腾讯云

AI大模型应用发展研究报告

电信运营商与云服务商的合作探索

腾讯云小程序

腾讯云计算（北京）有限责任公司

编委会

编委成员

主编：

张晋、栗蔚

编委（排名不分先后）：

秦若毅、周锐、吴炳文、马飞、苏越、赵伟博、桑柳

参编单位：

腾讯云计算（北京）有限责任公司

中国信息通信研究院云计算与大数据研究所

中国通信标准化协会

目录 / CONTENTS

1. 百模大战，电信运营商入局AI大模型 02

1.1 人工智能研究持续深入，大模型再掀浪潮 03
1.2 AI大模型市场规模持续增长，国内外呈现混战格局 03
1.2.1 海外企业占据先发优势，AI大模型已经多轮迭代 04
1.2.2 国内企业紧抓发展机遇，通用、专用、开源、闭源全面发展 05
1.3 大模型建设方持续多元化发展，电信运营商走出“体系化”建设道路 08

2. 优势互补，电信运营商与云服务商在竞争中探索合作共赢新局面 10

2.1 电信运营商AI大模型发挥通信业语料优势，用语音大模型打开市场 11
2.2 云服务商AI大模型发展发挥快速迭代落地优势，积累丰富市场反馈 13
2.3 “1+3+N”合作体系，云服务商全面助力电信运营商发展AI大模型 14

3. 一集群三路线，云服务商助力电信运营商进行软硬兼备的AI大模型建设储备 16

3.1 云服务商支持高效算力集群建设 17
3.1.1 算力集群建设与发展面临的挑战 17
3.1.2 构建高效算力集群的关键技术 20
3.2 云服务商打造三大软件合作路线 25
3.2.1 行业智算云+标准化应用：合力推广开箱即用的大模型软件 25
3.2.2 私有云集成+标准化组件：合力承建私域化的知识增强型应用 25
3.2.3 项目总集成+智算技术底座：合力支持按需定制的客户大模型 26
3.3 运营商和云服务商的融合共建价值 26
3.3.1 同质化的硬件堆叠难以保证竞争中的优势 26
3.3.2 运营商优质资源和云服务商最佳实践的结合 27

4. N个场景，云服务商支持电信运营商构建AI大模型场景化解决方案 28

4.1 企业知识应用场景 29
4.2 视联网内容分析场景 30
4.3 增值内容创作场景 33
4.4 客户服务场景 35
4.5 DICT合作场景 37

5. 电信运营商大模型应用案例 40

5.1 强强联合共建大模型算力集群 41
5.2 帮助运营商提高视频分析能力 44
5.3 为5G视频彩铃提供内容制作能力 46
5.4 AI代码助手助力运营商研效提升 49
5.5 行业大模型拓展运营商CH端场景 51

6. 电信运营商大模型发展展望 52

6.1 技术演进，大模型建设与应用不断探索高效率、高精度、高适用性 53
6.2 应用创新，电信运营商大模型要紧抓行业内、外痛点，打造差异化竞争力 53
6.3 跨领域协同，电信运营商与其他产业角色优势互补，谋求双赢 53

目录 / CONTENTS

1. 百模大战，电信运营商入局AI大模型 02

1.1 人工智能研究持续深入，大模型再掀浪潮 03
1.2 AI大模型市场规模持续增长，国内外呈现混战格局 03
1.2.1 海外企业占据先发优势，AI大模型已经多轮迭代 04
1.2.2 国内企业紧抓发展机遇，通用、专用、开源、闭源全面发展 05
1.3 大模型建设方持续多元化发展，电信运营商走出“体系化”建设道路 08

2. 优势互补，电信运营商与云服务商在竞争中探索合作共赢新局面 10

2.1 电信运营商AI大模型发挥通信业语料优势，用语音大模型打开市场 11
2.2 云服务商AI大模型发展发挥快速迭代落地优势，积累丰富市场反馈 13
2.3 “1+3+N”合作体系，云服务商全面助力电信运营商发展AI大模型 14

3. 一集群三路线，云服务商助力电信运营商进行软硬兼备的AI大模型建设储备 16

3.1 云服务商支持高效算力集群建设 17
3.1.1 算力集群建设与发展面临的挑战 17
3.1.2 构建高效算力集群的关键技术 20
3.2 云服务商打造三大软件合作路线 25
3.2.1 行业智算云+标准化应用：合力推广开箱即用的大模型软件 25
3.2.2 私有云集成+标准化组件：合力承建私域化的知识增强型应用 25
3.2.3 项目总集成+智算技术底座：合力支持按需定制的客户大模型 26
3.3 运营商和云服务商的融合共建价值 26
3.3.1 同质化的硬件堆叠难以保证竞争中的优势 26
3.3.2 运营商优质资源和云服务商最佳实践的结合 27

4. N个场景，云服务商支持电信运营商构建AI大模型场景化解决方案 28

4.1 企业知识应用场景 29
4.2 视联网内容分析场景 30
4.3 增值内容创作场景 33
4.4 客户服务场景 35
4.5 DICT合作场景 37

5. 电信运营商大模型应用案例 40

5.1 强强联合共建大模型算力集群 41
5.2 帮助运营商提高视频分析能力 44
5.3 为5G视频彩铃提供内容制作能力 46
5.4 AI代码助手助力运营商研效提升 49
5.5 行业大模型拓展运营商CH端场景 51

6. 电信运营商大模型发展展望 52

6.1 技术演进，大模型建设与应用不断探索高效率、高精度、高适用性 53
6.2 应用创新，电信运营商大模型要紧抓行业内、外痛点，打造差异化竞争力 53
6.3 跨领域协同，电信运营商与其他产业角色优势互补，谋求双赢 53

前言 / FOREWORD

OpenAI 2022年底发布ChatGPT再度引爆人工智能的全球研究热潮，各国纷纷投入或加强对AI大模型的研究，其中中国、美国成果频出，引领产业发展。

从市场格局来看，海外企业占据大模型先发优势，几大巨头科技企业及个别人工智能企业已经完成几轮AI大模型迭代，性能不断提升；国内AI大模型建设方“百花齐放”，依托技术优势及业务经验，在构建通用大模型的同时，打造具备企业业务特色、行业特色的专用AI大模型，并在AI大模型开源领域积极贡献。

从产业角色来看，海外AI大模型建设方以科技企业为主，其他类型企业陆续入局；国内建设方不断多元化发展，已形成云服务商、电信运营商、设备厂商、互联网厂商等多种产业角色共建的局面。

多产业角色入局带来了竞争，同时也带来了协同合作的机会。

本报告对云服务商与电信运营商在此背景下的竞合展开研究，旨在发挥二者的技术、产品、应用优势，提出“1+3+N”合作体系：联合打造1个适合AI大模型培育的算力集群；沿着标准软件研发、标准模型能力增强与定制化模型精调3条路线开展技术攻关；从企业知识应用、视联网内容分析、增值内容创作、客服场景、DICT合作等N个场景形成解决方案。

从AI大模型设计、建设到应用，探索云服务商全流程支持电信运营商AI大模型发展的机遇的措施。

未来，腾讯云与中国信通院将持续关注AI大模型发展与应用，继续深入研究云服务商、运营商等产业角色的合作，在人工智能的浪潮下发挥不同产业角色的优势，通过产业协同合作共同助力生态建设。

本报告内容仍有诸多不足，恳请各界批评指正。

01

百模大战，电信运营商入局AI大模型

人工智能历经数十年的发展，大模型的出现再度引发学术界、产业界的关注，该领域表现出的巨大发展潜力吸引了国内外数百家企业投入其中，工业和信息化部数据显示截至 2023 年年底，我国累计发布了 200 多个人工智能大模型。与以往人工智能的研究不同，大模型由于其更靠近产业应用，商业化路径较短，加上广阔的应用空间，吸引了众多类型的企业成为建设方，形成了如今“百模大战”的局面，电信运营商也早早入局，依托其产业优势，打造从硬件到软件体系化的大模型产品。

数据来源：微信公众号“工信微报”2024年2月27日《数字经济时代如何推进中国工业现代化？一组数据为你全面展现》

前言 / FOREWORD

OpenAI 2022年底发布ChatGPT再度引爆人工智能的全球研究热潮，各国纷纷投入或加强对AI大模型的研究，其中中国、美国成果频出，引领产业发展。

从市场格局来看，海外企业占据大模型先发优势，几大巨头科技企业及个别人工智能企业已经完成几轮AI大模型迭代，性能不断提升；国内AI大模型建设方“百花齐放”，依托技术优势及业务经验，在构建通用大模型的同时，打造具备企业业务特色、行业特色的专用AI大模型，并在AI大模型开源领域积极贡献。

从产业角色来看，海外AI大模型建设方以科技企业为主，其他类型企业陆续入局；国内建设方不断多元化发展，已形成云服务商、电信运营商、设备厂商、互联网厂商等多种产业角色共建的局面。

多产业角色入局带来了竞争，同时也带来了协同合作的机会。

本报告对云服务商与电信运营商在此背景下的竞合展开研究，旨在发挥二者的技术、产品、应用优势，提出“1+3+N”合作体系：联合打造1个适合AI大模型培育的算力集群；沿着标准软件研发、标准模型能力增强与定制化模型精调3条路线开展技术攻关；从企业知识应用、视联网内容分析、增值内容创作、客服场景、DICT合作等N个场景形成解决方案。

从AI大模型设计、建设到应用，探索云服务商全流程支持电信运营商AI大模型发展的机遇的措施。

未来，腾讯云与中国信通院将持续关注AI大模型发展与应用，继续深入研究云服务商、运营商等产业角色的合作，在人工智能的浪潮下发挥不同产业角色的优势，通过产业协同合作共同助力生态建设。

本报告内容仍有诸多不足，恳请各界批评指正。

01

百模大战，电信运营商入局AI大模型

人工智能历经数十年的发展，大模型的出现再度引发学术界、产业界的关注，该领域表现出的巨大发展潜力吸引了国内外数百家企业投入其中，工业和信息化部数据显示截至 2023 年年底，我国累计发布了 200 多个人工智能大模型。与以往人工智能的研究不同，大模型由于其更靠近产业应用，商业化路径较短，加上广阔的应用空间，吸引了众多类型的企业成为建设方，形成了如今“百模大战”的局面，电信运营商也早早入局，依托其产业优势，打造从硬件到软件体系化的大模型产品。

数据来源：微信公众号“工信微报”2024年2月27日《数字经济时代如何推进中国工业现代化？一组数据为你全面展现》

1.1 人工智能研究持续深入，大模型再掀浪潮

人工智能（Artificial Intelligence，AI）技术发展带来的变革已经深入人类的生产生活，近年来，各国也在逐渐加强对人工智能技术研究的重视程度，提升通过人工智能等新技术发展国家科技水平的战略地位。中国信息通信研究院《人工智能白皮书（2022年）》提到，自2016年起，先后有40余个国家和地区将推动人工智能发展上升到国家战略高度，越来越多的国家认可发展人工智能技术对提升国家全球竞争力的关键作用。例如欧盟2021年发布的《2030数字指南针：欧洲数字十年之路》将人工智能的使用程度作为欧洲企业数字化转型目标，指出到2030年，75%的欧洲企业使用云计算、大数据和人工智能；美国将人工智能作为未来产业进行重点布局，陆续发布《2020年未来产业法案》、《关于加强美国未来产业领导地位的建议》等确保在人工智能领域的投入以保障其在科技领域的竞争力；中国在《“十四五”数字经济发展规划》中强调要增强人工智能等关键技术的创新能力。

人工智能连续多年作为各国的研究热点，2022年底OpenAI发布ChatGPT3.5将人工智能的研究热度再度推上了新的高潮，自此，AI大模型的研究成为了人工智能领域的关键技术之一。

AI大模型，通常指具备十亿以上参数和复杂计算结构的人工智能机器学习模型，由深度神经网络构建而成，具备六大特点：模型大小上，可达数百GB；数据级规模上，需要TB甚至PB级的数据支持训练；训练时长上，通常需要成百上千个GPU训练几周甚至几个月；模型精度上，拥有相比作为早年研究热点的小模型更强的表达能力和更加准确的推理能力；推理能力上，能够支持文本、图像、音频、视频等多模态数据；智能化程度上，体现出了更近似人类的归纳和思考能力。

对AI大模型的研究意义深远，从技术的演进来看，一方面AI大模型的探索与实验将积累大量的大规模计算经验，驱动算力设备、集群、服务等整个产业链发展及完善其大规模计算支持能力，另一方面AI大模型在文本、图像等多模态数据的推理、生成上表现出的良好性能，推动技术跨学科、跨行业融合创新，展示出巨大的发展潜力；从应用的价值来看，一方面AI大模型支持应用以自然语言交互的方式提供服务，为产业应用提供了新的发展思路与机会，另一方面AI大模型有望将人力释放至资源更紧缺的岗位，同时可能衍生出新的产业角色与岗位，可能带动新一轮的产业变革。

AI大模型的发展与应用价值，近两年吸引了产业、学术、应用等领域内多家企业的关注与投入，企业纷纷开发各自的大模型及相关应用，逐渐形成了“百模大战”的格局。

1.2 AI 大模型市场规模持续增长，国内外呈现混战格局

据IDC预测，全球AI计算市场规模将从2022年的195亿美元增长到2026年的346.6美元，其中生成式AI计算市场规模将从2022年的8.2亿美元增长到2026年的109.9亿美元。在全球范围内，中美两国在AI大模型的研发上发挥着引领作用。2024全球数字经济大会上发布的最新数据显示，全球AI大模型已经超1300个，中国研发的AI大模型数量仅次于美国，位列全球第二，中美发布的AI大模型数量占全球的80%。其他国家也正加速发展人工智能产业及AI大模型，形成国内外AI大模型百花齐放、竞相发展的局面。

1.2.1 海外企业占据先发优势，AI大模型已经多轮迭代

海外AI大模型市场由美国主导，Google、Meta等互联网巨头企业与OpenAI及少部分人工智能初创企业共同引领行业发展，并与微软等企业“强强联合”进行应用验证，各模型已形成“多强竞合”的局面。

研发方面，以美国科技企业为主的多家企业陆续推出各自代表性的产品，经过市场验证后进行多次能力升级：

01 OpenAI

2022年11月，美国OpenAI发布人工智能技术驱动的自然语言处理工具ChatGPT-3.5，5天注册用户即破100万，被微软创始人比尔·盖兹称为可以与1980年图形用户界面相提并论的革命性技术；2023年3月，推出性能更优的ChatGPT-4，并于同年11月发布拟人化程度更高的ChatGPT-4-Turbo；2024年2月，发布人工智能文生视频大模型Sora；2024年5月，OpenAI发布了可支持跨模态综合理解和生成能力的ChatGPT-4o。

