2024年数字经济专题：大国经济体系下_人工智能领航数字经济新阶段

一、未来已来，人工智能正式开启第四次工业革命

（一）第四次工业革命概述，融合技术正在掀起历史性的技术浪潮

人工智能是核心技术催化剂，颠覆性技术之间的融合产生协同效应，带来生产力的爆发，成为第四次工业革命的主要推动力。人类文明已经历了三次工业革命，第一次是18世纪中叶以蒸汽机为代表的机械化革命，第二次是19世纪中叶以电力、内燃机为代表的电气化革命，第三次是20世纪中叶以信息技术为代表的自动化革命。当前我们正在进入第四次工业革命—智能化革命，以人工智能（神经网络/下一代云/智能设备）、公有区块链（加密货币/智能合约/数字钱包）、多组学测序（精准治疗/多组分技术/可编程生物学）、储能（自动驾驶/节能电池）和机器人（可复用火箭/自适应机器人/3D打印）为代表的五类颠覆性技术正在融合，融合产生的协同效应将发挥更大作用，推动生产力的发展。其中，人工智能是核心技术催化剂，与其他四类技术的融合范围最广、评分最高。

融合技术带来的实际 GDP 增速将远超第一次和第二次工业革命，AI 对经济 增长的贡献突出。据 ARK 估算，在全球范围内，未来 7 年的实际 GDP 增速将达 到 7%以上，而过去 125 年的平均增速只有 3%。AI 作为核心技术催化剂，对经济增长的贡献突出。据 ARK 估算，引进 AI 后，实际 GDP 在 2023 至 2030 年间有望 累积增长 130%。原因是在 AI 的赋能下，一些行业的生产效率和成本发生了巨大 的变化。比如：机器人与 AI 融合后，可以在非结构化环境中低成本高效地工作， 2030 年有望带来 24 万亿美元的经济效益；自动驾驶出租车与生成式 AI 融合提升 了安全性，到 2030 年有望广泛应用而使每英里成本低至 0.25 美分，创造一个 11 万亿美元的潜在市场；而 AI 软件直接提高了知识型工作者的生产力，2030 年有 望提升生产力至 2.5 倍，若软件价值量按 10%计算，则有望产生 13 万亿美元的经 济效益。

AGI 时代有望加速到来。近期，英伟达创始人 CEO 黄仁勋和谷歌 DeepMind CEO 哈萨比斯对 AGI 的到来时间进行了预测，他们的观点相似。哈萨比斯认为， AGI 最早可能在 2030 年出现，而黄仁勋则认为通用人工智能可能在五年内实现。 据 ARK 分析，根据赖特定律，加速计算硬件的改进将使 AI 相对计算单元（RCU） 的生产成本每年降低 53%，而算法模型的增强可以进一步带来每年 47%的训练成

本下降。换而言之，到 2030 年，硬件和软件的融合可以使人工智能训练成本以 每年 75%的速度下降。人工智能模型的训练成本下降将进一步加速其能力的迭代， AGI 有望加速到来。

（二）Sora 发布标志 AGI 系统有望超预期提前到来

1．DiTs 算法赋能 AIGC，Sora 开启文生视频新纪元

北京时间 2 月 16 日，OpenAI 发布“文生视频”大模型 Sora。可生成一分钟 的高保真视频，并配有 48 个生成案例及技术报告，能够通过自然语言指令生成 长达 60 秒的高清流畅视频，在生成视频长度、清晰度、连贯性、多镜头切换方 面都有显著提升。官方发布的技术报告指出，视频生成模型将是构建“世界通用 模拟器”的重要途径。

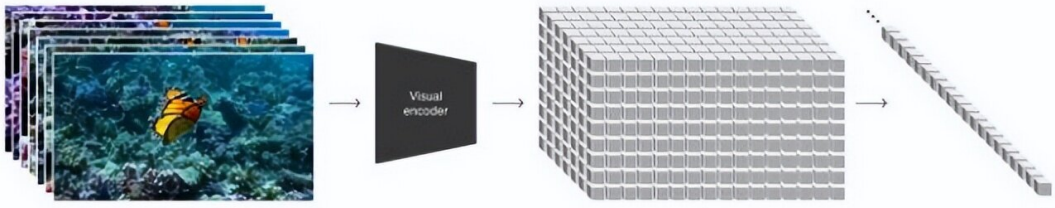
本质上，Sora 基于 Diffusion Transformers（DiTs）构建，并使用 DALL-E3 的 重捕获技术。研究表明，DiTs 相较于传统语义分割网络架构（U-Net）在模型大 小上更具可扩展性，并能生成更高质量的 2D 及 3D 图像。

OpenAI 曾采用循环网络、生成对抗网络、自回归 Transformer 和扩散模型等 方法对视频数据进行生成建模，但生成视频存在视觉数据受限、视频时长较短或 视频尺寸局限等问题。通过从大语言模型（LLM）中汲取灵感，类似 GPT

将自然语言转换为文本 tokens，Sora 可将视觉数据转换为 patches(视觉编码块)。通过此

种方式，即可实现在不同类型的视频和图像上训练生成模型。 通过视频压缩网络可降低视觉数据的维度，该网络将原始视频作为输入，并 在时间和空间上进行压缩，Sora 在压缩空间中接受训练并生成视频；另外，相应 的解码器模型可将生成的对象映射回像素空间。Sora 基于 patches 对可变分辨率、 持续时间和纵横比的视频和图像进行训练；亦可以通过在适当大小的网格中排列 随机初始化的 patches 来控制生成视频的大小。

图6：将视频压缩到一个较低维空间，再将其分解为 patches



资料来源：OpenAI，中国银河证券研究院

2．Sora 功能多样化，生成质量远超同类大模型

Sora 可以将简短文本描述转换成一分钟流畅视频，相对于 Runway、Pika、Stable Video 等同类大模型提升了几个代级。1）生成视频长度：Runway、Pika 等

传统文生视频大模型平均时长在 3-5 秒，Runway 用户可以最多延长视频长度至 16 秒，Sora 生成视频长度相对传统视频生成工具提升 15-20 倍；2）视频质量显 著提升：Sora 生成视频分辨率可达 1920 x 1080；3）可实现多镜头切换：可以理解和模拟运动中的物理规律，可以实现复杂的运动相机模拟；4）视频连贯性与 稳定性出色：在建模能力上表现更好，可以依赖关系进行建模，能初步理解并模 拟物理运动规律；5）高可拓展性：支持多种数据格式输入，具备实现文生视频、 图生视频、向前或向后视频扩展能力，同时支持视频连接。

