股票简称：海天瑞声　　股票代码：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

2023 年度向特定对象发行 A 股股票

募集资金使用可行性分析报告（修订稿）

二〇二三年十月

释　义

除非文义另有所指，下列简称具有如下含义：

专业术语		
人工智能/AI	指	英文为 Artificial Intelligence，缩写为 AI，是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门技术科学。
生成式人工智能/生成式 AI	指	生成式人工智能（Generative AI）是使用机器学习和深度学习算法，从现有数据中学习并生成新内容的一种人工智能技术。它可以生成各种形式的数据，如图像、音频、视频和文本，并在虚拟现实、自然语言处理、艺术创作和模拟实验等领域有广泛应用。
大模型	指	大模型就是 Foundation Model（基础模型），指通过在大规模宽泛的数据上进行训练后能适应一系列下游任务的模型。
数据生产垂直大模型	指	海天瑞声数据生产垂直大模型，是公司利用大模型理论和算法，自主打造的面向人工智能数据基础服务的技术底座，实现面向多个下游数据任务的设计与处理等规则及内容的多项生成能力模型。
多模态	指	多模态机器学习（Multi-modal Machine Learning）是一种机器学习方法，用于处理多种类型的数据，如图像、文本、音频等。它通过将来自不同模态的数据进行整合和联合建模，以提取和学习不同模态之间的相关信息和特征。多模态数据是指用于多模态机器学习训练使用的跨形态数据类型。
深度学习	指	英文为 Deep Learning，缩写为 DL，是从机器学习中的人工神经网络发展而来的多层人工神经网络和它的训练方法是一种基于人工神经网络的机器学习，其中多层处理被用来从数据中逐步提取更高层次的特征。
人类反馈强化学习	指	英文为 Reinforcement Learning from Human Feedback，缩写为 RLHF，是一种机器学习方法，旨在使智能系统从环境中学习，通过引入“奖励”和“惩罚”信号，让系统自行探索环境并学习最佳行为策略，以最大化某种特定目标。
人工智能基础数据、训练数据、人工智能基础数据服务	指	人工智能基础数据、训练数据均指通过采集与处理等步骤形成的、结构化的、可供人工智能算法模型训练使用的数据；人工智能基础数据服务指为 AI 算法训练及优化提供的数据集设计、数据采集/获取、清洗、标注/优化、评测等服务。
计算机视觉	指	英文为 Computer Vision，缩写为 CV，是一门研究如何使机器“看”的科学，是指用摄影机和电脑代替人眼对目标进行识别、跟踪和测量等。
自然语言处理	指	英文为 Natural Language Processing，缩写为 NLP，是计算机科学领域与人工智能领域中的一个重要方向，研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。

AIGC	指	AI Generated Content，是指利用人工智能技术来生成内容。
ChatGPT	指	美国 OpenAI 研发的聊天机器人程序，ChatGPT 是人工智能技术驱动的自然语言处理工具，它能够通过理解和学习人类的语言来进行对话，还能根据聊天的上下文进行互动，真正像人类一样来聊天交流，甚至能完成撰写邮件、视频脚本、文案、翻译、代码，写论文等任务。
GB	指	千兆字节，是数字信息存储容量的一个单位。它被用来衡量文件、文档和数据存储设备的大小。一个千兆字节是 1024 兆字节。
TB	指	太字节，是数字信息存储容量的一个单位。一个太字节是 1024 个千兆字节。
ZB	指	Zettabyte，是数字信息存储容量的一个单位。一个 Zetta 字节是 100 万太字节。它代表了一个巨大的数据存储容量，通常在讨论大规模数据中心、全球数据流量或主要组织或行业的存储需求时使用。
Transformer 技术	指	Transformer 技术是一种基于自注意力机制的神经网络架构，广泛应用于自然语言处理和计算机视觉等领域。Transformer 技术在机器翻译、文本生成、语言理解等任务中取得了显著的成果，并成为目前自然语言处理领域的主流模型。
OpenAI	指	OpenAI 是一家美国人工智能研究实验室和公司，在人工智能的各个领域进行前沿研究。
OCR	指	是指电子设备（例如扫描仪或数码相机）检查纸上打印的字符，通过检测暗、亮的模式确定其形状，然后用字符识别方法将形状翻译成计算机文字的过程。
Prompt		提示文本，提供模型完成下游任务所需的上下文等信息，以帮助模型更好地理解任务需求，从而更好地完成下游任务。
预训练/Pre-training	指	预训练是通过在大规模数据集上进行初始训练，学习通用特征和表示的机器学习方法。预训练模型可用于后续任务，并在多个领域应用广泛。
模型微调/Fine-tune	指	微调，是一种常用于预训练模型的模型优化技术，基于已经训练好的预训练模型进行微调，使模型能够更好地适应目标任务。
奖励模型/Reward Model	指	奖励模型（Reward Model）是在强化学习中使用的一种关键概念。它定义了在特定任务中，智能体所追求的目标和奖励的方式。奖励模型通过指定在智能体采取行动后，所获得的奖励信号来引导智能体的学习和决策过程。奖励模型可以基于任务的目标设定，对智能体的行为进行正向奖励或负向惩罚，从而影响智能体的策略优化和决策选择。
智算中心	指	全称人工智能计算中心，指为人工智能（AI）开发和应用提供所需算力服务、数据服务和算法服务的机构。
Stable Diffusion	指	由 Stability AI 公司（英国一家人工智能企业）于 2022 年推出的图文生成模型，用于从提示信息的自然语言描述中生成图像。

一、本次募集资金使用计划

北京海天瑞声科技股份有限公司（以下简称“公司”）2023 年度拟向特定对象发行 A 股股票（以下简称“本次发行”），募集资金总额不超过人民币 66,600.00 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	35,507.25
2	数据生产垂直大模型研发项目	40,651.64	31,092.75
合计		**78,989.00**	**66,600.00**

在本次发行募集资金到位前，公司将根据募集资金投资项目的实际情况，以自筹资金先行投入，并在募集资金到位后按照相关法律、法规规定的程序予以置换。募集资金到位后，若扣除发行费用后的实际募集资金净额少于拟投入募集资金总额，在本次发行募集资金投资项目范围内，公司将根据实际募集资金数额，按照项目的轻重缓急等情况，调整并决定募集资金的具体投资项目、优先顺序及各项目的具体投资金额，募集资金不足部分由公司自筹解决 。

