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：2024-016

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	东吴证券 黄诗涛
会议时间	2024 年 10 月 10 日
会议地点	公司会议室
上市公司接待人员姓名	证券部总经理 张哲
投资者关系活动主要内容介绍	1、数据标注行业未来会有什么样的发展趋势？ 首先是更加智能化，即通过拓展算法覆盖的场景以及算法预识别的准确率等，持续提升机器参与程度以及人机协作效率，降低数据处理成本。 其次，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，积极响应监管政策的企业将具有更强的市场竞争力。 此外，随着境内、外企业的全球化扩张成为确定性

	趋势以及各类客户群体扩张步伐加速，多语种能力作为支撑企业顺利出海的核心要素之一，重要意义更加凸显，具有强大语言研究能力的数据服务企业将获得更多商业机会。 另外，随着数据服务向多元化、多类型、多场景持续发展，充足、稳定且高质量的数据处理团队储备、以及更加智能化的资源配置能力，将成为数据高效、稳定交付的重要保障。 2、公司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。

	3、训练数据的生产过程是什么样的？ 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） ① 设计——训练数据集结构设计 在设计环节中，通过考虑算法模型的具体应用领域、应用场景以及预期实现的训练效果，反过来确定训练数据集内的数据类型、数量、比例分布等，相应确定原料数据的采集要求，为后续采集工作奠定基础。以语音识别、语音合成领域的训练数据集为例，在原料数据的采集环节，发音人（被采集对象）需要朗读公司提供的基础语料，并用指定的录音设备录制以形成原料音频数据。因此，在设计阶段，公司就需要考虑如何设计基础语料，才能使得容量有限的训练数据集能够覆盖尽可能多的自然语言现象，如覆盖更多的发音习惯、语言特点、句长分布，达到更好的音素平衡效果等，从而使得算法模型获得更好的训练结果。 ②采集——获取原料数据 根据此前设计好的训练数据集结构及数据量目标，制定原料数据采集方案并开展原始数据采集工作。采集过程所涉及的主要考虑因素包括： A.数据量方面：需根据成品训练数据集的目标数据量，预留少量冗余。在实际采集过程中，由于可能发生少量录音不合格的损耗情况，通常会在总采集数据量中预留少量冗余，从而略大于最终要交付的数据量，以备替换偶然出现的不合格录音数据。 B.数据属性方面：在采集环节中，根据客户算法模型应用的目标场景、领域等个性化需求，采集特定原料数据。以语音识别训练数据为例，在采集环节中，通常

	需要根据语音识别模型的语种/方言类别、目标应用场景（安静、噪音；家居、车载等），相应定义寻找符合要求的发音人，在合适的采集场景下由发音人朗读、或自然说出录制语音片段，生产原料音频数据。以语音合成训练数据为例，通常需要根据客户对拟合成的语音的风格（温柔、甜美、科技感等）、年龄（成人、儿童）、性别、语种、口音等方面的具体需求寻找发音人，并组织发音人按照前期设计完成的音素集、语料库等资料进行朗读，录制生成原料音频数据。此外，由于语音合成训练数据的录制对信噪比、底噪、录音棚混响时间等参数、指标和录音设备的要求很高，通常需要在专业级别的录音棚中完成录制工作。 ③加工——数据标注 通过公司 ADS 和 VDS 平台，对语音、文本、图片等原料数据进行标注，使其成为结构化可被算法识别和学习的专业训练数据集。该环节中，公司通常会应用相关算法模型，通过算法完成预识别和预标注，可以显著提高数据标注效率，降低标注成本。 ④质检——各环节数据质量检测 质检环节会渗透在整个训练数据的全生产流程，具体包括： A.在前端采集环节，公司开发的采集工具可对原始数据质量进行即时质检，不符合要求的原始数据不被计入采集数据之中； B.在中端加工环节，公司运用自动标注工具+人工校对检验的方式对数据加工情况进行检查，提升加工效率和准确度； C.在后端大规模质检环节，公司运用全自动校验技术，实现大规模训练数据集的质检需求。

	4、客户对训练数据是否有持续需求？ 客户对训练数据本身的需求是会长期持续的。 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，如此周而复始。因此，客户对训练数据的需求是持续的，且随着应用 AI 技术的场景越来越多，各种场景的数据集需求会兴起，带来的是训练数据的需求会越来越大。 5、决定智能驾驶数据业务市场需求空间的因素有哪些？未来智能驾驶的数据需求如何？ 智能驾驶数据业务的市场需求主要与三个要素相关：1）车厂的车型及传感器丰富度。通常来说，不同车型、不同传感器会有不同的硬件配置方案，继而需要不同的数据解决方案，因此车型/传感器等硬件配置的多样性程度将会直接影响所需数据解决方案的数量；2）量产车数量。量产车的数量决定了整个的训练数据需求基数的大小；3）智能驾驶级别的逐渐提升。智能驾驶级别和渗透率的提升决定了数据处理场景的种类和体量。 这三个要素对训练数据需求的影响是相互叠加的。公司预测，随着智能驾驶相关政策的推出以及单车成本的不断下降，智能驾驶的商业化进程将加速，在上述三个因素的共同作用下，数据处理需求将呈现指数级增长趋势。

	6、强化学习阶段的数据服务，今年有何进展或者变化？ 目前来看，随着各大模型的陆续上线，强化学习环节的整体数据需求在逐渐攀升，并在具体标注任务上呈现如下趋势： （1）逐渐向更多垂类拓展（例如，法律、金融、医疗）； （2）强化学习标注的评价/评分指标变得更为丰富，会要求标注人员从更多维度对模型的问答进行评判和打分； （3）由单模态向多模态转变：23 年主要的标注需求集中在文本类标注，今年开始逐步向多模态拓展（例如，文本-视频、文本-图像等）。 7、大模型领域的数据标注是否用到了自动化的方式？ 目前来看，大模型领域的标注任务主要集中在 SFT（有监督微调）、以及 RLHF（强化学习）等环节，具体标注方式包括分类、改写、评分、创作等，以上标注类型均主要依赖人工进行标注，需要标注人员对问题或答案的质量、类型等进行逐一判断或拟写，目前部分项目已经引入了算法自动化预标注策略来提升人工标注与校对的效率。 8、公司未来发展规划是什么？ 公司将自身发展战略定位为以下三个方向： （1）全球化业务：为更好把握国际市场需求，公司将推出一项更为全面的出海战略，涵盖技术创新、品牌升级、体系构建、市场推广等，全面加速全球市场的拓展。公司还将建立一个海外技术研发体系，紧跟全球 AI 的发展动态，并积极开发与海外新兴技术相

	适应的 AI 数据解决方案，以不断增强公司在国际市场的竞争力。 （2）智能驾驶业务：公司将积极把握智能驾驶领域的发展良机，继续升级自动驾驶数据平台 DOTS-AD；同时，不断完善算法技术，提高人机交互的数据处理效率及实现规模化效应；此外，公司会继续加强数据安全管理体系建设，确保数据处理流程的安全与合规；并进一步有效利用已获得的测绘资质，延伸数据服务范围，以提升公司智能驾驶业务的毛利水平。 （3）新兴业务探索—大模型、数据要素：公司将持续探索围绕大模型所需数据相关服务，通过前沿技术跟踪研究，开展以预训练、强化学习为代表的多元化数据获取、高阶垂向拓展等方向的数据服务能力建设；此外，还将探索以数据治理、数据交易、数据处理等为核心的数据要素领域，力争将数据要素创新业务打造成为具有潜在高增长价值的新兴业务板块。
附件清单（如有）	
日期	2024 年 10 月 10 日