上海燧原科技股份有限公司

（中国（上海）自由贸易试验区临港新片区业盛路 188 号 A-522 室）

首次公开发行股票并在科创板上市

申请文件的第二轮审核问询函

之回复报告

保荐机构（主承销商）

中信证券股份有限公司
CITIC Securities Company Limited

广东省深圳市福田区中心三路 8 号卓越时代广场(二期)北座

上海燧原科技股份有限公司

首次公开发行股票并在科创板上市

申请文件的第二轮审核问询函之回复报告

上海证券交易所：

贵所于 2026 年 4 月 16 日出具的《关于上海燧原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二轮审核问询函》（以下简称“《审核问询函》”）收悉，上海燧原科技股份有限公司（以下简称“公司”“发行人”或“燧原科技”）、中信证券股份有限公司（以下简称“保荐机构”或“保荐人”）、北京市中伦律师事务所（以下简称“中伦律师”或“发行人律师”）及毕马威华振会计师事务所（特殊普通合伙）（以下简称“毕马威会计师”或“申报会计师”）对反馈意见中的问题进行了落实，现对审核问询函回复如下，请审核。

如无特别说明，本回复报告中的简称或名词的释义与《上海燧原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”）中的相同。本问询回复中所列出的数据可能因四舍五入原因与根据本问询回复中所列示的相关单项数据直接相加之和在尾数上略有差异。

黑体（不加粗）：	**反馈意见所列问题**
宋体（不加粗）：	对反馈意见所列问题的回复
楷体（加粗）：	**对招股说明书（申报稿）的修改**
楷体（不加粗）：	对招股说明书（申报稿）的引用

目 录

1 关于营业收入

根据申报材料及问询回复：（1）报告期内，城市A二期项目已向发行人复购产品，与城市B项目已签署新投资合作协议；（2）Buy&Sell模式下，是否存在客户指定服务器供应商的情形，如有，说明该类模式下公司销售智算系统是否满足总额法确认收入的条件；（3）2024年末，随着库龄结构老化，发行人对无订单覆盖的二代产品进一步计提存货跌价准备。2024年末存货主要系用于成都高新电子智算集群业务的产品，因该项目执行周期较长，部分产品截至年末尚未完成交付，但已有明确销售合同支持，未计提跌价准备。

请发行人披露：（1）城市A二期项目向发行人复购产品的类型、数量、金额等情况，发行人与城市B项目签署新投资合作协议的具体约定，预计采购产品的类型、数量及预计实现的收入、利润、实现时间，结合智算中心项目向发行人复购产品及新投资合作情况，说明该等复购及新投资合作交易的可持续性；（2）Buy&Sell模式下是否存在客户指定服务器供应商的情形，如存在，说明该模式下公司销售智算系统是否符合总额法确认收入的条件；（3）结合2024年末有明确销售合同支持且未计提存货跌价准备存货的后续收入实现情况，说明2024年末相关存货的可变现净值是否低于其存货账面价值，以及2024年末该类存货跌价准备计提是否充分。

请保荐机构、申报会计师简要概括核查过程，并发表明确意见。

回复：

一、城市 A 二期项目向发行人复购产品的类型、数量、金额等情况，发行人与城市 B 项目签署新投资合作协议的具体约定，预计采购产品的类型、数量及预计实现的收入、利润、实现时间，结合智算中心项目向发行人复购产品及新投资合作情况，说明该等复购及新投资合作交易的可持续性

（一）城市 A 二期项目向发行人复购产品的类型、数量、金额等情况

城市 A 二期项目为城市 A 一期项目的复购项目，主要系公司与城市 A 政府达成战略合作关系，城市 A 作为国家八大算力枢纽节点之一，两次智算中心项目均采购公司产品。城市 A 一期项目系公司向客户 I 销售二代 AI 加速卡，客户 I 组装为智算系统，由其下属子公司作为项目业主方和运营方；城市 A 二期项目总集成商、业主方和运营方均为客户 E，2024 年 11 月发行人与客户 E 签订《产品销售合同》，复购产品为发

行人 313 套三代智算系统产品，产品合同含税金额为 28,060.45 万元，发行人已于 2024 年 12 月完成产品交付验收。

（二）发行人与城市 B 项目签署新投资合作协议的具体约定，预计采购产品的类型、数量及预计实现的收入、利润、实现时间

1、城市 B 项目签署新投资合作协议的具体约定，预计采购产品的类型、数量

2025 年 6 月，城市 H 人民政府（甲方）、合作伙伴 D（乙方）、发行人（丙方）签署《投资协议》。其中，合作伙伴 D 为城市 B 智算中心项目业主方的控股股东。

《投资协议》约定：乙方将在甲方所属园区投资城市 H 智算中心项目，同时乙方后续将在城市 H 工业园区注册项目投资建设及运营主体公司，项目总投资预计 5 亿元，预计投资建设算力规模不低于 4000P，发行人作为核心算力设备提供企业，发挥自身的技术优势，提供设备相关的技术支持。2025 年 7 月 10 日，乙方已于城市 H 注册成立合作伙伴 E。

2、城市 B 项目新投资合作协议预计实现的收入、利润、实现时间

城市 B 项目新投资协议预计实现收入约 1.8-2 亿元，具体时间及利润实现情况需结合未来公司的业务拓展战略和双方谈判情况考虑综合确定。一方面，公司未来将聚焦头部互联网客户群体。相较于其他类型客户，互联网客户对 AI 芯片具有单笔采购规模更大、复购持续性更强、年度需求稳步提升的特征，长期深耕头部互联网客户为公司核心发展战略。另一方面，发行人已参与多个智算中心项目，对智算中心项目的需求匹配、实施节奏、风险把控均具备成熟实操经验，发行人后续将结合市场和下游客户需求等情况合理推进新投资协议的产品落地。

（三）结合智算中心项目向发行人复购产品及新投资合作情况，说明该等复购及新投资合作交易的可持续性

城市 A 智算中心项目二期的复购，主要系发行人 2023 年向客户 I 销售 AI 加速卡，客户 I 在国家算力枢纽之一城市 A 地区建设了当地首个大规模国产化算力集群城市 A 一期项目并适配了发行人 AI 加速卡。该项目的实施一方面有力的支持了城市 A 地方算力基础设施建设，另一方面，形成了良好的首单市场示范效应，为城市 A 区域后续吸引其他业主方投资建设智算中心及公司进一步拓展当地市场奠定了坚实基础。

公司参与的城市 B 智算中心项目为城市 B 重点引进的算力项目，城市 H 人民政

府希望引入城市 B 智算中心项目的成功经验，复制到能源丰富的西部大省并形成规模效应，为城市 H 及全国的人工智能企业提供普惠绿色算力服务。

综上，发行人在智算中心落地方面积累了丰富的经验，已在算力领域建立起良好的市场口碑与客户认可度，为智算中心相关业务的持续拓展及客户复购可持续性奠定了有利基础。未来公司将重点支持头部互联网公司的业务需求，并结合实际经营情况择优布局具备实施条件的优质智算中心项目，智算中心的复购性落地将以项目质量、实施效益与商业可行性为前提，合理有序推进。

