关于北京海天瑞声科技股份有限公司

首次公开发行股票并在科创板上市申请文件

第四轮审核问询函的回复

保荐机构（主承销商）

华泰联合证券有限责任公司
HUATAI UNITED SECURITIES CO., LTD

（深圳市福田区中心区中心广场香港中旅大厦）

上海证券交易所：

根据贵所于 2019 年 6 月 25 日印发的上证科审（审核）〔2019〕316 号《关于北京海天瑞声科技股份有限公司首次公开发行股票并在科创板上市申请文件的第四轮审核问询函》（以下简称“问询函”）的要求，华泰联合证券有限责任公司（以下简称“华泰联合”或“保荐机构”）作为北京海天瑞声科技股份有限公司（以下简称“海天瑞声”、“发行人”或“公司”）首次公开发行股票并在科创板上市的保荐机构（主承销商），会同发行人和瑞华会计师事务所（特殊普通合伙）（以下简称“瑞华会计师”）等相关各方，本着勤勉尽责、诚实守信的原则，就问询函所提问题逐项进行认真讨论、核查与落实，并逐项进行了回复说明。具体回复内容附后。

说明：

1、如无特别说明，本回复中使用的简称或名词释义与《北京海天瑞声科技股份有限公司首次公开发行股票并在科创板上市招股说明书（申报稿）》（以下简称“招股说明书”）一致。涉及招股说明书补充披露或修改的内容已在招股说明书中以**楷体加粗**方式列示。

2、本回复中若出现总计数尾与所列值和不符的情况，均为四舍若出现总计数尾与所列值和不符的情况，均为四舍若出现总计数尾与所列值和不符的情况，均为四舍五入所致。

3、本回复中涉及的我国、我国经济以及行业的事实、预测和统计，包括本公司的市场份额等信息，来源于一般认为可靠的各种公开信息渠道。本公司从上述来源转载或摘录信息时，已保持了合理的谨慎，但是由于编制方法可能存在潜在偏差，或市场管理存在差异，或基于其他原因，此等信息可能与国内或国外所编制的其他资料不一致。

目 录

问题 1、关于核心技术

“回复材料显示，发行人的核心技术具备较高壁垒、难为同行业公司或上下游突破。发行人注重在相关领域的核心技术积累构建及核心保护，但并不单纯依赖以申请专利的形式对核心技术进行保护。

请发行人区分数据库产品开发和基础研发两类补充披露研发费用的内容构成、金额及占比。

请发行人：（1）用浅白语言说明发行人与同行业竞争对手相比的竞争优势以及较难为同行业公司或上下游行业突破的理由；（2）进一步说明发行人的核心技术未采取发明专利形式进行保护的理由，是否属于行业惯例；（3）发行人的核心技术是否为通用技术，是否具备新颖性；（4）发行人在数据库设计领域的全面性、专业性优势的衡量指标为发行人有能力设计覆盖多语种/方言、多场景、多领域的，采集方案更为复杂的数据库产品，说明上述衡量指标是否合理以及具体的语种、场景、领域、方案的复杂性及其相应的技术难度；（5）结合公司某些典型的产品举例说明发行人在语音语言学基础研究、多语种多模态数据库设计技术、数据同步技术、大数据驱动的高效数据处理技术、分布式高性能自动校验技术等核心技术方面的运用及技术壁垒；（6）说明各类人工智能训练数据的数据库结构，与发行人的数据库结构比较差异情况，发行人提供定制服务所涉及的数据库是否为发行人所设计并提供相关依据；（7）说明与纳税申报表中“加计扣除”研发费用之间的差异情况并逐项解释原因；（8）分析基础研发工作对核心技术尤其是数据库设计技术上的贡献并说明相关依据。

请保荐机构核查并发表明确意见。”

答复：

请发行人区分数据库产品开发和基础研发两类补充披露研发费用的内容构成、金额及占比。

发行人已在招股说明书“第八节 财务会计信息与管理层分析/九、经营成果分析/（四）期间费用分析/3、研发费用分析”中补充披露下述内容：

发行人的研发费用按数据库产品开发、基础研发两类拆分情况具体如下：

分类	2018 年		2017 年		2016 年	
	金额（万元）	占比	金额（万元）	占比	金额（万元）	占比
数据库产品开发	1,264.60	46.25%	1,455.04	57.56%	1,540.10	70.81%
基础研发	1,469.94	53.75%	1,072.95	42.44%	634.82	29.19%
合计	2,734.53	100.00%	2,527.99	100.00%	2,174.92	100.00%

数据库产品开发相关研发费用主要为数据服务费支出（即数据库产品开发过程所需的、非核心技术环节的生数据采集、标记服务支出）、数据库产品开发人员薪酬等，内容构成、金额、占比情况具体如下：

内容构成	2018 年		2017 年		2016 年	
	金额（万元）	占比	金额（万元）	占比	金额（万元）	占比
数据服务费	957.04	75.68%	1,196.38	82.22%	1,248.79	81.09%
职工薪酬	225.37	17.82%	204.40	14.05%	211.87	13.76%
折旧与摊销	24.59	1.94%	31.30	2.15%	37.87	2.46%
其他	57.60	4.56%	22.96	1.58%	41.57	2.70%
合计	1,264.60	100.00%	1,455.04	100.00%	1,540.10	100.00%

基础研发相关研发费用主要为基础研究研发人员薪酬，以及发行人为根据语种扩展的特殊要求而聘请外部语言学家所发生的语言研究劳务费，内容构成、金额、占比情况具体如下：

内容构成	2018 年		2017 年		2016 年	
	金额（万元）	占比	金额（万元）	占比	金额（万元）	占比
职工薪酬	1,148.87	78.16%	876.78	81.72%	510.28	80.38%
语言研究	175.46	11.94%	166.77	15.54%	98.66	15.54%
折旧与摊销	120.60	8.20%	14.62	1.36%	7.51	1.18%
其他	25.01	1.70%	14.78	1.38%	18.37	2.89%
合计	1,469.94	100.00%	1,072.95	100.00%	634.82	100.00%

（1）用浅白语言说明发行人与同行业竞争对手相比的竞争优势以及较难为同行业公司或上下游行业突破的理由

一、发行人相比同行业竞争对手的竞争优势

在核心技术领域，发行人相比同行业竞争对手的优势主要体现在：

（一）专业经验及核心技术积累

发行人自 2005 年起开始专注于数据资源开发，是我国最早专业从事人工智能数据资源产品服务研发与销售的主要企业之一。发行人在人工智能数据服务产业深耕多年，始终秉承基础研究与实际应用紧密结合的原则，持续开展基础研发等研究创新工作。

在数据资源开发相关算法、技术领域，发行人的专业研发团队结合多年数据资源开发经验需求，积累下 12 项核心技术（详见招股说明书“第六节 业务和技术/（二）发行人的核心技术、应用情况及先进性”），尤其在多语种的语言语音学基础研究和高效数据处理技术方面积累下 5 项具备较强专业性、较高技术壁垒，较难为同行业公司或上下游行业突破的核心技术——语音语言学基础研究、多语种多模态数据库设计技术、数据同步技术、大数据驱动的高效数据处理技术及分布式高性能自动校验技术。

在数据处理工具、平台相关领域，发行人自主开发了一体化数据处理技术支撑平台，嵌入数据资源开发各环节所需的工具、软件模块，持续将数据资源开发相关算法、技术的基础研究成果运用至具体工具/平台之中，并结合市场及内部数据资源开发需求的变动持续调整技术应用、打磨技术细节、优化工具/平台，提升开发效率、服务质量及数据安全性，从而为客户提供高效率的数据资源定制服务、高质量的数据库产品和高水准的数据资源应用相关服务。

（二）资源积累和覆盖能力

受益于前述专业技术积累以及多年专业客户服务经验，发行人已积累下超过 500 个自有知识产权的数据库产品，覆盖智能语音、计算机视觉及自然语言三大领域；产品/服务可覆盖 130 余个语种/方言，涉及生活交流、客服、家居、办公、行车、普通环境、噪声等多种特定场景中的语言现象和视觉呈现，可应用于智能助理、智慧交通、智能搜索、智能家居、自动驾驶等 16 类终端应用场景。

上述资源优势根植于发行人超过 14 年的数据资源开发经验和资源积累，需要长时间的行业深耕和持续投入才能形成。经量化比较，发行人在客户数量、获得行业奖项数量等市场指标维度，数据库数量、覆盖语种/方言数量、采集方案

复杂度（录音场景类别、采集设备种类、采集通道覆盖数）等技术维度，以及计算机软件著作权数量、在申请专利数量等知识产权维度，与国内同行业主要竞争对手公开信息渠道列示披露的信息相比存在明显优势（发行人数据库产品与同行业竞争对手对比的具体情况请参见第三轮审核问询函回复问题 1（3）、（4）之相关内容）。

（三）具备维持前述竞争优势的有利条件

1、人才优势——公司高管及核心技术人员具有深厚的技术背景和丰富的行业经验

公司高管及核心人员大多毕业于清华、北大、中科大、复旦等一流院校，大部分曾在中科院、微软、IBM、英特尔等业内领先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。这些核心人员将其在业内优秀企业积累的技术经验和管理经验应用于公司实践，并通过传帮带培养了一批专业而精干的中层技术团队，有效提升了企业的技术水平和规范化运作水平，实现了对客户的快速响应、高品质交付。

2、客户优势——客户覆盖人工智能领域的知名机构，有利于发行人积极服务并跟进人工智能领域的前沿进展

经过多年的发展，发行人与国内外诸多知名机构形成了长期稳定的合作关系，赢得了众多大型优质客户的认可，例如阿里巴巴、腾讯、百度、微软、三星等大型科技公司，科大讯飞、商汤科技、云知声、海康威视等人工智能企业，以及中国科学院、清华大学等科研机构。发行人通过深耕数据资源服务领域，伴随了众多国内客户在人工智能领域特别是智能语音领域的开拓、成长，为其提供了全球语种语音数据资源和高质量的本土服务，降低了其对国外同类数据资源的依赖，形成了强有力的客户粘性和信任关系。

发行人提供的数据资源产品服务通常应用于人工智能行业下游客户在算法模型开发、训练、拓展方面的前沿需求。通过了解客户需求、响应客户需求至最终交付向客户交付产品服务，发行人积极参与并服务了下游客户在人工智能领域的前沿开发，跟进了业界技术的最新发展成果，把握了未来技术发展趋势，不断推动自身基础研究和产品研发水平的提升。

二、各竞争优势较难为同行业公司或上下游行业突破的理由

发行人在国内起步较早，拥有稳定的、技术背景深厚、行业经验丰富的管理团队、研发团队以及语言学家咨询团队等，在各类语种/方言、应用场景、领域的数据资源开发方面具备先发优势、经验优势，有能力将积累的资源、经验应用于多语种语言语音学基础研究和高效数据处理技术等方面，构建了壁垒较强的核心技术体系。在此基础上，发行人通过持续投入形成了良好的客户规模、数据库产品规模、语言及应用场景覆盖能力。

发行人的专业经验及核心技术积累优势，以及数据库产品资源积累、语种/方言覆盖能力优势均根植于发行人数据资源开发经验和资源积累，需要较长时间的行业深耕和持续研发投入形成，上下游企业、同行业公司在短期内难以较易突破，具体如下：

慧听科技、标贝科技等境内同行业企业也专业从事数据资源开发，并掌握一定的类似技术，但与发行人相比，其在语种/方言覆盖能力、经验资源积累、数据量处理积累方面存在一定差距，弥补前述差距需要其投入较多时间、资源，难以在短期内突破发行人通过上述优势构建的核心技术壁垒。

下游行业主要为人工智能技术层及应用层相关企业，其因自身算法模型开发、训练所需，亦掌握一定的语音语言学基础研究知识、数据库设计技术、数据处理技术、数据质检技术等，但相关技术通常侧重于其算法模型所覆盖的特定语种/方言、应用领域，与发行人等专业从事数据资源开发的企业相比，因在产业链专业化分工侧重性不同，其在数据资源开发处理领域的专业化程度、全面性、数据处理量、实践经验、流程把控能力等方面存在一定差距，难以在短期内突破发行人通过上述优势构建的核心技术壁垒和资源壁垒。

上游供应商人力资源服务公司、技术服务公司等主要参与发行人等数据资源提供商进行数据资源开发过程所涉及的生数据采集（使用发行人提供的采集工具进行录音、拍照、编写文字等）和数据标记的具体执行（使用发行人自主研发的标注软件、工具集和平台对已分配的数据包进行数据属性及特征标记等）等非核心环节，不掌握发行人所具备的数据资源开发相关核心技术，难以突破发行人构建的核心技术壁垒。

三、小结

综上，发行人与同行业竞争对手相比的竞争优势主要体现在专业经验及核心技术积累、资源积累和覆盖能力方面，且发行人具备维持前述竞争优势的有利条件。发行人的竞争优势根植于发行人在行业中经营实践多年的业务经验及核心技术积累、人才积累、客户服务经验积累，因而较难为同行业公司或上下游行业突破。

（2）进一步说明发行人的核心技术未采取发明专利形式进行保护的理由，是否属于行业惯例；

一、发行人未针对全部核心技术采取发明专利形式进行保护的原因

（一）历史经营期间，发行人主要采取制度、技术角度的保护措施对核心技术进行保护

发行人根据数据资源开发过程中的相关需求，持续研发和创新，积累下相关核心技术。前述核心技术主要服务或应用于数据资源相关产品服务的内部开发及生产过程，因而在发行人历史经营期间，一方面考虑到相关技术的应用不直接体现于发行人的产品和服务中，对外泄露的风险较小；另一方面出于保护核心技术、减少技术公开程度的考虑，未主要采取发明专利形式对相关技术进行保护，而是主要通过制度、技术角度的保护措施进行保护（参见第三轮审核问询函回复问题1（1）之相关内容）。

（二）当前，顺应行业发展趋势，发行人综合考虑发明专利的保护效果及核心技术的特性，选择是否采取发明专利形式对核心技术进行保护

近几年，随着业务持续发展壮大、核心技术体系逐步发展完善，发行人进一步加强了核心技术保护体系的建设，开始将发明专利保护措施纳入综合性核心技术保护体系之中，对技术类型适合申请专利、生产研发受专利申请相关信息公开的影响可控的核心技术，申请发明专利进行保护。

截至本问询函回复出具日，发行人已提交 8 项发明专利申请，其中 7 项已进入实质审核阶段。上述专利申请中的 7 项从属于发行人核心技术项下的细分技术及研发成果。

（三）未来，发行人将持续完善核心技术保护体系，加强对核心技术的保护

随着数据资源行业持续发展，市场竞争、技术竞争愈发激烈，为保持行业领先地位，发行人在技术创新领域持续投入，并综合考虑技术类型、发明专利的保护效果等，从制度、技术、法律多角度出发，采取多重手段对核心技术进行综合保护。

二、发行人未针对全部核心技术采取发明专利形式进行保护符合行业惯例

发行人的同行业主要竞争对手申请/获得发明专利情况的情况具体如下：

项目	海天瑞声	Appen	慧听科技	标贝科技
正在申请/已授权的发明专利数量	8 项中国境内专利（均正在申请中）	2 项（均已授权）	2 项中国境内专利（均正在申请中）	3 项中国境内专利（均正在申请中）
在申请/已获得的发明专利名称	“语音对齐方法及装置音素误标注的检测方法和装置”、“语音检测方法和装置”、“韵律标注方法、装置和设备”、“中英混语料生成办法、装置、电子设备及存储介质”、“语料选取处理方法、装置、设备及计算机可读存储介质”、“音频质量评估方法、装置、电子设备及存储介质”、“基于麦克风的音频处理方法和装置”	“Document processor and associate method”、“Email document parsing method and apparatus”	“一种去中心化的数据生态系统”、“一种多设备录音的自动切分对齐方法”	“语音合成系统的评测方法和系统”、“一种声音数据管理方法、系统及存储介质”、“一种声音检索方法、装置、系统及存储介质”

数据来源：截至 2019 年 6 月，国家知识产权局中国及多国专利审查信息查询平台（http://cpquery.sipo.gov.cn/）、中国版权保护中心 CPCC 微平台等公开信息查询渠道及第三方机构可查询信息。

截至本问询函回复出具日，发行人同行业主要竞争对手 Appen、慧听、标贝科技在申请/已获得的发明专利数量均较少。Appen 已取得的发明专利主要针对自然语言处理相关的文本处理、语义分析；慧听科技提交的发明专利申请主要针对数据生态系统、语音数据对齐切分；标贝科技提交的发明专利申请主要针对语音合成系统的评测、声音数据管理和存储、声音数据检索。结合上述同行业主要竞争对手的数据资源覆盖领域、数据资源开发通用过程等因素合理分析，其已申

请/取得的发明专利均未完全覆盖全部核心技术。

发行人及同行业主要竞争对手均主要从事数据资源产品服务的研发与销售，核心技术主要涉及数据资源开发的各个内部生产环节，在对外交付的产品服务中不会直接体现前述核心技术的应用情况，发行人及同行业主要竞争对手申请/已获得的发明专利申请均未覆盖全部核心技术，发行人的举措符合行业惯例。

（3）发行人的核心技术是否为通用技术，是否具备新颖性；

一、发行人的核心技术不是通用技术

（一）通用技术无法直接满足生产经营需求

信息技术领域的通用技术为一般性的技术应用提供了基础，但其所能实现的基础功能与最终实际运用目标之间通常存在差距。在数据资源领域，不同语种/方言、领域、场景的数据资源在设计、采集、处理（标注）、质检等数据资源开发的各环节均具备特异性特点，通用技术无法直接满足上述特异性、实践性需求，需要根据生产经营需求引入具体技术。

（二）发行人的核心技术以通用技术为理论基础，经原始创新和自主研发积累

发行人以众多通用技术为理论基础，在长期实践工作中总结具体需求和实际难点，通过持续开展基础研究、在实践中持续进行原始创新和自主研发，并反复迭代解决方案，实现了具体技术的落地化、可用性研发，打造出了系列核心技术。例如：发行人将文本正则化技术在多语言数据库设计中落地，与其他技术一起构建了语料设计体系；根据数据资源开发过程的具体需要，自主研发大量软件、工具、平台，有机整合各项技术，构建了发行人的核心技术体系和专业壁垒。

二、发行人的核心技术具备新颖性

发行人的各项核心技术均是其原始创新、自主研发的成果。发行人通过整合各项核心技术，研发数据资源开发所需的软件、工具、平台，构建了专业优势及技术壁垒（其中 5 项核心技术具备较强的技术壁垒，具体参见第二轮审核问询函回复问题 3（1）之相关内容）。

此外，发行人在多个核心技术领域均提交了发明专利申请。截至本问询函回

复出具日，发行人已提交 8 项发明专利申请，其中 7 项已进入实质审查阶段。

发行人核心技术的新颖性主要体现在下述两个方面：

（一）实现了各领域技术的创新性、有机性结合

发行人将多领域技术创新性、有记性地结合在一起，构建为适应生产经营所需的新技术，提升了数据资源开发过程各环节的生产效率。

以发行人的核心技术“基于语音特征的韵律预测技术”为例：通常情况下，韵律标注环节需要引入有经验的标注员，通过收听语音并人工标记不同级别的停顿位置得到韵律信息；而发行人有机结合了语音识别技术和自然语言处理领域的文本分析技术，经过原始创新和自主研发构建了韵律预测技术，一方面对语音数据资源进行分析计算生成韵律预测结果，另一方面从语义角度对韵律进行预测，最终通过分析上述两个角度的韵律预测结果得到更为可靠的韵律预测结果。通过运用上述技术，发行人将实际韵律标注过程的工作效率有效提高了 25%。

以发行人的核心技术“混合语言语料设计技术”为例：该技术由基础数据的通用特征提取技术和大数据统计技术结合创新构建而成。具体而言，在设计大规模中英混合语料时，不仅需要使数据资源覆盖中文三音素，还要同时考虑英文音素覆盖情况。发行人将自然语言处理领域常见的 n 元语法模型升级为中英混合的语法模型，并应用于文本制作过程中；同时利用自身多年积累的大量语料资源，基于统计结果构建概率模型，并将上述两个模型相结合，最终形成可基于中英词典自动生成混合语料的技术，将其运用于中英混合语料的制作过程中，满足数据库设计所需。

