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：2025-006

投资者关系活动类别	□特定对象调研　　√分析师会议 □媒体采访　　□业绩说明会 □新闻发布会　　□路演活动 □现场参观　　□电话会议 □其他（请文字说明其他活动内容）
参与单位名称及人员姓名	信达澳亚 徐聪、孔文彬 天风证券 缪欣君 志开投资　王龙 华泰保兴基金　张立晨 中金公司　冯达 中邮创业基金　武若愚 银华基金 郭磊、同颖茜 中金资管　彭祺程 建信基金　江映德 广发基金北京区　周智硕 易知投资　程伟庆 国寿安保　严堃 正圆投资　吴皓月 亚太财险　朱军宁 上海兆顺　钱之皓
会议时间	2025 年 5 月 6 日 2025 年 5 月 7 日

	2025 年 5 月 8 日
会议地点	线上交流、现场交流
上市公司接待人员姓名	董事会秘书 张哲
投资者关系活动主要内容介绍	1、2025 年第一季度，公司收入增长的驱动因素是什么？ 随着多模态大模型的快速迭代及行业应用渗透提速，公司计算机视觉业务和自然语言业务分别同比实现高速增长。其中，在国家对“AI+数据要素”政策同步发力的背景下，以运营商、互联网平台公司为代表的大型客户持续加码高质量图像/视频等多模态数据采购，为其通用多模态大模型训练提供有力支撑；同时，政务、法律合规等场景应用的落地，带动场景类文本数据需求快速增加。在全球化布局方面，公司去年在东南亚新增建设的数据交付体系已进入爬坡运营阶段，通过拓展海外定制服务市场，不仅带来了可观的增量收入，并有望成为海外业务扩展新的战略支点。上述因素，共同驱动公司 2025 年第一季度营业收入显著增长。 2、公司 2025 年第一季度收入结构中，语音、视觉、文本的占比分别是多少？ 2025 年第一季度，公司收入结构呈现阶段性显著变化：视觉业务占比超过 45%，首次超越语音业务占比，文本业务占比约 10%。这一结构性转变主要源于，在大模型逐步向多模态领域发展的大背景下，公司一季度头部客户的数据需求集中体现在计算机视觉领域，带动视觉业务收入实现快速增长。展望全年，各业务方向的收入占比预计也将呈现动态变化。 3、训练特定垂向领域的大模型所需的数据，主要来源于哪里？

	目前来看，训练垂直领域大模型的核心数据来源可分为三类：公开数据、客户自有数据和垂直场景定向采集数据。其中，公开数据（如互联网知识库、开源数据集和行业标准文档）可以为模型提供基础数据支撑；客户自有数据和定向采集数据则针对具体业务场景进行专项优化。值得注意的是，这些原始数据必须经过专业处理流程才能投入使用，主要包括：1）数据清洗与标准化；2）格式转换（如语音转文本）；3）领域专家标注与校验。以智能病历系统开发为例，数据加工流程包括：首先将门诊录音转为文本数据，再由医学专家进行专业校对并提取关键临床信息，最终生成结构化电子病历。这一过程高度依赖专业领域知识，需要大量临床医师参与质量把控。正因如此，在垂直领域大模型训练中，专业数据服务商扮演着双重角色：既是特定领域高质量数据的提供方，也是专业数据加工服务的提供商。 4、国内大模型数据的来源有哪些？ 根据艾瑞咨询的调研报告，目前大模型训练主要有5类数据来源：分别是，可以公开获取的数据（可直接利用下载的数据，如来自高校、社区的免费共享数据）、网络爬虫数据、采购数据（通过第三方数据服务商）、大模型应用及客户合作数据以及企业自有数据。 未来，随着AI产业、数据要素产业的进一步交融发展，预计公共数据的有序开放流通也将带来更大规模的大模型数据供给。 5、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？ （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。 6、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。 7、公司的核心竞争力主要体现在哪？ （1）公司的业务模式是服务产品双模式，且产品化贡献显著，是收入和毛利的主要来源，标准化数据集的研、产、销体系是公司从业多年探索出来的业务

	模式，其复用性为公司的规模化和高利润率提供了保障。而保持这样的能力需要具备对行业需求的强判断力和较强的资金实力。截至2024年12月末，公司已积累超过1,700个自有知识产权的训练数据标准化产品，数据库存量稳居全球企业前列。 （2）技术平台能力：公司历来重视技术的研发，近年来更是加大研发投入的力度，全面提升公司的算法能力、平台能力、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效率，这样能够做大规模、提升效率、降低成本。 （3）供应链资源管理能力：公司通过长期建设的供应链体系，保障资源的获取，未来，公司会进一步加大供应链资源平台的建设，使人员管理、采标资源分配、质量检验、远程工作等各方面的能力得到显著提升，为客群拓展提供有力支撑。 （4）数据安全及合规能力：数据安全及合规能力已经成为了衡量品牌数据服务商综合能力的重要指标。公司在多年数据风险识别和管理实践中，已形成了较为成熟的安全、合规管理体系。 8、公司的主要竞争对手有哪些？ 从短期来看，公司竞对仍是传统模式下的数据服务公司，国内的主要竞争对手是一些品牌数据提供商，如数据堂、标贝以及一些新兴公司；国外的主要竞争对手是Appen。 与竞争对手相比，海天瑞声自身还是存在显著的竞争优势的，如丰富的产品积累、成熟的数据处理技术和平台、全球化的供应链管理能力等等。另外，从公司创业历程看，由于长期与国际性科技企业合作，对数据安全和合规的重视是深入到公司运作的方方面

	面的。而数据安全和合规是需要投入较高的成本建设的，在日益完善的法律环境下，这方面的投入为公司带来了新的竞争壁垒，也将会为公司未来在垂直行业和政企业务拓展形成有利优势。 但从长期来看，随着训练数据需求逐渐向高品质、规模化、行业化方向转变，基于自身持续研发能力建设的数据生产智能化程度将成为数据服务商的核心竞争力，因此，未来诸如 Scale AI 这类具有更强技术属性的同业公司将成为海天的主要竞争对手，为此海天自身已经开始在研发、人才等方面大规模持续投入，为未来竞争提前布局。 9、公司与运营商的合作进展如何？ 在国家"AI+数据要素"战略的指引下，尤其是国务院国资委连续两年开年启动部署中央企业“AI+”专项行动以来，以运营商为代表的重点央企自 2024 年起加速布局通用+垂向大模型研发，带动了高质量图像、视频等训练数据的规模化采购需求。公司凭借在数据领域的核心优势，已快速成为运营商类客户重要的数据服务供应商。未来，随着以运营商为代表的重点央企在多模态大模型方向的持续加码，以及其基座大模型在更多传统行业的应用落地，预计相关数据需求将进一步增长，为公司收入带来持续的增长动能。
附件清单（如有）	
日期	2025 年 5 月 8 日