3．Sora 发布标志 AGI 时代正加速到来，带动训练及推理算力需求增长

由于目前 Sora 处在初级阶段，训练数据集和参数规模有限，仍存在一些不足 之处。对于 Sora 当前存在的弱点，OpenAI 指出它可能难以准确模拟复杂场景的 物理原理，并且可能无法理解因果关系。该模型还可能混淆提示的空间细节，例 如混淆左右；或可能难以精确描述随着时间推移发生的事件，例如遵循特定的相 机轨迹。

AIGC 领域难度排序由易到难依次是文本、图像、音频、视频。未来训练数 据集将会数以万倍的增长，模型参数量也将不断提升，目前来看 Sora 训练所需算 力不及 GPT-4 等大语言模型，伴随 Sora 大模型不断迭代调优、训练数据集规模

逐渐扩大，我们认为，Sora 发布标志着 AGI 时代正加速到来：1）短期来看，模 型迭代优化、训练数据集增大将快速带动训练端算力需求；2）长期来看，Sora 技术逐渐成熟带动下游 AI 应用百花齐放，包括为影视制作、游戏开发、教育培训、广告及社交等场景应用带来颠覆，推理端需求将厚积薄发。

（三）国内 Kimi 加速迭代，AI 应用元年百花齐放

北京时间 3 月 18 日，Moonshot（月之暗面）宣布在大模型上下文窗口技术上 取得新的突破，其自研的 Kimi 智能助手已支持 200 万字超长无损上下文，并已 开启产品内测。 Kimi 智能助手在去年 10 月发布，支持 20 万汉字无损级别上下文输入，是当 时 AI 消费级产品支持上下文文本长度记录保持者。我们认为，Kimi 智能助手迭 代速度超预期，推动应用端加速落地，Kimi 智能助手目前已经支持 200 万文字超 长无损上下文，对比目前主流大模型：1）谷歌近期发布的 Gemini 1.5 pro 支持 100 万 token 输入；2）Claude 3 支持 20 万 token 输入；3）GPT-4 Turbo 支持 12.8 万 token 输入。我们依旧坚定年初观点，2024 年将是 AI 应用元年，Kimi 智能助手宣 布大模型进入“长文本时代”，长文本能力也将是通往 AGI 进程中的关键之一， Kimi 智能助手将是又一里程碑。

Kimi 智能助手支持多种应用场景，生成速度提升 3 倍之多。Kimi 智能助手 去年 10 月发布的版本仅支持 20

万上下文输入，时隔 3 个月，Moonshot 为 Kimi 智能助手提供了更多数据源，本次迭代升级使 Kimi 基于出色的长上下文处理能力 帮助用户解锁更多应用场景，比如专业学术论文的翻译和理解、辅助分析法律问 题、一次性整理几十张发票、快速理解 API 开发文档、快速筛选符合条件的简历 等。当面对一个问题时，Kimi 智能助手会尝试不同的方向搜索并据此做出回答。 在回答速度上也有提升，Moonshot 工程副总裁表示，基于 Infra 层的优化，Kimi 智能助手生成速度较去年 10 月份提升了 3 倍。 Kimi 智能助手访问量持续飙升，加速 AI 应用元年进程。根据 SimilarWeb 数 据显示，去年 12 月 Kimi 的周访问量还在 10 万次上下，到了 1 月下旬才突破 40 万，但是从春节开始访问量疾速攀升，至今周访问量已经超过 160 万次，2 月访 问量增长 107.6%，仅次于百度文心一言与阿里通义千问（访问量均下降超 30%）。 我们认为，2024AI 应用元年有两个条件：1）大模型达到可使用状态：这点从 Kimi 用户的如潮好评中可以看出。2）大模型公众可触达：目前 Kimi 已经面向全社会 开放使用。

（四）AIGC 到 AGI 将带来哪些颠覆式革命

通用人工智能（AGI）是指具备类似于人类思考能力，能够适应广泛领域并

解决多种问题的机器智能，是人工智能研究的重要目标之一。而狭义人工智能则指已取得显著进展但局限于特定领域的人工智能，例如语音识别、机器视觉等。 目前我们处于狭义人工智能相对成熟、通用人工智能乍现的阶段，GPT-4 等大语 言模型及 Sora 等多模态模型被认为是通向通用人工智能的重要潜在路径，并且这 一进程在逐渐加速。“超长文本”和“超强模拟物理运动”能力将是 AGI 时代关键。我们认为大 模型时代将会沿着两条路线继续演绎，一是支持超长上下文能力大模型，二是模 拟世界、物理运动规律的多模态能力。

多模态加速模拟文本、图像及视频，未来模拟物理运动规律将成为现实

多模态模型是指能够处理不同类型数据，并将其融合进行综合理解的人工智 能模型。这种模型能够更全面地理解和处理真实世界中复杂多样的信息，从而进 一步提升大型模型的迁移学习能力。多模态技术的发展在人工智能领域具有重要意义。 当前，单模态的人工智能模型，如处理文本、语音、图片等的模型，已经相 对成熟。而大型模型正在向多模态信息融合的方向快速发展：重要模型诞生以及 GPT-4 等模型的图像处理能力提升。大型模型不仅限于文字和图像的处理，也开 始拓展到音频、视频等领域，未来甚至有望延伸到包括味道等其他信号。

我们认为未来通用人工智能重点将变革以下领域：

1．具身智能成为 AI 发展新形态，机器人将取代大部分工种

具身智能作为人工智能发展的一个重要分支，是指那些可以感知、并与物理 世界进行交互、具有自主决策和行动能力的人工智能系统。这些智能体能够以主 人公的视角感受物理世界，并通过与环境的交互结合自我学习来理解和改变客观世界，机器人将成为具身最优载体。 未来机器人产业将持续快速发展，迎来战略机遇期。具身智能作为两大领域 交叉的核心应用，有望在未来取得快速发展。它将推动智能体具备更多自主规划、 决策、行动和执行的能力，实现人工智能的进一步进阶。

AI 与机器人结合能充分提高生产效率，更好应对各种复杂任务。在流水作业 以及生产线上，机器人可以根据大模型提供的实时数据分析结果，对生产流程进 行自动优化，提高生产效率和产品质量。并且能快速学习新技能，有效解放劳动力，降低劳动力工作时长，未来人类将更多时间应用在线上处理工作任务，一周 工作 3-4 天或将成为现实。