02 Google

除了OpenAI，Google在AI大模型的发展历史上同样是不可或缺的角色，早在2017年，Google即发布NLP模型Transformer，是目前大部分AI大模型训练时采用的网络结构，通过引入Self-Attention机制来提高模型训练速度，为后续大语言模型的升级迭代奠定了基础；2018年，发布大规模与训练模型BERT；2022年8月，Google推出高级语言学习模型PaLM；2023年5月，Google发布PaLM2与基于大模型的聊天机器人Bard；2023年12月，Google母公司Alphabet下设的人工智能实验室DeepMind发布AI大模型Gemini；2024年5月，Google I/O开发者大会上发布了由升级后Gemini模型驱动的AI助手项目Project Astra、对标Sora的文生视频模型Veo，以及在硬件方面发布的第六代Tensor处理器单元（TPU）Trillium芯片；同年8月，Google收购以利用大模型生成各种人物和角色风格对话为主业的人工智能初创公司Character.AI，反映出海外大模型市场正在展现出的市场格局变化趋势——大企业对具备创新能力、相关技术的中小企业展开并购。

03 Meta

2024年3月，Meta 将其大模型Llama 3正式开源，并宣称可通过AWS、Google Cloud、Azure等平台提供给开发者，同时获得了NVIDIA、AMD、Intel等硬件平台的支持，是大模型开源历程上重要的里程碑。

04 初创公司及研究机构

2023年3月，美国人工智能初创公司Anthropic对标ChatGPT、Gemini等产品，发布具有高级推理、视觉分析、代码生成、多模态等能力的大型语言模型家族Claude，2024年3月，Claude3系列发布，在数据和复杂任务理解等方面均有超出ChatGPT-4和Gemini 1.0 Ultra 的表现；2023年9月，法国企业Mistral AI推出了基座大模型Mistral 7B，并在基准测试中得到相当于或超出ChatGPT-3.5的性能表现；2024年8月，韩国LG集团新推出的人工智能研究机构 LG AI Research 首次公开了AI大模型 Exa One 3.0 轻量版作为开放源代码，该模型对韩语的应答做了针对性优化。

应用方面，微软率先拥抱AI大模型，其中，搜索引擎Bing将作为ChatGPT的默认搜索引擎，Azure AI Studio 将支持GPT-4o、Phi-3-vision、Llama 3、TimeGen-1等市面上主流的1600多种开/闭源大模型，帮助开发人员快速找到自己最想要的模型；推出Windows Copilot Runtime，支持包含DirectML、ONNX Runtime、PyTorch、WebNN在内的多个AI框架以及工具链如Olive、Visual Studio Code的AI工具包等，可以帮助开发人员快速引入自己的大模型，并在Windows硬件生态系统的广泛范围内扩展其AI应用。微软在大模型应用领域的探索，为全球大模型应用提供了有效的参考。

政策方面，美国在推进AI大模型发展的同时关注人工智能尤其是AI大模型潜在的内容安全风险，并发布《关于安全、可靠、可信地开发和使用人工智能的行政命令》等文件以对上述问题进行监管；英国科技部长于2023年4月宣布将斥资一亿英镑建立新的“基础模型工作组”，以开发能带来全球竞争力的AI方案；韩国2023年公布了人工智能大模型竞争力提升方案，聚焦非英语圈的全球市场，力争推进韩文人工智能大模型成为全球专业人工智能领域的领跑者。

1.1 人工智能研究持续深入，大模型再掀浪潮

人工智能（Artificial Intelligence，AI）技术发展带来的变革已经深入人类的生产生活，近年来，各国也在逐渐加强对人工智能技术研究的重视程度，提升通过人工智能等新技术发展国家科技水平的战略地位。中国信息通信研究院《人工智能白皮书（2022年）》提到，自2016年起，先后有40余个国家和地区将推动人工智能发展上升到国家战略高度，越来越多的国家认可发展人工智能技术对提升国家全球竞争力的关键作用。例如欧盟2021年发布的《2030数字指南针：欧洲数字十年之路》将人工智能的使用程度作为欧洲企业数字化转型目标，指出到2030年，75%的欧洲企业使用云计算、大数据和人工智能；美国将人工智能作为未来产业进行重点布局，陆续发布《2020年未来产业法案》、《关于加强美国未来产业领导地位的建议》等确保在人工智能领域的投入以保障其在科技领域的竞争力；中国在《“十四五”数字经济发展规划》中强调要增强人工智能等关键技术的创新能力。

人工智能连续多年作为各国的研究热点，2022年底OpenAI发布ChatGPT3.5将人工智能的研究热度再度推上了新的高潮，自此，AI大模型的研究成为了人工智能领域的关键技术之一。

AI大模型，通常指具备十亿以上参数和复杂计算结构的人工智能机器学习模型，由深度神经网络构建而成，具备六大特点：模型大小上，可达数百GB；数据级规模上，需要TB甚至PB级的数据支持训练；训练时长上，通常需要成百上千个GPU训练几周甚至几个月；模型精度上，拥有相比作为早年研究热点的小模型更强的表达能力和更加准确的推理能力；推理能力上，能够支持文本、图像、音频、视频等多模态数据；智能化程度上，体现出了更近似人类的归纳和思考能力。

对AI大模型的研究意义深远，从技术的演进来看，一方面AI大模型的探索与实验将积累大量的大规模计算经验，驱动算力设备、集群、服务等整个产业链发展及完善其大规模计算支持能力，另一方面AI大模型在文本、图像等多模态数据的推理、生成上表现出的良好性能，推动技术跨学科、跨行业融合创新，展示出巨大的发展潜力；从应用的价值来看，一方面AI大模型支持应用以自然语言交互的方式提供服务，为产业应用提供了新的发展思路与机会，另一方面AI大模型有望将人力释放至资源更紧缺的岗位，同时可能衍生出新的产业角色与岗位，可能带动新一轮的产业变革。

AI大模型的发展与应用价值，近两年吸引了产业、学术、应用等领域内多家企业的关注与投入，企业纷纷开发各自的大模型及相关应用，逐渐形成了“百模大战”的格局。

1.2 AI 大模型市场规模持续增长，国内外呈现混战格局

据IDC预测，全球AI计算市场规模将从2022年的195亿美元增长到2026年的346.6美元，其中生成式AI计算市场规模将从2022年的8.2亿美元增长到2026年的109.9亿美元。在全球范围内，中美两国在AI大模型的研发上发挥着引领作用。2024全球数字经济大会上发布的最新数据显示，全球AI大模型已经超1300个，中国研发的AI大模型数量仅次于美国，位列全球第二，中美发布的AI大模型数量占全球的80%。其他国家也正加速发展人工智能产业及AI大模型，形成国内外AI大模型百花齐放、竞相发展的局面。

1.2.1 海外企业占据先发优势，AI大模型已经多轮迭代

海外AI大模型市场由美国主导，Google、Meta等互联网巨头企业与OpenAI及少部分人工智能初创企业共同引领行业发展，并与微软等企业“强强联合”进行应用验证，各模型已形成“多强竞合”的局面。

研发方面，以美国科技企业为主的多家企业陆续推出各自代表性的产品，经过市场验证后进行多次能力升级：

01 OpenAI

2022年11月，美国OpenAI发布人工智能技术驱动的自然语言处理工具ChatGPT-3.5，5天注册用户即破100万，被微软创始人比尔·盖兹称为可以与1980年图形用户界面相提并论的革命性技术；2023年3月，推出性能更优的ChatGPT-4，并于同年11月发布拟人化程度更高的ChatGPT-4-Turbo；2024年2月，发布人工智能文生视频大模型Sora；2024年5月，OpenAI发布了可支持跨模态综合理解和生成能力的ChatGPT-4o。

02 Google

除了OpenAI，Google在AI大模型的发展历史上同样是不可或缺的角色，早在2017年，Google即发布NLP模型Transformer，是目前大部分AI大模型训练时采用的网络结构，通过引入Self-Attention机制来提高模型训练速度，为后续大语言模型的升级迭代奠定了基础；2018年，发布大规模与训练模型BERT；2022年8月，Google推出高级语言学习模型PaLM；2023年5月，Google发布PaLM2与基于大模型的聊天机器人Bard；2023年12月，Google母公司Alphabet下设的人工智能实验室DeepMind发布AI大模型Gemini；2024年5月，Google I/O开发者大会上发布了由升级后Gemini模型驱动的AI助手项目Project Astra、对标Sora的文生视频模型Veo，以及在硬件方面发布的第六代Tensor处理器单元（TPU）Trillium芯片；同年8月，Google收购以利用大模型生成各种人物和角色风格对话为主业的人工智能初创公司Character.AI，反映出海外大模型市场正在展现出的市场格局变化趋势——大企业对具备创新能力、相关技术的中小企业展开并购。

03 Meta

2024年3月，Meta 将其大模型Llama 3正式开源，并宣称可通过AWS、Google Cloud、Azure等平台提供给开发者，同时获得了NVIDIA、AMD、Intel等硬件平台的支持，是大模型开源历程上重要的里程碑。

04 初创公司及研究机构

2023年3月，美国人工智能初创公司Anthropic对标ChatGPT、Gemini等产品，发布具有高级推理、视觉分析、代码生成、多模态等能力的大型语言模型家族Claude，2024年3月，Claude3系列发布，在数据和复杂任务理解等方面均有超出ChatGPT-4和Gemini 1.0 Ultra 的表现；2023年9月，法国企业Mistral AI推出了基座大模型Mistral 7B，并在基准测试中得到相当于或超出ChatGPT-3.5的性能表现；2024年8月，韩国LG集团新推出的人工智能研究机构 LG AI Research 首次公开了AI大模型 Exa One 3.0 轻量版作为开放源代码，该模型对韩语的应答做了针对性优化。

应用方面，微软率先拥抱AI大模型，其中，搜索引擎Bing将作为ChatGPT的默认搜索引擎，Azure AI Studio 将支持GPT-4o、Phi-3-vision、Llama 3、TimeGen-1等市面上主流的1600多种开/闭源大模型，帮助开发人员快速找到自己最想要的模型；推出Windows Copilot Runtime，支持包含DirectML、ONNX Runtime、PyTorch、WebNN在内的多个AI框架以及工具链如Olive、Visual Studio Code的AI工具包等，可以帮助开发人员快速引入自己的大模型，并在Windows硬件生态系统的广泛范围内扩展其AI应用。微软在大模型应用领域的探索，为全球大模型应用提供了有效的参考。

政策方面，美国在推进AI大模型发展的同时关注人工智能尤其是AI大模型潜在的内容安全风险，并发布《关于安全、可靠、可信地开发和使用人工智能的行政命令》等文件以对上述问题进行监管；英国科技部长于2023年4月宣布将斥资一亿英镑建立新的“基础模型工作组”，以开发能带来全球竞争力的AI方案；韩国2023年公布了人工智能大模型竞争力提升方案，聚焦非英语圈的全球市场，力争推进韩文人工智能大模型成为全球专业人工智能领域的领跑者。

▲ 海外AI大模型发布情况

1.2.2 国内企业紧抓发展机遇，通用、专用、开源、闭源全面发展

中国积极响应全球大模型技术的发展趋势，高校、研究院所等科研机构,互联网企业，人工智能企业，电信企业等行业用户均不同程度地投入进AI大模型的建设中，陆续发布或开源基础AI大模型及具备行业属性、业务属性的AI大模型能力及产品，全面布局通用、专用、开源、闭源AI大模型发展路线。

研发方面，科研机构率先解决中国AI大模型“从0-1”的问题，互联网企业、电信企业等不同类型的企业纷纷入局为中国AI大模型研究生态注入动力。

01 科研机构

2021年3月，经过北京大学、清华大学、中国科学院等机构多个AI团队联合攻关，北京智源人工智能研究院发布中国第一个超大规模智能模型“悟道1.0”，并于同年6月发布了“悟道2.0”，达到1.75万亿参数，2023年6月，“悟道3.0”问世，并开源“悟道 · 天鹰”（Aquila）语言大模型系列和“悟道 · 视界”视觉大模型系列，与多个高校和科研院所合作构建FlagEval（天秤）开源大模型评测体系与开放平台；2023年2月，复旦大学发布大语言模型MOSS，并开源其研究成果；2023年3月，清华大学与智谱AI联合研发的大语言模型ChatGLM-6B正式发布，针对中文问答进行了优化，2024年1月，清华大学发布新一代基座大模型GLM-4，性能及中文能力逼近GPT-4，同年4月，清华大学联合北京生数科技有限公司共同研发视频大模型Vidu，是自Sora发布之后全球率先取得重大突破的视频大模型。

02 互联网企业及云服务商

中国大、中互联网型企业或其云服务公司基本均发布了各自的AI大模型及其配套的产品，并结合各自的产业优势、业务侧重推出适用于垂直场景的专用大模型。2021年，华为云发布盘古系列超大规模与训练模型；2023年3月，百度正式发布文心一言；4月，阿里巴巴发布“通义千问”语言模型，商汤发布“日日新SenseNova”大模型体系并陆续迭代5个版本；5月，科大讯飞发布“星火认知”大模型，迄今已经迭代5个版本；6月，360集团召开“360智脑大模型4.0”发布会，并宣称其已经接入360旗下产品全家桶；7月，华为云发布“盘古3.0”大模型，覆盖基础、行业及场景三层架构，京东推出“言犀”大模型，致力于服务零售、物流等产业，网易有道发布教育领域垂直大模型“子曰”；8月，抖音集团对外测试AI对话产品“豆包”，并发布“云雀”语言模型；9月，腾讯正式发布通用大语言模型“混元”，目前已经全面开源。2024年，除了已经发布的大模型不断迭代之外，7月，快手还发布了视频生成大模型“可灵”、图像生成大模型“可图”等系列大模型。