（二）实现了在特定领域、场景的创新性运用

发行人将多项核心技术创新地应用于特定领域、场景中。以“相似说话人自动筛查技术”为例，该技术中的说话人辨识技术通常应用至基于语音信息的自然人身份识别验证等情境中，发行人创新性地将此类技术和语音识别技术相结合，运用至数据资源开发过程之中，用于衡量多个采集自多个说话人的语音数据的相似程度、帮助筛查发现相似说话人，提升数据资源处理效率和产品质量。

三、小结

综上，发行人的核心技术不是通用技术，而是在通用技术的基础上经过原始创新和自主研发所得，具备新颖性。

（4）发行人在数据库设计领域的全面性、专业性优势的衡量指标为发行人有能力设计覆盖多语种/方言、多场景、多领域的，采集方案更为复杂的数据库产品，说明上述衡量指标是否合理以及具体的语种、场景、领域、方案的复杂性及其相应的技术难度；

一、以语种/方言、场景、领域、采集方案作为衡量指标具备合理性

发行人提供的数据资源经过标注及结构化等优化处理，主要用于人工智能算法模型的训练，不同算法模型适用的具体语种/方言、场景、领域存在差异，所需的训练数据资源也相应存在差异。数据库设计的专业性主要体现在数据库内容、分布等能否符合特定语种/方言、场景、领域的算法模型训练所需，全面性则体现在数据库内容、分布覆盖的语种/方言、场景、领域等的种类数量。将语种/方言、场景、领域作为衡量数据库设计全面性、专业性优势的衡量指标具备合理性。

二、各衡量指标的复杂性及技术难度说明

（一）语种/方言

全球的语种/方言具备多样性，各语种/方言在文字、发音等体系上具备自身特点，因此在设计对应的智能语音数据资源的过程中，需要根据具体语种/方言设计相应的录音文本、采取特异化的处理方式。

以语音数据库为例，其设计目标是在既定数据库容量下，覆盖尽可能多的语言语音现象，同时使各种语言现象的分布情况能够更好地模拟实际场景中的分布情况。语言包括文字、语音两个方面，各语种/方言的文字、语音学都各有其特色，数据库设计方案需与具体语种/方言的特色相符。以中文、阿拉伯语、法语等为例，对具体语种/方言的特异性介绍如下：

1、中文：语言现象由汉字、拼音表达。汉字与拼音存在对应关系，但也存在同一汉字在不同的语境、上下文中具备不同含义或不同发音的现象。

2、阿拉伯语：包含 28 个字母，文字从右向左书写；字母根据所在单词中首、

中、末位置的不同存在 84 种不同的字形变化；此外，阿拉伯文字在书写字母时存在元音缺失的特色，因而任一阿拉伯语单词均存在至少 8 种发音可能，因此在对阿拉伯语文本进行统计、分析等处理时，需要进行元音恢复以预测准确发音。

3、法语：在发音上存在“连诵”和“元音延长”等现象，人工智能算法模型欲实现通过文本预测发音，则需要针对上述现象设计专门算法。

不同的语种/方言存在不同的语音、语言学现象以及发音、拼读、语义特点，因此智能语音算法模型需要针对特定的语种/方言设计特异性算法，并使用符合相关语种/方言特色的数据资源进行训练。

为设计出满足上述要求的智能语音数据库资源，发行人在数据库设计时所需的文本设计基本过程如下：

（1）深入研究对应语种/方言的语言音韵学，依据国际音标使用规范，结合计算机使用的便捷性，设计出易于计算机处理的各语言音素集标准；

（2）研究语言文字与音素集的关系，并考虑各语种/方言的特有语言现象，设计相应的字音转换规则，研发发音预测算法以预测文本的发音；

（3）根据语种/方言特色开展语音数据库的文本设计工作，在覆盖对应语种/方言绝大多数语言现象的同时，尽可能地使设计出的文本中语言现象的分布符合该语种/方言一般的语言现象的分布情况。

因此，针对各语种/方言进行文本设计等需要综合运用语言学、语音学、计算机知识，并通过研发数据结构与算法实现对特定语种/方言语言语料的筛选，具备复杂性和技术难度。

（二）场景

人工智能算法模型根据其适用的具体语种、场景、领域存在差异，为使算法模型及应用在目标使用场景下获取更优的应用效果，算法模型训练使用的数据资源需要贴合实际情境需求、反映目标场景特点。

以智能语音算法模型为例，其训练数据库的采集场景根据最终算法模型应用场景不同，可能包括录音棚、室内、室外街道、餐厅、车内等环境。在室内场景中，需要根据噪声情况、采集设备、采集通道数、采集人员年龄性别口音等多方

面因素对数据库进行设计。而在车内场景中，除考虑上述因素以外，还需要考虑汽车的品牌与型号、发音人所坐位置、录音设备部署位置、车内噪声产生原因（车窗、讲话、广播、空调等）等因素，经合理组合设计出特定的噪声环境。不同场景数据库设计所需考虑的具体因素不同，无法直接复用其它场景的设计方案，需要进行针对性设计，具备复杂性和技术难度。

（三）领域

智能语音、计算机视觉、自然语言等不同领域的人工智能算法模型所处理的信息数据类型、拟实现的应用目标及方式存在较大差异。相应地，不同领域的训练数据库设计需要处理各自领域的差异性问题，具备复杂性和技术难度。具体举例如下：

①智能语音数据库设计：需要考虑噪声环境（包括安静环境，逐渐扩展为低噪声、高噪声、混合噪声）、采集设备（单通道，逐步扩展为多通道、麦克风阵列，乃至手环、音箱等智能设备）、录音方式（手持、近耳，远场）等多样性因素，以及具体数据资源的应用领域及类型（数字识别，逐渐变为连续语音识别，趋向对话识别、情感识别）等多样性因素。

②计算机视觉数据库设计：需要综合考虑数据采集地点、天气、光线、道路、交通状况等影响因素的多样性；在人脸数据库设计时则需要考虑肤色、年龄、性别、附属物、环境等影响因素的分布。

此外，设计、开发各领域数据资源，还需要针对不同领域开发对应的采集软件、数据处理流程和质检方案。因此，数据库覆盖领域数量也是数据库设计相关专业能力的反映指标。

（四）采集方案复杂度

数据资源提供商根据算法模型对应的语种/方言、场景、领域类型以及客户的其他个性化需求设计了多种多样的训练数据库，也需要相应制定适宜的、可行的采集方案以实现前述数据库设计的特定效果。采集方案依据语种/方言、场景、领域的差异在设计上也具备复杂性及技术难度。因而采集方案复杂度也是数据库设计相关专业能力的反映指标。

（5）结合公司某些典型的产品举例说明发行人在语音语言学基础研究、多语种多模态数据库设计技术、数据同步技术、大数据驱动的高效数据处理技术、分布式高性能自动校验技术等核心技术方面的运用及技术壁垒；

结合发行人的典型产品印地语语音数据库（手机）（King-ASR-282-2）、中文知性女声语音数据库（King-TTS-031）、山东重口音普通话语音数据库（手机）（King-ASR-384-12）及中文温柔女声语音数据库（King-TTS-042）对发行人在语音语言学基础研究、多语种多模态数据库设计技术、数据同步技术、大数据驱动的高效数据处理技术、分布式高性能自动校验技术等核心技术方面的运用及技术壁垒情况举例说明如下：

核心技术	典型产品示例	典型产品示例对应的核心技术运用举例	技术壁垒情况说明
语音语言学基础研究 举例： ■ 基于词典和模型的发音预测技术	**印地语语音数据库（手机）（King-ASR-282-2）：** King-ASR-282-2 是发行人针对印度印地语手机端语音识别应用开发的大规模数据库产品，收录了 1,021 名发音人共计 1,892 小时的语音数据。	King-ASR-282-2 数据库包含 15 万句不重复的印地语文本及其对应的录音数据，在录音文本设计过程中，发行人运用**语音语言学基础研究**项下的**基于词典和模型的发音预测技术**预测文本发音，具体体现在： 对于已收录于词典池中的单词，通过查找方式直接提取其发音；对于词典池未覆盖的单词，运用通过语音语言学基础研究所得的发音预测模型与规则对其发音进行了预测。 以 King-ASR-282-2 的语种印地语为例，其文本与发音不存在一对一的对应关系，发行人需要在设计录音文本时预测对应文本的正确发音。受益于印地语领域的语音语言学基础研究，发音人积累有印地语音素集及发音规则，并运用基于词典和模型的发音预测技术，建立了有效的统计模型，可批量预测录音文本对应的发音情况。	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。 发行人在语音语言学基础研究领域积累的、具有较高壁垒和相对竞争力的核心技术主要体现在基于词典与模型的发音预测技术之中。 该项技术使得发行人可针对超过 130 个语种/方言提供独立的发音词典产品服务或与智能语音数据库配套的发音词典。
多语种多模态数据库设计技术	**中文知性女声语音库（King-TTS-031）：**	发行人运用**多语种多模态数据库设计技术**设计了 King-TTS-031 对应的录音文本，具体如下：	多语种多模态数据库设计技术项下的的混合语言语料设计技术主要运用于

举例： ■ 多语种音素均衡语料设计技术 ■ 混合语言语料设计技术	King-TTS-031 是发行人研发的汉语女声语音数据库产品，应用于语音合成系统，该数据库的发音人音色知性，录音环境为录音棚，总语音数据时长 45.8 小时，包含 15,293 句中文文本、7,149 句英文文本以及 2,505 句中英混文本。	1、利用**多语种多模态数据库设计技术**项下的**混合语言语料设计技术**设计了包含中文汉字、英文字母及标点符号的中英文混合语料。 2、利用**多语种多模态数据库设计技术**项下的**语音语言学基础研究**预测录音文本对应的发音：对于已收录于词典池中的单词，通过查找方式直接提取其发音；对于词典池未覆盖的单词，运用通过语音语言学基础研究所得的发音预测模型与规则对其发音进行了预测。 3、利用**多语种多模态数据库设计技术**项下的**多语种音素均衡语料设计技术**，从已完成发音预测的中英混录音文本数据中筛选出指定规模的录音文本。	数据资源开发的设计阶段，该技术可基于原始中英混语料的特征提取，通过转化纯中文语料生成中英混合语料，并兼顾易读性和常见性，从而设计出具备良好中英文混合效果的数据产品。 发行人已针对该项技术提交了 1 项发明专利申请。
数据同步技术 举例： ■ 音频对齐技术	**山东重口音普通话语音数据库（手机）（King-ASR-384-12）：** King-ASR-384-12 是发行人针对山东普通话手机端语音识别应用开发的典型数据库产品。该产品共有三个通道，分别由三个不同的手机同时采集，每个通道采集约 542 个小时的语音，对应句子数达 500,866 句。	在 King-ASR-384-12 的制作过程中，需要对录制的 50 多万组、每组三句话的语音数据进行比对和对齐，工作量巨大，采用纯人工操作难以较快完成。 发行人通过运用核心技术**数据同步技术**项下的**音频对齐技术**，对上述 50 多万组语音数据进行了并行处理及对齐，高效地完成了前述音频文件的同步需求。	在数据库采集、开发过程中，存在大量多个设备同时采集数据（多通道数据库）的需求，受到设备的物理限制，原始数据难以完全同步。 发行人在数据同步技术领域的技术壁垒主要体现在自主研发的音频对齐技术之中，该技术可通过计算多个语音特征向量的自相关系数解决了多个音频文件自动对齐的问题，提升了数据库的质量。 发行人已针对该项技术提交了 1 项专利保护申请。
大数据驱动的高效数据处理技术 举例： ■ 音素边界自动预测技术	**中文女声甜美语音数据库（King-TTS-042）：** King-TTS-042 是发行人设计开发的有代表性的中文女声语音数据库，用于甜美特点的语音合成系	King-TTS-042 包含 178,302 个韵律标注点，数量巨大；发行人运用自主研发的**大数据驱动的高效数据处理技术**项下的**基于语音特征的韵律预测技术**，在基于自然语言处理技术的文本预测方式基础上加以研发创新，有	发行人充分结合熟练掌握的基础算法与多年积累的数据资源，持续优化多个算法模型，并将其运用于众多数据库产品和服务的开发过程中，提高了工作效率，降低了数据处理的成本。

■ 基于语音特征的韵律预测技术	统。该产品包含 27,091 句句子，共包含约 178,302 个韵律标注点。	效提升了韵律预测的准确率，提升了韵律预标注的效率、降低了人工韵律标注成本。	发行人在大数据驱动的高效数据处理技术中构建的技术壁垒和竞争优势主要体现在自主研发的音素边界自动预测技术及基于语音特征的韵律预测技术之中。 发行人已针对该项技术提交了 1 项专利保护申请。
分布式高性能自动校验技术 举例： ■ 相似说话人自动筛查技术 ■ 音素标注正确率校验技术	**山东重口音普通话语音数据库（手机）（King-ASR-384-12）：** King-ASR-384-12 是发行人针对山东普通话手机端语音识别应用开发的典型数据库产品。 该产品收录了 1,004 个不同发音人的录制语音信息，各发音人平均录制的语音信息时长约半小时，对应语句数量各约 500 句。	发音人的多样性是 King-ASR-384-12 应用于语音识别领域的基本特点。在 King-ASR-384-12 的制作过程中，需要对该库中一千多名发音人做相似性筛查，确认各语音数据和对应发音人的归属关系，避免收录的来自同一发音人的语音数据过多，影响数据库质量。 在对 King-ASR-384-12 发音人进行相似性筛查的过程中，采用常规的身份核查手段及全人工听判筛查的成本均较高，采用随机抽样筛查则无法保证高正确率。发行人运用了自主研发的**分布式高性能自动校验技术**项下的**相似说话人自动筛查技术**，通过建立 GMM 模型，对任意两个说话人做相似度打分，筛选出疑似相同说话人，再交由人工检查，减少人工校验工作量并提升了人工校验的正确率。	为保证数据产品和服务质量，发行人制定了一整套完整有效的质检流程，积累了涵盖语音识别等多项人工智能算法的全自动校验技术。同时，发行人专门部署了一套分布式高性能集群系统，可实现大规模数据库产品的质检需求。发行人在该技术中构建的技术壁垒和竞争优势主要体现在自主研发的相似说话人自动筛查技术及音素标注正确率校验技术领域。 发行人已针对前述两项细分技术分别提交了 1 项发明专利申请。

（6）说明各类人工智能训练数据的数据库结构，与发行人的数据库结构比较差异情况，发行人提供定制服务所涉及的数据库是否为发行人所设计并提供相关依据；

一、人工智能数据库结构情况说明

对于人工智能训练数据库而言，结构差异主要体现在数据库内容、数据分布等方面，为了实现更好的算法模型训练效果，需要结合应用场景、领域、运用目标等有针对性地设计。具体举例如下：

①语音识别算法模型：包含足够数据量、分布比重的数字串的训练数据库可实现更好的数字语音识别效果；包含足够数据量的唤醒词的训练数据库可以更好地提高算法模型的唤醒词识别率。

②语音合成算法模型：包含更多样化发音数据的数据库可以合成更多语音。

二、发行人数据库结构与同行业主要竞争对手公开披露的数据库对比情况

将发行人自有知识产权的数据库产品与同行业主要竞争对手公开披露的数据库信息进行对比，各类数据库在结构方面的差异情况具体如下：

1、智能语音数据库

对于智能语音类数据库而言，其在内容、分布方面的结构属性体现在文本内容、噪声环境、录音通道等方面。文本内容指智能语音数据库对应的录音文本覆盖的具体内容或类别，覆盖种类越多则该数据库可针对更广泛的应用领域和场景实现更好的算法模型训练效果；噪声环境指智能语音数据库覆盖的噪声环境类型数量及分布情况，智能语音数据库覆盖的噪声环境越多，则智能语音识别算法模型经训练后可实现更多环境下的可靠识别率；录音通道指智能语音数据库中的语音数据来源的录音设备数量，通道数越多则该智能语音数据库可适配的录音设备就越多。

选取较有代表性的两类智能语音数据库，与同行业主要竞争对手在数据库结构方面比较情况如下：

（1）通用语音识别数据库

数据库结构属性	海天瑞声	Appen	慧听科技	标贝科技
单一数据库覆盖的文本类型数量	1-18 类不等	1~7 类不等	1~3 类不等	1 类
文本类型	拼读词、命令词、数字串、自然数、时间、日期、人名、地名、歌曲名、长句等	数字串、自然数、人名地名、命令词、长句	命令短语、普通短语、日常短句	拼读词、命令词、数字串、自然数、时间、日期、人名、地名、歌曲名等
单一数据库覆盖的噪声环境类别数量（个）	1~6	1~5	1~3	未公开披露
噪声环境类型	办公室、家庭、街边、饭馆、车内、商场	办公室、家庭、街边、车内、其他	办公室/宿舍/家、餐厅/咖啡厅、街道	未公开披露
单一数据库覆盖的录音通道数量	1~4	1~4	1~2	1

注：Appen、慧听、标贝的智能语音数据库情况据其公开披露的数据库信息整理。

（2）车载语音识别数据库

数据库结构属性	海天瑞声	Appen	慧听科技	标贝科技
各数据库的文本内容类型	41 类	4 类	未披露具体信息	36 类
文本内容类型示例	数字串、街道名称、控制词、地名、命令词、英语、数字串、时间、日期、度量衡、长句等	数字串、街道名称、控制词、长句	未披露具体信息	导航找音乐，城市导航，机车系统控制，查火车，查酒店，查餐厅，查地铁，路况查询等
各数据库覆盖的噪声环境类别数量	7	2	5	7
覆盖车载噪声环境具体类别	怠速、低速、低速噪音、高速、高速噪音、城市道路、城市道路噪音	怠速、高速	高速路况、低速路况、停车待机、车窗打开、车窗关闭	怠速、低速、低速噪音、高速、高速噪音、城市道路、城市道路噪音
各数据库覆盖的录音通道数量	4	4	7	1

注：Appen、慧听、标贝的智能语音数据库情况据其公开披露的数据库信息整理。

2、计算机视觉

计算机视觉数据库主要是对现实生活中的人体（包含人脸和身体姿态等）、物体（包含车辆等）、生活场景中的文字图片等图像信息等进行采集和标注所形成的数据库。以人脸数据库为例，计算机视觉数据库的设计结构差异主要体现在

人脸图像角度、光线、背景，被采集人脸对应的肤色、年龄、性别，人脸图像的标注点数等方面。同行业主要竞争对手公开披露的计算机视觉数据库数量及相关信息较少，未披露数据库结构相关属性信息，无法直接对具体数据库的细节结构信息进行比较。

3、自然语言数据库

自然语言数据库主要是对现实生活中的文本类数据进行采集标注所形成的的数据库。自然语言数据库类型较多，例如文本来源可为新闻、短信、聊天记录、小说、诗歌、翻译句对等；标注点种类也较为繁多，如分词标注、命名实体标注、词性标注、句法结构标注、情感标注、极性标注、领域标注、意图标注等。不同自然语音数据库的结构差异主要体现在文本来源、内容、标注属性等方面。同行业主要竞争对手公开信息披露的自然语言数据库数量、信息均较少，未披露数据库结构相关属性信息，因此无法直接对具体数据库的细节结构信息进行比较。

综上，与同行业主要竞争对手公开披露、可在数据库结构维度加以比较的智能语音类数据库产品相对比，发行人的数据库产品在结构方面的差异主要体现在覆盖的噪声类型、录音通道数量、录音文本内容类型等方面，以通用语音识别数据库及车载语音识别数据库这两类典型的智能语音类数据库为例，发行人的数据库结构覆盖丰富的录音文本内容、噪音环境类别、录音通道数量，具备广泛性、强适用性。

三、发行人提供定制服务所涉及的数据库的设计权责承担情况及依据

根据客户的具体需求，在不同项目中，发行人提供定制服务所涉的数据库设计权责是否由发行人承担存在差异，具体如下：

序号	定制服务情境	情境示例	所涉及的数据库是否由发行人设计	依据
1	客户提供数据，由发行人进行后续处理（标记），形成数据库	以智能语音领域无采集的单一化标注项目为例，发行人仅对客户提供的数据承担标注工作，无需进行数据库设计工作。	否	业务协议/合同及项目交互记录——发行人与客户进行协商，并在合作协议或项目交互记录中约定定制服务的具体内容，明确双方权责，确定是否
2	发行人根据客户需求提供数据库设计、采集、标注、质检全链条服务，	以智能语音领域的有采集的单一化标注及多样化标注项目为例，在这两类项目下，除客户已限定特殊录音文本类的项目外，	是	

