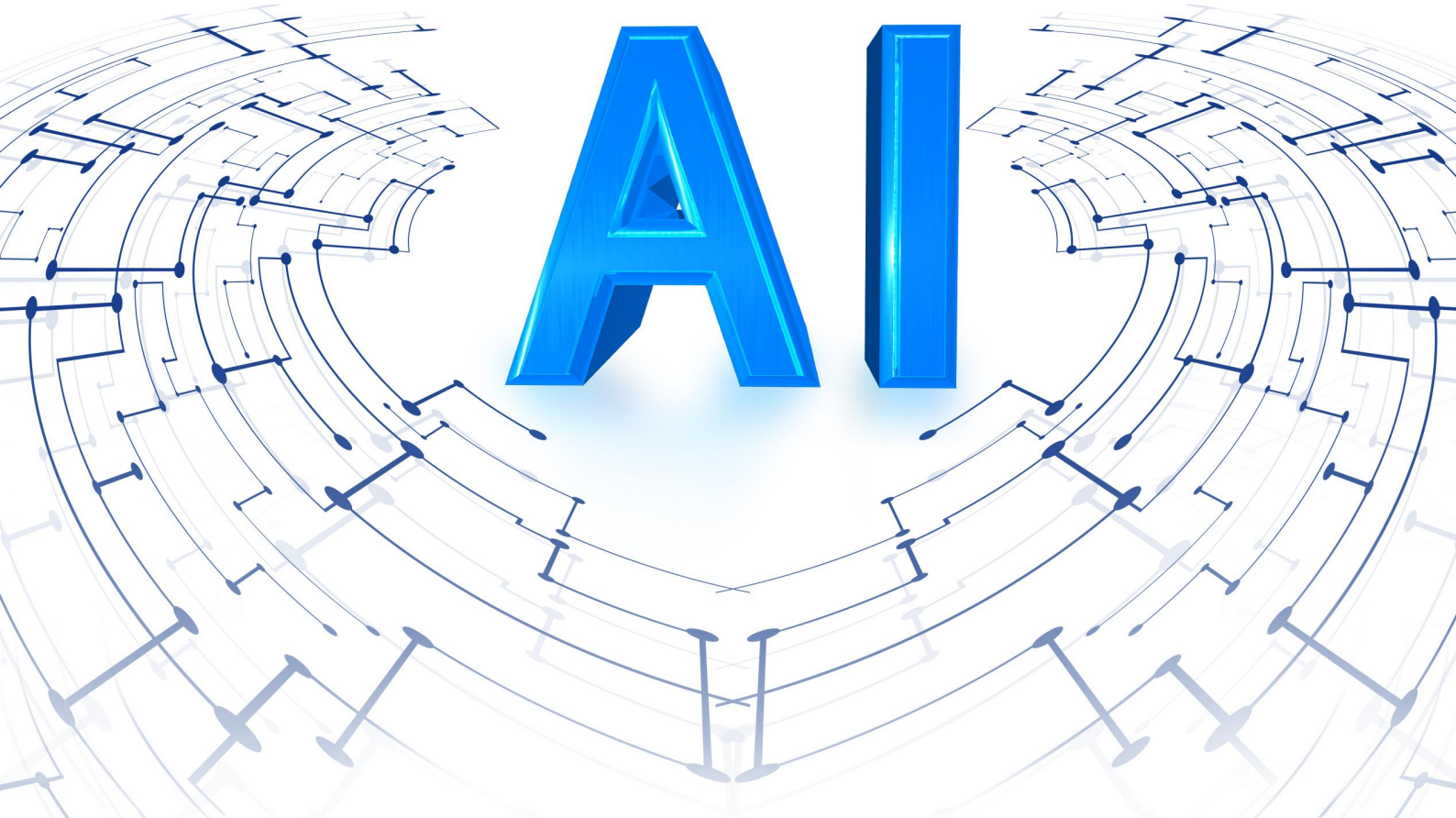
开启智能新时代

2024年中国AI大模型产业发展报告



人民网 people.cn 财经研究院

至顶科技 ZHIDING

2024年3月

前　言

伴随人工智能技术的加速演进，AI 大模型已成为全球科技竞争的新高地、未来产业的新赛道、经济发展的新引擎，发展潜力大、应用前景广。近年来，我国高度重视人工智能的发展，将其上升为国家战略，出台一系列扶持政策和规划，为 AI 大模型产业发展创造了良好的环境。当前，通用大模型、行业大模型、端侧大模型如雨后春笋般涌现，大模型产业的应用落地将进一步提速。作为新一代人工智能产业的核心驱动力，AI 大模型广泛赋能我国经济社会的多个领域，打开迈向通用人工智能的大门，推动新一轮的科技革命与产业变革。在大模型盛行的时代，产业发展到何种阶段，遇到何种挑战，未来将走向何方，这些都是需要面对的问题，亟需社会各界共同努力。

在此背景下，人民网财经研究院、至顶科技联合发布《开启智能新时代：2024 年中国 AI 大模型产业发展报告》，报告对于 AI 大模型产业发展背景、产业发展现状、典型案例、挑战及未来趋势等方面进行了系统全面的梳理，为政府部门、行业从业者以及社会公众更好了解 AI 大模型产业提供参考。

专家寄语

2023 年人工智能大模型全面爆发，给科技创新、生产生活带来重大变革、机遇和挑战。全球大模型竞争日趋激烈，众多国产大模型脱颖而出。大模型与电力、零售、出版等传统行业的成功融合，展现对传统产业改造提升的潜力。大模型赋能金融、医疗等行业提质增效，对推动新质生产力快速发展起到重要作用。大模型持续健康发展，需要政策法规保驾护航，满足隐私保护、数据安全等多方面要求。期待未来大模型持续深耕技术创新并服务于各行各业，为全社会全方位地注入高质量发展的新动能。

李君 传播内容认知全国重点实验室专职副主任

AI 大模型的出现，使得利用人工智能技术来生成内容，从“可用”跨越到“好用”。生产内容是所有行业共有的需求，如今大模型已经在电商、影视、传媒等领域被规模应用。大模型的商业化需要供需双方同时发力：供给侧来看，以 Transformer 为代表的根技术存在显著成本问题，当前大模型还有进一步压缩成本、提高性价比的空间；需求侧来看，企业高效应用 AI 大模型的必然前提是，投入大量资金、人力、时间以提升企业自身数字化程度。未来，人工智能生成内容从“好用”到“高效”，也许会再经历一次或多次技术范式的颠覆。

王蕴韬 中国信息通信研究院人工智能研究中心副总工程师

2024 年，多重利好因素将推动大模型快速发展，首先是“人工智能+”行动等来自政府层面的有力支持，其次用户提升生活、工作效率的需求激增，再加上科技公司加大对 AI 领域投入资金、人力、技术研发，各环节协同支撑大模型发展。当前大模型产业也面临挑战，包括算力分散不足、Transformer 结构是否为最优的疑问、领域数据稀缺、缺少现象级应用的问题。就产业趋势而言，投入基础模型训练的公司未来可能会大幅减少，转而更多的公司会去寻找应用场景和爆款应用。vivo 结合自研大模型端侧化、矩阵化的技术优势并且会聚焦手机行业的应用经验，利用大模型重构手机各类功能，找到落地场景，普惠更多用户。

周围 vivo 副总裁、vivo AI 全球研究院院长

人民网 people.cn 财经研究院

目　录

第一章 扬帆起航：中国 AI 大模型产业发展背景

1.1 中国 AI 大模型产业发展政策驱动力

近年来，我国始终高度重视人工智能发展机遇和顶层设计，发布多项人工智能支持政策，国务院于 2017 年发布《新一代人工智能发展规划》。科技部等六部门也于 2022 年印发《关于加快场景创新 以人工智能高水平应用促进经济高质量发展的指导意见》对规划进行落实。2024 年《政府工作报告》中提出开展“人工智能+”行动。伴随人工智能领域中大模型技术的快速发展，我国各地方政府出台相关支持政策，加快大模型产业的持续发展。当前，北京、深圳、杭州、成都、福建、安徽、上海、广东等地均发布了关于 AI 大模型的相关政策。具体来看，北京着力推动大模型相关技术创新，构建高效协同的大模型技术产业生态；上海强调打造具备国际竞争力的大模型；深圳重点支持打造基于国内外芯片和算法的开源通用大模型，支持重点企业持续研发和迭代商用通用大模型；安徽从资源方面着手吸引大模型企业入驻；成都着力推动大模型相关技术创新，重点研发和迭代 CV 大模型、NLP 大模型、多模态大模型等领域大模型以及医疗、金融、商务、交通等行业大模型；杭州支持头部企业开展多模态通用大模型关键技术攻关、中小企业深耕垂直领域做精专用模型。

2023 年以来我国各地出台的大模型产业相关政策

发布时间	发布机构	政策标题	政策内容
2023 年 5 月	北京市人民政府	《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025 年）》	支持创新主体重点突破分布式高效深度学习框架、大模型新型基础架构等基础平台技术。着力推动大模型相关技术创新。构建高效协同的大模型技术产业生态。建设大模型算法及工具开源开放平台，构建完整大模型技术创新体系。组建全栈国产化人工智能创新联合体，搭建基于国产软硬件的人工智能训练和服务基础设施，研发全栈国产化的生成式大模型，逐步形成自主可控的人工智能技术体系和产业生态。