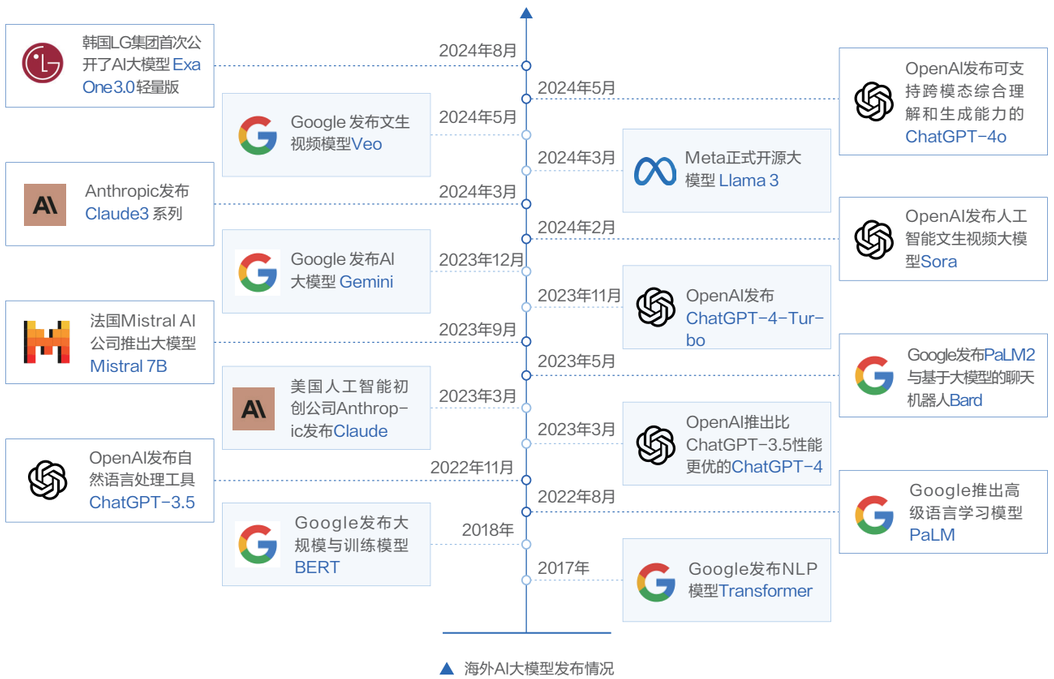
03 人工智能企业

近年来中国涌现出了一批人工智能企业，在AI大模型领域展示出了强大的创新力与竞争力。2023年6月，百川智能正式推出中英文训练大模型Baichuan-7B，并在随后的一年中陆续发布或开源了Baichuan2-7B、Baichuan2-13B和Baichuan-53B。2023年9月，北京月之暗面科技有限公司推出基于“Moonshot”大模型的智能助手Kimi，主要用于学术论文的翻译理解、法律问题辅助分析、API文档快速理解等场景，在长文本处理上表现出了良好的性能。Kimi已经在2024年开始探索商业化路线及个人用户的付费意愿。斑头雁（杭州）智能科技有限责任公司推出BetterYeah企业级AI应用开发平台，采用集成的思路内置ChatGLM、通义千问等国内外AI大模型。

应用方面，相较于海外的企业合作模式，中国的AI大模型应用具有更加明显的行业属性。医疗行业，AI大模型应用于病例的辅助编写、CT的智能辅助诊断、药物研发等，例如晶泰科技的XpeedPlay平台利用大模型技术，超高速生成苗头抗体，加速了药物的研发流程；交通行业，AI大模型帮助交通道路规划、辅助自动驾驶等，例如智己汽车采用生成式大模型帮助汽车理解用户意图，打造人车路多模态交互方式；教育行业，AI大模型参与知识问答、内容创作、个性化教学等，例如猿辅导利用AI大模型辅助进行学情分析、阅读对话、计算解题等教育能力。

政策方面，自AI大模型研究热度升高以来，中国及其多个省份陆续发布相关政策、措施，以支持和监管AI大模型的研究与应用，助力AI大模型发展并有效发挥正向的价值。2023年3月，工业和信息化部等四部门印发《新产业标准化领航工程实施方案（2023—2035年）》指出前瞻布局生成式人工智能等未来产业标准研究。7月，国家互联网信息办公室等部门联合发布《生成式人工智能服务管理暂行办法》，提出国家坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励生成式人工智能创新发展，对生成式人工智能服务实行包容审慎和分类分级监管。北京市2023年发布《北京市促进通用人工智能创新发展的若干措施》，上海市发布《上海市推动人工智能大模型创新发展若干措施（2023-2025年）》，广东、浙江等省份均发布相关文件，鼓励企业结合地方优势开展AI大模型研究工作。

尽管目前中国AI大模型技术仍处于研发和迭代的初期阶段，模型的性能、价值还有待接受市场的检验。目前AI大模型的研究和应用探索已经显现广阔的发展前景，中国企业的探索和创新能力使得中国有望在这场AI大模型的科技竞争中保持长久的竞争力。



▲ 海外AI大模型发布情况

1.2.2 国内企业紧抓发展机遇，通用、专用、开源、闭源全面发展

中国积极响应全球大模型技术的发展趋势，高校、研究院所等科研机构,互联网企业，人工智能企业，电信企业等行业用户均不同程度地投入进AI大模型的建设中，陆续发布或开源基础AI大模型及具备行业属性、业务属性的AI大模型能力及产品，全面布局通用、专用、开源、闭源AI大模型发展路线。

研发方面，科研机构率先解决中国AI大模型“从0-1”的问题，互联网企业、电信企业等不同类型的企业纷纷入局为中国AI大模型研究生态注入动力。

01 科研机构

2021年3月，经过北京大学、清华大学、中国科学院等机构多个AI团队联合攻关，北京智源人工智能研究院发布中国第一个超大规模智能模型“悟道1.0”，并于同年6月发布了“悟道2.0”，达到1.75万亿参数，2023年6月，“悟道3.0”问世，并开源“悟道 · 天鹰”（Aquila）语言大模型系列和“悟道 · 视界”视觉大模型系列，与多个高校和科研院所合作构建FlagEval（天秤）开源大模型评测体系与开放平台；2023年2月，复旦大学发布大语言模型MOSS，并开源其研究成果；2023年3月，清华大学与智谱AI联合研发的大语言模型ChatGLM-6B正式发布，针对中文问答进行了优化，2024年1月，清华大学发布新一代基座大模型GLM-4，性能及中文能力逼近GPT-4，同年4月，清华大学联合北京生数科技有限公司共同研发视频大模型Vidu，是自Sora发布之后全球率先取得重大突破的视频大模型。

02 互联网企业及云服务商

中国大、中互联网型企业或其云服务公司基本均发布了各自的AI大模型及其配套的产品，并结合各自的产业优势、业务侧重推出适用于垂直场景的专用大模型。2021年，华为云发布盘古系列超大规模与训练模型；2023年3月，百度正式发布文心一言；4月，阿里巴巴发布“通义千问”语言模型，商汤发布“日日新SenseNova”大模型体系并陆续迭代5个版本；5月，科大讯飞发布“星火认知”大模型，迄今已经迭代5个版本；6月，360集团召开“360智脑大模型4.0”发布会，并宣称其已经接入360旗下产品全家桶；7月，华为云发布“盘古3.0”大模型，覆盖基础、行业及场景三层架构，京东推出“言犀”大模型，致力于服务零售、物流等产业，网易有道发布教育领域垂直大模型“子曰”；8月，抖音集团对外测试AI对话产品“豆包”，并发布“云雀”语言模型；9月，腾讯正式发布通用大语言模型“混元”，目前已经全面开源。2024年，除了已经发布的大模型不断迭代之外，7月，快手还发布了视频生成大模型“可灵”、图像生成大模型“可图”等系列大模型。

03 人工智能企业

近年来中国涌现出了一批人工智能企业，在AI大模型领域展示出了强大的创新力与竞争力。2023年6月，百川智能正式推出中英文训练大模型Baichuan-7B，并在随后的一年中陆续发布或开源了Baichuan2-7B、Baichuan2-13B和Baichuan-53B。2023年9月，北京月之暗面科技有限公司推出基于“Moonshot”大模型的智能助手Kimi，主要用于学术论文的翻译理解、法律问题辅助分析、API文档快速理解等场景，在长文本处理上表现出了良好的性能。Kimi已经在2024年开始探索商业化路线及个人用户的付费意愿。斑头雁（杭州）智能科技有限责任公司推出BetterYeah企业级AI应用开发平台，采用集成的思路内置ChatGLM、通义千问等国内外AI大模型。

应用方面，相较于海外的企业合作模式，中国的AI大模型应用具有更加明显的行业属性。医疗行业，AI大模型应用于病例的辅助编写、CT的智能辅助诊断、药物研发等，例如晶泰科技的XpeedPlay平台利用大模型技术，超高速生成苗头抗体，加速了药物的研发流程；交通行业，AI大模型帮助交通道路规划、辅助自动驾驶等，例如智己汽车采用生成式大模型帮助汽车理解用户意图，打造人车路多模态交互方式；教育行业，AI大模型参与知识问答、内容创作、个性化教学等，例如猿辅导利用AI大模型辅助进行学情分析、阅读对话、计算解题等教育能力。

政策方面，自AI大模型研究热度升高以来，中国及其多个省份陆续发布相关政策、措施，以支持和监管AI大模型的研究与应用，助力AI大模型发展并有效发挥正向的价值。2023年3月，工业和信息化部等四部门印发《新产业标准化领航工程实施方案（2023—2035年）》指出前瞻布局生成式人工智能等未来产业标准研究。7月，国家互联网信息办公室等部门联合发布《生成式人工智能服务管理暂行办法》，提出国家坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励生成式人工智能创新发展，对生成式人工智能服务实行包容审慎和分类分级监管。北京市2023年发布《北京市促进通用人工智能创新发展的若干措施》，上海市发布《上海市推动人工智能大模型创新发展若干措施（2023-2025年）》，广东、浙江等省份均发布相关文件，鼓励企业结合地方优势开展AI大模型研究工作。

尽管目前中国AI大模型技术仍处于研发和迭代的初期阶段，模型的性能、价值还有待接受市场的检验。目前AI大模型的研究和应用探索已经显现广阔的发展前景，中国企业的探索和创新能力使得中国有望在这场AI大模型的科技竞争中保持长久的竞争力。

▲ 中国AI大模型发布情况

1.3 大模型建设方持续多元化发展，电信运营商走出“体系化”建设道路

在中国人工智能领域，除了互联网及云服务商之外，电信运营商正依托其丰富的算力、网络等基础设施以及多年的用户、数据基础，逐步发展互联网、云服务等业务，也在AI大模型领域持续投入，相继发布研究成果，成为AI大模型建设方的重要力量。

中国移动

中国移动在去年7月举办的“世界人工智能大会”上正式发布“九天”人工智能大模型，包括“九天·海算政务大模型”和“九天·客服大模型”。2024年5月，在中国移动人工智能生态大会上发布了“万千百”智能基座、AI+产品及应用、三大人工智能基地等AI+行动，表示中国移动已经具备万卡级智算集群、千亿多模态大模型以及较为完善的生态的平台，体现出中国移动在人工智能及AI大模型领域从软件到硬件建设的全面战略布局。同年7月，在2024世界人工智能大会暨人工智能全球治理高级别会议（简称：“2024世界人工智能大会”）上，中国移动继续表示要走出一条融入行业、开放共享的“体系化人工智能”发展道路。

中国电信

中国电信在去年11月举办的“数字科技生态大会”上发布了千亿参数星辰大模型。近年来，中国电信持续布局“1+N+M”星辰大模型产品体系，包括 1 个通用基础大模型、N 个行业大模型数量、M 个自用大模型，覆盖了语义、语音、视觉、多模态四大能力，并实现开源，面向政务、教育、交通等垂直领域推出12个行业大模型，面向网络运营、经营分析、代码研发等内部生产经营推出9个自用大模型。中国电信表示将持续坚持网是基础、云为核心、把握人工智能发展方向、创新产品和服务供给，以“网+云+AI+应用”满足千家万户、千行百业的数字化需求。在2024世界人工智能大会上，中国电信进一步表示要把握以人工智能为代表的新一轮科技革命和产业变革机遇，推动企业从传统电信运营商向服务型、科技型、安全型企业转型。

中国联通

中国联通在去年6月的上海世界移动通信大会（MWC）上发布了专为电信增值业务设计的垂直行业大模型“鸿湖图文大模型1.0”，初探大模型工程化和赋能应用的可行路径；2024年，中国联通在世界移动通信大会上发布了“1+1+M”联通元景大模型体系，包括1套基础大模型、1个大模型底座和M种行业大模型，借鉴动物智能演化规律、人类职业技能形成规律以及行业发展规律，布局中国联通人工智能体系。在2024年6月举办的世界移动通信大会上，中国联通继续表示积极落实“人工智能+”行动，推动AI与产业的深度融合。

▲ 中国AI大模型发布情况

1.3 大模型建设方持续多元化发展，电信运营商走出“体系化”建设道路

在中国人工智能领域，除了互联网及云服务商之外，电信运营商正依托其丰富的算力、网络等基础设施以及多年的用户、数据基础，逐步发展互联网、云服务等业务，也在AI大模型领域持续投入，相继发布研究成果，成为AI大模型建设方的重要力量。

中国移动

中国移动在去年7月举办的“世界人工智能大会”上正式发布“九天”人工智能大模型，包括“九天·海算政务大模型”和“九天·客服大模型”。2024年5月，在中国移动人工智能生态大会上发布了“万千百”智能基座、AI+产品及应用、三大人工智能基地等AI+行动，表示中国移动已经具备万卡级智算集群、千亿多模态大模型以及较为完善的生态的平台，体现出中国移动在人工智能及AI大模型领域从软件到硬件建设的全面战略布局。同年7月，在2024世界人工智能大会暨人工智能全球治理高级别会议（简称：“2024世界人工智能大会”）上，中国移动继续表示要走出一条融入行业、开放共享的“体系化人工智能”发展道路。

中国电信

中国电信在去年11月举办的“数字科技生态大会”上发布了千亿参数星辰大模型。近年来，中国电信持续布局“1+N+M”星辰大模型产品体系，包括 1 个通用基础大模型、N 个行业大模型数量、M 个自用大模型，覆盖了语义、语音、视觉、多模态四大能力，并实现开源，面向政务、教育、交通等垂直领域推出12个行业大模型，面向网络运营、经营分析、代码研发等内部生产经营推出9个自用大模型。中国电信表示将持续坚持网是基础、云为核心、把握人工智能发展方向、创新产品和服务供给，以“网+云+AI+应用”满足千家万户、千行百业的数字化需求。在2024世界人工智能大会上，中国电信进一步表示要把握以人工智能为代表的新一轮科技革命和产业变革机遇，推动企业从传统电信运营商向服务型、科技型、安全型企业转型。

中国联通

中国联通在去年6月的上海世界移动通信大会（MWC）上发布了专为电信增值业务设计的垂直行业大模型“鸿湖图文大模型1.0”，初探大模型工程化和赋能应用的可行路径；2024年，中国联通在世界移动通信大会上发布了“1+1+M”联通元景大模型体系，包括1套基础大模型、1个大模型底座和M种行业大模型，借鉴动物智能演化规律、人类职业技能形成规律以及行业发展规律，布局中国联通人工智能体系。在2024年6月举办的世界移动通信大会上，中国联通继续表示积极落实“人工智能+”行动，推动AI与产业的深度融合。

随着5G、云计算和大数据等技术的不断成熟与普及，电信运营商的AI大模型在智能客服、智慧城市、工业互联网等多个领域存在巨大的应用潜力，为其海量用户带来更加丰富、便捷、智能的服务体验。在这场智能化的浪潮中，电信运营商无疑将成为推动社会进步和科技创新的关键力量。