	形成数据库	发行人均需承担录音文本设计、采集方案设计等数据库设计工作。		需由发行人承担定制服务相关数据库的设计工作。

报告期内，发行人的营业收入主要来源于发行人承担数据库设计工作的业务类别，具体如下：

项目	2018 年		2017 年		2016 年	
	金额（万元）	占营业收入比重	金额（万元）	占营业收入比重	金额（万元）	占营业收入比重
数据资源定制服务	**12,369.55**	**64.20%**	**6,297.23**	**52.89%**	**4,471.89**	**53.09%**
——发行人承担数据库设计	5,395.28	28.00%	2,351.22	19.75%	1,650.61	19.60%
——发行人不承担数据库设计	6,974.27	36.20%	3,946.01	33.14%	2,821.28	33.50%
其中：客户提供数据	5,251.67	27.26%	2,521.99	21.18%	2,146.92	25.49%
客户提供录音文本	1,471.12	7.64%	928.03	7.79%	435.06	5.17%
数据库产品（数据库设计均由发行人承担）	**6,601.67**	**34.27%**	**5,489.31**	**46.10%**	**3,826.77**	**45.43%**
数据资源定制服务及数据库产品						
发行人承担数据库设计	11,996.95	62.27%	7,840.54	65.85%	5,477.38	65.03%
发行人不承担数据库设计	6,974.27	36.20%	3,946.01	33.14%	2,821.28	33.50%
合计	**18,971.22**	**98.47%**	**11,786.54**	**98.99%**	**8,298.66**	**98.53%**

（7）说明与纳税申报表中“加计扣除”研发费用之间的差异情况并逐项解释原因；

一、发行人财务报表中的研发费用（简称“研发费用”）与纳税申报表中“加计扣除”研发费用（简称“加计扣除金额”）之间的差异情况主要系部分研发费用未申请加计扣除所致

报告期内，发行人研发费用与加计扣除金额的差异情况如下：

单位：万元

项目	公式	2018 年	2017 年	2016 年
加计扣除金额	A	1,342.80	955.56	403.52
研发费用	B	2,734.53	2,527.99	2,174.92
差异金额	C=B-A	1,391.73	1,572.43	1,771.40
其中：				
未申请加计扣除的研发费用		1,438.90	1,572.43	1,771.40

内部交易合并抵消影响（注①）		-47.17	-	-

注①：上表中研发费用按抵消发行人合并范围内母子公司之间交易后的金额列示，加计扣除金额按照各个主体直接加总计算列示，合并抵消是造成研发费用与加计扣除金额存在差异的原因之一。

报告期内，发行人研发费用大于加计扣除金额，主要是部分研发费用未申请加计扣除所致，逐项列示如下：

单位：万元

项目	2018 年	2017 年	2016 年
1.数据库产品开发费用	1,264.60	1,455.04	1,540.10
2.合并范围内亏损主体研发费用未申请加计扣除	74.17	117.39	-
3.因口径差异不可申请加计扣除的研发费用	100.13	-	4.16
4.即征即退收入形成的费用（注②）	-	-	227.14
合计	**1,438.90**	**1,572.43**	**1,771.40**

注②：即征即退收入形成的费用指北京信审东审税务师事务所有限责任公司（简称“税务师”）对发行人 2016 年度研发费用加计扣除项目进行审核并出具《研发费用加计扣除鉴证报告》中的增值税即征即退收入金额。其中，135.22 万元于 2016 年当年收到，91.92 万元于 2016 年已申请退税但当年尚未收到。

二、部分研发费用未申请加计扣除的原因

（一）数据库产品开发费用按税务相关规定不得加计扣除

根据《财政部关于企业加强研发费用财务管理的若干意见》（财企〔2007〕194 号）规定，企业研发费用（即原“技术开发费”）指企业在产品、技术、材料、工艺、标准的研究、开发过程中发生的各项费用。发行人数据库产品开发费用属于在产品开发过程中发生的费用，属于研发费用的范畴。

根据《国家税务总局关于研发费用税前加计扣除归集范围有关问题的公告》（国家税务总局公告 2017 年第 40 号）规定，企业研发活动直接形成产品或作为组成部分形成的产品对外销售的，研发费用中对应的材料费用不得加计扣除。

发行人的数据库产品在开发完成后，可多次对外授权销售并形成销售收入；数据库产品开发费用中约 80%左右的支出类型为数据服务费，包括生数据采集和标记服务费——发行人的数据库产品开发费用属于研发费用，性质上主要为研发费用中的材料费用，根据上述规定中要求不得加计扣除。

（二）合并范围内亏损主体的研发费用未申请加计扣除

发行人全资子公司北京中瑞智科技有限公司（简称“中瑞智”）2017 年、2018 年未实现盈利，应纳税所得额为负数。与此同时，中瑞智研发费用金额不大，申请加计扣除对税负影响较小，故未申请享受研发费用加计扣除。

（三）研发费用的归集口径与加计扣除口径存在差异，部分基础研发费用不可申请加计扣除

研发费用归集与加计扣除分别属于会计核算口径范畴和税务口径范畴。会计核算口径由《财政部关于企业加强研发费用财务管理的若干意见》（财企〔2007〕194 号）规范。加计扣除税务口径由《财政部 国家税务总局 科技部关于完善研究开发费用税前加计扣除政策的通知》（财税〔2015〕119 号）、《国家税务总局关于企业研究开发费用税前加计扣除政策有关问题的公告》（国家税务总局公告 2015 年第 97 号）和《国家税务总局关于研发费用税前加计扣除归集范围有关问题的公告》（国家税务总局公告 2017 年第 40 号）规范。加计扣除税务口径允许扣除的研发费用范围采取的是列举方式，即政策规定中没有列举的加计扣除项目，不可以享受加计扣除优惠。

可以享受加计扣除的研发费用范围包括：直接从事研发活动人员的人工费用（工资薪金、基本养老保险费、基本医疗保险费、失业保险费、工伤保险费、生育保险费和住房公积金，以及外聘研发人员的劳务费用）；研发活动直接投入费用（材料、燃料、检验等费用）；用于研发活动的折旧费用（仪器、设备的折旧费）；用于研发活动的无形资产摊销费用；新产品设计试验费用；与研发活动直接相关的其他费用（如技术图书资料费、资料翻译费、专家咨询费、高新科技研发保险费，研发成果的检索、分析、评议、论证、鉴定、评审、评估、验收费用，知识产权的申请费、注册费、代理费，差旅费、会议费等。此类费用总额不得超过可加计扣除研发费用总额的 10%。）

报告期内，发行人发生的与研发活动相关的房屋租赁费、工会经费和职工教育经费等费用符合研发费用会计核算的口径，但不属于上述规定列举的加计扣除范围，因此不满足加计扣除条件，故未申请加计扣除。

另根据财税〔2015〕119 号文件规定，企业委托外部机构或个人进行研发活动所发生的费用，按照费用实际发生额的 80%计入委托方研发费用并计算加计扣

除。2018 年，发行人发生委托研发费用 47.17 万元，按照 80%计算加计扣除金额，剩余 20%部分即 9.44 万元因不符合加计扣除条件，未申请加计扣除。

报告期内，以上因素导致的未申请加计扣除的具体影响如下：

单位：万元

项目	2018 年	2017 年	2016 年
房屋租赁费	66.63	-	-
工会经费和职工教育经费	24.06	-	-
委托研发费用的 20%部分	9.44	-	-
其他相关费用	-	-	4.16
小计	**100.13**	**-**	**4.16**

注：发行人 2018 年开始为研发部门设立专属办公场地，并对工会经费、教育经费按部门单独进行核算。

（四）根据相关优惠政策将即征即退收入作为不征税收入处理

根据《财政部 国家税务总局关于进一步鼓励软件产业和集成电路产业发展企业所得税政策的通知》（财税〔2012〕27 号）文件规定，符合条件的软件企业按照《财政部 国家税务总局关于软件产品增值税政策的通知》（财税〔2011〕100 号）规定取得的即征即退增值税款，由企业专项用于软件产品研发和扩大再生产并单独进行核算，可以作为不征税收入，在计算应纳税所得额时从收入总额中减除。另根据《国家税务总局关于企业研究开发费用税前加计扣除政策有关问题的公告》（国家税务总局公告 2015 年第 97 号）（五）财政性资金的处理中规定，企业取得作为不征税收入处理的财政性资金用于研发活动所形成的费用或无形资产，不得计算加计扣除或摊销。

根据以上规定，2016 年，发行人将即征即退收入作为不征税收入处理；同时，即征即退收入用于研发活动所形成的费用不得计算加计扣除。故发行人 2016 年未将征即退收入用于研发活动所形成的费用申请加计扣除。

根据税务师出具的《研发费用加计扣除鉴证报告》，发行人研发费用中未申请加计扣除的金额共计 227.14 万元，较发行人 2016 年确认为其他收益-增值税即征即退收入的 135.22 万元高 91.92 万元，主要系因：《研发费用加计扣除鉴证报告》中确认的增值税即征即退收入按照截至 2016 年 12 月 31 日发行人已申请即征即退的收入总额 227.14 万元进行确认；而发行人按照实际收到退税款的金额

135.22 万元在财务报表中确认其他收益-增值税即征即退收入——发行人已申请即征即退但尚未收到的91.92万元退税款项已在第二轮审核问询函回复问题9（4）的加计扣除计算过程中列入其他相关费用。

（8）分析基础研发工作对核心技术尤其是数据库设计技术上的贡献并说明相关依据。

一、基础研发工作对核心技术尤其是数据库设计技术的贡献

发行人的基础研发工作主要包含数据资源开发相关的算法、技术的基础研究及工具、平台研发两个层次，各层次工作及其对发行人数据库设计等核心技术的贡献如下：

（一）数据资源开发相关算法、技术的基础研究

1、人工智能技术研究

发行人在研发中心下专门设有新技术研究部，聘有多名智能语音、计算机视觉和自然语言处理领域的资深研究员，紧跟人工智能各领域最前沿的算法知识和应用动态，通过原始创新、自主研发积累掌握了语音识别算法、语音合成算法、计算机视觉算法等人工智能算法技术。

前述基础研究对发行人数据库设计等核心技术的贡献体现在：

（1）发行人通过运用人工智能技术及算法模型（例如语音识别算法项下的语音数据库质量预估技术，语音合成算法项下的语音合成数据库评估技术、语音合成系统评测技术等），测试、检验了数据库对算法模型的训练效果，反向指导发行人优化数据库的内容、结构及采集方案设计等，提升了发行人的数据库设计能力。

（2）发行人将人工智能技术及算法模型应用于数据资源开发流程之中，构建了数据资源开发相关的核心技术，提升了生产效率和服务质量。以韵律标注环节为例，发行人通过有机结合语音识别技术和自然语言处理领域的文本分析技术并加以创新研发，构建了韵律预测技术，将实际韵律标注过程的工作效率有效提高了 25%。

（3）此外，该领域的基础研究使得发行人可以更好地把握人工智能技术的发展方向，得以更深入地理解下游客户对数据资源的运用逻辑和需求痛点，从而为客户提供更加优质的产品和服务。

2、语音语言学研究

发行人设有专门的语言研究部门，负责基础语音语言学研究工作，通过自主开展语音语言学基础研究并辅以与语言学家合作的形式，积累语音语言学领域的研究成果，拓展、构建特定语种/方言的音素集、发音规则及发音词典等；并在语种/方言覆盖能力拓展方面持续开发创新，建立起完整高效的发音词典开发流程。

该项技术研究对发行人数据库设计等核心技术的贡献主要体现在：

（1）音素集/发音规则/发音词典等语音语言学研究成果及相关技术积累：该项研究使得发行人在语音语言学基础研究领域构建了核心技术，如基于词典与模型的发音预测技术等；在语种/方言覆盖能力上形成了突出优势，可提供 130 余个语种/方言的发音词典，是发行人在智能语音数据资源领域的主要壁垒及核心技术之一。同时，该项研究使发行人掌握了成熟的词典构建技术和构建流程，能够保证高质量发音词典的持续、稳定研发。

（2）与多模态多通道数据采集技术、大数据驱动的高效数据处理技术等其他技术结合，构建发行人的多语种核心技术能力，如多语言分布式文本处理技术、多语种多模态数据库设计技术、多语言手写体数据采集技术、多语种拼写检查技术等。

3、数据库设计算法等其他技术研究

数据库设计是构建高质量的人工智能算法模型训练数据资源的关键环节。发行人的新技术研究部也承担数据设计相关技术的基础研究开发工作，致力于与其他部门协同合作，持续构建、完善核心技术体系，提升数据库设计、开发能力，致力于以既定的数据量为客户实现更好的算法模型训练效果。

前述基础研究对发行人数据设计等核心技术的贡献主要体现在：通过布局、开展数据库设计等相关的基础研究，发行人掌握并积累了多语种音素均衡语料设计技术、混合语言语料设计技术等具有壁垒的核心技术，以及数据库设计算法项下的 n 元语法模型训练与优化技术、文本正则化技术、基于语言模型的文本易读性评测技术等，提升了数据库设计水平和能力。此外，发行人也应客户需求的提升，在多语种、多场景、多领域等维度积累理论和实践经验，自主研发数据库设计算法，并将相关成果运用于多个定制服务、产品开发项目中，持续升级迭代，最终形成具有竞争优势的核心技术，提升数据库设计能力。

（二）数据资源开发相关工具、平台的研发工作

发行人在研发中心下设有软件开发部，负责持续优化发行人整体技术架构体系，并根据数据资源开发过程的实际需要持续开发、完善各业务环节所使用的工

具、平台，致力于将其他算法、技术的基础研究成果运用至具体工具、平台之中，推进基础研究成果的实际应用，提升数据资源开发效率和质量；并将技术、研发成果应用中所获的经验及成果以数据或日志的形式反馈至相关技术、成果的研发团队，形成积极有效的研发工作闭环体系，促进核心技术的持续积累和提升。

该类基础研发对发行人数据设计等核心技术体系的贡献主要体现在：构建了发行人的一体化数据处理技术支撑平台，整合贯通了数据资源开发过程相关的设计、采集、处理（标注）、质检等业务环节，将项目管理、质量控制、数据安全的相关需求模块化、工具化、流程化、规范化、体系化并嵌入至一体化数据处理技术支撑平台中，充分提高了数据资源的开发效率及质量控制水平。

二、基础研发工作对核心技术尤其是数据库设计技术贡献的相关依据

（一）发行人高度重视基础研发工作，持续投入研发人员及研发支出

截至 2018 年末，发行人已构建下 21 人的基础研发团队；2016-2018 年，发行人在基础研发领域的研发支出分别达到 634.82 万元、1,072.95 万元及 1,469.94 万元，各年投入持续提升。

（二）基础研发构建了核心技术体系，积累下丰富的研发成果

发行人从多年数据资源开发的具体经验、需求出发，持续开展基础研发工作，在设计、采集、处理（标注）、质检等数据资源开发的各个环节持续积累核心技术和工具、平台。

基础研发构建了包括数据库设计等在内的核心技术体系，搭建了 “一体化人工智能数据处理技术支撑平台”，与发行人积累的经验优势、资源优势一起构建了发行人的竞争壁垒。

基础研发为发行人积累下了丰厚的研发成果：截至本问询函回复出具日，发行人已取得 98 项计算机软件著作权，覆盖多项核心技术成果；已提交 8 项发明专利申请（覆盖 7 项核心技术项下技术），其中 7 项已进入实质审查阶段。

三、小结

综上，基础研发构建了发行人包括数据库设计在内的核心技术体系，搭建了

“一体化人工智能数据处理技术支撑平台”，为发行人积累下丰厚的研发成果，并与发行人积累的经验优势、资源优势一起构建了发行人的竞争壁垒。

请保荐机构核查并发表明确意见。

针对上述事项，保荐机构执行的核查程序如下：

1、通过查询公开信息、访谈发行人客户及所在行业专家、访谈发行人管理团队及业务人员等手段了解发行人的市场地位、技术实力、竞争优势相关信息，了解数据库设计专业性、全面性的衡量指标。

2、访谈发行人的管理团队、研发团队，了解发行人核心技术的积累历程、相关发明专利申请情况、核心技术在具体产品中的应用情况；了解发行人基础研发工作主要内容及其对数据库设计技术等核心技术的贡献；查阅发行人的发明专利申请文件、计算机软件著作权证书等。

3、取得发行人与核心技术保密相关的制度文件并核查相关制度的执行情况，了解发行人对核心技术采取的保护措施情况。

4、通过公开信息查询等方式，获取同行业主要竞争对手公开披露的数据库结构情况并与发行人数据库产品进行对比分析。

5、查阅有关增值税即征即退、研发费用加计扣除的相关法规政策，检查发行人相关处理是否符合相关法规政策的要求；复核发行人增值税即征即退计算过程、计算依据及计算方法的合理性；核查发行人财务报表中的研发费用与纳税申报表中“加计扣除”研发费用之间的差异情况。

经核查，保荐机构认为：

1、发行人与同行业竞争对手相比的竞争优势主要体现在专业经验及核心技术积累、资源积累和覆盖能力方面，且发行人具备维持前述竞争优势的有利条件。发行人的竞争优势根植于发行人在行业中经营实践多年的业务经验及核心技术积累、人才积累、客户服务经验积累，因而较难为同行业公司或上下游行业突破。

2、发行人未针对全部核心技术采取发明专利形式进行保护的原因主要是：历史经营期间，发行人主要采取制度、技术角度的保护措施对核心技术进行保护；

当前，顺应行业发展趋势，发行人结合发明专利的保护效果、核心技术的特性，综合选择是否采取发明专利形式对核心技术进行保护；未来，发行人将持续完善核心技术保护体系，不断加强对核心技术的保护。发行人未针对全部核心技术采取发明专利形式进行保护符合行业惯例。

3、发行人的核心技术不是通用技术，而是在通用技术的基础上经过原始创新和自主研发所得，具备新颖性。

4、发行人选取的数据库设计全面性、专业性指标具备合理性，在具体语种、场景、领域、方案方面具备复杂性和技术难度。

5、发行人将语音语言学基础研究、多语种多模态数据库设计技术、数据同步技术、大数据驱动的高效数据处理技术、分布式高性能自动校验技术等核心技术方面充分运用于具体产品研发、生产中，上述技术具备技术壁垒。

6、与同行业主要竞争对手公开披露的、可在数据库结构维度加以比较的智能语音类数据库产品相对比，发行人的数据库产品在数据库结构方面的差异主要体现在覆盖的噪声类型、录音通道数量、录音文本内容类型等方面。以通用语音识别数据库及车载语音识别数据库这两类典型的智能语音类数据库为例，发行人的数据库结构覆盖丰富的录音文本内容、噪音环境类别、录音通道数量，具备广泛性、强适用性。

7、发行人财务报表中的研发费用与纳税申报表中“加计扣除”研发费用之间的差异情况主要系部分研发费用未申请加计扣除所致，差异原因具备合理性。

8、基础研发构建了发行人包括数据库设计等在内的核心技术体系，构建下发行人的“一体化人工智能数据处理技术支撑平台”，为发行人积累下丰厚的研发成果，并与发行人多年经营积累的经验优势、资源优势一起，构建了发行人的竞争壁垒。

问题 2、关于募集资金用途

发行人本次拟募集资金7.2亿元，募投项目的实施地点为北京市海淀区，公司拟在北京市海淀区中关村、上地区域附近购置房产用于各项目的研发和办公场地。

请发行人：（1）区分场地费、人员薪酬、数据购置和开发费等类别量化说明募集资金的具体细分用途及合理性，并结合周边土地价格、人员工资水平、工时等量化分析募集资金金额的合理性；（2）结合购置房产等募集资金使用后的情况预测房产折旧、员工薪酬等对发行人未来成本、利润的具体影响；（3）说明募集资金数额和投资项目与企业现有生产经营规模、财务状况、技术水平和管理能力、在手订单及未来订单获取能力等是否相适应及依据，并对公司募集大额资金后的管理和消化能力做风险提示；（4）募集资金中用于扩大现有产品产能的部分，结合现有各类产品在报告期内的产能、产量、销量、产销率、销售区域，项目达产后各类产品新增的产能、产量，以及本行业的发展趋势、有关产品的市场容量、主要竞争对手等情况对项目的市场前景进行详细的分析论证；募集资金中用于新产品开发生产的，发行人应结合新产品的市场容量、主要竞争对手、行业发展趋势、技术保障、项目投产后新增产能情况，对项目的市场前景进行详细的分析论证；（5）说明募集资金投资中场地购置费用、场地费用的用途区别，结合目前公司场地费用情况说明其合理性；（6）说明如何准确区分募集资金投资金额用于“天籁”自主研发产品扩建项目、一体化技术支撑平台建设项目等项目的同类费用；（7）发行人于2018年末的资产总额为2.1亿元，本次拟募集资金7.2亿元远大于公司资产总额和收入规模，进一步说明公司募集资金总额的合理性，并对上述事项做针对性风险揭示及重大事项提示。