2．脑机接口成为新的创新交互方式

脑机接口作为一种新型的人机交互方式，在医学、教育、游戏等领域有着广

阔的应用前景。脑机接口技术将脑电信号转化为计算机可识别的数据，并通过计算机的处理和反馈，实现了人与机器的无缝互动。目前脑机接口技术正从“学术科学探索”走向“应用转化落地”。脑机接口 系统性能指标包括响应时间、识别正确率、可输出指令数量和菲茨吞吐量，可用 性指标包括易用性、长效性、鲁棒性、安全性和互操作性，易用性指标包括准备 时长、轻便性和舒适性。

经过数十年的科学探索与技术论证，脑机接口已从科幻成为科学，处于从科 学研究到产业落地的关键时期。就脑机接口目前的发展情况，在今后一段时间，脑机接口的基础学科研究和应用落地都将得到长足发展，从而有望促进脑机接口市场规模不断扩大。另一方面，人类与人工智能之间的交互方式也在不断升级，脑机接口有望成 为下一代人机交互方式。当前，脑机接口技术正在突破人类的生理界限，为残障 人士提供了前所未有的可能性。

3．治疗罕见病与精准治疗成为现实

未来 AGI 可以应用的领域众多，其中绕不开人类生物工程。我们认为，医疗 是 AGI 落地的最佳场景之一，大模型、多模态以及垂类大模型将更加广泛结合并 应用在药物研发、诊断、影像、治疗等细分环节。随着各类医疗大模型的加速迭代与演化，医疗大模型商业化前景有望进一步 打开。例如

，谷歌的 Med-PaLM2、微软子公司 Nuance 的 DAXExpress 等医疗大模

型已经在医疗领域得到应用，并取得了一定的商业化成果。在最近的一项研究中，Med-PaLM2 在 USMLE 问题上的准确率达到 85.4%，与参试专家的水平相当。这使得 Med-PaLM2 成为第一个在 USMLE 问题上达到专家级表现的人工智能系统。1）可以帮助医生更快速、更准确地进行诊断。它可以通过分析大量的病例和医学文献，提供对疾病的诊断和治疗建议。这有助于减少医疗错误和误诊的风险；还可以实现疾病的早期发现和治疗，从而改善患者的治疗效果并挽救生命。2）可以帮助医生节省时间和精力，使他们能够更专注于与患者的沟通和治疗。在 Med-PaLM 2 一项实验中，通过对超过 50 万个医学图像进行分析，成功预测了肺癌的发生率。Med-PaLM 2 还可以通过自动化完成目前由医生执行的许多日常工作，使他们能够有更充足的时间专注于为患者提供服务，可以缩短患者的等待时间并提高患者满意度。3）可以与其他医疗设备和系统进行集成，可以与智能手环和智能手表等设备进行连接，实时监测患者的生理参数，并提供相应的建议和警告。还可以与电子病历系统等其他医疗系统进行集成，从而实现更加智能化和高效化的医疗服务。

未来 AGI 时代，AI 赋能医疗有广阔前景。可以应用的领域很多，其中绕不开人类生物工程。我们认为，医疗是 AGI 落地的最佳场景之一，大模型、多模态

以及垂类大模型将更加广泛结合并应用在药物研发、诊断、影像、治疗等细分环 节，很多罕见病及疑难杂症将逐渐被治愈。

二、人工智能是核心技术催化剂，领航数字经济新阶段

（一）“适度超前”建设算力体系背后的财政货币支撑体系

1. 央行的“科技金融”、“数字金融”货币政策框架

1.1 结构性货币政策支持科技创新和数字经济发展

中央经济工作会议将“稳中求进、以进促稳、先立后破”作为 2024 年宏观 政策基调。在此背景下，货币政策“灵活适度，精准有效”，总量和结构政策双 重发力，既托底总量增长又推动结构改革，支持防范化解宏观风险，着力营造良 好的货币金融环境，高质量服务实体经济。“精准”即为强调信贷的方向引导。预计未来结构性货币政策工具将在货币 投放中扮演更加重要的角色，加大对重大战略、重点领域和薄弱环节的支持力度。 聚焦“五篇大文章”，即科技金融、绿色金融、普惠金融、养老金融、数字金融 五篇大文章。

中国人民银行货币政策司课题组近期文章《结构性货币政策助力做好“五篇 大文章”》中详细阐述了结构性货币政策工具的作用。“结构性货币政策是指在 市场配置资源基础上，设计适当激励机制，引导资金流向经济特定领域的货币政

策。结构性货币政策主要发挥结构功能，通过建立激励相容机制，将中央银行资 金与金融机构对特定领域和行业的信贷投放挂钩，发挥精准滴灌实体经济的独特 优势。结构性货币政策也有总量效应，通过投放基础货币，保持银行体系流动性 合理充裕，支持信贷平稳增长”。 截至 2023 年末，结构性货币政策工具余额 7.5 万亿元，比上年末增加约 1 万 亿元，占人民银行总资产的 16.4%。其中为“科技金融”和“数字金融”已创设 的结构性货币政策工具共 2 个，分别为“科技创新再贷款”和“设备更新改造专 项再贷款”。2022 年 4 月，人民银行创设科技创新再贷款，支持金融机构加大对 科技创新企业的信贷支持力度；2022 年 9 月，创设设备更新改造专项再贷款，支 持金融机构向制造业、社会服务领域和中小微企业、个体工商户等设备更新改造 提供贷款。根据央行披露，截至 2023 年第三季度末，科技创新再贷款 4000 亿元 额度全部用完，支持金融机构向科技企业累计发放贷款 1.69 万亿元；设备更新改 造专项再贷款累计发放 1694 亿元，支持新型基础设施和产业数字化转型等设备 更新改造。

2023 年科创企业贷款的获贷率提升，增速高于各项贷款增速。2023 年获得 贷款支持的科技型中小企业 21.2 万家，获贷率 46.8%，获贷率提升 2.1pct。科技 型中小企业本外币贷款

余额 2.45 万亿元，同比增长 21.9%，比上年末低 3.8pct， 比同期各项贷款增速高 11.8pct。2023

年获得贷款支持的高新技术企业 21.75 万家，获贷率为 54.2%，提升 0.8pct。高新技术企业本外币贷款余额 13.64 万亿元，同比 增长 15.3%，比上年末低 0.8pct，比同期各项贷款增速高 5.2pct。未来央行可能会继续在“科技金融”、“数字金融”领域创设相关结构性货 币政策工具，丰富政策工具箱，引导商业银行信贷直达实体经济，引导信贷投放 方向的作用，将为建设算力体系提供资金支持。

