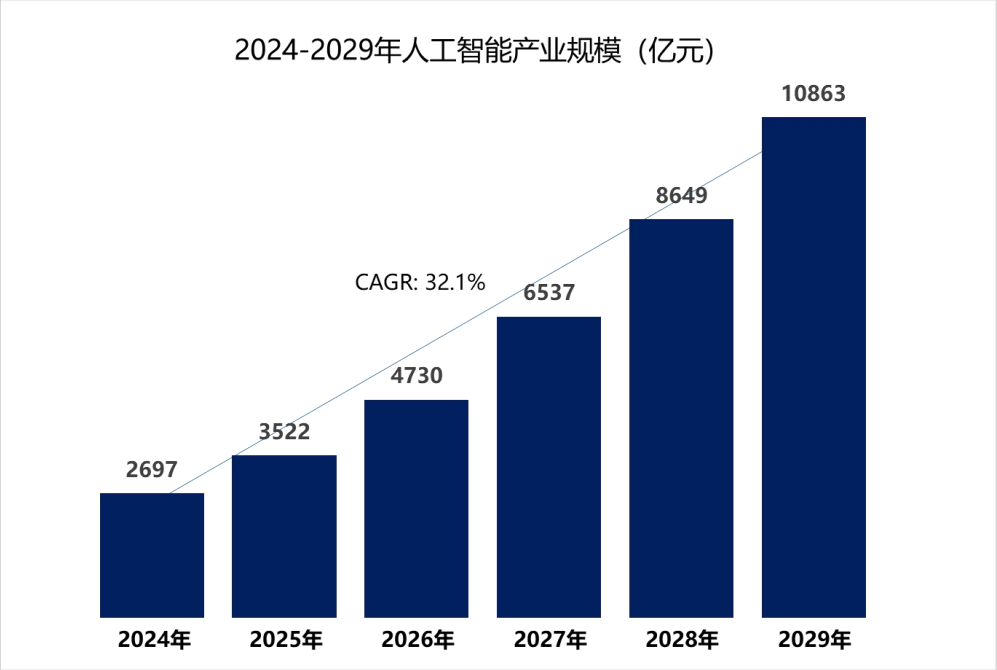
应用层面，技术平权加速 AI 向千行百业渗透。大模型正向金融、医疗、制造等核心领域深度赋能，智能风控、智慧医疗、智能制造等应用场景持续丰富，推动各行业效率提升与模式创新。

总体来看，在政策引导、技术迭代与商业落地的正向循环下，全球 AI 产业正加速迈向规模化、价值化发展的新阶段，迎来前所未有的战略机遇。根据国际数据公司（IDC）的数据，预计全球人工智能(AI)IT 总投资规模在 2028 年增至 8,159 亿美元，2024 年至 2028 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，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。



数据来源：艾瑞咨询

（2）训练数据作为 AI 发展的“燃料”作用更加凸显，成为大模型竞赛中的重要决定因素

算法、算力、数据是 AI 三大核心要素。当前，产业正经历从“以模型为中心”向“以数据为中心”的范式转变。算法端，主流大模型纷纷开源，技术壁垒持续降低；算力端，以 DeepSeek 为代表的架构创新大幅降低了训练成本，算力不再构成发展瓶颈。在此背景下，训练数据的重要性被进一步放大，从“辅助燃料”升级为“核心引擎”。

