
腾讯研究院
上海交通大学
SHANGHAI JIAO TONG UNIVERSITY
腾讯优图
腾讯云智能
2024大模型十大趋势
走进“机器外脑”时代

人工智能正在迅速发展，大模型技术正成为赋能各行各业的关键。

从算力底座、智力增强到人机协作，大模型正在重塑人类社会，成为我们可依赖的"外脑"。

目录CONTENTS

编委会

序言

PREFACE

走进“机器外脑”时代

序言1

PREFACE

司晓

腾讯副总裁 腾讯研究院院长

继ChatGPT开启大语言模型引领的新一轮人工智能革命以来，我们持续见证了人工智能领域技术的加速迭代，在过去的一年里众多公司如Google、Midjourney、Adobe以及Inflection等，都推出了创新的模型和产品，标志着大模型技术的成熟和大规模应用的开始。今年2月，Sora的出现再次震撼了技术界，预示着我们可能很快就会见证更多令人兴奋的技术突破。过去半年，我们以日为单位更新“AI每日动态”，这可以充分反映出来，AI技术的发展日新月异，以日来统计的话也是毫不过分的。

在海量GPU和新一代大模型的加持下，人工智能在三个方向上有了实质性的飞跃。第一是推理能力。大语言模型为人工智能带来了所未有的推理能力，极大地扩展了机器的认知边界。这种推理能力的跃迁得益于LLM在理解和生成自然语言方面的巨大进步。它们能够解析复杂的文本，提取关键信息，进行逻辑推理，并生成连贯、有见地的回应。这使得LLM能够处理各种知识密集型任务，如法律分析、市场研究、科学发现等，为个人和企业提供了强大的智能支持。以往人类智力难以企及的科学探索高地，都可以在AI的帮助下实现。例如，英伟达的“地球 2 号”项目，旨在创建地球的数字孪生体。模拟整个地球的运行，以预测其未来变化。通过这样的模拟，可以更好地预防灾难，更深入地理解气候变化的影响，从而更好地适应这些变化。通过这样的模拟，可以更好地预防灾难，更深入地理解气候变化的影响，从而让我们能够更好地适应这些变化。随着更高级的推理智能被开发出来，各行各业都将有机会拥有“机器之心”。AI 将引领新的服务模式，即“智力即服务”（IQaaS），该模式的一个重要特征将是机器的推理能力以在云端的方式、通过大模型提供给用户，“AI数字员工”将进一步成为现实。大模型使机器不再仅仅是执行简单任务的工具，而是成为了人类的“智力外脑”。

第二个方面是创意的生成。AI技术，尤其是AIGC，正迅速成为创意产业的一股颠覆性力量，为创意工作者提供了前所未有的生产力提升。今年2月，Sora的问世不仅是技术界的一次震撼，更是对未来创新潜力的一次大胆展示。AIGC技术通过文生文、文生图、文生视频等多种形式，使得创作、设计、分析等任务变得更加高效和易于实现。Sora和SUNO等现象级产品的出现，标志着AI生成内容的质量和多样性达到了新的高度。它们不仅让普通人能够创作出接近专业水准的音乐和视频作品，而且正在快速改变媒体、影视和音乐行业的生态。这些技术的普及，降低了专业技能训练的门槛，使得创意表达更加通用化。现在，只要有创意想法，人们就可以利用AI这个强大的“创意外脑”，将灵感转化为现实。AI的这种能力，不仅为专业创意工作者提供了强大的辅助工具，也为普通爱好者打开了创作大门，使他们能够轻松实现自己的创意愿景。随着AI技术的不断进步，我们可以预见，创意产业将迎来一个更加多元、开放和创新的新时代。

另一个方向属于广义的情感陪伴。Dan模式的全网爆火，不仅展示了AI在情绪理解与表达上的巨大进步，更凸显了其与人类情感交流的无缝对接。GPT4o等高级AI系统的自然交互体验，让人与机器的界限变得模糊，仿佛科幻电影《Her》中的情感故事正在逐步成为现实。

AI技术在满足人类情感需求方面展现出巨大潜力，扮演起了人们的“情感外脑”。AI聊天机器人提供的心理咨询服务，以其24/7的不间断陪伴，为需要帮助的人们提供了及时的情绪支持和专业建议。在儿童领域，智能玩具不仅陪伴孩子们成长，更通过情感交互，培养孩子们的情感认知和社交技能。随着情感智能技术的不断成熟，数字生命的议题也日渐升温。一些创新尝试正在探索如何利用数字技术复刻已故亲人，为生者提供缅怀与思念的渠道。尽管这一领域还面临着诸多法律和伦理挑战，但其在情感陪伴方面的应用前景无疑为AI赋予了新的温度和深度。AI不再仅仅是冷冰冰的生产力工具，它正在成为人类情感世界中的一个温暖伙伴。随着技术的不断发展和应用的不断拓展，我们有理由相信，AI将在人类的情感生活中扮演越来越重要的角色，为人们带来更多的陪伴与慰藉。

PRE-FACE1

PRE-FACE1

在本报告中，腾讯研究院基于科技行业发展和腾讯自身在AI领域的深耕，提出了10个关键性的趋势，试图理解全世界范围内正在发生的AI巨变。与往年一样，我们从技术、应用、社会三个方面来预测AI给经济社会带来的影响。我们正在进入一个“机器外脑”时代。加速技术为大模型行业的发展提供了算力的保障。随着大模型与人机协作的深入，个体创作的门槛进一步降低，越来越多的个体借助大模型外脑成为“斜杠青年”、“超级生产者”，甚至开启自己的“一人企业”。端侧模型的优化将大幅提升提升移动设备的体验，开启全新的人机交互方式。在工业领域，多模态通用感知技术正在提升生产力，而游戏与大模型的共生关系为Agent训练提供了新的舞台。开源模型的成熟，为技术共享与创新提供了强大的生态支持。最后，人机对齐成为确保大模型安全与治理的核心议题，指引着我们走向一个更加智能、高效和安全的未来。

这十大趋势共同勾勒出一个由大模型驱动的新未来。在这个未知和无限可能的时代，我们正在目睹AI如何将创意转化为现实，如何让个性化服务触手可及，以及如何为传统行业注入新的活力。AI让智力资源平权化，意味着无论背景或资源如何，每个人都有机会借助AI外脑实现自己的创意与梦想。这一变革不仅降低了创新的门槛，也为社会各阶层带来了前所未有的机遇。只要你拥有创新的想法并善于利用AI这一强大的外脑，即使在资源有限的情况下，也有可能以低成本创造出令人瞩目的成就。让我们一起走进这个“机器外脑”时代，见证人类能力的再次飞跃。

“人物-行为-场景”一体化的AI新范式

序言2

PREFACE

马利庄

上海交通大学特聘教授 人工智能研究院副院长

人物-行为-场景一体化视觉表达与理解是未来人工智能的重要研究方向。随着生成式人工智能以及通用人工智能大模型技术的发展，赋予了智能体感知理解、任务思考、持续学习的一系列能力，并通过直接的物理交互满足人类的各种需求。因而，在未来智能体可以承担更多的体力劳动和重复性任务，而人类则可以更加专注于创造性和思维类工作。其中，人物-行为-场景一体化视觉表达与理解是具身智能、智能生成等人工智能的核心基础，是链接物理世界的关键，一系列顶尖高校以及公司人员都已经下场研究这一新的AI范式。

斯坦福大学李飞飞教授创建的公司就利用类似人类的视觉数据处理方式，使人工智能能够进行高级推理。她曾在温哥华TED演讲中表示，其研究涉及一种可以合理地推断出图像和文字在三维环境中样子的算法，并根据这些预测采取行动，这种算法概念叫做“空间智能”。为了解释这一概念，她展示了一张猫伸出爪子将玻璃杯推向桌子边缘的图片。她表示，在一瞬间，人类大脑可以评估这个玻璃杯的几何形状，三维空间中的位置，它与桌子、猫和所有其他东西的关系，然后预测会发生什么，并采取行动加以阻止。她说：“大自然创造了一个以空间智能为动力的观察和行动的良性循环。”她还补充说，她所在的斯坦福大学实验室正在尝试教计算机“如何在三维世界中行动”，例如，使用大型语言模型让一个机械臂根据口头指令执行开门、做三明治等任务。

英伟达CEO黄仁勋此前在多个场合强调了一体化视觉表达的重要性，并预测人形机器人将成为未来主流产品。英伟达近期发布人形机器人通用基础模型Project GR00T，希望能让机器人拥有更聪明的“大脑”。由Project GR00T驱动的机器人能够理解自然语言，并通过观察人类行为来模仿人类动作。

PRE-FACE2

2024年5月，以“大模型具身智能”为主题的松山湖科学会议上，近40位院士专家围绕主题分享最新技术趋势和突破性进展。波士顿咨询公司(BCG)预测，到2030年，智能机器人系统可能给全球经济带来约4万—6万亿美元的年增长价值。

当前对人物理解的研究依旧是机器视觉的核心，但需要从单纯的人脸识别、动作识别等人物视觉技术逐渐转换为与场景交互的一体化表达范式。

例如，高速动态场景中自动驾驶系统无法理解周围环境中人和物的多变行为意图，容易引发严重的交通安全事故;服务机器人无法预测儿童的意图，也成为家庭的安全隐患。其核心问题是人物行为具有多样性和歧义性，同样的行为在不同的场景下具有不同的含义，行为意图的歧义性必须通过时序序列分析才能进一步消除。因此，必须研究时序数据进行人物-行为-场景一体化视觉表达，而这就需要对多模态的数据进行联合分析。

图灵奖得主Hinton教授在5月访谈中就表示多模态学习可以使模型更好地理解空间事物，因为仅从语言角度来看很难理解这些空间事物。当让模型成为多模态时，如果让它既能做视觉，又能伸手抓东西，并能拿起物体并翻转它们等等，多模态模型就会更好地理解物体。

随着diffusion、视频生成大模型的不断发展，真实物理世界的规则先验将成为未来视觉、人工智能研究重点。相较于ChatGPT、图文大模型等生成式人工智能在低维空间探索世界，Sora等视频生成式人工智能开始初步在三维空间模仿真实世界，并以人更容易接受的视频形式加以展示，这样高精度仿真世界投影的出现，也展示出了算力以及算法的进步。Sora生成的视频令人惊讶更多在于它大颗粒度上符合受众对真实物理世界的观察与体悟，让人感觉如“亲眼所见”。其背后是对相关物理规律，如近大远小、自由落体等进行深度挖掘、数字化后的成果。然而，众多权威学者和业内专家发声强调，Sora在二维视觉信息的传播与时空维度的表现力上虽然独树一帜，但并未达到对真实世界的全貌进行全面刻画和模拟的高度，尚未形成严格意义上的世界模型。但瑕不掩瑜，能够生成看起来像是在三维环境中自然移动和互动的视频，已经可以看作是人工智能“虚拟创世”趋势的关键节点。

PRE-FACE2

在AIGC发展的时间线上，总体是从单模态到多模态，从小模型到大模型这样的越来越复杂化和智能化的过程。AIGC模型的基本逻辑是从多模态的数据集，通过训练生成的大模型，服务于相应的各类应用任务。数据集包含文本、图像、语音、视频、结构化数据、3D信号等等。大模型通过训练来进行生成式的选择，然后不断的加以扩展。这个生成和扩展是需要大规模数据或知识的积累，就像一个人，行万里路以后具备了丰富经验，脑子里有很多经验与知识。但最后还有一步，还是要有一些专家或公认的权威人士对它做强化训练，通过强化学习等生成合理可用的AI模型。最后一步非常重要，如果纯是AI生成的内容，逻辑上可能会混乱，通过强化学习，提高它的精准度，并加以约束使之符合社会伦理、政策法规等。适应的任务范围包括知识检索、文本生成、音频制作、视频制作、科学研究等等，内容是非常广阔。

为适应数智时代数字内容智能化生产趋势，2024年4月17日，国家人力资源社会保障部等九部门就联合发布《加快数字人才培育支撑数字经济发展行动方案(2024—2026年)》，旨在通过规划数字人才未来的“成长地图”和培育体系，夯实数字经济高质量发展的“人才底座”，发挥数字人才支撑数字经济的基础性作用。

人工智能是新一轮科技革命和产业变革的重要驱动力量。其中，具身智能是场景理解感知、逻辑思考、行动决策三者有机智能融合的机器或系统，是人工智能在物理世界的进一步延伸。当今数智时代，能够以十分之一的成本实现千百倍内容生产速度的AIGC(生成式人工智能)，正越来越多地参与到数字内容的创意性生成，AIGC可以说将成为了未来互联网的内容生成基础设施，内容生产需求迈入强需求、视频化、拼创意的螺旋式升级阶段。

具身智能、智能生成等人工智能技术的发展既需要一套人物-行为-场景一体化表达范式，同时也需求大规模数据或知识的积累以及专业的强化训练。大规模数据或知识的积累这就隐藏着数据以及训练量的规模法则，微软全球前副总裁姜大昕就认为在肉眼可见的未来，至少还有十万亿和百万亿两个数量级。通过大规模数据以及训练集成了一体化表达范式的人工智能通用模型也可以在AIGC大模型、具身智能等应用中大放异彩，从而让机器能够更多地承担冗余重复的工作，释放更多的时间让人类能够更加享受创造性的工作以及高品质的生活。

共创、共建、共享智能美好未来

序言3

PREFACE

吴运声

腾讯云副总裁 腾讯云智能负责人 腾讯优图实验室负责人

过去的几十年，于中国实体产业来说是不平凡的，也是令人尊敬的。他们实现了技术与体验的数次跃迁，站在了信息化、数字化、智能化与开放化的“四化”前沿，并继续深入探索如何充分运用AI大模型、云计算、大数据等数字技术和产品，全方位重塑自身业务流程、商业模式与组织架构，迈进以用户体验为中心、业务快速迭代、健康可持续的新发展阶段。

作为千行百业的亲密合作伙伴，腾讯云有幸近距离见证了实体产业数字化、智能化转型道路上的点点滴滴，与每个客户一道深刻体会行业的沧桑巨变，体会数字经济时代对业务上、组织上和思维上的莫大影响。

近年来，腾讯云也从数字新基建、数字新连接及场景新服务等维度入手，以长期主义心态，不断夯实云计算、大模型等产品技术能力，发挥触达亿万用户的连接能力，与合作伙伴共建开放、健康、安全的数字生态，助力实体经济高质量发展。

过去一年里，我们发布了全链路自研的混元大模型，在国内率先采用混合专家模型 (MoE) 结构。目前，混元已经在腾讯内部600多个业务和场景中落地测试。例如，腾讯会议就基于混元推出AI小助手，通过简单自然的指令，就可以完成发言提醒、观点总结、会议纪要等能力，大幅度提升会议效率。

同时，我们面向ToB企业用户也推出了行业大模型，基于腾讯云TI平台和混元大模型基座，以高浓度的行业数据，加强模型对行业专业知识的理解；结合搜索增强与实时查询能力，提升模型解决产业问题的实时性、准确度、安全性等能力。目前，也已经在金融、医疗、教育、汽车、能源等20多个行业落地。