二、Buy&Sell 模式下是否存在客户指定服务器供应商的情形，如存在，说明该模式下公司销售智算系统是否符合总额法确认收入的条件

报告期内，公司 Buy&Sell 模式下不存在客户指定服务器供应商情形，具体如下：

首先，公司在 Buy&Sell 模式下向总集成商交付智算系统产品，双方签署的合同中未约定由客户指定服务器供应商的相关条款，总集成商仅对公司交付的智算系统产品履行验收义务；

其次，公司智算系统产品需依托服务器等硬件完成集成。公司新一代 AI 加速卡量产后，基于商务条件、历史合作情况等因素自主选择大型服务器厂商开展适配工作，适配完成的服务器机型纳入公司产品体系，公司已与多家主流服务器厂商建立适配合作，用以保障不同智算中心项目的客户需求；

最后，公司在取得智算系统产品订单后，公司再结合交付周期、交付区域及其他商务因素，综合选定一家或多家服务器厂商进行采购，确保最终交付的智算系统产品满足客户提出的技术指标与功能要求。

三、结合 2024 年末有明确销售合同支持且未计提存货跌价准备存货的后续收入实现情况，说明 2024 年末相关存货的可变现净值是否低于其存货账面价值，以及 2024 年末该类存货跌价准备计提是否充分

2024 年末，发行人对二代 AI 加速卡对应的在产品和库存商品中无销售订单/合同支持的部分足额计提存货跌价准备，对于已有明确销售合同支持的存货，则基于合同价格、相关税费测算其可变现净值不低于账面价值，故未计提跌价准备，具备合理性。针对后续期间未按原计划实现销售的二代卡相关存货，发行人已于 2025 年全额计提跌价准备，具体详见本问询回复之“4 关于存货”。

四、中介机构核查意见

（一）保荐机构、申报会计师进行了如下核查：

1、查阅城市 A 二期相关销售合同，了解相关销售合同明细情况，查阅城市 B 项目新投资合作协议，了解相关协议具体约定；

2、查阅行业研究报告、同行业公司信息披露文件以及访谈发行人管理层，了解行业经营模式、下游客户需求变化、同行业可比公司经营情况等，分析发行人预计实现收入的具体依据，以及复购交易的可行性；

3、获取发行人 Buy&Sell 模式下的销售、采购合同等相关业务凭证，询问发行人管理层，了解发行人该等业务模式流程及商业合理性；查阅了行业内其他公司的公开披露文件，评价公司会计处理是否符合企业会计准则的相关要求；

4、获取发行人 2024 年末二代产品已签订销售合同、商务项目的立项资料，了解发行人为销售预期而准备的存货具体情况；

5、询问发行人商务团队，了解相关项目存货的后续预计交付时点及可行性，了解后续销售情况，了解发行人于 2024 年末作出无需计提存货跌价准备的判断依据；走访成都高新电子集群项目智算中心现场，观察实际设备运行情况，检查验收单、销售发票等收入确认支持性文件；

6、询问发行人财务人员，了解存货进行可变现净值的测算过程、依据，获取并检查存货跌价准备计算表，结合企业会计准则评价存货跌价准备计提是否充分、审慎。

（二）经核查，保荐机构、申报会计师认为：

1、发行人拥有丰富的智算中心项目实施经验、良好的市场口碑与客户认可度，为智算业务持续拓展及客户复购奠定坚实基础。未来公司将聚焦头部互联网企业需求，结合经营实际择优储备优质项目，以项目质量、实施效益及商业可行性为核心，合理推进智算中心项目复购落地；

2、Buy&Sell 模式下，发行人不存在客户在指定服务器供应商的情形；

3、2024 年末，发行人对二代产品存货跌价准备的计提是充分、审慎的，符合《企业会计准则》的规定。对于有明确销售合同支持的存货，发行人基于合同价格等可靠依据判断其可变现净值高于存货账面成本，因而不计提跌价准备具有合理性；对于无

销售合同覆盖的存货，发行人已全额计提跌价准备，计提充分。

2 关于产品及业务

根据申报材料及问询回复：（1）英伟达软硬件生态训练的AI模型部署到非英伟达的软硬件生态均需进行迁移适配；原生GPGPU架构、非GPGPU架构（DSA架构）面对互联网应用场景均需进行针对性软件适配；（2）元宝、豆包、千问等AI对话类应用以及新的AI智能体应用对推理算力形成大规模需求；（3）公司第四代产品为训推一体AI芯片，目前未规模化量产出货，少量国产厂商进入腾讯供应链，相关产品主要用于适配难度及生态粘性相对较小的生成式AI类场景；（4）AI模型训练场景需要性能参数更高的AI加速卡，推理AI加速卡侧重面对海量用户的低延迟、高吞吐量和低功耗特性；报告期内公司主要销售推理卡产品，该产品毛利率低于同代际训练产品毛利率是行业惯例。

请发行人：（1）结合GPGPU与非GPGPU的分类、DSA与ASIC的差异、不同架构AI加速芯片与主流应用场景的适配性，说明发行人采用的非GPGPU（DSA）架构是否能够有效适配新模型、新场景；（2）用于训练、推理的AI加速芯片产品的价格、毛利率差异原因及技术难度差异情况；AI加速芯片应用场景及生态粘性的划分依据，用于大模型训练、推理的技术难度与市场规模；结合发行人推理产品比例高的形成原因、目前及未来产品预计在大模型训练、推理方面的应用及比例情况，说明发行人产品结构是否符合行业发展趋势。

请保荐机构简要概括核查过程，并发表明确意见。

回复：

一、结合 **GPGPU** 与非 **GPGPU** 的分类、**DSA** 与 **ASIC** 的差异、不同架构 **AI** 加速芯片与主流应用场景的适配性，说明发行人采用的非 **GPGPU**（**DSA**）架构是否能有效适配新模型、新场景

（一）**GPGPU** 与非 **GPGPU** 的分类、**DSA** 与 **ASIC** 的差异

1、GPGPU 与非 **GPGPU** 的分类

AI 芯片从硬件结构可分为 GPGPU 与非 GPGPU 两大类。

早期 GPU 芯片产品主要应用于图形渲染场景，硬件核心为专用于图形渲染和着色的处理单元模块。2008 年，英伟达将 GPU 芯片拓展至科学计算领域，

为进一步增加产品应用场景的泛化能力（并行计算的应用场景包括图形渲染、AI 矩阵计算、科学计算等多个领域），在产品中着重增加通用并行计算单元模块的数量（即 CUDA Core，可一次实现乘加法操作的最基础运算硬件单元）。2013 年，随着 AI 行业兴起，考虑提升 AI 矩阵运算效率，英伟达在其芯片产品中又加入了专用于 AI 矩阵计算的加速单元模块（即 Tensor Core，该硬件单元专用于 AI 矩阵计算，相应场景运行效率超过 CUDA Core）。根据官网公开资料，英伟达主流的 H100 芯片内部包含超过 16,000 个 CUDA Core 和超过 500 个 Tensor Core。在 AI 芯片领域，业界将兼顾 AI 计算、科学计算、仿真模拟等多场景算力需求，硬件冗余度较高、场景通用性较强的芯片称为 GPGPU 芯片。