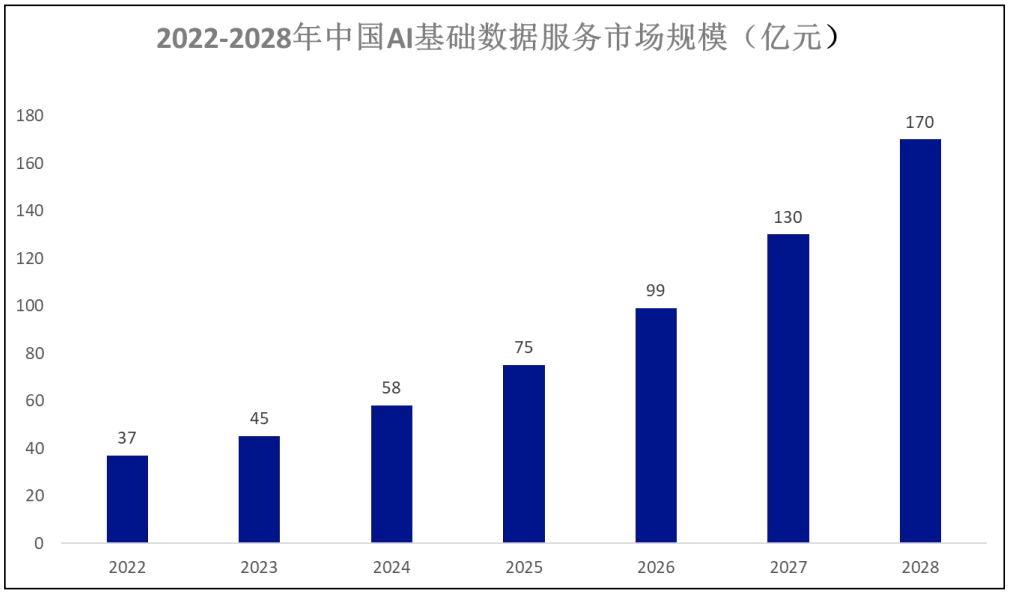
高质量数据直接决定模型能力上限，能显著提升推断可靠性并减少幻觉现象。当前，大模型发展正面临严峻的“数据墙”——高质量数据短缺已成为 AI 规模化落地的关键制约。业内普遍反映，诸多行业大模型未达预期，根源在于数据基础薄弱。因此，数据已成为各国发展 AI 产业的关键胜负手。

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，对高质量训练数据的需求持续攀升。国务院《关于深入实施“人工智能+”行动的意见》及国家数据局相关方案明确提出，将持续加强高质量数据集建设，重点布局多模态、具身智能、推理思维链及长视频数据等方向。在产业和政策双轮驱动下，中国 AI 基础数据服务市场进入加速增长通道。

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）数据要素价值加快释放，数据产业已成为数字经济发展新增长点

a. 政策驱动持续加码，数据制度不断完善

国家数据要素市场化配置改革已进入系统深化阶段。2024年1月，财政部《企业数据资源相关会计处理暂行规定》正式施行，数据资产入表从自选动作转变为规定动作。同年，国家数据局等17部门联合印发《“数据要素×”三年行动计划（2024—2026年）》（国数政策〔2023〕11号），选取工业制造、金融服务、医疗健康等12个行业和领域，推动发挥数据要素乘数效应。《关于促进数据产业高质量发展的指导意见》《关于促进数据标注产业高质量发展的实施意见》等21项政策和指导意见陆续发布，明确到2029年数据产业规模年均复合增长率超15%。从数据基础制度建设年到数据改革攻坚年，国家数据局进一步明确将2026年定调为“数据价值释放年”，加强高质量数据建设，持续支撑人工智能创新发展。

b.市场活力加速迸发，产业布局多点开花

在政策与需求双重驱动下，数据要素市场规模稳步扩大。高质量数据集建设成为“数据要素X”和“人工智能+”两大行动的“焊接点”，国家数据局推动成都、长沙、保定、沈阳等7个城市率先开展承接国家数据标注任务城市建设，先行先试探索产业发展经验，随后，呼和浩特、武汉、南宁等新一批城市也陆续开展强基扩容、标注攻坚、应用赋能等数据标注产业攻坚行动，推动数据标注创新试验区建设。“人工智能+”行动到哪里，高质量数据集的建设和推广就到哪里的发展势头强劲。

c.技术创新持续突破，流通底座日益夯实

核心技术迭代持续赋能数据要素市场化。数据标注领域，大模型辅助自动化标注、生成式AI融入标注流水线，推动人机协同智能化升级，2025-2026年全球数据标注解决方案市场年复合增长率达24.3%。可信数据空间建设进入规模化实践阶段，首批遴选的63个国家级试点项目已全面启动，覆盖国民经济32个行业大类，服务900余个具体应用场景，吸引了近7万家市场主体参与，数据流通利用的基础设施体系正在加速完善。数据要素的流通与利用成本持续降低，技术创新正为数据要素市场化配置构筑起日益坚实的支撑体系。