我们也看到，还有很多企业非常期待将大模型能力快速应用于生产、销售和服务。这需要模块化的大模型PaaS工具，大幅降低开发门槛，缩短从模型到应用的距离。因此，围绕文本、图片、视频三种信息主要载体，我们推出了三款PaaS产品，——“大模型知识引擎”、“大模型图像创作引擎”和“大模型视频创作引擎”，打造大模型原生工具链，助力企业在知识服务、图像和视频创作上提质提效。

PRE-FACE3

从通用模型到行业模型到模型开发工具再到即插即用的模型产品，我们一直以“全自研、高可用、强安全”的产品思路，去助力广大用户提效、去尽可能地降低技术使用门槛、去加速AI模型普惠。而这背后，也离不开腾讯在人工智能、大模型方面的投入与积累：过去五年，腾讯在人工智能领域申请专利超过10000项，居全球互联网行业榜首。腾讯优图实验室拥有1600多项人工智能相关专利，发表顶会论文800多篇，多次在国际权威比赛中创造世界纪录。

可以说，人工智能正在迅速发展，大模型技术也正成为赋能各行各业的关键。从算力底座、智力增强到人机协作，大模型正在重塑人类社会，成为我们可依赖的"外脑"。今天，我也很开心看到《2024AI大模型十大趋势——走进“机器外脑”时代》白皮书发布，报告中所呈现的内容方向精准且富有前瞻性，深入剖析了大模型发展的可能方向和应用影响。

比如，其中有一节谈到多模态AIGC会重塑内容产业生态，我是基本认同的。多模态大模型的技术路线是一条充满创新与突破的道路，它融合了多种模态的数据，如文本、图像、音频等，通过复杂的算法和强大的计算能力，挖掘出数据背后隐藏的模式和规律。这种融合不仅极大地丰富了模型对世界的理解和表达能力，还为解决复杂问题提供了全新的思路和方法。

其价值更是不可估量。比如在医疗领域，能够辅助医生进行更精准的诊断；在工业生产中，提升生产效率和质量；在文化创作领域，激发无限的创意灵感。多模态大模型正在成为推动社会进步和发展的强大引擎，为人类创造更美好的未来奠定了坚实的基础。

这份报告，既凝聚了腾讯云与各方在AI模型方面的洞察与互动，也引发了更多全新的讨论和大胆的畅想。我们希望，该报告能对正在探索人工智能、大模型发展的从业者们有所启发，也希望广大读者给予我们宝贵的反馈意见，期望后续与更多同行者一道推进科技赋能与产业创新，共创、共建、共享智能美好未来。

趋势1

算力底座：迈向十万卡集群量变，速度和效率双提升

作者：刘莫闲

生成式 AI 的迅猛演进，推动 AI 基础设施（AI Infra）加速发展，增长趋势将从大模型专业领域延伸至各行业领域，AI Infra“质量双螺旋”的发展模式将逐步形成，单集群从万卡“量变”至十万卡的同时，集成、互联和分布式将成为AI Infra“质变”破局的三板斧。

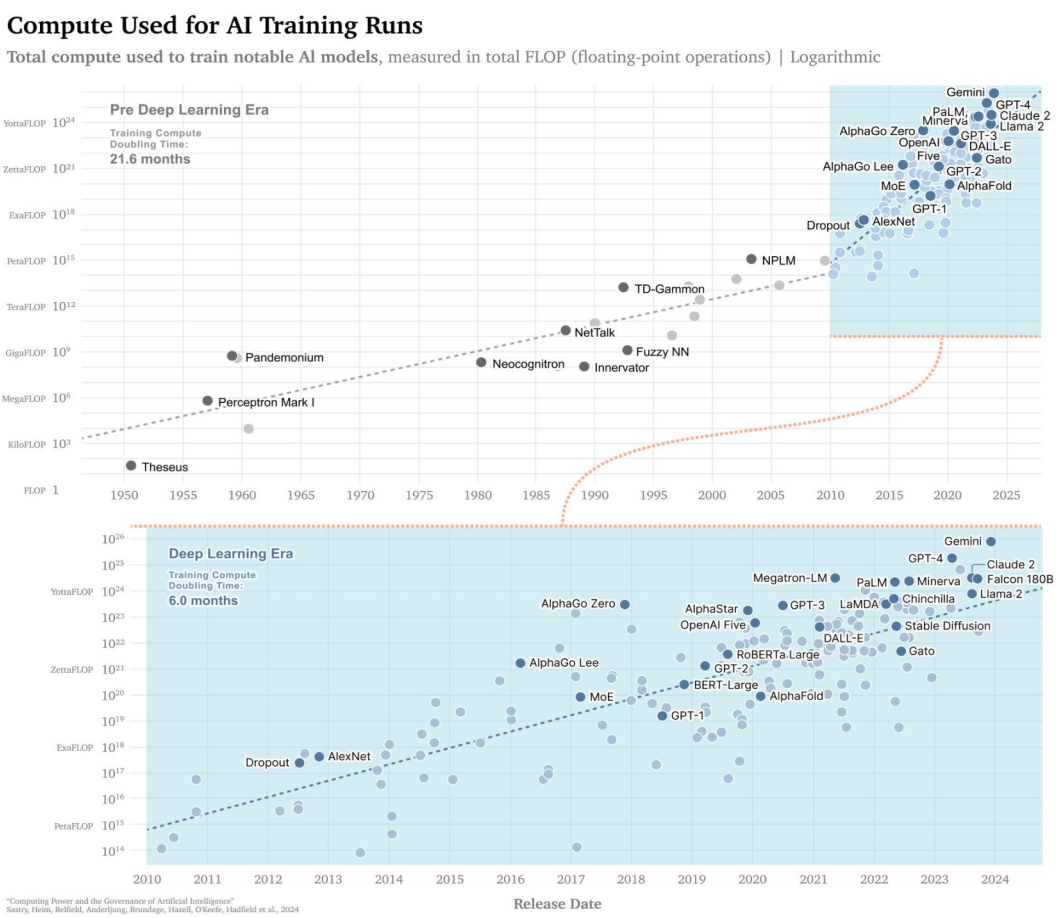
生成式 AI的演进也在推动它的底层基础 - 人工智能基础设施（AI Infrastructure，简称 AI Infra）技术的进步和建设的持续增长。由于生成式 AI 技术迭代和商业化探索仍在加速进行，AI Infra 短期的发展总体呈现供需两旺的形势。

一般的，AI Infra是指支撑 AI 大模型开发、部署和管理的软硬件工具组合。国际上，AI Infra 通常会被划分为 5 层，自下向上分别是：算力设施、基础大模型、数据和存储、模型开发和部署、以及监测与对齐。

当前，生成式人工智能的发展仍处于初期阶段，行业对 AI infra 的需求也相对初级，主要集中在算力设施层，未来将发展需求将逐步覆盖其他层级。而随着算力基础设施建设的规模进一步扩大，技术迭代逐步深入，人们对算力设施层关注的焦点，正在从单一对“量”的追求，向“质”、“量”兼顾演变。换句话说，未来AI算力基础设施的发展，将在更大规模加速卡集群容量、和更高算力利用率及计算能效之间交替进化、相互促进。AI Infra “质量双螺旋”的发展模式逐步显现，并向上层延伸。

人工智能基础设施供需两旺，增长趋势向行业企业延伸

生成式AI算力需求惊人，AI服务器市场增长预期明确。相关研究报告显示，自2012年以来，AI大模型训练的算力呈指数级增长，每3.4个月翻一倍。这意味着，从2012到2018年，AI算力增长了超过30万倍。与2012年的模型相比，2020年提出的模型需要600万倍的计算能力。预计这个增长还会继续快速提升。



图：历史上主要 AI 模型训练的算力用量趋势

图片来源：https://blog.heim.xyz/training-compute-thresholds/

以OpenAI为例，自ChatGPT发布已来，GPT等大模型能力快速和持续的提升，也得益于大规模AI加速算力对于模型训练的支撑：根据公开信息，OpenAI训练GPT-4大约需要25000张A100训练90-100天，训练GPT-MoE-1.8T需要8000张H100训练90天，训练Sora预计需要4200-10500张H100训练1个月，而训练GPT-5可能需要多达50000张 H100。

在Scaling Laws的指引下，越来越多的AI科技企业对于大模型更强能力的追求，正在引发更多的、对于更大规模、更高算力的AI Infra的惊人需求。

Meta到今年底前计划再获得35万个H100 GPU，并宣布将资本支出大幅提高到350-400亿美元；OpenAI和微软正在制定一个新的数据中心项目计划，投资可能高达 1000亿美元，其中还包含一个名叫“星际之门”的AI超级计算机。此外，亚马逊、微软等云服务商也在计划数年内投入近百亿美元增加其在全球各地的超大规模云计算和AI基础设施，以匹配迅速增长的大模型建设和服务需求。

IDC预计，全球人工智能硬件市场（服务器），将从2022年的195 亿美元增长到2026年的347亿美元，五年年复合增长率达17.3%；其中，用于运行生成式人工智能的服务器市场规模在整体人工智能服务器市场的占比将从2023年的11.9%增长至2026年的31.7%。

图：全球人工智能服务器市场规模预测2022-2026

来源：IDC《2022-2023 全球计算力指数评估报告》

AI Infra建设需求向行业企业延展，制造业对于AI Infra的建设最积极。全球各大科技巨头对AI基础设施的投入充分体现了他们对AI发展前景的乐观预期和决心。这一趋势不仅将加速人工智能技术的突破和应用落地，也将为相关产业链带来巨大的发展机遇。

根据微软与Forrester Consulting联合进行的《人工智能基础设施现状2024年度调查》报告显示，在受访的 1500多名各行业和地区的商业领袖中，56%的人反映所在企业并没有良好的基础设施来支持AI相关的业务工作；41% 的人认为人工智能基础设施是现阶段企业发展人工智能时最需要帮助的领域；而43%的人主张积极主动的发展自己的人工智能基础设施战略，其中最积极主动的行业包括：制造业，金融，软件，零售和卫生保健。

AI算力设施“量”变，集群规模将从万卡迈向十万卡

生成式AI的训练集群规模，以步入万卡量级。按照Scaling Laws的基本逻辑，拥有越大的模型参数，人工智能大模型的能力就会越强。同时，训练大模型所需要的算力集群规模也就越大，这样才能在合理的时间内完成大模型的训练。

从国内外头部的大模型训练情况得知，目前训练大模型所需要的单集群规模，已经从千卡上升至万卡。也就是说，训练一个大模型所需要的算力基础设施，需要10000张以上的AI加速卡集成在一个AI智算集群才能完成：OpenAI已在训练GPT-MoE-1.8T时使用了8000卡集群；Meta于2022年公布了拥有1.6万张A100 GPU的超级计算集群 RSC，并于2024年初公布了2个24576张Nvidia H100集群，支持Llama3的训练；国内头部科技企业也陆续实现万卡集群来训练大模型；

万卡集群的实现和运行，是各层级软硬件紧密耦合和持续调试的复杂系统工程。万卡集群并不是简单的将 AI 加速卡在硬件层面进行单纯的连接和堆砌，还要能够基于计算任务进行统一调度和管理，以实现大模型训练和推理的算力集群。实现稳定运行的万卡集群，并有效支撑大模型训练并非易事，总体看技术上会面临硬件和软件两个层面的挑战：

首先，构建万卡级别的超大规模集群本身就是一项极其复杂的系统工程。集群中成千上万的高性能计算单元需要以极高的带宽和极低的延迟进行互联，对网络拓扑、传输协议、线缆布线等都提出了苛刻的要求。同时,高密度部署还面临着散热和供电的巨大压力。现有的计算机网络和数据中心技术需要全面升级,才能满足万卡集群的苛刻要求。

其次,在软件层面,实现高效的分布式并行训练也面临重重挑战。传统的数据并行和模型并行范式在万卡尺度下将遇到通信瓶颈和负载不均衡等问题。需要全新的混合并行范式和任务调度机制,在最小化通信开销的同时实现高效的并行计算。分布式训练框架还需要内置故障检测和恢复机制,确保系统能容忍局部节点的失效。此外,高效的分布式优化算法,自动混合精度训练,以及针对大规模异构集群的资源管理和任务编排技术,都亟待突破。

国内外A加速卡的发展呈现多元化发展趋势，不同芯片架构、不同品牌、不同型号的 AI加速卡都将有可能成为万卡集群算力调度的一部分。如何将异构加速卡进行统一虚拟化、调度管理、并执行模型训练更是需要持续面对的技术挑战。突破这一难题，不仅需要考虑硬件层面的兼容问题，还要考虑不同并行计算、分布式训练等软件系统的相互兼容和融合，以及软硬件集成运行时的稳定性和可扩展性等问题。腾讯云在这方面做了必要的技术储备，除了适配国际上不同型号GPU外，还针对国内AI加速卡进行适配，为市场提供多样化算力选择。

下一代AI大模型的训练推动十万卡集群的探索。目前，国内外头部科技企业、云服务商以及科研机构正在逐步解决万卡集群建设和运行中的诸多难题，万卡集群的建设已在加速实现，并且在逐步迈向十万卡集群水平。在国内，多个万卡及以上规模的AI算力集群正在陆续建设；腾讯基于自研的高性能网络星脉，以及新一代算力集群HCC，可以支持10万卡GPU的超大计算规模；国际上，马斯克创立的xAI公司训练Grok2将采用约2万卡的H100集群。马斯克进一步透露计划建造由10万张H100组成的超级算力集群，用于 Grok3 的训练。

AI算力设施发展“质”、“量”兼顾，集成、网联和分布式将成破局三板斧

AI算力设施应激式发展的副作用显现，全球呼吁“质”、“量”兼顾的可持续计算。生成式AI的进展速度不断的加快，不断给我们带来对技术革新的惊喜和赞叹， Scaling Laws逐步也成为业界共识。然而，“大力出奇迹”的粗旷发展思路，也导致了全球 AI Infra 建设的应激式响应，除了 AI 加速卡等器件的价格上涨外，激增的高功率智算集群给社会、环境以及投资企业带来的负面影响也在陆续显现，并有加剧的风险。

AI Infra 的未来发展，需要“质”、“量”兼顾。包括国际经济论坛（WEF）、英特尔、英伟达、IBM、谷歌等在内的众多国际组织和企业纷纷呼吁“可持续计算”的发展，在不断提升计算能力的同时，重视提高算力设施的利用率（Model FLOPs Utilization，简称MFU）和能效（每瓦电能所实现的AI运算次数），控制和降低AI infra对能耗、环境等方面的负面影响，从而在AI发展持续递增的行业背景下，为“量”的进一步增长提供发展空间。

集成、网联和分布式将成为 AI Infra可持续发展的破局三板斧。可持续计算的具体实现几乎涉及 AI Infra从底层物理器件到上层模型算法的所有方面，各种技术更新和优化措施的效果和周期也都不尽相同。当前，AI Infra “质”变所面临的基础问题，是算力集群的高能耗和低能效。 而从解决问题的关键性和经济性两方面来看，硬件持续集成、高性能网络互联、以及分布式训练优化将可能成为破局的三个技术路线。

硬件的持续集成，主要指从芯片、加速卡到模块和机柜等各个硬件层面元器件、组件的迭代并持续集成，这也将是未来一段时间 AI Infra 核心硬件系统主要演变路线之一。