2023 年 5 月	北京市人民政府办公厅	《北京市促进通用人工智能创新发展的若干措施》	**高效推动新增算力基础设施建设：** 加快推动海淀区、朝阳区建设北京人工智能公共算力中心、北京数字经济算力中心，形成规模化先进算力供给能力，支撑千亿级参数量的大型语言模型、大型视觉模型、多模态大模型、科学计算大模型、大规模精细神经网络模拟仿真模型、脑启发神经网络等研发。 **开展大模型创新算法及关键技术研究：** 围绕模型构建、训练、调优对齐、推理部署等环节，积极探索基础模型架构创新，研究大模型高效并行训练技术和认知推理、指令学习、人类意图对齐等调优方法，研发支持百亿参数模型推理的高效压缩和端侧部署技术，形成完整高效的技术体系，鼓励开源技术生态建设。
2023 年 5 月	中共深圳市委办公厅、深圳市人民政府办公厅	《深圳市加快推动人工智能高质量发展高水平应用行动方案（2023—2024 年）》	重点支持打造基于国内外芯片和算法的开源通用大模型；支持重点企业持续研发和迭代商用通用大模型；鼓励大模型企业联合生态伙伴加强大模型插件及相关软硬件研发，推动大模型与现有的操作系统、软件、智能硬件打通、互嵌。
2023 年 7 月	杭州市人民政府办公厅	《杭州市人民政府办公厅关于加快推进人工智能产业创新发展的实施意见》	到 2025 年，基本形成“高算力+强算法+大数据”的产业生态，将我市打造成为全国算力成本洼地、模型输出源地、数据共享高地，人工智能创新应用水平全国领先、国际先进。算力设施先进泛在，算力供给普惠高效，全市可开放算力规模在使用半精度输出输入(FP16)下达到 5000 千万亿次浮点指令/秒(PFLOPS)以上，高性能算力占比达到 60%以上。模型创新应用领跑全国，培育性能达到国际先进水平的通用大模型 1 个、具有行业重大影响力的专用模型 10 个。支持头部企业开展多模态通用大模型关键技术攻关、中小企业深耕垂直领域做精专用模型，鼓励相关技术和算法开源开放，形成“1+N+X”的协同创新、双向赋能产业生态。
2023 年 8 月	成都市经济和	《成都市加快大模	支持企业与科研机构开展数据与知识深度联合学习、大规模

	信息化局、成都市新经济发展委员会	型创新应用推进人工智能产业高质量发展的若干措施》	认知与推理、可控内容生成等关键算法研发，着力推动大模型相关技术创新，重点研发和迭代 CV 大模型、NLP 大模型、多模态大模型等领域大模型，以及医疗、金融、商务、交通等行业大模型。
2023 年 9 月	福建省人民政府办公厅	《福建省人民政府办公厅关于印发福建省促进人工智能产业发展十条措施的通知》	以普惠算力降低人工智能企业研发成本，支撑快速增长的算力需求，促进自然语言，多模态认知等超大规模智能模型开发训练。
2023 年 10 月	安徽省人民政府	《安徽省人民政府关于印发打造通用人工智能产业创新和应用高地若干政策的通知》	对在皖落户的通用及行业大模型企业、跨领域应用企业、新兴算力企业、安全人工智能企业等，优先匹配算力、数据、场景、基金、场地等要素资源。
2023 年 10 月	上海市经济和信息化委员会、上海市发展和改革委员会等五部门	《上海市推动人工智能大模型创新发展若干措施（2023-2025 年）》	实施大模型创新扶持计划。支持引进高水平创新企业，支持本市创新主体打造具有国际竞争力的大模型，鼓励形成数据飞轮，加速模型迭代，对取得重大成果的予以专项奖励。实施大模型示范应用推进计划。重点支持在智能制造、生物医药、集成电路、智能化教育教学、科技金融、设计创意、自动驾驶、机器人、数字政府等领域构建示范应用场景，打造标杆性大模型产品和服务。
2023 年 11 月	广东省人民政府	《广东省人民政府关于加快建设通用人工智能产业创新引领地的实施意见》	围绕基础架构、训练算法、调优对齐、推理部署等环节，研发千亿级参数的人工智能通用大模型，形成自主可控的大模型完整技术体系。聚焦智能经济、智能社会等行业创新场景，研发具有多模态数据、知识深度融合的垂直领域大模型，支撑多任务复杂场景行业应用。

制表：报告组根据公开信息整理

1.2 AI 大模型产业发展技术驱动力

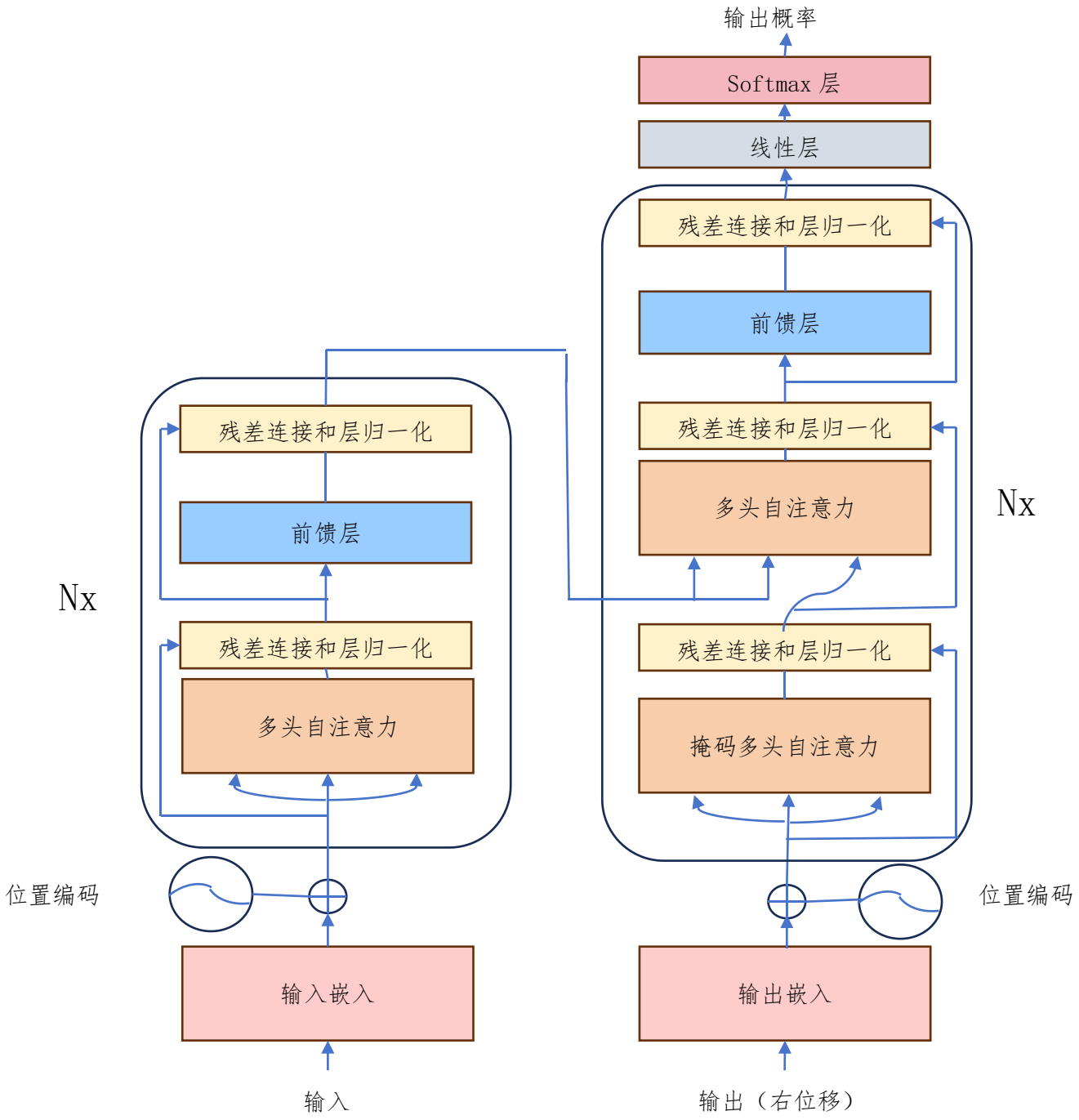
近年来，AI 大模型得到快速发展，当前大模型热潮主要由语言大模型相关技术引领。语言大模型通过在海量无标注数据上进行大规模预训练，让模型学习大量知识并进行指令微调，从而获得面向多任务的通用求解能力。2017 年，Google 提出基于自注意力机制的神经网络结构——Transformer 架构，奠定了大模型预训练算法架构的基础。2018 年，OpenAI 和 Google 分别发布了 GPT-1 与 BERT 大模型，预训练大模型成为自然语言处理领域的主流。2022 年，OpenAI 推出 ChatGPT，其拥有强大的自然语言交互与生成能力。2023 年，OpenAI 多模态预训练大模型 GPT-4 发布，其具备多模态理解与多类型内容生成能力。2024 年，OpenAI 发布视频生成大模型 Sora，提出时空碎片和扩散 Transformer 技术，大模型的多模态生成能力的进一步成熟。本部分将从经典 Transformer 架构出发，通过全面梳理基于人类反馈强化学习、指令微调、提示学习等相关大模型技术，体现技术对于产业发展的带动作用。

1.2.1 Transformer 架构

Transformer 架构是目前语言大模型采用的主流架构，于 2017 年由 Google 提出，其主要思想是通过自注意力机制获取输入序列的全局信息，并将这些信息通过网络层进行传递，Transformer 架构的优势在于特征提取能力和并行计算效率。

Transformer 架构主要由输入部分、多层编码器、多层解码器以及输出部分组成。其中，输入部分包括源文本嵌入层、位置编码器；编码器部分由 N 个编码器层堆叠而成；解码器部分由 N 个解码器层堆叠而成；输出部分包括线性层和 Softmax 层。

Transformer 架构图



制图：报告组根据公开信息整理

自注意力机制作为 Transformer 模型的核心组件，其允许模型在处理序列数据时，对每个词位置的输入进行加权求和，得到一个全局的上下文表示。在计算自注意力时，模型首先将输入序列进行线性变换，得到 Q（查询）、K（键）和 V（值）三个向量。然后，通过计算 Q 和 K 的点积，并应用 Softmax 函数，得到每

个位置的权重。最后，将权重与 V 向量相乘，得到自注意力的输出。为提高模型的表达能力，Transformer 模型采用了多头自注意力机制，这意味着模型在同一时间关注来自不同表示子空间的注意力信息。多头自注意力的实现方法是将输入序列分成多个组，每个组使用一个独立的权重矩阵进行线性变换，并计算自注意力。最终，自注意力的输出被拼接起来，并通过一个线性层得到最终的输出表示。在计算自注意力和多头自注意力之后，Transformer 模型使用前馈神经网络对输入序列进行变换。前馈神经网络由多个全连接层组成，每个全连接层都使用 ReLU 激活函数。前馈神经网络的作用是对输入序列进行非线性变换，以捕捉更复杂的特征。