若本次发行募集资金总额因监管政策变化或发行注册文件的要求予以调整的，则届时将相应调整。

二、募集资金投资项目基本情况及可行性分析

（一）AI 大模型训练数据集建设项目

1、项目背景

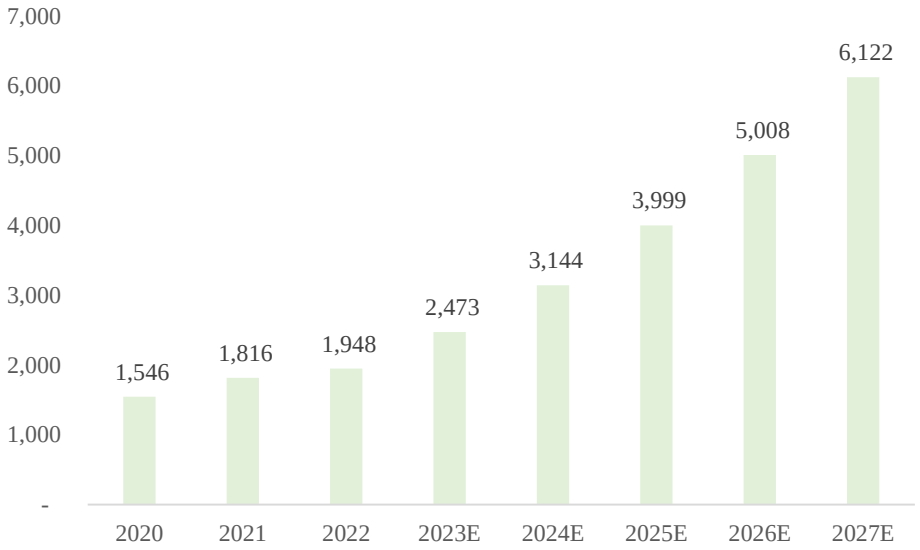
（1）人工智能产业加速发展，基础数据服务业规模不断提升

随着互联网、云计算、物联网以及穿戴设备的发展，数据要素价值不断释放，数字化转型成为大势所趋。人工智能作为数字经济发展的底层核心技术之一，成为数字经济发展的重要战略抓手。2022 年底 ChatGPT 的出现，掀起又一波人工智能发展热潮，以无监督学习模式为代表的预训练+人类反馈强化学习所构成的大模型技术路线的落地意味着人工智能开启发展新范式，基础模型能力通过预训

练及基于人类反馈的强化学习得到不断解锁，以解决海量开放式任务。

根据艾瑞咨询数据，2022 年中国人工智能产业规模为 1,948 亿元，预计 2027 年市场规模将达到 6,122 亿元，年复合增长率为 25.6%，主要与智算中心建设以及大模型训练等需求拉动的 AI 芯片市场、无接触服务需求拉动的智能机器人及对话式 AI 市场等快速增长相关。

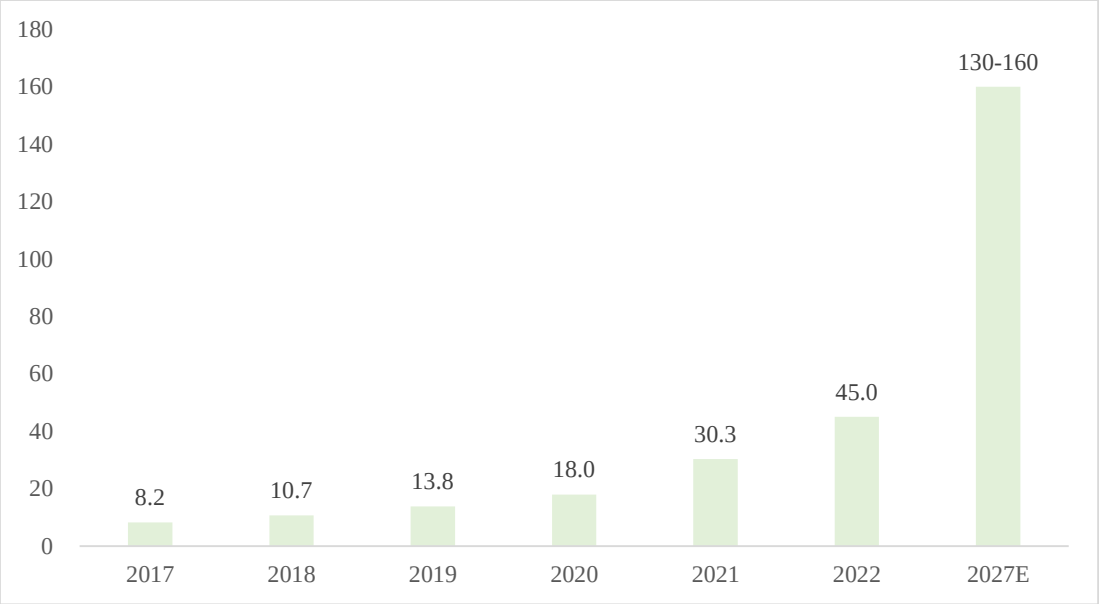
图 1　2020-2027 年中国人工智能产业规划（单位：亿元）



数据来源：艾瑞咨询

在人工智能产业链中，算法、算力和数据共同构成技术发展的三大核心要素。过去十年，人工智能产业以算法为中心，随着算法趋于开源，数据的重要性愈发凸显。在人工智能模型从技术理论到应用落地的过程中，需要依赖大量的训练数据，相较于以模型为中心的训练方法，以海量数据为中心的训练方法能够提升模型推断结论的可靠性。万亿 GB 量级的数据随着互联网、云计算、物联网、大数据等发展源源不断地产生，但数据质量参差不齐，对海量的复杂数据进行深入挖掘、输出，进而激活和释放数据的深层价值也成为数据市场的发展重点。根据德勤数据，2022 年中国人工智能基础数据服务市场规模为 45 亿元，2027 年规模将达到 130-160 亿元，年复合增长率为 23.6%-28.9%。

图 2　2017-2027 年中国人工智能基础数据服务市场规模（单位：亿元）



数据来源：德勤

（2）大模型技术发展带来 AI 范式变革，催生新型数据服务需求，进一步提升数据市场空间

人工智能大模型由于其强大的通用性能以及泛化能力正在加快人工智能发展路径，在大幅增强人工智能体验感的同时降低再开发门槛，使得人工智能产业具备在各实体产业快速落地发展的潜能。

随着大模型技术的发展，算法训练对数据的依赖程度逐渐加深，催生了新的数据需求和新的数据服务模式。一方面，数据的质量以及数据清洗的工程化能力会显著拉开大模型预训练阶段的效果差距；另一方面，预期更多模型将采用类强化学习模式来进行特定领域或特定方向上的优化迭代，以使得机器能够以更加接近于人类期望的方式提供答案输出。未来数据处理将不再局限在传统的有监督学习下的定向采集与精细化标注，而将叠加数据规模化获取、清洗以及类强化学习等方向。

此外，随着深度学习技术的不断突破，人工智能发展已经进入 2.0 时代，AI 技术与传统产业的融合将成为数字经济时代的新发展趋势。大模型算法训练需求正逐渐从通用基础能力建设向垂直领域拓展，数据需求向专业化方向发展。目前，AI 技术在金融、医疗、工业等传统行业中的渗透率和应用不断提升，展现出可观的商业价值和较强的发展潜力。为加速实现 AI 产业化落地，行业将衍生出更多垂直场景的数据需求，大模型将通过不断学习各个专业领域的行业高质量数据，