高性能网络互联，是组建大规模算力集群的关键技术，主要解决不同芯片单元、加速卡、节点以及机柜乃至集群等各级计算系统之间的高性能数据交换。大模型训练一般需要TB每秒级别的互联带宽和毫秒级的延迟标准。不断提升的网络互联技术，一方面有助于提高系统集成度，使十万卡甚至更大规模的集群得以实现，另一方面也提高数据交换效率，降低能耗。腾讯自研的高性能网络解决方案“星脉”，专为大模型训练等大规模并行计算场景打造，采用自研端网协同协议TiTa，支持基于RDMA的计算节点间互联，最高带宽可达3.2TB/s，最大支持单集群10万卡的组网性能。

提升大规模分布式训练的计算效率一直是该领域核心问题，分布式训练框架便是关键的 AI Infra 环节。分布式训练框架是将大模型训练任务进行分解和并行策略指定、并进一步调度和管理AI算力集群按策略执行训练任务的关键软件系统。合适的分布式训练框架和持续的针对性软硬件系统调优，可实现更高的算力利用率，节省训练算力成本。腾讯自研Angel机器学习平台面向大模型训练，在预训练、模型精调、强化学习多个阶段进行优化，相比业界开源框架，可以用更少的资源训练更大的模型，训练速度是主流框架的2.6倍。

集成、网络互联和分布式训练优化将为 AI Infra 向质量兼顾的新发展阶段打开局面，与此同时单晶元芯片、液冷、分布式数据库、神经形态计算等各其他层面的持续优化和技术创新也会在未来几年取得新的进展，推动 AI 基础设施持续进化。

趋势2

推理分析：LLM带来推理能力跃迁，推动智力即服务

作者：徐思彦

大型语言模型（LLM）为人工智能带来了所未有的推理能力，极大地扩展了机器的认知边界。它们不再仅仅是执行简单任务的工具，而是成为了人类的"智力外脑"，能够提供深入的分析、创造性的解决方案和复杂的决策支持。这种推理能力的跃迁得益于LLM在理解和生成自然语言方面的巨大进步。它们能够解析复杂的文本，提取关键信息，进行逻辑推理，并生成连贯、有见地的回应。这使得LLM能够处理各种知识密集型任务，如法律分析、市场研究、科学发现等，为个人和企业提供了强大的智能支持。

思维链的生成

与以往的人工智能相比，大语言模型的显著特征是推理能力的强大表现。推理能力是指模型在处理信息时，能够进行逻辑推导、分析和解决问题的能力。通常体现在能够对复杂问题的理解、对信息的整合以及在给定上下文中生成合理、连贯和有说服力的回答。

如同人类学习语言一样，AI大模型通过大量数据的学习和模仿，逐渐构建起丰富而高效的模型。在训练阶段，大模型通过深度学习技术，通过多层神经网络，对接收输入的海量数据进行学习和优化，并通过学习调整模型的参数，使其能够对输入数据进行准确的预测。推理（Inference)阶段则建立在训练完成的基础上，将训练好的模型应用于新的、未见过的数据。模型利用先前学到的规律进行预测、分类或生成新内容，使得AI在实际应用中能够做出有意义的决策。这些模型利用深度学习架构，如Transformer，来捕捉文本中的长距离依赖关系，并通过注意力机制聚焦于输入序列中与任务最相关的部分。此外，通过使用启发式算法如贪婪解码、束搜索或思维链（Chain of Thought）Prompting等策略，LLM能够生成连贯且逻辑性强的文本，展现出在复杂问题上的推理能力。

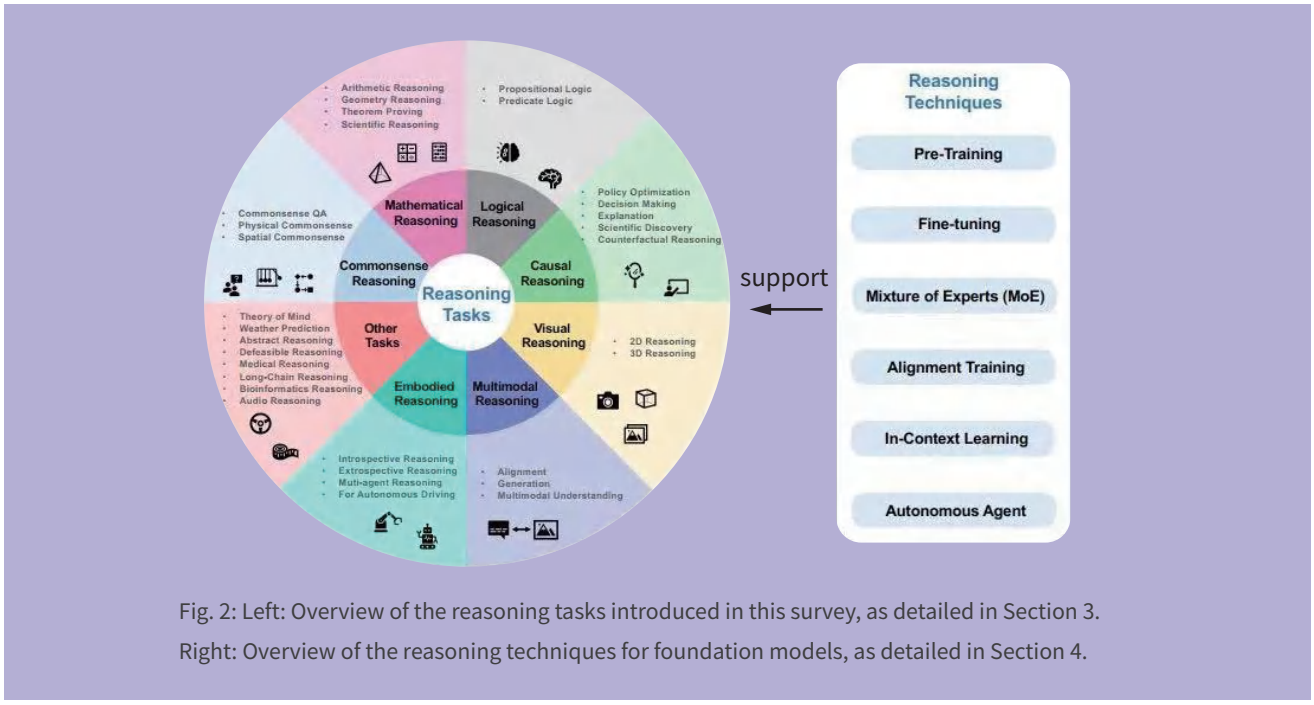


Fig. 2: Left: Overview of the reasoning tasks introduced in this survey, as detailed in Section 3. Right: Overview of the reasoning techniques for foundation models, as detailed in Section 4.

机器外脑开启“智力即服务（IQaaS）新模式

更高的算力与更好的模型的加持下，机器不再仅仅是执行简单任务的工具，而是成为了人类的"智力外脑"，能够提供深入的分析、创造性的解决方案和复杂的决策支持。它们能够识别问题的核心要素，构建逻辑框架，并通过逐步推演得出结论。这种能力使得大模型在法律、医疗咨询、科学研究等高知识密度领域展现出巨大潜力，为用户提供了更为精准和深入的服务。过去几十年我们一直在追求更快的计算机，但现在和未来，我们将追求更强大的大脑。Andrej Karpathy 提出了“软件2.0”，即传统软件开发转变到以AI模型为核心的新时代。未来，我们将依赖于无尽的算力和多样化的AI模型来构建“机器之心”，这将使我们能够实现前所未有的智能服务和应用。

AI将引领新的服务模式，即“智力即服务”，它通过云端平台为用户提供了一种灵活、高效的人工智能使用方式。过去的SaaS服务通常按每个账户定价，本质上是以员工成本为基准，以提高员工的生产力。在大模型时代，直接出售工作成果开辟来了新的垂直机会。LLMs为创业公司提供了一个机会发掘以前受到软件的市场推广和定价限制而无法涉足的领域。为此，创业者应考虑的不仅仅是出售软件以提高端用户的生产力，而是思考如何出售工作需求包本身。

与传统的本地部署相比，智力即服务模式允许用户根据实际需求快速调整资源，无需大量前期投资即可获得专业的AI能力。企业和个人可以根据自己的具体需求，选择相应的LLM服务，如文本分析、情感分析、机器翻译等。这些服务以需求包的形式被调用，并且可以轻松集成到现有的工作流程中，提高效率，优化决策过程，并创造新的商业价值。例如，企业可以把优化营销结果的目标拆分成“利用LLM进行市场趋势分析，预测消费者行为，制定更有针对性的营销策略”，然后通过“智力即服务”获取这些服务。这种模式不仅降低了技术门槛，还通过持续的更新和优化，保证了服务的先进性和可靠性。

此外，大模型的专业化和定制化服务能力，使得不同行业的企业都能获得符合自身特定需求的智能解决方案。这促进了企业运营效率的提升，支持了基于数据分析的决策制定，同时激发了创新和新产品开发。智力即服务还强调了成本效益和安全性，帮助企业优化成本结构，同时确保用户数据的安全和隐私。随着大模型技术的不断进步，智力即服务正在成为推动各行各业数字化转型的重要力量。

机器外脑助力个体成为“超级生产者”

对于个体而言，大模型技术迭代加速、衍生的AI原生产品层出不穷，并非单纯是AI对人的能力的替代。LLM的推理能力也为个人提供了巨大的便利。无论是学术研究、创意写作还是日常决策，人们都可以借助LLM获取深入的见解和建议。个体借助大模型应用、通过与AI进行协作，能够有效拓展能力边界，在生活和工作场景中成为一名擅长“人机协作”、充满创意和效率的超级生产者。随着大模型技术向多模态、端侧智能和Agents（智能体）等前沿方向演进，其在创作领域的应用潜力将持续赋能个体进行更高效、更具创造性的创作。未来，我们将迎来一批具备以下特质的超级生产者：熟练掌握与AI协作的方式；具备跨学科和跨领域融合的思维；擅长建立和拥有个人品牌和网络：拥有强大的个人品牌和广泛的专业网络；具有创新动力和能力；拥有技术伦理意识。从这一视角出发，超级生产者也有潜力成为未来零工经济的主力。

随着技术的不断进步，LLM的推理能力将变得更加强大和精细。我们可以预见，未来，iQaaS使人类的推理能力得以在云端实现，智力将有望变成像电力一样的公共服务获取。这不仅将极大地推动社会的整体智力水平，也将为个人和企业带来更多的发展机遇和创新可能。

趋势3

创意生成：AIGC应用爆发，降低专业创作门槛

作者：陈楚仪 王鹏

在这个精神追求引领物质需求的时代，AI的进步与社会文化的演变紧密相连，专注音乐和视频生成的AI平台应运而生，为热爱创作的“斜杠青年”们提供了更低门槛的工具，创建了自我表达和创意释放的新社区。大模型的崛起并非仅仅是人工智能对人类能力的替代，更开启了人与AI协作的全新篇章。

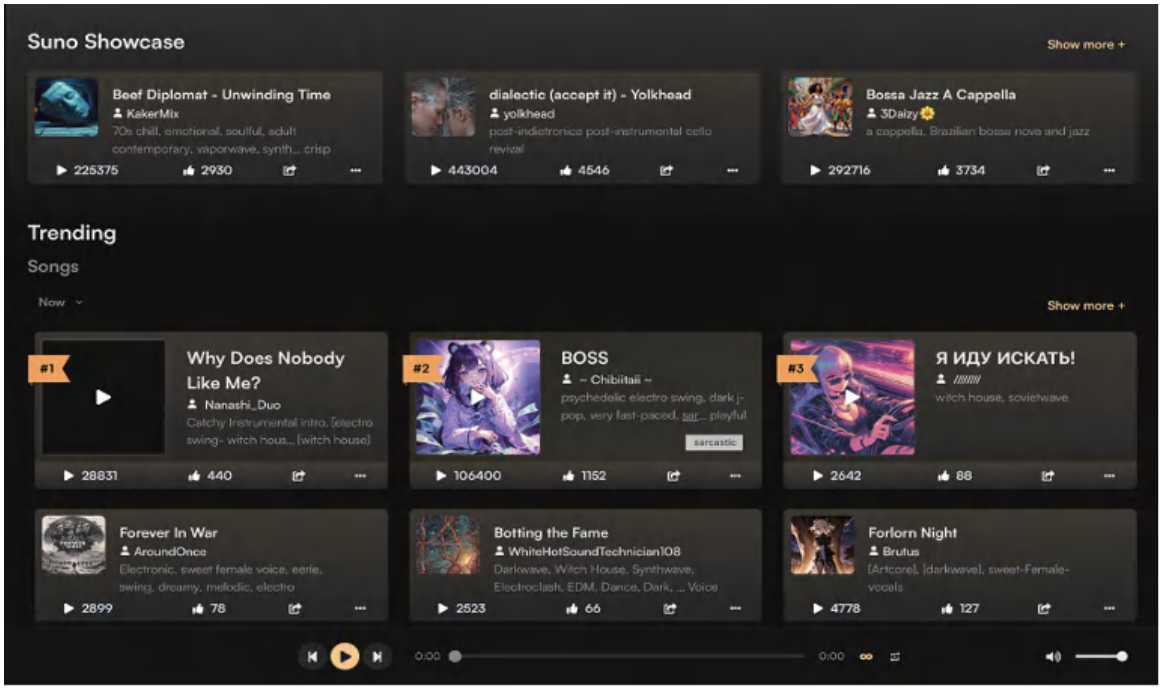
在内容创作与创意生产这一领域，大模型正以前所未有的方式重塑行业格局：（1）多模态内容：大模型可以生成涵盖文本、图像、视频等多种媒体形式的内容，提供更丰富的感官体验，提高信息的传达效率和吸引力。（2）个性定制：通过分析用户的行为、偏好和反馈，大模型可根据用户需求生成个性化、高度相关的内容,提高创作效率。（3）创意激发：AI可以分析大量的艺术作品，找出常见的主题和模式，然后提出创新的组合,为创作者提供灵感和创意灵感。（4）整合协作：大模型整合不同创作者的输入，实现跨界协作。

音乐生成模型拓展个体创作能力，降低专业创作门槛

“斜杠青年”在当今时代已成为一种多重身份和多样技能融合的生活态度和职业选择，而大模型会进一步推动艺术创作领域的“斜杠青年”群体壮大。一方面，个体通过大模型技术所带来的低成本创作工具，激发创造力和拓展创作能力边界，从而超越单一角色的限制，在社交圈层拥抱多元化身份。另一方面，AI艺术平台为向往“诗和远方”的普通个体提供了一个展示自我审美的表达空间。艺术创作领域的人机互动与人机协作，深刻影响了艺术创作方式，共同推动个体能力的延展和社会文化的演进。

正是在这种背景下，AI技术在音乐创作领域的应用显得尤为突出。音乐创作的动机，是音乐创作过程中的灵感来源和动力，也是作曲家创造音乐作品的核心要素。通过Suno、Udio、Elevenlabs这些AI音乐生成平台，没有任何音乐专业背景的个人也可以将自己的灵感或音乐动机转化为prompt（指令）形式，输入创作要求，选择音乐风格，借助生成模型创作出媲美专业水准的作品。对于普通人而言，这些平台简化了音乐制作的流程，降低了参与音乐创作的门槛，极大地拓宽了艺术创作的广度和包容度。对于专业创作人而言，AI音乐生成平台也是实验不同的音乐风格和旋律组合、发现新的创作路径的生态圈。