请保荐机构核查并发表明确意见。

答复：

（1）区分场地费、人员薪酬、数据购置和开发费等类别量化说明募集资金的具体细分用途及合理性，并结合周边土地价格、人员工资水平、工时等量化分析募集资金金额的合理性

一、募集资金按类别划分的具体用途

发行人募集资金投资项目按类别划分的具体用途如下：

具体用途	“天籁”自主研发产品扩建项目	一体化技术支撑平台建设项目	研发中心建设项目	业务管理平台建设项目	补充流动资金
场地费用	共计 8,072.10 万元（1,251 平方米，每平方米购置单价 6 万元，装修单价平均 0.45 万元/㎡），其中 1,775.10 万元（291 平方米）用于办公场地，6,297 万元（960 平方米）用于采集场地	共计 6,073.40 万元（944 平方米，每平方米单价 6 万元，装修单价平均 0.43 万元/㎡），其中 3,623.40 万元（594 平方米）用于办公场地，2,450 万元（350 平方米）用于实验室场地	共计 5,193.00 万元（850 平方米，每平方米单价 6 万元，装修单价平均 0.11 万元/㎡），全部用于办公场地	共计 1,525.00 万元（250 平方米，每平方米单价 6 万元，装修单价平均 0.10 万元/㎡），全部用于办公场地	-
人员薪酬	共计 2,857.00 万元，计划投入人员 60 人，其中利用现有员工 18 人，新增员工 42 人	共计 4,323.00 万元，计划投入人员 35 人，其中利用现有员工 11 人，新增员工 24 人	共计 6,826.50 万元，计划投入人员 45 人，其中利用现有员工 3 人，新增员工 42 人	共计 1,537.50 万元，计划投入人员 25 人，其中利用现有员工 9 人，新增员工 16 人	-
数据购置和开发费	共计 6,241.20 万元，用于采购开发数据库产品所需的数据服务	-	-	-	-
设备、软件购置费用	共计 2,192.25 万元，用于购买服务器、车辆、电脑、办公软件等各类项目所需设备和软件	共计 4,537.35 万元，用于购买服务器、电脑、办公软件等各类项目所需设备和软件	共计 3,900.70 万元，用于购买服务器、电脑、办公软件等各类项目所需设备和软件	共计 200.16 万元，用于购买服务器、备份硬盘、电脑、办公软件等各类项目所需设备和软件	-
其他	共计 2,665.59 万元，包括基本预备费和铺底流动资金	共计 6,114.06 万元，包括基本预备费和铺底流动资金	共计 318.40 万元，为基本预备费	共计 65.25 万元，为基本预备费	共计 10,000 万元，为补充流动资金

二、募集资金具体用途及金额的合理性分析

（一）场地费用

1、场地面积总体扩张的合理性

发行人本次募投项目所需场地包括与人员相关的办公场地和业务实施相关的采集场地、研发活动相关的实验场地等。截至 2018 年底，发行人办公、业务实施、研发场地所用租赁面积合计 2,401 ㎡。为了满足业务快速发展的需求，发行人于 2018 年底于海淀区知春路 1 号购置办公用房 343.77 ㎡，用于子公司中瑞智的研发和办公。但目前发行人场地已较为拥挤，研发、办公场地严重不足，部分项目实施场地需要采取在其他地点临时租赁的方式补充。本次募投项目按计划实施后，发行人员工人数将从 2018 年末的 127 人增加至 251 人，相应地募集资金拟合计购置场地 3,295 ㎡，研发和办公环境得到一定改善，项目实施场地更加充足。

2、场地费用金额的合理性

募集资金中的场地费用根据各投资项目所需面积和市场平均单价测算。场地费用分为场地购置费用和场地装修费用。

办公场地面积根据项目计划投入人员数量和人均办公面积确定，人均办公面积根据具体人员的职能和需求情况确定。“天籁”自主研发产品扩建项目计划投入人员主要为负责项目管理的技术人员，所需办公场地主要为办公工位，按 5 平方米/人左右规划；一体化技术支撑平台建设项目和研发中心建设项目计划投入人员为研发人员，所需办公场地主要为办公工位、设备工位、会议室、机房等，按 18 平方米/人左右规划；业务管理平台建设项目一定程度上可以分享其他项目规划的会议室、机房等公共区域面积，因此按 10 平方米/人左右规划。以上规划同时考虑了项目实施的基本场地需要及公共空间需要，具备合理性。

采集场地、实验室等场地面积和数量根据实际业务需求确定。“天籁”自主研发产品扩建项目规划 19 个录音室和视频录制室，合计面积约 960 ㎡，以满足开发 137 个各类数据库产品的专业录制需求，包括家居、远场环境录音室，播放音乐、桌面、手机、普通安静室内录音室，文学读物类录音室，视频录制室，普

通录音室以及监控室等；一体化技术支撑平台建设项目规划实验室 5 个，合计面积约 350 ㎡，包括音频处理实验室、全消音室、图像视频录制测试实验室等，促使技术研发满足客户日益多样化的数据资源定制需求。

募投项目实施场地的市场平均购置单价按 6 万元/㎡确定，参考以下两方面因素，具有合理性：

（1）发行人 2018 年 12 月购置学院国际大厦 343.77 ㎡，实际购置单价 6.12 万元/㎡，学院国际大厦所在地为北京市海淀区中关村区域，与发行人募投项目拟实施区域一致；

（2）根据发行人与房产中介/部分物业的询价情况，清华科技园科技大厦购置单价在 6.5 万元/㎡左右，中关村辉煌时代大厦购置单价在 5.5 万元/㎡左右，因此平均购置单价按照 6 万元/㎡计算。

平均装修单价根据场地用途的不同，为 1,000-10,000 元/平方米不等，例如一般办公场地为 1,000 元/平方米，多样化标注语音采集场地和部分实验场地由于对隔音、降噪等方面的采集环境要求较高，为 10,000 元/平方米（包括专业录音设备）。

（二）人员薪酬

募集资金中的人员薪酬根据各投资项目计划投入人员和人均薪酬测算，包括项目两年建设期内的人员薪酬。

1、人员数量方面，“天籁”自主研发产品扩建项目按照项目规划的数据库产品开发数量和报告期内发行人的人均产出能力，确定投入人员数量；一体化技术支撑平台建设项目、研发中心建设项目、业务管理平台建设项目按照项目实际需求配置岗位，确定投入人员数量。

报告期内，发行人营业收入平均同比增长率 51.58%以此作为预测增长率，预测发行人募投项目建设期（2019 年和 2020 年）的营业收入。根据募集资金投资项目计划投入的人员情况，测算募投项目实施后的人均收入情况如下：

项目	项目建设期

预测营业收入平均值（万元）	36,735.93
预计平均员工数量（人）	237.50
预测人均收入（万元/人）	154.68

按报告期各期末发行人员工总数计算，报告期内员工人均营业收入分别达86.83 万元、109.24 万元、151.70 万元。募投项目实施后人均收入与 2018 年较为接近，不存在人均收入明显下降的情形。发行人通过实施募投项目，将有效提升技术能力和业务管理能力，从而为人均收入的长期增长奠定良好基础。

2、人均薪酬方面，发行人根据 2018 年人员薪酬情况和 2019 年预算情况，确定项目实施第一年的人均薪酬，并按照 5%的涨幅确定第二年的人均薪酬。人均薪酬根据人员的具体岗位有所差异。

发行人 2018 年研发人员平均薪酬和募投项目实施第一年根据计划投入的研发人员薪酬和人员数量测算的平均薪酬情况如下：

单位：万元

人员类型	募投项目实施第一年	2018 年
高级研发人员	100.91	-
基础研发人员	44.78	35.43

发行人拟通过募投项目强化自身研发实力和技术能力，基础研发人员预计平均薪酬有所上升，与市场水平相比仍处于合理范围，同时计划引入对发行人技术体系和研发工作具备重要作用的较为资深的高级研发人员，人均薪酬较高，具备合理性。

（三）数据购置和开发费

募集资金中的数据购置和开发费根据“天籁”自主研发产品扩建项目计划产出的数据库产品预计投入成本情况，按照发行人报告期内数据库产品支出中数据服务费平均占比 80%左右的比例测算所需的数据购置和开发费。

数据购置和开发费共计 6,241.20 万元，分两年投入，第一年预计投入 2,242.80 万元，相较于 2018 年研发费用中数据服务费 957.04 万元，增长 134%。该增幅一方面反映了发行人完善自主知识产权数据库产品体系、加强高毛利贡献业务的

布局；另一方面体现了发行人对市场的预判，随着行业的发展，算法技术对数据资源的要求更加趋向规模化和特征化，因此规划募投项目建设的数据库产品单库规模较大，同时多为多样化标注语音、自由对话类语音、稀缺语种类语音、儿童类语音、自动驾驶视觉、3D 人脸视觉、稀缺语种词典等采集、标注难度较高的类型。以上综合导致所需的数据服务费支出较高，具备合理性。

（四）设备、软件购置费用

募集资金中的设备、软件购置费用主要为各类研发项目需要的服务器设备和相关软件，其次为“天籁”自主研发产品扩建项目中的采集测试设备和软件。购置计划符合募投项目定位，购置金额根据各项目实际需求和发行人市场询价结果测算，具备合理性。

（五）预备费和铺底流动资金

基本预备费按建设投资的 2%计提，用于项目建设过程中不可预见费用的支出；铺底流动资金为各项目在建设期所需流动资金，按照分项详细估算法进行估算。

（六）补充流动资金

发行人以本次募集资金 1 亿元补充流动资金主要用于满足未来运营所需资金缺口，并以剩余部分为未来公司战略规划的实施提供资金支持。补充流动资金的合理性分析请参见首轮审核问询函回复问题 33（3）。

（2）结合购置房产等募集资金使用后的情况预测房产折旧、员工薪酬等对发行人未来成本、利润的具体影响

根据发行人募集资金的具体用途，在募集资金使用的两年中，员工薪酬、各类固定资产采购所带来的折旧、各类无形资产采购所带来的摊销对发行人成本、费用的影响情况如下：

单位：万元

第一年				
项目	人工支出	固定资产折旧	无形资产摊销	合计
“天籁”自主研发产	1,087.40	490.99	83.99	**1,662.37**

品扩建项目				
一体化技术支撑平台建设项目	1,635.00	456.14	136.81	**2,227.95**
研发中心建设项目	3,330.00	372.50	132.34	**3,834.84**
业务管理平台建设项目	750.00	68.81	16.55	**835.35**
合计	**6,802.40**	**1,388.43**	**369.69**	**8,560.52**
第二年				
项目	**人工支出**	**固定资产折旧**	**无形资产摊销**	**合计**
“天籁”自主研发产品扩建项目	1,769.60	663.15	130.46	**2,563.21**
一体化技术支撑平台建设项目	2,688.00	791.27	261.94	**3,741.21**
研发中心建设项目	3,496.50	628.62	257.84	**4,382.95**
业务管理平台建设项目	787.50	78.58	25.56	**891.64**
合计	**8,741.60**	**2,161.61**	**675.80**	**11,579.01**

注：上述折旧摊销年限中，房屋场地类固定资产折旧年限为 20 年，电子设备类固定资产折旧年限为 3 年，运输设备类固定资产折旧年限为 5 年，软件类无形资产摊销年限为 5 年。

募集资金投资项目实施过程中所带来的人工支出、新增的大量固定资产和无形资产所带来的折旧摊销，对发行人未来成本费用和利润影响金额较大。本次募集资金投资项目与发行人主营业务紧密相关，具备较强的可行性，发行人预计主营业务收入的增长可以消化本次募投项目新增的折旧摊销等费用支出。但如果行业或市场环境发生重大不利变化，募投项目无法实现预期收益，则募投项目折旧摊销等费用支出的增加可能导致公司利润出现一定程度的下滑。发行人已在招股说明书“第四节 风险因素/十六、募集资金金额较大及募投项目实施的风险”中对上述风险因素进行了披露。

（3）说明募集资金数额和投资项目与企业现有生产经营规模、财务状况、技术水平和管理能力、在手订单及未来订单获取能力等是否相适应及依据，并对公司募集大额资金后的管理和消化能力做风险提示

一、募集资金数额和投资项目与发行人现有生产经营规模、财务状况、技术水平和管理能力、在手订单及未来订单获取能力等相适应的依据

（一）募集资金数额和投资项目与发行人现有生产经营规模和财务状况相

适应的依据

发行人经过多年经营积累，拥有超过 500 个自主知识产权可授权使用数据库，并向下游客户提供了累计 2,000 余个定制数据资源库及相关服务。这些产品和服务可支持超过 130 余个语种和方言，可覆盖生活交流、客服、家居、办公、行车、普通环境、噪声等多种特定场景中的语言现象和视觉呈现，构建成独具特色的数据资源集合，已应用于智能助理、智慧交通、智能搜索、智能家居、自动驾驶等16类应用领域。报告期内，发行人营业收入分别为8,422.86万元、11,907.09万元和 19,265.77 万元，扣除非经常性损益后的净利润分别为 1,598.26 万元、3,270.27 万元和 6,206.30 万元，经营业绩保持了快速增长，盈利能力较强。同时，按报告期各期末发行人员工总数计算，报告期内员工人均营业收入分别达 86.83 万元、109.24 万元、151.70 万元，人均净利润分别达 10.61 万元、31.33 万元、52.87 万元，人均营业收入、人均净利润的年均复合增长率分别达到 32.17%、123.25%，人均效益同样保持了快速增长。

行业的进一步深入发展和市场竞争的日益加剧，对发行人产品服务的丰富程度和运营效率提出了更高的要求。但受制于资金实力较弱、规模较小的因素影响，发行人的业务发展已收到一定程度的制约。例如，2018 年度由于部分公司客户随着其人工智能业务的深入发展，数据资源定制化服务需求增长迅速，公司自身数据开发资源优先满足客户的定制化需求，从而使得数据库产品的收入增速放缓。

发行人本次拟募集资金 72,542.46 万元，用于“天籁”自主研发产品扩建项目、一体化技术支撑平台建设项目、研发中心建设项目、业务管理平台建设项目及补充流动资金，围绕发行人主营业务，进一步扩充产品数量，强化技术支撑能力和业务平台管理能力，充实运营资金，从而有效提升发行人经营规模，增强经营业绩和盈利能力，适应行业发展和市场竞争的变化。

因此，本次募集资金数额和投资项目与发行人现有生产经营规模和财务状况相适应。

（二）募集资金数额和投资项目与发行人技术水平和管理能力相适应的依

据

发行人自设立起即进入人工智能数据资源开发及服务行业，具有十几年的行业经验，通过长期的业务实践和创新积累了包括数据库设计、数据采集、数据处理、质量控制等环节在内的多项具体核心技术，拥有软件著作权 98 个，自行研发并投入使用的多种核心软件工具和平台类工具，可覆盖语音、文本、图像等各类数据的开发，实现了数据库从设计、开发、采集、标注到质检的全流程技术支持，能够有效支撑本次募集资金投资项目的建设。

发行人作为多语种、跨领域的数据资源及相关数据服务的提供商， 每年承接大量的项目，项目种类繁多、涉及领域广泛、各项目流程较多且需要公司内部多个部门协同完成。多年的经营积累使得发行人建立了较为完善的内部管理体制，相关管理团队人员具备较为丰富的管理经验，可以保证募集资金投资项目的顺利实施。

因此，本次募集资金数额和投资项目与发行人技术水平和管理能力相适应。

（三）募集资金数额和投资项目与发行人在手订单及未来订单获取能力相适应的依据

本次募集资金规模和具体投资项目，计划新开发 137 个数据库，并建设一体化技术支撑平台、研发中心和业务管理平台，提升业务运营整体效率，从而进一步丰富发行人数据库资源储备，增强发行人产品服务的提供能力。

发行人在数据资源服务领域深耕多年，目前与人工智能产业链上的各类机构都建立了长期的战略合作伙伴关系，主要合作伙伴包括阿里巴巴、腾讯、百度、微软、三星等大型科技公司，科大讯飞、商汤科技、云知声、海康威视等人工智能企业，以及中国科学院、清华大学等科研机构。发行人报告期内在手订单年复合增长率为 46%，呈现快速增长趋势。未来随着人工智能产业发展对数据资源的需求持续增长，发行人基于上述长期合作的客户群体，有望继续保持较强的订单获取能力。

因此，本次募集资金数额和投资项目与发行人在手订单及未来订单获取能力相适应。

二、发行人募集大额资金后的管理和消化能力的风险提示

发行人已在招股说明书“第四节 风险因素/十六、募集资金金额较大及募投项目实施的风险”和“重大事项提示/二、特别风险因素/（七）募集资金金额较大及募投项目实施的风险”中对以下内容进行了补充披露：

本次募集资金金额及投资项目综合考虑了行业和市场状况、技术水平及发展趋势、场地、设备和人员等因素，并对其可行性进行了充分论证，具备合理性。但由于公司为轻资产型公司，且处于成长阶段，公司资产和收入规模相对较小，而本次募集资金金额相对于公司资产和收入规模较大，大额募集资金到位后的管理和消化对公司各方面经营管理能力和资产运营能力均提出了更高的要求。

同时，募投项目实施过程中将新增大量固定资产、无形资产、研发投入，年新增折旧摊销等费用金额较大。如果未来行业或市场环境发生难以预期的不利变化，或由于发行人管理能力、资产运营能力不足等原因对募集资金投资项目的按期实施及完全达产造成不利影响，将导致募投项目的实施进度及经济效益的实现存在较大不确定性，且募投项目折旧摊销等费用支出的增加可能导致公司利润出现一定程度的下滑甚至亏损。

（4）募集资金中用于扩大现有产品产能的部分，结合现有各类产品在报告期内的产能、产量、销量、产销率、销售区域，项目达产后各类产品新增的产能、产量，以及本行业的发展趋势、有关产品的市场容量、主要竞争对手等情况对项目的市场前景进行详细的分析论证；募集资金中用于新产品开发生产的，发行人应结合新产品的市场容量、主要竞争对手、行业发展趋势、技术保障、项目投产后新增产能情况，对项目的市场前景进行详细的分析论证

发行人本次募集资金投资项目不涉及新产品开发生产，“天籁”自主研发产品扩建项目属于扩大现有产品产能。

一、现有各类产品在报告期内的产能、产量、销量、产销率、销售区域及项目达产后各类产品的产能、产量

发行人不属于生产制造型企业，其产出并不主要依赖于固定资产。发行人主

要根据客户订单和自身的研发计划，相应配置项目人员并采购所需的生数据采集、标记服务，因此不适用于产能指标。

发行人数据库产品主要为智能语音数据库产品，现有智能语音数据库产品在报告期内的产量、销量、产销率、销售区域情况如下：

智能语音数据库产品	2018 年	2017 年	2016 年
产量（小时）	19,815.40	14,367.52	22,928.50
销量（小时）	115,732.12	68,520.75	51,053.55
境内销量（小时）	69,238.98	52,529.42	30,592.15
境外销量（小时）	46,493.14	15,991.33	20,461.40
产销率	584.05%	476.91%	222.66%

注：产销率=销量/产量，由于发行人数据库产品可以重复授权，因此存在销量大于产量的情况。

募集资金投资项目实施后，智能语音数据库产品的预计产量情况如下：

智能语音数据库产品	第一年	第二年
产量（小时）	约为 25,800	约为 38,600

发行人报告期内智能语音数据库产品销售情况良好，募集资金投资项目达产后智能语音数据库产品产量相较于报告期内产量有所增长，是发行人扩大经营规模、提升产品能力的重要举措。随着更多的具有特色的数据库产品达产，在较高产销率支撑的情况下，有助于发行人保持经营活力。