实现更广阔的垂向拓展。

与此同时，国内科技互联网巨头纷纷布局多模态大模型，基于公开数据及自身特有数据训练多模态大模型，多模态数据集需求快速提升，多模态成为大模型时代下新发展范式。

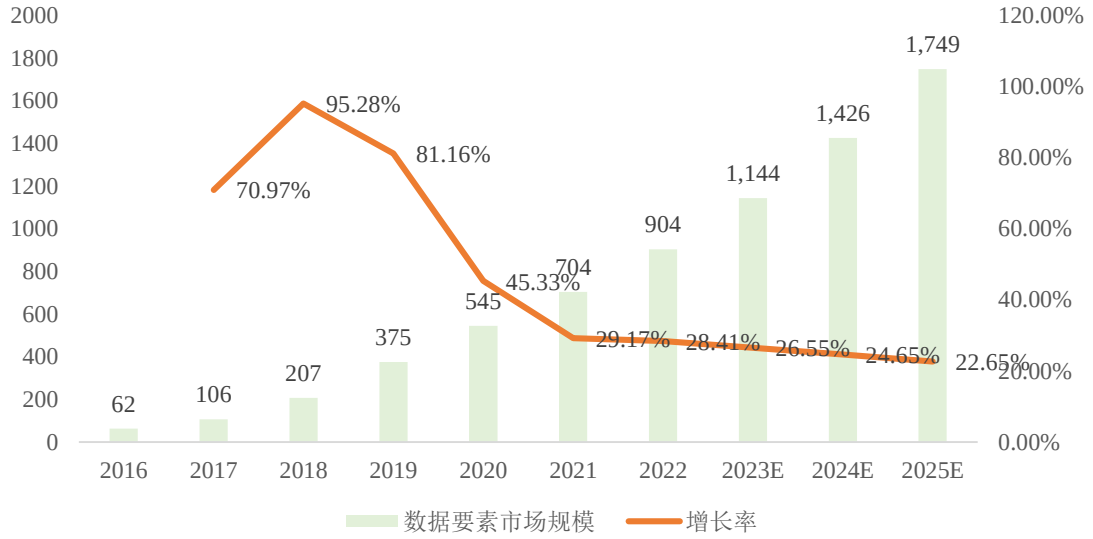
（3）各地推动数据基础制度建设，数据要素市场迎来新发展机遇

近年来，我国数字经济蓬勃发展，数据要素因具有基础性战略资源和关键性生产要素的双重属性，相关市场规模持续增长。尤其在《中共中央、国务院关于构建数据基础制度更好发挥数据要素作用的意见》出台后，我国系统性布局了数据基础制度体系的“四梁八柱”，加速了数据流通交易和数据要素市场发展，进一步推动了公共数据、企业数据、个人数据合规高效流通使用。为更好响应中央号召，北京、上海、广州、深圳、杭州等地数据政策陆续出台，逐步构建了多层次、多元化数据要素市场生态体系。

以北京为例，《北京市促进通用人工智能创新发展的若干措施》和《关于推进北京市数据专区建设的指导意见》指出，北京市要加快建设“数据基础制度先行先试示范区”（以下简称“先行先试示范区”），探索打造数据训练基地，归集高质量基础训练数据集，推动数据要素高水平开放，提升本市人工智能数据标注库规模和质量，并建设针对重大领域、重点区域或特定场景建设专题数据区域，吸纳市场主体和数据、技术、资本等多元要素参与。北京市陆续出台的多项文件旨在打破数据壁垒，推动数据融合利用，加快推动公共数据开放，促进数据要素流通，激发数字市场创新活力，释放和发展数字化生产力，打造多层级数据要素市场，成为具有竞争力和影响力的数字产业集群。按照“政府引导、市场运作、创新引领、安全可控”的原则，“先行先试示范区”有望成为国际领先的数据要素高效流通核心枢纽。

根据国家工信安全发展研究中心数据，2022 年我国数据要素市场规模为 904 亿元，预计到 2025 年将达到 1,749 亿元左右，2020 年-2025 年年复合增长率为 26.26%，数据要素将成为赋能中国数字经济发展的重要驱动力量。

图 3 中国数据要素市场规模及预测（亿元）

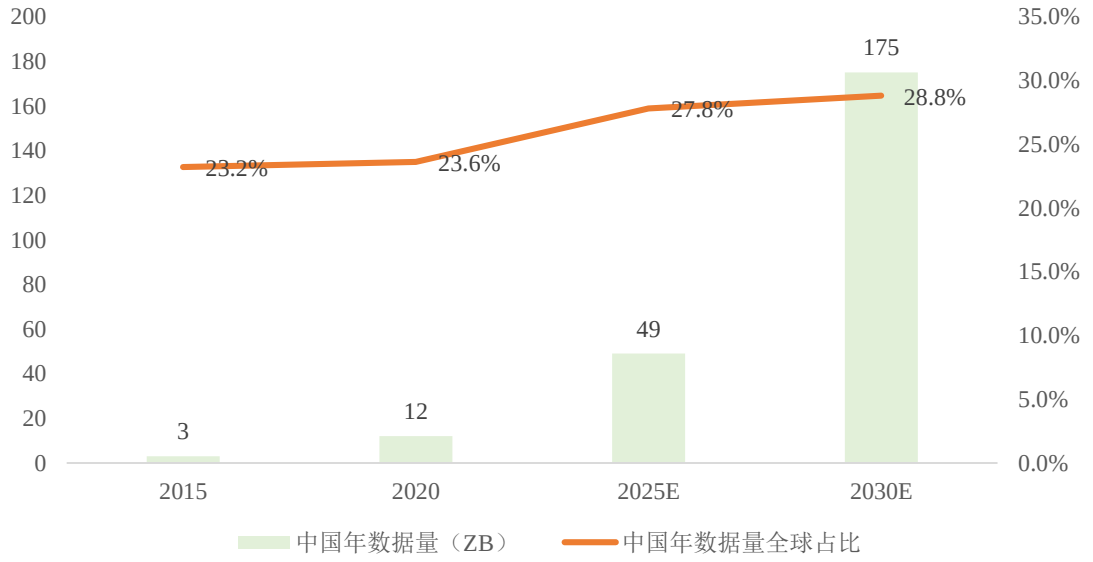


数据来源：国家工业信息安全发展研究中心，中国信息通信研究院，华泰研究

（4）我国拥有海量数据资源，但数据质量仍面临严峻挑战，成为行业亟待解决的问题

我国各行业数据资源较为丰富，根据艾瑞咨询数据，2015 年-2030 年中国数据量规模由 3ZB 将增长至 175ZB，预计 2030 年中国数据量约占全球的 28.8%，年复合增长率约为 31%。