此外，这些平台采用类似MusicLM的架构和TTS（语音合成）技术，将音乐生成、人声合成和歌词创作等多项技术融为一体，将音乐创作和表演推向了全新的高度，让普通人不仅能“作曲作词”，还能当“音乐制作人”。技术发展还为新型AIGC音乐创作社区的形成提供了新的契机，目前Suno平台已汇集了包括获得格莱美得奖艺术家在内的1000万用户。



图：个人制作的音乐可以在Suno上线得到用户围观、点评，推动AI音乐社区形成

多模态AI改变视频内容生产模式

以Sora为代表的多模态大模型的推出，标志着视频生成技术进入了一个新的阶段。它巧妙地将视频内容重构为时空补丁，借鉴了大语言模型处理多样化文本的技术，为视频和图像内容的生成提供了一个高度可扩展和有效的表示。当前，众多AIGC产品已经能够制作出长达一分钟、高保真的视频，初步展现了模拟现实与数字世界的复杂场景。此外，主流产品还能够接受图像或视频等其他形式的输入，执行一系列的图像和视频编辑任务，如创建完美循环的视频、为静态图像添加动画以及在时间轴上延伸视频等多样化编辑任务。AIGC视频生成工具（如Sora、Pika、Runway）可能重塑“策划—筹备—拍摄—后期”的传统视频制作范式，带来输入提示词即可生成视频的“提示交互式”新范式。Adobe公司已经在主流的视频编辑工具Premiere中植入了Sora 等文生视频模型。

在Sora出现之后，DiT架构大大提速了相关领域的技术进展，类似技术和产品层出不穷。近期生数科技和快手可灵，都已经在时长和效果上达到甚至超过 Sora 水平，而且可灵已经可以直接申请使用。视频生成技术的应用正在逐步扩展到多媒体内容创作、游戏开发、虚拟现实等领域。这些技术的进一步发展和优化，不仅将重塑内容创作生态，也将重新定义我们与数字世界的互动方式。

发布日期	2023.11	2023.11	2023.11	2023.11	2023.12	2024.2	2024.4.22	2024.4.27	2024.6.6	2024.6.6
公司	Runway	Pika Labs	StablilityAI	Meta	斯坦福+谷歌	OpenAI	抖音	生数科技	快手	极佳科技
产品	Gen-2	Pika1.0	Stable VideoDiffusion	Emu Vide	W.A.L.T.	Sora	即梦 Dreamina	Vidu	可灵 Kling	视界一粟 YiSu
时长	4s	3-7s	2-4 s	4s	3s	60s	3s	16s	120s	16s
特点	可延长至18s	运动笔刷	开源	扩散模型	Transformer	物理规律	易申请，免费	U-ViT首个DiT 架构	申请可用	可端侧运行
是否已可用	是	是	是	否	否	否	是	否	是	否

2024年5月14日，腾讯宣布其混元文生图大模型进行全面升级并开源，这一举措为企业和个人开发者提供了免费商用的可能性。作为业内首个中文原生的DiT（Diffusion Models with Transformers）架构文生图开源模型，混元文生图大模型支持中英文双语输入及理解，拥有高达15亿的参数量。

混元文生图大模型采用了与Sora模型一致的DiT架构，即Hunyuan-DiT架构，这不仅使其能够支持文生图，也为视频等多模态视觉内容的生成奠定了基础。在性能上，采用Hunyuan-DiT架构的腾讯混元文生图大模型超越了开源的Stable Diffusion模型，成为当前效果最为卓越的开源文生图模型之一，其整体能力达到国际领先水平。同时，它还与原有的SD生态保持了良好的兼容性，使得开发者可以以较低的成本进行迁移和应用。腾讯混元文生图大模型的开源，为构建以中文为核心的文生图开源生态系统提供了强有力的支持，有望催生出更多具有本土特色的原生插件，进一步推动中文文生图和视频生成技术的研发和应用。

多模态AIGC技术将从以下方向影响视频内容生产模式，使得个体有机会创作低成本、高质量的视频产品，也为影视领域的“超级生产者”诞生带来新的机遇：

（1）效率革命：AIGC技术通过自动化算法极大地加速了从概念构思到成品制作的整个视频生产流程，显著减少了传统视频制作中的时间消耗和繁琐步骤。例如，Sora模型能够根据文本提示生成长达一分钟的视频，这在传统制作流程中可能需要数天甚至数周的时间来完成。

(2)成本优化：AIGC技术的应用减少了对专业团队的依赖，通过智能化的视频生成和编辑，大幅降低了人力和试错成本，使得高质量视频内容的制作变得更加经济高效。

(3)个性化定制：AIGC技术利用用户数据分析，实现了内容生产的个性化定制，为用户带来更加贴合个人偏好的视频体验，同时为企业提供了更精准的市场定位和增强的用户粘性。

多模态AIGC技术改变了视频内容生产模式，使创作者不再受限于传统影视制作中对高度专业化流程和团队的依赖。其影响是复杂的：一方面，AI所带来的制作门槛降低，将会让高度专业化的影视制作行业受到冲击，大量专业化人士引以为傲的技能可能会被部分技术替代。另一方面，借助生成技术，创作团队能够更加专注于创意和情感的表达，以及故事的叙述，甚至创造出前所未有的听觉和视觉效果以及叙事手法。2024年3月，全球首部完全由AI生成的长篇电影终结者2翻拍版《Our T2 Remake》发布，影片由50位艺术家花费3个月的时间完全使用AI技术创作而成，所呈现的画面展示了AI在视频制作中的巨大潜力。

图：AI生成的长篇电影《Our T2 Remake》

展望未来，大模型不仅提升了现有行业的效率和产出质量，还极大程度降低艺术创作的门槛，加速“超级生产者时代”到来。Sora、Suno等现象级产品，让普通人有可能创作出媲美专业水准的作品，并且快速渗透到媒体、影视、音乐行业，为创作注入前所未有的发展动力和想象空间。基于大模型的软件和平台不仅仅是技术的应用，也是个体追求个性化表达和自我价值实现的新场域，更是技术、社会和文化的聚合点。

趋势4

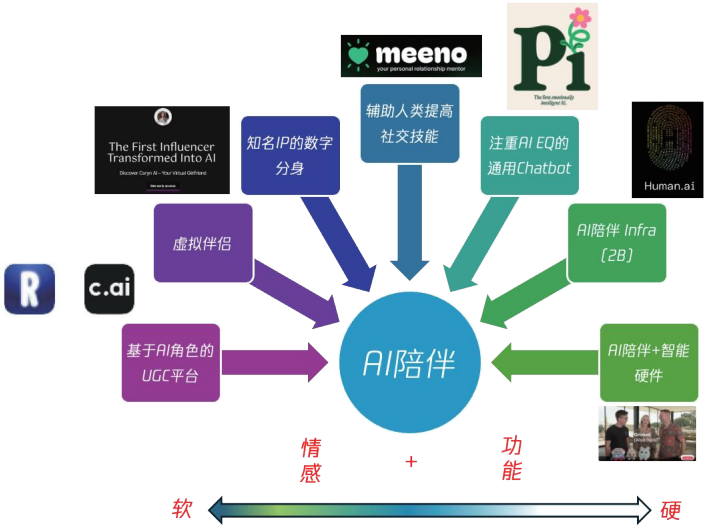
情绪智能：多模态大模型赋予机器情感价值，打开人机陪伴市场

作者：白惠天

兼具EQ与IQ的大模型将在未来2-3年内打开人机陪伴市场。基于最新的AI模型如GPT-4o和Gemini 1.5 Pro，未来AI陪伴将通过流式语音识别、多模态AI和情感计算等技术极大地提升互动体验。在个性化方面，AI通过深度分析用户的情感和行为提供精准服务；在共创方面，AI能够实时理解和回应用户的意图，共同创造一个可交互的丰富世界；在平权方面，多语言支持和辅助技术使不同背景的用户无障碍交流。为实现情感交互和多样性表达，如何做好大模型的长期记忆和持久性是需要突破的核心技术难点，其中涉及两个环节：一是记忆系统的建立，另一个是“回忆策略”的设定。此外，数据隐私、算法偏见和心理依赖等伦理和隐私问题也需要得到充分关注和解决，以确保技术的公平性和安全性。预测未来人机陪伴市场将从以互动游戏、兴趣社区为主的年轻人市场，进一步破圈到包括各年龄层的更广泛用户群体，通过不同级别的情感理解、发散性、准确性、物理交互水平解决用户的多元化陪伴诉求。

兼具情商（EQ）与智商（IQ）的大模型将在未来2-3年内打开人机陪伴市场。据a16z 3月发布的Top50AI产品榜单，与去年9月的榜单相比，AI陪伴产品在Top50的占比由2个大幅增至10个。包括Character.ai（估值50亿美元），星野/Talkie（估值25亿美元），Pi（估值40亿美元），小冰（估值20亿美元）等在内的独角兽不仅估值涨幅明显，用户参与度也显著高于其他AI应用。例如Character.ai目前月访问量超2亿次，流量仅次于ChatGPT和Gemini。

由于用户的需求和动机不同，AI陪伴的产品形态千差万别。譬如是否需要一个具体的形象、是否需要配合硬件使用、是独立应用还是与现有社交平台打通、是否在提供情感价值的同时满足心理辅导、1对1教育等实际性功能……不论聚焦于哪个垂类，面向未来的人机陪伴产品有如下共性特征：真正地理解用户，并在此基础上做到足够的个性化。



流式语音识别、多模态AI和情感计算等领域的突破为AI陪伴奠定了技术基础。基于最新的AI模型如GPT-4o、Claude 3和Gemini 1.5 Pro等，大模型流式语音识别的响应时间已缩短至200-300毫秒，情感识别准确率可达80-90%，并且可同时处理文本、语音、图像和视频多模态信息。未来随着上述技术的持续迭代，AI陪伴有望进一步提升互动体验，变得越来越有人情味。具体来说，一个能够提供情感价值的AI陪伴将兼具个性化、人机共创和普惠平权三大特征。

AI陪伴首先是个性化的。情商（EQ）是高度个性化的概念——不同的人对情感和社交动态的理解有所不同。精调（Fine-tuning）可以实现根据特定用户群体（比如儿童）或任务领域（比如恋爱）的需求来调整模型，使其更加符合个性化的场景和需求。目前的AI陪伴产品能够记住用户的偏好、兴趣和情感，根据对话场景实时调整互动方式，并通过个性化的语音、图片等等把对于用户的理解嵌入进产品体验中。未来随着实时情感分析、自适应学习等技术的进步，AI将能够更准确、高效地理解用户的情绪状态，提供细致入微的服务。例如在多模态情感计算的加持下，AI有望进一步打通多种感官，通过结合视觉（如表情）、语音（如语调）、文本等多种感知方式来更全面地理解人类的情感状态。再结合可穿戴设备所采集的心率、皮肤电位等信息，AI陪伴有望做到比用户更懂用户。甚至是先己一步，以最适合用户的方式提供情感慰藉与支持，实现自然、贴心的人机交互体验。

在高质量陪伴的过程中，核心的互动内容一定是人机共创的。理解用户、实现个性化需要丰富的上下文内容，为此，就要设定情景让AI与用户创造一些独特的共同经历——例如，一个引人入胜的故事、一个丰富可探索的世界、一个充满神秘感需要用户不断挖掘的角色……形态上也可能是文本、音频、视频、游戏或者这些模态的结合。目前的AI陪伴产品利用生成对抗网络（GANs）生成高质量的创意内容（例如Replika用户可以与AI一起编写小说或剧本），除此之外，用户的行为能够真实地对产品中的世界产生影响，并且这些影响是可以通过一些具象化的方式展现出来的，比如某些游戏产品中设置的抽卡环节等。未来，通过增强自然语言处理和多模态整合，AI陪伴产品将进一步提升用户与AI共同创作的体验。特别是流式语音识别技术的进步使AI能够实时理解和回应用户的语音输入，提升创作效率。这种技术进步将使用户能够更便捷地与AI合作完成复杂的创意项目。

最后，AI陪伴有望弥合数字鸿沟，以低门槛的形式走进寻常百姓家。过去，长尾群体（如小语种、少数族裔、残障人士、低收入群体）的诉求很难被看到，针对他们的陪伴往往成本高企、触不可及。目前的AI陪伴产品主要通过如下技术手段多语言推动技术普惠与平权：语言方面，实时翻译技术通过流式语音识别技术消除了语言障碍，促进了跨文化交流，为不同国别、种族、语种背景的用户提供了平等的交流机会。此外，公平算法的应用确保了AI在训练和应用过程中的无偏见，进一步提升了不同背景用户的使用体验。辅助技术，如语音识别和手语翻译等，帮助有特殊需求的用户顺畅使用AI产品，提升了用户的包容性和可达性。这些技术的发展不仅关注于提高翻译效率和准确性，还包括了对特定群体如留守儿童、LGBTQ社区的支持（例如AI心理健康支持），打击歧视和偏见。尽管技术的发展为实现社交平权提供了工具，但仍存在一些挑战，譬如“算法黑箱”持续存在、性别中立翻译问题在跨语言环境中的应用仍然是一个未被充分探索的问题。

尽管大模型的双商（IQ&EQ）都在不断提高，解决好长期记忆和持久性问题仍是尚待攻克的技术难点。由于大模型本身不具备长期记忆，目前大多数 AI 陪伴产品只能进行片段式的对话交互，每次对话都是从头开始的，而这恰恰是陪伴类产品的大忌。用户在与AI互动时会期望AI能够记住他们之前的对话和互动内容，从而提供连贯和有意义的回应。否则人机交互就无法连续进行，个性化服务的精准性也会大打折扣。目前哪怕在头部捏崽类产品中，伴随着对话的深入也会出现“脱皮”“下皮”等问题，即人物忘记用户创建的初始人设特征。这严重影响了人机之间建立深层次亲密关系的能力。

目前长期记忆的处理方法还存在很多不自然的地方，其中涉及两个环节：一是记忆系统的建立，另一个是“回忆策略”的设定。记忆体系建立方面，简单地把跟用户的对话历史（含文字、图像、视频甚至表单、PPT等）堆到长文本中，作为后续交流的“Memory”，这种处理方式会造成信息丢失的问题。因为过去的普遍做法是各类信息/实体先转换为文字再映射到同一向量空间中，以提示词（prompt）的方式输入给大语言模型（LLM）。转文字模态会导致信息丢失或扭曲。随着GPT-4o等模型的出现，记忆系统建立方式发生了系统性变化：海量信息通过多模态解析（multi-modal understanding）技术采用统一嵌入（unified embedding）的方式直接映射到多模态模型（multi-modal model）。这在一定程度上可以减少转换中的信息损耗，但也并不完美，因为这还是机器的记忆方法，而不够像人。人类会对过去的经历进行整理和抽象，形成各种各样的“总结”，还受到当时的情感、情绪等因素的影响。这些“总结”会作为长期记忆留存在大脑中，目前围绕大模型记忆的许多探索都在做类似的工作。回忆策略设定方面，长期记忆所形成的“总结”如何在后续对话中被“唤醒”（recall），根据当下任务的需求去检索相关的内容进行调用，并形成某种启发或者共鸣——这是目前大模型正在攻关的难题。比如，当你的聊天对象很伤心需要安慰时，你是调用过往开心的记忆还是不开心的记忆来安慰他?根据场景和用户对象不同，选择很可能不同，这背后涉及到很复杂的调用策略。

