编者按 人工智能是新一轮科技革命和产业变革的重要驱动力量。能否抓住人工智能发展的重大历史机遇，关乎产业未来和国家前途命运。今起，本版推出“AI风向标”系列报道，关注人工智能前沿动态，探讨AI发展趋势，建言产业健康发展。

未来大模型：更多能 更轻量 更亲民

AI风向标①

◎本报记者 **崔 爽**

多模态将成为大模型标配，未来，大模型不仅能听会说，还能看懂图片和视频，更能识别理解。

去年起，全球掀起以大模型为代表的生成式人工智能行业浪潮，人工智能迈向全面应用新阶段。国外以OpenAI、微软为代表，谷歌、Meta等巨头一个不落，国内头部科技企业则悉数入场。

近日，国家数据局局长刘烈宏表示，中国10亿参数规模以上的大模型数量已超100个，行业大模型深度赋能电子信息、医疗、交通等领域，形成上百种应用模式，赋能千行百业。

大模型技术日新月异，产业化脚步追星赶月。在这个热闹夹杂争议的行业，不同的技术路线和选择正在延展开来。

多模态正成标配

360集团董事长周鸿祎在关于2024大模型发展的十大趋势判断中明确表示，多模态将成为大模型标配。未来，大模型不仅能听会说，还能看懂图片和视频，更能识别理解。

中国科学院自动化研究所研究员刘静在《GPT-4对多模态大模型在多模态理解、生成、交互上的启发》一文中谈到，模态是指表达或感知事物的方式，例如人类的听觉、视觉、触觉等。在人工智能领域，多模态是指利用多种模态的信息来进行感知和理解。多模态技术可以让机器同时接收并处理不同模态信息，如文本、图像、音频等，从而提高机器感知和理解能力。

刘静进一步解释，相比传统的单模态大模型，多模态大模型更加符合人类的多渠道认知方式。它能将不同模态信息相互补充，提高信息的完整性和准确性，从而更好地应对复杂环境、场景和任务。如在语音识别中，多模态大模型可以结合语音和文本信息，让机器更准确地识别和理解语音内容。在图像识别中，图像和文本信息的结合可以让机器更深入地理解图像内容。

刘静介绍，落实到应用层面，多模态能使机器更好地理解人类的意图和需求，提供精准、个性化服务。例如，在目前智能技术已经深入落地的客服领域，多模态大模型可以根据用户的文本和语音信息，综合理解用户问题并提供解决方案。在智能家居领域，多模态大模型通过与智能家居设备的交互，可以根据用户需求智能调整家居环境。在医疗领域，多模态大模型则可以融合医学图像、病历文本等信息，辅助医生进行疾病诊断、制定治疗方案。

结合国内大模型产业布局，多模态大模型是近年主流厂商普遍选择的发力方向。如“紫东太初”2.0全模态大模型可实现文本、图片、语音、视频、3D点云、传感信号等不同模态的统一表征和学习。

落地应用越来越“轻”

为提升大模型性能、增进用户体验，大模型正以肉眼可见的速度越来越“大”。开源模型的参数从几百亿增加到几千亿，闭源模型也在沿着尺度定律路径不断升级，在千亿参数的基础上向着万亿参数攀登。但与此同时，人工智能产业算力吃紧，难以满足大模型参数规模的指数级增长。

为加速大模型落地应用，人工智能领域正尝试对大模型进行轻量化，通过打造更小型、更高效、成本更低的模型吸引行业客户，让模型在更广泛的场景有更高应用价值。

大模型轻量化通过降低模型的计算复杂度和内存占用，实现模型性能与资源消耗的平衡。这不仅有助于提高计算效率，降低运行时资源消耗，还可以减轻计算系统的硬件和软件压力，提高系统的可靠性和稳定性。此外，轻量化还有助于提高模型部署的灵活性和可扩展性，为模型在各种场景下的应用提供更多可能性。

轻量化主要通过模型压缩来实现。模型压缩是指在不降低模型性能的前提下，减小模型的计算复杂度和内存占用。模型压缩的方法有很多，剪枝、低秩分解等是业界常用的方法。具体来看，剪枝是指删除一些不必要或冗余的参数，低秩分解是指将高维数据或高维模型分解为低维数据或低维模型。这些方法都是通过降低模型的复杂度，减少参数和计算量，来达到让模型“更轻”的目的。

