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：2023-004

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	广发基金 邱世磊、李阳 龙猛投资 张瀑 中泰证券 苏怡 东吴证券 张文佳 招商证券 孟林 天风证券 缪欣君 招商基金 陈西中 东吴证券 王紫敬、王世杰 博时基金 黄继晨、于福江
会议时间	2023 年 2 月 15 日 2023 年 2 月 16 日 2023 年 2 月 17 日
会议地点	海天瑞声会议室
上市公司接待人员姓名	董事会秘书：吕思遥 证券事务代表：张哲 投资者关系负责人：袁璐
投资者关系活动主要内	1、ChatGPT 是人工智能的又一次革命性创新，5

容介绍	天之内用户突破 100 万，请问公司如何看待 ChatGPT 的发展以及其对公司业务的影响？ 我们很欣喜看到人工智能取得新的突破，迎来新的发展阶段，但同时我们应理性看待技术发展，技术从兴起到广泛应用落地仍需要产业链上企业持续的努力和探索。从目前公司实际情况来看，短期内暂未看到由 AIGC 带来的大幅订单增长，但公司会持续关注该领域最新发展，抢抓相关领域的新发展机遇。 2、请问 OpenAI 是否是公司客户？国内 BAT 都在做自己的大模型，公司是否有为其提供过大模型相关的训练数据？ OpenAI 不是公司客户。BAT 一直以来都是的公司的重要客户，公司为其提供智能语音、计算机视觉以及自然语言等各类数据集产品或服务，但出售的数据集具体应用领域还请以客户的相关发布为准。 3、公司是否从事与算法相关的业务？OpenAI 是否是公司的客户？ChatGPT 将会给海天业务带来怎样的影响？ 公司专注于为包括 AI 技术公司在内的 AI 产业链各类机构提供算法模型开发训练所需的专业数据集，业务与从事人工智能算法及应用开发的企业有比较大的区别。截止目前，公司未与 OpenAI 开展合作，其 ChatGPT 的产品和服务未给公司带来业务收入，该领域对数据需求的发展趋势有待观察。 4、ChatGPT 代表的大模型或 AIGC 的快速发展会对公司产生什么影响？AIGC 相关业务是否会起量？ 公司也注意到 ChatGPT 等 AIGC 类话题近期在国内、国外产业界引发了大量关注和讨论，公司自身始终秉承冷静、理性、专业的态度看待包括 ChatGPT 等

	在内的各类产业界新现象对公司业务所能产生的实质影响，公司认为整个AIGC领域未来将保持长期向上发展趋势，但其发展速度、阶段性效果等需要冷静分析、避免短期盲目过热，公司始终坚信需要回归到行业及公司业务基本面看待各类新现象所能产生的价值，无论行业发展浪潮处于何种阶段，公司应本着实事就是的态度，专注业务本身，真正提高自身的核心竞争力。从目前公司实际情况来看，短期内暂未看到由AIGC带来的大幅订单增长，公司会持续关注该领域最新发展，抢抓相关领域的新发展机遇。 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、智能驾驶的市场竞争格局如何？ 智能驾驶市场主要参与者有品牌数据服务商，客户自建团队以及一些中小服务商。从目前行业格局来

	看，品牌服务商占据较大比例的市场份额。根据海天观察，在品牌数据服务商里，Appen 和百度智能云数据众包在该领域实现较早布局、处于领先；海天从去年开始发力该领域，并已经实现了第四代智能驾驶标注平台的上线，未来将会通过持续提升平台和算法的能力、拓展客户资源，加速培育能力，力争未来在该领域实现高速增长。 7、智能驾驶行业的核心竞争力是什么？ 智能驾驶数据领域的核心竞争力主要体现在三个方面，分别是平台能力、算法能力以及数据安全能力。 平台能力是数据标注能力的基石。平台功能点覆盖的丰富度是评价平台水平的核心要素，目前同时具备 2D 标注、3D 点云标注、2D-3D 联合标注以及 3D 语义分割标注的供应商比较有限，能以最快速度覆盖更多功能需求的数据服务商将能更好掌握智能驾驶数据市场的主动权以及议价能力。 第二个核心要素是算法能力。平台的智能化程度越高，对人的依赖程度越低，在提高平台的生产效率的同时可以大幅降低生产成本。 第三个要素是数据安全能力。智能驾驶数据不同于传统的语音类数据，由于其采集图像涉及大量的地理及个人隐私信息，为更好防范数据安全风险，国家近年密集出台相关法律法规，要求数据流转链条上各类企业必须做好充分的数据安全保障。去年 8 月底，自然资源部发布《关于促进智能网联汽车发展维护测绘地理信息安全的通知》，《通知》明确说明将对数据服务全链条进行监管，包括采集、标注处理等在内的各类业务形态均被纳入监管范畴，且明确规定内资