02

优势互补，电信运营商与云服务商在竞争中探索合作共赢新局面

多样产业角色入局共建 AI 大模型，带来的首要影响是市场竞争格局日益激烈。与众多新技术的发展路径类似，企业在扩张市场广度的同时进行市场深度的拓展，即通过技术合作、行业融合、产业协同等方式探索更多的应用领域、应用方式、社会价值。电信运营商与云服务商均已形成各自的 AI 大模型研究基础与研究体系，发布具备各自企业特征的 AI 大模型产品，在大模型发展的下一阶段，电信运营商与云服务商正探索如何通过合作有效调动起双方优势，打造合作共赢的新局面。

随着5G、云计算和大数据等技术的不断成熟与普及，电信运营商的AI大模型在智能客服、智慧城市、工业互联网等多个领域存在巨大的应用潜力，为其海量用户带来更加丰富、便捷、智能的服务体验。在这场智能化的浪潮中，电信运营商无疑将成为推动社会进步和科技创新的关键力量。

02

优势互补，电信运营商与云服务商在竞争中探索合作共赢新局面

多样产业角色入局共建 AI 大模型，带来的首要影响是市场竞争格局日益激烈。与众多新技术的发展路径类似，企业在扩张市场广度的同时进行市场深度的拓展，即通过技术合作、行业融合、产业协同等方式探索更多的应用领域、应用方式、社会价值。电信运营商与云服务商均已形成各自的 AI 大模型研究基础与研究体系，发布具备各自企业特征的 AI 大模型产品，在大模型发展的下一阶段，电信运营商与云服务商正探索如何通过合作有效调动起双方优势，打造合作共赢的新局面。

2.1 电信运营商 AI 大模型发挥通信业语料优势，用语音大模型打开市场

01 中国移动九天大模型

九天大模型是中国移动在AI领域的重要成果，具有安全可信、广泛支持技术创新产品等显著特点。中国移动当前已经基于九天大模型推出了二十余个行业大模型，覆盖政府治理、工业生产、民生服务和通信特色等领域。

其中，九天 · 客服大模型和九天 · 海算政务大模型已经在中国移动集团内外部客户中落地应用。例如，九天 · 网络大模型助力多省网络调参效率提升30%，性能问题工单效率提升80%以上；九天 · 客服大模型在中国移动10086在线客服场景规模化应用，是业界首个将大模型用于超大规模客服生产系统的工程化案例。

2024年7月，中国移动正式开源139亿参数语言大模型，开源内容包括模型权重、微调代码、推理代码，是业界首批直接面向行业构建的基础大模型之一。该模型采用纯解码结构+多专家的算法架构，训练数据中融合29个行业、占比达10%的行业数据。此外，中国移动还发布了AI+5G新通话，AI+办公等23个AI软硬产品，AI+工业、AI+交通等20个行业应用。

02 中国电信星辰大模型

- 星辰语义大模型

中国电信研发的星辰语义大模型利用1.5万亿Tokens的中英文高质量语料进行训练，并提出缓解多轮幻觉的解决方案，将“幻觉率”降低40%。2024年，中国电信接连开源7B、12B、52B的星辰语义大模型版本，并开放高质量清洗数据集。这一举措不仅降低了大模型开发门槛，还为国内AI生态的发展注入了新活力。

星辰大模型主要聚焦于行业应用场景，例如：在政务服务领域：通过智能分析和处理，提高政务服务的效率和质量；在智慧城市领域，为城市运营和管理提供决策支持、优化资源配置；在经营分析领域，帮助企业进行市场趋势分析、客户行为预测；在公文写作领域，长文写作任务的有效采纳率高达85.7%，极大提升了文档处理的效率。

- 星辰语音大模型

星辰语音大模型是中国电信推出的创新语音识别工具，基于无标签的30万小时多方言语音数据进行训练，旨在打破传统的语言识别局限，为各种地方方言提供精准的识别能力。该模型能够理解并识别包括粤语、上海话、四川话、温州话在内的30多种方言，开启了人工智能在多语言识别领域的全新篇章。

星辰语音大模型提供基础和大型两个版本的开源预训练模型，总参数量分别为0.09亿和0.3亿。该模型可应用在不同地区方言的用户沟通场景：智能家居领域，让智能设备能听懂不同地区的用户指令，提高用户体验；客服系统领域，提升多区域客户服务的响应效率和准确度；教育领域，促进跨地域的在线语言教学和学习；社交媒体领域，增强语音社交平台的语言包容性；医疗保健领域，辅助医疗咨询，优化普通话运用不熟练的用户群体体验。

目前，星辰语音大模型已在福建、江西、广西、北京、内蒙古等地的中国电信万号智能客服试点应用，实现日均处理约200万通电话，应对30余种方言。此外，星辰语音大模型还落地多地市的12345平台，助力政务工作智能化升级。

03 中国联通元景大模型

中国联通的元景大模型已经发展到2.0版本，实现了基座能力、MaaS平台、安全能力以及行业应用的四项能力升级，形成“更易定制、更懂行业、更加可信”的鲜明特色。

元景大模型已经发布2040亿参数的元景多模态大模型、元景文生图大模型、元景语音大模型三大基础模型。目前，万亿参数MoE大模型已基本训练完成，并积极探索从松耦合多模态到原生多模态的演进路线，提升模型在多模态对话时的流畅度。

应用落地方面，元景大模型已在政务、企业、个人与家庭等应用场景实现“人工智能+”，形成了35+行业大模型和100余个标杆应用，并在诸多场景助力提质增效。例如，辽宁12345省级平台的全部坐席已全面接入政务热线大模型，显著提升了咨询业务处理的效率和准确率。

结合电信运营商已发布的AI大模型来看，其建设与应用呈现以下特点：

① 运营商作为国家智算建设的排头兵、主要建设方，拥有可服务各行各业的广泛智算基础设施。

② 电信运营商在通信行业拥有巨大的语料优势。例如中国电信基于多种方言的客服通话训练的星辰语音大模型，充分体现数据的核心竞争力；中国移动的九天网络大模型，则是基于网络运维积累的大量数据进行训练。

③ 缺乏普适性的C端应用入口，电信运营商在C端大模型领域市场竞争力相对较弱，但对于视频彩铃、5G通话等通信行业特征显著的业务领域，运营商已有较多积累与试验。

④ 基于在行业DICT集成领域的绝对优势，电信运营商大模型均主打行业大模型应用。一方面积极构建面向不同行业的基础大模型，另一方面积极打造MaaS平台，通过生态组件的组合满足行业客户的需要。

2.1 电信运营商 AI 大模型发挥通信业语料优势，用语音大模型打开市场

01 中国移动九天大模型

九天大模型是中国移动在AI领域的重要成果，具有安全可信、广泛支持技术创新产品等显著特点。中国移动当前已经基于九天大模型推出了二十余个行业大模型，覆盖政府治理、工业生产、民生服务和通信特色等领域。

其中，九天 · 客服大模型和九天 · 海算政务大模型已经在中国移动集团内外部客户中落地应用。例如，九天 · 网络大模型助力多省网络调参效率提升30%，性能问题工单效率提升80%以上；九天 · 客服大模型在中国移动10086在线客服场景规模化应用，是业界首个将大模型用于超大规模客服生产系统的工程化案例。

2024年7月，中国移动正式开源139亿参数语言大模型，开源内容包括模型权重、微调代码、推理代码，是业界首批直接面向行业构建的基础大模型之一。该模型采用纯解码结构+多专家的算法架构，训练数据中融合29个行业、占比达10%的行业数据。此外，中国移动还发布了AI+5G新通话，AI+办公等23个AI软硬产品，AI+工业、AI+交通等20个行业应用。

02 中国电信星辰大模型

- 星辰语义大模型

中国电信研发的星辰语义大模型利用1.5万亿Tokens的中英文高质量语料进行训练，并提出缓解多轮幻觉的解决方案，将“幻觉率”降低40%。2024年，中国电信接连开源7B、12B、52B的星辰语义大模型版本，并开放高质量清洗数据集。这一举措不仅降低了大模型开发门槛，还为国内AI生态的发展注入了新活力。

星辰大模型主要聚焦于行业应用场景，例如：在政务服务领域：通过智能分析和处理，提高政务服务的效率和质量；在智慧城市领域，为城市运营和管理提供决策支持、优化资源配置；在经营分析领域，帮助企业进行市场趋势分析、客户行为预测；在公文写作领域，长文写作任务的有效采纳率高达85.7%，极大提升了文档处理的效率。

- 星辰语音大模型

星辰语音大模型是中国电信推出的创新语音识别工具，基于无标签的30万小时多方言语音数据进行训练，旨在打破传统的语言识别局限，为各种地方方言提供精准的识别能力。该模型能够理解并识别包括粤语、上海话、四川话、温州话在内的30多种方言，开启了人工智能在多语言识别领域的全新篇章。

星辰语音大模型提供基础和大型两个版本的开源预训练模型，总参数量分别为0.09亿和0.3亿。该模型可应用在不同地区方言的用户沟通场景：智能家居领域，让智能设备能听懂不同地区的用户指令，提高用户体验；客服系统领域，提升多区域客户服务的响应效率和准确度；教育领域，促进跨地域的在线语言教学和学习；社交媒体领域，增强语音社交平台的语言包容性；医疗保健领域，辅助医疗咨询，优化普通话运用不熟练的用户群体体验。

目前，星辰语音大模型已在福建、江西、广西、北京、内蒙古等地的中国电信万号智能客服试点应用，实现日均处理约200万通电话，应对30余种方言。此外，星辰语音大模型还落地多地市的12345平台，助力政务工作智能化升级。

03 中国联通元景大模型

中国联通的元景大模型已经发展到2.0版本，实现了基座能力、MaaS平台、安全能力以及行业应用的四项能力升级，形成“更易定制、更懂行业、更加可信”的鲜明特色。

元景大模型已经发布2040亿参数的元景多模态大模型、元景文生图大模型、元景语音大模型三大基础模型。目前，万亿参数MoE大模型已基本训练完成，并积极探索从松耦合多模态到原生多模态的演进路线，提升模型在多模态对话时的流畅度。

应用落地方面，元景大模型已在政务、企业、个人与家庭等应用场景实现“人工智能+”，形成了35+行业大模型和100余个标杆应用，并在诸多场景助力提质增效。例如，辽宁12345省级平台的全部坐席已全面接入政务热线大模型，显著提升了咨询业务处理的效率和准确率。

结合电信运营商已发布的AI大模型来看，其建设与应用呈现以下特点：

① 运营商作为国家智算建设的排头兵、主要建设方，拥有可服务各行各业的广泛智算基础设施。

② 电信运营商在通信行业拥有巨大的语料优势。例如中国电信基于多种方言的客服通话训练的星辰语音大模型，充分体现数据的核心竞争力；中国移动的九天网络大模型，则是基于网络运维积累的大量数据进行训练。

③ 缺乏普适性的C端应用入口，电信运营商在C端大模型领域市场竞争力相对较弱，但对于视频彩铃、5G通话等通信行业特征显著的业务领域，运营商已有较多积累与试验。

④ 基于在行业DICT集成领域的绝对优势，电信运营商大模型均主打行业大模型应用。一方面积极构建面向不同行业的基础大模型，另一方面积极打造MaaS平台，通过生态组件的组合满足行业客户的需要。

2.2 云服务商 AI 大模型发展发挥快速迭代落地优势，积累丰富市场反馈

01 腾讯

腾讯混元大语言模型由腾讯云主要进行研发与技术支撑，在多个NLP基准测试中取得了优异的成绩。此外，混元开源中文DiT文生图模型，在人像真实感和场景真实感方面展现显著的优势。目前混元大模型已经在腾讯内部的600多个业务场景中进行测试：

- 腾讯文档基于混元大模型支持了数十种文本创作场景，如一键生成标准格式文本、自然语言生成函数以及基于表格内容生成图表等功能，极大地提高了用户的工作效率和创作体验。
- 腾讯会议基于混元大模型推出AI小助手，与会者可以通过简单的自然语言指令完成会议信息提取、内容分析、会议纪要整理等工作。AI小助手还可以根据会议讨论的内容生成会议总结，自动提取关键点。
- 在游戏领域，混元大模型被用于创建智能的游戏NPC，提供更丰富的交互体验。游戏中的NPC可以根据玩家的行为和选择做出不同的反应，参与更复杂的故事情节发展，增强游戏的真实感和沉浸感。
- 腾讯广告平台利用混元大模型推出了一站式的AI广告创意平台“腾讯广告妙思”，为广告主提供快速生成创意素材的能力，实现全链路流程自动化。通过智能分析和个性化推荐，混元大模型帮助广告主提高广告的点击率和转化率，降低广告投放成本。

目前，腾讯通过腾讯云平台对外开放混元大模型的API服务接口，同时积极探索混元大模型在教育、医疗、金融等行业的应用场景。

02 阿里巴巴

阿里云通义千问大语言模型已经深入应用于阿里巴巴的电商、金融、物流等数字生态系统中。通义千问被集成到智能客服系统中以提供个性化服务，还用于自动生成商品描述和支持跨境电商的多语言翻译，推动了业务的创新和效率提升。阿里巴巴集团利用通义千问在智能客服、文本摘要、内容创作和机器翻译等多个领域取得了显著成果，极大提升了服务效率和用户满意度。

阿里巴巴已将通义千问的部分模型进行了开源，包括参数量从5亿到1,100亿不等的大模型，在2024年上半年的SuperCLUE报告中，通义千问的开源模型Qwen2-72B成为排名第一的中国大模型，也是全球最强的开源模型之一，引领全球的开源生态。

阿里巴巴通过“百炼”平台支持企业用户接入通义千问，为企业提供定制化的解决方案。目前，阿里巴巴集团正在持续扩展通义千问在企业级市场的应用，并积极探索其在国际市场的应用潜力，旨在服务全球客户。例如，通义千问赋能奥运史上首个大模型应用，用于比赛解说场景。

03 百度

百度智能云推出的文心一言是百度集团的大型预训练语言模型。除了通过APP提供服务以外，文心一言大模型还覆盖了百度的众多产品和服务生态：文心一言已经接入百度搜索，为客户提供更为丰富、更具交互性的搜索结果，目前已有10%的搜索流量通过文心一言的模型生成；文心一言已经通过百度智能云对外开放文心大模型4.0 Turbo的API服务。

相对电信运营商，云服务商AI大模型的建设更加注重开源能力，同时发挥其快速研发、快速论证的能力，具有以下特征：

- 云服务商AI大模型注重开源贡献，例如腾讯开源了MimicMotion、ToonCrafter、Aniportrait、SEED-Story、GFPGAN、Photo-Maker等众多大模型相关组件，字节跳动则有AnimateDiff-Lightning、Depth Anything、PuLID、MTVQA等开源贡献。
- 云服务商各自拥有不同领域的语料，因此大模型的优势能力有所不同。例如腾讯混元在网文创作方面的能力优势、字节豆包在娱乐领域的能力优势、百度文心一言在搜索领域的能力优势。
- 云服务商大模型的重点场景是赋能自有的互联网应用，目前均有成规模的C端大模型服务。例如腾讯会议、豆包APP、百度搜索等都是与大模型相结合的典型应用。
- 互联网企业在大模型行业应用方面均有所规划，但当前成果的丰富度不如C端应用。当前主要聚焦于广告创作、代码助手、智能客服、教育培训等领域，仍需持续扩展行业应用的深度和广度。