非 GPGPU 架构主要包括 ASIC 架构和 FPGA 芯片。在 AI 领域，ASIC（Application-Specific IC）是一个广义概念，是针对 AI 计算这一特定领域进行优化的处理器架构，其核心在相同的芯片面积上，通过牺牲场景通用性来提升特定应用场景下的效率。与 GPGPU 架构不同，该架构弱化通用加速单元配置，强化 AI 计算专用加速单元，将有限的芯片面积及晶体管数量更多地配置予 AI 计算需要的加速单元。因此，在 AI 芯片领域，ASIC 不再兼顾 GPGPU 的多场景兼容属性，专用于 AI 计算场景，在硬件层面也对 AI 计算场景进行优化，进而使得加速效率更高。FPGA 芯片是另一种产品，该类芯片不需要专门设计，专门流片，厂商直接购买现成 FPGA 芯片并对其烧录编程实现特定功能的验证或小批量试产，并非主流。

2、DSA 与 ASIC 的差异

结合上文，AI 芯片领域，ASIC 芯片是一个广义概念。最严格意义的 ASIC 芯片等同于场景完全定制化，即完全 Application-Specific，例如加拿大 Taalas 公司，专门设计此类 ASIC 芯片，其 HC1 芯片直接将 Llama 特定大模型固化在芯片硬件架构，只能用于该模型推理应用，但效率极高。如更换模型算法，必须重新流片；国内部分端侧专用于图像识别、语音唤醒等传统 AI 模型的 ASIC 芯片也属于这一类别；DSA 架构（Domain Specific Architecture）是 ASIC 芯片概念的发展，其基于 AI 并行计算这一 Domain 进行加速优化，但并不固化在 AI 并行计算的某个特定场景或模型（即并不 Application-Specific，可适用于 Domain 内的多个 Application），这需要公司硬件架构既具备可编程性和通用性，兼顾

AI 计算领域的算法、模型以及用户场景迭代；又要确保 AI 并行计算的效率和性能。公司 AI 芯片采用 DSA 架构，舍弃了非 AI 并行计算领域的通用性，但具备满足 AI 并行计算领域的通用性。具体硬件架构上，公司芯片包含了多个专用于 AI 矩阵计算的 GCU-CARA 加速单元（类似英伟达芯片的 Tensor Core 单元），每个加速单元均可独立编程。

考虑 AI 算法和模型迭代，主流 AI 芯片厂商，尤其是云端领域所采用的 ASIC 路径通常为 DSA 架构。谷歌 TPU、亚马逊 Trainium、华为昇腾、昆仑芯、寒武纪以及发行人的产品均属于 DSA 架构的 AI 芯片。

（二）不同架构 AI 加速芯片与主流应用场景的适配性，采用的 DSA 架构是否能有效适配新模型、新场景

1、不同架构 AI 加速芯片与主流应用场景的适配性

在 AI 领域，GPGPU 架构、DSA 架构以及固定不可编程的狭义 ASIC 芯片在场景适配性上存在显著差异：

GPGPU 架构通过强化通用并行计算单元形成了成熟的通用编程生态，可覆盖 AI 计算、科学计算、图形渲染等多个场景，但针对 AI 计算这一特定领域场景的效率存在一定损耗。

固定功能完全定制化的 ASIC 芯片因硬件电路完全固化，仅支持预设算子、模型与计算精度，无可编程与迭代能力，对当前 AI 主流应用场景的适配性较差，亦无大规模商用部署的案例。

公司采用的 DSA 架构在芯片设计阶段即对 AI 领域的核心算法、模型和应用场景进行加速优化，将所涉及的核心计算、互联和存储优化特性“硬化”至芯片架构，兼具专用加速效率与可编程迭代能力，适配性显著优于狭义 ASIC 架构，同时在 AI 场景下的效率表现优于通用 GPGPU。

基于高盛全球投资研究部的模型预测，在 AI 服务器中的非 GPGPU 架构（即 DSA 架构）AI 芯片的出货占比将呈现明确上升趋势，其与 GPGPU 架构的比例预计将从 2025 年的 38%：62%，增长至 2027 年的 45%：55%。不断增长的市场中 DSA 架构份额不断扩大，说明其对主流应用场景的良好适配能力。

2、采用的 DSA 架构是否能有效适配新模型、新场景

当前全球主流 AI 大模型均基于 Transformer 底层架构构建，该底层架构的 AI 大模型运行过程主要包含数据处理、注意力计算、内容生成三大核心计算环节。上述计算流程在硬件层面，对应 AI 芯片计算单元、存储单元和互联单元构成的微架构体系；在软件层面，对应支撑硬件高效运行的编程模型、计算加速库、通信加速库等基础软件栈。软硬件协同配合是共同实现 AI 模型的高效训练与推理，是决定 AI 芯片性能与生态适配能力的核心关键。

基于上述技术逻辑，具备可编程架构、可通过软件栈灵活调度硬件资源的 DSA 架构芯片能够全面覆盖 AI 模型运行所需的各类核心单元，包括支持 FP32/BF16/FP16/FP8 等精度的计算单元、SRAM 高性能存储单元、高速数据搬运单元以及 GCU-LARE 等高速互联接口，配套 TopsRider 软件栈可充分调度硬件算力，实现软硬件协同优化，使 AI 芯片具备 AI 模型的运算能力，并具备持续适配新模型、新场景的基础能力。

产业实践中，新兴应用场景的诞生本质是新业务需求驱动下衍生出新的大模型算法，新场景的核心算力需求最终均需通过对对应新模型的高效支撑来实现，因此 AI 芯片对新场景的适配能力核心是对新模型的适配能力。当一款全新的 AI 算法出现，外部客户的开发团队或公司的开发团队均可通过自研的驭算 TopsRider AI 计算及编程软件平台，开发其所需要的计算算子或通信算子并集成到加速库当中，这些算子在芯片内进行一系列的数据处理和计算，最终输出目标结果。基于这一过程，公司第三代和第四代产品能够快速适配支持最近推出的 DeepSeek V4 Pro/Flash、阿里 Qwen、混元 Hy3 preview 等新模型。以报告期内公司第三代 AI 加速卡 S60 适配 DeepSeek R1 为例，S60 于 2024 年 7 月量产，DeepSeek R1 于 2025 年 1 月发布，公司研发团队通过模型解析与软硬件协同适配，完整实现 INT4/INT8 低精度量化、MLA 多头潜在注意力、标准 Attention、All-to-All 全互联通信及分布式并行等全部核心能力，不仅完成了新模型适配，更实现了该模型在单机 8 卡、多机 16 卡及 32 卡集群等规模化推理新场景下的稳定部署。同时，公司可结合客户不同行业、不同业务形态的新场景需求，通过联合软件优化快速完成适配，进一步提升对新兴 AI 应用场景覆盖效率。

综上，从技术架构原理与实际落地案例均可证明，公司采用的 DSA 架构

AI 芯片具备优异的迭代适配能力，能高效适配各类新发布大模型，也可依托模型适配的核心技术能力，全面适配 AI 领域各类新兴应用场景。

二、用于训练、推理的 AI 加速芯片产品的价格、毛利率差异原因及技术难度差异情况；AI 加速芯片应用场景及生态粘性的划分依据，用于大模型训练、推理的技术难度与市场规模；结合发行人推理产品比例较高的形成原因、目前及未来产品预计在大模型训练、推理方面的应用及比例情况，说明发行人产品结构是否符合行业发展趋势