	企业需获得测绘资质才能从事测绘相关活动（外商投资企业则不能申请测绘资质）。可以看出数据安全的重要性更加凸显，未来不具备相关数据安全能力的供应商将逐渐被市场淘汰。 8、智能驾驶的舱外数据的来源是什么？ 总体来讲，舱外数据来源可以分为两类：一部分是由客户提供的真实路采数据；另一方面，数据服务商在获得测绘资质并形成路采能力后，其自身可以通过自主设计、搭建路采方案，并进行上路采集。智能驾驶车外业务领域的数据采集难度相对较高，而海天瑞声多年来积累的丰富的项目管理经验以及在供应链资源都能够起到积极作用。 9、未来智能驾驶会不会有相关的数据产品？ 基于目前智能驾驶的技术发展态势，各个客户的技术方案多有不同，例如技术路线、车型、传感器选型、部署位置、数据处理的需求等等都有各自的要求，因此现阶段智能驾驶训练数据需求仍以定制化需求为主。 随着去年测绘资质的获得，公司拥有了上路采集数据的准入资格，也使得生产自有产权舱外数据集产品成为可能，具备了产品化开发的基础条件。未来，公司将持续洞察市场共性需求，择机进行智能驾驶相关产品的开发。 10、公司智能驾驶客户有哪些？ 受益于智能驾驶业务蓬勃发展以及公司在该领域的强力布局，截至目前，公司已服务超过 40 家智能驾驶领域客户，覆盖传统车企、新势力车企、智能驾驶技术公司等。目前公司也在该领域进行持续的客户拓展，进一步加固客户储备，迎接行业爆发。

	11、智能驾驶业务毛利展望如何？ 目前，公司智能驾驶数据业务以舱外的视觉需求为主，主要向客户提供定制类标注服务，整体业务毛利水平维持在20%-30%左右，部分项目的毛利水平可以达到40%左右。未来，公司一方面会基于新获得的乙级测绘资质进一步拓展采标一体的综合数据解决方案，通过设计、采集、加工等环节的综合服务，进一步提高智能驾驶业务的溢价能力和毛利水平；另一方面，公司正在强力布局算法中台，目前已研发完成的第四代智能驾驶标注平台已在十种算法框架上进行模型拓展，可服务于车道线、车前障碍物、视频跟踪目标等各场景标注需求，在公开测试集上的准确率也达到了较高水平。未来公司将通过算法能力搭建，大幅降低标注成本，预计可为智能驾驶业务带来更为可观的毛利提升。 12、了解到海天的数据产品已经入驻北数所，并开始了实际的数据交易，想请问公司入驻数据交易所对公司数据交易带来哪些变化?除了交易，公司是否可以参与到交易流通的其他业务环节？ 2021年3月北数所成立之初，海天就受邀加入了其牵头成立的北京国际数据交易联盟，并在2021年9月至10月上线了若干款数据产品。北数所的数据交易平台，为海天等数据服务商搭建和扩充了数据交易渠道，通过平台实现点对面，极大的扩充了数据产品的辐射范围。相信未来，随着交易平台的逐渐完善和影响范围的持续扩大，通过北数所数据交易平台的买家将陆续增多，公司将获得更高的客流量入口。 此外，随着国家近年来对数字经济的重视，国家层面大力培育和发展数据要素市场，北京、上海、深