同时，随着计算机视觉和自然语言领域的快速发展，发行人这两类数据库产品的产量规模也将在募集资金投资项目达产后逐步扩大，进一步丰富发行人产品在人工智能数据资源服务行业的应用领域。

二、募集资金投资项目扩大现有产品产能的市场前景

（一）行业发展趋势和市场规模情况

经过多年的发展，人工智能在深度学习、海量数据和高性能计算的支撑下，现已进入产业化应用初期。2017 年，全球人工智能产业规模达到 2,307 亿元，预计 2020 年全球人工智能市场规模将达 6,800 亿元。同时，我国正在成为世界人工智能领域的新增长极。（行业情况请参见招股说明书“第六节 业务与技术/二、发行人所处行业的基本情况和竞争状况/（三）所属行业发展情况和未

来发展趋势”）

大规模经标注、可学习的结构化数据推动人工智能应用系统性能的持续提升。因此随着人工智能行业的迅速发展和规模增长，发行人所处的人工智能细分领域——数据资源产品与服务的需求量也将保持快速增长。根据Mckinsey Global Institute研究结果：在大多数应用场景中有效使用神经网络，需要大量的经标注训练数据和有效的计算基础设施，其中深度学习模型对训练数据的需求量更为显著。要发挥人工智能的完整潜能，则需要包括图像、视频及语音在内的多种数据资源。人工智能技术要求算法、模型根据潜在的应用场景变化持续更新，相应地，人工智能算法模型所使用的训练数据也需要定期更新，具体而言，约1/3的算法模型需要至少每月更新，约1/4的算法模型需要至少每日更新。人工智能算法模型的持续更新需要将衍生出大量的训练数据资源需求，各领域数据资源未来仍有较大的需求空间。

（二）市场竞争格局和发行人市场地位

自2015-2016年前后，人工智能技术进入广泛应用期和高速发展期。随着算法、算力技术水平的提升，围绕数据的采集、分析、处理产生了众多的企业。行业内尚未有形成垄断的龙头企业，发行人是我国最早专业从事数据资源产品和服务研发与销售的主要企业之一，技术实力、业务规模均居行业领先地位。

发行人自设立起即进入数据资源开发及服务行业，在行业内深耕十几年，通过多年的业务实践和创新积累了大量核心技术。以上技术有效地支撑了公司在数据库设计、采集、处理、质控等数据资源处理过程中的自主创新。此外，发行人围绕数据库的设计、数据采集、标注、质量控制和数据安全等环节，自主开发打造了“一体化数据处理技术支撑平台”、“数据资源开发一站式解决方案”，广泛应用于智能语音、计算机视觉、自然语言处理等各个技术领域，引领公司打造行业领先的数据资源产品与服务，在核心技术和管理能力领域形成了较强的竞争壁垒。

综上所述，发行人所处行业处于高速发展期，发行人具备较强的市场地位，募集资金投资项目的市场前景良好。

（5）说明募集资金投资中场地购置费用、场地费用的用途区别，结合目前公司场地费用情况说明其合理性

募集资金投资中场地费用包含场地购置费用和场地装修费用，分别用于购置场地和购置后的装修支出。

募集资金用途中的场地费用合理性说明请参见本问题（1）。

此外，与租赁办公场地相比，购置办公场地对公司财务影响相对更小，相关测算[1]如下：

项目（简称）	“天籁”自主研发产品扩建项目	一体化技术支撑平台建设项目	研发中心建设项目	业务管理平台建设项目
预计投入面积	1,251.00	944.00	850.00	250.00
年平均租金	456.62	344.56	310.25	91.25
年折旧	324.12	244.58	220.23	64.77

经测算，发行人采取购置方式比租赁方式的成本更低。发行人拟以募集资金购置房产的用途为研发和办公，从经济效益上测算优于同等地理位置租赁场地。购置场地后，发行人募投项目得以实施，研发和办公环境将得到改善。发行人以募集资金购置房具有合理性。

（6）说明如何准确区分募集资金投资金额用于“天籁”自主研发产品扩建项目、一体化技术支撑平台建设项目等项目的同类费用

场地费用根据各项目所对应人员和业务实际使用的场地区域和面积进行区分并计入项目投资金额，其中人员所涉及的办公场地根据人员归属部门进行区分，各部门办公场地采用区域相对独立划分的方式；采集场地、实验室等场地根据业务经营中的实际用途进行区分，例如各类采集室面积全部归入“天籁”自主研发产品扩建项目，所有实验室面积全部归入一体化技术支撑平台建设项目。

人员薪酬直接根据各项目所对应人员所属岗位进行区分并计入项目投资金

[1]以上测算假设：①购置房产价格按发行人现有办公场地附近写字楼售价约 6 万元/㎡测算，折旧年限按 20 年测算。②租赁房产价格按发行人现有办公场地租赁价格约 10 元/㎡/天测算。

额。不同募投项目对人员岗位属性的要求均有所区别，在人员入职时岗位即已确定；在人员薪酬计算时会对不同部门、不同岗位进行汇总，基于人员薪酬计算表，可以准确区分不同募投项目的人员薪酬投入。

设备、软件购置费用根据各项目实际使用设备、软件的情况进行区分并计入项目投资金额。设备及软件在购置入库时，发行人会记录购置部门、购置用途、领用部门、领用时间等信息，据此作为项目间区分的依据。一般情况下，不会出现不同项目混用同一设备或软件的情形。如果出现，可按当年领用天数占全年总天数进行分摊，分别计入不同项目。

（7）发行人于 2018 年末的资产总额为 2.1 亿元，本次拟募集资金 7.2 亿元远大于公司资产总额和收入规模，进一步说明公司募集资金总额的合理性，并对上述事项做针对性风险揭示及重大事项提示

一、发行人资产和收入规模较小符合自身业务形态特征和行业特点

发行人为轻资产型科技企业，主营业务为人工智能数据资源产品和相关服务的研发与销售，处于成长期，资产和收入规模尚相对较小。报告期内，发行人呈现出资产周转率和净资产收益率较高、流动性较强的特点，一定程度上体现出发行人较高的运营能力和盈利能力，因此发行人可以在体量有限的情况下，充分运用募集资金，借助自身的高成长性产生良好效益。发行人资产、收入规模及运营能力、盈利能力指标与其他软件和信息技术服务业科创板拟上市公司及 A 股上市公司整体水平的比较情况，请参见首轮审核问询函回复问题 33（6）和第二轮审核问询函回复问题 14（3）。

二、发行人募集资金具备合理用途，与发行人实际经营情况相适应

发行人本次募集资金拟用于“天籁”自主研发产品扩建项目、一体化技术支撑平台建设项目、研发中心建设项目、业务管理平台建设项目及补充流动资金。募集资金投资项目均围绕发行人主营业务进行，符合发行人的发展战略。

“天籁”自主研发产品扩建项目将完善发行人数据库产品体系，拓展发行人数据库产品覆盖场景和领域，加快对客户需求的响应速度；一体化技术支撑平台建设项目将进一步升级与丰富发行人提供数据资源服务及开发数据资源产品的

数据处理技术和工具，提升发行人的数据资源开发服务效率，进而增强发行人的核心竞争力；研发中心建设项目将增加发行人的前瞻性技术储备，通过人工智能技术手段的运用提升产品服务质量；业务管理平台建设项目有利于提升发行人的运营管理效率、增强业务执行能力，提升产能及服务质量。补充流动资金可优化发行人资本结构，提升发行人抗风险能力。发行人本次募集资金投资项目具体细分用途和金额的合理性分析请参见本问题（1）。

发行人本次募集资金数额和投资项目与发行人实际经营情况相适应，项目的建设实施具备可行性，具体分析请参见本问题（3）。

三、发行人募集资金金额较大的风险提示

发行人已在招股说明书“第四节 风险因素/十六、募集资金金额较大及募投项目实施的风险”和“重大事项提示/二、特别风险因素/（七）募集资金金额较大及募投项目实施的风险”中对以下内容进行了补充披露：

本次募集资金金额及投资项目综合考虑了行业和市场状况、技术水平及发展趋势、场地、设备和人员等因素，并对其可行性进行了充分论证，具备合理性。但由于公司为轻资产型公司，且处于成长阶段，公司资产和收入规模相对较小，而本次募集资金金额相对于公司资产和收入规模较大，大额募集资金到位后的管理和消化对公司各方面经营管理能力和资产运营能力均提出了更高的要求。

同时，募投项目实施过程中将新增大量固定资产、无形资产、研发投入，年新增折旧摊销等费用金额较大。如果未来行业或市场环境发生难以预期的不利变化，或由于发行人管理能力、资产运营能力不足等原因对募集资金投资项目的按期实施及完全达产造成不利影响，将导致募投项目的实施进度及经济效益的实现存在较大不确定性，且募投项目折旧摊销等费用支出的增加可能导致公司利润出现一定程度的下滑甚至亏损。

请保荐机构核查并发表明确意见。

经核查，保荐机构认为：发行人募集资金金额和具体用途具备合理性；如本次募集资金投资项目按预期实现效益，发行人预计主营业务收入的增长可以消化

本次募投项目新增的折旧摊销等费用支出，同时发行人已在招股说明书中披露了相关风险；发行人募集资金数额和投资项目与企业现有生产经营规模、财务状况、技术水平和管理能力、在手订单及未来订单获取能力相适应，发行人已在招股说明书中补充披露了募集大额资金后的管理和消化能力的风险提示；发行人募集资金投资项目具备良好的市场前景；发行人已在招股说明书中补充披露了募集资金金额较大的风险提示。

问题 3、关于采购业务

根据专项核查报告，2016年及2017年发行人未保留业务数据，2018年发行人仅保留了部分业务数据，中介机构对采购框架合同、采购订单资料、经双方盖章确认的验收结算单、收到的发票、支付采购结算款项的银行回单进行了核查并对劳务公司进行了函证。

请发行人补充披露各类业务经营过程中采购和销售业务的定价方式和过程。

请发行人说明：（1）结合与客户的保密条款约定，说明发行人未保留业务数据的原因及合理性，发行人是否保存了与经营业务相关的日志数据，如何保证生产经营过程可追溯，进一步说明发行人与财务报表相关的内部控制是否健全并有效执行；（2）发行人对劳务公司提供服务的人数和工作量如何进行复核和确认，定量分析实际服务人数和工作量与结算数据的对比情况、结算调整及差异原因，验收结算确认单中相关结算信息的依据和来源，如何保证实际采购业务与财务信息一致。

请保荐机构和申报会计师对上述事项进行核查并发表明确意见，说明：（1）在没有保存业务数据的情况下如何对发行人数据服务费支出进行核查，如何保证核查结论的合理和准确；（2）对劳务公司提供服务的真实性进行核查的过程、程序和范围，包括但不限于核查个人的身份登记记录、个人提供服务后确认的签字凭证、劳务公司向个人付款的相关凭证和完税凭据等。

答复：

第一部分：请发行人补充披露各类业务经营过程中采购和销售业务的定价方式和过程

报告期内，发行人主要的业务类型为数据资源定制服务和数据库产品，其销售和采购业务的定价方式和过程已在招股说明书“第六节 业务和技术/（三）主要经营模式/3、采购模式 及4、销售模式”补充披露：

1、销售业务的定价方式和过程

（1）数据资源定制服务

定价方式：销售价格以定制项目的预估成本为基础，结合项目技术难度、复杂程度、时限要求等与客户协商确定。例如，采集环境和场景、采集对象年龄结构、语种等因素，都会在定价中予以考虑。

定价过程：确定需求→评估成本→评估技术难度、复杂程度、时限要求等因素→确定报价→与客户协商价格→确定价格、签署合同。

（2）数据库产品

定价方式：综合考虑数据库产品的开发支出、市场需求程度、未来估计重复销售频率等因素，向客户报价，并与客户协商确定销售价格。数据库产品通常以单个数据库为单位进行定价，定价比较灵活。

定价过程：确定需求→参考产品开发成本→评估市场需求程度、未来估计重复销售频率等等因素→确定价格区间→与客户协商价格→确定价格、签署合同。

2、采购业务的定价方式和过程

数据资源定制服务、数据库产品的数据服务采购模式无差异。按照采集、标注两个环节，采购业务的定价方式和过程如下：

（1）采集

定价方式：采集环节采购的是音频、图像、文本等未经处理的生数据的采集服务。发行人向供应商采购采集服务的采购总价=结算数据量×采集单价。采集单价根据市场上类似服务的一般价格，并参考采集对象资源的稀缺性来确定。例如，外语种、儿童或老人、行车环境等要求较为特殊的采集项目，采集资源有一定稀缺性，则采集单价高于一般项目。结算数据量根据生数据采集服务的来源和内容，单位也有所不同，常见单位包括人、字/词、张等。

定价过程：确定采集方案→向供应商发出采集需求→综合考虑市场价格和该供应商的过往采购价格，与供应商协商价格→确定采集价格。

（2）标注

定价方式：标注环节采购的是给生数据做人工标记的服务。发行人向供应

商采购标记服务的采购总价=有效工时×标记单价。标记单价根据市场上类似服务价格，并参考标记任务的难度来确定。有效工时根据发行人对标产比的测速结果，与标记人员达成一致后确定。有效工时=交付数据量×标产比。

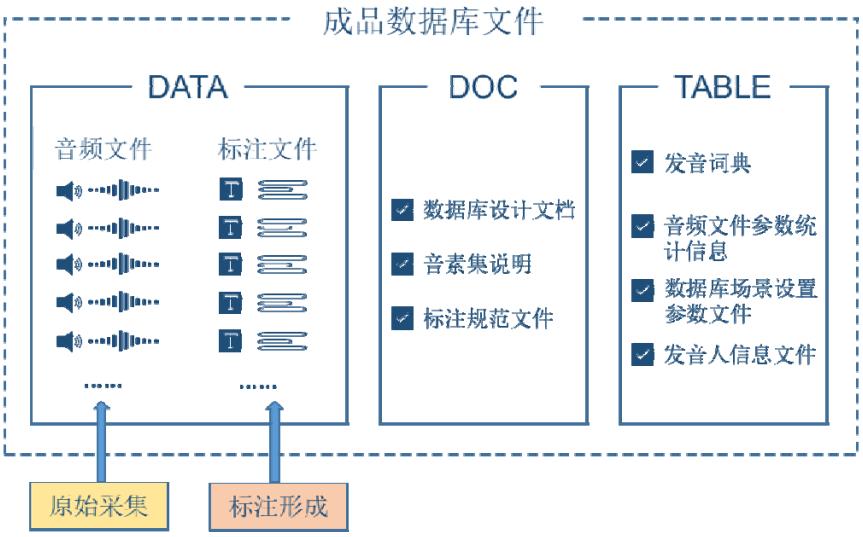
定价过程：确定标注方案→向供应商发出标记需求→综合考虑市场价格和该供应商的过往采购价格，与供应商协商价格；确定项目标产比→确定标记价格。

第二部分：发行人对采购模式的进一步说明

一、发行人数据库的特点

无论是数据资源定制服务还是数据库产品，发行人最终向客户交付的均为数据库文件。发行人完整的生产环节包括：数据库设计→采集→标注→质检，从而形成最终可交付的成品数据库。

以智能语音数据库为例，发行人成品数据库由三部分组成：一是“DATA”文件，为采集的原始音频数据文件和对应的标注文件；二是“DOC”文件，为数据库规范文件；三是“TABLE”文件，为数据库特征信息文件。数据库的具体结构如下图所示：



由上图可以看出，发行人的数据库具有如下特点：

1、数据库结构简单透明，由原始采集形成的音频文件、与音频文件对应的

带有时间戳的标注文件，在数据库中均独立存储；

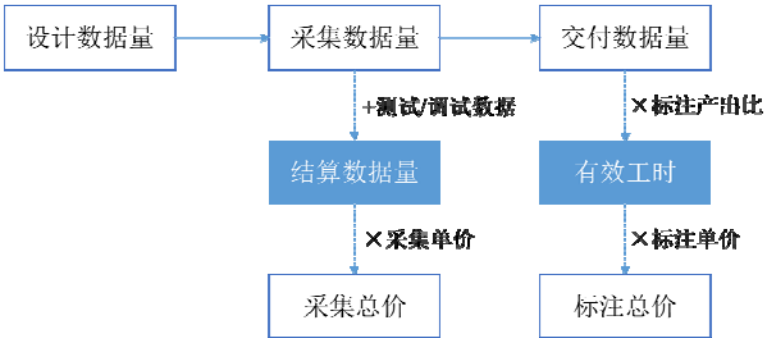
2、交付数据量决定采购数据量，可从成品数据库向原始采集数据追溯；

3、包含了项目绝大部分业务数据信息，包括数据库设计、采集、标注等各个环节。

二、发行人数据服务采购量的确定机制

发行人的数据服务采购涉及采集、标注两个业务环节。其中采集环节采购的是音频、图像、文本等未经处理的原始数据；标注环节采购的是给生数据做人工标记的服务。对于数据采集，发行人以结算数据量为单位与供应商结算；对于标记服务，发行人以有效工时为单位与供应商结算。

发行人数据资源定制服务和数据库产品[2]采购各环节的数据量和有效工时之间的关系，以及与数据采购支出的关系如下图所示：



①采集数据量：指发行人通过采集环节获取的、可进入标注环节的合格数据的数量。

②交付数据量：指经过标注、质检环节，可作为成品交付客户的数据的数量。

③结算数据量：指发行人与供应商确认并进行结算的数据的数量。

④标注产出比（标产比）：指标注环节，每产出1小时的交付数据量，需要花费的标记工作的时间。

⑤有效工时：指标注环节，标记人员的有效工作时长，有效工时=交付数据量×标产比。

三、发行人采购模式的特点

[2]发行人主要产品和服务分为数据资源定制服务、数据库产品、数据资源相关的应用服务三类，但数据资源相关的应用服务规模较小，报告期各年收入、毛利占比均不超过 2%，对主营业务不构成重要影响。

发行人成品数据库的以上特点，使得可以通过数据处理进度即可对采集和标记服务的采购进行有效的监控与核算。

（一）采集

成品数据库中的音频文件即为原始采集的合格的音频文件；后续数据处理并非编辑音频文件，而是向数据集中增加标注文本、数据规范文件、数据库特征信息文件，故理论上原始采集音频文件的数量与成品数据库文件中的音频文件数量是相同的。实际生产过程中，总采集数据量通常会留出少量冗余，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；而发行人通过技术手段、商务约定和严格的项目管理，将冗余量控制在较低范围内。例如，某“150人中文纯净人声录音棚录音项目”，客户订单要求录音人数为 150 人，实际采集人数为 152 人。

因此，数据库需要交付多少数据量，采集多少数据量并留出少量冗余即可，不存在额外采集更多数据量的经济必要性。从上述业务逻辑出发，发行人通过采集数据量与交付给客户并验收的数据量、与供应商结算数据量之间的一致性，来保证采购采集服务的完整性。

（二）标注

原始音频文件采集完毕后，进入标注环节。该环节采购的服务为标记服务，其工作成果体现为标注文件。采集形成的音频文件可以对应到发音人的人数以及录音时长，但标注文件本身无法直接反映出标记的工作量，因此需以某指标为基准进行标准工作量的转换。实际经营中，发行人以交付数据库的数据量为基础，以标产比为转换单位，两者相乘得出标记工作的有效工时。例如，前述“150人中文纯净人声录音棚录音项目”的标记工作，经试标注测速后，标产比确定为2.7（即标注完成1小时音频数据，需要花费2.7个小时的工作量）；假设某标记人员完成了10小时的音频数据标记，则其有效工时核定为10×2.7=27小时。

四、发行人数据服务采购量控制和核算的具体模式

以数据资源定制服务为例，以下对发行人数据服务采购量的控制和核算的具体模式进行说明。数据库产品与数据资源定制服务相比，对外采购所涉及的采集

和标注环节的业务模式无差异。

（一）发行人采购生数据采集服务的具体模式

1、客户合同/订单明确约定交付数据量，发行人根据约定交付数据量采购采集和标记服务

发行人与客户签订的业务合同/订单中会对交付数量进行明确的约定。例如有采集的单一化标注语音业务合同中约定发音人人数和总时长，多样化标注语音合同中约定字/词数或句数。发行人向客户交付数据时，需要满足约定的交付数量并由客户验收通过。

