# 人工智能大模型发展对电信运营商机遇分析及对策研究

## Research on the Development of Artificial Intelligence Big Models and Analysis of Opportunities for Telecom Operators and Countermeasures

殷凯凯

Kaikai Yin

中国电信股份有限公司北京分公司　中国·北京 100032

Beijing Branch, China Telecom Company Limited, Beijing, 100032, China

**摘　要：**以 ChatGPT 为代表的 AI 大模型应用快速发展，为信息通信行业带来新的机遇和挑战。论文介绍了人工智能大模型的概念、发展阶段、应用场景、发展趋势与挑战，分析了人工智能大模型发展对电信运营商企业带来的机遇与挑战。并在此基础上提出了电信运营商应从算力基础设施建设、布局算力服务、探索构建大模型即服务（MaaS）产业生态三个方面采取积极行动，拥抱大模型发展趋势，最后提出相关工作建议。

**Abstract:** The rapid development of AI big model applications represented by ChatGPT has brought new opportunities and challenges to the information and communication industry. The paper introduces the concept, development stages, application scenarios, development trends and challenges of artificial intelligence big models, and analyzes the opportunities and challenges that the development of artificial intelligence big models brings to telecommunications operators. On this basis, it is proposed that telecommunications operators should take active actions from three aspects: computing infrastructure construction, layout of computing services, and exploration of building a MaaS industry ecosystem, embracing the trend of large-scale model development, and finally proposing relevant work suggestions.

**关键词：**人工智能；大模型；算力；智算中心；大模型即服务（MaaS）

**Keywords:** Artificial Intelligence; large mode; computing power; intelligent computing center; MaaS

**DOI:** 10.12346/csai.v2i1.9110

## 1 引言

2022 年年底 OpenAI 公司的大语言模型应用 ChatGPT 一经发布，便在社会上引发大模型和人工智能热潮，目前 AI 大模型已经成为产学研各界关注的焦点。截至 2023 年 8 月全球发布大模型数量超过 200 个，中国超过 100 个。AI 大模型呈现快速发展之势，并在多个领域政务、金融、医疗、教育、办公、交通、应急等多个行业广泛应用。随着 AIGC（人工智能生成内容）、大模型等算力新应用、新业态不断涌现，全社会对算力、数据、模型算法的需求快速增长。论文总结 AI 大模型相关概念、现状和发展趋势，分析了给电信运营商带来的机遇，最后给出应对策略和发展建议。

## 2 AI 大模型发展现状及趋势

### 2.1 AI 大模型概念界定

AI 模型可分为中小模型和大模型，中小模型参数规模通常在几万到几百万不等，大模型参数量一般超过 10 亿。大模型，也可以称之为基础模型，是人工智能预训练大模型的统称 [1]。大模型狭义上指基于深度学习算法进行训练的自然语言处理（NLP）模型，主要应用于自然语言理解和生成等领域，广义上还包括机器视觉（CV）大模型、多模态大模型和科学计算大模型等。图 1 梳理了 AI 模型、大模型、大语言模型和 ChatGPT 的关系。

【作者简介】殷凯凯，硕士，从事数据中心、智算中心等算力基础设施规划、建设、运营管理，算力服务管理等研究。

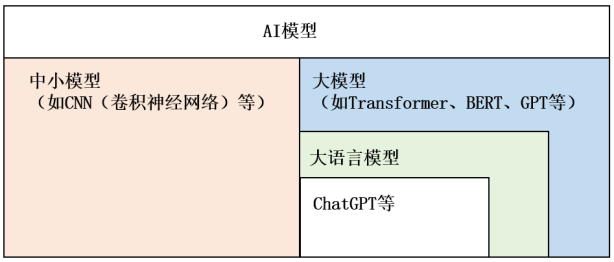


**图 1 AI 模型、大模型、大语言模型和 ChatGPT 关系示意图**

### 2.2 AI 大模型发展阶段

大模型发展主要经历了三个阶段，分别是早期萌芽期、沉淀积累期和快速演进期。第一个阶段早期萌芽期（1950—2005）是以卷积神经网络（CNN）为代表的模型阶段。卷积神经网络和深度学习技术的发展对后续深度学习框架的迭代及大模型发展奠定了坚实的基础。第二阶段沉淀积累期（2006—2019 年）是以 Transformer 为代表的全新神经网络模型阶段。2017 年 6 月，谷歌机器翻译团队提出采用新型的自注意力机制的深度学习模型——Transformer 来进行机器翻译任务，并且取得了很好的效果 [2]。奠定了大模型预训练算法架构的基础，进而引发了大模型的创新性发展。第三阶段快速演进期（2020—）是以 GPT 为代表的预训练大模型阶段。此阶段由 2020 年 OpenAI 公司推出了 GPT-3 开始，随后 ChatGPT 和 GPT-4 相继推出，在全球范围内引发人工智能和 AI 大模型热潮。图 2 梳理了从 2017 年至 2023 年全球主要大模型的发展历程。