	圳等地陆续成立了数据交易所，促进数据要素交易和流通。海天瑞声也积极加入了北京、上海数据交易所，成为首批数据服务商。未来海天瑞声也将依托国家政策和各领域平台建设，积极探索拓展服务边界，在数据交易流通中更多环节发挥更大的作用和价值。 13、海外业务的营收增长点在哪？ 公司境外收入的主要增长点来自于多语种相关的智能语音以及计算机视觉类业务。 随着AI在全球的快速发展，海外越来越多的科技企业以及互联网企业正在加速进行全球化扩张，为更好实施其发展战略，已释放出快速增长的多语种数据需求，例如将全球化扩张作为其收入增长核心动力的海外科技互联网企业、将AI及元宇宙作为其重点发展方向的大型科技企业等，都已释放出大量的多语种语音需求以及多语种OCR需求。 海天也已在多语种方面加快布局，通过规模化的多语种产品研发投入，精准对接海外客户需求。此外，为进一步撬动更大的境外市场需求，公司将增设海外本土销售团队，并通过多维营销方式增强海外客户触达，提升客户服务体验，力争实现海外市场收入在未来保持良好增长态势。 14、境外业务的毛利率为什么会比境内业务高？ 首先，公司境外业务当中标准化数据集产品的销售占比相对更高一些，而标准化产品的销售毛利率为100%，远大于定制服务毛利水平。此外，相比于境内客户，境外客户更认同数据服务商的综合能力及品牌价值、价格敏感度相对较低。以上两个因素综合导致境外业务较高的毛利水平。 15、数据行业的竞争态势如何？未来市场份额是

	越来越集中么？ 目前来看，数据服务市场主要由品牌数据服务商、客户自建团队以及一些中小数据服务商构成。 未来，公司预判整个数据服务市场将进行重新洗牌，集中度将进一步提升。市场各类主体将会通过在技术研发投入、资源能力建设等主要方面的竞争，逐步淘汰掉那些研发能力弱、资源势力差的品牌服务商和中小玩家。此外，国家对于数据安全及合规要求的进一步趋严，会将那些不具备数据安全合规能力或尚未进行此方面布局的企业逐渐淘汰出局。 在客户自建团队部分，出于其自身对数据和业务的敏感性、保密性需求，可能会与品牌服务商长期共存。 16、公司采集业务往往涉及大量终端人，请问公司是否需要获得终端人的授权？ 是的，对于业务中所采集的终端人个人信息，我们按照《个人信息保护法》《数据安全法》等法律要求，依法依规进行采集。法律要求获得授权同意的，我们会事先取得合法有效的授权，以此来保护其个人信息安全及相关合法权益。因此，公司在开展涉及个人信息采集的业务时，会根据所适用的法律要求，并结合项目具体情况，事先准备好授权文件，供终端人了解项目情况及其所享有的权利，终端人了解了授权文件的内容、同意作出授权并签署授权文件后，公司才会开始相关采集作业。 17、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。 18、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？客户会如何选择？未来的发展趋势如何？ 区别：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。 产品+服务的组合一直是公司向市场提供的综合解

	决方案，是一个整体，服务于不同客户的不同研发阶段需求，其收入贡献比例在各年间也呈现较为一致的趋势。而产品+服务带来的数据积累，也哺育了公司的数据处理平台和相关算法不断提升，努力达到数据处理场景下的行业最优。 未来，如果把垂直行业数据这个大领域放进来考虑，那么先期，更高要求的定制化服务业务的占比可能会逐渐上升，以智能驾驶为例，客户对于数据服务商的主流需求其实是一体化、闭环式的数据解决方案，这就需要类似于海天瑞声这样的数据服务商有能力为客户提供从数据采集、处理到训练、仿真、测试、验证的完整闭环服务，以满足客户的数据处理量更大、数据处理的迭代频次更高等需求特点。但在定制化服务提供过程中，公司将发挥在语音领域一样的特点，提取标准化需求，在垂直行业领域也构建建设自身产品体系的能力，形成有价值的行业数据集。 19、请问公司在训练数据领域具体有哪些竞争优势？ 经过多年发展与积累，公司逐步构建起了在行业内的竞争壁垒，核心竞争力主要体现在： （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。截至目前，公司已积累超过1,050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。公司全方位做好数据风险管控工作，通过了业内重要的 ISO/IEC 27001 体系认证、ISO27701 个人隐私信息安全管理体系认证，形成了具有自身特色的数据安全与隐私保护整体解决方案。目前，公司符合 GDPR、《数据安全法》、《个人信息保护法》等一系列国际通用与国内法律法规的管理规范要求，获得了业务领域合作客户的高度认可。 20、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。 21、公司毛利率为什么比较高？ 近年来，公司毛利始终维持在 64%-68%区间，高毛利主要受益于公司收入结构中产品占比较高，产品为一次开发可多次销售的数据集，产品的毛利为 100%，因此产品的销售是公司毛利的重要来源。此外，公司拥有大量的海外客户，相比于境内客户，境

