

证券代码：688787

证券简称：海天瑞声

## 北京海天瑞声科技股份有限公司

### 投资者关系活动记录表

编号：2025-004

<b>投资者关系活动类别</b>	<input type="checkbox"/> 特定对象调研 <input checked="" type="checkbox"/> 分析师会议 <input type="checkbox"/> 媒体采访 <input type="checkbox"/> 业绩说明会 <input type="checkbox"/> 新闻发布会 <input type="checkbox"/> 路演活动 <input type="checkbox"/> 现场参观 <input type="checkbox"/> 电话会议 <input type="checkbox"/> 其他（请文字说明其他活动内容）
<b>参与单位名称及人员姓名</b>	信达澳亚基金 何鑫 Auspice Terry 君茂资本 陈晓燊 景林资产 徐伟 平安证券 窦泽云 华泰证券 郭雅丽 阳光资管 王鹏 上海世诚 陈家琳 博时基金 许振辉 Triata Capital Joseph She 前海开源基金 夏英妍 Stillpoint Investments Julien Jing 百川财富 唐琪 五矿证券 罗芳 国海富兰克林基金 王志欣 华夏基金 彭锐哲 珠海万方资产 陈江

	<p>金建投资 赖昕</p> <p>中信资管 贾可训</p> <p>富达基金 陈月桥</p> <p>谢诺辰阳 郭乐颖</p> <p>农银汇理 邢军亮</p> <p>CPE 源峰 周清逸</p> <p>鹏华基金 陈大烨</p> <p>大成基金 方向、徐一清</p>
会议时间	<p>2025 年 3 月 13 日</p> <p>2025 年 3 月 19 日</p> <p>2025 年 3 月 24 日</p>
会议地点	现场交流
上市公司接待人员姓名	投资者关系 袁璐
投资者关系活动主要内容介绍	<p>1、2024 年收入增长的驱动因素是什么？</p> <p>受益于大模型技术的快速发展以及应用场景不断落地，以智能终端厂商、科技互联网公司为代表的国内外科技巨头纷纷加大多模态数据投入以支撑其智能终端、内容生成等领域的 AI 能力建设。在此背景下，以多语种、多音色为代表的智能语音业务需求、以及以指令微调、偏好对齐等为代表的自然语言业务需求同比均呈现大幅增长，整体上驱动公司营业收入同比显著增加。</p> <p>2、DeepSeek 出来后，对数据需求的影响如何？是否会降低 AI 行业对数据的需求？</p> <p>（1）Deepseek 推出了一系列模型，其中 V3 模型依然使用了预训练、以及 SFT 等训练方式，其中预训练阶段的 token 使用量达到了 14.8T，远超 GPT4 等同类可比大模型预训练阶段的数据使用量，且在后训练阶段也使用了一定规模的标注数据，这也更加说明海量以及</p>

高质量数据对于基础模型能力提升的重要意义。

(2) 关于让大家震撼的 R1 模型，基于目前的公开信息来看，其部分优势体现在推理类任务上，尤其是那些具备较强的规则性、可以推导的任务类型上，确实不需要大量的人工标注，但是对于其他领域（尤其是更为广阔的垂向领域）的复杂问题，依然需要观察，我们认为高阶的数据专家的参与依然非常重要。

(3) 此外，数据质量不仅影响模型获取和表达知识的能力，还决定了模型生成内容的风格和准确性，帮助 DeepSeek 实现了在输出端的文采能力提升。

其一，高质量数据可以提升模型表达和推理能力。优质数据包含准确、连贯且富有表现力的语言样本。例如，包含 CoT 数据可以引导模型在推理时进行反思，进而在生成回答时展现出清晰的逻辑和优美的语言表达。这正是 DeepSeek 模型能够生成既准确又具有华丽文风的关键因素之一。

其二，高质量数据可以降低噪音和确保一致性。数据中的错误、噪音或不一致信息会导致模型生成内容出现语法或逻辑问题。高质量的数据则能有效减少这些问题，使模型更好地学习到语言规律，从而提高整体生成质量。

其三，高质量数据可以提升泛化能力。数据的多样性和全面性使得模型在面对不同领域和任务时都能生成高质量的回答。丰富且准确的样本帮助模型在多种场景下自如切换风格，无论是精炼的技术解答还是文采斐然的创意写作，都能游刃有余。

(4) 往未来看，Deepseek 模型的出现，有望进一步助推模型向产业端发展，真正让大模型技术深入渗透到各个行业中，这一过程中必将凸显专业知识的直要

性，需要更多数据、以及数据专家的参与，因此我们看好并期待未来大模型在各行业百花齐放的局面。

3、未来 AI 数据如果自动标注了，会对公司及行业产生影响吗？

首先，AI 一定不会实现完全的自动化标注，因为机器如果想要持续演进，使其更接近于人类的判断和理解，就一定需要人类作为引导，通过人工标注帮助其完成新知识的学习，所以只要人工智能在持续发展和进化就一定需要人类参与，即无法达到完全的自动化标注。

另一方面，更加智能化的人机协作模式一直是数据服务行业的发展趋势，同时也是数据服务企业的核心竞争能力之一，自动化标注的核心不是完全替代人类，而是提高人机协作效率，海天瑞声近年来在研发领域持续加大投入，不断提升公司数据生产的智能化水平，并据此形成规模效应、实现降本增效。