未来3-5年，情感陪伴类AI产品能否在打开市场之后持续破圈，关键就在于长期记忆问题。这个卡点一旦突破，AI陪伴可以显著提升个性化服务和人机交互共创内容的质量。这不仅提高了用户体验，还增强了用户对AI的依赖和情感连接。通过记住用户的偏好、兴趣和互动历史，AI能够提供更加贴心、连续和个性化的服务，满足用户的多样化需求。

当然除了技术上的痛点，数据隐私、算法偏见和心理依赖等伦理和隐私问题也需要得到充分关注，以确保AI时代人机陪伴的公平性和安全性。如果上述问题能够得到妥善解决，有理由预测未来人机陪伴市场将从以互动游戏、兴趣社区为主的年轻人市场，进一步破圈到包括各年龄层的更广泛用户群体（K-12、大学生、年轻打工人、中年人、老年人），通过不同级别的情感理解、发散性、准确性、物理交互水平解决用户的多元化陪伴诉求。

趋势5

智能制造：多模态大模型技术升级，提升工业新质生产力

作者：刘俊 刘永 汪铖杰

2024年被普遍认为是大模型应用落地的元年，而工业场景将会是大模型的重要战场。工业生产包含复杂的流程，为AI落地提供了丰富的场景。未来多模态大模型有望与当前普遍使用的专用小模型互补融合，并深度赋能工业制造的各个环节，从而推动生产制造的提质增效。例如在研发设计、生产和管理等环节的应用有助于提升生产效率、产品品质和任务处理能力。大模型与产业深入融合以及多模态的混合交互模式的出现，有望重构智能制造系统并推动工业智能化。另一方面，随着场景数据的整合和积累，多模态大模型的感知和理解能力将进一步升级，以满足生产制造中的个性化需求。如提升垂直场景中超微小目标感知及超细粒度语义理解，强化对复杂多模态Prompt的理解能力等。未来借助PEFT等技术的发展，大模型+小样本数据适配的模式将成为模型更新新范式，极大地降低了专属数据量需求。应用中“大模型+工业？”的落地模式将迎来爆发，从而推动工业产业变革，助力人类社会迈向更高层次智能化发展。

工业场景是AI大模型的重要战场，未来5~10年最大的机会

2024年，各大公司推出强大理解能力的多模态大模型，将引领人工智能技术创新和应用，工业场景将成为多模态大模型的最佳实践场地。随着GPT-4o,Gemini1.5Pro，LLaVA1.6的发布，基于Transformer架构和海量数据训练的多模态大模型再次点燃AGI，其对文本，图像等多模态输入的支持和强大的理解能力也象征着人工智能迈向通用人工智能（AGI）的新阶段。随着工业智能化的推进，大模型有望重构智能制造的系统，为工业智能化提供新动力。

工业场景将成为大模型最佳的“练兵场”，据专业机构统计，2023年我国全部工业增加值约40万亿元，而当前多模态大模型在应用中部署仅占了8%，未来存在巨大的上升空间。工业生产，整个流程主要分为研发，生产，管理等环节，为AI落地提供了丰富的场景。在当前阶段工业AI的应用主要以专用的小模型为主，专用小模型在具体的某个细分场景会针对性地收集较大的标注数据，如在工业质检领域通常每类缺陷需要收集百张以上的产品图，然后对该场景进行精细化调优，从而保证了模型的准确性和稳定性。而专用小模型定制化的设计模式制约了其在应用中的进一步渗透，其在研发和交付过程中依赖较长周期的数据采集及较复杂的个性化定制的业务逻辑功能，导致模型通常无法在其他场景通用。多模态大模型借助强大的理解能力，泛化能力和丰富的多模态输入的支持，有望和小模型进一步互补融合，为工业制造注入新的动力。未来、AI应用技术会加快往工业领域迁移，通过场景适配和多种部署形式，最终实现工业大模型的落地赋能。

多模态大模型深度赋能智能制造，将进一步推动生产制造的提质增效。在大模型的驱动下，工业场景下的丰富数据有望进一步整合并开源，如MvTec，Real-IAD等，为多模态大模型奠定并丰富数据基础，促进行业大模型的快速发展，同时反哺生产制造的提质增效。在研发设计环节，大模型可以通过对结构化数据以及产品结构的理解，能辅助产品设计图的布局优化，参数校核并可进行动态仿真，进而缩短研发周期，提升研发设计的效率。在生产环节，大模型强大的理解能力和迁移学习能力可以对不同生产场景中的产品质量缺陷，零部件装配，工作人员作业规范进行高效的视觉感知，进而提升产品生产的品质，降低质检人员数量，并对厂区安全合规全方位监控。例如在缺陷检测中，通常需要大量的质检人员对产品质量缺陷进行识别并分拣，存在人力成本高、效率低、检测结果一致性弱等问题。大模型的强大泛化能力和迁移学习能力能进一步强化机器视觉的能力，未来只需少量数据进行场景适配调优，即可达到客户要求。在管理环节，文本、文档、图片甚至视频数据会进一步整合，从而强化多模态大模型的能力，能全方位对异常图像，故障数据，文档数据进行分析理解并自动化处理，快速提升任务处理的效率。

大模型的落地，助力重构智能制造系统，推进工厂智能化，提升新质生产力。工业发展是一个逐步演化的过程，经历了机械化，电气自动化，信息化等阶段，正处于从数字化迈向智能化的阶段。在迈向智能化的过程中，工业和各类创新技术的融合进而对工业生产体系进行智能化升级和改造，以提高生产效率，降低成本，提高产品品质和生产安全。当前，工业制造累积了大量的数据，为大模型提供了良好的基础条件。随着高质量数据的积累，大模型的理解能力会大大提升，首先将会在生产的各个环节得到大量的部署和应用。在多模态输入的加持下，文本、图像、语音等混合交互模式进一步提升生产效率，智能化感知和交互也将重构整个制造管理体系。未来有望实现“智能感知，智能决策，智能执行”的全新智能化工厂。

技术加快向工业领域的迁移，多模态大模型的能力升级，更好满足个性化需求

目前国内外推出的主流多模态模型乃是基于自然场景数据集的基础模型，通常能理解的物体和语义粒度较粗，未来需要在垂直场景实现超微小目标感知及超细粒度语义理解，如工业AI质检。识别目标的细微程度远超日常所见物体，实际尺寸往往精细到亚微米级别，其在图像中的占比甚至仅为百万分之一，这要求多模态大模型需具备高度精准的像素级识别能力。另一方面，工业场景落地场景复杂要求高，部分场景数据有限，未来需强化对复杂多模态Prompt的理解并提升数据的利用效率以最大化多模态模型的理解能力。

工业多模态大模型对超微小目标的感知和超细粒度语义理解能力有望快速提升。在工业生产的各个生产环节，为满足生产的要求，要感知的目标非常小，语义描述也趋于超细粒度描述。得益于现有强大的开源大语言模型以及大量自然场景下的图像-文本数据集，目前自然场景的多模态大模型通过将视觉特征嵌入到大语言模型的语义空间，借助预训练大语言模型的泛化能力对多模态的输入有较好的理解。但是受限于输入图像分辨率和计算资源的限制，通常处理的图片分辨率在百万像素以下，其对图片中主要物体、如大于100x100pixel的目标感知能力较强，而局部细节的感知能力不强，对超小目标和超细粒度语义理解能力较弱。为了缓解这个问题，Wisconsin–Madison、微软公司、清华大学等相继发布了LLaVA-1.6，LLaVA-Next，LLaVA-UHD系列方法，通过切分高分辨率输入图像成多个低分辨率的子区域，并将高分辨率输入图像下采样到低分辨率的版本，随后将这些低分辨率图像一同输入到大语言模型以提升模型对图像局部的感知和理解，使得多模态大模型对小目标的感知能力有一定的提升。随着上述多模态技术的更新迭代，以及更多细粒度的图像-文本数据集的增加，未来多模态大模型对超小目标的感知及细粒度语义理解能力有望显著提升。

未来将增强模型对复杂多模态Prompt的理解能力。现有的自然场景多模态大模型的Prompt指令数据集主要包含少量的人工标注图像-文本数据以及借助GPT-4V或Gemini-Pro等多模态大模型来合成的大量图像-文本数据，通过大量自然场景的多模态指令数据集进行指令微调能够实现较强的多模态Prompt理解能力。尽管在工业场景下能够获取到少量人工标注的图像-文本数据，但是现有的GPT-4V或Gemini-Pro等闭源领先的多模态大模型尚不具备在垂直场景下的精细理解和感知能力，未来在工业高质量数据积累下，有望将结合工业多模态大模型能力进行Prompt生成并针对性的进行技术改进，进而提升对复杂多模态Prompt的理解能力。

基础模型+小样本数据适配成为模型落地的新范式。在生产制造中，由于生产工艺的不同、引进设备及视觉方案不同，文件及描述规范不同等因素，AI在实际落地中普遍表现出“需求个性化”。针对某个细分场景，专用小模型优化需要收集较多的数据并需定制化开发业务逻辑，收集数据周期较长，通常需要持续8周及以上，某些场景甚至难以收集数据，导致指标提升慢交付效率低难以匹配生产节奏的要求。大模型在虽然有较强的理解能力，但由于缺少具体的场景数据，导致其无法充分捕捉到某个细分场景的专属特征，比如专属名词、专属描述、专属物体及形态，这种专业知识的匮乏使得大模型在应对工业流程优化、机器视觉缺陷检测、设备故障等专业问题时会有所缺陷，难以提供精确、可靠的解决方案，无法满足工业现场的个性化要求。大模型真正融入行业应用，需要进一步适配，来解决大模型的“不懂专属场景”的局限。近年来，随着PEFT（如IA3、Adapters、Soft Prompts、LoRA）技术的发展，所需的专属数据量能降低90%。未来高效利用有限数据来适配细分场景的技术会进一步得到发展。同时，高效的多模态大模型和轻量化部署也将成为落地应用中探索的重点。

工业人工智能探索日益活跃，未来“大模型+?”的落地模式会迎来爆发。比如“大模型+工业设备”，有望提升工业设备的协同性和智能化并驱动实现具身智能；“大模型+工业软件”能进一步提升查找、设计、识别的效率和精度；“大模型+工业知识图谱”带来新的交互方式并提高知识问答的准确性和效率。未来大模型技术将继续引领工业产业变革，推动人类社会向更高层次的智能化发展迈进。

趋势6

游戏环境：大模型与游戏共振共生，打造Agent最佳训练场

作者：王枢

大语言模型与游戏环境的相结合，正在为AI Agent训练打造最佳训练场。游戏不仅能为AI Agent训练提与现实世界类似的虚拟环境，还能为AI Agent训练提供清晰、可量化的评估规则，大幅提升技术迭代与测试效率。当前，包括OpenAI、DeepMind等在内的前沿研究团队都选取游戏作为AI Agent训练场景，致力于在不同类型的游戏场景中的打造通用AI Agent。

大模型与游戏共振共生，不断加速技术迭代与应用创新。未来2-3年内，基于游戏环境训练通用AI Agent将成为行业趋势，游戏将成为AI Agent训练的重要试验场。在大模型和游戏环境的加持下，AI Agent将有望实现决策和泛化能力的突破。

技术试验场：基于游戏环境的通用AI Agent实践

Google SIMA带来“AI智能体的ChatGPT时刻”

在人工智能领域，AI智能体（AI Agent）是指能够观察环境、作出决策并执行行动，以实现特定目标或任务的系统。AI智能体可以是软件形式，如聊天机器人、推荐系统、游戏AI等，也可以是集成到物理设备中的，如自动驾驶汽车、机器人等。

2024年3月13日，Google DeepMind团队发布名为SIMA（Scalable Instructable Multiworld Agent）的AI智能体（AI Agent），将其定义为一种在多重3D虚拟世界中可扩展、可指导的通用游戏智能体。DeepMind研究人员评估了SIMA 按照指令完成近1500个具体游戏内（in-game）任务的能力，涵盖导航（例如左转）、对象交互（爬梯子）和菜单使用（打开地图）等，并且在多个游戏环境中都表现出了高于同类智能体的性能水平。因其具有强大的自然语言理解和迁移学习的能力，不少研究人员将它的出现视为“智能体的ChatGPT时刻”。

游戏是Google SIMA训练是的重要试验场

游戏是人工智能（AI）系统的重要试验场，与现实世界一样，游戏也是一种丰富的学习环境，具有反应灵敏的实时设置和不断变化的目标。DeepMind团队选取了9款当下流行的3D网络游戏和4个基于Unity引擎制作的3D场景作为SIMA智能体的训练环境，并从游戏中收集了大量人类玩家的行为和操作数据，用以训练智能体。从围棋人工智能AlphaGO和AlphaZero，到基于游戏《星际争霸2》的AlphaStar，再到如今基于大语言模型的SIMA，DeepMind团队一直在基于游戏环境进行通用智能体的测试和研究，在他们看来，智能体在游戏环境中训练出的决策和行动能力，有望能够迁移到现实世界的场景中，为孵化通用人工智能提供新思路和新实践。

图：Google SIMA项目概述

SIMA项目是DeepMind团队在通用人工智能（AGI）研究领域的一个重要里程碑。与该团队之前发布的游戏智能体相比，SIMA训练的目的不在于击败人类玩家或在游戏内取得高分，而是学会在各种游戏环境中遵从人类发出的自然语言指令，并作出与指令一致的行为。SIMA在训练中引入了大语言模型，整个训练过程都遵循语言优先的规则，所有的训练行为都由自然语言直接驱动。也就是说，SIMA 既不需要访问游戏的源代码，也不需要定制的 API。它只需要两个输入：屏幕上的图像信息，以及用户提供的自然语言指令，即可使用键盘和鼠标控制游戏中的角色执行这些指令。DeepMind 创始人及CEO德米斯·哈萨比斯（Demis Hassabis）在采访中表示，“将大语言模型、AI智能体训练与游戏环境相结合的这个领域，有着巨大的发展前景，DeepMind未来将持续加大对该领域的研究投入。”

基于游戏环境训练通用AI Agent已经成为业内共识

早在SIMA发布之前，业内已经存在着多个通用游戏智能体研究项目，其中比较有代表性的项目分别是DeepMind发布的Gato，以及英伟达发布的Minedojo，这两个项目分别对应着人工智能研究中的两类不同思路：解决足够多的任务或解决一个足够复杂的任务。

Gato使用了类GPT的大语言模型架构，其训练材料包括图像、文本、机械臂关节数据以及其他多模态数据集(multimodal dataset)，可游玩雅达利系列游戏（Atari Games），并可操控真实的机器人手臂堆叠积木。微软在2023年3月的一篇研究中指出，Gato这类融合了多模态信息的大模型，极有可能诞生出初期的智能。