（一）训练、推理的 AI 加速芯片产品的价格、毛利率差异原因及技术难度差异情况

1、训练产品的性能要求、技术难度和硬件配置高于推理产品

AI 训练卡与 AI 推理卡在应用场景、设计目标与核心诉求存在本质差异，主要系训练卡作为大模型研发的基础设施，追求极致性能与高稳定性，而推理卡作为大模型服务的交付载体，更注重部署成本与场景适配效率。技术难度上，呈现 AI 训练卡高于 AI 推理卡的行业特征，在硬件设计、技术实现等层面均存在差异，具体如下：

对比维度	AI 训练卡	AI 推理卡
设计目标	追求极致性能与集群协同稳定性，优先保障训练效率与精度收敛	追求极致性价比与场景适配效率，支撑下游客户业务的低延迟、高并发交付，优先平衡成本与业务体验
硬件特性	算力：高算力，数据精度以 FP8/BF16 为主，具备一定的 FP32 算力； 显存：大容量、高带宽； 互联：高互联，低延迟，对卡间互联带宽有较高要求	算力：较高算力，数据精度以 FP8 为主，未来向 FP4 等更低精度发展； 显存：根据不同模型参数量差异化配置。千亿以上参数大模型要求大容量、高带宽；百亿及以下参数大模型要求中等容量及带宽； 互联：千亿以上参数大模型要求高速互联，低延迟，其他参数规模模型对互联能力要求偏低
技术难度	技术难度高，核心在于： 1、对软件生态成熟度要求高，要求算子库、算子算法开发工具链、调试调优工具链等完整 2、对硬件性能要求高，包括单卡性能和集群性能，需对多卡互联通信能力进行极致优化 3、对软硬件协同稳定性要求高，要求超低故障率	技术难度中等，核心在于： 1、对软件生态成熟度要求弱于训练场景，关注硬件算力供应商的技术支持高响应度和适配优化能力 2、对硬件性价比要求较高，需针对具体应用场景提供极致性价比部署方案； 3、对稳定性和超低故障率的容忍度高于训练场景

2、从发行人角度，相同代际的训练产品单价和毛利率高于推理产品

报告期内，发行人不同代际不同类别产品毛利率对比如下：

单位：万元

公司	产品	2025 年度		2024 年度		2023 年度	
		单价	毛利率	单价	毛利率	单价	毛利率
发行人	二代训练卡	-	-	3.61	73.67%	1.90	44.62%
	二代推理卡	-	-	1.84	58.76%	0.79	9.27%
	三代推理卡	1.31	34.53%	1.39	39.86%	-	-
	四代训推一体模组	5.03	-40.06%	-	-	-	-

从上表可见，报告期内发行人第二代产品存在训练卡和推理卡两类，训练卡单价及毛利率均高于推理卡。以训练卡T20和推理卡I20为例，不同加速卡的芯片均包含数个自研的第二代GCU-CARA加速单元，均基于同一工艺制程制造，处于同一代际。但鉴于训练场景客户要求性能、推理场景客户要求性价比，不同功能加速卡的内含芯片需要基于客户需求分别设计，芯片和板卡规格均存在差异。

显存方面：训练卡芯片包含4颗存储颗粒，推理卡芯片仅包含2颗存储颗粒，训练卡芯片的内存带宽是推理卡芯片的2倍；

互联方面：训练卡配备了公司第二代GCU-LARE互联接口，可实现卡间高速互联，二代推理卡并未配置相应互联接口；

功耗方面：训练卡芯片性能更强、PCB板载的电子元器件更多、功耗更大，配置了液冷冷板方案；推理卡功耗更低，仅配置风冷方案。

除硬件成本差异外，国产AI加速卡的定价主要以客户整体部署的性价比为核心考量，导致训练卡与推理卡的毛利率存在差异。推理卡在互联、存储等关键配置上进行了裁剪，导致在性能上存在受限。尤其是在大参数模型推理场景下，受互联、存储等性能的制约，客户往往需要部署更多数量的推理加速卡，并同步增加服务器等配套设备的数量以满足业务需求，进一步推高了客户整体部署成本。客户因此对推理卡的单卡性价比提出了更高要求，压低了推理卡单价。这是推理卡的毛利率会显著低于同代际训练卡或训推一体产品的本质原因。

综上，训练卡因硬件配置更高、物料成本更高、技术复杂度更高，定价高

于同代际推理卡具有合理性。同时，国产AI加速卡的定价主要以整体部署的性价比为核心考量，推理卡因单卡性能受限导致客户整体部署成本上升，客户会进一步压低单价，毛利率低于训练卡亦具有合理性。

3、同行业公司对比来看，训练产品单价和毛利率同样高于推理产品

报告期内，同行业可比公司不同产品单价和毛利率具体情况如下：

单位：万元

公司	产品	2025 年度		2024 年度		2023 年度	
		单价	毛利率	单价	毛利率	单价	毛利率
寒武纪	智能芯片及加速卡（训练/训推一体为主）	5.52	55.22%	3.01	56.69%	2.74	60.63%
摩尔线程	AI 智算板卡（训练/训推一体为主）	5.80	90.72%	6.02	90.70%	-	-
沐曦股份	训推一体 GPU 板卡	3.89	56.21%	4.69	63.50%	5.69	65.48%
	智算推理 GPU 板卡	0.30	6.37%	0.68	10.09%	0.67	5.37%
天数智芯	训练产品	3.61	64.22%	3.86	60.24%	3.18	53.21%
	推理产品	1.17	39.19%	1.02	46.67%	0.80	35.77%

数据来源：可比公司年报及招股说明书等。
注 1：2025 年度数据中，发行人、寒武纪、天数智芯为 2025 年全年数据，摩尔线程因 2025 年报未披露 AI 智算板卡单独数据，为 2025 年 1-6 月数据，沐曦股份因 2025 年报未披露产品单价，为 2025 年 1-3 月数据。壁仞科技未披露销量，无法计算单价，上表未列示。
注 2：寒武纪智能芯片及加速卡单价和毛利率系通过其年度报告披露云端产品线、边缘端产品线和智能计算集群系统合计收入除以同期智能芯片及板卡销售量粗略估算。天数智芯 2025 年训练产品和推理产品销售量系根据年度业绩报告披露销售量增长率、招股说明书披露 2024 年度销售额和销售单价进行粗略估算。

随着AI模型从传统AI模型向AI大模型发生技术范式转变，尤其是LLAMA、DeepSeek等开源大模型推出，自研AI大模型数量迅速收敛，同时AI大模型推理需求爆发。AI大模型推理分为预填充和解码两个场景，其中解码场景的需求体量更大，相较于预填充场景更需要AI加速卡具备大显存和内存带宽。考虑客户对推理场景的性价比要求，国内外AI加速卡厂商纷纷推出物料方案成本更低、性价比更高的推理产品方案，如上表沐曦股份的智算推理板卡、天数智芯的推理产品、发行人的三代推理卡和英伟达早期的L20均属于这一方案产品。鉴于上述相关产品未采用价格更贵的物料方案，该情况与公司二代训练卡和推理卡不同，相应专用推理卡物料成本进一步压缩，单价低于同期的训练/训推一体产品具有合理性。由于训练/训推一体产品的配置更高、性能更强，同时客户考虑自身部署成本压低推理卡价格，相应训练/训推一体产品毛利率更高具有合理性。