总体来看，电信运营商倾向于发挥智算、网络等基建优势、项目集成优势和行业影响力优势，从行业应用入手发展大模型应用，同时注重参与相关技术、应用的标准化建设；云服务商倾向于借助资金和技术优势快速迭代和优化模型，嵌入既有业务生态系统快速实现服务集成与市场验证，借助市场反馈探索发展方向，同时从C端及小B端应用切入拉动行业应用的发展。

2.3 “1+3+N”合作体系，云服务商全面助力电信运营商发展 AI 大模型

结合云服务商与电信运营商的优势领域分析，二者在行业大模型领域具有更加广阔的合作空间。

技术优势互补

电信运营商可以依托云服务商的技术积累来增强AI大模型能力；云服务商可以借助运营商的智算及网络资源来满足遍布全国的行业客户需求。

行业应用互补

在某些特定行业中，如智慧城市、工业互联网等，运营商和云服务商可以联手打造综合解决方案，合作开发面向特定行业的垂直应用，比如智能客服、智能制造等。

生态力量互补

双方可以通过共建开放平台的方式，促进技术和服务的相互融合；共同推进行业标准的制订，促进大模型的标准化和规范化发展。

根据电信运营商和云服务商的优势能力，二者的合作可以形成“1+3+N”体系；

1—联合打造1个适合AI大模型培育的算力集群。

2.2 云服务商 AI 大模型发展发挥快速迭代落地优势，积累丰富市场反馈

01 腾讯

腾讯混元大语言模型由腾讯云主要进行研发与技术支撑，在多个NLP基准测试中取得了优异的成绩。此外，混元开源中文DiT文生图模型，在人像真实感和场景真实感方面展现显著的优势。目前混元大模型已经在腾讯内部的600多个业务场景中进行测试：

- 腾讯文档基于混元大模型支持了数十种文本创作场景，如一键生成标准格式文本、自然语言生成函数以及基于表格内容生成图表等功能，极大地提高了用户的工作效率和创作体验。
- 腾讯会议基于混元大模型推出AI小助手，与会者可以通过简单的自然语言指令完成会议信息提取、内容分析、会议纪要整理等工作。AI小助手还可以根据会议讨论的内容生成会议总结，自动提取关键点。
- 在游戏领域，混元大模型被用于创建智能的游戏NPC，提供更丰富的交互体验。游戏中的NPC可以根据玩家的行为和选择做出不同的反应，参与更复杂的故事情节发展，增强游戏的真实感和沉浸感。
- 腾讯广告平台利用混元大模型推出了一站式的AI广告创意平台“腾讯广告妙思”，为广告主提供快速生成创意素材的能力，实现全链路流程自动化。通过智能分析和个性化推荐，混元大模型帮助广告主提高广告的点击率和转化率，降低广告投放成本。

目前，腾讯通过腾讯云平台对外开放混元大模型的API服务接口，同时积极探索混元大模型在教育、医疗、金融等行业的应用场景。

02 阿里巴巴

阿里云通义千问大语言模型已经深入应用于阿里巴巴的电商、金融、物流等数字生态系统中。通义千问被集成到智能客服系统中以提供个性化服务，还用于自动生成商品描述和支持跨境电商的多语言翻译，推动了业务的创新和效率提升。阿里巴巴集团利用通义千问在智能客服、文本摘要、内容创作和机器翻译等多个领域取得了显著成果，极大提升了服务效率和用户满意度。

阿里巴巴已将通义千问的部分模型进行了开源，包括参数量从5亿到1,100亿不等的大模型，在2024年上半年的SuperCLUE报告中，通义千问的开源模型Qwen2-72B成为排名第一的中国大模型，也是全球最强的开源模型之一，引领全球的开源生态。

阿里巴巴通过“百炼”平台支持企业用户接入通义千问，为企业提供定制化的解决方案。目前，阿里巴巴集团正在持续扩展通义千问在企业级市场的应用，并积极探索其在国际市场的应用潜力，旨在服务全球客户。例如，通义千问赋能奥运史上首个大模型应用，用于比赛解说场景。

03 百度

百度智能云推出的文心一言是百度集团的大型预训练语言模型。除了通过APP提供服务以外，文心一言大模型还覆盖了百度的众多产品和服务生态：文心一言已经接入百度搜索，为客户提供更为丰富、更具交互性的搜索结果，目前已有10%的搜索流量通过文心一言的模型生成；文心一言已经通过百度智能云对外开放文心大模型4.0 Turbo的API服务。

相对电信运营商，云服务商AI大模型的建设更加注重开源能力，同时发挥其快速研发、快速论证的能力，具有以下特征：

- 云服务商AI大模型注重开源贡献，例如腾讯开源了MimicMotion、ToonCrafter、Aniportrait、SEED-Story、GFPGAN、Photo-Maker等众多大模型相关组件，字节跳动则有AnimateDiff-Lightning、Depth Anything、PuLID、MTVQA等开源贡献。
- 云服务商各自拥有不同领域的语料，因此大模型的优势能力有所不同。例如腾讯混元在网文创作方面的能力优势、字节豆包在娱乐领域的能力优势、百度文心一言在搜索领域的能力优势。
- 云服务商大模型的重点场景是赋能自有的互联网应用，目前均有成规模的C端大模型服务。例如腾讯会议、豆包APP、百度搜索等都是与大模型相结合的典型应用。
- 互联网企业在大模型行业应用方面均有所规划，但当前成果的丰富度不如C端应用。当前主要聚焦于广告创作、代码助手、智能客服、教育培训等领域，仍需持续扩展行业应用的深度和广度。

总体来看，电信运营商倾向于发挥智算、网络等基建优势、项目集成优势和行业影响力优势，从行业应用入手发展大模型应用，同时注重参与相关技术、应用的标准化建设；云服务商倾向于借助资金和技术优势快速迭代和优化模型，嵌入既有业务生态系统快速实现服务集成与市场验证，借助市场反馈探索发展方向，同时从C端及小B端应用切入拉动行业应用的发展。

2.3 “1+3+N”合作体系，云服务商全面助力电信运营商发展 AI 大模型

结合云服务商与电信运营商的优势领域分析，二者在行业大模型领域具有更加广阔的合作空间。

技术优势互补

电信运营商可以依托云服务商的技术积累来增强AI大模型能力；云服务商可以借助运营商的智算及网络资源来满足遍布全国的行业客户需求。

行业应用互补

在某些特定行业中，如智慧城市、工业互联网等，运营商和云服务商可以联手打造综合解决方案，合作开发面向特定行业的垂直应用，比如智能客服、智能制造等。

生态力量互补

双方可以通过共建开放平台的方式，促进技术和服务的相互融合；共同推进行业标准的制订，促进大模型的标准化和规范化发展。

根据电信运营商和云服务商的优势能力，二者的合作可以形成“1+3+N”体系；

1—联合打造1个适合AI大模型培育的算力集群。

3—形成3条AI大模型研发合作路线：

- 标准软件路线：企业直接采购开箱即用的软件完成落地，例如企业直接购买代码助手、文生图工具、办公助手等应用，开箱即用。
- 标准模型能力增强路线：企业需要参与提示词工程等环节来增强模型能力，优化模型输出结果。例如企业基于私有化部署的大模型、RAG技术搭建私有化知识引擎构建起企业自有的、数据不外泄的智能客服系统、经营分析系统。
- 定制化模型精调训练路线：企业需要完成二次模型训练工作。例如基于行业基础大模型进行精调，训练出企业专属的金融大模型、出行大模型、医药大模型等。

N—形成N个场景化落地解决方案。

从AI大模型设计、建设到应用，全流程支持电信运营商AI大模型发展。

03

一集群三路线，云服务商助力电信运营商进行软硬兼备的AI大模型建设储备

AI 大模型的出现对算力需求带来了指数级的增长。OpenAI 发布的 GPT-3 模型包含 1750 亿个参数，需要进行数千万次的计算操作来完成一次推理任务。如何将训练好的模型推广与应用，并最终赋能社会生产及产生社会价值，需要在硬件建设及软件设计阶段就做好准备。

从硬件来看，算力集群能力的高低决定了 AI 大模型能否在合理的时间内完成训练，决定了模型投入市场的周期以及它们在实际应用中的性能表现。从软件来看，软件建设和设计路线，决定了软件未来的应用范围和使用体验。云服务商从自身建设经验出发，结合电信运营商面临的挑战及优势能力，支持运营商打造适用于 AI 大模型训练的算力集群，形成助力电信运营商完成 AI 大模型设计、应用的软件研发合作路线。

3—形成3条AI大模型研发合作路线：

- 标准软件路线：企业直接采购开箱即用的软件完成落地，例如企业直接购买代码助手、文生图工具、办公助手等应用，开箱即用。
- 标准模型能力增强路线：企业需要参与提示词工程等环节来增强模型能力，优化模型输出结果。例如企业基于私有化部署的大模型、RAG技术搭建私有化知识引擎构建起企业自有的、数据不外泄的智能客服系统、经营分析系统。
- 定制化模型精调训练路线：企业需要完成二次模型训练工作。例如基于行业基础大模型进行精调，训练出企业专属的金融大模型、出行大模型、医药大模型等。

N—形成N个场景化落地解决方案。

从AI大模型设计、建设到应用，全流程支持电信运营商AI大模型发展。

03

一集群三路线，云服务商助力电信运营商进行软硬兼备的AI大模型建设储备

AI 大模型的出现对算力需求带来了指数级的增长。OpenAI 发布的 GPT-3 模型包含 1750 亿个参数，需要进行数千万次的计算操作来完成一次推理任务。如何将训练好的模型推广与应用，并最终赋能社会生产及产生社会价值，需要在硬件建设及软件设计阶段就做好准备。

从硬件来看，算力集群能力的高低决定了 AI 大模型能否在合理的时间内完成训练，决定了模型投入市场的周期以及它们在实际应用中的性能表现。从软件来看，软件建设和设计路线，决定了软件未来的应用范围和使用体验。云服务商从自身建设经验出发，结合电信运营商面临的挑战及优势能力，支持运营商打造适用于 AI 大模型训练的算力集群，形成助力电信运营商完成 AI 大模型设计、应用的软件研发合作路线。

3.1 云服务商支持高效算力集群建设

AI大模型时代，模型参数的指数级增长，远超遵循摩尔定律的硬件增长速度，在训练成本、数据需求、训练时间、模型训练方式等方面都出现了显著的变化。

训练成本上升

随着模型参数的增加，训练成本也显著上升。一方面，更多的计算资源意味着更高的硬件成本和能源消耗。另一方面，更大的模型需要更多的存储空间来保存模型参数，这也增加了存储成本。此外，训练更大的模型通常需要更多的时间，这也意味着需要支付更多的计算资源使用费用。因此，训练成本的上升成为了制约深度学习模型发展的一个重要因素。

数据需求的增加

更大的模型通常需要更多的数据来训练，以避免过拟合。这是因为更大的模型具有更强的拟合能力，如果没有足够的数据来训练，模型可能会学习到数据中的噪声，而不是数据的内在规律。因此，数据需求的增加不仅增加了数据收集的成本，也增加了数据清洗和预处理的复杂性。在某些领域，如医疗、金融等，获取大量高质量的数据是非常困难的，这进一步加剧了数据需求的问题。

训练时间的延长

随着模型参数的增加和数据集的扩大，训练时间也显著延长。训练一个大型深度学习模型可能需要几天甚至几周的时间。这种长时间的训练过程不仅延缓了模型的迭代和实验的速度，也增加了研究人员的等待时间，降低了研究效率。此外，长时间的训练过程也增加了模型训练过程中出现意外情况的风险，如硬件故障、数据错误等。

模型训练的分布式和并行化

为了应对模型规模的增长，采用模型训练的分布式和并行化技术。通过将模型训练任务分配到多个计算节点上，可以有效地利用更多的计算资源，加速模型的训练过程。然而，分布式和并行化训练也带来了一些新的挑战，如通信开销、负载均衡、模型同步等问题。

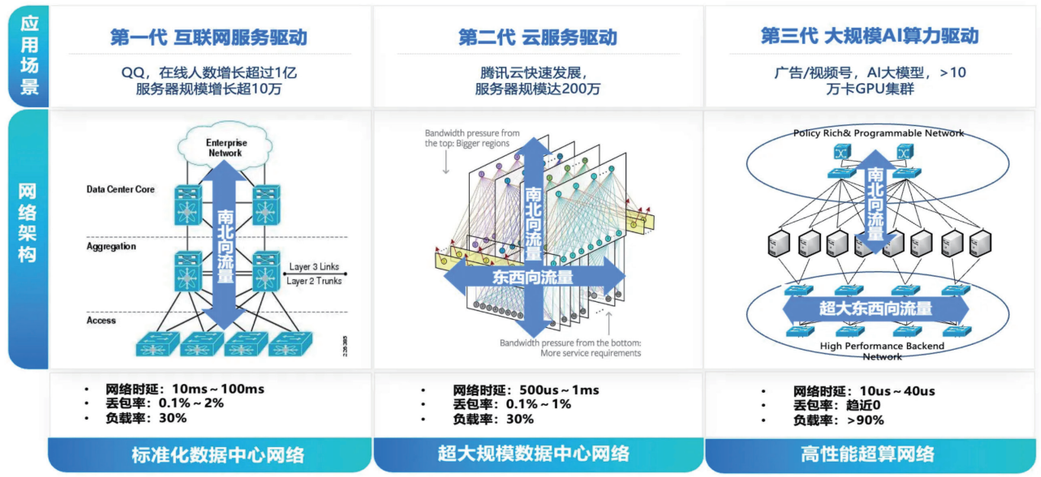
上述问题的出现为AI大模型算力集群的建设带来了众多挑战。

3.1.1 算力集群建设与发展面临的挑战

算力集群发展面临的挑战主要体现在网络、存储和协同三个方面。

01 网络方面，算力集群规模的快速扩张，对网络通信提出了更高的要求。

数据中心网络从传统互联网服务向智算服务演进。基于数据中心网络服务的应用以及网络流量的主要特征，发展大致可以划分为以下几个阶段：



▲ 数据中心网络演进

第一代由互联网服务驱动。以QQ等互联网应用为代表的DCN建设，网络流量特征体现为用户与服务器之间的南北流量占主导；这一阶段主要使用了商用网络设备，搭建标准化数据中心网络。

第二代由云服务驱动。随着大数据和云计算的兴起，服务器之间的东西向流量逐渐增多，云租户对网络产生了虚拟化和隔离的要求。数据中心网络架构逐渐演变为同时承载南北向和东西向流量的云网络架构。