轻量化不仅有助于提高计算效率，降低运行时资源消耗，还可以减轻计算系统的硬件和软件压力，提高系统的可靠性和稳定性。

对于通用大模型服务垂直领域、行业大模型落地应用来说，轻量化更是关键步骤——通过合理的数据准备、模型选择、超参数设置和训练验证过程，可以使模型更好地适应特定领域，提高模型性能。

联想集团副总裁、联想研究院技术战略与创新平台总经理王茜莺曾表示，用好大模型主要有三个要素，第一是场景化，第二是隐私化，第三是轻量化。轻量化不仅是将个人大模型进行压缩并部署到用户设备上，还包括企业内部的轻量化，甚至云端的轻量化。轻量化意味着使用成本的降低。

端侧成厂商新“角力场”

去年以来，大模型正逐步走向“智能终端侧”，国内外一些厂商纷纷宣布加速推进大模型在移动终端的部署。端侧大模型，正成为行业热词。

所谓端侧，是指模型服务不部署在云端，而直接存储在终端内部的芯片中，利用芯片算力生成结果。这样的大模型服务不需要联网，数据也不需要被上传到云端。

相比于云端大模型，端侧大模型的优势主要在于：无需云端处理信息，安全隐私性更好；不消耗云成本，高频使用下的成本更可控；弱网和无网环境下也可以使用，用户无需担心云端服务器宕机，交互体验更稳定。

中国工程院院士邬贺铨表示，通过模型压缩和定制人工智能芯片提升算力，将具有百亿参数大模型的推理能力嵌入手机，可以提供低成本、低时延、高安全的智能助手服务。

相比ChatGPT、Midjourney等人工智能应用依赖云端服务器提供服务，端侧大模型主打在本地实现智能化。甚至有厂商提出，让每个人在手机上都拥有“个人大模型”。

目前，在中国手机市场份额排名前五的企业中，除了苹果公司，其余均已发布自有端侧大模型产品。手机厂商提出的包括通话记录自动生成、人工智能图像处理等应用场景，日渐成为消费者的日常。

端侧大模型同样是个人电脑产品发布时的高频词。在相关宣传中，端侧大模型不仅能够提升日常使用及办公效率，还是保护隐私和数据安全的最优解。

艾媒咨询CEO兼首席分析师张毅表示，人工智能可能成为今年度的手机新卖点，尤其是手机的社交价值功能呈现方面，将有更多故事可讲。对于手机厂商而言，大模型是公认的未来发展赛道和方向。尽管现在人工智能在智能手机终端的应用仍较为初级，但未来的普及和推广仍值得期待。

不过，现有技术条件下，要做好端侧大模型并不容易。目前条件下，手机性能远远不足以支撑大模型运行。对于大模型来说，参数量是模型能力的直观体现，如今云端大模型的参数量均在千亿级别，而手机端最高参数规模则不过百亿。

联想集团董事长兼CEO杨元庆认为，未来十年是人工智能的十年，人工智能将改造所有业务。将来人人都会有自己的智能体，后者通过交互学习越来越懂用户，最终成为每个人的“人工智能双胞胎”。

但他也不讳言市场的不确定性。在他看来，内置个人智能体需要更高配置，如果这些对用户不是刚需，可能就不需要这么高配置。

国际数据公司认为，未来人工智能终端将在创作与创意、游戏和虚拟世界、语音合成与转换、视觉和图像处理、聊天机器人等十大领域广泛落地发展。

2024年度 5G轻量化贯通行动启动

科技日报讯（记者**崔爽**）记者4月22日获悉，为加快推进5G轻量化（RedCap）商用进程，打通5G RedCap标准、网络、芯片、模组、终端、应用等关键环节，近日，工业和信息化部印发通知，部署开展2024年度5G轻量化（RedCap）贯通行动。

行动具体包括七方面重点工作，分别为标准筑基，实现5G RedCap技术标准贯通；网络先行，完成5G RedCap网络贯通；能力升级，加快5G RedCap芯片模组贯通；产品丰富，推动5G RedCap终端贯通；示范带动，强化5G RedCap应用场景贯通；安全护航，促进5G RedCap安全能力贯通；强化保障，确保5G RedCap全面贯通。