1.2.2 AI 语言大模型关键技术

AI 语言大模型关键技术主要涉及基于人类反馈强化学习、指令微调、模型提示等相关技术。

（1） 基于人类反馈强化学习

基于人类反馈强化学习(Reinforcement Learning from Human Feedback, RLHF)，是指将人类标注者引入到大模型的学习过程中，训练与人类偏好对齐的奖励模型，进而有效指导语言大模型的训练，使得模型能够更好地遵循用户意图，生成符合用户偏好的内容。

基于人类反馈强化学习具体包括以下几个步骤：

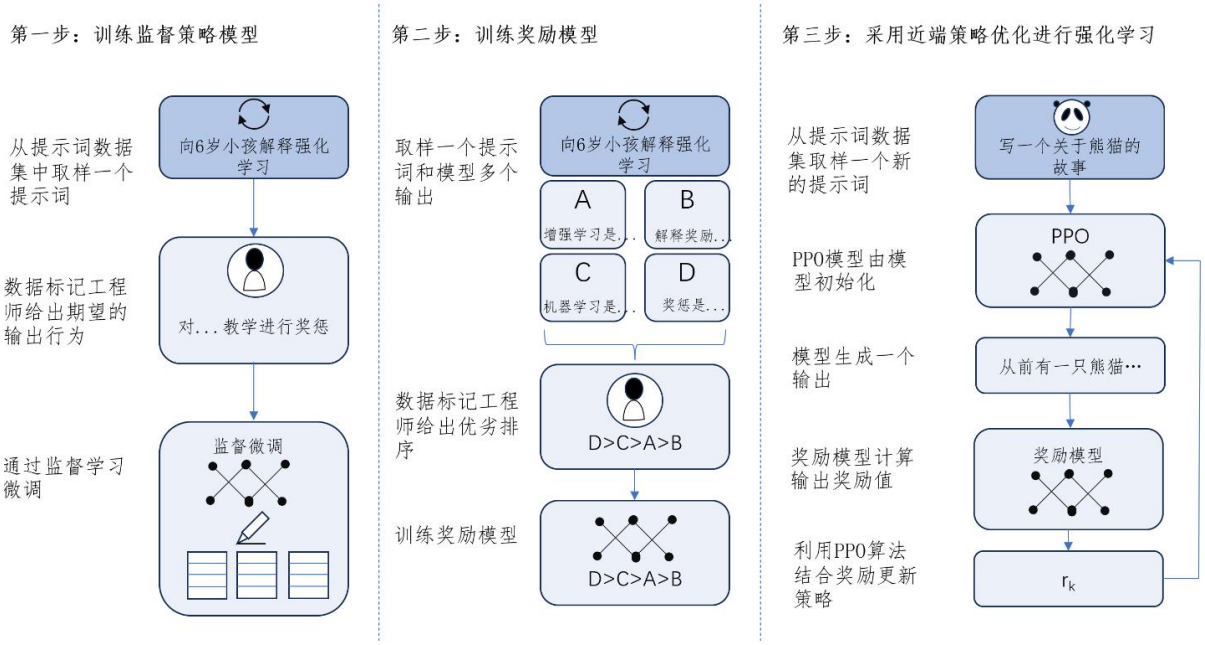
1）训练监督策略模型：使用监督学习或无监督学习的方法，对一个预训练的语言模型进行训练，通过给予特定奖励或惩罚引导 AI 模型的行为，使其能够根据给定的输入预测输出或行为。

2）训练奖励模型：让标记员参与提供有关模型输出结果的反馈，对模型生成的多个输出或行为的质量或正确性进行排名或评分，这些反馈被转换为奖励信号，用于后续的强化学习过程。

3）采用近端策略优化进行强化学习：先通过监督学习策略生成近端策略优化（PPO）模型，经过奖励机制反馈最优结果后，再将结果用于优化和迭代 PPO

模型参数。具体而言，在 PPO 模型训练过程中，智能系统通过尝试不同的行为，并根据每个行为获得的奖励来评估其质量，智能系统逐步改进行为策略。

基于人类反馈强化学习示意图



制图：报告组根据公开信息整理

（2） 指令微调

指令微调（Instruction Tuning），是一种帮助语言大模型实现人类语言指令遵循的能力，在零样本设置中泛化到未知任务上的学习方法。指令微调是让语言大模型理解人类指令并按照指令要求完成任务，即在给定指令提示的情况下给出特定的回应。指令微调可被视为有监督微调（Supervised Fine-Tuning，SFT）的一种特殊形式，但两者目标有所差别。SFT 是一种使用标记数据对预训练模型进行微调的过程，以便模型能够更好地执行特定任务，而指令微调是一种通过在（指令，输出）对的数据集上进一步训练大型语言模型（LLMs）的过程，以增强 LLMs 的能力和可控性。指令微调的特殊之处在于其数据集的结构，即由人类指令和期望的输出组成的配对，这种结构使得指令微调专注于让模型理解和遵循人类指令。

（3） 模型提示

通过大规模文本数据预训练之后的语言大模型具备作为通用任务求解器的潜在能力，这些能力在执行特定任务时可能不会显式地展示出来，在大模型输入中设计合适的语言指令提示有助于激发这些能力，称为模型提示技术。典型的模型提示技术包括指令提示和思维链提示。

指令提示（Instruction Prompt）。OpenAI 在 GPT-3 中首次提出上下文提示，并发现 GPT-3 在特定领域少样本提示下能够达到人类水平，证明在低资源场景下非常有效。指令提示的核心思想是避免强制语言大模型适应下游任务，而通过提供“提示（Prompt）”来给数据嵌入额外的上下文以重新组织下游任务，使之看起来更像是在语言大模型预训练过程中解决的问题。

思维链提示（Chain of Thought，CoT）。推理的过程通常涉及多个推论步骤，通过多步推理允许产生可验证的输出，可以提高黑盒模型的可解释性。思维链是一种提示技术，已被广泛用于激发语言大模型的多步推理能力，被鼓励语言大模型生成解决问题的中间推理链，类似于人类使用深思熟虑的过程来执行复杂的任务。在思维链提示中，中间自然语言推理步骤的例子取代少样本提示中的〈输入，输出〉对，形成〈输入，思维链，输出〉三元组结构。思维链被认为是语言大模型的“涌现能力”，通常只有模型参数规模增大到一定程度后才采用思维链能力。激活语言大模型的思维链能力方法，在提示中给出逐步的推理演示作为推理的条件，每个演示都包含一个问题和一个通向最终答案的推理链。

1.3 中国 AI 大模型产业发展市场驱动力

中国 AI 大模型产业发展源于多领域的广泛需求，例如来自办公、制造、金融、医疗、政务等场景中降本增效、生产自动化、降低风险、提高诊断准确率、提高政务服务效率等诉求。相关领域的创新和发展共同推动着中国 AI 大模型产业的蓬勃发展，预示着未来更广阔的市场前景。

1.3.1 办公场景

近年来，随着文字、语音、图像等处理能力跃迁，大模型摇身变为“助理”走入办公室和会议室，结合传统软件使得办公和会议智能化。

基于大模型的智能办公产品满足日常办公场景中文案生成、PPT 美化、数据分析等各种需求。仅通过自然语言交互，用户便可将繁琐的文字、演示、数据处理工作交给“助理”，用节约的时间做更有创意的事情。智能文档负责协助用户构建文章大纲、一键生成模板、生成内容、优化表达、处理和理解文档；智能演示承担自动排版美化、生成演讲备注、一键生成幻灯片等任务；智能表格通过对话即可生成公式、数据批量处理、自动生成表格。

智能会议方面，大模型可从会议策划、同声传译、会议记录等各环节赋能。会议策划场景大模型根据会议主题等提示词，自动生成会议环节、会议分论坛、会议时间、会议预算等完整策划内容；在大模型能力加持下，同声传译的准确性、及时性和多语言能力得到显著提升；通过大模型处理后，结构清晰、要点明确的会议记录结果使得会后回顾更加高效。

1.3.2 制造场景

人工智能崛起引领制造行业的深刻变革，改变研发设计、生产制造、供应链管理等流程。大模型+EDA/CAE/CAD，将传统研发设计软件效率进一步提升。大模型助力数字孪生和机器人，获得强大的感知场景和执行任务能力。大模型融合供应链管理，实现工厂管理的智能化转型。

在研发设计阶段，以大模型+EDA 为例，利用云端扩展性实现设计自动化，并确保设计在电气方面准确无误，同时简化系统设计流程，缩短 PCB 设计周转时间。企业借此缩短研发周期、降低研发成本、提升行业竞争力；生产制造中，利用 AIGC 和数字孪生技术，可模拟真实生产环境派出虚拟人代替工人进行危险、故障排查，或是通过仿真设备操作场景，完成沉浸式作业教学。拥有大模型功能的机器人凭借机器视觉技术，可执行路径规划、物体识别等任务；大模型集成于供应链管理系统中，能重构数字化办公流程，通过自然语言指令实现人机交互，推动企业进行更高效的管理决策、更便捷的数据分析与可视化，在需求端及时预测需求达到降本增效的目的，在仓库和物流端实现智能调度、智能跟踪和智能预警。

1.3.3 金融场景

金融行业存在前、中和后台的业务划分，在数字经济时代的浪潮中，相关业务已被大模型全局赋能提升效率。以银行为例，对话机器人、虚拟助理已经逐渐出现在个性化服务、电子营销、金融欺诈检测、信贷支持等服务场景中。

个性化服务方面，银行大模型以客户数据为依据，为客户提供定制的财务和产品计划；电子营销方面，大模型根据客户行为偏好生成个性化电子邮件；金融欺诈检测方面，大模型赋能专业人员检索大量数据识别欺诈行为；信贷支持方面，大模型通过分析海量生产生活和信用数据，为信贷部门人员生成高质量的信贷方案建议，减少银行贷款收益损失。

1.3.4 医疗场景

得益于近年来医疗大模型的不断迭代，复杂的医疗数据分析任务得以解决。由于患者行为数据的独特性，大模型通过个性化设计，满足患者“千人千面”的医疗服务需求，应用于智慧影像、智慧手术、智慧健康等领域。

智慧影像覆盖 CT、MR、DR、US、DSA、钼靶等医疗影像场景，为患者进行早期检测、诊断及健康风险评估；智慧手术功能大幅提高患者病情评价准确度，打牢术前风险评估、术中手术规划、术后预后估计的基础；智慧健康则作为一般患