1.2 提升直接融资占比，构建覆盖科技型企业全生命周期的金融服务体系

央行 2003 年 4 季度货币执行报告首次提出“合理把握债券与信贷两个最大 融资市场的关系”，扩大直接融资，社融结构实现再平衡。提升直接融资的要求 与当下中国经济结构转型的新方向相匹配。报告专栏 1《准确把握货币信贷供需规律和新特点》提出“先进制造、科技创新、绿色低碳、数字经济等新兴产业蓬 勃发展，这些新动能领域与直接融资的金融支持模式更为适配，也会对贷款形成 良性替代”。央行行长在十四届全国人大二次会议经济主题记者会上强调“在宏 观层面，要加强顶层设计和系统筹划。比如，科技金融方面，科技型企业一般会 经历种子期、初创期、成长期、成熟期不同的阶段，企业成长周期的不同阶段，对金融需求有不同的特点。在科技型企业成长的早期，更多需要风险投资、创新

创业投资基金的介入，目前还是一个薄弱环节；金融机构对科技型企业风险评估 能力，也需要进一步提升，下一步需要着力补齐短板，构建覆盖科技型企业全生 命周期的金融服务体系”。 目前，交易商协会创设了科创类融资产品工具箱，并在 2022 年将其升级为 了科创票据。科创票据用于支持科创类企业以及非科创类企业的科技创新发展行 为。2022 年上交所和深交所正式落地了科创债。 因此，除了通过结构性货币政策工具为建设算力体系提供间接融资支持外， 对于相关企业的债券融资等直接融资的支持也会进一步提升，试图构建构建覆盖 科技型企业全生命周期的金融服务体系。

2. 财政政策支持重点首次转向现代化产业体系建设

从去年年底的中央经济工作会议到今年两会期间的政府工作报告，均把“加 快现代化产业体系”建设放在首要目标，与此同时持续强调要增强宏观政策取向 一致性。我们认为，在当前经济同时面临逆周期和结构性调节的关键时期，财政政策配合国有资本的政策调整将是有效发挥举国体制优势，支持数字经济和科技 转型的重要力量。其实，过去以来的历次经济周期中，积极的财政政策均起到了 关键性作用，力挽经济于狂澜。但我们也注意到，过去几轮积极财政政策主要支持的方向为传统基建产业，而现如今财政政策支持重心转向科技创新，需要通过 何种方式调整？

2.1 过去三轮积极财政政策主要投向传统产业

我国分别于 1998 年、2008 年以及 2013 年开启过三轮积极财政政策，其中前 两次均是由外部因素导致的需求冲击加剧了经济波动，继而使用扩张性财政政策 予以对冲。始于 2013 年的第三轮积极财政政策主要是由于传统产能的供给过剩 引起，财政政策方面主要以结构性政策为主，但由于土地财政等预算外广义财政 的存在，使得传统产能并未完全出清。由此，过去不同历史时期财政政策的主要 支持工具及资金投向也有所区别： 1998 年首轮财政支持经济转型：以狭义政府赤字和财政支出政策为主，促进 城镇化、工业化转型。受到亚洲金融危机的影响 1997 年 6 月至 1999 年 12 月 PPI 连续 31 个月在负值区间运行，且 CPI 处于负值区间。未来应对外需冲击，同时 促进我国城镇化和工业化的转型发展，我国分别于 1998 年至 1999 年两年期间多 次增发了共计 1600 亿元的国债，需要注意的是 1998 年我国 GDP 水平仅 8.5 万亿 元。当时我国刚刚实行改革开放不久，如图 15 所示，1994 年我国公路及民航建 设刚刚起步，国内基础设施建设具有大幅扩张的空间及较高的投资回报率，相对 而言资本是稀缺的。为此，1998 年新增国债由中央政府向国有商业银行定向增发 进行募资。

2008 年第二轮财政支持经济转型：以政府性基金扩张为主，催生地产与基建 快速发展。2008 年全球金融危机影响，中国国内需求遭受大幅冲击，PPI 负值运 行 12 个月，CPI 同样落入负值区间。为稳定经济增长，同时推动一批利长远项目建设，我国实施“四万亿”投资计划。资金来源方面，据国家发改委公布的“4 万亿”投资明细，四万亿投资的资金构成是 1.18 万亿中央预算内投资和 2.82 万 亿配套资金。资金来源是中央财政赤字、地方财政、地方债（财政部代理发行）、 政策性贷款、企业债和中期票据、银行贷款，以及民间投资。其中财政预算内较 上一轮积极财政中，多增了政府性基金收入中的土地收入增长。2009 年和 2010 年的政府性基金收入分别为 1.83 万亿元和 3.57 万亿元，较上年分别增长了 17.26% 和 95.15%，其中超过 70%专项用于土地开发及建设领域的投资和补偿。资金投向 方面，主要是重大基础设施建设、灾后重建、保障房建设、民生工程等。

2013 年后第三轮积极财政：以结构性政策配合完成过剩产能出清，但实际开 启了广义财政扩张。2008 年以来广义财政的扩张导致传统产能过剩，经济结构性 矛盾和隐性债务风险不断加剧。理论上此时财政政策应该以结构性政策配合货币政策完成“供给出清”，我们可以看到基建的投资缺口在 2015 年之后开始显著 提升，如图 19

所示。但实际上由于地方政府考核目标仍在，地方在稳经济过程 中实际开启了广义财政的扩张，而“宽货币、紧信用”的货币政策使得“宽”出 的货币进了城投和地产，”紧“的信用又使城投承担了较高的融资成本，造成了 一定隐性债务。

2.2 新一轮积极财政支持重点转向现代化产业体系建设

2023 年中央经济工作会议及 2024 年政府工作报告均将“现代化产业体系建 设”放在首位，与之对应的 2024 年财政预算草案中，对于主要财政收支政策也 做出了重要调整。根据过往几年预算草案来看，每年基本会公布 7-8 项主要政策，其基本规律是：后几项政策相对固定（第四是乡村振兴、第五是生态环境、第六 是民生、第七是国防、外交、政法），但前三项政策的内容和排序往往指明当年 重点方向。例如 2021 年首要政策是“推动创新发展和产业升级”、2022 年首要 政策目标是保市场主体、2023 年为扩大内需。 而 2024 年主要财政收支政策中的前两项均和科创相关，第一是支持加快现 代化产业体系建设，第二是支持深入实施科教兴国战略。去年的首要目标“扩大 内需”已经移居第三。据此可以得出的结论是：新一轮财政逆周期调控不同于以 往几次扩张，更加聚焦经济转型和科技创新。