为确保数据质量，发行人采集环节的采集数据量可能略多于交付数据量，以便在剔除不符合要求的数据后仍满足客户交付数量的要求；同时由于发行人启动采集项目前可能进行少量的采集测试、调试采集设备和软件等原因，可能造成结算数据量略大于采集数据量。发行人最终与供应商的结算数据量与向客户的交付数据量之间的差额称为采集损耗，其占总供应商结算数据量的比例称为采集损耗率。发行人通过技术手段、商务约定和严格的项目管理，将采集损耗率控制在较低水平，所需冗余采集量小，不存在额外补充采集较大规模数据的情形。

2、发行人采集损耗率较低，较小的冗余采集量即可满足业务需要

在上述生产模式决定下，当一个项目的交付数据量确定时，发行人的采集损耗率是反映采集数据量的合理性和完整性的主要指标。

（1）发行人的采集模式自身决定了较低的损耗率

①对于有采集的单一化标注语音，只需被采集人按照给定文本或预设场景，根据既定的录音流程进行录音即可，难度较低，原始采集的音频文件的初始合格率本身即比较高，因此不存在原始采集数据大比例不合格的情况。

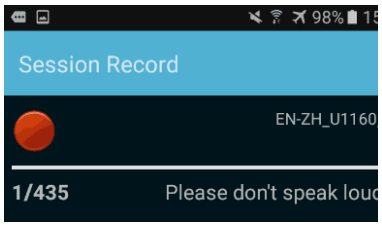
②对于多样化标注语音，由于设计阶段的语料设计和采集阶段的发音人录音要求均较高，且发行人与供应商在合同中对发音人录音事项明确了质量要求，只有录音合格后才会与供应商结算发音人费用，因此后续环节基本不存在录音不合格但仍需与供应商结算该不合格部分的发音人费用的情况。

（2）发行人控制损耗率的措施

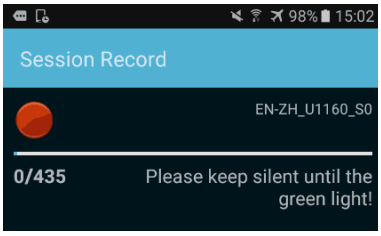
发行人以不超过 5%的内部标准，对采集损耗率进行严格的控制。具体措施如下：

①发行人通过技术手段降低采集损耗率

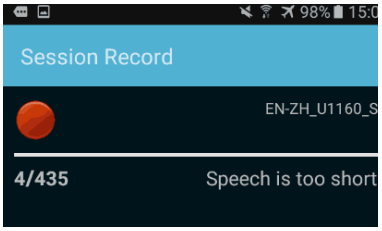
发行人采用了技术手段严格控制采集阶段的不合格数据总量。以单一化标注语音项目的采集为例，发行人在自行开发的各录音软件中均内嵌了质量控制模块，可以通过对语音信号的一些特征（如音量、背景噪声水平等）进行判断，在声音录制的过程中给出反馈，例如音量过大或者过小等，从而可以实时提示发音人对不合格的声音进行重新录制，合格后才能进入下一条语音的录制。软件界面举例如下图所示：



图一：音量过大



图二：未保留足够静音



图三：语音段太短

因此，发行人在采集环节所取得的声音文件在语音信号质量方面已基本满足后续处理要求，从而能够有效降低后续处理环节中初始采集数量的损耗比例。

②发行人通过商务约定和严格的项目管理来降低采集损耗率

同时，发行人也通过商务约定和严格的项目管理对数据质量进行控制。例如，对于有采集的单一化标注项目，在采集阶段开始前，发行人向供应商提出明确的被采集人特征要求，例如发音人的口音地区和轻重、年龄、性别要求等，在采集环节会对数据进行检查。

③小结

通过以上两种方式，发行人在采集环节中对于所取得数据的数量和质量均有严格把控，确保采集数据量一般仅略高于最终产成品的交付数据量，后续环节的

损耗率较低，不存对较多冗余数据量的需求。

（二）发行人采购标记服务的具体情况

对于标注环节，决定标注有效时长的主要因素为标产比。发行人自测标产比来核定标记人员的有效工时，并按该有效工时与供应商进行结算。

1、发行人按标产比折算有效工时，并以其作为采购单位依据的合理性

（1）采用标产比的方式，可以避免标记人员工时统计不准确、工作时长虚增，导致支出核算不准确的问题。此外，采取标产比的方法，有利于督促供应商加强标记人员的管理、提高标记效率。

（2）发行人单位产出的数据服务费支出为量和价相乘的结果。数据服务费的差异主要体现为所消耗有效工时数量的差异（如生数据类型不同、标注环节多少不同），采取标产比的方法即体现了上述差异，因此可以更为准确地反映项目特征与业务实质。

2、发行人确定标产比的方式

发行人通过自行测量的方式确定标产比，并按照自测标产比折算有效工时，依此与供应商进行结算。具体方法为：

（1）发行人将待标注数据拆分为若干任务后，选取少量任务由标记人员进行试标记，测量在连续工作状态下的标产比。

（2）确定速率后与标记人员达成一致，按照标记人员的实际产出数据量在自测速率下折算为有效工时，与供应商进行结算。

除上述自测标产比情形外，还存在少量的客户约定标产比的情形。该情形下，发行人与客户书面约定总交付数据量、标注工时每小时单价和标产比。

发行人单个项目以固定的标产比进行折算。但由于发行人业务定制性较强，各项目所需处理的数据资源类型差异较大，因此各项目的标产比存在一定范围的差异。报告期内，发行人主要业务的标产比保持稳定，变动原因合理，不存在异常波动，具体情况请参见第二轮审核问询函回复问题 7（8）。

综上所述，发行人采集数据量、有效工时与交付数据量直接相关，以交付量确定采购量。采购数据与成品数据具有一一对应关系，采集损耗率控制在较低水平。标记服务的工作量按照标产比折算为有效工时，标产比在报告期内保持稳定，不存在异常波动。

第三部分：请发行人说明：

（1）结合与客户的保密条款约定，说明发行人未保留业务数据的原因及合理性，发行人是否保存了与经营业务相关的日志数据，如何保证生产经营过程可追溯，进一步说明发行人与财务报表相关的内部控制是否健全并有效执行；

一、发行人未保留业务数据的原因及合理性

发行人的业务数据可分为两类：一是发行人生产的数据库本身，包含了采集数据量、交付数据量、标注文本数量以及其他数据库特征信息，其中数据资源定制服务在数据库交付验收后，不再长期保留成品数据库文件，数据库产品业务的成品数据库予以保留；二是生产过程数据，主要以项目采集信息表、项目周报、客户沟通邮件、客户验收报告等线下文档的形式记录，并予以保留。

（一）发行人不保留定制服务数据库的原因及合理性

对于数据资源定制服务，发行人在数据库交付验收后，不再长期保留成品数据库文件。其原因及合理性分析如下：

1、数据资源定制服务完成后，数据库所有权归属客户，发行人不应保留

发行人数据资源定制服务所形成的数据库所有权在销售时已经转移至客户，公司不再拥有该等数据资源的所有权，如果仍然长时间保留备份，则存在相关信息失密和侵权的风险。因此，发行人对定制服务数据不予保留。

2、重要客户对于服务完成后不得留存数据有明确要求

发行人的客户包括大型科技公司、人工智能企业和科研机构等，高度重视数据安全和保密工作，并对其供应商的数据安全和保密作出严格规定，对于发行人定制服务涉及的数据均有相应的保密要求；部分重要客户通过合同或者专项文件、供应商管理制度等形式对服务完成后的数据留存做出了明确要求，具体如下：

序号	客户名称	相关文件名称	相关条款内容
1	微软	SSPA（Supplier Security and Privacy Assurance，供应商安全和隐私保证）	“#13 确保 Microsoft 个人数据和机密数据的保留时间不超过执行相关服务所需的时间，法律要求继续保留 Microsoft 个人数据和/或机密数据的情况除外。 #14 确保 Microsoft 自行决定 Microsoft 个人数据或机密数据在供应商的处理或控制下归还给 Microsoft，或应 Microsoft 的请求在服务执行后进行销毁。 必须在应用程序内制定相关流程以确保从应用程序中安全删除数据（无论是由用户明确删除，还是根据数据期限之类的其他诱因删除）。 如果需要销毁 Microsoft 个人数据或机密数据，供应商必须烧毁、粉碎或撕毁包含 Microsoft 个人数据和/或机密数据的实物资产，以便无法读取或重建相关信息。”
2	阿里巴巴集团	阿里巴巴集团外包安全管理和人员行为规范细则	“1.5.1 乙方人员不得将文档（包括但不限于打印件、复印件、各种甲方内部资料等）带出驻场或供应商场地，如涉及到纸质文档必须用碎纸机销毁。 1.5.2 外包项目结束后，乙方必须将所有交付物（包括但不限于代码、文档、图片等）移交给甲方，不得私自保留拷贝。”
3	百度时代网络技术（北京）有限公司	语音数据标注服务框架协议（第二期）	“4.1 本合同项下的一切数据资源（包含原始数据和标注数据资源，下同）的所有权利（包括但不限于知识产权）归甲方享有。乙方不得对数据资源以及数据资源享有的知识产权进行任何形式的使用，包括但不限于转让、销售、公开或许可第三方使用或用于本合同以外的任何目的。乙方英语数据提交验收合格后对数据的声音文件及文本文件进行删除，不得备份。”
4	腾讯科技（深圳）有限公司	AI 语音素材采购框架协议	“2.所有包含保密信息的或与之有关或相关的材料，无论是否由披露方所提供，包括但不限于废弃图纸、文件碎片、照相底片或计算机输入或输出信息，并且包括任何种类的全部复印件和复制件，乙方应在收到甲方书面要求或相关订单履行完成（以较早者为准）之日起七（7）天内，在合理可行的范围内立即归还或销毁。自此，接收方不得再为可目的使用保密信息。”
5	我国某大型科技公司	EAI 语义引擎项目采购说明书	“接受方应立即返还或销毁所有根据‘本协议’所接受的披露方的‘保密信息’，包括但不限于以任何形式存在的保密信息的原件、复印件、复制品及对“保密信息”的概述摘要，并向披露方提供已经返还或销毁“保密信息”的书面确认。”
6	某国际消费电子产品厂商	CONFIDENTIALITY AGREEMENT（保密协议）	“7.Within ten (10) business days of receipt of Discloser's written request, and at Discloser’s option, Recipient will cither return to Discloser all tangible Confidential Information, including but not limited to all electronic files, documentation, notes, plans, drawings, and copies thereof, or will provide Discloser with written certification that all such tangible Confidential Information of Discloser has been destroyed.” （在收到披露方的书面请求后十（10）个工作日内，并且

序号	客户名称	相关文件名称	相关条款内容
			根据披露方的选择，接收方应将任何形式的机密信息返回给披露者，包括但不限于所有电子文件、文档、注释、计划、图纸及其副本，或向披露方提供所有任何形式的机密信息已被销毁的书面证明。）

3、发行人数据资源定制服务数据库包含了客户的个性化保密信息

发行人定制化数据库的结构参见本问题回复之“（2）/一、发行人产品与生产模式的特点”。以智能语音数据库为例，定制化数据库中记录的音频文件、标注文件、数据库设计信息、被采集对象信息、音频/图像参数和数据库参数等信息，均属于客户的个性化信息，为客户的保密信息。

综上所述，发行人定制化数据库交付验收后所有权转移给客户，重要客户对于服务完成后不得留存数据有明确要求，而定制化数据库包含了客户的个性化保密信息，因此发行人不留存定制化数据库是合理的。

（二）对于数据库产品，发行人对成品数据库予以保留，并不删除

数据库产品为发行人根据对人工智能算法模型应用领域、行业发展趋势、市场需求等的评估和研判设计开发，开发完成后授权给客户使用，销售完成后，数据库产品的所有权不发生转移，仍归发行人所有。因此，数据库产品在销售后发行人予以保留，并不删除。

二、发行人是否保存了与经营业务相关的日志数据，如何保证生产经营过程可追溯

（一）发行人未在系统中保存与经营相关的日志数据，但不影响发行人对采购业务量的监控和核算

发行人的“一体化人工智能数据处理技术支撑平台”为一系列采集和标注软件及工具的集合，业务数据随着生数据采集和数据处理的过程记录和保存于成品数据库中，未以日志数据的形式存在，未在信息系统中保留经营业务相关的日志数据。

发行人依据采集数据量和有效工时与供应商结算数据服务采购费用，两者均与交付给客户的数据量直接相关，在项目执行过程中，已完成的采集数据量可通

过音频文件个数和长度、图片文件个数、文本文件的条数等直观看出，已完成的有效工时可通过已处理的数据量乘以标产比的计算过程得出，故发行人通过数据处理进度即可对采集和标注服务的采购进行有效的监控。例如“云南500人1500小时重口音手机录音项目”，在项目执行过程中的某时点，服务器上已形成300人/900小时的音频文件及其对应标注文件，则发行人按300人/900小时为依据进行采集数据量和有效工时的核算，并不需要依赖于信息系统中记录的日志数据。

因此，发行人尽管未保留与经营业务相关的日志数据，但不影响对采购业务量的监控和核算。

（二）发行人通过线下文档的形式记录和保留生产过程数据，生产过程可追溯

除了线上数据，发行人的生产过程数据还以项目采集信息表、项目周报、客户沟通邮件、客户验收确认文件等线下文档形式记录和保留。

项目采集信息表：为发行人进行音频采集时，对每个发音人信息的具体记录（如性别、年龄、国籍、来自省份或城市、语言或口音等）。

项目周报：为项目执行当中，各业务负责人对项目进展进行定期汇总并向总经理、主管业务的副总经理汇报，相关信息包括项目编号、内容、客户、项目负责人、交付时限、当前状态、已完工比例等。所发送周报均保存于发行人163企业邮箱系统内，为第三方系统，不存在篡改的可能性。

客户沟通邮件：对于数据资源定制服务，在项目执行过程中，发行人与客户针对文本设计、标产比、项目进度等，进行交流与确认。客户沟通邮件通过163企业邮件系统完成，不存在篡改的可能性。

客户验收确认文件：成品数据库完成质检向客户提交；客户验收通过后，向发行人发送验收确认文件，或在客户系统中完成验收。客户验收确认通过企业邮件系统或客户系统完成，不存在篡改的可能性。

此外，对于数据库产品，开发完成后，纳入发行人的产品列表，包括产品说明文档、产品文件等主要信息全部归档备份管理。

上述文件覆盖了发行人数据资源定制服务和数据库开发的生产过程，保证发行人的生产过程可追溯。

（三）进一步说明发行人与财务报表相关的内部控制是否健全并有效执行

1、发行人与财务报表相关的内部控制健全，并得到有效执行

综上所述，发行人的采集、标注数据服务采购与交付给客户的数据量紧密挂钩。发行人采集损耗率保持较低水平，保证了采集服务的采购结算量可通过对比拟交付数据量得到有效控制；发行人采取以交付数据量为基数，用标产比来折算有效工时的方式来核算标记工作量，而标产比为实际测试得到或由外部客户约定，保证了标记服务的采购结算量可通过对对比交付数据量并控制标产比的合理性而得到有效控制。

在上述业务特点的基础上，发行人对数据服务采购的核算建立了如下内部控制流程。发行人与财务报表相关的采购核算内部控制健全，并得到有效执行。

业务管理系统流程

财务系统流程

注：标*为关键控制点

1. 根据合同确认履约订单、一一对应建立项目并录入系统；对于数据库产品，根据产品开发决策立项、项目信息录入系统

2. 履约订单与项目一一对应，确保可按项目对收入、支出进行独立归集和核算；每个项目进行项目预算申请及审核，通过后立项成功

3. 项目负责人监控项目执行，并记录采集数据量、有效工时等关键业务信息

4. 依据项目过程记录的关键信息，将结算数据量、有效工时录入业务管理系统，进行支出报销申请；
复核人员复核关键业务信息，确保系统中数据服务采购量的完整、准确

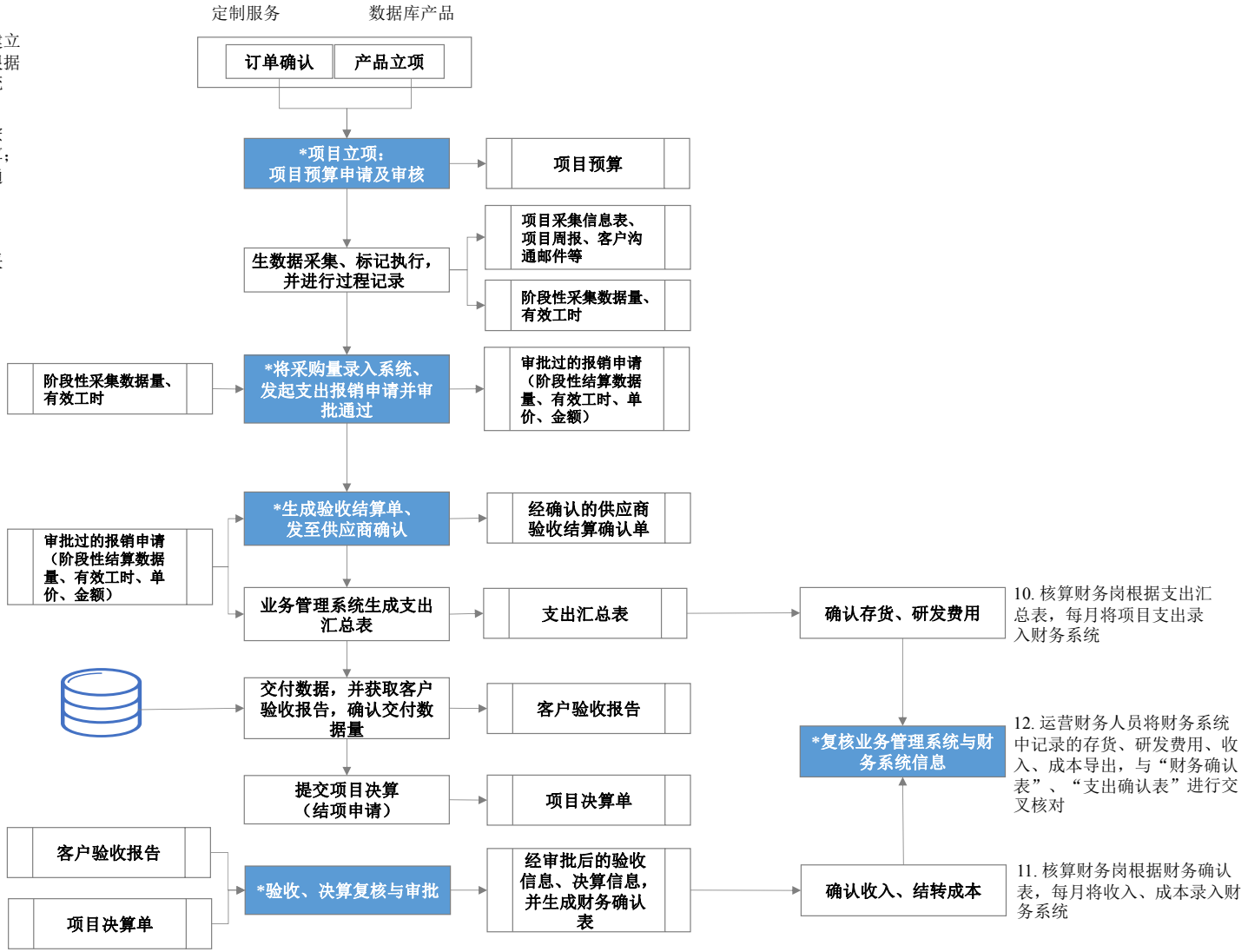
5. 根据支出报销审批信息生成验收结算单，并发送给供应商，由其确认后形成双方结算依据