第三代由大规模算力驱动。随着AI大模型的出现，采用高性能计算网络，东西向、南北向流量的分离架构。构建了独立的超大带宽、符合AI训练流量特征的网络架构，满足超强算力对网络性能的新需求。

AI大模型对网络性能需求：大带宽、高负载、零丢包的无损网络。

AIGC的火爆带来AI大模型参数量从亿级到万亿级的飙升。为支撑海量数据的大规模训练，大量服务器通过高速网络组成算力集群，互联互通，共同完成训练任务。

大集群不等于大算力，相反，GPU集群越大，产生的额外通信损耗越多。大带宽、高利用率、信息无损，是AI大模型时代网络面临的重要挑战。

3.1 云服务商支持高效算力集群建设

AI大模型时代，模型参数的指数级增长，远超遵循摩尔定律的硬件增长速度，在训练成本、数据需求、训练时间、模型训练方式等方面都出现了显著的变化。

训练成本上升

随着模型参数的增加，训练成本也显著上升。一方面，更多的计算资源意味着更高的硬件成本和能源消耗。另一方面，更大的模型需要更多的存储空间来保存模型参数，这也增加了存储成本。此外，训练更大的模型通常需要更多的时间，这也意味着需要支付更多的计算资源使用费用。因此，训练成本的上升成为了制约深度学习模型发展的一个重要因素。

数据需求的增加

更大的模型通常需要更多的数据来训练，以避免过拟合。这是因为更大的模型具有更强的拟合能力，如果没有足够的数据来训练，模型可能会学习到数据中的噪声，而不是数据的内在规律。因此，数据需求的增加不仅增加了数据收集的成本，也增加了数据清洗和预处理的复杂性。在某些领域，如医疗、金融等，获取大量高质量的数据是非常困难的，这进一步加剧了数据需求的问题。

训练时间的延长

随着模型参数的增加和数据集的扩大，训练时间也显著延长。训练一个大型深度学习模型可能需要几天甚至几周的时间。这种长时间的训练过程不仅延缓了模型的迭代和实验的速度，也增加了研究人员的等待时间，降低了研究效率。此外，长时间的训练过程也增加了模型训练过程中出现意外情况的风险，如硬件故障、数据错误等。

模型训练的分布式和并行化

为了应对模型规模的增长，采用模型训练的分布式和并行化技术。通过将模型训练任务分配到多个计算节点上，可以有效地利用更多的计算资源，加速模型的训练过程。然而，分布式和并行化训练也带来了一些新的挑战，如通信开销、负载均衡、模型同步等问题。

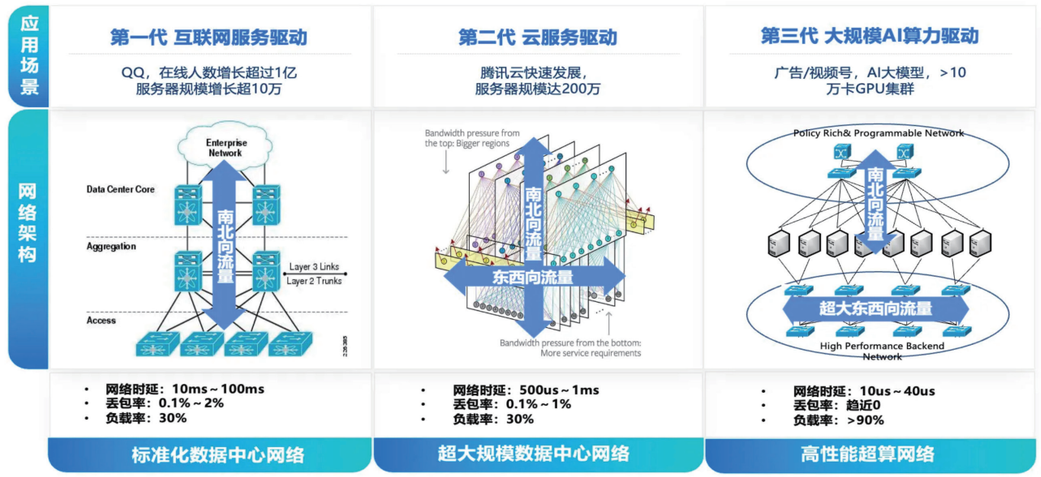
上述问题的出现为AI大模型算力集群的建设带来了众多挑战。

3.1.1 算力集群建设与发展面临的挑战

算力集群发展面临的挑战主要体现在网络、存储和协同三个方面。

01 网络方面，算力集群规模的快速扩张，对网络通信提出了更高的要求。

数据中心网络从传统互联网服务向智算服务演进。基于数据中心网络服务的应用以及网络流量的主要特征，发展大致可以划分为以下几个阶段：



▲ 数据中心网络演进

第一代由互联网服务驱动。以QQ等互联网应用为代表的DCN建设，网络流量特征体现为用户与服务器之间的南北流量占主导；这一阶段主要使用了商用网络设备，搭建标准化数据中心网络。

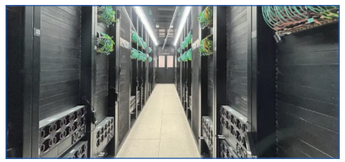
第二代由云服务驱动。随着大数据和云计算的兴起，服务器之间的东西向流量逐渐增多，云租户对网络产生了虚拟化和隔离的要求。数据中心网络架构逐渐演变为同时承载南北向和东西向流量的云网络架构。

第三代由大规模算力驱动。随着AI大模型的出现，采用高性能计算网络，东西向、南北向流量的分离架构。构建了独立的超大带宽、符合AI训练流量特征的网络架构，满足超强算力对网络性能的新需求。

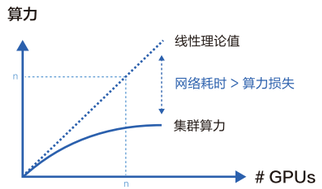
AI大模型对网络性能需求：大带宽、高负载、零丢包的无损网络。

AIGC的火爆带来AI大模型参数量从亿级到万亿级的飙升。为支撑海量数据的大规模训练，大量服务器通过高速网络组成算力集群，互联互通，共同完成训练任务。

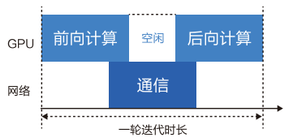
大集群不等于大算力，相反，GPU集群越大，产生的额外通信损耗越多。大带宽、高利用率、信息无损，是AI大模型时代网络面临的重要挑战。



大集群 ≠ 大算力

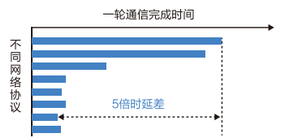


后向计算



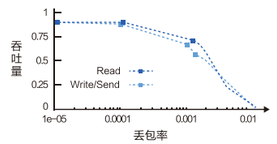
训练过程中通信占比最大可达50%

时延影响



不同网络架构的时延差可达5倍

时延影响



0.1%丢包会造成50%算力损失

▲ 大集群和大算力的对比

千亿、万亿参数规模的大模型，训练过程中通信占比最大可达50%，传统低速网络的带宽远远无法支撑。同时，传统网络协议容易导致网络拥塞、高延时和丢包，而仅0.1%的网络丢包就可能导致30%−50%的算力损失，最终造成算力资源的严重浪费。AI大模型对网络规模、性能、可靠性、成本、运营能力要求方面提出了更高的要求。

02 存储方面，全链路的海量数据处理效率影响大模型的训练效率。

在大模型场景下，对存储的吞吐量、时延、可靠性、成本等，提出了更高的要求，主要包括：

数据吞吐量需求高：大模型训练需要处理和存储大量的数据。这要求存储系统具有高吞吐量，以满足快速数据访问的需求。

低延迟访问：模型训练过程中，参数更新和梯度计算需要快速响应。存储系统的访问延迟直接影响到训练的效率。

数据一致性：在分布式训练环境中，需要确保不同计算节点访问的数据是一致的，避免因数据不一致导致训练错误。

可扩展性：随着模型规模的增长，存储需求也在不断增加。存储系统需要具备良好的可扩展性，以适应不断增长的数据量。

存储成本：存储大量数据需要昂贵的存储硬件和维护成本，如何在保证性能的同时控制成本是一个问题。

备份和恢复：大模型训练过程中需要定期保存模型的状态（Checkpointing），以便在发生故障时能够快速恢复训练状态。这要求存储系统支持高效的数据备份和恢复机制。

智能化存储：利用AI技术优化存储资源的分配和数据管理，提高存储效率和降低成本。

解决这些挑战需要综合考虑存储硬件、软件架构、数据管理策略以及与计算资源的协同设计。随着技术的发展，需要有新的存储解决方案和优化技术的不断涌现，以应对大模型带来的存储挑战。

03 协同方面，大集群的生产效率遵循“木桶效应”，易形成短板。

大集群需要各种类型的计算单元、存储单元、网络单元的高效协同，任何一个短板都将成为集群效率的瓶颈。

多级存储访存带宽不一致。在大模型训练中，激活值、梯度位于GPU中，模型的FP16/FP32参数、优化器状态位于CPU中甚至位于SSD中，模型的前向和反向在GPU上进行运算，而参数更新在CPU做运算，这就需要频繁的进行内存显存以及SSD之间的访问，多级存储访问带宽的不一致很容易导致硬件资源闲置，如何减少硬件资源的闲置时间是大模型训练优化的一大挑战。

模型状态冗余存储。大模型训练时的模型状态存储于CPU中，在模型训练过程中会不断拷贝到GPU，这就导致模型状态同时存储于CPU和GPU中，这种冗余存储是对本就捉肘见襟的单机存储空间是一种严重浪费。

内存碎片过多。大模型拥有巨量的模型状态，单张GPU卡不能完全放置所有模型状态，在训练过程中模型状态被顺序在CPU和GPU之间交换，这种交换导致GPU显存的频繁分配和释放，此外大模型训练过程中海量的Activation也需要频繁分配和释放显存，显存频繁分配和释放会产生大量的显存碎片。

带宽利用率低。在大模型分布式训练中多机之间会进行各类参数通信，单机内部存在模型状态的数据传输。通信以及数据传输带宽利用率低是训练框架在分布式训练中最常见的问题。

3.1.2 构建高效算力集群的关键技术

云服务商通过在高性能智算网络、高性能存储及高效协同框架三个方面提供支持，针对性的解决电信运营商算力集群建设面临的问题。

高性能智算网络底座：让大模型训练的网络通信“交通顺畅”。

当前在大模型算力集群中普遍采用IB（InfiniBand）网络和RoCE（RDMA over Converged Ethernet）网络两种技术路线。随着集群需求和规模的扩大，算力集群的组网能力已经从设备级往业务级的运营运维演进。

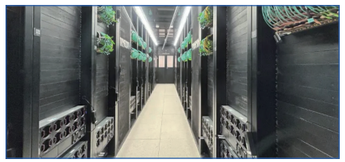
InfiniBand网络以其高性能、低延迟和高可靠性，在高性能计算组网中得到广泛应用。其容易上手，但技术体系封闭，主要基于网络设备级能力，不利于深度运营。

RoCE依赖于各厂商提供的技术方案成熟度，随着越来越多的头部互联网及AI厂商越来越多采用基于以太网络的业务级运营运维路线构建RDMA网络，在业界的最佳实践中其在性能上得到有效的保证。另一方面，其进一步扩展了网络底座能力的边界，已从单纯网络设备级的运营运维能力延伸到云网端协同，再进一步扩展到业务级的运营运维能力。比如，Meta在最新的Llama3大模型训练环境采用以太网协议RoCEv2，而腾讯的混元大模型基于RoCEv2构建的智能高性能网络，在高效训练中实现业务0中断。

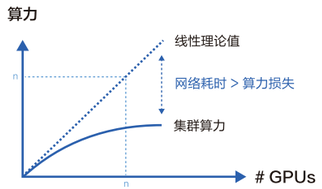
01 构建大模型训练新一代算力网络系统的技术要求

面向AI大模型训练需要构建的新一代算力网络系统，基于新的网络架构、协议栈、运营需要，构建为大模型、多租户提供更大规模、更高性能、更加可靠的算力网络底座。向上支撑AI大模型训练框架和业务，向下将分散的计算、存储资源高效地整合互通，从而为算力集群提供更好的网络底座。

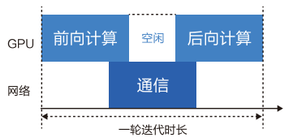
随着千亿甚至万亿大模型的上线，算力集群的规模从千卡发展到万卡级别，对网络性能提出了更高的要求：



大集群 ≠ 大算力

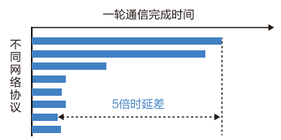


后向计算



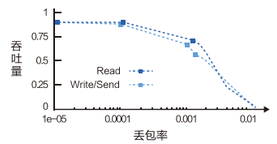
训练过程中通信占比最大可达50%

时延影响



不同网络架构的时延差可达5倍

时延影响



0.1%丢包会造成50%算力损失

▲ 大集群和大算力的对比

千亿、万亿参数规模的大模型，训练过程中通信占比最大可达50%，传统低速网络的带宽远远无法支撑。同时，传统网络协议容易导致网络拥塞、高延时和丢包，而仅0.1%的网络丢包就可能导致30%−50%的算力损失，最终造成算力资源的严重浪费。AI大模型对网络规模、性能、可靠性、成本、运营能力要求方面提出了更高的要求。

02 存储方面，全链路的海量数据处理效率影响大模型的训练效率。

在大模型场景下，对存储的吞吐量、时延、可靠性、成本等，提出了更高的要求，主要包括：

数据吞吐量需求高：大模型训练需要处理和存储大量的数据。这要求存储系统具有高吞吐量，以满足快速数据访问的需求。

低延迟访问：模型训练过程中，参数更新和梯度计算需要快速响应。存储系统的访问延迟直接影响到训练的效率。

数据一致性：在分布式训练环境中，需要确保不同计算节点访问的数据是一致的，避免因数据不一致导致训练错误。

可扩展性：随着模型规模的增长，存储需求也在不断增加。存储系统需要具备良好的可扩展性，以适应不断增长的数据量。

存储成本：存储大量数据需要昂贵的存储硬件和维护成本，如何在保证性能的同时控制成本是一个问题。

备份和恢复：大模型训练过程中需要定期保存模型的状态（Checkpointing），以便在发生故障时能够快速恢复训练状态。这要求存储系统支持高效的数据备份和恢复机制。

智能化存储：利用AI技术优化存储资源的分配和数据管理，提高存储效率和降低成本。

解决这些挑战需要综合考虑存储硬件、软件架构、数据管理策略以及与计算资源的协同设计。随着技术的发展，需要有新的存储解决方案和优化技术的不断涌现，以应对大模型带来的存储挑战。