（二）AI 加速芯片应用场景及生态粘性的划分依据，用于大模型训练、推理的技术难度与市场规模

AI 加速卡应用场景及生态粘性的划分，核心依据为所服务的 AI 模型类型以及与 AI 加速卡软硬件生态的绑定程度，主要划分为传统 AI 模型、搜索广告推荐类 AI 模型、生成式文字及多模态 AI 大模型三大类，不同场景下的技术难度、与生态粘性情况具体如下：

场景类别	技术要求
传统 AI 模型	**生态粘性：**传统 AI 模型参数量较小，一般为百万到千万级参数规模，但涉及的模型种类繁多，不同模型差异大，需针对特定场景下的特定模型进行深度优化，迁移成本高，生态粘性较强 **技术难度：**以推理场景为主，计算节点以单卡为主，要求加速卡具备虚拟化特性（即单张加速卡可以同时部署多个传统 AI 模型），并满足高并发、低延时的需求；对硬件计算精度要求中等，主要需求为 FP16 精度的算力，少部分情况需要支持 FP32 精度算力 **需求特点：**对 AI 加速卡性价比敏感，强调低成本、高性价比
搜索广告推荐类 AI 模型	**生态粘性：**搜索广告推荐类 AI 模型参数量中等，一般为百亿级参数规模，相应模型和算子需要根据场景需求进行深度优化，平台迁移成本高，生态粘性较强，是互联网厂商创收业务的核心技术基础 **技术难度：**以推理场景为主，对 AI 计算的稳定性、高并发、低延迟和高精度要求严苛，芯片架构及软件栈需针对性地对 Top K Elements 等搜索广告推荐类模型常用的排序算子进行特定优化；对计算精度要求较高，硬件侧要求同时支持 FP32 及 FP16 精度算力 **需求特点：**对 AI 加速卡性价比敏感，强调低成本、高性价比
生成式文字及多模态 AI 大模型	**生态粘性：**生成式文字及多模态 AI 大模型参数一般为千亿到数万亿参数规模，相应模型属于新生应用，与客户已有业务场景融合度、生态粘性和适配难度目前低于前两类场景，但不断加深，该类模型的算力需求目前处于供不应求状态 **技术难度：** （1）训练：对软件生态成熟度要求高，要求算子库、算子算法开发工具链、调试调优工具链等完整。强调单卡性能和集群性能，需具备极强且稳定的多卡互联通信能力，并对计算精度要求较高； （2）推理：推理场景更注重性价比、高并发的数据吞吐处理能力、高算力和低延迟。一般要求加速卡原生支持 FP8 等低精度数据格式，同时支持超节点，对训练和推理性能都有显著提升。推理场景还可进一步细分为两类： ①对话等大流量高并发场景下，一般采用完整参数大模型，即千亿到万亿参数规模，该场景分为预填充和解码两个阶段，算力需求以解码阶段为核心。解码阶段偏重 AI 加速卡大显存容量、高显存带宽；预填充阶段偏重 AI 加速卡较高算力密度。该场景下通常需要单台到多台 AI 服务器完成单一模型的推理部署，要求 AI 加速卡具备高互联带宽性能 ②搜索、语料处理、B 端应用等中低并发场景，一般采用蒸馏模型，一般为百亿及以下参数规模。该场景下通常由单卡完成蒸馏大模型的预填充和解码阶段，对算力、显存容量、显存带宽等要求中等，互联能力要求相对偏弱，对 AI 加速卡的性价比敏感

	③此外，对于 AI 对话场景也存在针对输入问题不同复杂度，考虑性价比自动识别调用完整参数大模型或蒸馏模型的业务场景优化 **需求特点：**AI 推理卡更加注重低成本、高性价比；AI 训练卡更加注重高算力、高互联和高稳定性，以及软件栈的算子完整性

市场规模方面，传统 AI 模型和搜索广告推荐类 AI 模型应用市场是当前人工智能市场的主要构成，未来需求稳步增长。根据 IDC 发布的《全球人工智能和生成式人工智能支出指南》，2024 年中国人工智能市场投资总规模为 355 亿美元，其中非生成式 AI 市场（主要为传统 AI 模型和搜索广告推荐类 AI 模型市场）投资规模约为 290 亿美元，占比为 81.8%，预计在 2029 年将达到约 656 亿美元，期间复合增长率为 17.7%；生成式文字及多模态 AI 大模型应用需求则呈现高速增长态势，尤其是随着 AI 对话、AI 多模态应用以及 Open Claw 等 AI Agent 应用的爆发，该类模型场景的算力需求持续提升。根据 IDC 发布的《全球人工智能和生成式人工智能支出指南》，2024 年中国生成式 AI 投资规模约为 65 亿美元，预计在 2029 年将达到约 458 亿美元，5 年复合增长率为 48.0%。

（三）结合发行人推理产品比例高形成原因、目前及未来产品预计在大模型训练、推理方面的应用及比例情况，说明发行人产品结构是否符合行业发展趋势

1、发行人推理产品比例较高的形成原因

报告期内，公司 AI 加速卡及模组中推理产品占比较高，主要系公司总体采用“训练产品逐步探索，推理产品迅速推进”的产品迭代路径。该策略是基于训练领域技术复杂度与系统要求极高，而推理场景商业化落地更快，能快速实现自我造血以反哺训练产品持续投入的务实选择。

公司第三代产品立项时，生成式大模型仍停留在小参数阶段且未大规模应用，国内 AI 应用市场主要为传统 AI 模型及搜广推 AI 模型，因此公司基于第三代芯片仅推出了面向推理场景的 S60 推理加速卡，未推出同代训练卡。

此外，推理需求爆发也验证了公司产品路线选择。根据灼识咨询数据，2025 年互联网行业 AI 加速卡需求中推理占比约 48%，非互联网行业推理需求占比高达约 70%，公司推理产品比例较高契合当前市场需求特征。

2、目前及未来产品预计在大模型训练、推理方面的应用及比例情况

报告期内，公司产品主要用于客户推理场景。但随着 AI 技术演进，大模型参数量不断提升，大参数大模型的推理应用对显存带宽、卡间互联、低延迟提出新的要求，类似公司第四代的训推一体产品将逐步成为行业主流，一方面训推一体产品可复用硬件架构、软件栈与生态，大幅降低研发、交付与运维成本；另一方面客户智算中心普遍训练与推理混合部署，训推一体产品可以提升资源利用率与场景适配性。