5G RedCap是国际标准组织3GPP（第三代合作伙伴计划）定义的一种5G技术。它通过减少终端带宽、收发天线数量、降低调制阶数等方式，降低终端成本和功耗，有利于5G商用网络的规模普及和落地应用。业界初步估算，轻量化技术可使5G终端复杂度和成本下降50%—65%。2022年6月，RedCap在3GPP R17阶段实现标准冻结。

去年10月16日，工业和信息化部发布《关于推进5G轻量化（RedCap）技术演进和应用创新发展的通知》，提出到2025年，全国县级以上城市实现5G RedCap规模覆盖，连接数实现千万级增长。

根据相关部署，此次贯通行动将积极推进5G RedCap标准进程，2024年9月前完成基于3GPP R17版本的5G RedCap行业标准制定，构建涵盖基站、终端、通用模组等设备的全系列测试标准体系。开展面向R18版本5G RedCap演进技术研究，推动5G RedCap技术持续演进。在网络能力建设方面，鼓励重点城市已建5G基站完成5G RedCap升级，新建5G基站支持5G RedCap，2024年12月前实现超100个地级及以上城市城区连续覆盖，并按需向县城城区延伸覆盖，满足可穿戴设备、智慧汽车等移动场景的应用需求。在终端产品研发方面，围绕工业网关、摄像头、自动导引运输车等推出超100款产品，满足电力、工业、安防等领域的应用需求。着力突破5G RedCap在智能手表等消费类电子产品和车载终端设备的研发创新。

据悉，各运营商正积极推动5G RedCap应用创新和商用部署。中国移动支持5G RedCap的5G基站总规模已超10万个，覆盖全国52个城市，实现城区连续覆盖，今年将持续扩大部署规模，实现全国县级市以上连续覆盖。中国电信日前联手中国联通在浙江、贵州、广东、河南、上海等5省市现网环境下完成全频段、全制式、全场景5G RedCap商用验证，启动百城规模商用进程。

世界首个中药全产业链大模型发布

科技日报讯（记者**操秀英**）记者4月22日获悉，由成都中医药大学、北京百度网讯科技有限公司、太极集团有限公司、天府中药城等单位联合开发的全球首个中药全产业链大模型——本草智库近日在第二届“千种本草基因组计划”研讨会上发布。

据介绍，本草智库大模型基于中国工程院院士、成都中医药大学首席教授陈士林团队本草基因组学的研究成果构建。团队在建立千种药用植物基因组数据库、药用植物新品种选育、合成生物学、濒危药用植物就地保护及迁地保护等方面获得系列重要成果。

陈士林介绍，本草智库汇集了1500万条中药材基原物种基因信息、3000余万条中药成分与靶点互作信息、400余万个化合物等中药研究底层核心数据，同时融合团队主编的一系列中药领域权威专著精华，形成了覆盖中药全产业链的2000余万个实体和超20亿个关系对知识图谱。该模型以千亿级参数规模文心大模型为支撑，通过指令微调和检索增强生成技术，具备中药知识提取与生成、中药垂直领域解决方案输出、中药产业一站式数字化服务三大功能，实现了中药研究底层核心数据与中药全产业链关键环节有机结合。这一模型有助于提升中药基础研究和产业整体效率及质量水平，可为中药全产业链各关键环节提供精准决策支持，从而优化生产流程，提高产品质量，确保药品安全。

大模型安全领域两项国际标准发布 全球AI安全评估测试有了新基准

◎本报记者 **崔 爽**

第27届联合国科技大会期间，在以“塑造AI的未来”为主题的AI边会上，国际组织世界数字技术院（WDTA）发布了一系列突破性成果，包括《生成式人工智能应用安全测试标准》和《大语言模型安全测试方法》两项国际标准。这是国际组织首次就大模型安全领域发布国际标准，代表全球人工智能安全评估和测试有了新基准。