03 协同方面，大集群的生产效率遵循“木桶效应”，易形成短板。

大集群需要各种类型的计算单元、存储单元、网络单元的高效协同，任何一个短板都将成为集群效率的瓶颈。

多级存储访存带宽不一致。在大模型训练中，激活值、梯度位于GPU中，模型的FP16/FP32参数、优化器状态位于CPU中甚至位于SSD中，模型的前向和反向在GPU上进行运算，而参数更新在CPU做运算，这就需要频繁的进行内存显存以及SSD之间的访问，多级存储访问带宽的不一致很容易导致硬件资源闲置，如何减少硬件资源的闲置时间是大模型训练优化的一大挑战。

模型状态冗余存储。大模型训练时的模型状态存储于CPU中，在模型训练过程中会不断拷贝到GPU，这就导致模型状态同时存储于CPU和GPU中，这种冗余存储是对本就捉肘见襟的单机存储空间是一种严重浪费。

内存碎片过多。大模型拥有巨量的模型状态，单张GPU卡不能完全放置所有模型状态，在训练过程中模型状态被顺序在CPU和GPU之间交换，这种交换导致GPU显存的频繁分配和释放，此外大模型训练过程中海量的Activation也需要频繁分配和释放显存，显存频繁分配和释放会产生大量的显存碎片。

带宽利用率低。在大模型分布式训练中多机之间会进行各类参数通信，单机内部存在模型状态的数据传输。通信以及数据传输带宽利用率低是训练框架在分布式训练中最常见的问题。

3.1.2 构建高效算力集群的关键技术

云服务商通过在高性能智算网络、高性能存储及高效协同框架三个方面提供支持，针对性的解决电信运营商算力集群建设面临的问题。

高性能智算网络底座：让大模型训练的网络通信“交通顺畅”。

当前在大模型算力集群中普遍采用IB（InfiniBand）网络和RoCE（RDMA over Converged Ethernet）网络两种技术路线。随着集群需求和规模的扩大，算力集群的组网能力已经从设备级往业务级的运营运维演进。

InfiniBand网络以其高性能、低延迟和高可靠性，在高性能计算组网中得到广泛应用。其容易上手，但技术体系封闭，主要基于网络设备级能力，不利于深度运营。

RoCE依赖于各厂商提供的技术方案成熟度，随着越来越多的头部互联网及AI厂商越来越多采用基于以太网络的业务级运营运维路线构建RDMA网络，在业界的最佳实践中其在性能上得到有效的保证。另一方面，其进一步扩展了网络底座能力的边界，已从单纯网络设备级的运营运维能力延伸到云网端协同，再进一步扩展到业务级的运营运维能力。比如，Meta在最新的Llama3大模型训练环境采用以太网协议RoCEv2，而腾讯的混元大模型基于RoCEv2构建的智能高性能网络，在高效训练中实现业务0中断。

01 构建大模型训练新一代算力网络系统的技术要求

面向AI大模型训练需要构建的新一代算力网络系统，基于新的网络架构、协议栈、运营需要，构建为大模型、多租户提供更大规模、更高性能、更加可靠的算力网络底座。向上支撑AI大模型训练框架和业务，向下将分散的计算、存储资源高效地整合互通，从而为算力集群提供更好的网络底座。

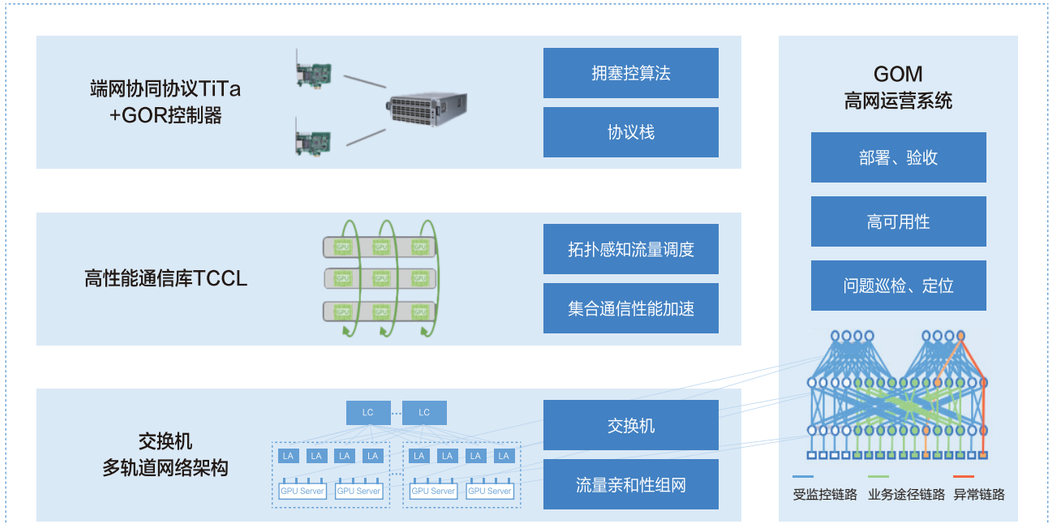
随着千亿甚至万亿大模型的上线，算力集群的规模从千卡发展到万卡级别，对网络性能提出了更高的要求：

大规模，高带宽：集群规模几千到几万卡，单卡接入速率400G。支持不同训练任务混跑、不同GPU卡的混部。而传统DCN 100G接入，单POD规模1.5k。

高性能，90%负载下零丢包：GPU训练是瞬时吞吐90%，且通信对丢包敏感，0.1%丢包损失30%～50%算力，需要网络做到传输无损。传统DCN利用率<40%，丢包率0.1%～1%。

高可用，零中断：一旦网络中断，任务重启需要约1.5小时，需要减少网络中断。

02 大模型训练新一代算力网络系统框架



▲ 大模型训练新一代算力网络系统框架

大模型训练新一代算力网络系统在大规模组网上，采用无阻塞胖树（Fat-Tree）拓扑，分为Block-Pod-Cluster三级。

- Block是最小单元，包括256个GPU。
- Pod是典型集群规模，包括16~64个Block，即4096~16384个GPU。
- 多个Block可以组成Cluster。1个Cluster最大支持16个Pod，即1个cluster可支持65536~262144个GPU。最高26万个GPU的规模能够满足当前训练需求。

在流量调控上，已经出现相关产品对通信流量路径进行优化，引入了“多轨道流量聚合架构”，提升集群的通信效率。

在可编程与协议上，借助拥塞算法，例如端网协议栈TiTa的拥塞控制算法，实时监测并调整网络拥塞，实现高负载下的零丢包，集群通信效率达90%以上。

在软件加速上，通过高性能通信库TCCL的网络拓扑感知能力实现GPU之间通信路径的最优规划，可减少50%~80%流量绕路，可最大化释放底层的网络硬件能力。

AI大模型GPU集群强稳定性，要求高性能超算网络业务级运营运维能力，在运营系统上，提供全局规划与视图、训练任务分析、业务问题诊断、拥塞调度消除等能力，通过自动化部署与验收，高精度网络测量与监控，将交付周期缩短50%；提供了毫秒级时延度量，分钟级自愈运营系统，出现故障时可1分钟发现、3分钟定位、5分钟自愈，实现数千卡级规模在整个训练周期（数十天）无网络问题导致的训练中断。

通过以上的优化，可在集合通信性能方面，AllReduce性能的网络负载率达到90%以上，将大模型网络训练中的通信占比降到6%左右，通过拓扑亲和性，跨LA流量通信占比减少75%，相比于IB和RoCE网络，对AI大模型的计算需求进行针对性的优化与适配。

高性能存储：既要快、又要省。

随着多模态技术的进化和落地应用的逐渐爆发，从语言和图像为主的GPT，到视频生成模型Sora，大模型参数正在指数级增长。Sora为主的多模态技术，更会让需要处理的数据量急剧增加，而这才刚刚是视频生成模型的1.0时代。

01 构建大模型训练新一代存储底座的技术要求

参数越大，对云存储的需求就会越高，包括云存储的数据量以及吞吐量等，如果云存储能力不能够满足大模型的需求，则会直接影响到大模型的训练速度和推理效率。

要保持大模型的更新频率和速度，就要保证整个大模型数据训练过程的高效，其中某一个环节出现问题，就可能会拉长整个训练时长，增加训练成本。因此，作为整个大模型数据训练的数据底座，云存储的重要性日益凸显。

存储作为数据的载体，现如今已经不仅仅只承担“存”的作用，更需要打通数据从“存”到“用”的最后一公里。

在AI大模型数据训练的4个环节中，存储需要提供的具体能力，包括：

- 数据采集阶段，需要一个大容量、低成本、高可靠的数据存储底座。
- 数据清洗阶段，需要提供更多协议的支持，以及至少GB甚至TB级的数据访问性能。
- 数据训练阶段，作为大模型训练的关键环节，则需要一个TB级的带宽存储保证训练过程中Checkpoint能够快速保存，以便于保障训练的连续性和提升CPU的有效使用时长，也需要存储提供百万级IOPS能力，来保证训练时海量小样本读取不会成为训练瓶颈。
- 数据应用阶段，则需要存储提供比较丰富的数据审核能力，来满足鉴黄、鉴暴等安全合规的诉求，保证大模型生成的内容以合法、合规的方式使用。

02 大模型训练新一代存储底座框架

基于以上4个阶段，大模型训练新一代存储底座框架提供包括对象存储、高性能并行文件存储、数据加速器和数据智能处理服务，满足不同场景下的存储需求，可将大模型的数据清洗和训练效率提升1倍。

在数据采集环节，采用可支持单集群管理百EB级别存储规模的对象存储服务，借助数据加速器提升数据访问性能，获取高达数TBps的读取带宽，提供亚毫秒级的数据访问延迟、百万级的IOPS和TBps级别的吞吐能力。

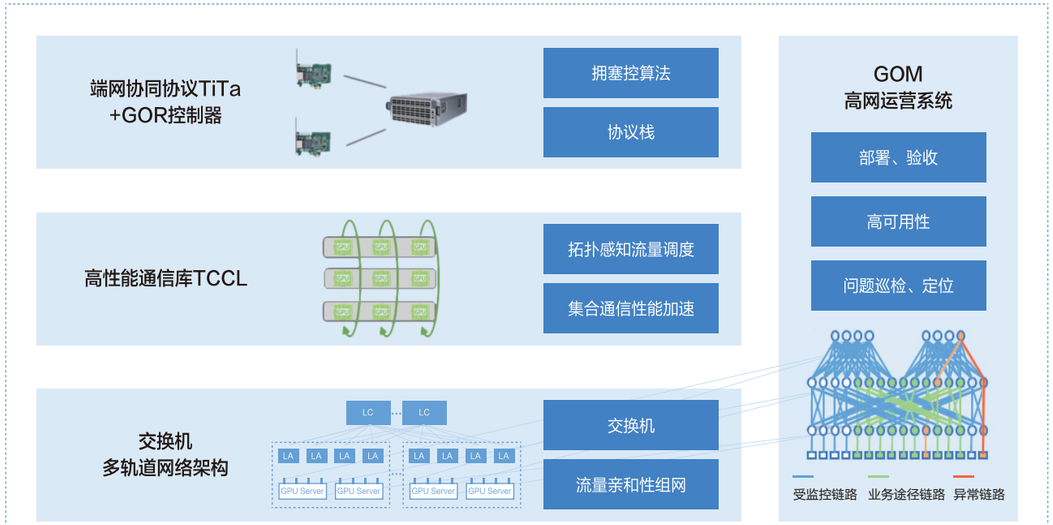
大模型的训练则更加耗时，短则数周、长则数月，这期间如果任何一个CPU/GPU的节点掉线，都会导致整个训练前功尽弃。此时则需要将保存的Checkpoint时间缩短到越短越好，但数千上万个节点都需要保存Checkpoint，这就对文件存储的读写吞吐提出了非常高的要求。

大规模，高带宽：集群规模几千到几万卡，单卡接入速率400G。支持不同训练任务混跑、不同GPU卡的混部。而传统DCN 100G接入，单POD规模1.5k。

高性能，90%负载下零丢包：GPU训练是瞬时吞吐90%，且通信对丢包敏感，0.1%丢包损失30%～50%算力，需要网络做到传输无损。传统DCN利用率<40%，丢包率0.1%～1%。

高可用，零中断：一旦网络中断，任务重启需要约1.5小时，需要减少网络中断。

02 大模型训练新一代算力网络系统框架



▲ 大模型训练新一代算力网络系统框架

大模型训练新一代算力网络系统在大规模组网上，采用无阻塞胖树（Fat-Tree）拓扑，分为Block-Pod-Cluster三级。

- Block是最小单元，包括256个GPU。
- Pod是典型集群规模，包括16~64个Block，即4096~16384个GPU。
- 多个Block可以组成Cluster。1个Cluster最大支持16个Pod，即1个cluster可支持65536~262144个GPU。最高26万个GPU的规模能够满足当前训练需求。

在流量调控上，已经出现相关产品对通信流量路径进行优化，引入了“多轨道流量聚合架构”，提升集群的通信效率。

在可编程与协议上，借助拥塞算法，例如端网协议栈TiTa的拥塞控制算法，实时监测并调整网络拥塞，实现高负载下的零丢包，集群通信效率达90%以上。

在软件加速上，通过高性能通信库TCCL的网络拓扑感知能力实现GPU之间通信路径的最优规划，可减少50%~80%流量绕路，可最大化释放底层的网络硬件能力。

AI大模型GPU集群强稳定性，要求高性能超算网络业务级运营运维能力，在运营系统上，提供全局规划与视图、训练任务分析、业务问题诊断、拥塞调度消除等能力，通过自动化部署与验收，高精度网络测量与监控，将交付周期缩短50%；提供了毫秒级时延度量，分钟级自愈运营系统，出现故障时可1分钟发现、3分钟定位、5分钟自愈，实现数千卡级规模在整个训练周期（数十天）无网络问题导致的训练中断。

通过以上的优化，可在集合通信性能方面，AllReduce性能的网络负载率达到90%以上，将大模型网络训练中的通信占比降到6%左右，通过拓扑亲和性，跨LA流量通信占比减少75%，相比于IB和RoCE网络，对AI大模型的计算需求进行针对性的优化与适配。

高性能存储：既要快、又要省。

随着多模态技术的进化和落地应用的逐渐爆发，从语言和图像为主的GPT，到视频生成模型Sora，大模型参数正在指数级增长。Sora为主的多模态技术，更会让需要处理的数据量急剧增加，而这才刚刚是视频生成模型的1.0时代。

01 构建大模型训练新一代存储底座的技术要求

参数越大，对云存储的需求就会越高，包括云存储的数据量以及吞吐量等，如果云存储能力不能够满足大模型的需求，则会直接影响到大模型的训练速度和推理效率。

要保持大模型的更新频率和速度，就要保证整个大模型数据训练过程的高效，其中某一个环节出现问题，就可能会拉长整个训练时长，增加训练成本。因此，作为整个大模型数据训练的数据底座，云存储的重要性日益凸显。

