证券代码： 002195　　证券简称：岩山科技

上海岩山科技股份有限公司

投资者关系活动记录表

编号：2024-002

投资者关系活动类别	√ 特定对象调研　□ 分析师会议 □ 媒体采访　□ 业绩说明会 □ 新闻发布会　□ 路演活动 □ 现场参观 √ 其他（券商策略会）
活动参与人员	长江证券路畅、甬兴证券童非、东吴证券王紫敬、王世杰、张文佳及东吴证券策略会议所邀请的投资者
时间	2024 年 3 月 15 日（周五） 下午 14:00~15:00、15:30~16:30
地点	公司会议室；东吴证券策略会会议现场
形式	现场方式
上市公司接待人员姓名	1、公司董事、副总经理兼董事会秘书邱俊祺 2、公司旗下岩思类脑人工智能研究院首席科学家李孟博士 3、公司旗下上海岩芯数智人工智能科技有限公司 CEO 刘凡平
交流内容及具体问答记录	主要内容如下： **一、岩思类脑人工智能研究院简介** 公司在 2023 年半年度报告中对全资子公司上海岩思类脑人工智能研究院有限公司（以下简称“岩思类脑研究院”）的基本情况进行了披露。岩思类脑研究院是公司在承继了控股股东多年在类脑人工智能领域的研究成果基础上，于 2023 年 8 月成立。岩思类脑研究院致力于开展大脑内部状态解析与调控、深度生成式大脑信号解码算法、非器质性重大脑疾病的诊断和干预等前沿领域的研究。 岩思类脑研究院以脑电大数据与脑电大模型为核心技术底座，面向脑科学和人工智能领域的前瞻性研究，开展脑机接口解码算法与系统、非器质性脑疾病的诊断和评估、大脑内在状态调控等方向的科学研究和产品开发，推动研究成果商业化落地。 目前类脑研究院的研究工作由哈佛大学博士后研究员、德国马克普朗克学会研究科学家、中国科学院上海微系统与信息技术所研究员李孟博士领衔。

	李孟博士的主要研究成果包括：深耕大脑神经解码（斑马鱼、啮齿类动物和人类大脑）、类脑计算和原生脑计算（脑机接口）等前沿领域。在哈佛大学工作期间，破解了全球首例斑马鱼全脑十万量级神经元网络，相关研究论文《Internal state dynamics shape brainwide activity and foraging behaviour》发表于顶级学术期刊 Nature，并被 Nature 杂志以“News and Views”和“News Feature”形式进行单独评论和报道，是脑科学与人工智能交叉领域的里程碑式工作。 **二、RockAI（岩芯数智）业务简介** 在 2023 年半年报中，公司已经披露了在 AIGC 领域的布局情况。为进一步推进相关业务发展，公司已于 2023 年 6 月专门成立了上海岩芯数智人工智能科技有限公司（以下简称“RockAI”或“岩芯数智”）。 RockAI 是以认知智能为基础，专注于自然语言理解、人机交互等核心技术的创新型企业。RockAI 从零开始完全自主研发、并于 2024 年 1 月发布了国内首个非 Attention 机制的通用大模型——“Yan 1.0 模型”。 经对比实验验证（实验情况详见后文），在该等实验情况下 Yan 1.0 模型拥有相较于同等参数 Transformer 架构 LLaMA 2 模型更高的训练推理效率、吞吐量及记忆能力，更低的机器幻觉表达，同时支持 CPU 无损运行并 100%支持私有化应用。 2024 年 RockAI 将围绕 Yan 架构持续加强核心算法创新及迭代升级，着力构建 Yan2.0 大模型，Yan2.0 模型将会融合文字、音频、视频等多模态，以应用于更广泛的业务。 **三、介绍环节** **1、公司在类脑人工智能、脑机接口领域是如何布局的？** 类脑人工智能的目标是利用最新的脑科学与人工智能技术及工具，通过破译生物大脑的结构和功能，绘制大脑功能、结构和信息处理图谱，从微观、介观和宏观水平加深对生物大脑工作原理的理解，并构建模拟生物大脑的人工神经网络系统，最终达到“认识脑、保护脑和模拟脑”的目标。脑机接口技术是类脑人工智能研究的一个细分领域，旨在打破大脑与外界信息交互瓶颈，是实现人机交互、人机交融的必由之路。