综上，数据要素正从支撑性资源转变为基础性生产要素，政策、市场、技术协同推进，深度融入企业经营与产业升级，将成为未来十年最重要的新兴生产要素之一。

（4）训练数据领域的未来发展趋势

随着DeepSeek、Gemini等成为现象级应用，以及AI手机、具身智能等终端加速落地，大模型技术正驱动数据需求发生深刻变革。

a.多模态大模型成为主流，驱动多模态数据需求爆发式增长

大模型正从单模态向多模态范式加速演进。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、视频等不同形式的数据，使AI具备更接近人类的全维度认知能力。这种演进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和文本信息，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，通过模拟现实中的视觉推理逻辑，训练AI建立视觉-语言联合表征能力。实践证明，数据质量与多样性直接决定多模态模型的能力上限。随着多模态数据生态的完善，AI的感知与认知能力将实现新跨越。

b. 大模型从“规模驱动”转向“推理驱动”，思维链（CoT）数据成为关键突破口

随着模型参数量逼近实用天花板，传统Scaling Law的边际收益正在递减。单纯堆算力与参数已难以解决逻辑、数学等复杂推理任务，行业正加速向“推理驱动”范式转型——让模型从直觉式“快思考”转向逻辑式“慢思考”。

2025年，DeepSeek R1的推出验证了这一路径的可行性。其核心创新在于思维链（Chain-of-Thought, CoT）技术：通过将复杂问题拆解为多步可追溯的推理步骤，得以模拟人类的分步思考过程，显著提升逻辑一致性与答案可解释性。这一技术突破使CoT数据从“可选项”变为“必选项”。

对于数据服务商而言，CoT数据的供给能力将成为衡量专业水准的关键标尺。率先建立专家标注体系、掌握复杂推理数据生产方法论的企业，将在大模型下一阶段的竞争中占据核心生态位。

c.从通用到垂直，高质量行业数据需求显著提升

DeepSeek等开源模型的高性能、低成本加速了AI应用普及，推动AI从通用助手向行业专家和AI Agent演进。医疗、法律、金融等垂直领域对专业数据的需求激增：医疗大模型要求标注人员具备医学知识，法律模型需理解法条与判例逻辑。同时，AI Agent需要理解用户指令并执行订餐、行程规划等复杂任务，对多轮交互、任务拆解类数据提出新要求。对于数据服务商而言，上述变化意味着核心竞争力正在重构：不再仅仅是“数据产量”的比拼，更是行业理解深度、专家资源网络、复杂任务拆解能力的综合考验。能够为垂直领域提供“数据+知识”一体化解决方案的企业，将在AI产业深水区占据不可替代的位置。

d.具身智能浪潮来袭，数据供给瓶颈亟待突破

具身智能被广泛视为通往AGI的关键一跃。2025年，特斯拉Optimus、Figure 01等机器人加速从实验室走向工厂与家庭。与纯软件AI不同，具身智能要求模型理解并交互于真实物理世界——这一跨越带来了根本性的数据挑战。行业依赖以下种数据来源：互联网开源数据集（规模有

限，仅百万级）、虚拟合成、动作捕捉、第一人称视角（如头戴相机记录人类操作）、UMI（通用操作接口）、真机遥操。虚拟合成数据虽可批量生成，但“仿真到现实”的差距始终存在——物理引擎无法完美模拟摩擦力、形变、光照变化等复杂因素。真机遥操采集（如人类通过VR设备远程操控机器人）能产出最高质量的数据，但单条数据采集成本高达数十元，难以快速规模化。目前，混合式数据策略成为行业共识，即用第一人称视角、UMI数据、虚拟数据等进行预训练、用真实数据精调。