图 4　2015-2030 年中国数据量规模及全球占比



数据来源：艾瑞咨询

虽然中国数据资源丰富，但由于数据挖掘不足，以及大量数据无法在市场上

自由流通等原因，优质中文数据集仍然稀缺。以 ChatGPT 为例，其模型训练数据中，中文数据来源不足千分之一。目前，国内头部科技企业主要基于公开数据集以及自身特有的数据进行大模型训练，但由于中文优质数据质量以及数据资源的制约，国内大模型的能力与以 ChatGPT 为代表的国际大模型相比仍存在一定差距。

国内缺乏高质量数据集的主要原因包括当前国内数据挖掘和数据治理的力度不足、资金投入较大；数据流通与数据安全保障措施不够健全；国内市场缺乏开源意识，大量数据无法在市场上自由流通；国内相关公司成立较晚，数据积累较少；学术领域中文数据集受重视程度低以及国产数据集市场影响力及普及度较低等。从原始数据到可被应用的数据集产品，需要经历数据集结构设计、数据获取、数据处理（包括数据清洗、数据标注/优化等）等过程，以形成可供使用的优质数据集，国内数据服务市场的发展有助于缓解中文数据集数量不足和质量欠佳等问题。

2、项目基本情况

大模型训练数据具备如下三个特点，具体而言：一是数据规模大，根据 DeepMind 论文《Training Compute-Optimal Large Language Models》，模型参数规模预训练数据的 Token 数最佳比例在 1:20，要充分训练一个千亿规模的模型，至少需要 TB 级的训练数据；二是数据质量高，在模型训练之前，需要依赖专业团队对数据进行清洗等预处理，防止数据中的噪声对模型的训练产生不良影响，在一些特定的任务中，还需根据不同目的对模型训练数据进行过滤；三是数据类型丰富，多领域的数据是大模型具备通用 AI 能力的关键，需从不同渠道收集各种训练数据，包括各类垂直领域数据、多语言数据、翻译类平行语料、多轮对话数据、代码库和题库等。

基于以上特点，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、多样化的个人数据资源、企业数据资源、公共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“数据基础制度先行先试”区域在基础制度、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，协助实现个人数据、企业数据、公共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：①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）；②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。

3、项目建设必要性

（1）本项目建设是响应国家建立数据基础制度，落实北京建设“先行先试示范区”的必然选择

党的十八大以来，习近平总书记屡次强调建设数字中国以及构建数据要素的重要性，并明确指出数据是新的生产要素，是基础性资源和战略性资源，也是重要生产力。为进一步推动国家数字经济发展，发挥数据要素在经济发展中的重要价值，我国推出《中共中央、国务院关于构建数据基础制度更好发挥数据要素作用的意见》，从顶层设计角度，在数据产权、流通交易、收益分配、安全治理等方面构建了数据发展的基础制度和规划纲要，以促进数据合规高效流通使用，充分发挥中国海量数据规模和丰富应用场景优势，赋能实体经济，激活数据要素的潜能。

北京市则率先开展国家数据基础制度“先行先试示范区”建设， 2023 年 5 月

发布的《北京市促进通用人工智能创新发展的若干措施》指出，充分发挥政府引导作用和创新平台催化作用，整合创新资源，加强要素配置，营造创新生态，提升高质量数据要素供给能力，归集高质量基础训练数据集，2023 年 7 月发布的《关于更好发挥数据要素作用进一步加快发展数字经济的实施意见》提出，“发展数据生产服务业，支持企业开展数据采集、清洗加工、存储计算、数据分析、数据标注、数据训练等数据生产服务，支持企业研发建设数据生产线，推进数据生产自动化”。公司作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，有义务和责任积极响应北京建设“先行先试示范区”的号召，通过本项目的实施有效助力数据要素市场培育，推动数字经济创新发展，为北京市加快建设全球数字经济标杆城市提供助力。

（2）本项目建设是践行国家规范生成式人工智能产品要求的重要举措

生成式人工智能产品因其复杂性可能带来社会风险、技术伦理风险、企业商业秘密和个人信息泄露风险、虚假信息风险、知识产权侵权风险及其他潜在风险。为了更好地促进生成式人工智能技术健康发展和规范应用，国家网信办等七部门于 2023 年 7 月出台了《生成式人工智能服务管理暂行办法》，该办法从内容合规、数据来源合法性、知识产权及商业秘密保护、虚假信息防范等方面，对生成式人工智能产品提出了全方位的合规要求。该办法明确提出，“提供者应当依法开展预训练、优化训练等训练数据处理活动”、“采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性”。

根据前述规定，数据获取、数据处理的高标准意味着数据获取难度及处理成本将大幅增加，以预训练阶段为例，由于大量数据来源应合法合规，需投入大量成本完成数据获取。因此，出于成本与数据集质量的平衡性考量，在大模型训练中，大模型厂商通常会选择与专业的第三方数据集厂商合作，由专业第三方提供的合规、高质量数据集或相关解决方案将成为践行国家规范生成式人工智能产品要求的重要举措。

（3）本项目建设是支撑大模型训练，提升大模型输出能力的有效方式

随着人工智能应用场景日益丰富、产品智能化要求的不断提升，数据需求逐渐向海量、高质量、多元化方向演进。从自然数据源简单收集、获取的数据资源，

通常无法直接满足大模型的训练需求，需经专业化的数据分类设计、清洗、加工处理，形成相应的工程化数据，以供大模型训练使用。一般而言，符合大模型训练标准的数据需具备质量高、规模大、样本丰富等三个特点。首先，海量具有无毒害性、公平性等高质量特征的数据集能够提高模型效果（例如，精度与可解释性），并且减少收敛到最优解的时间；其次，在强化学习阶段，原始数据由于存在信息量低、含有噪声或需补齐等问题，使用前需要进行数据对齐等诸多微调操作，优秀的指令数据集能够帮助大模型更好的泛化适配更多下游任务。再次，数据丰富程度能够显著提高大模型的泛化能力，减少过拟合情况的发生，达到更优的模型效果。

当前国内数据资源虽然丰富，但优质的中文大模型训练数据仍然稀缺，中文大模型训练数据数量与质量，受国内产业环境、数据积累程度、数据运营生态等因素影响，与全球领先国家仍存在一定差距，使得国内大模型难以拥有足够专业的数据资源进行训练。本项目通过提供覆盖预训练、强化学习及应用拓展阶段的海量、高质量专业数据集，更好的支撑大模型训练，提升大模型输出能力。