	外客户更认同数据服务商的综合能力及品牌价值、价格敏感度相对较低，因此境外定制服务的毛利相对较高。算法能力和平台能力也是高毛利的重要来源，公司历来重视技术研发，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，提升数据标注效率，降低成本，提升毛利。 22、语言学研究的具体作用和价值是什么？目前公司语言学研究的最新进展是什么？ 语音语言学领域的专业知识是构建高质量语音识别算法和语音合成算法的关键要素。以语音合成为例，在语音合成系统中，发音词典提供了从单词到音素之间的映射关系，将语言模型建模单位解构为声学模型的建模单元，为后续合成发音奠定基础。语音合成系统接收到文本信息后，首先运用发音词典对其进行语言处理、韵律处理，将文本（单词、字符等）转换并解构为一系列对应的发音符号（类似于国际音标）；随后，系统中的语音合成器接收到前述发音符号，运用语音库合成转换为语音对外输出，最终实现文本到语音的语音合成过程。可见，高质量的发音词典在语音合成系统中具备重要作用。由上述示例可以看出，要获得高准确率的语音合成算法模型，就要求智能语音训练数据结构中包含高质量的发音词典。要在大词汇量的连续语音交互中正确、合理运用智能语音相关的语言模型、语法及词法模型，则必须有效地运用计算语言学方面的基础知识和研究成果。语音语言学领域的基础研究成果和专业知识构建了发音规则、发音词典的形成基础，进而为构建高准确率的语

	音识别、合成训练数据提供了条件。 公司在语音语言学基础研究方面有深厚积累：公司建立了成熟的发音词典构建流程、积累了深厚的语音语言学基础研究成果。截至目前，公司的产品/服务已覆盖 190 个语种/方言，已积累下超过 100 个语种/方言的发音词典，累计词条数超过 1,000 万条，可构建高质量的智能语音训练数据。 23、公司未来是否会做 SaaS 服务？ 目前，公司的商业模式还是集中在向客户群体提供数据产品、服务。随着公司自研数据处理平台以及算法能力的日趋成熟、完善，公司不排除未来会进行包括 SaaS 服务模式在内的新型业务模式的推广，我们会根据市场的实际需求及行业变化动态优化、扩展商业模式，以保障公司业务持续向前发展。 24、客户是否会自建数据团队？ 客户自建团队这种现象对于公司来讲并不陌生，有一些客户通过自建团队主要解决其自身的部分敏感数据需求，但受专业化分工的影响，客户仍然会大量购买数据服务提供商的数据，以充实其算法模型训练的规模性需求。相较于客户自建团队，海天瑞声历来都是对接众多大型科技公司、头部人工智能企业、科研院所等，获得的信息是广泛的，项目经验丰富，同时积累了大量的 know-how，对数据的理解更广、更深刻，同时我们搭建了成熟的数据处理算法平台，通过更高效的人机交互实现降本增效，保证数据质量的同时能有效降低成本，为客户提供更高性价比的训练数据产品/服务。 25、科大讯飞是公司的竞争对手吗？ 科大讯飞是公司多年来的优质客户，公司给科大

	讯飞提供的产品及服务主要集中在智能语音领域，包括多语种的语音识别数据集产品或语音识别数据定制服务。
附件清单（如有）	
日期	2023 年 2 月 28 日