“随着人工智能系统，特别是大语言模型成为社会各方面不可或缺的一部分，以一个全面的标准来解决它们的安全挑战变得至关重要。”WDTA人工智能安全可信负责任工作组组长黄连金介绍，此次发布的标准汇集了全球人工智能安全领域的专家智慧，填补了大语言模型和生成式人工智能应用方面安全测试领域的空白，为业界提供了统一的测试框架和明确的测试方法，有助于提高人工智能系统安全性，促进技术负责任发展，增强公众信任。

随着人工智能系统，特别是大语言模型成为社会各方面不可或缺的一部分，以一个全面的标准来解决它们的安全挑战变得至关重要。

记者了解到，此次发布的两项国际标准是大模型及生成式人工智能应用方面的安全测试标准。

《生成式人工智能应用安全测试标准》以WDTA为牵头单位，为测试和验证生成式人工智能应用的安全性提供了框架。它定义了人工智能应用程序架构每一层的测试和验证范围，包括基础模型选择、嵌入和矢量数据库等，确保人工智能应用各方面都经过严格的安全性和合规性评估，有利于保障其在整个生命周期内免受威胁和漏洞侵害。

《大语言模型安全测试方法》以蚂蚁集团为牵头单位，为大模型本身的安全性评估提供了一套全面、严谨且实操性强的结构性方案。它提出了大语言模型安全风险分类、攻击分类分级方法以及测试方法，并给出四种不同攻击强度的攻击手法分类标准，提供了严格的评估指标和测试程序。它全面测试大语言模型抵御敌对攻击的能力，使开发人员和组织能够识别和缓解潜在漏洞，有利于提高使用大语言模型构建的人工智能系统的安全性和可靠性。

“一方面，生成式人工智能将释放巨大生产力。另一方面，我们也要对它带来的新风险高度警惕。大型科技公司应在促进生成式人工智能安全和负责任的发展中发挥关键作用，利用其资源、专业知识和影响力，构建一个优先考虑安全、隐私和道德的生态系统。”作为标准参与单位代表，蚂蚁集团机器智能部总经理、蚂蚁安全实验室首席科学家王维强在会议发言中说。他进一步解释，可通过制定行业标准与指南，为开发和部署生成式人工智能系统的开发者和机构提供清晰指导；投入研发并开放保障生成式人工智能安全的工具，形成产业共治格局等。

记者了解到，蚂蚁集团从2015年起积极投入可信人工智能技术研究，目前已建立了大模型综合安全治理体系。集团还自研了业界首个大模型安全一体化解决方案“蚁天鉴”，用于人工智能生成内容的安全性和真实性评测、大模型智能化风控、可解释性检测等。此次发布的《大语言模型安全测评方法》，便是基于“蚁天鉴”人工智能安全检测体系的应用实践，与全球生态伙伴交流编制而成。

Create 2024百度AI开发者大会在深圳举行

科技日报讯（记者**罗云鹏**）记者4月22日获悉，Create 2024百度AI开发者大会近日在广东省深圳市举行。

大会演讲环节，百度创始人、董事长兼CEO李彦宏发布三个AI开发工具：智能体开发工具AgentBuilder、AI原生应用开发工具AppBuilder、各种尺寸模型定制工具ModelBuilder。

“AI正在掀起一场创造力革命。未来，开发应用就像拍个短视频一样简单。人人都是开发者，人人都是创造者。”李彦宏说。

大会现场，李彦宏展示了用AgentBuilder打造的智能体案例。“今天，每一个商家、每一个客户，都能够在百度拥有专属智能体。”李彦宏说，“整个过程完全不需要编程，通过类似提示词的信息输入和简单的几步操作调优，就能迅速生成一个智能体。”

在应用开发方面，借助AppBuilder提前封装和预置的各种组件和框架，开发者仅需使用自然语言，即可开发出一个AI原生应用，并能将其便捷地发布到各类业务环境中。

ModelBuilder则可以根据开发者需求定制任意尺寸模型，并根据细分场景对模型进一步精调。

大会当日，文心大模型4.0工具版正式发布。据悉，相比一年前，文心大模型的算法训练效率提升到了原来的5.1倍，周均训练有效率达到98.8%，推理性能提升了105倍，推理成本降到了原来的1%。

另悉，大会还设置了深度论坛、AI公开课、AI互动体验区、AI音乐会等环节。

本版图片由视觉中国提供