6. 业务管理系统根据系统记录的经审批的支出报销信息，自动生成支出汇总表，按项目、任务记录每笔支出

7. 各阶段数据开发完毕后，形成数据库并交付给客户，客户确认交付数据量与数据质量无误后，出具验收报告

8. 项目完成且所有支出报销完毕后，申请结项

9. 复核人员复核验收、决算信息，并审批通过



针对上述内部控制流程中的关键控制点所实施的控制措施如下：

（1）流程2——项目立项：项目预算申请及审核

项目预算申请及审核，是发行人保证成本完整性、准确性的重要控制点，一般项目为三级审批制，重大项目为五级审批制。主要复核环节如下：

①将预计采集数据量与履约订单约定的交付数据量之间进行核对，保证预算采购量的真实、完整、准确；②对比同类项目间的采集产出比、标注产出比，保证预计采集数据量和有效工时的合理、完整；③综合考虑市场价格和该供应商的过往采购价格，对单价进行复核，保证预算中采集、标记单价的合理性；④根据履约订单金额、预算支出金额，测算预算毛利率，复核项目盈利能力的合理性。

（2）流程4——将采购量录入系统、发起支出报销申请并审批通过

针对生数据采集环节，复核人员针对采集数据量、结算数据量、交付数据量进行交叉比对，如存在重大差异，需由项目负责人员解释原因；如与预算对比出现重大差异，需经业务分管负责人、财务负责人审批：

①设计数据量对比采集数据量，是否总体基本一致；②音频文件个数对比采集数据量，是否一致；③采集数据量对比结算数据量，是否存在异常结算情况；④结算数据量对比预算采购量，是否存在较大偏离；⑤复核实际采集产出比，是否与商务约定或预算方案一致；⑥采集单价对比预算单价。

针对标注环节，复核人员针对标注产出比、有效工时进行合理性复核；如有效工时与预算对比出现重大差异，需经业务分管负责人、财务负责人审批：

①对比采集数据量，复核标注工作覆盖的数据量是否完整；②标记产出比对比预算，是否出现重大偏离；③有效工时的计算是否准确；④标记单价对比预算单价。

（3）流程5——生成验收结算单、发至供应商确认

①根据支出报销审批信息生成验收结算单，并发送给供应商；②由供应商确认后盖章，形成双方结算依据。

（4）流程9——验收、决算复核与审批

复核验收、决算信息，保证验收金额的真实、准确，项目支出的准确、完整。复核内容主要包括：

①交付数据量对比验收数据量；②验收量对比履约订单中的销售量；③验收单价对比履约订单中销售单价；④对比预算，复核项目决算中覆盖的环节、各环节的采购数量是否完整；⑤决算支出对比预算支出，出现重大偏离需提交项目决算分析表，解释原因。

（5）流程12——复核业务管理系统与财务系统信息

①项目采购支出、收入、成本等关键信息通过业务管理系统自动汇总并生成支出汇总表、财务确认表；②财务系统经过人工录入后，生成存货、研发费用、收入、成本等信息；③由专门人员将两个系统的记录导出后进行复核比对，如有差异，调查原因，并进行相应修订。

2、发行人进一步完善相关内部控制手段的措施

发行人尽管建立了较为完善的线下内部控制以保证数据服务采购的真实、准确、完整，但未能充分利用信息系统实现全部生产过程信息的记录和留存。为了进一步提升精细化管理水平、完善相关内部控制手段、提升内部控制的有效性，发行人将加强业务信息系统建设，实现对生产过程数据的自动化记录和保存。拟采取如下改进措施：

①进一步完善人员信息管理模块，包括管理人员、公司研发人员、供应商管理人员、被采集人、标记人员等。保证所有人员的个人信息在系统内的完整、真实、可追溯。

②进一步完善日志管理模块，涵盖数据资源开发的全流程，对于与经营业务相关的日志数据，做到集中处理、冗余备份、实时监控。

③新增数据分析模块，包括对业务数据、日志数据、用户行为数据等多维度数据做全面定时定量分析，提供对业务情况、财务情况的交叉验证能力。

④加强支出报销申请与项目决算申请中的复核机制。除目前执行的复核措施外，由项目负责人提交上述系统记载信息（包括但不限于采集数据量、测试标产

比等）与实际支出报销信息的匹配说明，由主管部门负责人及运营财务部审批。

⑤实现业务管理系统信息向财务系统自动传输，生成记账凭证，避免信息导入财务系统过程中人为操作失误的风险。

（2）发行人对劳务公司提供服务的人数和工作量如何进行复核和确认，定量分析实际服务人数和工作量与结算数据的对比情况、结算调整及差异原因，验收结算确认单中相关结算信息的依据和来源，如何保证实际采购业务与财务信息一致。

一、发行人对劳务公司提供服务人数和工作量进行复核和确认的方式

发行人对劳务公司提供服务的人数和工作量进行复核和确认的方式如下：

1、项目负责人员根据已完成的采集数据量/已发生的有效工时，在业务管理系统中填报支出申请；

2、部门经理对支出申请进行审核，交叉比对已完成数据量及其他业务信息；如果支出与预算出现较大偏差，需经业务分管负责人审批；

3、财务部门对支出申请进行审核，交叉比对业务信息；如果支出与预算出现较大偏差，需经财务负责人审批；

4、支出申请审批通过，业务管理系统生成供应商验收结算确认单；验收结算单发送给供应商，供应商确认后盖章并发回。

上述复核和确认的具体过程和方式，请参见本问题回复“第三部分/（2）/（三）进一步说明发行人与财务报表相关的内部控制是否健全并有效执行”中发行人内部控制的相关内容。

二、实际采集人数和标注工作量与结算数据量的对比情况、结算调整及差异原因

1、实际采集人数与结算数据量的对比情况

发行人有采集的单一化标注项目类型涉及的采集人数较多。与实际采集人数相比，结算人数略多于实际采集人数，主要是由于：（1）正式采集前，发行人

会根据项目情况采集少量数据进行测试，相应的测试人数不作为有效的实际采集人数，但该部分数据量需要与供应商结算；（2）由于发行人的采集设备和软件问题导致所采集的数据不符合要求，则不计入实际采集人数，但该部分数据量需要与供应商结算。

根据对发行人采集环节的核查情况，按报告期内各年数据资源定制服务和当期首次销售数据库产品收入合计占比50%以上的标准抽取项目，其合计结算人数分别为19,946人、21,360人、31,910人，超过实际采集人数的数量分别为334人、352人、812人，占实际结算人数的比例分别为1.67%、1.65%、2.54%。结算人数不低于实际采集人数，且相差较小。

2、实际标注工作量与结算数据量的对比情况

发行人按交付数据量×标产比来确定标记人员的有效工时，并按此与供应商结算相关数据服务费用，因此实际标注工作量与结算数据量之间不存在差异。

三、供应商验收结算单相关信息的来源

供应商验收结算确认单中的相关结算信息主要为单价、数量、服务费、管理服务费和结算总金额。其中，服务费=单价×数量，管理服务费=服务费×管理服务费率，结算总金额=服务费+管理服务费。

单价：请参见前文本问题“请发行人补充披露各类业务经营过程中采购和销售业务的定价方式和过程”之回复。

数量：①对于采集数量信息，发行人项目负责人员根据已完成采集的数据量对应的采集服务支出填报，经过部门经理、主管领导和财务部门复核审批；供应商核对确认后盖章。②对于标记工作量信息，项目负责人员根据已完成标注的数据量×标产比得到的有效工时数量填报，经过部门经理、主管领导和财务部门复核审批；供应商核对确认后盖章。

四、保证实际采购业务和财务信息一致的方法

（一）业务管理系统中记录的实际采购业务信息完整、准确

如前所述，发行人采集数据量、有效工时均与交付数据量直接挂钩，保证了

业务管理系统记录的实际采购业务信息是完整、准确的。

发行人在内部管理中，通过预算管理、支出管理和项目决算管理，从项目计划到实施完毕，始终通过交叉复核的方式监控采集、结算、交付数据量的一致性；通过测速管理、同类项目标产比比较分析等措施保证标产比的合理性；通过交付数据量的核对，保证有效工时计算的准确性。从而保证业务管理系统中的采购业务信息是完整、准确的。

（二）财务系统中的采购相关信息与业务管理系统中的采购业务信息一致

项目采购支出信息通过业务管理系统自动汇总并生成支出汇总表，人工录入财务系统后，生成存货、研发费用等信息。由专门人员将两个系统的记录导出后，在表格内进行总额的逐一比对，如有差异，调查原因，并进行相应修订。从而保证采购相关财务信息与业务信息的一致性。

保证采购业务与财务信息一致性的具体内控措施请参见本问题回复“第三部分/（2）/（三）进一步说明发行人与财务报表相关的内部控制是否健全并有效执行”的相关内容。

第四部分：保荐机构和申报会计师核查情况

保荐机构和申报会计师对上述事项进行核查并发表明确意见：

经核查，保荐机构和申报会计师认为：发行人未保留数据资源定制服务数据库的原因合理；发行人未保存与经营业务相关的日志文件，主要通过线下控制手段保证生产经营过程的可追溯；发行人与财务报表相关的业务采购内部控制健全并得到有效执行，并制定了进一步完善内部控制的有效措施；发行人实际采购业务与财务信息一致。

保荐机构和申报会计师的核查方式和核查过程：

（1）在没有保存业务数据的情况下如何对发行人数据服务费支出进行核查，如何保证核查结论的合理和准确；

如前所述，发行人定制化数据在交付之后不予保留，但数据库产品仍然保留；生产过程数据通过项目采集信息表、项目周报、客户沟通邮件、客户验收确认文

件等线下形式予以保留。保荐机构和申报会计师结合对发行人生产模式和产品特点的理解，通过留存数据对生产模式进行了验证，并通过抽样的方式对数据采集的完整性、标产比的准确性和稳定性等方面进行了详细复核比对；对终端被采集人和标记人员进行了电话抽样访问，并抽查了被采集人签字记录、供应商向被采集人和标记人员的付款记录等，核查发行人采购数据服务的真实性。

此外，保荐机构和申报会计师通过供应商函证、走访，确认发行人采购数据服务费真实、准确；查阅了发行人实际控制人、董监高等关联方银行流水，获取了供应商的相关说明和承诺，确认发行人不存在其他方代垫成本费用的情形，确认发行人采购数据服务费的完整性；对发行人报告期内的毛利率、采集产出比、标注产出比进行了详细的分析性复核，核查发行人采购数据服务费的完整性。

具体核查方式和核查过程如下：

一、关于发行人生产模式的有效性验证以及相关核查

保荐机构通过核查发行人的生产模式和采购数量确定机制的合理性，以及复核交付验收数据量、采集数据量、有效工时等主要指标的一致性，来验证数据服务采购的完整性。由于智能语音业务为报告期内发行人最主要的业务领域，报告期内各年收入占比均超过 80%，故本次核查以智能语音业务为主。同时对于数据库产品，由于数据库音频文件在发行人服务器上予以保留，因此保荐机构对数据库产品发音人音频文件本身的真实、有效性也进行了核查。

（一）对发行人采集和标注环节业务数据的核查

保荐机构和申报会计师对数据资源定制服务项目和数据库产品的抽查比例情况如下：

单位：万元

	2018 年	2017 年	2016 年
①抽查项目收入	5,726.43	3,511.87	2,950.80
②发行人总营业收入	19,265.77	11,907.09	8,422.86
占比（①/②）	29.72%	29.49%	35.03%
③发行人不含重复销售的数据库产品的营业收入	13,877.05	8,312.39	6,649.69
占比（①/③）	41.27%	42.25%	44.37%

④抽查项目成本（定制服务）+产品研发费用（数据库产品）	2,936.03	2,031.21	1,916.56
⑤发行人营业成本+产品研发费用	8,051.38	4,758.51	4,227.56
占比（④/⑤）	36.47%	42.69%	45.33%

注：（1）由于重复销售的数据库无成本或研发费用发生，故上表除列示抽查项目收入对总营业收入的占比外，还列示了对不含重复销售的数据库产品的营业收入的占比。（2）由于数据库产品的开发支出计入研发费用，故将数据库产品的研发费用与定制服务的成本合并计算。

1、采集环节的核查

（1）对有采集的单一化标注语音项目采集人员数量完整性执行的核查程序

核查方式：

①抽查发行人报告期内项目的客户合同或订单、客户验收文件、项目采集信息表、供应商结算单等业务和财务资料，统计核对客户验收发音人总人数、供应商采集总人数、供应商结算总人数等数据的匹配关系；

②统计分析供应商结算总人数与客户验收发音人总人数总人数的差额占前者的比例（“采集损耗率”），统计分析采集损耗率在报告期内的稳定性、合理性，验证分析采集支出的准确性、完整性；

③随机抽取发行人服务器上存储的数据库产品音频文件进行试听，验证音频文件有效性，同时对抽取的数据库产品发音人数量与该数据库产品的采集数据量对比分析。

有采集的单一化标注语音项目采集环节的核查情况：

单位：人

项目年份	抽查项目数量（个）	客户验收发音人数	供应商采集总人数	供应商结算总人数	采集损耗率
2018 年	86	30,285	31,038	31,581	4.10%
2017 年	55	20,826	21,350	21,730	4.16%
2016 年	28	19,355	19,837	20,198	4.17%

注：采集损耗率=（供应商结算总人数-客户验收发音人总人数）/供应商结算总人数。

2016 年至 2018 年，按当期数据资源定制服务和当期首次销售数据库产品收

入合计占比 50%以上并尽量覆盖各种语种、场景和采集设备[3]的标准，分别抽查了 28 个、55 个、86 个有采集的单一化标注语音项目。供应商采集总人数指采集数据量，与项目采集信息表（包括采集人员姓名、性别、年龄、地区、联系方式、采集文本编号、采集场景、采集日期等信息）总人数一致；抽查项目的采集损耗率区间为 4.10%至 4.17%；供应商结算总人数为发行人与供应商之间的结算数据量，样本结算总人数略大于供应商采集总人数。经核查，上述关键数据量指标基本一致，采集损耗率稳定且保持较低水平。

2016 年至 2018 年，在各年度随机抽取了 5 个存储在服务器上的数据库产品音频文件进行试听，经核查，服务器上存储的数据库产品音频文件真实、有效，抽样数据库产品发音人数量与该数据库产品的供应商结算数据量基本一致。

核查结论：

有采集的单一化标注语音项目的供应商结算总人数（结算数据量）不低于供应商采集总人数（采集数据量），供应商采集总人数不低于客户验收总人数（交付数据量），符合实际情况，不存在已采集未结算的数据量；同时采集损耗率在报告期内基本保持稳定，且维持较低水平，所需冗余采集量较小；发行人服务器存储的数据库产品音频文件真实、有效，抽样数据库产品的发音人数量与该数据库产品的供应商结算人数基本一致。

（2）对多样化标注语音项目采集字词、句数量完整性进行的核查程序

核查方式：

①抽查报告期内项目向供应商采集的字词、句数量，与发行人设计文本字词数量，以及交付客户并验收字词数量进行核对，计算供应商结算字词数对客户验收字词数的损耗率；

②抽样听取了部分多样化标注语音数据库产品的音频文件，并与其对应的语料文本内容核对一致性，验证多样化标注语音项目采集支出的完整性、真实性。

[3] 语种：选取中文、英语、德语、韩语、法语、西班牙语等主要语种/方言；场景：覆盖车载、桌面、录音棚等常见场景；采集设备：覆盖手机、音箱、麦克风等录音通道。

多样化标注语音项目采集环节的核查情况：

项目年份	抽查项目数量（个）	客户验收字词数	公司设计文本字词数	供应商结算字词数	采集损耗率	抽样音频文件是否与字词文本一致
2018 年	18	4,411,026	4,418,198	4,427,664	0.38%	是
2017 年	17	3,952,738	3,955,516	4,000,342	1.19%	是
2016 年	10	2,436,716	2,446,166	2,466,766	1.22%	是

注：采集损耗率=（供应商结算字词数-客户验收字词数）/供应商结算字词数

2016 年至 2018 年，按当期数据资源定制服务和当期首次销售的数据库产品合计收入占比 50%以上并尽量覆盖各种语种、发音人类型、细分标注环节[4]的标准，分别抽查了 10 个、17 个、18 个多样化标注语音项目。样本项目的公司设计文本字词数和供应商采购字词数比客户验收字词数略多，主要原因是公司为保证客户验收合格率和客户满意度，会额外设计并采集少量的备用字词录音。样本项目中，供应商采购字词数对客户验收字词数的损耗率为 0.38%至 1.22%，报告期内基本稳定。

由于部分多样化标注语音项目在启动前期，发行人会有发音人遴选的过程，即安排多个发音人按相同的设计文本录制小样，并最终确定一个发音人。参与遴选但未被选中的发音人，发行人仍需与供应商结算相应的发音人费用。由于上述原因，使得供应商结算的采购字词数比设计文本字词数略多。

核查结论：

对于多样化标注语音数据库，公司设计文本字词数大于客户验收的字词数，公司向供应商采集录音的字词数大于客户验收的字词数，符合实际情形，不存在已采集未结算的数据量；公司设计文本字词数对客户验收字词数覆盖率、供应商采购字词数对客户验收字词数覆盖率在报告期内基本保持稳定，且维持较低水平，所需冗余采集量较小；数据库音频文件真实、有效；多样化标注语音项目采集字词、句数量准确、完整。

[4] 语种：选取中文、英文、西班牙语、葡萄牙语等主要语种/方言；发音人方面：覆盖成年人、儿童及男性、女性；标注环节：覆盖了韵律、词性、音素边界标注、校对等标注环节。

2、标注环节的核查

（1）对自测标产比执行的核查程序

核查方式：

抽查智能语音业务部分项目，统计供应商完成标注、质检等业务环节的有效工作时长，计算标产比，现场观摩同类型项目标产比，验证标产比的合理性。

①有采集的单一化标注语音的核查情况：

项目年份	抽查项目数量（个）	客户验收小时数（小时）	供应商完成工时数（小时）	标注产出比
2018 年	86	42,700	129,080	3.02
2017 年	55	19,234	103,389	5.38
2016 年	28	21,121	59,690	2.83

注：标注产出比=供应商完成工时数/客户验收小时数。

2016 年至 2018 年，仍选取采集环节抽取的样本，样本平均标注产出比分别是 2.83、5.38 和 3.02。

2017 年抽样项目的平均标注产出比较高，主要因为当期抽取的 55 个项目中有 8 个自由对话项目，其中还包括泰米尔语、古吉拉特语、泰卢固语等标注难度较高的印度小语种自由对话项目。自由对话类项目在采集环节无预设文本，由发音人自由发挥，公司需要先进行人工文本生成，再进行文本与声音一致性检查、声音事件分类标注等标注环节，因此相对有预设文本类项目而言标注复杂度更高，标注用时更长，因此提升了 2017 年抽样项目的平均标注产出比。

保荐机构和申报会计师现场抽取了 5 个有采集的单一化标注语音中文项目，涵盖了中文和外语种，对标记工作独立进行了现场监督和测速，测得上述项目有预设文本类的项目标注产出比区间为 2.89~3.01，自由对话类项目标注产出比区间为 16.23~19.30，与抽查项目的有采集的单一化标注语音标注产出比基本一致，并且自由对话类项目标注难度较大，标注产出比较高。

核查结论：

保荐机构抽取了有采集的单一化标注语音项目，对其标注环节独立进行了监

督测速，与将测速结果与样本的标注产出比进行比对，两者基本一致，同时自由对话类项目标注难度较大，标注产出比较高，符合公司业务实际情况。

②无采集的单一化标注语音标注环节的核查情况

项目年份	抽查项目数量（个）	客户验收小时数（小时）	供应商完成工时数（小时）	标注产出比
2018 年	50	54,287	680,140	12.53
2017 年	50	24,378	295,423	12.12
2016 年	49	30,152	353,136	11.71

注：标注产出比=供应商完成工时数/客户验收小时数。

公司报告期内无采集的单一化标注语音项目没有数据库产品，只有数据资源定制服务收入，2016 年至 2018 年，按当期数据资源定制服务收入占比 50%以上并尽量覆盖各种语种、音频长度、主题[5]的标准分别抽查了 49 个、50 个、50 个无采集的单一化标注语音项目，标注产出比区间为 11.71~12.53，报告期内基本稳定，呈小幅上升趋势的原因是由于长句音频项目逐年增多，长句通常语速较快且变化更为复杂，人物角色和领域较为复杂，多为客服等领域，因而难度较高，增加了语音标注的难度。

保荐机构和申报会计师现场抽取了无采集的单一化标注语音的 3 个短音频项目、3 个长音频项目，按标注、质检等不同环节分别进行测速，汇总各环节标注时长后，除以加工的音频文件时长，得出该项目的标注产出比。分别测得上述短音频项目的标注产出比区间为 10.29~11.00，长音频项目的标注产出比区间为 14.42~15.00，与抽查项目的无采集的单一化标注语音标注产出比区间基本一致。