（4）本项目建设符合公司“夯实传统业务，探索新型业务”的战略目标

为更好实现公司业务发展战略，公司在保障人工智能基础数据业务稳健发展的同时，不断探索寻求新的业绩增长点。如前文所述，数字经济时代下，数据要素市场发展前景广阔，大模型等人工智能技术已成为国家科技发展的重要抓手，但国内数据仍存在数据质量差、各领域数据无法流通等问题制约了人工智能行业的发展。公司将基于过往的数据服务经验，结合行业前沿需求，积极拓展大模型训练数据服务领域，力争将大模型训练数据等创新业务打造成为具有潜在高增长价值的新型业务板块。

4、项目建设可行性

（1）数据要素政策红利持续释放，利好政策支撑数据服务产业发展

国家高度重视数字经济发展，而数据要素作为数字经济深化发展的核心引擎重要性更加凸显，多项政策密集出台为本项目的顺利实施提供了政策保障，具体内容如下：

表 1：数据要素相关政策

序号	发布时间	颁布主体	主要行业政策及法律法规	相关内容
1	2023 年 7 月	中共北京市委、北京市人民政府	关于更好发挥数据要素作用进一步加快发展数字经济的实施意见	发展数据生产服务业，支持企业开展数据采集、清洗加工、存储计算、数据分析、数据标注、数据训练等数据生产服务，支持企业研发建设数据生产线，推进数据生产自动化。培育人工智能生成内容产业发展，发展人工智能生成语音、图像和自然语言等内容，丰富合成数据供给。 打造数据基础制度综合改革试验田，支持北京经济技术开发区等开展数据基础制度先行先试，打造政策高地、可信空间和数据工场。通过物理集中和逻辑汇通相结合的方式，导入工业、金融、能源、科研、商贸、电信、交通、医疗、教育等领域数据资源，促进数据跨行业融合应用，切实激活数据要素资源。
2	2023 年 7 月	国家互联网信息办公室等七部门	《生成式人工智能服务管理暂行办法》	生成式人工智能服务提供者（以下称提供者）应当依法开展预训练、优化训练等训练数据处理活动，遵守以下规定：（一）使用具有合法来源的数据和基础模型；（二）涉及知识产权的，不得侵害他人依法享有的知识产权；（三）涉及个人信息的，应当取得个人同意或者符合法律、行政法规规定的其他情形；（四）采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性；（五）《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》等法律、行政法规的其他有关规定和有关主管部门的相关监管要求。
3	2023 年 5 月	北京市科学技术委员会、中关村科技园区管理委员会	北京市促进通用人工智能创新发展的若干措施	归集高质量基础训练数据集：组织有关机构整合、清洗中文预训练数据，形成安全合规的开放基础训练数据集；持续扩展多模态数据来源，建设高质量的文字、图片、音频、视频等大模型预训练语料库。谋划建设数据训练基地：加快建设数据基础制度先行先试示范区，探索打造数据训练基地，推动数据要素高水平开放，提升本市人工智能数据标注库规模和质量。
4	2022 年 12 月	中共中央、国务院	《关于构建数据基础制度更好发	数据作为新型生产要素，是数字化、网络化、智能化的基础，维护国家数据安全，促进数

			挥数据要素作用的意见》	据合规高效流通使用。
5	2022 年 11 月	北京市人大常委会	《北京市数字经济促进条例》	从立法层面，加强数字基础设施建设，培育数据要素市场，规范公共数据的汇聚、清洗、共享、开放、应用和评估管理机制，开展公共数据专区授权运营。
6	2022 年 1 月	国务院	《“十四五”数字经济发展规划》	强化高质量数据要素供给、加快数据要素市场化流通、创新数据要素开发利用机制等重点任务举措
7	2021 年 3 月	十三届全国人大四次会议	《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》	加强关键数字技术创新应用，建设重点行业人工智能数据集，发展算法推理训练场景。

（2）大模型驱动人工智能发展全面提速，新型训练数据服务具备市场空间

随着人工智能大模型技术的发展，行业对数据的依赖程度逐步加深。本项目产出的大模型训练数据集拟显著改善大模型训练中，包括预训练数据获取、清洗、强化学习调优、对齐、应用阶段评测等各个阶段的数据规模与质量问题。该类数据集将有效提升行业内面向大模型训练数据集的类别和质量，并保障数据来源与处理合法合规，也将发挥规模化运营的优势，平衡数据集成本与市场效益，实现基于大模型通用能力和垂直领域数据的支撑和训练学习，协助实现个人数据、企业数据、公共数据等各类高价值数据资源汇聚。本项目与公司多年发展中持续运行的商业模式相契合，市场空间广阔，具备可行性。

（3）公司具备较强的数据生产及服务等综合能力，为项目实施奠定基础

①公司拥有深度学习的技术储备，为新业务提供技术支撑

自 2005 年以来，公司始终致力于为 AI 深度学习提供算法模型开发训练所需的专业数据集，提升模型推断结论的可靠性。公司现已积累较为完备的综合性、一体化数据处理平台及工具体系，覆盖智能语音、计算机视觉、自然语言等全业态领域，可服务于市面上绝大多数数据处理需求。截至 2023 年 6 月 30 日，公司已取得 34 项专利和 164 项计算机软件著作权，覆盖平台工具开发、算法研究、

产品设计等多方面。此外，公司还设置了 AI+研发部门，前瞻性挖掘和布局新兴市场需求，抢占市场先机。

公司现有的深度学习模型数据主要是通过定向采集、精细化标注实现，即通过打标签的方式将数据类别、位置、性状、结构等信息进行精细化标注，提供给深度学习模型进行学习。大模型的训练则需要以海量数据为基础，对数据的缺失值、异常值、格式等进行清洗处理，通过高效的、多元化的、专业的人类反馈不断强化和优化模型训练，提升大模型与用户交互过程中的反馈质量。公司可将现有业务的技术储备复用到大模型业务中，将深度学习数据集生产中积累的能力延伸使用至大模型数据集生产。

②公司具有丰富的、多领域数据集产品生产经验，为新业务奠定经验基础

公司的标准化数据集产品是公司区别于众多竞争对手以定制化服务为主的特有商业模式，在多语种及多音色语音数据集和发音词典、动作捕捉等多模态数据集、以及多语种 OCR 和手写体数据集等方面积累了丰富的标准化产品资源。截至 2022 年 12 月 31 日，公司拥有智能语音数据集产品储备 927 个、计算机视觉数据集产品储备 125 个、自然语言数据集产品储备 282 个。经过多年积累，公司已向下游客户提供了累计约 6,000 次/个定制或标准化训练数据集，覆盖个人助手、语音输入、智能家居、智能客服、机器人、语音导航、智能播报、语音翻译、移动社交、虚拟人、智能驾驶、智慧金融、智慧交通、智慧城市、机器翻译、智能问答、信息提取、情感分析、OCR 识别等 19 类创新应用领域，构建出独具特色的训练数据资源及服务能力集群，公司在标准化数据集产品的能力获得市场认可，并为后续标准化数据产品生产奠定扎实基础。