对于数据服务商而言，具身智能是一个全新的蓝海市场。当前行业仍处于“数据荒”阶段，率先建立物理世界数据采集、以及仿真数据能力的数据服务企业，将在具身智能时代占据更强的先发优势。

e.数据安全法规密集落地，合规能力成核心竞争力

近年，《数据安全法》《个人信息保护法》《网络数据安全管理条例》等法律法规相继实施。2025年，国家进一步强化数据出境安全评估、生成式AI内容标识等要求。对于数据服务企业，数据安全与合规能力已成为核心评价维度。能够持续跟踪法律变化、建立成熟安全管理体系、坚持发展与安全并重的企业，将具备更强的市场竞争力。

2.3.1.2 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

随着大模型训练从“以模型为中心”转向“以数据为中心”，头部客户群体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、合理的人机协同。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，多模态数据、CoT 数据、具身智能数据等新型数据类型的涌现，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落地并向更多垂直场景渗透，同时受中国企业出海需求、国外企业全球拓展两方面支撑，客户在多语种、多音色等方面的需求持续提升。多语种数据标注需兼顾发音、语法及文化背景差异。此外，情感标签、语调标记、韵律特征等细粒度语音标注需求日益增加，要求数据服务商在音素集构建、发音词典编制、跨语种迁移学习等基础研究领域具备深厚积累。只有在这一领域长期投入、具备系统性语音语言学研发能力的服务商，才能满足客户在多语种、多场景下的多元化数据需求。

因此，市场上仅有极少数企业通过长期自主研发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。

2.3.2 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	262 人	1,185 人	312 人（截至 2025 年 6 月	未公开披露

公司	海天瑞声	Appen	数据堂	标贝科技
			30日）	
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；国央企，如中国移动等；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1,200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。	专注于智能语音交互和AI数据服务，打造多场景应用的语音交互方案，包括通用场景的语音合成和语音识别，以及TTS音色定制，声音复刻，情感合成等语音技术产品；AI数据业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、跨模态的数据采集和标注服务，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	1,877个	超过700个	超过1,000个	188个

公司	海天瑞声	Appen	数据堂	标贝科技
练数据集数量				
语种/方言覆盖能力	超过300个	超过290个	超过100个	40余个
已取得专利授权	42项（40项发明专利、1项实用新型专利及1项外观设计专利）	未公开披露	63项	36项
计算机软件著作权数量	192项	未公开披露	259项	65项
综合能力				
数据安全能力	北京市规划和自然资源委员会行政许可乙级测绘资质；ISO27001信息安全管理体系认证、ISO27701隐私信息管理体系认证、ISO42001人工智能管理体系认证证书、ISO20000信息技术服务管理体系；国家信息系统安全等级保护三级备案；CMMI成熟度3级认证证书；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	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”、北京市科学技术进步奖二等奖等多个国家或市级重要奖项、北京数字经济企业	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

公司	海天瑞声	Appen	数据堂	标贝科技
	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、国家数据局数据标注优秀案例、国家数据局高质量数据集典型案例、全国第一批数据标注产业伙伴、北京市2025年数字经济标杆企业、《财富》中国科技50强			

注1：Appen、数据堂、标贝科技数据：除特别标注外，均为2025年1-12月/截至2025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：海天瑞声数据：为2025年1-12月/截至2025年12月31日数据。

2.3.3 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。DeepSeek官方已公开R1的完整训练路径，将全过程拆解为冷启动、推理导向RL、拒绝采样再微调、对齐导向RL四步，其中冷启动阶段正是使用数千条能体现思考过程的CoT数据对模型进行监督微调，才使得R1在AIME 2025测试中，准确率由70%提升至87.5%。在医疗影像诊断、法律文书推理、金融风控等专业领域，融入分步推理过程的CoT数据，可使模型掌握从问题解析到结论验证的完整认知闭环，提升专业任务准确性和可解释性。因此，在大模型向垂直领域拓展时，高质量的多领域CoT数据需求预期将快速增加，并成为推动AI技术发展的关键因素。