者的贴身健康助手，通过小程序等便捷方式为患者提供高质量导诊服务和个性化健康建议。

1.3.5 政务场景

在办公、制造、金融、医疗场景得到助力的同时，政务场景下的效率、信息参考范围、经验共享、规范性等常见痛点也获得大模型能力加持得以解决。

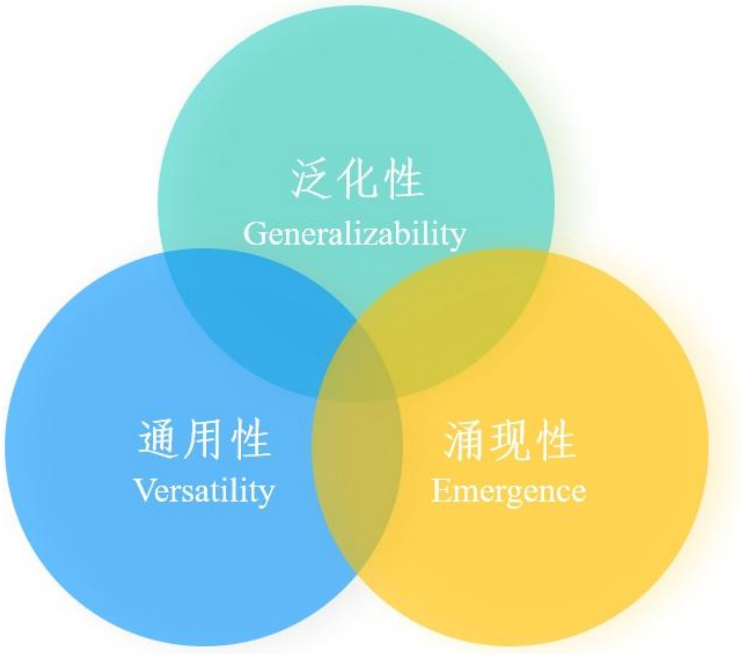
为提升效率，大模型利用自动化的政策检索、政策比对解决海量政策参考、人工分析比对的耗时问题；为缩小信息参考范围，政策撰写助手结合政策数据权威白名单，并接入政策全量库，避免不可靠信息来源引发舆论风险；为提高政策管理经验共享，大模型引入政策经验知识库，提升政务业务理解和政策管理能力；为规范政策撰写，政务大模型凭借规范化生成、检查功能维护成果的规范性、权威性。

第二章 百舸争流：中国 AI 大模型产业现状及典型案例

2.1 AI 大模型主要特征

AI 大模型具有泛化性(知识迁移到新领域)、通用性(不局限于特定领域)以及涌现性(产生预料之外的新能力)特征。以 ChatGPT 为代表的 AI 大模型因其具有巨量参数和深度网络结构,能学习并理解更多的特征和模式,从而在处理复杂任务时展现强大的自然语言理解、意图识别、推理、内容生成等能力,同时具有通用问题求解能力,被视作通往通用人工智能的重要路径。

AI 大模型的三大特征：泛化性、通用性、涌现性



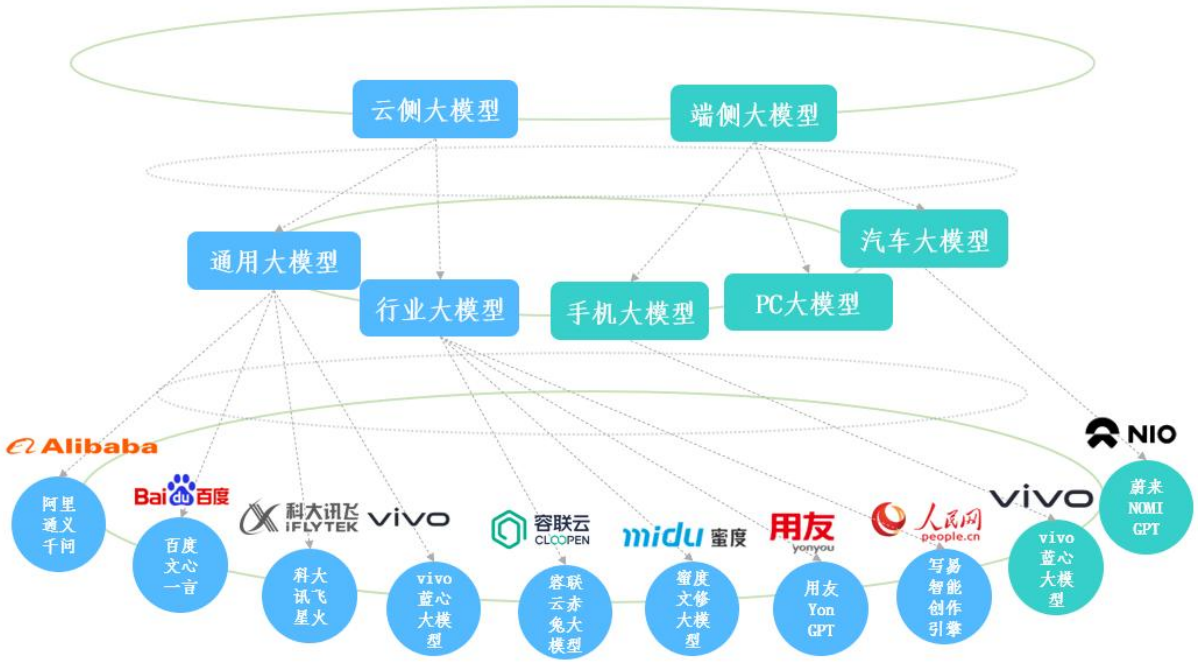
制图：报告组绘制

2.2 AI 大模型主要类型

按照部署方式划分，AI 大模型主要分为云侧大模型和端侧大模型两类。云侧大模型由于部署在云端，其拥有更大的参数规模、更多的算力资源以及海量的数据存储需求等特点；端侧大模型通常部署在手机、PC 等终端上，具有参数规模小、本地化运行、隐私保护强等特点。

具体而言，**云侧大模型分为通用大模型和行业大模型；端侧大模型主要有手机大模型、PC 大模型**。从云侧大模型来看，通用大模型具有适用性广泛的特征，其训练数据涵盖多个领域，能够处理各种类型的任务，普适性较强。行业大模型具有专业性强的特点，针对特定行业（如金融、医疗、政务等）的需求进行模型训练，因而对特定领域具有更深的业务理解和场景应用能力。从端侧大模型来看，手机和 PC 大模型由于直接部署在设备终端，让用户体验到更加个性化和便捷的智能体验。

AI 大模型主要分为云侧大模型和端侧大模型两类



制图：报告组根据公开信息整理

当前，我国 AI 大模型产业呈现蓬勃发展的态势。伴随多家科技厂商推出的 AI 大模型落地商用，各类通用、行业以及端侧大模型已在多个领域取得了显著的成果，如在金融、医疗、政务等领域，AI 大模型已成为提升服务质量和效率的重要手段。我国具有代表性的通用 AI 大模型主要包含科大讯飞的讯飞星火认知大模型、百度公司的文心一言大模型、阿里巴巴的通义千问大模型等；行业 AI 大模型主要涵盖蜜度的文修大模型、容联云的赤兔大模型、用友的 YonGPT 大模型；同时具有云侧和端侧大模型的端云结合 AI 大模型主要有 vivo 的蓝心大模型；端侧 AI 大模型主要以蔚来的 NOMI GPT 大模型为代表。

中国 AI 大模型分类及典型案例

类别	AI 大模型功能	AI 大模型案例
AI 通用大模型	文本生成、语言理解、知识问答、逻辑推理、数学能力、代码能力、多模态能力等	科大讯飞—讯飞星火认知大模型 百度公司—文心一言大模型 阿里巴巴—通义千问大模型
AI 行业大模型	1. 金融：文档处理、知识问答、内容生成、辅助决策 2. 医疗：医学影像生成、知识问答、辅助决策 3. 政务：政策检索、知识问答、辅助决策 4. 电商：经营分析、商品推广、商品销售 5. 传媒：录音转写、新闻写作、视频剪辑	蜜度—文修大模型 容联云—赤兔大模型 用友—YonGPT 大模型 人民网—“写易”智能创作引擎
AI 端侧大模型	物体识别、语言理解	蔚来—NOMI GPT
AI 端云结合大模型	语义搜索、知识问答、文本创作、图片生成、智慧交互等	vivo—蓝心大模型

制图：报告组根据公开信息整理

2.3 中国 AI 通用大模型典型案例

案例一：科大讯飞—讯飞星火认知大模型

（1）大模型简介：

讯飞星火认知大模型是科大讯飞推出的新一代认知大模型，可实现基于自然对话方式的用户需求理解与任务执行。讯飞星火从赋能万物互联时代的人机交互、赋能知识学习与内容创作、提升数智化生产力三个方面展现其应用能力。**讯飞星火认知大模型具备七大核心能力：文本生成、语言理解、知识问答、逻辑推理、数学能力、代码能力和多模态能力。**

（2）大模型优势：

2024 年 1 月，讯飞星火认知大模型 V3.5 发布，七大核心能力全面提升。据科大讯飞宣称，数学、语言理解超 GPT-4 Turbo，在代码能力方面已经达到 GPT-4

Turbo 的 96%。另外，星火大模型在多模态理解方面达到 GPT-4V 的 91%，其中语音的多模态能力已经超过 GPT-4。

讯飞星火认知大模型 V3.5 七大能力

图片来源：讯飞星火认知大模型 V3.5 升级发布会

讯飞星火认知大模型 V3.5 从三个角度展示了模型能力的提升，赋能万物互联时代人机交互、赋能知识学习与内容创作、提升数智化生产力。**人机交互方面**，讯飞星火 V3.5 在语义理解、指令跟随和多轮对话中展现优异能力，在情绪感知和拟人合成方面也有出色表现；**知识学习与内容创作方面**，要素抽取、问题生成等底层能力的进步，能够帮助知识学习和内容创作领域，产生更丰富更有用的智能体。讯飞星火大模型能够结合外部知识进行合理拓展，做到“旁征博引”；**数智化生产力方面**，逻辑推理能力和时空推理能力并重，数学则是大模型的基础能力，而代码能力用于生成各种工具链接虚拟和现实世界，最后多模态能力也是机器人、工业、家庭等场景中必备的能力。讯飞星火 V3.5 在这些关键技术领域取得显著进步。