2.3 当前及未来一段时期财政如何支持科技创新发展？

从今年财政预算草案及相关新增政策来看，未来具体支持政策主要以下几方 面：

一是新增政府债务工具重点支持科技创新。一方面是专项债用途在以往的传 统基建和新能源建设的基础之上，新增了数字基础设施建设用途。另一方面，我 们看到今年两会提出了新型债务工具—超长期特别国债，在 3 月 6 日举行的十四届全国人大二次会议经济主题记者会上，国家发展和改革委员会主任郑栅洁指出 初步考虑，超长期特别国债将重点支持科技创新、城乡融合发展、区域协调发展、 粮食能源安全、人口高质量发展等领域建设。这些领域潜在建设需求巨大、投入周期长，现有资金渠道难以充分满足要求，亟需加大支持力度。其中首要支持领 域便是科技创新。

二是税收政策持续深化落实，加大研发费用扣除比例。即落实技术改造相关 投资税收优惠，落实研发费用加计扣除等政策，例如去年已将符合条件的集成电 路和工业母机企业研发费用税前加计扣除比例提高至 120%，将符合条件的研发费用税前加计扣除比例由 75%提高至 100%。预计在今年及未来一段时间将持续 贯彻落实以上税收支持政策，且有望逐步扩大政策支持范围和支持力度。

三是更好发挥国有资本和国有企业在科技创新中的引领、引导作用。今年 1

月份国务院印发了《关于进一步完善国有资本经营预算制度的意见》，本次新《意 见》在之前基础上对于“国有资本”的功能定义新增了“落实国家战略”，并居 于首位，如表 6 所示。而当前我国首要战略显然是科技创新和数字经济的发展。实际上，国有资本经营预算作为财政收支的重要组成部分，既是中国特色，也是 我国更好支持科技转型的体制优势。本次制度的修订实现了所有国有企业的预算 全覆盖，并加强了支出纪律约束，未来对于关键领域的国有资本的注入有望助力科技创新相关行业的快速发展。

四是专项产业基金支持。去年我国的产业基础再造和制造业高质量发展专项 资金增长了 20.3%，今年预算草案再安排专项资金 104 亿元。其中主要强化对制 造业企业技术改造的资金支持，落实技术改造相关投资税收优惠政策。深入实施首台（套）重大技术装备和首批次重点新材料应用保险补偿政策。优化产业投资 基金功能，鼓励发展创业投资、股权投资，充分运用市场化手段，支持集成电路、 新一代信息技术等产业加快发展。

五是教育和研发预算支出支持。今年财政预算草案中主要财政政策第二位是“科教兴国”，从支持加快建设高质量教育体系和推动高水平科技自立自强两方 面做了阐述，主要支持政策是加大教育支出和财政研发补贴。于此同时，对于今

年中央财政支出的预算安排中，科技支出和教育支出预算安排的支出增速分别为 10%和 5%，大幅高于去年的 2.9%和 1.9%，是今年中央各项财政支出中主要提升 的项目。

（二）大国体系下的中国数字经济产业已全面开启，预计在 2035 年达到 GDP 的 71.6%

总量法测算：国内数字经济预计 2035 年占 GDP 比将达到 71.60%。近年来，我国数 字经济整体实现量的合理增长，2022 年数字经济规模达到 50.2 万亿元，同比增加 4.68 万 亿元，首次突破 50 万亿元。2023 年，面对经济新的下行压力，各级政府、各类企业纷纷 把发展数字经济作为培育经济增长新动能、抢抓发展新机遇的重要路径手段，数字经济 发展活力持续释放，我国数字经济规模有望达到 54.6 万亿元，面对多方面不利因素，我 国数字经济仍保持强劲增长、凸显韧性，持续为国民经济稳增长保驾护航。 我国数据经济规模维持高位增长，增速连续 11 年高于名义 GDP 增速。2022 年，我 国防控取得重大胜利，经济发展环境得到改善，国内生产总值同比名义增长 5.3%， 数字经济规模达到 50.2 万亿元，同比名义增长 10.3%，高于 GDP 名义增速 4.98 个百分点。 自 2012 年以来，我国数字经济平均增速 15.9%，已连续 11 年显著高于 GDP 增速，数字 经济持续发挥经济“稳定器”、“加速器”作用。

我国数字经济占 GDP 比重持续提升，2030 年数字经济占比有望追上发达国家水平，2035 年有望位列全球首位。2022 年，我国数字经济占 GDP 比重为 41.86%，从 2023-2035 年的整体趋势及预测来看，中国数字经济占 GDP 的比重持续提升，我们预测 2030 年占比 达到 59.73%，有望追上发达国家平均水平，预计 2035 年占比将达到 71.60%。

（三）中美数字经济及人工智能产业要素发展对比

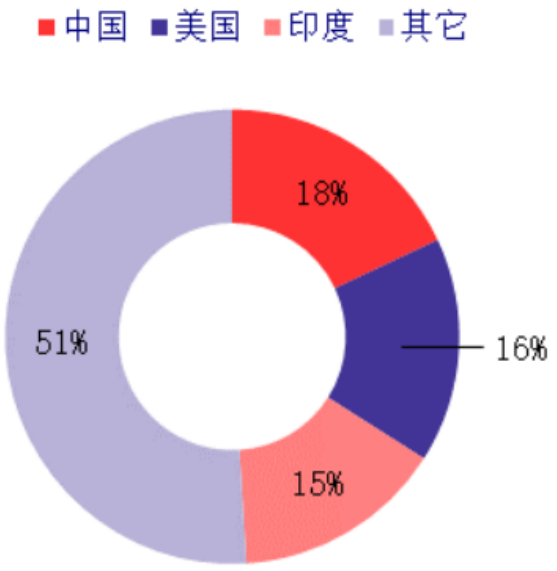
人工智能的三大基础要素为数据、算力和算法，目前中国在数据方面具有明显的大 国优势。首先，数据的底层是人和人的活动，因此发展主体（国家或者区域内）的人口 数量与质量对数据资源的“量”与“质”起到至关重要的影响。数量方面，目前世界人 口排名前列的国家或地区依次为印度 14.17 亿、中国 14.12 亿、欧盟 4.48 亿、美国 3.33 亿。据 IDC 统计，2022 年中国产生的数据规模达 23.3ZB，在全球占比达到 23%。伴随国 内数据要素成为国家战略，数据价值不断释放，到 2025 年中国人工智能产业在数据量方 面将具有相对优势，IDC 预计到 2025 年中国数据量规模达到 48.6ZB。质量方面，2022 年 中国人均 GDP 为 1.27 万美元，持平全球平均水平，但从互联网渗透率来看，中国达到 75.6%，明显高于世界平均水平 63%。