公司后续代际产品以训推一体产品为主，第四代产品是公司首款训推一体产品，主要面向生成式文字及多模态 AI 大模型的场景推理应用，是目前国内 AI 芯片中需求增速最快的领域，也是国内厂商在互联网厂商中的主要应用领域。训练领域主要包括基础模型预训练领域和模型微调领域（利用自身专有数据对完成预训练的基础 AI 模型微调训练），其中基础模型预训练是主要需求。国产 AI 加速卡在预训练领域的应用正从中小规模参数模型向中大规模参数模型加速推进，万亿参数大模型预训练市场目前仍由国际厂商产品主导。训练领域，公司应用持续推进，已基于第四代训推一体产品 L600 与客户联合完成了基于多个中小参数模型的预训练、微调和强化学习工作，并正在推进基于大参数系列模型的预训练工作。

未来，随着公司第四代训推一体产品 L600 规模化量产及超节点系统的交付，公司将持续拓展训练领域的应用，但受行业需求结构、供应链与技术迭代节奏影响，推理场景仍将是公司未来训推一体产品的核心应用场景，训练场景应用比例将稳步提升。

3、说明发行人产品结构是否符合行业发展趋势

首先，公司推理产品收入占比高与同行业可比公司存在差异具有合理性。

报告期内，公司与可比公司训练/训推一体加速卡及模组产品占当期 AI 计算加速卡及模组产品收入的比例如下：

公司	2025 年度	2024 年度	2023 年度
发行人	1.15%	0.54%	17.65%
寒武纪	未披露训练和推理产品细分数据，根据单价和公开披露信息，其报告期内云端产品线主要为训练/训推一体产品		
摩尔线程	未披露训练和推理产品细分数据，根据单价和公开披露信息推测，其报告期内		

	AI 智算板卡主要为训练/训推一体产品		
沐曦股份	98.73%	99.32%	52.92%
天数智芯	63.27%	72.89%	82.86%

数据来源：上市公司招股说明书和年度报告。

报告期内，发行人产品结构以推理产品为主，训练/训推一体产品收入占比低于可比公司，主要系公司第三代产品立项时，生成式大模型仍停留在小参数阶段且未大规模应用，国内 AI 应用市场主要为传统 AI 模型及搜广推 AI 模型，公司基于第三代仅推出了面向推理场景的 S60 推理加速卡，未推出同代训练卡。同时，第四代训推一体产品虽已经完成流片，但截至本回复报告出具日尚处于规模化量产阶段，尚未形成大规模收入贡献。

其次，公司产品迭代与国际主流厂商路线基本一致。

从行业主流产品路线看，英伟达的产品布局呈现高端训推一体产品面向超大规模模型训练，同时同步推出高性价比推理产品的特点。英伟达先后推出 H100、H200、B200、B300 等高端训练/训推一体产品面向超大规模模型训练，同时推出 H20、L20 等主打性价比推理场景的产品。公司 S60 推理卡主打高性价比推理场景，L600 训推一体产品面向大模型训练及推理。公司产品定位、场景分层逻辑与国际主流厂商路线基本契合，符合行业发展趋势。

综上，发行人当前以推理加速卡为主、同时稳步推进训推一体产品落地的产品结构，契合当前国内 AI 算力市场的阶段性需求特征，也与同业公司的产品布局方向一致，且长期产品路线与国际主流厂商的行业主流路线相契合，符合 AI 芯片行业的发展趋势。

三、中介机构核查意见

（一）保荐机构进行了如下核查：

1、访谈公司技术人员并查阅相关行业报告，了解 GPGPU 与非 GPGPU 的分类、DSA 与 ASIC 的差异情况，以及不同架构 AI 加速芯片的适配能力，分析 DSA 架构对于新模型、新场景的适配能力；

2、获取发行人收入成本明细表，查阅公司同行业可比公司招股说明书、反馈回复、定期报告等资料，了解公司与可比公司主要产品单价与毛利率情况；

访谈公司管理层，了解 AI 加速芯片应用场景及生态粘性的划分依据，公司推理产品比例较高的原因，目前及未来产品预计在大模型训练、推理方面的应用及比例情况；查阅相关行业报告，了解 AI 加速芯片用于大模型训练、推理的技术难度和市场规模；

（二）经核查，保荐机构认为：

1、考虑 AI 算法和模型迭代，主流 AI 芯片厂商所采用的 ASIC 路径通常为 DSA 架构。云端领域，AI 芯片市场基本可以分为 GPGPU 架构和 DSA 架构两大阵营。公司采用的 DSA 架构 AI 芯片具备优异的迭代适配能力，能高效适配各类新发布的大模型，也可依托模型适配的核心技术能力，全面适配 AI 领域各类新兴应用场景。

2、训练产品的性能要求、技术难度和硬件配置高于推理产品，无论从公司自身产品比较，还是同行业产品比较，训练产品单价和毛利率高于推理产品符合行业规律，具有合理性；AI 加速卡根据应用场景及生态粘性可划分为传统 AI 模型应用、搜广推 AI 模型应用和生成式及多模态 AI 大模型应用；大模型训练的技术难度更高，但大模型推理的市场规模更大；发行人报告期内推理产品收入占比较高具有商业合理性。随着 AI 技术演进，发行人基于趋势后续主要推出训推一体产品，生成式 AI 大模型推理场景仍将是公司未来训推一体产品的核心应用场景，训练场景应用比例将稳步提升。

3 关于与腾讯科技的关联交易

根据申报材料及问询回复：（1）报告期内，经过腾讯与发行人协商，将AVAP模式转为直接销售模式，腾讯向发行人支付部分预付款用于其上游供应链采购；（2）公司目前对潜在客户A、潜在客户B、潜在客户C、潜在客户F等非关联的头部互联网及非互联网运营商及行业客户在测试、签订框架合同、开始交付等方面取得了进展。

请发行人披露：（1）腾讯相关预付款用于上游供应链采购的具体安排，是否符合行业惯例；（2）区分互联网客户与非互联网客户，说明发行人与非关联方客户的合作进展情况，包括但不限于签署的框架合同，报告期内已交付的产品数量、期后预计供货量，预计供货量对应的收入及利润，灰度测试进展、预计开始供货时间等。

请保荐机构简要概括核查过程，并发表明确意见。

回复：

一、腾讯相关预付款用于上游供应链采购的具体安排，是否符合行业惯例

具体支付及采购安排已申请豁免披露。

AI 芯片设计公司向上游供应链支付大额预付款符合行业惯例，根据寒武纪 2025 年定增问询函回复披露，2024 年末和 2025 年一季度末其预付账款分别为 77,437.67 万元、97,333.00 万元，主要向原材料供应商和委外加工厂商支付。用于云端产品线采购备货。

下游客户向 AI 芯片设计公司支付预付款符合行业惯例，根据摩尔线程 2026 年 3 月披露的《关于签订日常经营重大合同的公告》，合同金额为 6.60 亿元，在合同签署后 2 日内向客户收取合同价款的 30%。

综上，腾讯相关预付款安排符合行业惯例，具有商业合理性。

二、区分互联网客户与非互联网客户，说明发行人与非关联方客户的合作进展情况，包括但不限于签署的框架合同，报告期内已交付的产品数量、期后预计供货量，预计供货量对应的收入及利润，灰度测试进展、预计开始供货时间等