（3）大模型应用：

讯飞星火七大能力的提升，实现了各类应用场景性能升级。**语言理解方面**，情感分析可以提取文本情感色彩更好了解内容观点和态度。文本摘要总结简洁准

确的摘要，快速理解文章的核心观点；**文本生成方面**，科大讯飞推出可以一键快速自动生成文档和 PPT 的办公产品——讯飞智文，主要功能有文档一键生成、AI 撰写助手、多语种文档生成、AI 自动配图、多种模板选择、演讲备注等；**知识问答方面**，讯飞星火对生活常识问答、医学知识问答、政策问答等任务“信手拈来”；**逻辑推理方面**，思维推理可以通过分析问题的前提条件和假设来推理出答案或解决方案，给出新的想法和见解。科学推理则使用已有的数据和信息进行推断、预测和验证等科学研究中的基本任务；**数学能力方面**，讯飞星火可以解决方程求解、立体几何、微积分、概率统计等数学问题；**代码能力方面**，讯飞星火能根据注释、函数名智能生成代码，支持逐行代码注释，还可以精准定位代码语法、逻辑错误，甚至可以智能生成单元测试数据；**多模态能力方面**，讯飞星火可根据用户上传图片返回准确的图片描述，或完成针对图片素材的问答，还可以凭借用户描述，生成期望的音频和视频。

案例二：百度公司—文心一言大模型

（1）大模型简介：

文心一言是百度研发的人工智能大语言模型产品，具备跨模态、跨语言的深度语义理解与生成能力，在文学创作、文案创作、搜索问答、多模态生成、数理逻辑推算等众多领域都能为用户提供高质量服务。**文心一言拥有四大基础能力：理解能力、生成能力、逻辑能力、记忆能力。**

（2）大模型优势：

2023 年 10 月发布的“文心大模型 4.0”，相比上一代文心大模型，四大能力显著升级，其中逻辑提升幅度是理解的 3 倍，记忆提升幅度是理解的 2 倍。**理解能力**方面，文心一言能听懂潜台词、复杂句式、专业术语、前后乱序、模糊意图等复杂提示词，也能胜任代码理解与调试任务；**生成能力**方面，文心一言能快速生成风格多样的文本、代码、图片、图表、视频，比如进行文案创作、制定生活计划、编写高质量代码；**逻辑能力**方面，文心一言能帮用户解决复杂的逻辑难题、困难的数学计算、重要的职业/生活决策、代码纠错、常识推理、逻辑校验、

立体几何、辩论灵感等；**记忆能力**方面，经过多轮对话后，文心一言依然能记住对话的重点，轻松胜任复杂问题、沉浸体验角色对话。

文心大模型 4.0 的能力提升源自相关举措：（1）在万卡算力上基于飞桨平台，通过集群基础设施和调度系统、飞桨框架的软硬协同优化，支持了大模型的稳定高效训练。（2）通过建设多维数据体系，形成了从数据挖掘、分析、合成、标注到评估闭环，充分提高数据的利用效率，大幅提升模型效果。（3）基于有监督精调、偏好学习、强化学习等技术进行多阶段对齐，保证了模型能够更好地与人类的判断和选择对齐。（4）利用可再生训练技术通过增量式的参数调优，有效节省了训练资源和时间，加快了模型迭代速度。

文心大模型 4.0 典型特征

图片来源：百度世界大会

（3）大模型应用：

文心大模型在文学创作、文案创作、搜索问答、多模态生成、数理逻辑推算等方面已有应用面向用户开放。**文学创作方面**，文心一言可以清晰地表达观点、传递情感，因此可以应用于小说、散文、诗歌等文学作品的创作中；**文案创作方面**，在商业领域，文心一言可以撰写商业计划、市场分析报告等商业文案，提供有力的文字支持。文心一言可以激发创意思维，为广告行业提供新的灵感和想法，可以帮助广告人员快速构思出吸引人的广告文案和宣传语；**搜索问答方面**，基于文心一言的聊天机器人可以与用户进行自然语言交互，理解用户的意图和需求，并提供相应的回答和建议。这种应用可以广泛应用于生活服务、教育辅导、客服

等领域；**多模态生成方面**，文心大模型支持图像生成和处理，可以根据用户需求生成图像或者对已有图像进行处理编辑。文心大模型还支持语音合成、语音识别和音频分类。文心大模型还能对视频数据进行处理，或将文本转化为动态图像序列完成视频分类、目标检测等任务；**数理逻辑推算方面**，文心大模型可以解决复杂的数学问题，也可以成为代码编写助手，比如百度基于文心大模型研制了智能代码助手 Comate，提供智能推荐、智能生成、智能问答等多种功能，支持多种编程语言和 IDE。

案例三：阿里巴巴—通义千问大模型

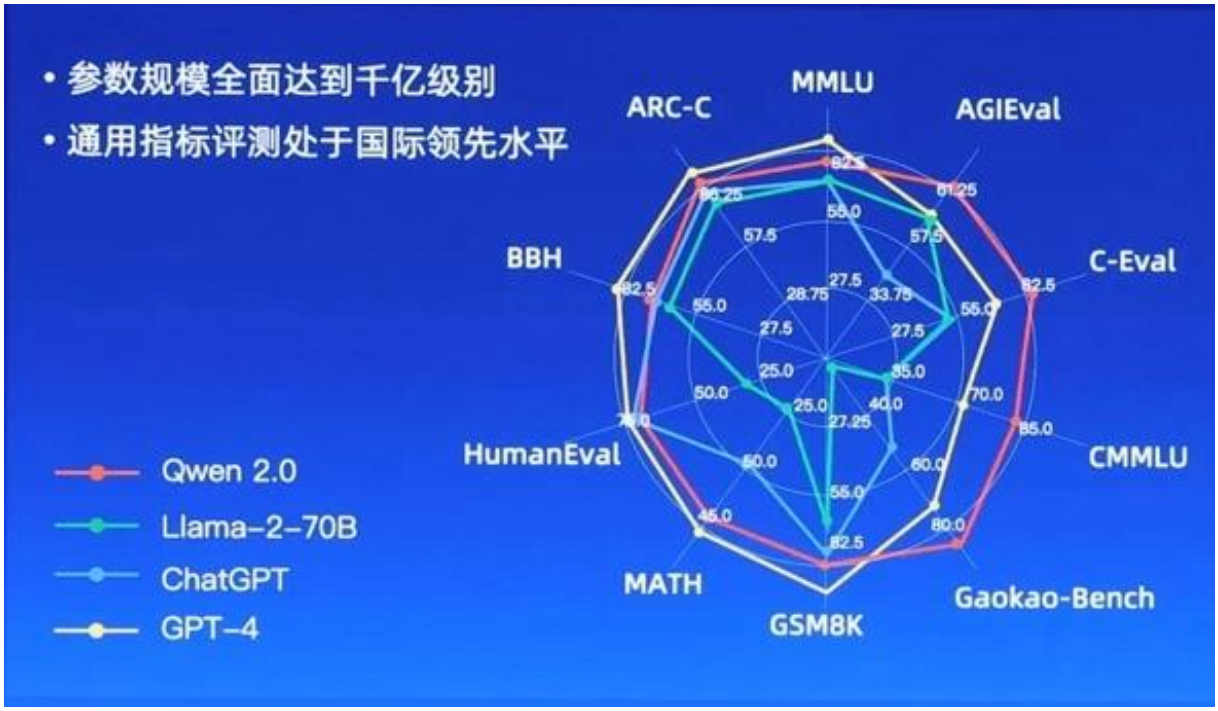
（1）大模型简介：

通义千问是阿里云研发的预训练语言模型，基于先进的自然语言处理技术（NLP），执行理解、生成和解释人类语言、图片和文档等任务。通义千问能在创意文案、办公助理、学习助手、趣味生活等方面为使用者提供丰富的交互体验。**通义千问具备中英文理解、数学推理、代码理解**等能力。

（2）大模型优势：

2023 年 10 月，**千亿级参数大模型通义千问 2.0 发布**，相比 1.0 版本，其在复杂指令理解、文学创作、通用数学、知识记忆、幻觉抵御等能力上均有显著提升。**中英文理解能力**是大语言模型理解和表达的基础能力，英语任务中，通义千问 2.0 的 MMLU（伯克利大学、哥伦比亚大学等联合发布）基准得分是 82.5。中文任务中，通义千问 2.0 在模型训练中学习了更多中文语料，在 C-EVAL（上海交大和清华联合研发的中文大语言模型测试集）基准上获得最高分；**数学推理**方面，在推理基准测试 GSM8K（OpenAI 发布的小学数学测试集）中，通义千问排名第二，展示了强大的计算和逻辑推理能力；**代码理解**方面，HumanEval（OpenAI 发布）测试衡量大模型理解和执行代码片段的能力，通义千问排名第三，这一能力是大模型在编程辅助、自动代码修复等场景的基础。**（注：得分排名信息均为各榜单 2023 年 10 月数据）**

通义千问 2.0 参数及指标评测



图片来源：阿里云公众号

（3）大模型应用：

通义千问目前主要应用于四个方向：创意文案、办公助理、学习助手、趣味生活。**创意文案应用**包括：“撰写营销文案”，输入产品介绍获得量身定制的金牌营销文案。“文章润色”能对用户提交的文章进行深度分析，挖掘其中表达不足之处，提供词汇句式变化建议。“直播带货剧本生成”，基于丰富的商品信息和用户需求，为电商主播提供生动有趣且具有营销力的脚本内容；**办公助理应用**于：“SWOT 分析”为用户提供全面、深入且精准的战略决策支持，从多元视角理解并评估内外部环境对特定项目的影响。“PPT 框架生成”，智能地为用户构建专业且逻辑清晰的 PPT 结构；**学习助手应用**包含了：“题目加工厂”，根据提供的专业以及学科领域进行高质量试题生成，大大节省了教师、家长以及教育机构在出题上的时间和精力。“学习计划站”可为用户提供个性化、系统化的学习路径规划，定制高效且科学的学习日程安排；**趣味生活应用**有：“会放飞的菜谱”，输入菜名逐步指导提供美食烹饪秘诀。“AI 健身教练”为用户制定专属健身计划。“写歌词”，根据用户提示的歌名写出生动歌词。