算力方面，中美两国的计算力与其他梯队国家相比有显著优势，但美国仍领先于中 国。由于超大规模互联网企业在算力投入上的大幅增长，2022 年美国算力指数从 77 分增 长到 82 分，尤其是一级指标计算能力增长显著，包括通用计算能力和 AI 计算能力。从 数据来看，2022 年，美国服务器市场规模达到 530 亿美元，同比增长 19.7%；人工智能 服务器市场规模达到 75 亿美元，同比增长 48.1%，是全球服务器市场增长的主要驱动力。 中国受阻于反复，2022 年算力投入有所放缓，但整体增速仍高于 GDP，算力指数从 70 分增长到 71 分。在这样的大背景下，中国整体服务器市场规模仍然保持 6.9%的正增 长，达 270 亿美元，占全球市场 25.0%，仅次于美国稳居第二。从服务器子市场来看，边 缘计算服务器和液冷服务器市场均呈现 20.0%以上的增长。

算法方面，我们主要通过人才规模、专利数、企业数量、融资规模和大模型这几个 角度来对比。1）人才方面，人工智能人才整体规模中国居首，但人工智能顶尖人才美国处领先地 位。中国人工智能人才规模凸显，LinkedIn 和猎聘的数据统计显示，全球当前累计 AI 人 才突破 100 万人。其中中国 AI 人才占全球 AI 人才总数的 18%，位居世界首位。美国和印度 AI 人才数量分居全球第二、三位，且均超过 15 万。根据清华大学 AMiner 的历年统 计数据，美国学者入选 AI2000

榜单年均超过 1,000 人次，数量最多且远高于其他国家，表明美国在 AI 顶尖人才方面具有领军地位。与之对比，中国人才数量位居第二，但差距仍然较大。2）专利方面，中国信通院发布的《全球数字经济白皮书（2023 年）》指出，在专利申请授权方面，2013-2023 年 Q3，全球 AI 专利申请量累计达 129 万，全球 AI 专利授权量累计超 51 万，中国 AI 专利申请量占全球 64%，位列全球第一，论文数也遥遥领先。3）融资方面，当前美国在人工智能企业数量和融资规模方面均占据显著优势。截至 2023 年 6 月底，全球人工智能企业共计 3.6 万家，美中英企业数量名列前茅。美国人工智能企业数量约为 1.3 万家，在全球占比达 34%，中国占比 16%，英国 7%，美中英三国的人工智能企业数量合计占全球的 56%。计算 2013 年以来的私人 AI 投资总额时，美国以 2489 亿美元的投资排名第一，其次是中国（951 亿美元）和英国（182 亿美元）。AI 企业和融资活动集中在美、中、英等国家。

图26：全球人工智能人才数量占比



资料来源：LinkedIn,猎聘，清华大学 AMiner 团队发布的 AI2000 学者榜单，智谱研究，尚普研究院，中国银河证券研究院

大模型方面，我们通过大模型数量和大模型表现来进行比较。大模型数量：根据赛 迪顾问数据，截至 2023 年 7 月，中国已累计发布 130 个大模型；国外共发布 138 个，其 中美国共 114 个。中美两国大模型合计数量占全球 90%以上，具有绝对优势。从大模型 参数来看，中美两国的代表性大模型里均是既有千亿参数的通用大模型，又有几十亿参 数的行业垂直大模型。但是，在影响力方面，中国缺少像 GPT-4、Gemini、Sora 这种具有 全球影响力的大模型，中国的行业大模型居多，占总数的 60%，商业、金融、医疗居多。

大模型表现的比较，我们主要参考国内较权威机构 SuperCLUE 在 23 年 7 月至 24 年 2 月的测评结果。测评是基于 4572 道中文评测题，可以看到在过去半年里国内领军大模型 在不断进步，与 GPT-4 的差距在不断缩小。24 年 2 月的测评结果，在中文领域国内领军 大模型的平均水平已经接近 GPT-4。SuperCLUE 还对全球大部分模型进行了测评，从 24 年 2 月的结果来看，国内在中文领域综合能力超过 GPT3.5 的模型有 13 个，文心一言 4.0、 GLM-4、通义千问 2.1 排名前三；国外模型的平均成绩为 57.83 分，国内模型平均成绩为 68.75 分。可以看出，国内大模型在中文领域的能力的平均水平已经超过国外大模型。

三、人工智能三要素共振，算力、算法、数据未来趋势推演

（一）算力：供需缺口加大，AI 服务器产业链分析

1．大模型时代智能算力渗透率持续提升，AI 服务器有望量价齐升

服务器处理大规模计算，主流架构分为 X86 和 ARM。服务器是高性能计算机，比普 通计算机运行更快、负载更高、价格更贵。主要功能包括运行网站、存储数据、运行程 序、数据分析、云计算开发、人工智能运算等，在网络中为其它客户机如 PC 机、智能手 机、ATM

等终端等大型设备提供计算或者应用服务，具有高速的CPU 运算能力、长时间 的可靠运行、强大的 I/O 外部数据吞吐能力以及更好的扩展性。

AI 服务器中用于运算和存储的芯片占服务器成本结构约70%，通用服务器用于运算 和存储芯片占服务器成本结构大约 50%左右。一台服务器主要硬件包括处理器、内存、 芯片组、I/O(RAID 卡、网卡、HBA 卡)、硬盘、机箱(电源、风扇)。以一台普通的服务器 生产成本为例，CPU 及芯片组大致占比 50%左右，内存大致占比 15%左右，外部存储大 致占比 10%左右，其他硬件占比 25%左右。其中机器学习型服务器中 GPU 成本占比达 72.8%。

全球服务器市场高增长，中国市场占比提升。根据 Statista 数据，2022 年全球服务器 市场规模达到 848.7 亿美元，同比增长 2.04%，中国服务器市场占比 30.16%，预计 2023 年全球服务器市场规模来到 907.8 亿美元，同比增长 6.96%，中国服务器市场占比 33.93%， 变化+3.77pct。我们认为，随着人工智能所需算力扩大，未来中国服务器市场有望进一步扩大。