（2）垂向领域数据需求快速增加，标注复杂度不断提升

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。麦肯锡调研显示，全球78%的组织已在日常

运营中使用某种AI工具，其中85%已将AI Agent集成至少一项工作流程。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know-how，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，丰富的垂类专家资源也至关重要。

（3）AI Agent技术路线快速演进，GUI数据与行为轨迹数据需求已开始呈现增长态势

以OpenClaw为代表的AI Agent开始大规模落地应用，标志着AI从被动响应工具向主动决策执行者的根本性跨越。AI Agent已具备明确的“感知-决策-执行”闭环能力，可应用于采购策略制定、工作流审批、工业设备操控等复杂场景。在技术路线上，GUI Agent路线加速走向成熟，突破传统API调用模式，使智能体能够像人类一样通过视觉识别“看”懂屏幕、利用模拟点击“操作”按钮，实现跨应用自动化操作。然而，GUI Agent的训练面临严峻的数据瓶颈——端到端训练需要海量高质量GUI交互数据，但手动大规模标注行动轨迹成本极高。为此，行业正积极探索从公开屏幕录制视频中自动挖掘训练数据的技术路径，有望大幅降低标注需求。同时，行为轨迹数据的采集与标注成为新焦点，数据服务商需大规模采集鼠标移动、点击、键盘输入、屏幕触摸等完整操作序列，并将宏观看似复杂的任务指令拆解为可训练的微观动作逻辑，为AI Agent训练提供结构化且具情境意义的数据支撑。在Agent迈向规模化落地的关键窗口期，具备多端（PC、移动）行为轨迹数据采集与标注能力的数据服务商将获得显著的差异化竞争优势。

（4）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，正受到越来越多的关注。2026年被行业公认为具身智能的“数据之年”，数据需求正呈指数级爆发——从Pi0的1万小时训练，到Gen-0的27万小时，头部具身大模型所需真机训练数据正逼近甚至超过百万小时级别。业内共识认为，具身模型真正收敛需几百万甚至数千万小时高质量训练数据，但当前国内各家具身智能公司数据总量仅约几十万小时，量级差距巨大。从政策层面看，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；北京、上海、深圳等城市已出台专项行动计划，通过资金与政策支持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，该能力的构建依赖海量来自“真实物理环境”的动态交互数据进行训练。目前数据获取仍面临成本高昂、场景覆盖有限等挑战，高质量具身智能数据市场正呈现巨大的供需缺口，需求旺盛，未来增长潜力巨大。

（5）数字经济发展催生新型数据服务模式

发展数字经济已经成为我国经济“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、国家数据发展研究院、世界数据组织（WDO）等职能部门和组织，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，以及以行业高质量数据集构建、可信数据空间建设运营、数据标注基地建设、数据平台开发运营、数据交易为代表的新产品、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

	2025年	2024年	本年比上年增减(%)	2023年
总资产	861,086,529.72	808,464,516.38	6.51	824,507,109.18
归属于上市公司股东的净资产	741,700,891.76	743,282,633.50	-0.21	782,293,983.51
营业收入	376,972,016.48	237,083,030.07	59.00	170,010,956.57
利润总额	12,816,291.30	10,594,740.28	20.97	-40,851,754.36
归属于上市公司股东的净利润	14,118,462.80	11,336,089.30	24.54	-30,385,187.56
归属于上市公司股东的扣除非经常性损益的净利润	10,785,222.84	4,973,475.19	116.85	-43,470,684.50
经营活动产生的现金流量净额	1,772,560.02	28,733,413.17	-93.83	-31,046,209.61
加权平均净资产收益率（%）	1.90	1.49	增加0.41个百分点	-3.77
基本每股收益（元／股）	0.24	0.19	26.32	-0.50
稀释每股收益（元／股）	0.24	0.19	26.32	-0.50
研发投入占营业收入的比例（%）	15.54	26.76	减少11.22个百分点	34.40