③公司已经服务全球众多科技巨头，为新业务拓展提供客户资源基础

公司自 2005 年成立以来，始终致力于挖掘行业客户需求，解决客户痛点，通过在智能语音、计算机视觉、自然语言等领域的技术积累，获得全球众多客户认可，包括阿里巴巴、腾讯、百度、科大讯飞、海康威视、字节跳动、微软、亚马逊、三星、中国科学院、清华大学等全球主流企业、教育科研机构以及政企机构。截至 2022 年底，公司累计服务客户数量已达到 810 家。公司的存量客户与新业务的客户重合程度较高，且存量客户群中的部分头部企业已输出或计划输出

其大模型产品与服务，为公司该项新业务拓展提供了客户资源基础。

④公司历来重视数据安全能力及合规体系建设，为新业务提供合规保障

公司一直以来非常重视数据安全能力及合规体系建设，数据安全管理工作获得市场认可。资质方面，公司拥有 ISO27001 信息安全管理体系认证、ISO27701 隐私信息管理体系认证、国家信息安全等级保护三级认证、北京市规划和自然资源委员会行政许可乙级测绘资质等。行业参与方面，公司入选中共中央网络安全和信息化委员会办公室“人工智能企业典型应用案例”，成为中国信通院数据安全推进计划成员单位，董事兼副总经理李科入选该计划数安智库专家，发表《AI 训练数据安全管理实践》等文章，为人工智能领域数据安全管理建言献策，并荣获数安智库 2022 年度优秀专家称号；公司根据实践经验总结、撰写的《人工智能基础数据业务之个人信息收集活动的合规审计》案例获选中国信通院、中国内审协会“全国首届数字化审计论坛”评选的“个人信息保护合规审计先锋实践案例”。

公司一直坚持安全与发展并重的原则，持续进行数据安全合规能力建设，建立了较强的数据合规体系并积累了丰富的数据合规实践经验，为大模型开展合规训练提供合规保障。

（4）公司实施本项目在经济效益和社会效益上具备可行性

基于谨慎测算，本项目内部收益率高于社会基准折现率，说明项目的经济效益较好，盈利能力较强。本项目生产的产品属于国家鼓励的行业发展方向，能够带动产业链上下游各企业协同发展，具备社会效益。

综上，从经济效益和社会效益分析来看，该项目具备较强可行性。

5、项目投资概算

本项目投资金额总量为 38,337.36 万元，投资明细主要包括场地购置及装修费用、设备购置费用、软件购置费用、数据资源采购、技术人员费用和铺底流动资金，具体投资金额如下：

表 1 本项目投资金额明细（单位：万元）

序	项目	金额	拟使用募集资金	占比	是否资本性

号			金额		支出
1	场地购置及装修	18,195.00	18,195.00	51.24%	是
2	设备购置费用	2,563.50	2,563.50	7.22%	是
3	软件购置费用	2,048.20	2,048.20	5.77%	是
4	技术人员费用	1,584.00	1,584.00	4.46%	否
5	数据资源采购	6,690.00	6,690.00	18.84%	否
6	铺底流动资金	7,256.66	4,426.55	12.47%	否
合计		**38,337.36**	**35,507.25**	**100.00%**	

6、项目实施主体及实施计划

（1）项目实施主体

本项目的实施主体为北京海天瑞声科技股份有限公司。

（2）项目实施计划

本项目建设期3年，具体进度安排如下表：

表 2 本项目实施计划

序号	时间安排	Y1				Y2				Y3			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1	场地购置及装修												
2	设备购置												
3	硬件软件平台搭建												
4	人员引进与培训												
5	AI 大模型训练数据集生产												
6	产品销售												

注：Y1、Y2、Y3 代表建设期年份，Q1、Q2、Q3、Q4 代表季度。

7、项目经济效益评价

本项目投资金额 38,337.36 万元，经测算，税后内部收益率为 16.82%，税后投资回收期（含三年建设期）为 5.89 年，经济效益良好。

上述测算不构成公司的盈利预测，测算结果不等于对公司未来利润做出保证，投资者不应据此进行投资决策，投资者据此进行投资决策造成损失的，公司不承担赔偿责任，请投资者予以关注。

8、项目批准情况

本项目已取得北京市投资项目备案证（备案证号：京技审项（备）[2023]150号）。本项目不同于常规生产性项目，不存在废气、废水、废渣等工业污染物，不属于根据《中华人民共和国环境影响评价法》和《建设项目环境影响评价分类管理名录》等相关法律法规需要进行环境影响评价的建设项目。因此，本项目无需进行项目环境影响评价，亦不需要取得环保主管部门对项目的审批文件。

（二）数据生产垂直大模型研发项目

1、项目背景

（1）受大模型技术驱动，全球人工智能产业进入加速发展期，快速提升大模型相关技术能力成为国家新兴科技发展战略

人工智能大模型因其良好的泛化性和迁移性，有助于推动人工智能进入大规模落地应用，已成为人工智能发展新赛道。同时其强大的理解和生成能力，将驱动人工智能技术加速与实体产业融合，并深刻改变未来人类的生活和工作方式，发展大模型技术成为全球各国比拼科技实力，提升经济效率，拉升经济增长的重要动能之一。目前，国际巨头纷纷布局以大模型为核心的通用人工智能产业，产业进入加速发展期。在这一信息技术重点领域，我国与国际巨头存在一定差距，正加速布局和应对。国内众多研究机构、企业积极研究生成式AI大模型技术的最优路径，并进行产品发布。近期，在国内科技及投资各领域的高度关注下，百度、商汤、阿里巴巴、华为、科大讯飞、360、京东、字节跳动等企业均有所行动。

我国在“十四五”期间，针对人工智能的未来发展陆续出台了相关指导方案和激励政策，对人工智能的整体发展方向和技术发展重点做出重要规划，同时提出加强算法创新与应用、推动算力基础设施建设、完善数据基础支撑体系等关键建议，倡导未来不断夯实产业发展新基础。全国各地亦陆续出台多项数据政策，其中，《北京市促进通用人工智能创新发展的若干措施》明确提出要“系统构建大

模型等通用人工智能技术体系：开展大模型创新算法及关键技术研究；加强大模型训练数据采集及治理工具研发；建设大模型评测开放服务平台；构建大模型基础软硬件体系。推动通用人工智能技术创新场景应用。”《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025 年）》提出“到 2025 年，人工智能基础理论研究取得突破；关键核心技术基本实现自主可控，其中部分技术与应用研究达到世界先进水平；人工智能高水平应用深度赋能实体经济，促进经济高质量发展”的目标，并进一步提出了“自然语言、通用视觉、多模态交互大模型等形成完整技术栈；生成式产品成为国内市场主流应用和生态平台”等具体目标。