2.4 中国 AI 行业大模型典型案例

案例四：容联云—赤兔大模型

（1）大模型简介：

赤兔大模型是容联云开发的面向企业应用的垂直行业多层次大语言模型，赋能企业搭建专属智能客服和数智化营销，完成从“降本增效”到“价值创造”的进化。丰富的智能应用为赤兔大模型能力保驾护航，包含会话洞察、业务话术、问答知识库、知识运用、数据分析、智能对话框架、流程管理。

（2）大模型优势：

赤兔大模型三个核心点分别是智能性、可控性和投产比。**智能性**方面是客户最关心的，首先能力是否足够丰富，能否解决以前不能解决的问题以及相关能力到底能做多好。**智能性方面，**包括检索增强、会话分析、逻辑推理、数据分析。**检索增强**是指在海量文档中快速定位到信息，经过整理给客户提供答案。**会话分析**能让模型在对话中发现多维度信息，包括情绪、立场、各种细节的意见，并且根据不同业务快速切换业务场景。**逻辑推理**体现在推荐话术的原因，投诉、预警的原因这种因果分析能力。**数据分析**体现在对数据更细致且自动化的分析，减少技术人员和业务人员的数据层面上的操作；**可控性方面**，赤兔大模型在道德、伦理、安全、风格、偏好上对齐，对话或话术生成时满足基本安全需求。另外让模型知道应该处理的知识范畴、知识边界，从而避免自由对话潜在的安全风险和资源消耗；**投产比方面**，大模型强大能力来源于大规模，而大规模需要大投入，合理的投产比是客户采取方案的底层逻辑。所以明确是否所有场景都需要大模型，AI 底座上，没有摒弃小模型，大小模型相配合完成对上层能力的输出。机制上合理调动分配，比如有的环节大模型靠后完成线下或离线的工作，有的环节大模型调动指挥小模型完成。

（3）大模型应用：

基于赤兔大模型，容联云发布了生成式应用“**容犀 Copilot**”。**容犀 Copilot 具备三大核心能力：大模型话术、智能知识库、会话洞察。大模型话术：**容犀 Copilot 后台一键快速对海量历史会话数据进行核对筛选，挑选出更佳话术并生成金牌话术，兼顾质与量的同时，挖掘出客户高频关注的问题，从问题中洞悉业务痛点；**智能知识库**：可以帮助企业从零开始、低成本地快速构建话术库，包括理解文档知识、知识快搜、智能问答等，大幅提升构建效率；**会话洞察**：高效便捷洞察每一通会话沟通情况，分析客户诉求，精准诊断问题并优化。回归实际业务本身，容犀 Copilot 深入金融行业细分场景，打造场景化客服助手，譬如分期挽留助手、荐卡挽留助手、投诉安抚助手等，实时辅助快速洞察客户需求，推荐更佳应答话术，诊断客户情绪变化，提醒措辞及注意事项。

容犀 Copilot 产品应用场景

图片来源：容联云公众号

案例五：蜜度—文修大模型

（1）大模型简介：

文修大模型是蜜度推出的一款聚焦于智能校对领域的大语言模型，基于蜜度在校对领域的知识和经验积累，为政务单位、新闻媒体、企业单位、学校机构、出版机构等专业用户提供更贴合使用场景的校对服务。**文修大模型具备校对能力**

强、速度快、匹配度高三大特点，更好地解决垂直行业的问题。

（2）大模型优势：

数字化时代，内容创作与传播速度惊人，信息准确无误地传达给公众尤为重要，蜜度文修大模型通过优秀的校对能力、高效的处理速度和高度的匹配度应对变局。校对能力方面，文修大模型以拼写错误、语义错误、语法错误为基础，以内容差错、常识校对差错、内容风险识别三大类校对类型，27 类细分类型为校对标准，有效满足出版行业、新闻行业的“三审三校”的校对规范和实际业务需求，提供诸如广告法检测、常识校对等更为全面的校对服务；校对速度方面，文修大模型几秒钟就能校对完一篇千字文章，几分钟即可校对一本 10 万字书稿。其快速校对的背后是实实在在的“学习能力”，文修大模型能迅速将人们短时间内难以学习消化的内容，转变成自身的校对能力，完成快速输出；匹配度方面，蜜度服务政府部门、媒体单位十余年，数十款智能应用产品及解决方案，覆盖政务部门、出版单位多个办公环节及场景，拥有成熟的流程服务，在洞察用户需求和使用场景方面拥有深厚的经验。

（3）大模型应用：

文修大模型满足政务单位、新闻媒体、企业单位、学校机构、出版机构的多行业场景应用需求。**政务单位领域**，赋能各级政务部门校对流程智能化，提供文字材料的内容错敏校对、修改提示和文本润色等服务，全力保障内容的准确性及严谨性，支持内网环境下校对，满足更高保密需求；**新闻媒体领域**，文修大模型深入新闻媒体工作的各个环节，对多模态内容进行多类错敏校对，帮助快速定位错误并高亮显示，让内容更加规范严谨，有效维护官方账号的公信力；同时提供文本润色服务，提高出稿速度，保障新闻时效性；**企业单位领域**，全流程切入企业办公场景，从内容纠错、提升文本质量等多方位出发，优化宣发内容，提高文案吸引力，助力营销效果显著提升；**学校机构领域**，针对学校机构的宣传材料、新媒体稿件、科研报告、学术论文等内容进行全面校审，有效降低文字错误率，保障学术严谨性。通过 AI 润色功能助力文章、报告、材料的起草、优化工作，有助于进一步提升学校传播力、影响力；出版机构领域，提供专业、便捷、高效

的内容筛查及文字质量把关服务，协助各出版机构高效处理多语言文本，降低内容差错概率，保障内容的规范性、准确性。

案例六：用友—YonGPT 大模型

（1）大模型简介：

YonGPT 是用友基于数字和智能技术服务企业和公共组织数智化的企业服务大模型。YonGPT 在企业服务领域的应用主要集中在 4 个方向：**智能化的业务运营、自然化的人机交互、智慧化的知识生成、语义化的应用生成。**

YonGPT 企业服务大模型整体架构图

图片来源：用友官网

（2）大模型优势：

用友人工智能研发团队基于大规模的商业应用数据，结合企业应用场景和领域经验，标记了大量的企业服务语料数据，形成丰富的企业服务大模型训练素材，并将业务知识与领域经验融入企业服务大模型，确保了 YonGPT 的专业性、实用

性及领先性。同时 YonGPT 通过**上下文记忆、知识库表索引、Prompt 工程、Agent 执行、通用工具集**等扩充大模型的存储记忆、适配应用和调度执行能力，形成体系化的企业服务大模型。YonGPT 还优化了企业服务大模型的训练效率和成本，集成了丰富的开发工具和优化算法，通过自有的数据管理、大模型精调、大模型评估优化、大模型推理和插件服务等功能，为大模型的构建和服务提供稳定且有效的支撑。

（3）大模型应用：

用友企业服务大模型 YonGPT 围绕四个方向推进模型训练和产品效果优化，提供深入到客户业务前端的全价值链、全场景的泛在智能和群体智能应用。**在智能化业务运营方面：**YonGPT 通过强大的数据分析和预测能力，深入洞察企业运营、识别潜在的业务风险和机会，并提供智能化的解决方案，从而提高经营决策水平和业务运营效率；**在自然化人机交互方面：**YonGPT 通过强大的自然语言处理技术和理解能力，使能企业应用和服务与用户进行自然而流畅的对话交流，以“人”为本的方式实现不同应用的调用、连接、组装，更自然、高效地完成工作；**在智慧化知识生成方面：**YonGPT 通过从海量数据和信息中提取、整合知识，生成新的、有价值的知识内容，涵盖了行业解决方案、专业领域知识分享，助力企业和用户全面利用自身知识的储备和积累，促进知识的传播和应用；**在语义化应用生成方面：**YonGPT 通过对用户需求、企业业务和数据特征的理解，可以自动生成具有语义化能力的应用程序，全方位提升企业个性化应用服务的创建效率。

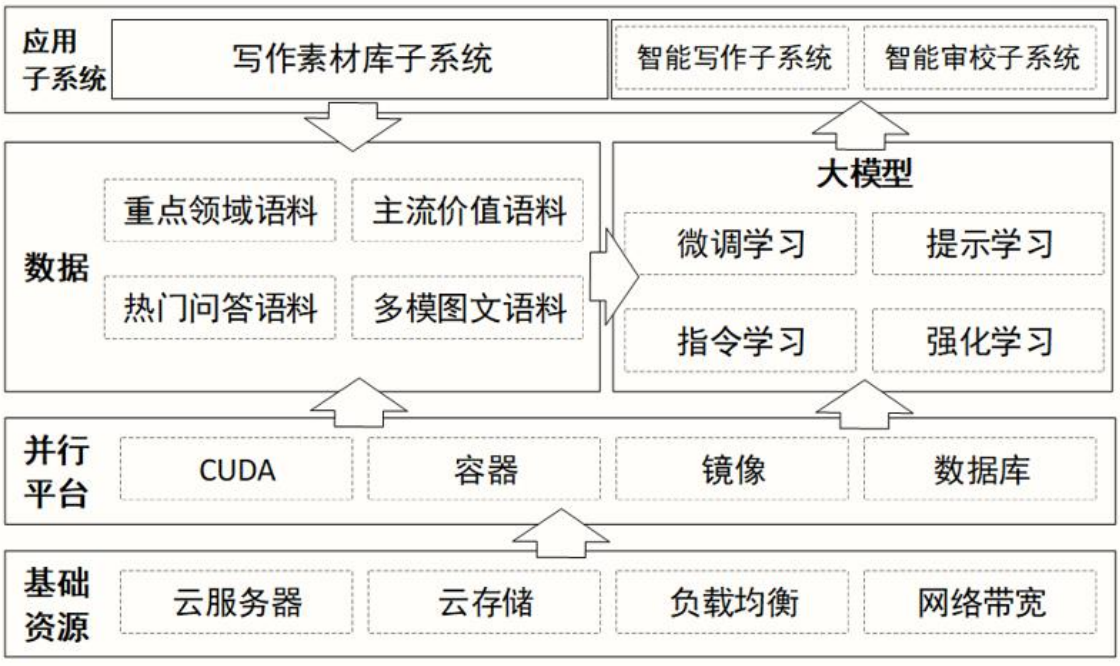
此外 YonGPT 在智能化场景服务中实现四个服务：**企业收入/利税经营智能分析**，可以实时掌控经营状况、快速洞察问题所在、精准预测企业效益、有效预见应对变化；**智能生单**，融合了丰富的供应链经验，通过“交互革新式”订单生成助手，实现快速智能生单，提高企业效能；**智能招聘**，帮助企业快速精准识别定位目标人才，从海量简历池中发现人才，通过 AI 互动优化应聘体验，实现选人、用人的精准决策；**智能大搜**，提供“沉浸式”搜索新体验，加速企业知识的价值化服务，并洞察用户需求、实现搜推一体，让知识赋能业务和组织。