核查结论：

保荐机构和申报会计师抽取了无采集的单一化标注语音项目，对其标注环节独立进行了监督测速，与将测速结果与样本的标注产出比进行比对，两者基本一致，不存在较大差异。

[5] 语种：选取中文、英文等主要语种/方言；长度：长句音频、短句音频；主题：覆盖会议、演讲、客服、呼叫中心等常见主题。

③多样化标注语音标注环节的核查情况

项目年份	抽查项目数量（个）	客户验收小时数（小时）	供应商有效标注工时数（小时）	标注产出比
2018 年	18	504	32,803	65.10
2017 年	17	412	35,018	85.20
2016 年	10	292	19,415	66.57

2016 年至 2018 年，仍选取采集环节抽取的样本，样本的标产比分别为 66.57、85.20 和 65.10。

2017 年多样化标注语音项目标注产出比较高的原因是：抽样项目中有 2 个项目分别为西班牙语和法语项目，由于文本中夹杂着大量的英语单词，需要先就西班牙语或法语进行标注，然后对英语进行标注，相当于标注工作需要重复两遍进行；另外抽样项目中有 1 个为马来西亚语项目，由于马来西亚语难度较高，该项目由马来西亚籍标注员和中国标注员分别校对了一次，合计标注时间较长。综上所述，由于抽样项目中部分外语种项目的标注难度较高、标注时间较长，提升了 2017 年抽样项目的平均标注产出比。

多样化标注语音项目的细分标注环节较多，包括韵律、词性、音素边界标注、校对，整体标产比为各环节标产比的累加，环节越多，所需标注时长越长。同时，中文和外语种的标注难度有显著差异，中文的标产比明显低于外语种。

保荐机构和申报会计师现场抽取了 6 个目前正在执行的多样化标注语音项目，分别按韵律标注、词性标注、音素边界标注、校对 4 个环节进行了测速，测得标产比情况如下：

明细环节	中文项目平均标产比	外语种项目平均标产比
韵律标注	8.12	10.54
词性标注	12.37	15.16
音素边界标注	36.55	52.34
校对	9.89	13.25

前述抽查的多样化标注语音项目按包含的明细标注环节可主要分为下表列示的两大类。将各类被抽项目的标产比与测速项目进行对比，情况如下：

项目包含的明细标注环节	样本均值 2018 年	样本均值 2017 年	样本均值 2016 年	中文项目测速标产比	外语种项目测速标产比

韵律+词性+音素边界+校对	74.21	87.76	88.81	66.92 注	91.29 注
韵律+音素边界+校对	59.02	70.75	70.74	54.55 注	76.14 注

注：数值为前表中相应语种的对应环节标产比累加

由上表可以看出，抽查的多样化标注语音项目标产比均介于测速得到的中文和外语种标产比数值之间。

核查结论：

保荐机构和申报会计师抽取了多样化标注语音项目，对其标注环节独立进行了监督测速，与将测速结果与样本的标注产出比进行比对，两者基本一致，不存在较大差异。

（2）对客户确认标产比情况执行的核查程序

核查方式：

客户确认标产比的情形主要存在于部分计算机视觉、自然语言项目，确认方式为邮件。保荐机构和申报会计师抽查了部分项目，查阅客户的速率确认邮件，统计供应商完成标注、质检等业务环节的有效工时；并用客户确认的标产比×交付数据量，计算得出客户认可的工作小时，与供应商完成的有效工作小时数进行对比分析，验证标注环节支出的合理性、完整性。

核查情况：

项目年份	抽查项目数量（个）	客户验收工时数（小时）	供应商有效标注工时数（小时）	标注产出比
2018 年	92	53,879	71,137	1.32
2017 年	47	27,402	34,740	1.27
2016 年	20	24,944	28,624	1.15

注：客户验收工时数=客户验收数据量×客户确认的工作速率；标注产出比=供应商有效标注工时数/客户验收工时数。

2016 年至 2018 年，有部分计算机视觉和自然语言项目按客户确认的标产比进行结算，按此类结算项目合计数量 50%以上的标准各期分别抽查了 20 个、47 个、92 个项目，抽查的项目中，标注产出比区间为 1.15 至 1.32，标注产出比高于 1 的原因主要是由于客户确认的工作量不包含质检，供应商的有效工时包含了

质检。报告期内标注产出比总体较为稳定，逐年略有上升的主要原因是随着客户对数据标注要求越来越高，质检的抽查比例提升，使得每小时产出所需总的标注工作小时数量增加所致。

核查结论：

对于客户确认标产比的项目，与供应商结算的有效工时大于按客户确认速率计算的工作小时，主要原因为与供应商结算的有效工时中额外包含了质检工时，不存在未结算的有效工时；抽样项目的标注产出比在报告期内总体较为稳定，标注产出比逐年略有上升原因具有合理性。

（二）项目周报的核查

公司管理人员为了有效监督各业务主要项目工作进度，形成了各业务负责人每周向公司管理人员汇报项目进度的项目周报制度，保荐机构和申报会计师在前述“（一）对发行人采集和标注环节业务数据的核查”抽查的项目中又按项目数量抽取了三年合计 30%以上的项目，对抽查项目是否在项目周报中归集并汇报的情况进行了核查，核查情况如下：

单位：个

业务类型	抽样项目数量	周报覆盖数量	占比
有采集的单一化标注语音项目	62	62	100.00%
无采集的单一化标注语音项目	47	47	100.00%
多样化标注语音项目	17	17	100.00%
计算机视觉和自然语言项目	50	49	98.00%
合计	176	175	99.43%

经核查，发行人通过项目周报的形式对绝大部分项目进行了监控和管理；周报以邮件的形式由业务负责人向发行人总经理、主管业务的副总经理发出，发行人使用 163 企业邮箱系统，不存在人为篡改邮件系统的可能，相关项目周报的控制流程真实、可追溯。

（三）对最终采集对象和标记人员的相关核查

对最终采集对象和标记人员的相关核查情况请参见“问题 3/（2）对劳务公司提供服务的真实性进行核查的过程、程序和范围，包括但不限于核查个人的身

份登记记录、个人提供服务后确认的签字凭证、劳务公司向个人付款的相关凭证和完税凭据等。”

二、供应商函证、实地走访

1、对数据服务供应商的函证情况

保荐机构和申报会计师对发行人报告期内主要数据服务供应商进行函证，核查采购金额和往来余额的真实性和准确性，各年发函比例如下，所发函证均取得回函，回函金额不存在差异；

单位：万元

年度	数据服务供应商发函情况			数据服务供应商回函情况	
	采购金额	发函金额	发函比例	回函金额	回函比例
2018 年	7,352.66	6,232.50	84.77%	6,232.50	100.00%
2017 年	4,057.31	3,181.51	78.41%	3,181.51	100.00%
2016 年	3,396.15	2,691.43	79.25%	2,691.43	100.00%

2、对数据服务供应商的实地走访情况

保荐机构和申报会计师对发行人报告期内主要数据服务供应商进行了实地走访，所走访的数据服务供应商的对应各年数据服务采购金额占各年数据服务采购总金额的比例超过90%，访谈过程中，主要就双方合作时间、合作方式、提供服务类型、各年交易规模、发行人采购占其同类销售占比、验收结算方式、与外部劳务人员签订协议情况等情况进行详细了解，核查采购交易的真实性、合理性，以及采购模式的可持续性。

经核查，发行人对主要数据服务供应商的采购真实、合作商业背景合理。

三、关于关联方是否代垫成本费用的核查

1、保荐机构和申报会计师获取了发行人报告期内主要数据服务供应商的工商资料和报告期内税务合规证明，并穿透获得其股东和其他关键人员、关联公司名单，检查与发行人、发行人关联方及其董事、监事、高级管理人员等关键人员是否存在关联关系。

2、保荐机构和申报会计师获取了发行人报告期内主要数据服务供应商出具

的说明和提供的财务数据，核查其与发行人之间采购定价是否公允，采购交易是否符合商业实质并为供应商带来商业收益，确认双方之间及双方股东、董事、监事、高级管理人员、关键经办人员之间是否存在关联关系，是否存在提供劳务服务以外的其他交易和资金往来并为发行人代垫成本费用的行为，核查发行人成本费用的完整性。

3、保荐机构和申报会计师获取了发行人实际控制人及其配偶、发行人总经理控制、担任董事或高级管理人员、具有重大影响的企业（包括注销或转让的）及其董事、监事、高级管理人员出具的说明，确认该等企业及其董事、监事、高级管理人员与发行人供应商之间是否在交易、资金往来或其他利益安排，是否存在为发行人代垫成本费用的行为，核查发行人成本费用的完整性。

经核查，发行人主要数据服务供应商与发行人、发行人关联方及其董事、监事、高级管理人员等关键人员不存在关联关系，发行人主要数据服务供应商与发行人之间采购定价公允，不存在除提供劳务服务以外的其他交易和资金往来，不存在其他方为发行人代垫成本费用的情况，报告期内发行人成本费用完整。

四、对发行人毛利率、采集产出比、标注产出比等经营指标进行分析性复核

保荐机构和申报会计师对发行人报告期内对营业收入和营业成本的匹配性进行了重点核查，对毛利率、采集产出比、标注产出比、单位价格和单位成本等关键经营指标进行了详细的分析性复核，通过高管访谈、查阅项目资料和财务资料、了解行业情况以判断波动是否符合行业趋势等方式对存在较大波动的情形进行了分析，获得了合理解释。相关分析内容参见第二轮审核问询函回复问题 7、10 和第三轮审核问询函回复问题 5、6 之相关内容。

经核查，发行人营业收入和营业成本具有较高的匹配性，相关指标波动的具体业务背景原因合理，有效验证了数据服务采购的完整性。

（2）对劳务公司提供服务的真实性进行核查的过程、程序和范围，包括但不限于核查个人的身份登记记录、个人提供服务后确认的签字凭证、劳务公司向个人付款的相关凭证和完税凭据等。

一、对劳务供应商提供服务真实性的总体核查情况

1、对发行人采购劳务服务的穿行核查

保荐机构和申报会计师对发行人报告期内向劳务供应商采购劳务服务进行了穿行核查，从账面记录的数据服务费采购出发，获取采购框架合同、采购订单资料、经双方盖章确认的验收结算单、收到的发票、支付采购结算款项的银行回单，检查供应商名称、验收服务的内容、验收时间、结算数量、结算单价、结算金额、支付金额等信息与发行人记录是否一致，核查采购交易的真实性、准确性。2016 年、2017 年和 2018 年，核查金额占发行人当期营业成本与产品研发费用之和的比例分别为 60.46%、67.95%和 74.19%。

2、对劳务供应商进行函证和走访

保荐机构和申报会计师对劳务供应商进行函证和走访的情况请参见前文“二、供应商函证、实地走访”。

3、对终端劳务服务人员的核查

保荐机构和申报会计师对劳务公司提供服务真实性进行了一系列核查，包括对提供劳务的采集和标注人员电话访谈、查阅由采集人员签字的现场采集登记表、核查劳务公司向个人付款的相关凭证和完税凭据等。具体核查内容、核查范围、核查情况如下。

二、对终端劳务服务人员进行穿透核查的情况

（一）对采集和标注人员电话访谈的劳务真实性核查

为了验证劳务公司提供服务的真实性，保荐机构和申报会计师在前述“（一）对发行人采集和标注环节业务数据的核查”抽查的项目中又按项目数量抽取了三年合计 30%以上的项目，合计超过 1,100 名采集和标注人员进行了电话访谈，请访谈对象对其是否真实参与了某项目的录音/标注工作、工作完成后是否收到报

酬、报酬金额或报酬单价的区间进行了确认。

2016 年至 2018 年，对采集和标注人员电话访谈的核查情况如下：

单位：个/人

业务类型	电话访谈项目数量	电话访谈人数	电话访谈有效回复人数	确认比例
有采集的单一化标注语音项目	62	857	648	75.61%
无采集的单一化标注语音项目	47	145	145	100%
多样化标注语音项目	17	51	51	100%
计算机视觉和自然语言项目	50	69	69	100%

其中有采集的单一化标注语音项目的确认比例低于其他类型项目，主要是由于该类型项目主要为临时性、一次性录音，并且有的发音人参与录音的时间距今较久，有部分发音人在电话中不愿配合回答问题。对于未确认的被访谈对象，保荐机构和会计师进一步核查了现场采集登记表，其中有 190 位被访谈对象在登记表中留有签字，占有采集的单一化标注语音项目电话访谈未有效回复人数的 90.91%。

（二）劳务公司向个人付款的相关凭证核查

保荐机构和申报会计师对前述“一、对采集和标注人员电话访谈的劳务真实性核查”中抽查的项目除进行电话访谈以外，又向劳务供应商索取了对应项目的付款明细表和银行回执等向终端劳务人员付款的相关凭证，核查是否向劳务人员如实支付相关费用；同时将劳务供应商付款明细表与发行人报销明细表进行核对，查验支付给劳务人员的费用与发行人报销金额是否一致。

核查情况：

单位：万元

业务类型	抽样项目数量	发行人支付给供应商的服务费[注]	供应商支付给个人的费用	供应商向个人支付费用占比
有采集的单一化标注语音项目	62	267.78	266.58	99.55%
无采集的单一化标注语音项目	47	700.49	696.59	99.44%
多样化标注语音项目	17	262.30	241.36	92.02%
计算机视觉和自然语言项目	50	112.48	111.75	99.35%

合计	176	1,343.05	1,316.28	98.01%

注：“发行人支付给供应商的服务费”不含向供应商支付的管理服务费。

经核查，上述抽样项目劳务公司支付明细表明细加总金额与发行人报销明细加总金额基本一致，抽样项目发行人合计支付给供应商服务费 1,343.05 万元，供应商支付给个人的费用合计 1,316.28 万元，抽样项目的供应商向个人支付费用整体占发行人支付给供应商服务费的比例为 98.01%。

上述抽样项目电话访谈和供应商付款凭证的抽查比例情况如下：

单位：万元

	2018 年	**2017 年**	**2016 年**
①抽查项目收入	2,659.81	819.55	655.11
②发行人总营业收入	19,265.77	11,907.09	8,422.86
占比（①/②）	13.81%	6.88%	7.78%
③发行人不含重复销售的数据库产品的营业收入	13,877.05	8,312.39	6,649.69
占比（①/③）	19.17%	9.86%	9.85%
④抽查项目成本（定制服务）+产品研发费用（数据库产品）	1,250.85	489.64	373.96
⑤发行人营业成本+产品研发费用	8,051.38	4,758.51	4,227.56
占比（④/⑤）	15.54%	10.29%	8.85%

注：（1）由于重复销售的数据库无成本或研发费用发生，故上表除列示抽查项目收入对总营业收入的占比外，还列示了对不含重复销售的数据库产品的营业收入的占比。（2）由于数据库产品的开发支出计入研发费用，故将数据库产品的研发费用与定制服务的成本合并计算。

（三）对个人的身份登记记录、个人提供服务后确认的签字凭证的核查

报告期内大部分时间，发行人以项目采集信息表电子表格的形式留存被采集对象的相关信息。2018 年开始，为了进一步加强对采集活动的管理，发行人开始要求被采集对象在现场采集时，需在纸质版的项目采集信息表上签字作为现场采集登记表，并有专人负责对现场采集登记表进行收集和整理。

保荐机构和申报会计师对前述抽查的2018年86个有采集的单一化标注项目的现场采集登记表留存情况进行了核查，经核查，其中 60 个项目保留了现场采集登记表，被采集对象在现场采集登记表上留有签字。上述项目对有采集的单一化标注业务的收入占比、对成本和产品研发费用之和的占比分别为 37.82%、29.79%。

（四）完税凭据的核查

劳务供应商主要为发行人提供生数据采集、标记服务。

1、采集服务

关于生数据采集业务，劳务供应商未提供相应被采集人的完税凭据。存在以下两种情形：

（1）有采集的单一化标注语音项目、计算机视觉类项目和自然语音类项目

该等项目特点为采集人数众多、人均结算单价较低。报告期内，有采集的单一化标注语音项目人均结算单价分别是 183.67 元、141.79 元和 202.28 元，因此为发行人提供此类服务的个体单次收入较低，平均而言未达到每次 800 元的劳务报酬个人所得税起征点。

（2）多样化标注语音项目

该等项目特点为采集人数较少，人均结算单价较高，发音人单次收入平均而言达到了个人所得税缴纳标准。报告期内，发行人为采购此类采集服务向境内劳务供应商支付的数据服务费金额分别为 216.86 万元、251.34 万元、336.45 万元。假设发音人均需以全部劳务服务收入为应纳税所得额，按照 20%的税率缴纳个人所得税，据此估算，劳务供应商需代扣代缴的个人所得税总额分别不超过 43.37 万元、50.27 万元、67.29 万元。

2、标记服务

发行人智能语音类、计算机视觉类、自然语言类业务均采购了劳务供应商的标记服务，根据标记人员与劳务供应商之间的关系可以分为以下两类：

（1）因发行人部分客户有持续订单需求，希望有相对较为稳定、熟练的标记人员为其提供服务。因此，针对该部分需求，劳务供应商与部分标记人员签署了正式劳动合同，并为其依法代扣代缴了个人所得税。保荐机构和申报会计师取得了目前仍在职的 33 名标记人员在主管税务机关开具的《个人所得税纳税清单》。

（2）部分标记人员为临时性兼职人员，单次劳务服务收入较低，且未与劳

务供应商签署正式劳动合同，劳务供应商未能提供相关完税凭据。保荐机构和申报会计师对前述“一、对采集和标注人员电话访谈的劳务真实性核查”抽查项目中采购标记服务支出的数据服务费总额、标记人员数量进行了统计，测算出报告期内平均单个标记人员的年均劳务收入低于 8,500 元，金额较低，推断标记人员涉及个人所得税纳税义务的可能性较小。

此外，保荐机构和申报会计师取得了报告期内前五大数据服务供应商提供的确认文件、中国境内数据服务供应商主管税务机关出具的证明，根据上述确认文件和证明，报告期内，关于数据服务提供商为发行人提供劳务涉及向劳务人员发放的劳务费报酬，数据服务提供商已依法代扣代缴个人所得税，不存在因此而受到税务机关处罚的情形。

核查意见：

为核查劳务供应商提供服务的真实性，保荐机构和申报会计师对采购劳务服务进行了穿行核查，对主要劳务供应商进行了走访和函证，并对终端劳务服务人员提供服务的情况进行了核查。对于终端劳务人员的核查，保荐机构和申报会计师抽取了一定数量的劳务服务人员进行了电话访谈，对访谈对象为公司提供劳务服务并取得报酬相关事项进行了确认；抽查了部分项目现场采集登记表，对其留存和签字情况进行了核实；抽取了一定数量的项目检查供应商向劳务服务人员的付款情况并进行了逐笔核对和加总复核，核实供应商向终端劳务服务人员付款的真实性和金额的准确性；获得了部分劳务服务人员的完税凭证。综合上述核查情况，保荐机构和申报会计师认为，发行人的劳务供应商向发行人提供的服务是真实的。

（本页无正文，为北京海天瑞声科技股份有限公司对《关于北京海天瑞声科技股份有限公司首次公开发行股票并在科创板上市申请文件第四轮审核问询函的回复》之签章页）

北京海天瑞声科技股份有限公司

2024年7月1日

（本页无正文，为华泰联合证券有限责任公司对《关于北京海天瑞声科技股份有限公司首次公开发行股票并在科创板上市申请文件第四轮审核问询函的回复》之签章页）

保荐代表人：

葛 青　　姚玉蓉

华泰联合证券有限责任公司

2019年7月7日

保荐机构总经理关于审核问询函回复报告的声明

本人已认真阅读北京海天瑞声科技股份有限公司本次审核问询函回复报告的全部内容，了解报告涉及问题的核查过程、本公司的内核和风险控制流程，确认本公司按照勤勉尽责原则履行核查程序，审核问询函回复报告不存在虚假记载、误导性陈述或者重大遗漏，并对上述文件的真实性、准确性、完整性、及时性承担相应法律责任。

保荐机构总经理：

江　禹

华泰联合证券有限责任公司

2019年7月1日