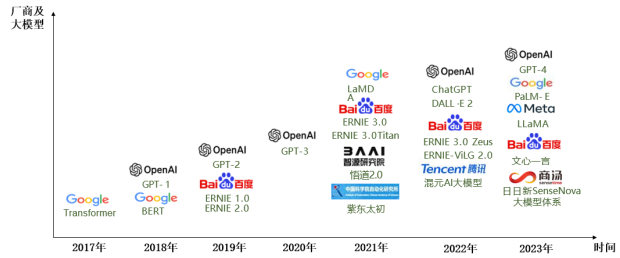


**图 2 2017 年至 2023 年全球主要大模型发展历程**

### 2.3 AI 大模型应用场景

根据公开信息，目前全球已经发布大模型超过 200 个，其中中美数量占比超过 90%。在国内来看，发布大模型的公司主要有四类，AI 科技公司、互联网公司、行业公司及科研学术机构。大模型主要典型应用有文本生成、音频生成、视频生成、图像生成和数字人等，并在办公、金融、医疗、教育、交通等领域开展落地应用。比如在办公领域，用户可以通过大模型进行文档摘要、自助翻译、文档分类、智慧办公等方面提高文档处理的效率。在金融领域，可通过大模型在市场营销、产品设计、风险管控、客户服务、运营支持等全面赋能金融机构，提升金融机构的服务效率。在医疗领域，AI 大模型在药物研发、预约就诊、预检分诊以及导诊，影像、综合辅助诊断及远程医疗等应用较为广泛。

### 2.4 大模型发展趋势和挑战

人工智能大模型呈现快速发展的态势，未来将涉及更多模态和场景，呈现四方面发展趋势：第一，从模型参数来看，整体参数规模还将持续增长。过去十余年，大语言模型参数规模增长数百万倍，目前千亿级参数规模的大模型成为主流。第二，从模型算法来看，模型算法会越来越聚焦，国内的大模型上除了采用 GPT 算法架构之外，还对 BERT、ALBERT. NEZHA 等进行了广泛的探索。不过在 GPT3.0 发布后，GPT 逐渐成为大模型的主流路线。第三，从模型算力来看，大模型发展需要大算力的支撑。大模型参数量的增加导致训练过程的计算需求呈现指数级增长。第四，从应用领域来看，通用大模型与行业大模型同步发展，未来除少数几家通用大模型外，在能源、金融、制造、教育、政务等不同领域行业大模型将发挥更大作用。

AI 大模型发展受算力、数据、算法、网络、能源、隐私及安全、知识产权等方面影响与制约。大模型需要大算力且投资成本高昂。GPT-3 训练需要上万块 A100 GPU 芯片，所耗费的算力大约为 3640 PF-days（即 1PetaFLOP/s 效率训练 3640 天）。据估算，GPT-3 训练成本约为 140 万美元，一些更大参数规模的大型语言模型约 1120 万美元。大模型应用运行功耗大，耗电多。另外，大模型训练及推理阶段对大规模算力组网、异构算力融合与调度、绿色低碳等技术提出很高要求。

## 3 AI 大模型发展带来机遇与挑战分析

### 3.1 大模型需要大算力，带来算力基础设施发展机遇

算力，通俗理解即计算能力。算力基础设施是集信息计算力、网络运载力、数据存储力于一体的新型信息基础设施 [3]。中国信通院将算力分为通用算力、智能算力、超算算力和边缘算力 [4]。与大模型发展息息相关的是智能算力，智能算力是以 GPU、FPGA 和 AI 芯片等输出的人工智能计算能力为主，具备渲染、推理和模拟能力，可面向智能驾驶、人脸识别、大模型等人工智能应用提供智算服务的一种算力服务形态 [5]。

AI 大模型的发展需要强大的智能算力支撑。智能算力以智算中心为承载体。智算中心是典型的算力基础设施，在国家政策和人工智能大模型新技术的双重驱动下迎来较快发展。政策方面，从中央到地方城市，近几年密集出台了一系列人工智能与算力基础设施支持政策，为加快推动算力基础设施规划建设指明方向。技术方面，受大模型、CPU、GPU 异构算力、算力网络等技术发展，各地掀起智算中心建设热潮。据 2023 年发布的《智能计算中心创新发展指南》显示，目前全国有超过 30 个城市正在建设智算中心。电信运营商企业作为信息基础设施建设及运营者迎来难得发展机遇。

### 3.2 带来算力服务发展机遇

算力服务可理解为以多样性算力为基础，以算力网络为

连接，以供给有效算力为目标的算力产业新领域，通过全新计算技术实现异构算力统一输出，并与云、大数据、AI（人工智能）等技术交叉融合，最终将算力、存储、网络等资源统一封装，以服务形式完成算力交付[6]。

电信运营商自身拥有优质网络、算力和云计算能力。在人工智能大模型时代，为把握发展机遇，应逐步探索向算力服务转型，积极推进算力服务及产品布局，为企业增长打造“第二增长曲线”，同时践行社会责任，推动算力服务“普惠化”“泛在化”“标准化”，使其成为社会公共资源。