	近期国内外脑机接口技术不断取得新进展，岩思类脑团队很早之前就已经认识到随着材料科学、信号处理、医疗设备的不断进步，可以采集到的脑电信号的数据量越来越庞大，如何从海量的数据中提取出所需颗粒度的信息，其中的脑电解码算法是脑机接口系统中急需突破的关键。 基于上述思考，岩思类脑跳过电极、芯片等硬件的研发，直接提前布局脑电大模型的构建和研发，从而可以适应现在及将来非侵入式、侵入式等多种方式获得的海量脑电神经网络数据，以脑电大模型为硬件赋能，从而达成实时、精准、高效的人机交互系统。因此，岩思类脑研究院当前重点开展大脑内部状态解析与调控、及脑电大模型的研究。 **2、目前岩思类脑已经开始尝试进行脑电大模型的预训练，请问与传统大数据相比，脑电大数据有哪些特点？与语言大模型类比，脑电大模型有哪些区别？** 与传统大数据相比，脑电大数据训练数据获取难度高，一般临床医学上通过侵入式方式获得的脑电数据更加精准；脑电大数据的时空复杂度高，大脑是一个三维空间，脑电数据既包含大脑皮层空间位置信息，又是时间维度上的连续信号；脑电大数据的预处理难度更大，需要按神经系统特性规律进行合理分割后才能token化。 与自然语言大模型类比，脑电大模型需要实现更高自由度、更细颗粒度的解码效果以及极强泛化性能，以实现跨样本、跨物种模型泛化移植的效果。当前侵入式脑机接口没有大范围应用的瓶颈之一在于模型的泛化性比较差，通常只是建立针对单个病人的脑电数据模型。但是脑电大模型可以通过采集海量的临床数据，提取其底层最本征的表达，进行脑机接口解码，未来目标包括实现在不同样本甚至不同物种之间进行移植。 **3、RockAI（岩芯数智）为什么要从零开始设计非 Attention 机制的 YAN 架构，而不是使用 ChatGPT、LLaMA、PaLM 等 Transformer 架构的大模型进行设计或调整？** Attention机制是一种能让模型对关键信息重点关注并充分学习吸收的技术，也就是把注意力集中放在重要的点上，而忽略其他不重要的因素。ChatGPT 等都利用了 Transformer 架构，其核心技术之一就是 Attention 机制。标准的

Attention 机制的计算复杂度为 O(n^2*d)（其中 n 表示序列长度、d 表示特征维度，^2 指平方）。标准 Attention 机制的复杂度随序列长度呈 2 次方增长。通常来说 Transformer 架构具有训练周期较长、应用成本过高、高机器幻觉表达等缺陷，在实际应用中需要的高算力和高成本让不少中小型企业望而却步。

针对 Transformer 架构的上述缺陷、以及不同行业对于高效能、低能耗 AI 大模型需求的不断增长，公司旗下岩芯数智研发团队意识到从零开始设计新架构的必要性，并于 2024 年 1 月推出了国内首个非 Attention 机制大模型—Yan 1.0 模型。