基于行业供给现状及公司自身产能、研发资源约束，公司依据供应链实际情况统筹优化非关联客户拓展策略与推进节奏，合理调配内部样卡资源、研发人力，优先聚焦潜在客户 A、潜在客户 D 等需求规模大、合作粘性强的核心客户；同时稳步推进潜在客户 F、潜在客户 B、潜在客户 H、潜在客户 C、潜在客户 E 等优质客户开拓，并取得积极进展，总体开拓进展情况如下：

类别	公司名称	灰度测试等具体进展	预计业绩	2026 年盈利预测是否考虑
互联网客户	潜在客户 A	第四代产品已通过前期硬件系统测试和模型匹配测试；进入确定灰度测试具体方案阶段，预计于 2026 年年内启动灰度测试	有望在 2026 年 12 月小规模交付，2027 年开始实现大批量交付，预计毛利率、供货量及对应收入已申请豁免披露（下同）	否
	潜在客户 F	第四代产品已通过硬件系统测试和模型匹配测试，进入确定灰度测试具体方案阶段，并预计于 2026 年年内启动灰度测试	有望在 2026 年 12 月小规模交付，2027 年实现大批量交付	否
	潜在客户 B	第四代产品已通过硬件系统测试和模型匹配测试，客户无灰度测试要求	有望在 2026 年 12 月小规模交付，2027 年分别实现大批量交付	否
	潜在客户 C	第四代产品已通过硬件系统测试，进入模型测试阶段。由于公司当前研发资源较为紧张，暂缓推进相关测试及商务需求沟通工作	-	否
	潜在客户 H	第三代产品已通过硬件系统测试和模型自测，正在进行系统级业务部门复测，预计于 2026 年内启动灰度测试	-	否
非互联网客户	潜在客户 D	第三代产品通过与合作伙伴 A 合作，已实现潜在客户 D 集采项目首次入围及中标并已完成中标后复测验收，合作伙伴 A 已与潜在客户 D 完成框架合同签署	预计 2026 年 6 月前向合作伙伴完成产品交付，已跑通了完整流程，后续与合作伙伴深度合作争取更多市场份额	否
		第四代产品正在与合作伙伴 A、合作伙伴 B 等合作方共同开发高密度 AI 服务器及超节点服务器，未来将依托合作伙伴在国内运营商 AI 服务器市场的领先地位及深厚渠道资源推进产品测试并进行市场开拓		否
	潜在客户 F	第三代产品已完成模型测试，测试结果满足客户要求，与合作伙伴 C 合作已于 2025 年向其实现数百张产品销售。未来公司将与合作伙伴 C、合作伙伴 B 等密切合作，争取在 2026 年实现更大批量销售	-	否

注：互联网厂商验证阶段一般分为三步，包括硬件系统测试、模型匹配测试和集群灰度测试，在通过硬件系统测试和模型匹配测试后，视集群灰度测试需要加速卡的数量进行小批量下单，全部验证通过后进行商务谈判并批量下单。

三、中介机构核查意见

（一）保荐机构、发行人律师和申报会计师进行了如下核查：

1、查阅发行人与腾讯签署的相关协议，腾讯向发行人预付款涉及的发票及银行回单；查阅发行人与上游供应链签署的协议，发行人向上游供应链支付预付款涉及的订单、发票、银行回单；访谈发行人管理层了解腾讯支付预付款用于上游供应链采购的具体安排；查阅同行业公司存在大额预付款安排的案例，判断存在大额预付款安排是否符合行业惯例；

2、访谈发行人管理层，了解发行人与非关联方客户的合作进展情况；查阅发行人与非关联客户合作的相关文件。

（二）经核查，保荐机构、发行人律师和申报会计师认为：

1、腾讯向发行人支付预付款的安排符合行业惯例；

2、公司非关联客户开拓进展良好，互联网客户中潜在客户 A、潜在客户 F、潜在客户 B 已完成硬件及模型匹配，正在推进灰度测试（其中潜在客户 B 无需灰度测试），预计在灰度测试完成后有望在 2026 年内完成小规模产品交付，在 2027 年实现大批量产品交付；非互联网客户中潜在客户 E 已实现数百张张第三代 AI 加速卡交付，潜在客户 D 公司的合作伙伴已与其完成框架合同签署，预计 2026 年内可实现收入，公司产品已进入潜在客户 D“采购短名单”，后续将与合作伙伴深度合作争取更多销售份额。

4 关于存货

根据申报材料：2024年末，随着库龄结构老化，发行人对无订单覆盖的二代产品进一步计提存货跌价准备。2024年末存货主要系用于成都高新电子的智算集群业务的产品，因该项目执行周期较长，部分产品截至年末尚未完成交付，但已有明确销售合同支持，未计提跌价准备。

请发行人说明：结合2024年末有明确销售合同支持且未计提存货跌价准备存货的后续收入实现情况，说明2024年末相关存货的可变现净值是否低于其存货账面价值，以及2024年末该类存货跌价准备计提是否充分。

请保荐机构、申报会计师说明核查依据、过程，并发表明确核查意见。

回复：

一、结合 2024 年末有明确销售合同支持且未计提存货跌价准备存货的后续收入实现情况，说明 2024 年末相关存货的可变现净值是否低于其存货账面价值,以及 2024 年末该类存货跌价准备计提是否充分

（一）2024 年末发行人二代 AI 加速卡对应的在产品和库存商品中无销售订单/合同支持的部分足额计提存货跌价准备，其余部分经跌价测试无需计提，具体如下：

单位：万元

项目	金额	存货跌价准备
无销售订单/合同（a）	9,977.42	9,977.42
有明确销售订单/合同（b）	9,440.56	-
——成都高新电子集群项目	8,596.87	-
——客户 Y	314.71	-
——客户 Z	528.98	-
质保备件、配件及其他（c）	618.12	-
在产品和库存商品合计余额（a+b+c）	**20,036.10**	**9,977.42**

（二）未计提存货跌价准备存货的后续收入实现情况，说明 2024 年末相关存货的可变现净值是否低于其存货账面价值。

1、成都高新电子集群项目

发行人作为总集成商为业主方搭建基于二代 AI 加速卡的智算集群项目，合同含

税价格为 18,012.52 万元，建设期限三年，采用分阶段交付和验收的项目模式。该项目 2025 年完成最终验收当期一次性确认收入，项目整体毛利率为 21.92%，客户已按约定节点回款，回款情况良好。截至 2024 年末，公司测算该项目对应二代 AI 加速卡对应存货可变现净值高于存货账面成本，2024 年末未计提存货跌价准备。

2、客户 Y

该客户计划筹建面向高等教育人才实训领域的智算中心项目，旨在培育国产 AI 加速卡使用生态，发行人二代 AI 加速卡性能能够完全满足该项目需求。发行人 2024 年 12 月与该客户的代理商签订正式销售合同，约定收到合同款后 5 日内发货，同时合同约定的二代推理卡及二代训练卡单价为发行人 2024 年同类产品平均售价的四至五折区间，单价差异主要系发行人在行业技术快速迭代背景下，为消化特定库存、同时争取该高等教育领域示范效应与长期生态价值而主动采取的商业谈判策略。