（2）人工智能大模型正处于产业发展转型关键期，垂直应用面临爆发

在大模型通用性、泛化性以及扩大人工智能应用范围的优势推动下，人工智能加快与各类产业的渗透和融合。人工智能大模型正处于打造商业模式，形成基础设施能力的关键时期，将从通用逐渐走向垂直领域，在基础模型之上的垂直行业应用也有望兴起。大模型在搜索、推荐、智能交互、生产流程变革、产业提效等场景已表现出了较大的潜力。例如，在金融领域，陆续产生了通过构建大语言模型等解读征信报告、实现交互式智能客服，为金融服务提质增效赋能。目前，国内相关机构及头部企业在深耕通用基础大模型研发之外，同时根据自身产业生态布局，打造垂直领域大模型，触达应用场景落地；其他具备模型自研能力的肩部厂商，亦基于开源模型或海量数据，打造垂向大模型，建立垂直行业的平台生态。

图 5 我国 AIGC 商业落地产业图谱如下图所示

来源：亿欧·TE《中国AIGC商用场景趋势捕捉指北》

由于大模型在垂直领域应用场景中，需要依赖垂直领域数据和行业know-how、应用场景和用户数据反哺以及一站式端到端工程化能力等。因此，为实现通用大模型对行业应用的赋能，需要相关领域机构或服务提供商基于通用大模型进行知识迁移，建设行业垂向大模型，实现其纵向业务价值。

（3）大模型对人工智能数据处理技术提出了新要求，该类技术的持续提升是支撑大模型长期发展、持续服务垂直应用的必备能力

目前人工智能进入大模型时代，大规模、高质量数据的重要性愈加凸显，并成为模型训练效果的核心支撑之一，但在数据前沿性及工程化技术方面依然充满挑战。长期来看，AI数据处理技术的持续拓新与发展是及时适应甚至超前引领大模型技术和应用发展的关键。

大模型研发的第一阶段，即预训练阶段，需要通过对海量未经标注数据进行学习，获得"基本的语言能力和通用知识"。虽无需标注，但这一阶段需要对海量数据进行清洗，清洗质量的好坏，会显著影响无监督学习的效果及大模型的精准性。在第二阶段，即强化学习阶段，需要加入人类反馈，人类以标注的方式对机器自学习后的判断进行调整，使得大模型的认知和人类认知进行对齐，亦构成大模型带来优质体验感的核心环节。

当前，业界已形成高度共识，即对于大模型训练来说，数据是模型训练质量的重要保障和核心要素。若要训练一个功能全面的高质量大模型，不仅需要持续获取大规模、高质量、多模态、多场景、多垂向的数据，更需具备持续迭代的高质量数据筛选、清洗等技术和指令、对齐、标注等策略，以不断提升包括预训练阶段、强化学习阶段中所需数据的质量，确保通用能力及各垂直应用能力的提升，为大模型精确性、通用性及泛化能力的实现奠定坚实基础。

2、项目基本情况

本项目建设目标为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。

大模型所需数据不同于传统有监督学习范式下的数据需求，数据规模量级大，且近年随着数据安全环境快速驱严，数据使用权限和范围受到更多的限定，因此大模型时代下的数据处理规则将显著区别于传统方式。此外，由于大模型训练数据本身具有更高的复杂性和多样性，其数据服务规则的设计难度也将指数级提升。因此，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，形成下载方案设计、清洗方案设计、指令方案设计、指令泛化与迁移、指令数据验证、多模态数据方案等多项生成能力，以及在上述方案下的原始数据及标注成果生成能力。

同时，为更好实现数据生产垂直大模型下的各类生成能力，公司将研发并引入预训练数据集设计与处理技术、指令数据集设计与处理技术、任务对齐与泛化技术、强化学习技术、Transformer技术、大模型训练框架技术、大模型训练相关底层工程技术、大模型评测技术等，夯实数据生产垂直大模型构建的基础。

此外，基于数据生产垂直大模型的核心能力，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 6 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

3、项目建设必要性

（1）本项目建设是公司落实国家科技创新发展战略的重要举措

人工智能是战略性新兴产业的重要组成部分，对我国经济发展和提升国家战略安全具有重要意义。在世界政治经济格局加速重构的影响下，未来逆全球化趋势仍将延续。全球产业合作格局重构、国际分工体系全面调整，关键环节的国际竞争将加剧，我国在关键核心技术上的问题愈发突出，战略性新兴产业的产业链安全稳定存在潜在隐患。因此，我国需要进一步集中优势资源，在重点领域加快突破一批关键核心技术，助力提升我国新兴产业的产业链关键环节、关键领域、关键产品的安全保障能力，保障国家战略安全。

公司是我国人工智能数据服务领域的龙头提供商，本项目以研发数据生产垂直大模型为核心，并基于该生产大模型对数据集生产的强大支撑能力，升级海天瑞声一体化技术支撑平台，持续以自主可控的技术与平台为我国人工智能技术与产业发展提供支撑。本项目的建设是公司落实国家科技创新发展战略的重要举措。

（2）本项目建设是巩固公司的核心技术壁垒，构建长期技术实力的必然手段

随着人工智能从深度学习阶段走向大模型阶段，对训练数据服务产生了新的需求，具体可分为预训练阶段和强化学习阶段：在预训练阶段，模型所需的数据量巨大；在强化学习阶段，模型所需的数据质量较高，并需要以相关领域 know-how 作为模型输入。此外，随着多模态大模型的不断发展，跨语音、文本和视频图像数据等多种类别的数据集需求将快速增加。

数据集生产能力和一体化技术支撑平台是公司核心技术的重要体现。目前ChatGPT 等模型执行通用生成任务的效果证明了大模型可具备数据生成能力。本项目的建设将基于公司在深度学习阶段数据集生产所积累的 know-how，自主研发数据生产垂直大模型，构建大模型数据处理技术通用化解决方案能力，实现完整、可持续迭代的大模型数据技术框架和数据策略，进一步提高公司在人工智能基础数据服务领域的智能化水平，巩固公司的核心技术壁垒，形成长期技术实力支撑。

（3）本项目建设是提升公司数据服务综合竞争力的有效途径

大模型训练数据集的生产流程包括设计、获取（模型生成）、清洗、标注、安全管理、质控评测等不同的环节。系统化的开发平台和专业化的软件处理工具对应对大模型时代的数据处理需求和全流程支撑至关重要。本项目有助于进一步优化公司的数据处理技术，促进数据资源处理经验的进一步沉淀，长期来看，可以大幅提高公司的数据处理能力、效率，提升服务范围和水平，适应人工智能发展的新阶段，获得有效长期的发展动力，进一步巩固和提升公司在数据服务领域的竞争力。