案例七：“写易”智能创作引擎

（1）大模型简介：

“写易”智能创作引擎是人民网推出的垂类写作大模型，依托自主研发的主流价值观大模型针对写作场景训练而成。“写易”智能创作引擎适合党政党媒、央企国企、学校医院等有日常阅读需求和写作需求的群体，提供专业权威、系统整体、持续更新的数智服务，从而更好地启发创作者的写作灵感。

“写易”智能创作引擎产品架构



图片来源：人民网

（2）大模型优势：

“写易”智能创作引擎深入挖掘用户需求，构建了“随查”“随写”“随审”的交互体验，具有创作高效、安全准确、内容丰富的特点。其主要功能有：**高效的智能创作引擎**：“写易”智能创作引擎依托于超过 25 万条的权威主流语料库，创新性地实现了基于检索增强的辅助生成能力。可通过标题自动生成优质核心内容，同时结合标题与观点等上下文信息，为用户生成相关性更强、准确性更高的

文章素材，从而显著提升写作效率；**专业的涉政内容审校**：1.原文引用检测：检测文本是否引用了重要讲话、重要政策文件，分析引用的规范性，同时给出原文的出处等溯源信息。2.规范表述审校：对领导人重要讲话、党和国家重大政策重要文件等文本进行规范性检测。3.人物信息审校：对文本中出现的人物姓名及职务信息等进行勘误检测。4.关键信息审核：检测文本中有标志性、代表性的重大事件、重要人物、重要敏感信息等内容。5.基础纠错：对文本中的错别字检测、标点符号错误检测，包括中文、英文拼写、成对标点、特殊数字与符号等错误；**内容丰富的写作素材知识库**：依托人民网和人民日报，配备强大的资料库，系统化整理信息，按照时间、内容、形式、图文音频等多维度进行分类汇总。及时同步《人民日报》每日的精选文章，实时提炼更新当日最优标题、最美佳句、最优词语等内容，为用户提供新鲜的写作素材。

（3）大模型应用：

“写易”智能创作引擎，服务于以国资央企、党政机关、事业单位、党媒党网、地方融媒体、教育系统、医疗系统等为重点的全行业客户，满足客户在不同场景中的**定制化高效写作需求。**“写易”智能创作引擎可结合客户数据库进行定制，具体而言，人民网以行业客户数据库语料为核心，《人民日报》内容为辅助，为客户定制化训练“写易”智能创作语言大模型，帮助提高写作能力、积累写作素材、规范写作格式。在写作过程中还可以为标题撰写、修辞使用、引用诗文和网言网语等提供丰富素材，帮助创作者启发灵感、提供思路，从而自动、高效地生成符合客户写作场景的高质量文章素材，为客户提供全维、全时、全域的智能化服务，助力工作总效率大幅提升。

2.5 中国 AI 端云结合大模型典型案例

案例八：vivo—蓝心大模型

（1）大模型简介：

蓝心大模型是行业首个在手机端运行的开源自研大模型，也是更适合中文用

户的中文开源大模型，其包含十亿、百亿、千亿三个参数量级，共 5 款自研大模型（10 亿、70 亿、700 亿、1300 亿和 1750 亿）。随着参数提升，蓝心大模型逐渐具备文本总结、语言理解、文本创作、知识问答、角色扮演、复杂逻辑推理、复杂任务编排等能力。基于蓝心大模型能力，vivo 开发出蓝心小 V 和蓝心千询两款手机端产品。

vivo 蓝心大模型矩阵

图片来源：vivo 官网（注：榜单信息为 23 年 11 月数据）

（2）大模型优势：

1）矩阵化优势

vivo 大模型矩阵具有不同参数量级、多种部署方式，可应用于不同使用场景，在满足用户手机体验的同时，优化大模型推理性能以及端侧部署时占用的手机内存、功耗。10 亿参数的**蓝心大模型（1B）**，是面向端侧场景打造的专业文本大模型，具备本地化的文本总结、图片风格化能力，适用于需要快速本地化摘要、生图的场景；70 亿参数的**蓝心大模型（7B）**，是面向手机打造的端云两用模型，有良好的上下文关联能力和任务拆解能力，在语言理解、文本创作等场景下表现优秀。蓝心大模型 7B 也是手机行业首家开源的大模型，实现 AI 普惠；700

亿参数的**蓝心大模型（70B）**，是 vivo 面向云端服务的主力模型，在角色扮演、知识问答等场景下表现优异，既有智能涌现，也能兼顾成本和性能。面向复杂任务，vivo 也推出了两款千亿参数模型，1300 亿和 1750 亿大模型 ，凭借更丰富的知识量提供更加专业的智能体验。

2）端侧优势

蓝心大模型 1B 和蓝心大模型 7B 可在手机端运行，展现了出词快、内存低、全天候、真安全的强大端侧运行优势。出词快：vivo 对手机端上的 1B 模型优化，测试出字速度极限可达 60 字每秒，远超人眼约 10-20 字每秒的阅读需求；**内存低：**vivo 优化模型在手机端的内存占用问题，蓝心大模型 1B 和蓝心大模型 7B 分别只需占用 1.3G 和 3.8G 手机内存。**全天候：**将蓝心大模型置于终端设备上可以减少数据延迟，并能够满足例如出差飞机、高铁等无网弱网场景下的大模型使用需求，使得一些应用可以全天候正常运行。**真安全：**vivo 实现端侧内容安全过滤模型，优化输入语料和大模型生成内容的合规性问题，端侧大模型有助于在本地处理数据形成闭环，减少了敏感信息通过网络传输的风险，为大模型应用场景落地移动终端提供安全基石。

3）算法优势

蓝心大模型具备三大算法优势：**强大的基础能力、精准的指令跟随以及正确的价值取向。**首先，**强大的基础能力是大模型的根基。**在预训练阶段，vivo 利用最前沿的 Transformer 架构，改良注意力机制、位置编码等关键模块，**采用混合精度训练以及梯度缩放策略缩短训练周期。**其次，**精准的指令跟随是大模型与用户交互的核心。**vivo 对于微调**采用“target only loss”方法，并通过聚类分析对指令进行适应性处理，**以更好地确保模型的均衡性。第三，**正确的价值取向是大模型的灵魂。**在强化学习阶段， vivo 建立了 300 余名专业人员组成的审核团队，制定了 200 余项的安全审查机制，对模型的输出进行筛查和标注。同时，vivo **采用离线采样策略和双重奖励模型等策略，**在安全性上有明显提升。

（3）大模型应用：

1）蓝心小 V

蓝心小 V 是 OriginOS 4 上搭载的一款全局智能辅助功能，支持超能语义搜索、超能问答、超能写作、超能创图、超能智慧交互。**超能语义搜索：**使用自然语言即可搜索手机中的照片、文档、日程等信息；**超能问答：**根据用户上传的文档，快速提供文档总结，也可根据文档内容快速问答，还可以是知识百科开放式问答；**超能写作：**基于用户的要求，结合 AI 能力给用户生成文本，如润色、扩写、总结、格式文本等；**超能创图：**1、文生图和图生图：基于用户文字描述或上传图片，生成目标图片。2、AI 路人消除（路人隐身）：上传包含路人的图片，通过对话消除路人，生成更为纯净的图片；**超能智慧交互：**1、智能识屏服务功能：一键识别屏幕上的文本、网页链接，提取有效信息；2、超直觉化的交互方式：交互更多元，语音、文字、点击、拖拽、悬浮形式，互动更轻松。

2）蓝心千询

vivo 基于蓝心大模型打造的全天侯 AI 私人小帮手——**蓝心千询**，覆盖 AI 对话和 AI 灵感两大核心应用场景。蓝心千询是手机行业首个大模型公开版免费 APP。**AI 对话模块**，蓝心千询支持“超能问答”以及“超能创图”两大功能，用户可以通过 AI 对话进行文本问答、开放问答或是基于文档的问答，以高效获取信息、知识。另外，无论写诗、AI 作画、创作歌词、撰写标题/活动方案，蓝心千询皆可胜任。**AI 灵感模块**，蓝心千询能够提供社交媒体文案创作、PPT 大纲生成、中英文本互译等功能，还设置有穿搭建议等有趣的灵感工具。灵感广场设置有不同场景下的灵感技能卡片，覆盖工作、学习生活中的创作场景。蓝心千询将基于卡片语境快速生成对应文案，为用户的工作、学习、生活提供广泛灵感支持。

第三章 大浪淘沙：中国 AI 大模型产业发展所面临的挑战

3.1 大模型产业遭遇算力瓶颈

随着 AI 大模型规模呈现指数级增长，训练大模型越发依赖高性能 AI 芯片。AI 大模型的训练速度、产出质量，都和算力直接相关，对于 GPT 这种大语言模型（LLM）来说，算力的要求更高，也决定了模型的“智商”。目前主要以英伟达的 A100、H100 为代表的高性能 AI 芯片应用到主流 AI 大模型的训练过程。以 ChatGPT 为例，微软 Azure 云服务为其提供了 1 万枚英伟达 A100 GPU，这个算力也正是国内云计算技术人士共识的 AI 大模型门槛。然而国内拥有 1 万枚 GPU 的企业很少，而且单枚 GPU 普遍弱于英伟达 A100。由于英伟达 A100 及以上性能 GPU 被列入管制清单，目前中国企业能获取的替代品为英伟达 A800，然而 A800 也存在缺货和溢价的情况。从我国自研 AI 芯片来看，中国本土的高性能芯片龙头以华为海思、寒武纪、地平线、昆仑芯等为代表。我国正在高性能芯片领域加大投入并取得极大进展，部分解决方案正替代英伟达成为一些大厂的选择。但国产芯片性能目前仍与国际顶尖水平存在一定差距。