4、从大模型解码层结构来看，Yan 架构与 Attention 机制模型区别如何？

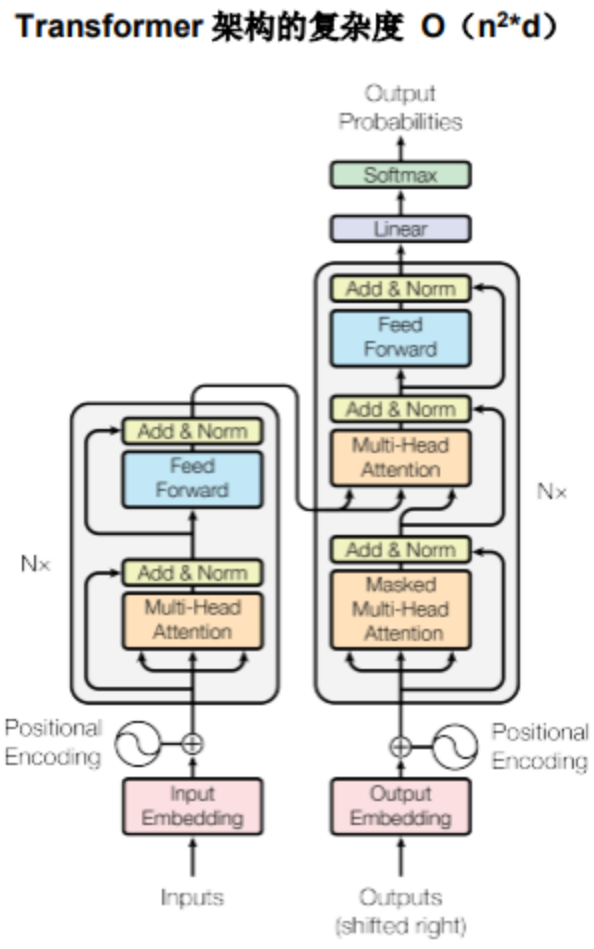

图 1 基于多头 Attention 机制的 Transformer 模型结构

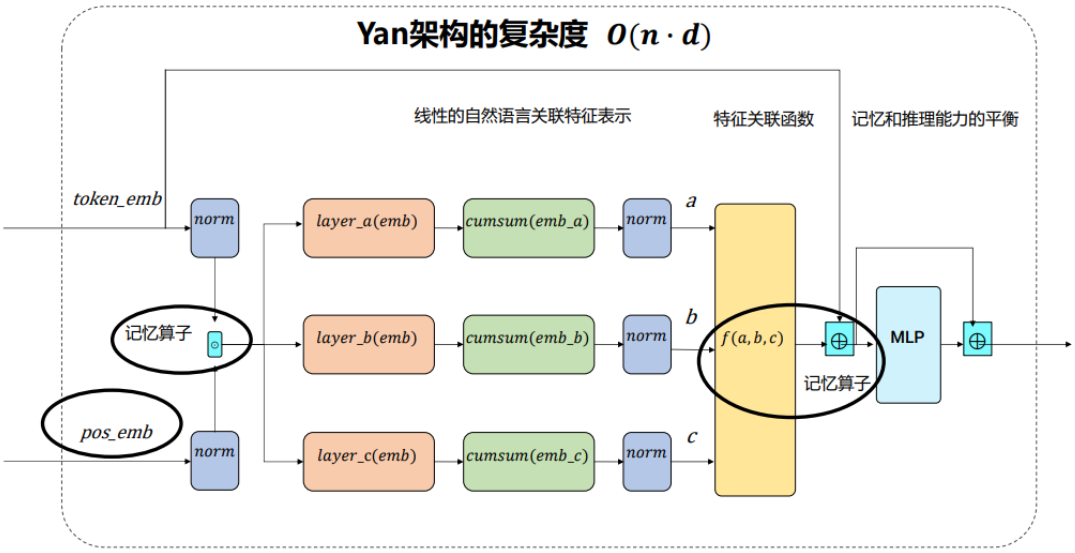

图 2 Yan 架构的大模型解码层结构

图 1 为基于多头 Attention 机制的 Transformer 大模型结构，标准的 Attention 机制的计算复杂度为 O(n^2*d)，其复杂度随序列长度呈 2 次方增长。这也就意味着，当输入序列长度增加时，计算成本和显存需求会快速增长。

从图 2 的 Yan 架构大模型解码层结构可以看到，Yan 架构没有基于 Attention 机制，也没有基于 RNN（指 Recurrent Neural Network，循环神经网络）等序列模型，而是通过完全自研的记忆算子及特征关联函数，将计算复杂度从标准 Attention 机制的 O(n^2*d)降低为 O(n*d)（线性复杂度），从而提高了 Yan 架构模型的训练效率和收敛速度。因此，Yan 架构模型具有训练速度快、推理成本低、记忆能力强等优势。