4、公司的业务是否存在规模效应？

公司业务是存在规模效应的，一方面随着公司在研发方面加大投入，自研平台的能力逐步提升，可以赋能数据处理过程中的人机协作朝着更加智能化的方向前进，这就使得公司进行更大规模的数据生产成为可能。同时，数据产品的积累、平台以及工具的研发，在公司业务规模逐渐上升的情况下，相关的研发费用、管理费用将被摊薄；

从成本端看，数据生产的成本还有很大的下沉空间，对于成本控制我们会在两方面进行持续投入：一方面是继续加大技术投入，采用更为合理的人机协同比例完成数据处理任务，降低人员投入，提高处理效率；另一方面是加强供应链资源管理能力，扩大资源供给，降低单位成本。

此外，数据集产品一直是我们公司所坚持的重点方向，公司开发大量通用型、复卖率高的标准化产品数据集，反复给公司带来利润，也能实现训练数据产品的规模化效应。

5、成本结构里最大的部分是什么？如何能够持续性的优化成本结构？

公司最大的成本就是原料数据采购费用，即：采集、标注成本。一方面，公司通过继续加大研发投入的力度，全面提升公司的算法能力、工程化能力，加深算法辅助能力与人工工作的结合，达到更佳的人机协同，这样能够做大规模、提升效率、降低成本；另一方面是加强供应链资源管理能力，扩大资源供给，降低单位成本。

6、公司提供的训练数据整体解决方案中，各个环节的技术难度如何？

首先，训练数据集的设计和原料数据采集环节是存在相当的技术难度的，比如语音类采集，文本设计是否贴合实际场景、如何实现最小采集量且确保场景覆盖丰富度等因素均是设计和采集环节需要考虑和解决的；在视觉类采集方面，复杂的人像采集、物体影像采集，同样具有如何设计合理的数据浓度达到最小成本最高训练效果，如果是垂直行业数据集的采集，例如交通行业内的自动驾驶领域，则存在准入资质、技术难度（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）等方面的门槛。

数据标注环节的难度在于面对大量的数据标注需求，如何快速的找到充足的资源，而且通过算法平台实现机器的辅助标注，并在人机协作过程中，寻找效率与质量的平衡，在提升数据标注效率、保证数据质量的同时降低成本。

7、标品化的产品数据集业务与定制化服务业务的区别是什么？

产品数据集是先于客户需求形成的模拟数据，是公司区别于其他竞争对手的一大特色，基于公司对市场的判断和通用化需求的提取能力，其属于是一次性投入、未来重复授权销售，对于公司的营收、毛利有着重要作用；而定制业务的需求来源是客户的定向化需求，有些定制业务的原始数据来源是客户提供的实网数据，公司提供纯加工的服务。

客户的 AI 产品在上线之前及初期，因为其自身尚未产生实网数据，通常需要采购模拟型数据集进行算法模型的训练，在产品上线并运行一段时间、产生大量实网数据之后，则会提供实网数据给到我们进行数据加工，加工的数据反哺到客户的产品上从而促进其产品的迭代、升级。之后，客户需要进行产品功能或语种的拓展，再次需要购买模拟数据集来支撑，后续再采购数据加工服务进行迭代。

8、训练数据产品和服务的定价模式、收费模式是什么样的？

定制服务定价模式：一般采用成本加成定价法。公司根据客户的具体服务需求预估项目成本，在预估成本的基础上，参考公司制定的指导毛利率水平，结合项目技术难度、复杂程度、时限要求等进行报价，并根据市场环境与客户协商，最终确定价格。

产品定价模式：一般采用需求导向定价法。公司综合考虑训练数据集的开发支出、市场需求程度、预计未来重复销售的频率等因素，制定产品标准价格及价格区间，在销售过程中，根据客户的实际需求情况，以价格区间为基础向客户报价，经双方协商确定最终销售价

	<p>格。训练数据产品通常以单个数据集为单位进行定价，定价比较灵活。</p> <p>9、定制数据逐渐积累，是否可以转化为自有的数据产品？</p> <p>客户定制服务涉及的训练数据在交付给客户并完成验收后，所有权完全转移给客户，海天瑞声是不能用于自身产品建设的，这一点是公司始终遵循的知识产权要求。</p> <p>在定制数据集的生产过程中，积累下来的经验、know-how 会帮助公司各方面能力的提升，例如工具平台因为处理了大量的定制数据集，使平台完善性有很大的增益，加强了公司的数据处理能力；再如，在一些情况下，公司在生产定制数据集时，也会根据对行业需求的判断，在保障数据权属划分清晰的前提下，利用团队管理、资源获取的便利性，同步安排额外的设计、采集和标注工作，完成产品数据集的开发。</p> <p>10、行业里的玩家增多，会不会出现价格战？</p> <p>是否存在价格竞争主要取决于该领域是否较为存在较高壁垒。在较为成熟的细分方向，比如中文智能语音数据领域，确实存在进入者增多、价格竞争的情况；但对于外语种领域，数据服务商则会有更高的议价空间。所以，未来公司将主攻有较高技术壁垒，存在较大毛利空间的细分场景，尽力避免价格竞争带来的过度消耗。</p> <p>此外，公司也将通过持续的专项研发投入及研发升级，进一步提高自研平台能力，通过智能化促进产能提升、效率提升、成本降低实现规模效应和盈利能力的提升。</p>
附件清单（如有）	

日期	2025 年 3 月 25 日
----	-----------------