4、项目建设可行性

（1）本项目建设符合政策要求和行业发展趋势

2023 年 7 月，国家互联网信息办公室等七部门公布《生成式人工智能服务管理暂行办法》，文件明确指出，“生成式人工智能服务提供者（以下称提供者）应当依法开展预训练、优化训练等训练数据处理活动”、“采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性”。该办法从政策层面对生成式人工智能的数据集提出了明确的合法、合规、合理、准确以及知识产权清晰的高要求。

但目前国内大模型的发展普遍存在数据来源不均衡、数据更新实时性弱、垂直类型数据不足、指令集质量欠佳且存在偏见等问题，由此导致大模型的效果、效率、合规性、合理性等方面亟待完善与提升，且在大模型持续发展过程中，部分问题的影响可能持续扩大。因此，建立一套完整、完善、可持续迭代的大模型

训练数据技术框架和数据策略，符合生成式人工智能技术与应用合规、高效发展的趋势。

（2）公司与现有客户、科研院所联系紧密，可确保项目技术框架明确、技术路线可行有效

公司自 2005 年成立以来，始终致力于挖掘行业客户需求，解决客户痛点，通过在智能语音、计算机视觉、自然语言等领域的技术积累，获得全球众多客户认可，截至 2022 年底，公司累计客户数量已达到 810 家。公司现有客户包括阿里巴巴、腾讯、百度、科大讯飞、海康威视、字节跳动、微软、亚马逊、三星、中国科学院、清华大学等全球主流企业、教育科研机构以及政企机构。

公司部分现有客户是当前大模型领域的积极实践者，通过与客户的长期合作，深度交流，能够第一时间获取大模型研发中数据痛点与需求，并可在持续交流反馈中不断修正本项目的建设方案。此外，公司也与科研院所和高校等开展深入合作，可引入外部专家资源，以保证技术路线的可行性。

（3）公司拥有深厚的技术沉淀和人才储备，具有完成本项目的技术基础

公司深耕行业近 20 年，拥有一支高素质的研发团队，公司高管及核心研发人员大多毕业于清华、北大、复旦等一流院校，大部分曾在微软、阿里巴巴、英特尔、IBM、中科院等业内领先的成熟企业与研究机构担任人工智能领域技术研发与管理的领导职务。截至 2023 年 6 月 30 日，公司研发人员达到 77 人，经验丰富的技术团队为本项目的执行提供了人才保证。

截至 2022 年底，公司拥有算法模型框架 16 个、算法模型数量超过 200 个，公司自然语言理解算法支持包括语义理解、情感分析和意图识别等能力，语音识别算法支持语种 58 个，计算机视觉算法支持几十大类、上百小类的物体识别。公司在智能语音、自然语言、计算机视觉领域均有多年算法积累，该等算法模型能够全面支撑公司多个领域数据生产活动的开展。

5、项目投资概算

本项目投资金额总量为 40,651.64 万元，投资明细主要包括场地购置及装修费用、设备购置费用、软件购置费用、研发人员费用和设备托管费用，具体投资金额如下：

表 3 本项目投资明细（单位：万元）

序号	项目	金额	拟使用募资金投资金额	占比	是否资本性支出
1	场地购置及装修	2,346.00	2,346.00	7.55%	是
2	设备购置费用	29,895.00	21,222.70	68.26%	是
3	软件购置费用	451.89	451.89	1.45%	是
4	研发人员费用	4,902.50	4,902.50	15.77%	否
5	设备托管费用	3,056.25	2,169.66	6.98%	否
合计		**40,651.64**	**31,092.75**	**100.00%**	

6、项目实施主体及实施计划

（1）项目实施主体

本项目的实施主体为北京海天瑞声科技股份有限公司。

（2）项目实施计划

本项目建设期3年，具体进度安排如下表：

表 4 本项目实施计划

序号	时间安排	Y1				Y2				Y3			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1	场地购置及装修												
2	设备购置												
3	硬件软件平台搭建												
4	人员引进与培训												
5	项目研发												

注：Y1、Y2、Y3 代表建设期年份，Q1、Q2、Q3、Q4 代表季度。

7、项目经济效益评价

本项目是公司落实发展战略，顺应行业发展趋势，支撑公司加速数据服务领域算法能力建设、持续构建 AI 产业核心竞争力的必要手段。本项目不直接产生效益，项目建成后将成为公司主营业务长期发展的技术底座。

8、项目批准情况

本项目已取得北京市投资项目备案证（备案证号：京技审项（备）[2023]149号）。本项目不同于常规生产性项目，不存在废气、废水、废渣等工业污染物，不属于根据《中华人民共和国环境影响评价法》和《建设项目环境影响评价分类管理名录》等相关法律法规需要进行环境影响评价的建设项目。因此，本项目无需进行项目环境影响评价，亦不需要取得环保主管部门对项目的审批文件。

三、本次向特定对象发行对公司经营管理和财务状况的影响

（一）对公司经营管理的影响

本次向特定对象发行募集资金扣除发行费用后，拟投资于“AI大模型训练数据集建设项目”和“数据生产垂直大模型研发项目”。“AI大模型训练数据集建设项目”通过建设应用于通用和特定垂直领域的AI大模型训练数据集提升行业内面向大模型训练数据集的类别和质量，“数据生产垂直大模型研发项目”以研发海天瑞声数据生产垂直大模型为核心，升级公司一体化技术支撑平台。本次募集资金项目是公司在现有主营业务基础上，结合市场需求和未来发展趋势，加大对公司核心主业重点产品及重要研究方向投资力度的体现，符合国家大力支持人工智能发展的产业政策以及公司整体战略发展方向，项目实施可以巩固和发展公司在行业中的竞争优势，具有良好的市场发展前景和经济效益，符合公司长期发展需求及股东利益。

（二）对公司财务状况的影响

本次向特定对象发行完成后，公司的资本实力进一步增强。公司的总资产和净资产规模均会有所增长，营运资金得到进一步充实。同时，公司资产负债率将相应下降，公司的资产结构将得到优化，有利于增强公司的偿债能力，提高公司抵御财务风险的能力。同时，公司的总股本也有所增加，且本次募投项目存在一定的建设周期，因此在项目实现效益前，公司净资产收益率、每股收益等财务指

标可能存在一定程度的摊薄。从中长期来看，随着本次募投项目的顺利实施以及募集资金的有效使用，项目效益的逐步释放将提升公司运营规模和经济效益，从而为公司和股东带来更好的投资回报并促进公司健康发展。

四、可行性分析结论

综上所述，本次向特定对象发行股票募集资金投资项目的建设符合国家产业发展规划政策，符合产业发展的需求，符合公司的战略发展目标。在人工智能产业进入以大模型为代表的新的发展时期，通过本次募集资金投资项目的实施，公司将建设一批市场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，公司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，巩固并增强公司综合竞争力，有利于公司可持续发展，符合全体股东的利益。因此，本次募集资金投资项目是必要的、可行的。

北京海天瑞声科技股份有限公司

董事会

2023年10月23日