MineDojo以《我的世界》游戏的玩家视频（YouTube）、百科（Wiki）和用户社区（Reddit）的资讯作为训练材料，能够在《我的世界》游戏中根据文字提示信息，完成各种不同任务，不仅能够完成简单的程序化任务(programmatic tasks)，还可以根据简单描述完成创造任务（creative tasks），例如根据描述建造一个图书馆等。

此外，在TED AI 2023演讲上，英伟达高级科学家Jim Fan也提出了基础模型（Foundation Agent）概念，再次强调了游戏环境对于通用AI模型训练的重要性。Jim Fan认为，AI研究的下一个前沿将是塑造一个可以在虚拟世界和现实世界里泛化，掌握广泛技能，控制许多身体，并能够泛化到多个环境中“基础模型”，而这个模型的训练，同样离不开游戏环境。在国内，腾讯也牵头构建起AI多智能体与复杂决策开放研究平台——开悟，依托腾讯AI Lab和《王者荣耀》在算法、算力、实验场景方面的核心优势，为学术研究人员和算法开发者提供国内领先的应用探索平台。

应用新场景：大模型助力游戏创作，提升内容创作效能

伴随着以Stable Diffusion、Transformer等生成式AI技术的成熟，AI技术也开始反向助力游戏以及更广泛的文化行业的内容创作，越来越多的从业者能够以更低成本生成图片、文字、音视频、NPC等数字资产，提升产品研发效能，进一步降低交互内容的制作门槛。

游戏公司应用生成式AI模型提升研发效能

《2024 Unity 游戏业报告》显示，在使用AI技术之后，有71%游戏工作室表示其研发和运营效能得到了提升。在游戏内容生产侧，生成式AI已被广泛应用于文本生成、2D美术创作、代码生成于检测、关卡设计生成等环节。在AI工具介入游戏美术工作流程之前，游戏美术工作者完成一张高质量插画图的时间大概在一周左右，在使用Stable Diffusion等生成式AI工具后，可将时间缩短至1天。

在降低不同类型工作者沟通成本方面，生成式AI也有着巨大应用空间。例如在游戏制作过程中，尤其是在对游戏美术风格进行定调和选型时，游戏策划和美术工作者之间的沟通往往需要耗费大量的时间成本，生成式AI工具的介入，能够帮助策划者快速将创意落地并呈现，极大降低沟通成本。

科技公司基于生成式AI能力重塑工具平台

科技公司英伟达、Unreal 和Unity等也纷纷布局生成式AI领域，将生成式AI能力聚合到游戏制作的工具平台中。英伟达于2023年6月发布了面向游戏开发者的AI工具平台NVIDIA ACE for Games，让游戏开发者可以在游戏中构建和部署定制化的语音、对话和动画等AI模型，极大提升游戏内容生产和制作效率；在GDC 2024上，NVIDIA和Inworld 联合公布了一项全新的数字人技术 Covert Protocol，基于该技术塑造的游戏NPC能够与玩家进行实时交互，并且能够基于互动内容，实时生成游戏玩法。

引擎公司Unity和Unreal相继发布基于生成式AI的新工具。Unity于2023年7月发布两款基于人工智能技术的新产品：Sentis 和Muse，据悉两款产品可将传统内容创作的效率提升十倍；Unreal则在自身引擎中集成了大量AIGC工具，如数字人制作工具Metahuman creator，尝试以人工智能技术加速创作高质量的角色及大规模场景生成效率。

国内科技公司也全面拥抱AI技术，用AI赋能内容制作工具，不断提升内容研发效率。以腾讯为例，腾讯AI Lab发布了自研游戏全生命周期AI引擎“GiiNEX”，该引擎借助腾讯自研生成式AI和决策AI模型，面向AI驱动的NPC、场景制作、内容生成等领域，可提供包括3D图形、动画、城市及音乐等多种AIGC能力。在GiiNEX引擎助力下，原本需要5天才能完成的城市建模任务，现在只需要25分钟即可完成，效率提升达百倍。

图:腾讯游戏AI引擎GiiNEX架构图

尽管当下的人工智能研究距离实现AGI还有相当长的路要走，但大语言模型与基于游戏环境的AI Agent训练，无疑为实现AGI开辟了新的可能性。随着训练环境的不断增加，游戏中的AI Agent或将具备对更复杂、更高级语言指令的理解和能力，人们有望创造出更为灵活、适应性更强、更接近人类智能的AI系统。在未来人工智能技术的创新发展过程中，应进一步重视游戏产业的科技价值，明确游戏作为人工智能技术“实验场”的角色定位，更好发挥游戏产业在技术创新、应用创新和跨域反哺中的作用，助力数实融合快速发展。

趋势7

移动革新：端侧模型带来智能设备与应用入口变革

作者：李瑞龙

端侧生态已成为科技大厂竞争的焦点，端侧大模型结合AI芯片和操作系统，正在构建出一套完整的技术体系。目前，全球科技巨头如Google、苹果、微软以及国内终端厂商都在积极探索这一领域。智能终端在接下来几年，鉴于终端瓶颈、AI agent、端侧真需求等方面的考虑，端+云的混合模型可能更加符合未来长期的发展趋势，其中云端模型承担主要的计算和存储任务，而端侧模型则专注于满足用户的隐私保护和数据安全需求。在交互入口方面，SaaS正在全面推进AI化，AI功能已被集成到各种应用中，从而提升用户体验；随着AI原生OS的发展，操作系统可能会发展成API直接调用的模式，减少对传统图形用户界面的依赖，当前APP的交互服务形态将发生变化，回归本源，服务的深层价值决定着用户去留及时长。

侧生态已成为各大厂的必争之地。从技术上来看，端侧大模型不是孤立的技术，而是跟AI芯片（GPU/TPU）、操作系统共同形成一个完整的技术体系，目前，无论是事实上无论是Google、苹果、微软还是国内终端厂商，都已经向这条路线迈进。从最关键的基础模型迭代方向来看，也从此前追求Scaling Law、万亿参数模型的路线，转变为对终端更为友好的多模态小参数模型转变，比如最新的微软加载GPT-4o，Phi-3-vision等等；

端+云结合的混合模型将是长期存在的主流

来自于端侧的硬件瓶颈。目前大模型想落地端侧，存在比较明确的“智力”门槛。以Llama3 8B的FP16版本模型为例，作为目前未过分量化稳定可用的版本，其模型大小为16G，就目前市面上的终端能力来看，仅旗舰PC可用；除GPU芯片之外，端侧的普及还需要解决一系列硬件生态的问题，包括电池、显存带宽、显存容量等等。硬件的迭代周期一般以年为单位，相比大模型发展来说可以说是龟速；其中电池技术还存在明显的能力上限，更是十年为单位的研发周期，最终或将需要继续优化芯片效率与能耗的方式来实现曲线救国来满足需求。

未来AI Agent优势或不在端侧体现。随着当前云计算的发展，用户的大量非隐私数据，有着逐步上云的趋势，比如大量的相册照片、非敏感数据，事实上在云端而不是在端侧，苹果iCloud、google drive管理着数十亿独立用户数据库；随着AI infra的进一步发展，未来高性能的算力、数据更多在云端运行与流转，完全本地化的「端侧模型」也许并不是一个高效的算力中枢，而更多是一个满足用户隐私保护、敏感数据安全需求的交互入口模型。更多的需求，需要更专业的AI Agent主动式工作流程，搭配云端最先进的模型来实现，且不断突破能力上限。

不是所有端侧都需要本地模型。当前有四类主要的终端设备：1）PC、手机终端；2）AR与VR设备；3）智能车；4）大量的收数设备，小微终端（包括各类摄像头、耳机、无人机，甚至是前期火热的AI PIN均属于此类设备），事实上目前最火热的不管是AI PIn和Rabbit R1这类小微设备，都不会是端侧大模型的第一代产品形态，这一系列收数设备，主要还是以API调用的方式来运行，而不需要端侧模型；本地化部署的优先级，除了考虑厂商需求，更需要考虑用户需求，智能车、AR与VR、PC、手机不同终端类型的「端侧模型」占比将出现分化。

领先于原生智能终端，终端交互体验提前变革

短期来看，虽然整体的能力输出，需要硬件生态的发展，与云上大模型配合，才能实现通用Agent、复杂COT/TOT、超长文本、多模太等高级功能，但这并不妨碍端侧大模型，成为各大基础模型厂商有效的进行价值兑现重要入口的这一趋势，变化已经在提速。

SaaS将全面推进AI化。距离真正的AI智能手机、PC的出现，或许还需要1-3年的时间。但当前各类SaaS服务的“含AI量”已经大幅提升。不管是满足既有需求的AI+应用，如产品+AI的应用，如bing、腾讯会议、Convas；以及AI原生应用，如妙鸭相机、perplexity.ai；还是创造新需求的AI应用，如Character.ai、Convai.ai等。现有应用积极+AI，AI原生应用加速涌现，AI化，大模型的终端使用赋能已经广泛的被应用于搜索、会议、文档、陪伴、游戏、个人助理、学习等诸多领域，随着SaaS的全面AI化，终端的AI交互体验正在提前兑现。

APP将面临AI原生OS不同的“去皮”风险，交互方式回归「本源」，直接调用API。从用户交互的角度，GUI图形用户界面将变身对话式Conversational UI，大家不需要再依赖屏幕GUI图形用户界面，而是可以由OS直接拉起服务，转向更加通用的用户界面。回顾过去，从最开始PC的DOS命令行，到Windows视窗，再到iOS的触控，微信的语音，Vison Pro的眼动、肢体动作等等，每一次新交互方式的诞生，几乎都带来新一个十年的应用变革。随着GPT-4o的发布，让大家看到了下一代交互方式的可能性：语音+多模态视觉（融合身体的各个输入器官，语言与肢体动作，环境等等，都可以用于交互），正预示着下一个十年的交互方式，即将到来，人类彻底的解放外部交互，将身体的任何一个器官用于交互（待攻克的还有脑机接口）。

应用入口变化可能性	低 内容生产	中 资源聚合	高 同质转售
高 交互目标在GUI之外	●长视频平台 ●地图	●O2O交易平台 ●电商平台	●分发平台[应用分发等] ●工具:理财、天气等APP
中 交互目标是物理实体在GUI的投射	●搜索引擎/浏览器【形态AI化】	●社交平台 ●内容社区	●资讯类APP【新闻等】 ●影视内容的分发平台
低 GUI即交互目的	●游戏	●流量APP【短视频等】	——

从后台运行的角度，APP本身也需要做出颠覆式的改变，大量的服务，不需要通过一个个的APP，而是直接调用API的方式。同样的，服务将延续一直以来的“去皮化”的过程，从浏览器/网页（“APP”的聚合）到APP（API的聚合），再到直接的 API提供服务，OS直连API，“没有中间商赚差价”。服务将回归本源，遵循第一性原理，要么提升两个实体目标的连接效率，要么辅助拓宽虚拟世界的边界。比如打车、即时社交（非线上延时社交，聊天工具，邮件等）等，可以实现快速的连接撮合；另一方面，如游戏、短视频等，将提供更多的交互性，如无限延展的元宇宙世界与情景剧情节变化等等。

总的来说，未来APP入口时长价值，取决于APP本身的资源与服务价值，纯聚合的流量或将进一步下沉到OS层面。标准化交易平台的用户时长价值将快速贬损，比如高铁票、机票等；其次是个性化交易平台，如电商、外卖、酒店等；最后是本身就是数字消费的交易平台，比如聊天平台、长短视频、在线游戏等等；下一个十年，APP入口需要摒弃原来中间商的思路，借助端云大模型能力的加持，提升用户的体验价值与时长价值。

趋势8

具身智能：人型机器人与大模型共同进化，为外脑提供“躯体”

作者：陈玉珑 李永露 张志忠

人型机器人作为人工智能的终极载体之一，凭借其类人形态和全身自由度，能够适应未经特别改造的人类环境，从而在各种生产和生活场景中发挥作用。人形机器人的发展依靠两大技术支柱：运动控制与任务训练。机器人本体运动控制即从传统的液压系统向更高效、更精确的电机驱动系统的转变，使得机器人的物理动作更加细腻和人性化；其次，大模型的应用，即结合先进的机器学习技术，尤其是在任务训练方面，大模型的利用极大提高了机器人的学习效率和执行复杂任务的能力。这两种技术的融合不仅推动了人形机器人的技术革新，也为其在实际应用中的广泛部署打开了可能。

这种技术的深入发展预示着人机共生的未来，其中人形机器人将在各行各业中发挥越来越重要的作用，从家庭服务到高风险的工业操作，都能见到它们高效、安全的身影。通过持续的技术革新和应用拓展，未来人形机器人有望在提高生活质量和工作效率方面发挥关键影响，进一步融入人类的日常生活中，成为不可或缺的助手。

人形机器人的发展历程中，运动控制和任务训练技术的突破已经开始重塑全球的工业、服务和社交交互领域。这些技术的提升不仅增强了机器人的功能性，也使它们能够更自然、更高效地与人类互动。人形机器人，以其类人形态和全身自由度而被设计，旨在无需对人类环境进行特别适配即可在其中自然地移动和操作。随着技术的进步，尤其是90年代计算能力和传感器技术的提升，人形机器人的行为能力得到显著增强，代表作如Honda的ASIMO在平滑行走和上下楼梯等方面取得了突破。进入2000年代后，更高级的认知功能和更精细的动作控制使得人形机器人开始广泛应用于服务业、娱乐和研究领域。最近几年，深度学习和大模型技术的应用使得人形机器人如Boston Dynamics的Atlas和Tesla的Optimus展示了前所未有的灵活性和适应能力，预示着人形机器人技术在未来社会中的广泛应用和深远影响。

运动控制关键技术进步促进机器人“大脑”运行

电机技术革新助力人形机器人实现高效精确的运动控制。人形机器人的组成复杂且精密，涵盖了模拟人类的腿、腰、手等硬件结构，这些结构使得机器人不仅能行走和搬运，还能执行如抓取等精细动作。特斯拉的Optimus和Boston Dynamics的Atlas代表了运动控制技术的最新进展。特斯拉的创新电机技术推动了机器人运动控制的电动化，通过将其在电动汽车领域积累的电机控制技术转化应用到机器人技术中，Optimus在行走、抓取及其他复杂动作的执行上展示了极高的精确度和流畅性。这种电机技术的应用不仅提升了机器人的操作效率，还降低了能耗，使得机器人更加适用于长时间的工作需求。