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	69,809,531.77	86,886,417.09	77,638,766.69	142,637,300.93
归属于上市公司股东的净利润	371,626.10	3,432,988.85	375,113.70	9,938,734.15
归属于上市公司股东的扣除非经常性损益后的净利润	-684,831.02	1,668,354.37	-3,178.31	9,804,877.80
经营活动产生的现金流量净额	-15,975,392.28	-17,775,854.87	13,260,193.87	22,263,613.30

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)	11,279
年度报告披露日前上一月末的普通股股东总数(户)	11,651
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	-363,831	11,773,784	19.52	0	无	0	境内自然人

宁波中毅安创业投资合伙企业（有限合伙）	-1,413,551	5,522,229	9.15	0	无	0	其他
中移投资控股有限责任公司	0	4,797,881	7.95	0	无	0	国有法人
北京清德投资中心（有限合伙）	0	2,824,448	4.68	0	无	0	其他
中国工商银行股份有限公司－诺安稳健回报灵活配置混合型证券投资基金	863,765	863,765	1.43	0	无	0	其他
宁波丰琬创业投资合伙企业（有限合伙）	-1,206,252	827,072	1.37	0	无	0	其他
全国社保基金四一四组合	100,000	713,694	1.18	0	无	0	其他
北京海天瑞声科技股份有限公司－2024 年员工持股计划	439,897	439,897	0.73	0	无	0	其他
国投证券股份有限公司－博时上证科创板人工智能交易型开放式指数证券投资基金	435,732	435,732	0.72	0	无	0	其他
江苏银行股份有限公司－诺安积极回报灵活配置混合型证券投资基金	见“注”	426,662	0.71	0	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，公司控股股东、实际控制人贺琳持有 100%股权的盐城创合投资管理有限公司为宁波中毅安创业投资合伙企业（有限合伙）的普通合伙人、执行事务合伙人，并持有宁波中毅安创业投资合伙企业（有限合伙）36.67%的出资；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

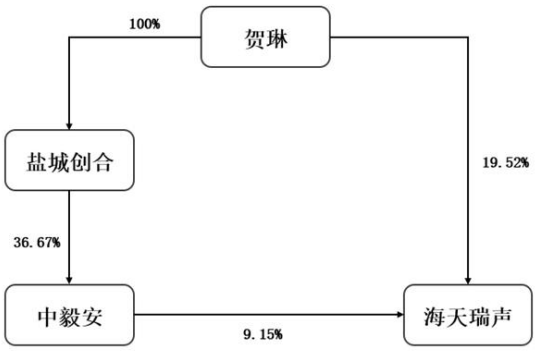
□适用 √不适用

截至报告期末表决权数量前十名股东情况表

□适用 √不适用

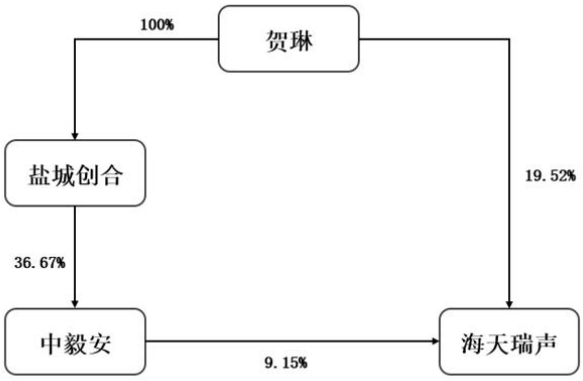
4.2 公司与控股股东之间的产权及控制关系的方框图

√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，公司实现营业收入 3.77 亿元，较上年同期增长 59.00%；归属于母公司所有者的净利润 1,411.85 万元，较上年同期增加 24.54%；归属于母公司所有者的扣除非经常性损益的净利润为 1,078.52 万元，较上年同期增加 116.85%；经营性现金流净额 177.26 万元，较上年同期下降 93.83%。截至报告期末，公司总资产为 8.61 亿元，较期初增加 6.51%；归属于母公司的所有

者权益为 7.42 亿元，较期初减少 0.21%。

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