总体而言，国内 AI 高性能芯片市场受进口限制和国内技术瓶颈的双重影响，大模型产业发展受到算力层面的一些制约。

3.2 主流大模型架构仍存在诸多局限

当前，主流 AI 大模型所使用的 Transformer 架构存在消耗算力资源大、占用内存储量多等局限性。

首先，Transformer 架构消耗的算力资源普遍较大。传统 Transformer 架构由于算法特性，计算量会随着上下文长度的增加呈平方级上升。假如用户输入的上下文增加 32 倍，计算量可能会增加 1000 倍以上。

其次，基于 Transformer 架构的大模型对存储设备的要求也更高。在训练过程中需要在内存中存储参数的当前值、梯度以及其他优化器状态。模型的参数越

多，所需的计算就越多，需要的存储空间就越大。如 1000 亿个参数的 Transformer 模型，存储这些参数就需要 400GB 的空间。

3.3 高质量的训练数据集仍需扩展

国内的 AI 大模型数据主要来自互联网、电商、社交、搜索等渠道，存在数据类型不全面，信息可信度不高等问题。整体来看，我国可用于大模型训练的中文数据库体量严重不足。如悟道语料库，其包括文本、图文和对话数据集，最大的仅 5TB，其中开源的文本部分仅为 200GB。另外一个开源的中文本数据集 CLUECorps 为 100G。相比之下，GPT-3 的训练数据量，以英语为主，达到 45TB。此外，国内大模型的数据还缺乏多数据源的调用，可供大模型训练的有效数据源呈现碎片化分散状态，如微信公众号的文章仅在搜狗引擎支持调用，而多数大模型如智谱清言在联网收集数据时无法直接调用微信公众号文章。

当前，政府部门的权威数据、大型企业掌握的行业或内部数据通常不对外公开。以阿里巴巴的“通义千问”大模型为例，训练数据来自公开来源的混合数据，中文语料主要来自知乎、百度百科、百度知道等公开网络数据，来源于政府及企业数据较少。未来，仍需构建高质量的 AI 大模型训练数据集，不断扩充数据源提高数据质量。

3.4 大模型爆款应用尚未出现

自 ChatGPT iOS 版本发布近十个月以来，该应用一直在下载量、用户支出和会话时长方面牢牢占据生成式人工智能应用下载量榜单前三名的位置。GPT4.0 推出后，已支持语音输入和输出，可以理解用户的基本自然语言语音指令并进行回应，也可以将生成的文本以语音形式播放出来。此外，OpenAI 于 2024 年 1 月上线包含超 300 万个应用的 GPT 商店 GPTs。GPTs 的应用被划分为“写作”“效率”“研究和分析”“编程”“教育”和“生活方式”等类目。如热门应用 Consensus 可以从 2 亿篇学术论文中进行搜索，并获得科学的答案；Grimoire 可以在用户填写基本信息后提供所需的 HTML、CSS 和 JavaScript 代码，创建网站（或其他）的编程应用。

相比而言，国内的AI大模型产业至今没有出现爆款级应用，原因在于尚未找到商业化思路，缺乏满足客户需求的个性化应用。我国大模型产业要推出爆款级应用，势必要在应用领域做深做细，让每一个用户都可以充分享受到大模型所带来的真正便利。

第四章 天阔云高：中国 AI 大模型产业趋势展望

4.1 AI 云侧与端侧大模型满足不同需求，C 端用户将成为端侧的主要客群

我国云侧大模型百花齐放数量众多，以百度文心一言、阿里通义千问、科大讯飞星火、腾讯混元等为代表。强大的算力和海量的训练数据库，支撑大语言模型高参数，云侧大模型能够提供语言理解、知识问答、数学推理、代码生成等能力。一方面，面向 C 端个人用户，云侧大模型提供智能问答、文本生成、图片生成、视频生成等功能。另一方面，面向 B 端企业用户，云侧大模型变革企业传统业务模式，提供营销、客服、会议记录、文本翻译、预算管理等个性化服务。

端侧大模型具有成本低、移动性强、数据安全等优势，主要应用在手机、PC 等终端设备上。端侧大模型主要面向 C 端用户，重塑传统个人设备的使用方式和习惯，提供手机文档搜索、智能识屏、图像创作、生活助手、出行助手等专属服务。成本方面，根据云侧大模型每次调用成本、用户数、用户使用频率不同，云侧大模型服务器每年成本可达数亿或数十亿，高昂的服务器支出成为各大厂商发展大模型的障碍。将大模型端侧化，能把一部分云端计算转移给终端，从而大大降低云端服务器成本。安全方面，由于端侧大模型数据保存在本地，个人数据不需要上传云端，个人隐私数据更加安全。丰富的使用场景、较低的模型成本、安全的隐私保护，使得未来大模型端侧化可能成为趋势。瑞银预计生成式 AI 智能手机出货量将从 2023 年的 5000 万部增长到 2027 年的 5.83 亿部，到 2027 年收入将达 5130 亿美元。未来面向广大 C 端用户的端侧大模型市场前景广阔。

4.2 AI 大模型趋于通用化与专用化，垂直行业将是大模型的主战场

通用大模型具有参数规模大、泛化能力强、多任务学习能力优等特点。通用大模型参数规模较大，达到数百亿甚至上千亿参数。通过大规模数据训练，通用

大模型能学习捕捉复杂规律和特征，对未见过的数据做出预测。通用大模型能理解学习多种任务，如文本总结、对话问答、逻辑推理等。通用大模型得益于大规模预训练和微调范式，可完成多领域任务，并具备多模态（包括文字、图像、语音、视频）理解和生成能力。

行业大模型适用于金融、政务、医疗等特定行业和领域，更好处理相关行业的特定任务。具体而言，金融大模型能帮助金融企业评估信用风险；政务大模型提供政务问答、公文撰写润色、内容审核；医疗大模型为医生和患者提供影像诊疗、手术评估、导诊服务。

与通用大模型相比，行业大模型具有专业性强、数据安全性高等特点，未来大模型真正的价值体现在更多行业及企业的应用落地层面。一方面，行业大模型将通用大模型用于形成多领域能力的资源集中于特定领域，模型参数相对较小，对于企业落地而言具有显著的成本优势。另一方面，行业大模型结合企业或机构内部数据，为 B 端用户的实际经营场景提供服务，能更加体现模型对于机构的降本增效作用。

4.3 AI 大模型将广泛开源，小型开发者可调用大模型能力提升开发效率

未来，大模型开源将成为趋势，一方面能降低大模型开发者的使用门槛，另一方面也能提高算法的透明度和可信度。从具体进展来看，2022年8月，清华大学开源中英双语预训练模型 GLM-130B，使用通用模型算法进行预训练。2023年6月，百川智能发布开源可商用大规模预训练语言模型 Baichuan-7B，支持中英双语。2023年10月，智谱AI 开源ChatGLM3系列模型。2023年11月，vivo开源70亿参数的大模型，向广大开发者提供了获取大模型技术的渠道。2023年12月，阿里云开源 Qwen-72B、Qwen-1.8B 和Qwen-AudioQwen大模型。随着大模型逐渐开源，将进一步助推AI大模型产业的创新发展。

小型开发者通过调用大模型能力，大幅提升编程效率，进一步推动 AI 应用落地。一方面，小型开发者可基于大模型进行项目、应用以及插件等开发工作，

不再局限于算力资源、无需进行复杂的模型训练、调参，轻松实现应用落地。另一方面，小型开发者利用大模型技术提升开发效率，通过在代码工具中集成大模型能力，辅助完成部分重复性工作，为开发人员提供量身定制的代码建议，还可以自动检测代码中的 Bug，并生成相应的测试用例，缩短工程师开发流程中的编码和纠错时间。

4.4 AI 高性能芯片不断升级，AI 大模型产业生态体系将不断完善

在大模型场景下，AI 高性能芯片主要用于大模型的训练环节，芯片性能的强弱直接影响大模型的性能和表现。在全球 AI 高性能芯片市场中，英伟达的芯片产品采用最前沿半导体工艺和创新 GPU 架构保持行业的领先地位。目前，英伟达的 A100 芯片在主流 AI 大模型训练中占据重要市场份额，H100 虽性能强劲但难以获取。AI 高性能芯片未来将不断迭代升级，持续推动大模型性能和能力的提升。在国内，AI 高性能芯片近年来发展速度加快。其中，华为昇腾主要包括 310 和 910 两款主力芯片，其中昇腾 910 采用了 7nm 工艺，最高可提供 256 TFLOPS 的 FP16 计算能力，其能效比在行业中处于领先水平。寒武纪是中国具有代表性的另一本土 AI 芯片厂商，公司先后推出了思元 290 和思元 370 芯片及相应的云端智能加速卡系列产品、训练整机。未来，随着全球 AI 高性能芯片不断迭代升级，也将持续推动大模型性能和能力的提升。

结语

AI 大模型将加快新质生产力发展，助力我国经济社会高质量发展

AI 大模型可以创造新价值、适应新产业、重塑新动能，是加快发展新质生产力的关键要素。AI 大模型作为当前人工智能领域的重要技术，是孕育新质生产力的沃土。新质生产力是创新起主导作用，摆脱传统经济增长方式、生产力发展路径，具有高科技、高效能、高质量特征，符合新发展理念的先进生产力质态，其由技术革命性突破、生产要素创新性配置、产业深度转型升级而催生。以劳动者、劳动资料、劳动对象及其优化组合的跃升为基本内涵，以全要素生产率大幅提升为核心标志。AI 大模型作为实现新质生产力发展的重要手段，可以推动多个领域的智能化升级，提高生产效率、降低生产成本、提升产业竞争力。随着中国经济进入高质量发展阶段，AI 大模型在催生新产业、新模式、新动能方面展现出巨大潜力，不仅支撑了经济社会的高质量发展，也符合《国家创新驱动发展战略纲要》所强调的创新驱动和产业升级要求。我国众多产业对于高质量发展的需求，将为大模型的落地应用提供场景支撑。随着人工智能技术的不断升级，大模型产业化应用也成为可能。以 vivo 为代表的科技企业发布的大模型为生产生活提供更多便利，带动商业模式创新，牵引产业升级，令人们生活更加美好。面对未来，我国需进一步加强资源与研发力量的统筹，强化大模型在发展中的场景牵引作用，促进经济社会的高质量发展，以实现大模型技术的高质量应用突破，驱动实体经济的蝶变和产业变革。