5、对相同参数量级的 Yan 模型和 Transformer 架构的大模型进行对照实验后，实验结果如何，是否能验证 Yan 架构的优势？

RockAI 对相同参数量级的 Yan 1.0 模型和 Transformer（对照实验中采用的 Transformer 是基于 HuggingFace LLaMA 2 的标准结构，同时开启了 flash-attn 的支持）架构模型分别用 1,000 万条中英平行语料，基于同样软硬件环境的 NVIDIA A800 服务器训练以及同样的迭代次数下进行了对照试验：

（1）训练效率方面，在上述对照实验环境下 Yan 1.0 模型的损失值要低于 Transformer 架构的 LLaMA 2 模型。在训练集上，训练数据收敛到相同的 loss

	（loss=3.0）时，Yan 1.0 模型仅需要 1.5 小时，而 LLaMA 2 模型却花费 10.5 小时，因此 Yan 1.0 模型的训练效率更高。 **（2）推理准确率方面，**在上述对照实验环境下 Yan 1.0 模型比 LLaMA 2 模型在训练集上的预测准确率高出 17%、在验证集上的预测准确率高出 13%。 **（3）显存占用方面，**基于同样的参数量级在单张 NVIDIA RTX 4090 24G 显卡上，当输出 token 的长度超出 2,600 时，LLaMA 2 模型会出现显存不足，进而无法完成推理；Yan 1.0 模型的显存使用始终稳定在 14G 左右，可以进行正常推理。Yan 1.0 模型的显存占用及成本比 LLaMA 2 模型更低。 **（4）记忆能力方面，**古诗是通过简短的字和词语表达丰富语境的一种体裁，token 之间的困惑度也高于现代文，这恰好可用于评测模型的记忆能力。在对照实验中分别用数十万条古诗数据进行续写训练，与 LLaMA 2 模型相比，Yan 1.0 能够更快的达到更好的收敛，以及更高的准确率。 **6、Yan 模型能够部署在办公电脑上吗？** 原生结构的 Yan 架构模型，在零压缩、零裁剪的情况下，依然能够流畅运行于主流消费级 CPU 设备，例如经训练后的模型可以部署在配置了 Intel i7、i5 CPU 的笔记本电脑或台式机，以及 M 系列芯片的 MacBookPro 等。 对比之下，70 亿（7B）参数量的 Transformer 却无法在上述 CPU 设备上直接运行，通常 Transformer 需要经过 8bit 甚至 4bit 的量化后才能正常运行，这不可避免的带来了推理精度的损失。 **7、OpenAI 发布了首个文生视频模型 Sora，请问贵公司是否也布局推进相关的新技术？** 2024 年 1 月 RockAI 发布的 Yan1.0 大模型以自然语言为主，尚不支持文生视频功能。目前 RockAI 正在研发 Yan2.0 大模型，Yan2.0 将会融合文字、音频、视频等多模态，以应用于更广泛的业务。Yan 2.0 模型预计将于 2024 年下半年推出，具体推出时间请以后续岩芯数智的发布为准。

	8、未来 Yan 模型有哪些应用潜力，商业化构想如何？ Yan 架构的模型也是通用大模型的一种，可适用于当前所有通用化大模型的商业化应用场景。Yan 架构模型现阶段商业化的重点主要在尝试为企业客户提供本地化应用和部署，满足客户对于数据隐私、安全及低成本部署上的需求，目前尚未形成规模收益。 未来，针对 to B 垂直领域，RockAI 希望能在低消耗、显存受限的情况下，打造基于 Yan 架构的专业生产力工具，解决更多低算力模型场景应用，如在网络连接不稳定或离线使用场景的应用等。 **9、请介绍一下公司拟收购的智能驾驶企业 Nullmax 纽劢科技的最新进展情况。** 公司拟增资并收购 Nullmax （Cayman） Limited 部分股权的事项目前正在向商委、外汇管理局等主管部门办理境外投资 ODI 审批手续中。相关进展请以公司公告为准。
关于本次活动是否涉及应披露重大信息的说明	不涉及
活动过程中所使用的演示文稿、提供的文档等附件（如有，可作为附件）	