2024 年末，发行人以已签订合同价格为基础，测算的可变现净值高于相关存货的账面成本。同时发行人实地考察客户正在机房改造等实质性资本投入，表明客户正在为项目落地执行前置工作。结合发行人与客户、代理商在年末的持续沟通，合同履行不存在重大不确定性的客观迹象。因此，基于资产负债表日已获得的全部信息，判断无需在期末对相关存货计提跌价准备。

2025 年 9 月，发行人获知客户因自身原因未完成合同预付款筹集，项目被迫终止。2025 年，公司对对应的二代卡存货全额计提存货跌价。2025 年后续情况发生不影响发行人在 2024 年末基于当时可获得的所有合理及可支持信息所作出的无需计提存货跌价准备的判断。

3、客户 Z

该客户筹备在深圳建设异构智算中心，计划部署 80%的英伟达产品和 20%的国产 AI 加速卡。由于客户在项目筹措时点未有明确的应用方案，且对国产 AI 加速卡性能需求不高，为推进项目并针对性消化二代 AI 加速卡库存，发行人于 2024 年 5 月与该项目代理商签订了为期一年的销售框架协议。该协议明确了合作意向、产品搭配方案（二代产品和三代产品组合销售），并提供了具有竞争力的意向售价，属于主动进行库存管理，具有合理的商业实质。

2024 年末，发行人基于当时可获得的所有信息，对该框架协议对应的存货进行

了审慎评估。由于该框架协议仍在有效期内，且根据发行人与客户的持续沟通，客户方明确表示仍在继续推动该项目计划，合同履行不存在重大不确定性的客观迹象。因此，管理层判断该存货存在合理的销售预期。由于该框架协议未约定不可撤销的固定价格，发行人参考具有可比性的客户 Y 的合同价格作为估计售价，测算的可变现净值高于存货账面成本。因此，基于资产负债表日已获得的全部信息，判断无需对相关存货计提跌价准备。

该框架协议于 2025 年 5 月到期，因客户侧前期准备工作延迟，机房改造等未完成，项目最终未能继续开展，协议未再续签。2025 年，公司对对应的二代卡存货全额计提存货跌价。2025 年后续情况发生不影响发行人在 2024 年末基于当时可获得的所有合理及可支持信息所作出的无需计提存货跌价准备的判断。

4、质保备件、配件

（1）智算系统配件：2024 年 12 月，因城市 D 智算中心项目业主方将部分二代卡智算系统要求更换为三代卡智算系统，发行人完成退货验收。相关退货存货中，二代 AI 加速卡因无明确销售预期已全额计提跌价，剩余 309.38 万元如服务器机箱、CPU 等通用型智算系统配件，其价值不受 AI 芯片技术迭代影响，且发行人将搭配换货的三代 AI 加速卡重新销售至对方。因此 2024 年末，发行人判断该部分退货配件不存在跌价风险，未计提跌价准备。该部分智算系统配件已于 2025 年 1 月全部实现销售。

（2）质保备件：截至 2024 年末，公司为二代产品预留的质保备件余额为 301.60 万元。发行人根据仍在质保期限内的产品销售数量，基于对产品全生命周期预计缺陷率的判断及行业经验确定预留比例，作为用于履行未来的质保义务的专用备件，未计提跌价，具有合理性。

（三）2024 年末，发行人对该类存货跌价准备的计提充分

综上所述，发行人于 2024 年末对存货计提的存货跌价准备充分、审慎，严格遵循《企业会计准则第 1 号——存货》及应用指南的相关规定。对于无销售合同覆盖的存货，发行人已全额计提跌价准备；对于已有合同支持的存货，其可变现净值的测算具有可靠依据，对于后续因客观环境变化而新增的存货减值风险，不影响 2024 年末会计估计的合理性。同时，截至 2025 年末，发行人对未按计划实现销售的二代卡相关存货已全额计提跌价准备。

二、中介机构核查意见

（一）保荐机构、申报会计师执行了如下核查程序：

1、获取发行人 2024 年末二代产品已签订销售合同、商务项目的立项资料，了解发行人为销售预期而准备的存货具体情况；

2、询问发行人商务团队，了解相关项目的后续预计交付时点及可行性，了解后续销售情况，了解发行人于 2024 年末作出无需计提存货跌价准备的判断依据；走访成都高新电子集群项目智算中心现场，观察实际设备运行情况，检查验收单、销售发票等收入确认支持性文件；

3、询问发行人财务人员，了解存货进行可变现净值的测算过程、依据，获取并检查存货跌价准备计算表，结合企业会计准则评价存货跌价准备计提是否充分、审慎。

（二）经核查，保荐机构、申报会计师认为：

2024 年末，发行人对二代产品存货跌价准备的计提是充分、审慎的，符合《企业会计准则》的规定。对于有明确销售合同支持的存货，发行人基于合同价格等可靠依据判断其可变现净值高于存货账面成本，因而不计提跌价准备具有合理性；对于无销售合同覆盖的存货，发行人已全额计提跌价准备，计提充分。后续情况变化不影响 2024 年末会计估计的合理性。

保荐机构总体意见

对本回复材料中的发行人回复（包括补充披露和说明的事项），本保荐机构均已进行核查，确认并保证其真实、完整、准确。

（以下无正文）

（此页无正文，为《关于上海燧原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二轮审核问询函的回复报告》之签章页）

上海燧原科技股份有限公司

2026 年 6 月 2 日

发行人董事长声明

本人已认真阅读上海燧原科技股份有限公司本次审核问询函的回复报告全部内容，确认本回复报告不存在虚假记载、误导性陈述或者重大遗漏，并对其真实性、准确性、完整性承担相应法律责任。

发行人董事长：

ZHAO LIDONG

上海燧原科技股份有限公司

2026 年 6 月 2 日

（此页无正文，为《关于上海燧原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二轮审核问询函的回复报告》之签章页）

保荐代表人：

张 欢　　陈 泽

中信证券股份有限公司

2026 年 6 月 2 日

保荐机构董事长、法定代表人声明

本人已认真阅读上海燧原科技股份有限公司本次审核问询函之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，审核问询函之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。

董事长、法定代表人：

张佑君

中信证券股份有限公司

2026 年 6 月 2 日

（本页无正文，为北京市中伦律师事务所关于《上海燧原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二轮审核问询函之回复报告》之律师签章页，我们仅对审核问询函中需要律师进行核查的事项发表核查意见）

北京市中伦律师事务所（盖章）

负责人：

张学兵

经办律师：

唐周俊

经办律师：

曹美璇

经办律师：

梁　晶

2026 年 6 月 2 日

本页为毕马威华振会计师事务所（特殊普通合伙）关于《关于上海燧原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二轮审核问询函的回复》（“问询函”）之会计师签章页。根据问询函的要求，我们仅对问询函中要求会计师核查的事项进行核查并发表核查意见。同时，我们对发行人申报财务报表审计的目标是对财务报表整体是否不存在由于舞弊或错误导致的重大错报获取合理保证，并不是对财务报表中的任何个别账户或项目的余额或金额、或个别附注单独发表意见。

毕马威华振会计师事务所（特殊普通合伙）　　中国注册会计师

徐侃瓴

杨瑾璐

中国 北京

2026年6月2日