证券简称：海天瑞声　　证券代码：688787

北京海天瑞声科技股份有限公司

（Beijing Haitian Ruisheng Science Technology Ltd.）

（北京市海淀区成府路 28 号 4-801）

海　天　瑞　声
DATAOCEAN AI

关于本次募集资金投向属于科技创新领域的说明

二〇二三年六月

北京海天瑞声科技股份有限公司

关于本次募集资金投向属于科技创新领域的说明

北京海天瑞声科技股份有限公司（以下简称“海天瑞声”或“公司”）根据《上市公司证券发行注册管理办法》（以下简称“《注册管理办法》”）等有关规定，结合公司 2023 年度向特定对象发行 A 股股票（以下简称“本次发行”）方案及实际情况，对本次发行募集资金投向是否属于科技创新领域进行了研究，制定了《北京海天瑞声科技股份有限公司关于本次募集资金投向属于科技创新领域的说明》（以下简称“本说明”），具体内容如下：

一、公司的主营业务

海天瑞声是一家从事 AI 训练数据的研发设计、生产及销售业务的公司，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。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。经过多年发展，海天瑞声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。

二、本次募集资金投向

本次向特定对象发行股票募集资金总额不超过 78,989.00 万元（含本数），扣除相关发行费用后的募集资金净额拟用于以下项目：

序号	项目名称	项目投资总额（万元）	拟投入募集资金额（万元）
1	AI 大模型训练数据集建设项目	38,337.36	38,337.36
2	数据生产垂直大模型研发项目	40,651.64	40,651.64
合计		**78,989.00**	**78,989.00**

注：项目名称最终以主管部门核准或备案名称为准

（一）AI 大模型训练数据集建设项目

本项目的实施主体为北京海天瑞声科技股份有限公司及或下属子公司。鉴于大模型训练数据通常具备数据规模大、数据质量高、数据类型丰富等特点，本项目拟建设 AI 大模型训练数据集，即生产用于通用型、及各种垂直领域大模型训练的海量、高品质数据集。本项目拟购置办公楼作为建设大模型训练数据研发生产基地，并购置数据采集、数据处理、数据存储和办公等软硬件设备，利用海量、高质量、多样化的公共数据资源、社会数据资源和稀缺性数据源，通过数据集设计、数据采集/获取、清洗/分类/标准化、标注/优化、评测等全流程的任务执行进行高质量大模型训练数据集建设。

本项目将充分利用“先行先试示范区”在基础制度、数据供给等方面的先行先试政策，采用多元化的方式获取大规模原始数据；利用工程化的数据处理技术进行预训练阶段的数据清洗；采用人类反馈强化学习模式，基于微调和奖励模型训练的方法，以人类撰写少量的典型问题和标准答案与深度学习阶段基础性标注相结合的模式，生产出市场适用性较强的大模型训练数据集。

本项目建成后，将提供可供大模型训练和评测的不少于 10 个品类的专业数据集，显著提升行业内面向大模型训练数据集的类别和质量，协助实现公共数据、社会数据等各类高价值数据资源汇聚，实现基于大模型通用能力和垂直领域数据的训练学习。本项目的数据集产品具体可分为三大类：

①通用及特定垂直领域的大语言模型训练数据集，包括但不限于：

A、中文大模型预训练语料数据集（含通用场景、特定场景、对话场景、指令集等）；

B、多语言大模型预训练语料数据集（含通用场景、对话场景、指令集等）。

②多模态大模型训练数据集：可应用于多语言图文大模型训练、多模态数字人训练、多语种语音大模型训练、全场景自动驾驶大模型训练等场景的跨模态数据集。

③大模型评测数据集：可应用于大模型的能力、任务、指标等方面的评测。

（二）数据生产垂直大模型研发项目

本项目建设目标为通过大模型基础研究，研发海天瑞声数据生产垂直大模型，并以海天瑞声数据生产垂直大模型为核心，升级海天瑞声一体化技术

支撑平台。本项目的实施主体为北京海天瑞声科技股份有限公司及/或下属子公司。

为应对大模型时代下数据规模量极大、复杂性和多样性高，数据服务规则设计难度指数级提升等诸多问题，且为更高效高质完成数据规则的规模化生产，公司将采用全栈自研的数据生产垂直大模型技术，辅助完成面向多个下游任务的数据设计与处理规则。同时，为更好实现数据生产垂直大模型的生成能力，公司将研发并引入多项新兴技术，夯实数据生产垂直大模型构建的基础。

此外，基于大模型的核心能力，项目还将升级海天瑞声一体化技术支撑平台，使其能够全面拥有大模型范式下的数据服务能力。通过嵌入预训练数据下载工具、预训练数据清洗工具、指令数据集筛选工具、指令数据集生成与调优工具、大模型评测数据集评测工具、大模型评测数据集质检工具、多模态数据集生产工具等模块，完成大模型的数据获取与处理工作，打造模型训练、模型评测的能力。

图：海天瑞声新一代基于数据生产垂直大模型的数据服务技术架构图

三、本次募集资金投向属于科技创新领域

（一）本次募集资金投向符合国家产业政策，主要投向科技创新领域

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。本次募集资金总额全部用于

AI大模型训练数据集生产和数据生产垂直大模型的研发，系围绕公司主营业务展开。根据国家统计局《战略性新兴产业分类（2018）》，公司所从事的训练数据生产业务属于“新一代信息技术产业—新兴软件和新型信息技术服务—新型信息技术服务—信息处理和存储支持服务—数据加工处理服务”行业，是国家重点支持的“新一代信息技术领域”的战略性新兴产业。因此，本次募集资金投资项目投向属于科技创新领域。

（二）本次募集资金投资项目将进一步提升公司科技创新水平

在人工智能产业进入以大模型为代表的新的发展时期，通过本次募投项目的实施，公司将建设一批适用性较强的大模型训练数据集，拓展潜在高增长价值的新型业务板块，并藉此进一步扩大公司业务规模；同时，以研发海天瑞声数据生产垂直大模型为核心，升级海天瑞声一体化技术支撑平台，研发并引入多项新兴技术，促进公司科技创新水平的不断提升，巩固公司的核心技术壁垒，构建长期技术实力支撑，从而进一步增强公司核心竞争力。

四、结论

综上所述，公司认为：公司本次发行募集资金投向属于科技创新领域，符合未来公司整体发展方向，有助于提高公司科技创新能力，强化公司科创属性，符合《上市公司证券发行注册管理办法》等有关规定的要求。

北京海天瑞声科技股份有限公司

董事会

2023 年 6 月 21 日