### 3.3 带来大模型即服务（MaaS）发展机遇

目前已有 AI 和云计算公司推出了大模型即服务。以美国人工智能研究公司 OpenAI 为例，OpenAI 在 2023 年 2 月 1 日推出 ChatGPT Plus 试点订阅计划以及 API 付费调取服务，这是典型的 MaaS 订阅制收费服务。国内腾讯云、阿里云等公司也推出了 MaaS 服务。为把握发展机遇，运营商企业也陆续发布大模型应用并探索建立大模型产业生态。中国电信在 2023 世界人工智能大会正式对外发布中大语言模型 TeleChat。中国移动于 2023 年 7 月 8 日正式发布九天·海算政务大模型和九天·客服大模型。中国联通发布“鸿湖图文大模型 1.0”，鸿湖图文大模型，可以实现文本生成图像、视频剪辑和图像生成图像等功能。

### 3.4 对算力网络的高带宽、低时延等提出更高要求

高质量算力网络是电信运营商提供算力服务的基础条件，也是影响客户感知的重要因素。算力网络需要满足高带宽、低时延、零丢包、超高稳定性和网络自动化部署等要求。目前业界一般采用 InfiniBand 或 RoCE 在智算中心内组网，提供超低时延无损算力网络，确保智算中心集群内训练 POD 间及计算、存储的高速互联。

## 4 大模型时代电信运营商应对策略探讨

### 4.1 布局智算中心，建设算力基础设施

运营商企业积极建设智算基础设施，满足未来大模型训练和推理需求是大势所趋。从布局选址方面，优先在“东数西算”八大枢纽节点布局集中化、大规模、低成本大型算力中心。其次是聚焦北京、上海、广州等一线城市，积极联合地方政府加快行业智能算力中心布局建设。从建设运营方面，面对智算中心的高能耗、高成本，传统数据中心机房环境、制冷方式很难满足，应加快液冷服务器等节能新设备和技术应用落地，建设绿色低碳算力基础设施。从训练和推理算力部署方面，首先是训练算力池，建议运营商企业集团层面统筹考虑智算中心的建设布局，在低成本园区统一建设大型公共训练池，在经济热点一线城市以满足行业客户需求为重点布局行业算力中心。其次是推理算力池，可在现有的云资源池中增加 GPU 算力池，具体部署位置可以按需规划部署。

### 4.2 积极推进算力服务及产品布局

推进算力服务及产品布局，开展算力资源租赁服务和平台服务是大模型时代运营商企业的重要市场业务。算力资源租赁是算力服务的初级模式，以算力基础资源（智算中心、算力网络、AI 服务器、GPU 卡等）租赁为主。平台运营是算力服务的进阶模式，以云计算管理服务平台和算力调度管理平台运营为主。云计算管理服务平台方面，主要提供 AI 服务器训练环境的预置能力、算力资源运维能力、资源管理能力（可以进行能力输出，纳管合作方或者客户的资源池）等 IaaS 和 PaaS 服务。算力调度管理平台方面，借助电信运营商的云公司能力提供自建 / 合作算力资源的统一管理、调度以及使用，实现一点开通、一点管理。

### 4.3 提供大模型即服务（MaaS）

电信运营商提供 MaaS 服务，需要重点聚焦应用场景和生态运营两个方面，坚定做大通用大模型和行业大模型生态。应用场景方面，响应国家政策，立足客户需求，重点做好政务、医疗、教育、金融、应急、文旅等行业场景解决方案。生态运营方面，持续扩大生态圈，结合生态合作伙伴能力，做好通用模型、行业模型、企业专属大模型的构建、训练、推理和生产变现。总体来说运营商 MaaS 服务挑战较大，目前应着力开展能力建设，提供技术水平，培养专家队伍，联合产业生态合作伙伴共同探索场景应用和商业模型。

## 5 结语

以 ChatGPT 为代表的人工智能大模型应用快速发展，开启了人工智能新一轮增长，并带来算力基础设施建设热潮和新型算力服务模式，给信息通信业带来新的机遇和挑战。对于电信运营商而言，需要积极把握行业发展趋势，结合自身资源、能力和技术优势，从建设算力基础设施、布局算力服务、打造大模型即服务产业生态三方面积极布局，拓展新赛道，打造业务新增长点，尽快形成第二增长曲线。

## 参考文献

[1] 刘亮,张琛,杨学燕.生成式人工智能技术对通信行业的影响研究[J].邮电设计技术,2023(7):1-7.

[2] Vaswani A , Shazeer N , Parmar N ,et al.Attention Is All You Need[J].arXiv, 2017.

[3] 工业和信息化部等六部门. 算力基础设施高质量发展行动计划[R/OL].(2023-10-08) https://www.gov.cn/zhengce/zhengceku/202310/content_6907900.htm.

[4] 2022中国算力大会.《中国算力白皮书(2022)》[R/OL].(2022-08-17)[2023-06-25]. https://www.odcc.org.cn/news/p-1559872438149832705.html.

[5] 郭亮.数据中心发展综述[J].信息通信技术与政策,2023,49(5):2-8.

[6] 中国信息通信研究院云计算与大数据研究所.中国算力服务研究报告(2023年)[R].(2023-07).