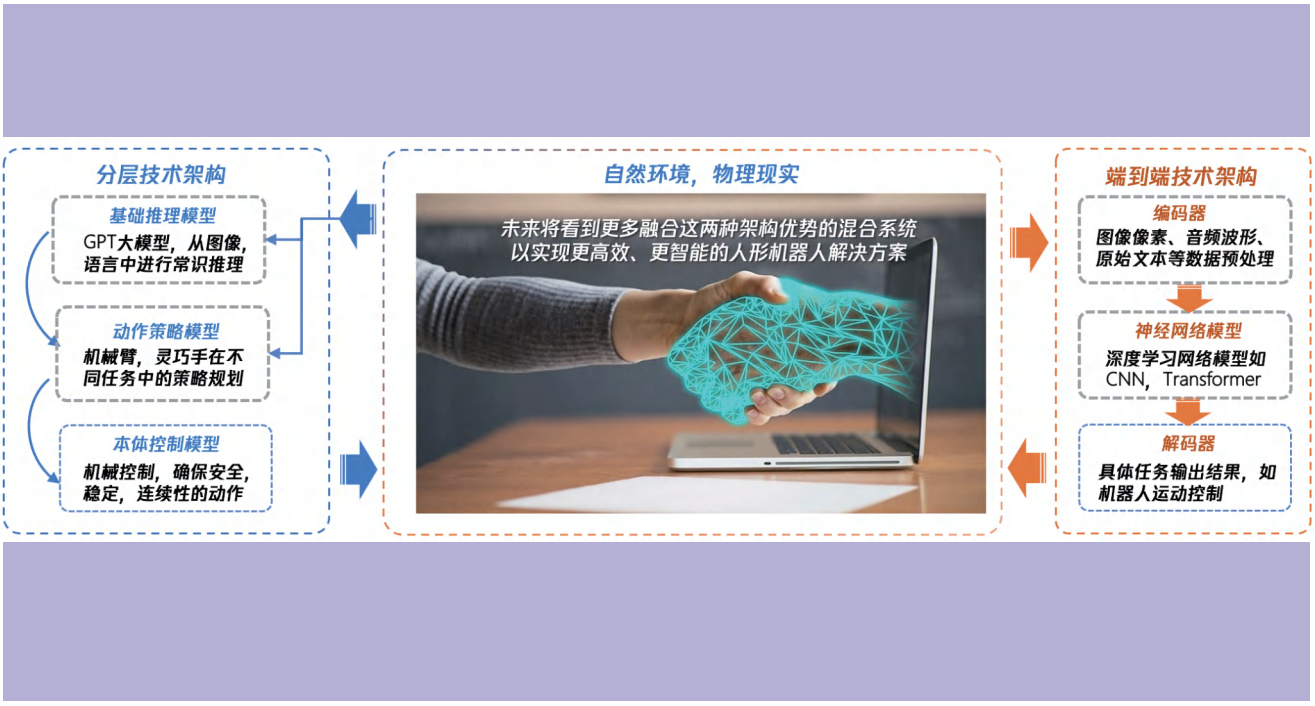
相比之下，Boston Dynamics的Atlas最初依靠液压技术执行复杂的运动控制任务，这一技术使得Atlas在执行高强度动作时具有出色的力量输出和稳定性。然而，液压系统的重量和复杂度使得机器人的敏捷性和应用范围受到限制。近年来，Boston Dynamics开始将Atlas的动力系统从液压转向更轻便、能效更高的电机驱动，这一转变显著提高了Atlas的动作灵活性和适应各种环境的能力。这些技术的革新不仅提高了机器人的物理性能，也为其在更广泛的实际应用场景中的部署奠定了基础。

运动控制关键技术进步促进机器人“大脑”运行，人形机器人将在广泛应用场景中发挥更大作用。优化的运动控制不仅提升了机器人的基础动作执行能力，更是智能化发展的基石。一旦运动控制达到一定的成熟度，机器人的“大脑”——即处理复杂认知和决策任务的AI系统的整合，就可以更加顺畅地进行。实际上，高效的运动控制为机器人提供了更多实验和学习新任务的可能性，这在救援、医疗和家庭服务等领域尤为重要。在未来，随着运动控制技术的持续进步和AI技术的深度融合，我们可以预见人形机器人将在更广泛的应用场景中，如自动化生产、灾难响应和个人助理等领域发挥更大的作用。通过进一步的技术创新和应用探索，人形机器人的发展将继续推动生产力的变革和社会生活的改进。

任务训练与大模型的结合

任务训练成为了人形机器人领域技术进步的另一个关键领域。在这一方面，OpenAI、Figure以及Nvidia等公司的创新尝试展示了如何通过大模型技术来提高机器人的学习效率和执行任务的能力。这些大模型通过处理庞大的数据集，使机器人在学习执行特定任务时更为高效。

系统架构的选择是决定人形机器人与大模型结合的开发和部署中的关键因素。目前，分层架构和端到端架构是两种主流的设计方法，各有其优势和应用场景。



分层架构，如OpenAI和Figure所采用的，通过将感知、决策和行动分成处在不同层次的模块来处理，增强了系统的稳定性和可维护性。这种结构允许各层之间进行信息流和命令流交互，便于单独优化和调试。

端到端架构（Any2Any）提倡一个统一的大模型直接从输入到输出学习任务，强调通过单一的神经网络处理从感知到动作的全部过程，以期达到更高的操作灵活性和效率。这种架构尝试通过直接学习输入到输出的映射来简化系统的训练和部署，适用于处理复杂的、动态的任务环境。英伟达的DrErureka项目和谷歌DeepMind的AIagent项目展示了端到端架构的强大能力。

选择分层架构还是端到端架构，取决于特定的应用需求、系统的复杂性及开发团队对错误容忍度的要求。随着机器学习和人工智能技术的进步，我们预计未来将看到更多融合这两种架构优势的混合系统，以实现更高效、更智能的人形机器人解决方案。

人形机器人与大模型机器学习的融合将引领技术革命，加速产业广泛落地

在大模型的帮助下，未来人形机器人将会应用在各个领域。人形机器人技术的核心，如高级运动控制和复杂任务执行能力，通过与大模型的结合，可以实现更快的学习速度和更高的操作精确性。大模型在自然语言处理领域的成功应用已经证明了其在处理复杂数据和学习复杂模式上的强大能力。将这些模型应用于人形机器人，可以极大地提高机器人在现实世界中处理未知情况和自主决策的能力。随着技术的成熟，人形机器人预计将在医疗、家庭服务、公共安全、教育以及娱乐等多个领域找到其应用。例如，在医疗领域，经过大模型训练的人形机器人可以执行精细的手术操作；在家庭服务领域，它们可以进行家务管理和提供老年护理；在公共安全领域，可以执行危险的搜救任务。这些应用的实现，将极大依赖于人形机器人与大模型的有效融合。人形机器人的部署将对劳动市场产生深远影响，其可以替代或协助人类工作，提高生产效率，同时减少工作场所的事故。这将导致劳动力结构的变化，要求重新考虑教育和培训系统，以适应新的技术需求。同时，它也提出了伦理和法律问题，如隐私保护和机器人权利的界定，这需要政策制定者、行业领导者和社会各界共同努力解决。

随着人形机器人技术的重要性日益增加，全球各国和地区都在加大投入，希望在这一领域占据领先地位。这种趋势不仅是技术竞赛，也反映了各国对未来社会形态——人机共生的预见和准备。同时，这也推动了国际间在标准制定、伦理考量和技术分享方面的合作，为人形机器人的健康发展和广泛应用奠定了基础。

预计随着大模型技术的不断进化和优化，未来人形机器人将在智能化和自主性上实现更大的飞跃。他们将能够更加自如地在复杂环境中工作，执行更多需要高级认知和决策能力的任务。这将极大地扩展人形机器人的应用范围，从而更深入地融入人类的日常生活和工作中，成为社会发展的重要推动力。总之，人形机器人技术的快速发展和大模型的集成正开启一个新时代，这不仅将改变机器人行业的面貌，也将影响我们的工作和生活方式，带来前所未有的变革。

趋势9

开源共享：开源生态实现降本普惠，推进外脑共享和迭代

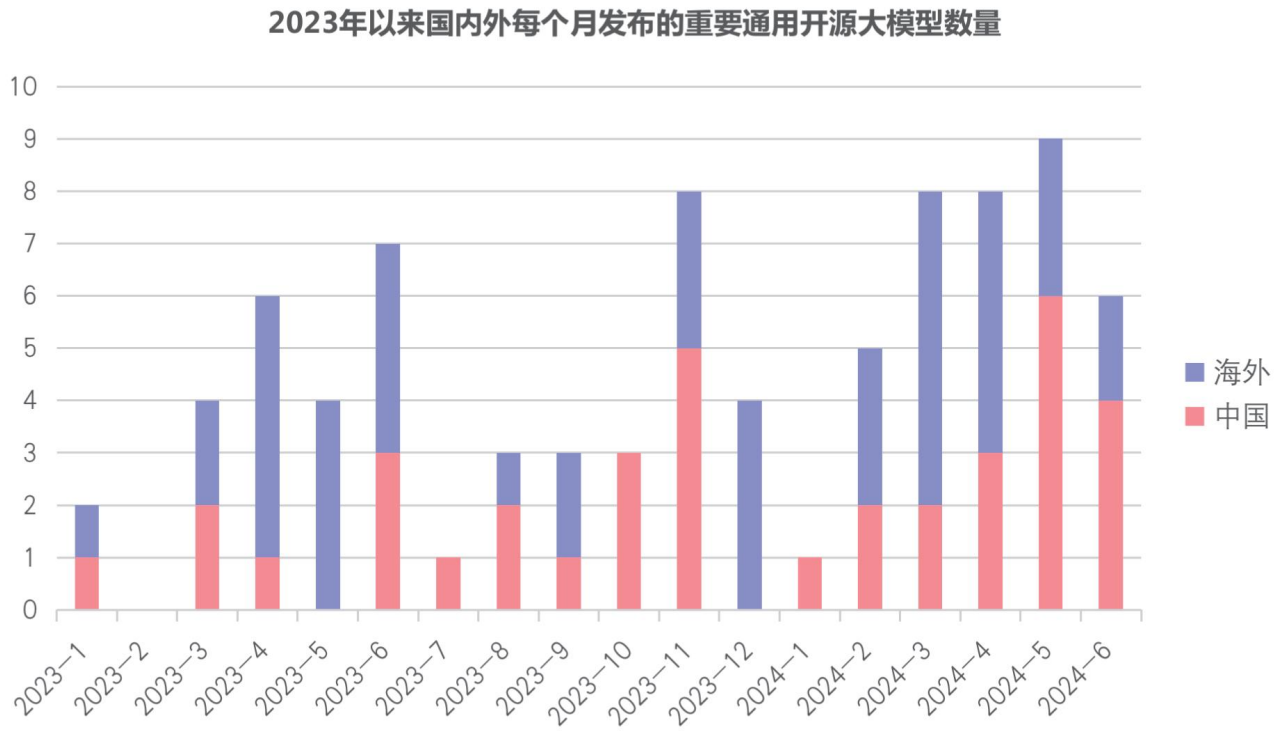
作者：袁晓辉

基于对国内外100多个开源大模型的分析，我们预计未来2-3年内，AI开源生态将迎来繁荣发展，随着开源大模型数据质量与多样性提升，大模型将实现规模缩减和质量提升，推进开源大模型从“可用”到“好用”的演变。开源社区的全球协作特性促使开发者共享资源，推动全球知识分享与技术协作。开源模型特别适合需求个性化、数据敏感的垂直行业应用，也为中小企业提供低成本、高效率的AI解决方案，助力更多商业场景的创新发展。此外，AI开源也为安全治理和人才培养提供了更好地条件。

AI开源生态更加繁荣，持续推动大模型质量持续提升

在未来2-3年内，AI开源生态预计将进入一个更加繁荣的阶段。随着开源大模型训练数据的质量和多样性的提升，我们不仅会看到开源大模型质量的持续提高和模型规模的有效缩减，还将看到具备差异化特色的大模型更多出现。目前开源大模型生态包括了大语言模型、多模态大模型、具身智能大模型，以及部署和应用工具等。开源生态的繁荣一方面推进开源大模型从“可用”到“好用”的演变，另一方面推进模型使用的软硬件成本的降低，通过开源生态的力量惠及更多元的使用群体和行业。

我们收集了2017年以来国内外100多个开源大模型，包括大语言模型、多模态大模型和具身智能大模型等。数据显示，从2023年1月到2024年6月，国内和国外的开源大模型发布数量都在增加。特别是在2024年初，国内发布的开源大模型数量显著上升。例如，2024年5月国内发布了6个开源大模型，说明国内的开源生态进入了一个更加繁荣的阶段。



图：2023年以来国内外每个月发布的重要通用开源大模型数量

数据来源：腾讯研究院收集整理

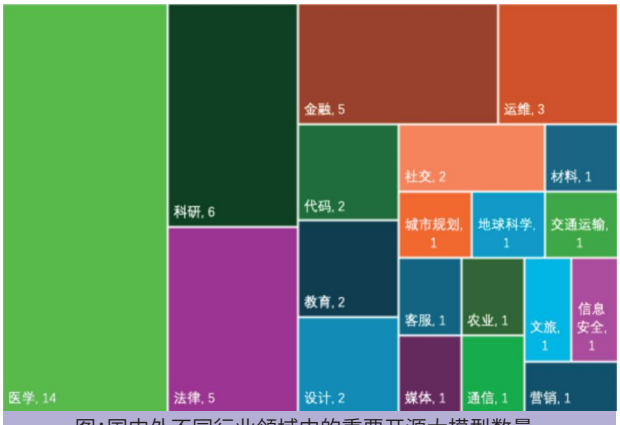
同时，这些数据印证了开源大模型质量和多样性的提升趋势。从模型参数规模来看，国内外开源大模型的参数规模正在逐渐增加，这表明模型的复杂度和能力在不断提升。例如，英伟达于2014年6月开源的Nemotron-4　340B参数规模达到3400亿参数，而国内AI企业深度求索于2014年5月开源的DeepSeek-V2模型参数规模也达到了2360亿参数，每个token激活的参数量为210亿。而不同开源模型的迭代也在持续降低训练成本，减少KV缓存，并提高最大生成吞吐量。与此同时，开源模型能处理的上下文长度也在不断提升，目前国内开源模型中，上下文长度最大的开源模型之一是零一万物的Yi-1.5-34B模型，可以提供200k上下文长度。这些数据表明开源生态不仅在数量上增长，在提供的模型质量和多样性上也有显著提升。

AI开源推进全球范围的开放创新

开源大模型推动社区驱动的开放创新，其全球协作特性将促进不同国家和地区开发者的共同工作，分享知识和经验。例如根据不完全统计，国内有超过50家机构贡献了开源大模型，包括大型互联网企业和硬件企业，如腾讯、阿里巴巴、华为、百度、字节、VIVO、昆仑万维，也包括大模型初创企业，如智谱华章、百川智能、零一万物、深度求索、元象、原始智能等，以及高校和科研院所高校，如智源研究院、上海人工智能实验室、北京大学、清华大学、香港中文大学、中山大学、香港科技大学等。而贡献开源大模型的海外机构数量超过100家，也同样包括了企业、高校和各类科研机构。这些开放的大模型，以及训练和微调方法，在推进产业界演进和迭代的同时，将极大助力学术和科研机构基于模型架构，探索更优的模型方案，进一步推进人工智能相关的理论研究。同时，大模型开源也将鼓励不同学科背景的机构和人才站到同一起跑线，广泛参与到大模型演进的生态中，更有助于激发集体智慧，推动人类的“机器外脑”高效迭代和进化。