人工智能时代 AI 服务器优势凸显。随着 AI 技术升级应用，CPU 的串行处理架构不 能满足 AI 时代的算力需求，企

业需要为人工智能、机器学习和深度学习建设全新的 IT 基础架构，逐渐由 CPU 密集型转向搭载 GPU、FPGA、ASIC

芯片的加速计算密集型机构，且越来越多地使用搭载 GPU、FPGA、ASIC 等加速卡的服务器。全球 AI 服务器市场预计 2025 年达到 1350 亿美元，未来三年 CAGR 有望超过 50%。研究机构 Aletheia 报告指出，预估 AI 服务器市场规模将在 2024 年翻倍、2025 年达到 1350 亿美元，是 2022 年规模的 4.5 倍。此外，根据 TrendForce 预测，2026 年全球 AI 服务器出 货量将进一步提升，2022 至 2026 年 CAGR 达 10.8%。

从搭载芯片种类上来看，目前全球以 GPU 服务器为主流。据 IDC 统计，2022 年全球 GPU 服务器出货量占比 87.3%，GPU 服务器销售额占比 89.5%。

中国 AI 服务器市场存量替换需求叠加增量需求，有望迎来量价齐升。根据中商产业 研究院数据及我们预测，中国 AI 服务器市场预计 2026 年市场规模超千亿元，未来三年 复合增长率 21.65%，预计 2026 年出货量 64.5 万台，未来三年复合增长率 15.26%。中国 AI 服务器受益于人工智能等相关新兴领域的应用以及“东数西算”政策下，云计算、超 算中心的蓬勃发展，数据计算、存储需求呈几何级增长，算力需求持续释放，AI 服务器 作为算力基础设备保持较快增速。1）存量来看：服务器平均寿命 3-5 年更换一次每年根 据算力需求使用需求变化产生比较明显的更新需求。2）增量来看：伴随人工智能浪潮以

及数字中国建设，未来对智能算力需求将持续爆发增长，且智能算力增长速度远超算力 总体增速，中国 AI 服务器市场将迎来爆发增长，占比将逐步提升。2021-2026 年我国 AI 服务器市场规模由亿 350 亿元增长至 1089.4 亿元，2021 至 2026 年 CAGR 为 20.83%。

2．Sora 等多模态加速应用端落地，推理服务器需求激增

多模态对推理算力需求指数级增长，推理服务器占比将持续提升。根据 IDC 预测， 2023 年 AI 服务器训练需求占比达 41.5%，随着大模型的应用，该比例在 2025 年将降低 至 39.2%；将 GPT-4 的推算结果作为训练需求，进一步推算 2023、2025 年推理需求最高 达 44081、48502PFlop/s-day。

单个 AI 应用如 ChatGPT 可以带动推理算力 66 亿美元需求。假设平均针对 20 字的提 问生成 200 字的响应，对应 267tokens，根据 OneFlow 的数据和《Scaling Laws for Neural Language Models》，在推理过程中每个 token 的计算成本约为 2*N Flops，其中 N 为模型 参数数量，则在 ChatGPT4 一万亿参数中每个 token 需算力 2 万亿 Flops。假定 GPT-4 训练 期间 FLOPS 利用率为 32%，则每人每次提问需要算力：2 万亿 *267tokens/32%=17PFlops。 据官网 9 月数据，ChatGPT 目前拥有超过 1 亿用户，每月产生 18 亿次访问量，假定每日 访问

量为 6000 万人次，每人提问 10 次，且假设一天平均分布，则每秒算力需求为 118EFlops，目前 AI 推理使用的主流 GPU 是

T4，提供混合精度算力 65TFlops，则需要 182 万个 T4GPU 可满足单日访问量，对应 22.75 万台 8*T4 服务器，一台 8*T4 服务器的价格 约为 29000 美元，则目前来看推理服务器的需求在 66 亿美元。

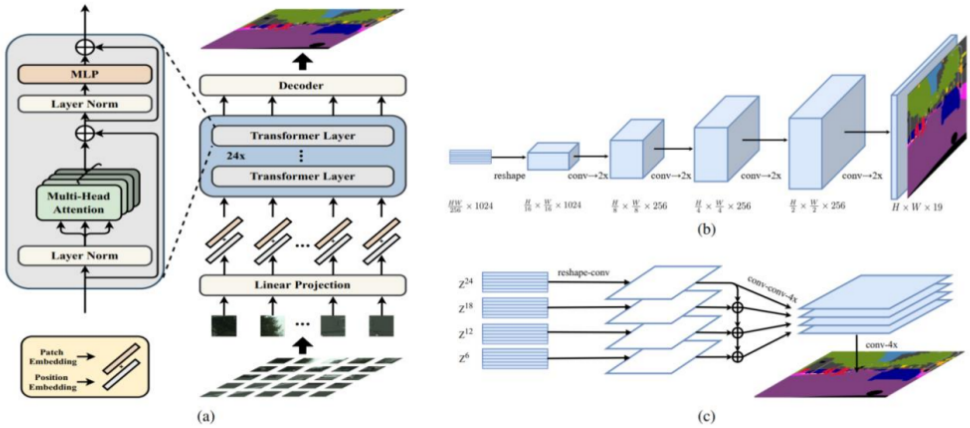
Sora 训练一次所需算力或可达到 2.6×10^24Flops，相当于 GPT-3 175B 的 8.2 倍。多 模态大模型在训练端算力需求通常在计算大语言模型算力需求通常与参数量及 token 数量 成正比，而 Sora 大模型中可以将 Patch 类比与大语言模型中 token，基于大语言模型计算算力需求方法框架及以下三大假设，对 Sora 算力需求进行分析测算。

假设一：Sora 训练数据集为 60 亿张图片，分辨率为 1980×1024；3500 万个视频， 每个视频平均时长为 30 秒，分辨率为 1980×1024，帧率为 60FPS。 根据阿里联合浙江大学、华中科技大学提出的文生视频模型 I2VGen-XL，研究人员 收集了大约 3500 万单镜头文本-视频对和 60 亿文本-图像对来优化模型。我们暂且保守 假设 Sora 训练数据集与 I2VGen-XL 相同，同时二维向量空间图片表示为 H×W×C（其 中 H 为长度，W 为宽度，C 为 RGB 颜色通道数，假设 C=3）。 我们估算 Sora 训练数据集中视频类数据 Patch 规模 =3500×10^4×60×30×3=1.89× 10^11；图片类数据 Patch

规模=60×10^8×1024×1980×3=3.65×10^16；训练数据集总 Patch=图片类数据 Patch+视频类数据 Patch=3.65×10^16。