AI开源助力更多商业场景发展

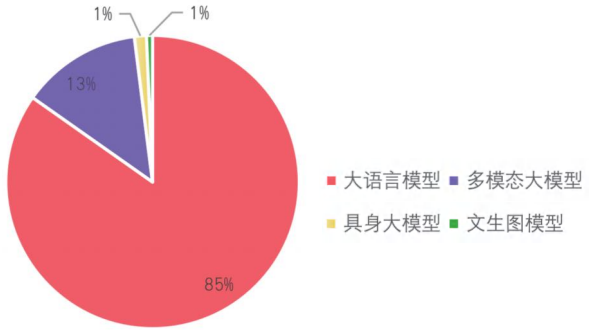
大模型厂商也将面临大模型开源和闭源的选择。在闭源模型性能无法与其他厂商显著拉开差异的背景下，一些厂商也将选择开源路线，来推进大模型性能的加速迭代和用户生态的尽快构建。特别是，在开源大模型跨越“好用”的门槛后，其灵活性将支持更多定制化的软硬件解决方案，有效利用客户的已有资源。比如在一些数据隐私和安全性要求高、行业知识更为集中的行业，如医疗、科研、法律、金融等垂直行业，或者是已经采购了模型训练硬件资源，希望降低后续使用模型的推理成本的企业，都更有可能选择使用开源模型。



图：国内外不同行业领域中的重要开源大模型数量

数据来源：腾讯研究院收集整理

此外，开源模型也将为中小企业提供前所未有的创新机遇，尤其是在与硬件结合的端侧应用上。这些企业可以利用开源模型以较低的成本快速部署AI解决方案，实现在本地硬件上的智能处理和即时响应。特别是，我们也看到越来越多的多模态大模型和具身智能大模型选择开源，比如腾讯于2024年5月开源的HunyuanDiT模型，是业内首个中文原生的DiT架构文生图开源模型，再比如智源研究院于2024年5月开源了具身智能大模型ASGrasp，在泛化抓取方面，实验成功率突破95%。质量更高且更丰富的开源大模型也会进一步加速大模型在更多商业场景的落地。



图：国内外重要开源大模型类型占比

数据来源：腾讯研究院收集整理

AI开源将促进安全治理和人才培养

最后，开源大模型及相关工具因其开放性，也将更有利于安全治理。热别是开源项目促进了不同学科领域的专家之间的合作，有助于从多角度审视和解决伦理和可解释性问题。例如，技术专家、伦理学家和法律专家可以共同工作，确保AI系统的决策过程不仅符合技术标准，也遵循伦理和法律规范。在推进AI技术在更加精准、高效的同时，更具备安全性和隐私保护能力。

此外，开源大模型也将在推动形成行业标准和人才培养方面发挥更大价值。目前很多学校都将开源大模型作为教育工具，帮助学生和新入门者理解复杂的AI概念和技术，通过实际操作和修改开源代码，帮助学生深入了解较为前沿的大模型工作原理，对培养下一代AI专家和解决复杂的安全问题至关重要。

趋势10

人机对齐：AI对齐是大模型产品的重要竞争力，也关乎通用人工智能的未来

作者：曹建峰

在大模型时代，随着AI模型具有越来越多的类人能力、越来越像人，不再被视为纯粹的被动工具，如何让AI模型的能力和行为和人类的价值、目标、伦理道德、意图等追求相一致，这个被称为AI对齐的问题变得越来越重要。人机对齐由此成为了AI发展的重要理念和技术实践。在实践层面，人机对齐是大模型产品成功的关键，也是实现通用人工智能（AGI）的前提。通过人机对齐我们可以构建更加实用、真诚、安全、无害的AI系统，确保智能向善。

大模型发展引发深度担忧，AI对齐成为关键议题

2023年以来，随着AI大模型的加速发展，相关的AI安全风险和控制问题引发全球关注。具体而言，以大模型为代表的新兴AI模型，不断推动AI技术迈向新的前沿，拉近人们与通用人工智能（AGI）之间的距离。但这也在一定程度上引发了一些业内人士对人工智能未来极端风险的担忧。除了数据保护、个人隐私、算法歧视和算法黑箱、虚假信息、模型网络安全等已有伦理问题，未来的AI大模型（即所谓的“前沿AI”）是否可能导致灾难性风险或极端风险的问题也得到了更多的关注。

换言之，不同于以往的任何技术，前沿AI技术主要会在三个核心的维度上，给个人和社会带来全新的风险挑战。其一，决策让渡中的风险，在经济社会活动的维度，AI和机器人会在越来越多的人类事务中替代人类进行决策和行动，这种决策让渡会带来新的风险，诸如AI幻觉、算法歧视、价值对齐、技术性失业、AI安全等。其二，情感替代中的风险，在人际/人机关的维度，AI和机器人已经并将持续深入介入人类情感领域，给人们提供情感陪伴价值，但却可能影响到人际交往，产生情感替代中的风险，导致人与人之间的真实联系被削弱甚至被取代。这种新型人机关系的伦理边界应该怎么确定？首要的原则是，人类AI交互必须促进人类联系和社会团结。真实的人类联系在AI时代将是弥足珍贵的。其三，人类增强中的风险，在人自身发展的维度上，AI和脑机接口等新技术，可能推动人类社会进入“后人类时代”，AI和脑机接口被用于增强、改造人类自身，未来人机深度融合后，人的身体、大脑、智力等都可能可以被AI技术改造，届时人会变成什么？这种人类增强会不会带来新形式的人类不平等？在这几个维度之外，还有技术滥用、恶用的风险（如deepfake），AI消耗大量能源对环境和可持续发展的挑战，科学界担忧的AI技术失控、威胁人类生存的灾难性风险，AI加速派与AI对齐派的发展理念分歧，等等。因此，负责任AI的理念和实践变得越来越重要。

在这种背景下，随着大模型能力的持续提升，以及日益变得更加通用化，如何让大模型的能力和行为与人类的价值、目标、意图、伦理道德等相一致，成为了大模型发展中的关键议题。严谨地讲，AI对齐（AI alignment）是AI安全和伦理领域的一个概念，其主要目的是将AI大模型打造成安全、真诚、有用、无害的智能助手，避免在与用户交互过程中出现潜在的负面影响或危害，例如输出有害内容、产生幻觉等。在大模型时代，AI对齐对于确保人类与人工智能交互过程中的安全与信任至关重要。现在的聊天机器人等大模型应用之所以能够游刃有余地应对用户的各种提问，而不至于产生太大负面影响，在很大程度上归功于AI对齐方面的实践。因此，AI对齐是现在的大模型的可用性和安全性的重要基础。

人类反馈方法和原则型AI方法推动AI对齐有效落地

在实践中，目前业界将AI对齐作为对AI大模型进行安全治理的重要思路，并在技术上取得了可观的效果，在很大程度上确保大模型部署和使用中的安全与信任。**AI对齐作为大模型研发过程中的一个重要环节，目前主要有两种AI对齐的方法。一种是自下而上的思路，也就是人类反馈的强化学习，需要用价值对齐的数据集对模型进行精调，**并由人类训练员对模型的输出进行评分，再通过强化学习的方式让模型学习人类的价值和偏好。在技术上，人类反馈强化学习（RLHF）包括初始模型训练、收集人类反馈、强化学习、迭代过程等步骤。**另一种是自上而下的思路，核心是把一套伦理原则输入给模型，并通过技术方法让模型对自己的输出进行评分，以使其输出符合这些原则。**

例如，OpenAI采取了人类反馈强化学习（RLHF）的对齐方法，Anthropic采取了AI反馈强化学习（RLAIF）的对齐方法即所谓的“原则型AI”（ConstitutionalAI），这些AI对齐方法殊途同归，都致力于将大模型打造成为安全、真诚、有用、无害的智能助手。以RLHF为例，RLHF在改进模型性能、提高模型的适应性、减少模型的偏见、增强模型的安全性等方面具有显著优势，包括减少模型在未来生产有害内容的可能性。OpenAI将RLHF算法发扬光大，ChatGPT籍此取得成功，能够在很大程度上输出有用的、可信的、无害的内容。除此之外，产业界还在探索对抗测试（红队测试）、模型评估、可解释AI方法、伦理审查、第三方服务等多元化的安全和治理措施，共同确保负责任AI的发展。

值得一提的是，对于可能具有灾难性风险的前沿AI模型和将来可能出现的超级AI，美国一些AI企业在探索专门的安全机制（例如Anthropic的负责任扩展政策、OpenAI的“预备”团队），其核心思路是对新研发的更先进模型进行系统性评估，只有在模型的风险低于一定的安全风险阈值时才会对外推出，否则将暂缓推出直至安全风险得到缓解。例如，Anthropic最新推出的Claude 3模型的风险级别被评估为ASL-2（按照Anthropic自己确定的AI安全级别），尚不具有显著的高度风险或极端风险，因此可直接向公众提供。

可见，OpenAI、Anthropic等美国主流AI企业通过在AI对齐上的相关探索和举措，建立起了其AI产品的市场竞争力，同时这些企业将AI对齐作为保障未来更强大的AI模型安全、负责任发展的核心要素，积极开展前沿探索。

AI对齐是大模型的必由之路，也关乎未来AGI的安全发展

可以说，价值对齐是当前AI大模型和未来通用人工智能（AGI）的必由之路，可以帮助解决AI大模型商业应用扩散过程中面临的难题，让大模型成为更加有用且安全可信的AI助手。例如，GPT-4在RLHF训练阶段，通过增加额外的安全奖励信号（safety reward signal）来减少有害的输出，这一方法产生了很好的效果，显著提升了诱出恶意行为和有害内容的难度。GPT-4相比之前的模型（如GPT-3.5）显著减少了幻觉、有害偏见和违法有害内容等问题。经过RLHF训练之后，GPT-4在相关真实性测试中得分比GPT-3.5高40%，响应禁止性内容请求的可能性比GPT-3.5降低了82%，并且能够更好地回应涉及敏感内容的用户请求。总之，RLHF算法可以为大语言模型建立必要的安全护栏，在大模型的强大性/涌现性和安全性/可靠性之间扮演着“平衡器”这一关键角色。

腾讯混元大模型等AI产品应用也积极重视大模型安全与伦理，把安全作为大模型产品的核心竞争力。腾讯研究院等团队今年初发布的《大模型安全与伦理研究报告2024——以负责任AI引领大模型创新》，系统性分享了大模型安全框架和相关安全措施，确保始终把安全作为大模型产品的首要事项。具体而言，腾讯混元大模型在prompt安全测评、大模型蓝军攻防演练、大模型源代码安全防护、大模型基础设施漏洞安全防护方面已经采取了全方位的实践做法。同时，对大模型系统设计、开发、测试、部署、运行、退役等环节涉及到的安全保护要求积极改进优化，并参与大模型安全的一系列国标，行标的标准制定。通过这一系列做法，腾讯混元大模型已经建立了比较完善的安全能力矩阵。

AI对齐在解决大模型的安全和信任问题上扮演着重要角色，能够实现安全与创新的有效平衡，需鼓励、支持大模型价值对齐的技术和管理措施，推动形成相关的政策指南、行业标准、技术规范等。总之，AI大模型未来会在更多场景中辅助人类甚至替代人类做出各种决策和行动，因此AI对齐是大模型的必由之路，这既关乎信任，也关乎控制。因为AI对齐不仅是将现在的大模型打造成更加安全、真诚、有用、无害的智能助手的核心举措，也关乎未来的AGI的安全，对于控制未来更加强大的AI模型的涌现风险至关重要。

创新者预见

青腾 TENCENTX

金沙江创投主管合伙人 **朱啸虎**

今年比较普遍的特点是都在积极拥抱AI，看怎么用AI赋能自己的产业，怎么用AI来降本增效，这是一个非常明显的趋势。我觉得不需要过度担忧被落下或追求高大上的概念，就在一个比较小的场景能尽快落地，尽快看到效果，这对很多企业来说更重要。尽快的商业化落地，先看到效果，先看到财务上的效果。然后再逐步跟着大模型、跟着整个技术的进步，扩张自己的场景，也会更有效一点。在具体的商业化场景上，真正深刻理解人，这是中国最擅长，也是中国最有价值的创业方向。

源码资本管理合伙人 **黄云刚**

一代人有一代人的机会，要相信热爱技术的年轻人。在生成式AI时代，技术、产品和基础设施正在同步快速发展，首先对快速变化的技术和趋势有很好的预判，再加上产品嗅觉和客户思维，才有机会做出杀手级的应用，而独特的产品才是胜出的关键。

真格基金管理合伙人 **戴雨森**

AI的基建周期会比预计更长，投入更大；于此同时，AI应用的短期落地很可能会让乐观的人失望，我们将再一次经历「短期乐观、长期悲观」的周期。但以史为鉴，长期来看我毫不怀疑 AI 会创造大量价值，并且创业公司会在其中扮演重要作用，尤其是全新的产品和商业模式，大概率在 AI 渗透率到达一定程度之后才会诞生。因此如何加速 AI 普惠大众，提高AI 产品的市场规模是我现在最关心的。

晶泰科技董事长 **温书豪** （青腾未来科技学堂校友）

我相信人工智能就像互联网这个行业，一定会沉淀出一些最重要的应用场景，特别是生物医药和材料这个领域，会产生巨大的商业价值和社会价值。生命科学有可能进入到一个新阶段，以前是科学驱动，有很多不确定性，现在因为数字化的基础很好了，逐渐可以转变成工程学驱动。材料领域也将进入AI时代。我们可以预见AI驱动发现的可控核聚变，室温超导的材料有望解决能源的终极问题。

美图公司、董事长兼CEO **吴欣鸿** 青腾未来产业学堂校友）

随着大模型的发展，AI应用将经历“点线面”三阶段。AI单点功能正逐步被串联成AI工作流。就像搭积木一样，AI会根据需要调用不同的功能，快速组成工作流，从而完成特定的任务。大模型也将基于工作流的数据反馈自动迭代。相信在不远的将来，AI工作流会进化成AI平台生态，便利每个人的工作和生活。

深言科技创始人 **岂凡超** （青腾未来产业学堂校友）

大模型将在两方面继续发展，一方面以OpenAI为代表的前沿企业和研究机构继续拓展模型规模，探寻Scaling Law的科学边界，另一方面更多的企业会面向特定任务或场景打造垂直模型，在效果、性能和成本实现更好的平衡，加快大模型的产品化和落地。前者从供给出发，后者从需求出发，更加贴合用户和场景，将爆发出更大、更多样化的机会。

达观数据创始人、董事长兼CEO **陈运文** （青腾未来科技学堂校友）

大模型技术会和垂直领域的知识深度融合，带来知识管理领域的革命性变革。未来，企业将通过智能化知识库实现精准信息检索，利用多模态数据处理丰富知识资源。随着大模型的持续学习和自我优化，知识库将不断进化，提供个性化决策支持，提升客户服务体验，并激发企业创新潜力。同时，知识共享与团队协作将因智能化系统而更加高效，推动跨领域知识交流。安全合规性将成为设计知识管理系统时的关键考量，确保数据保护与合理访问。大模型的专业化和领域定制化将进一步推动企业知识管理向更深层次的智能化和效率发展。垂直大模型和特定领域的专业知识的结合是推动AI智能化落地的关键力量，大模型和知识库的不断交融将形成对行业规律的深刻理解，能更加智能的处理专业任务，为各行各业带来革命性的变革。