假设二：Sora 中 Patch Size 为 16×16，将 Patch 转化为 token。 根据谷歌论文《AN IMAGE IS WORTH 16×16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE》，Transformer 的输入是一个序列，对于一张图像来说如果把 每个像素点当作一个 token，那就会需要相当旁大的计算量，该文则将图像划分为 16×16 大小的一个个 Patch，然后将每个 Patch 当作一个 token 组成一串序列作为 Transformer 的 输入，减少了计算成本。我们假设 Patch Size 为 16×16，通过将 Patch（N×P×P×C）转 换为 token，N 大小为 H×W/（P×P），每个 token 的大小为 P×P×C，P=16，通过计算 得到 token=3.65×10^16/（16×16）=1.43×10^14。

图44：Tranformer 语义分割方法



假设三：Sora 模型参数为 30B，训练一次所需总算力=模型参数量×token 数量×3×2。 根据 OpenAI 论文，T5

模型由于采用编码器-解码器模型，在向前和向后传播的过程 中只有一半 token 处于激活状态，而 BERT 与 GPT 基于 Transformer 的自然语言监督模型， 每个 token 都处于活跃状态，而每个 token 都在向前传播过程中涉及一次加法和一次乘法， 论文添加一个 3×的乘数来计算向后传递 的计算量，故推出 GPT 模型所需算力：训练所 需总算力=模型参数量×token 数量×3×2×训练轮数。通过上述公式我们计算得到 Sora 训练一次所需算力=30×10^8×1.43×10^14×3×2=2.6×10^24Flops。

根据上述测算，基于 Sora 参数量大概在 30 亿（待确认）水平，同时采用 I2VGen-XL 训练数据集水平进行估算，我们保守估计，Sora 训练一次所需算力或可达到 2.6×10^24Flops，相当于 GPT-3 175B 的 8.2 倍（测算采用参数和训练数据集规模会与实际有一 定出入）。

假设四：Sora 模型训练不考虑利用及其他成本，大约需在 1 万张 A100 上训练 154 天。 单张 A100 算力为 19.5TFlops，暂时不考虑模型训练利用率及其他训练成本，如果在 10000 张英伟达 A100 进行训练，所需时间 =2.6×10^24/(19.5×10^12×10000)/(24×60×60) ≈154 天。

3.“人工智能+”鼓励数字基础设施“适度超前”发展，国产服务器迎来曙光

2024 年 3 月 5 日，国务院总理李强在政府工作报告中提出，要大力推进现代化产业 体系建设，加快发展新质生产力。同时，深化大数据、人工智能等研发应用，开展“人 工智能+”行动，打造具有国际竞争力的数字产业集群。另一方面，报告中也提出，要深 入推进数字经济创新发展，制定支持数字经济高质量发展政策，积极推进数字产业化、 产业数字化，促进数字技术和实体经济深度融合。适度超前建设数字基础设施，加快形 成全国一体化算力体系。 政府工作报告中首次提出“人工智能+”行动，数字基础设施建设未来将成为经济核心抓手。在政府工作报告中被首次提出的“人工智能+”行动，正是推进数字产业化、产 业数字化的重要举措。另一方面，在实践过程中，政府和企业都需要加大对人工智能领 域的研发投入，提高对我国人工智能产业发展中技术创新、跨界人才等资源的储备力度。

“适度超前”建设算力基础设施，国产服务器大势所趋。报告中提出，适度超前建 设数字基础设施，加快形成全国一体化算力体系。我们认为，“适度超前”一方面将驱 动数字产业化加速发展，算力基础设施国产化提上日程，芯片、服务器国产化率将进一 步提升，建议重点关注上游算力基础设施如国产芯片、国产 AI 服务器、光模块、液冷等

细分赛道机会；另一方面，将持续推动产业数字化升级转型，算力基础设施将大范围赋 能千行百业，降本增效，为新质生产力发展提供新动能。 出口禁令影响海外供应，AI 服务器国产化进程提速。2023 年 10 月 17 日，美国商务 部工业和安全局（BIS）发布了针对芯片的出口禁令新规，更加严格的限制了中国购买重 要的高端芯片。一方面，从 ChatGPT 面世以来，国内各企业和研究院在短短半年多的时 间内先后推出了超过 130 款大模型，其中领跑玩家已经开始着手于将大模型应用于特定 场景，打造爆款应用。另一方面，为了构筑算力底座，各地政府纷纷上马智算中心建设， 铺设大数据时代的信息高速，推动产业创新升级，降低企业调用以大模型为代表的科技 成果的成本。根据华为昇腾计算业务总裁张迪煊在 2023 世界人工智能大会上的揭示，大 模型所需的算力相对于 2020 年预计将增长 500 倍，这个算力缺口正在不断扩大。A800、H800 被禁后，英伟达继续推出新款芯片，单卡性能 H20 弱于昇腾 910b。2023 年 11 月 9 日，相关报道称英伟达已开发出针对中国市场的最新改良版系列芯片 ——HGXH20、L20PCIe 和 L2PCIe。最新三款芯片是由 H100 改良而来，就单卡性能而言 H20 弱于昇腾 910b。

华为昇腾芯片为 AI 体系提供强大算力，昇腾 910b 单卡性能接近英伟达 A100。华为 昇腾芯片是华为发布的两款人工智能处理器，包含昇腾 310 用于推理和 910 用于训练，

均采用自家的达芬奇架构。昇腾 910 是一款高性能 AI 芯片，采用了 7nm 工艺制程，集成 了数千个达芬奇核心，能够提供高达 256TOPS 的算力，在业界其算力处于领先水平。昇 腾 310 是一款入门级 AI 芯片，采用了 12nm 工艺制程，集成了数百个达芬奇核心，能够 提供高达 8TOPS 的算力，适合用于边缘计算和物联网等应用场景。2023 年科大讯飞与华 为昇腾启动专项攻关，合力打造我国通用人工智能新底座，让国产大模型架构在自主创 新的软硬件基础之上，当前华为昇腾 910B 能力已经基本做到可对标英伟达 A100。

华为昇腾生态打开市场空间，国产算力产业链有望持续受益。我们认为，国内第一 批大模型厂商使用的基本都是英伟达 A100、A800 的芯片，因为英伟达构建了完善的 CUDA 生态，贸然换生态，意味着学习成本、试错成本、调试成本都会增加。目前华为基于“鲲 鹏+昇腾”双引擎正式全面启航计算战略，打造算力底座，未来趋势下，华为昇 腾市场份额将不断提升，产业链细分赛道上市公司有望持续受益。15%到全国产化是大概 率事件，国产化空间巨大。根据 IDC 数据，2022 年中国 AI 芯片出货量约 109 万张，其 中英伟达市占率约为 85%，华为在内的国产 AI 芯片市占率约为 15%，国产化仍有很大空 间。