存储作为数据的载体，现如今已经不仅仅只承担“存”的作用，更需要打通数据从“存”到“用”的最后一公里。

在AI大模型数据训练的4个环节中，存储需要提供的具体能力，包括：

- 数据采集阶段，需要一个大容量、低成本、高可靠的数据存储底座。
- 数据清洗阶段，需要提供更多协议的支持，以及至少GB甚至TB级的数据访问性能。
- 数据训练阶段，作为大模型训练的关键环节，则需要一个TB级的带宽存储保证训练过程中Checkpoint能够快速保存，以便于保障训练的连续性和提升CPU的有效使用时长，也需要存储提供百万级IOPS能力，来保证训练时海量小样本读取不会成为训练瓶颈。
- 数据应用阶段，则需要存储提供比较丰富的数据审核能力，来满足鉴黄、鉴暴等安全合规的诉求，保证大模型生成的内容以合法、合规的方式使用。

02 大模型训练新一代存储底座框架

基于以上4个阶段，大模型训练新一代存储底座框架提供包括对象存储、高性能并行文件存储、数据加速器和数据智能处理服务，满足不同场景下的存储需求，可将大模型的数据清洗和训练效率提升1倍。

在数据采集环节，采用可支持单集群管理百EB级别存储规模的对象存储服务，借助数据加速器提升数据访问性能，获取高达数TBps的读取带宽，提供亚毫秒级的数据访问延迟、百万级的IOPS和TBps级别的吞吐能力。

大模型的训练则更加耗时，短则数周、长则数月，这期间如果任何一个CPU/GPU的节点掉线，都会导致整个训练前功尽弃。此时则需要将保存的Checkpoint时间缩短到越短越好，但数千上万个节点都需要保存Checkpoint，这就对文件存储的读写吞吐提出了非常高的要求。

采用高性能并行文件存储服务，将读写吞吐能力从100GB直接升级至TiB/s级别，让3TB checkpoint 写入时间从10分钟，缩短至10秒内，时间降低90%，大幅提升大模型训练效率。目前已有服务可实现将元数据目录打散至所有存储节点上，提供线性扩张能力，从而实现文件打开、读取、删除的百万级IOPS能力。

应用阶段，大模型推理场景则对数据安全与可追溯性提出更高要求。数据智能处理服务例如一站式内容治理服务平台，可以对AI生成的内容进行一站式管理，提供图片隐式水印、AIGC内容审核、智能数据检索MetaInsight等能力。

基于大模型训练新一代存储底座，可有效提升大模型的训练、推理效率。

高效协同框架：优化大模型集群生产框架，消除训练短板。

万亿模型的模型训练仅装下一组参数和优化器状态便需要1.7TB以上的存储空间，这还不包括训练过程中产生的激活值所需的存储。在这样的背景下，大模型训练不仅受限于海量的算力,更受限于巨大的存储需求。如何以更少的算力更高效的训练大模型，除了基于IaaS底座性能的优化，同时需要进一步优化训练框架和集群的协作能力。

01 大模型训练框架优化的技术要求

为了以最小的成本训练大模型，需要将模型的参数、梯度、优化器状态以模型并行的方式切分到所有GPU，引入了显存内存统一存储视角，将存储容量的上界由内存扩容到内存+显存总和，有效的扩容和利用存储空间。

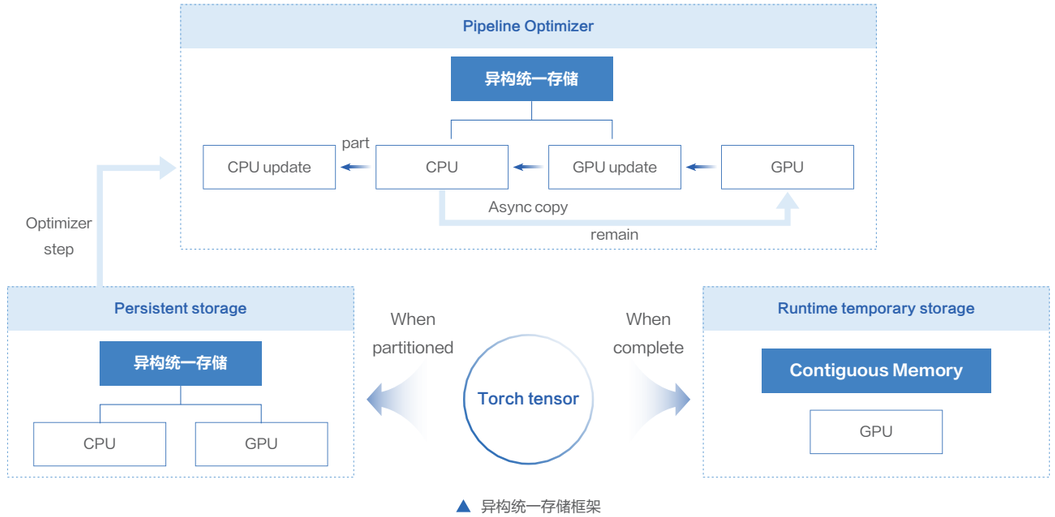
同时采用多流异步化，在GPU计算的同时进行数据IO和NCCL通信，使用异构流水线均衡设备间的负载，最大化提升整个系统的吞吐。

02 大模型训练框架优化

ZeRO-Cache是一款超大规模模型训练的利器，其通过统一视角去管理内存和显存，在去除模型状态冗余的同时扩大单个机器的可用存储空间上限，通过Contiguous Memory显存管理器管理模型参数的显存分配/释放进而减少显存碎片，通过多流均衡各个硬件资源的负载，通过引入SSD进一步扩展单机模型容量。

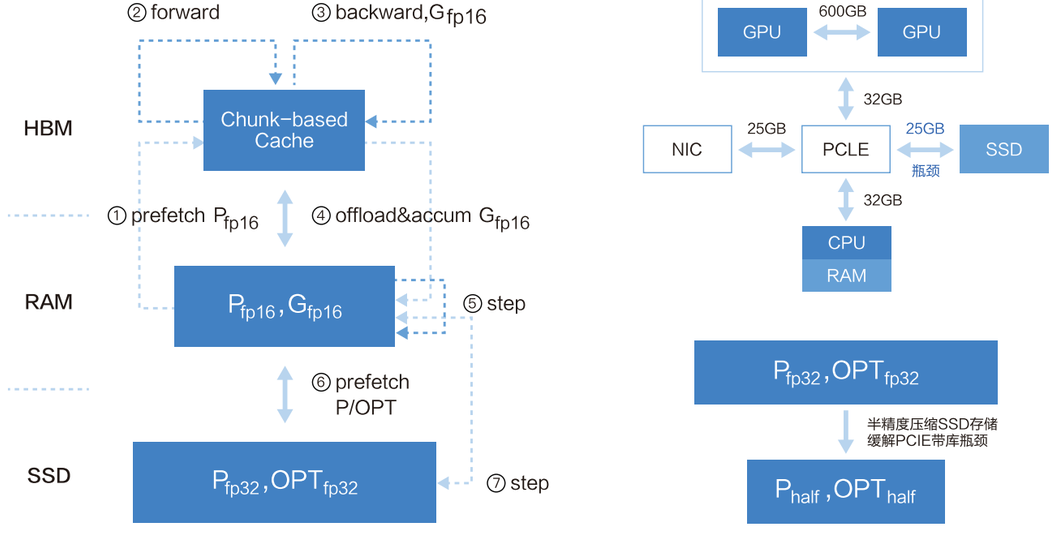
大模型训练时模型状态都位于CPU内存中，在训练时会拷贝到GPU显存，这就导致模型状态的冗余存储(CPU和GPU同时存在一份)，此外大模型的训练中会用到大量的pin memory，pin memory的使用会提升性能同时会导致物理内存的大量浪费，如何科学合理的使用pin memory是ZeRO-Cache着重要解决的问题。

ZeRO-Cache本着极致化去冗余的理念引入了chunk对内存和显存进行管理，保证所有模型状态只存储一份通常模型会存储在内存或显存上，ZeRO-Cache提出异构统一存储，采用内存和显存共同作为存储空间，击破了异构存储的壁垒，极大扩充了模型存储可用空间。



▲ 异构统一存储框架

为了更加低成本的扩展模型参数，ZeRO-Cache进一步引入了SSD作为三级存储，针对GPU高计算吞吐、高通信带宽和SSD低PCIE带宽之间的差异，ZeRO-Cache将所有fp16参数和梯度移动到内存中，让forward和backward的计算不受SSD低带宽影响，同时通过对优化器状态做半精度压缩来缓解SSD读写对性能的影响。



▲ ZeRO-Cache SSD三级存储框架

大模型训练过程中有大量的计算和通信，包括GPU计算、H2D和D2H单机通信、NCCL多机通信等，涉及的硬件有GPU、CPU、PCIE等。ZeRO-Cache为了最大化的利用硬件，多流异步化GPU计算、H2D和D2H单机通信、NCCL多机通信，参数预取采用用时同步机制，梯度后处理采用多buffer机制，优化器状态拷贝采用多流机制。

采用高性能并行文件存储服务，将读写吞吐能力从100GB直接升级至TiB/s级别，让3TB checkpoint 写入时间从10分钟，缩短至10秒内，时间降低90%，大幅提升大模型训练效率。目前已有服务可实现将元数据目录打散至所有存储节点上，提供线性扩张能力，从而实现文件打开、读取、删除的百万级IOPS能力。

应用阶段，大模型推理场景则对数据安全与可追溯性提出更高要求。数据智能处理服务例如一站式内容治理服务平台，可以对AI生成的内容进行一站式管理，提供图片隐式水印、AIGC内容审核、智能数据检索MetaInsight等能力。

基于大模型训练新一代存储底座，可有效提升大模型的训练、推理效率。

高效协同框架：优化大模型集群生产框架，消除训练短板。

万亿模型的模型训练仅装下一组参数和优化器状态便需要1.7TB以上的存储空间，这还不包括训练过程中产生的激活值所需的存储。在这样的背景下，大模型训练不仅受限于海量的算力,更受限于巨大的存储需求。如何以更少的算力更高效的训练大模型，除了基于IaaS底座性能的优化，同时需要进一步优化训练框架和集群的协作能力。

01 大模型训练框架优化的技术要求

为了以最小的成本训练大模型，需要将模型的参数、梯度、优化器状态以模型并行的方式切分到所有GPU，引入了显存内存统一存储视角，将存储容量的上界由内存扩容到内存+显存总和，有效的扩容和利用存储空间。

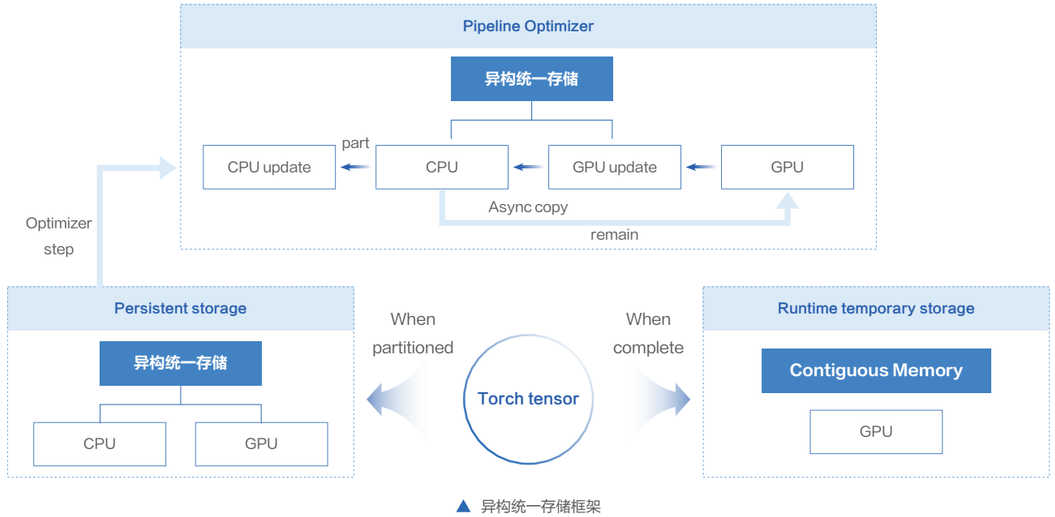
同时采用多流异步化，在GPU计算的同时进行数据IO和NCCL通信，使用异构流水线均衡设备间的负载，最大化提升整个系统的吞吐。

02 大模型训练框架优化

ZeRO-Cache是一款超大规模模型训练的利器，其通过统一视角去管理内存和显存，在去除模型状态冗余的同时扩大单个机器的可用存储空间上限，通过Contiguous Memory显存管理器管理模型参数的显存分配/释放进而减少显存碎片，通过多流均衡各个硬件资源的负载，通过引入SSD进一步扩展单机模型容量。

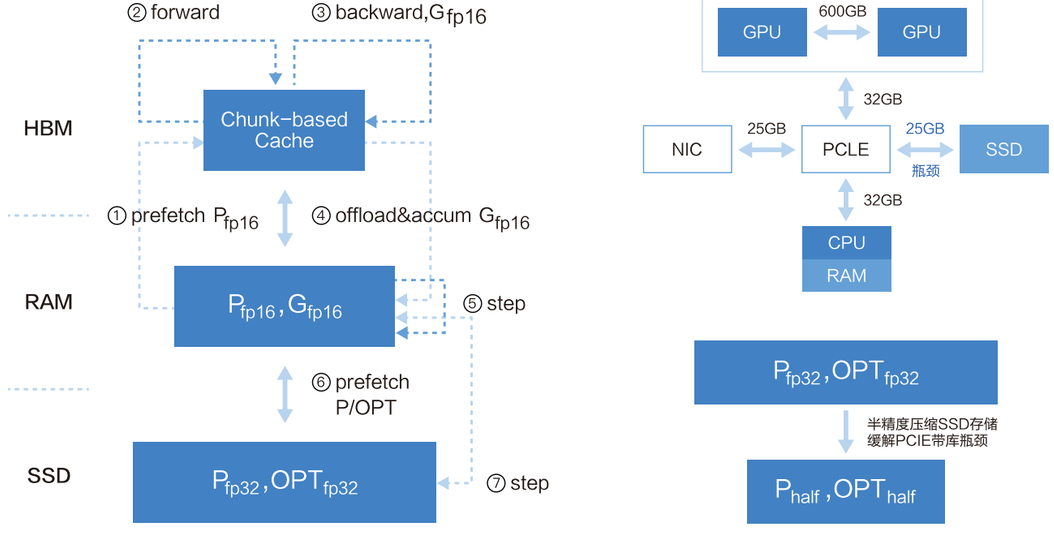
大模型训练时模型状态都位于CPU内存中，在训练时会拷贝到GPU显存，这就导致模型状态的冗余存储(CPU和GPU同时存在一份)，此外大模型的训练中会用到大量的pin memory，pin memory的使用会提升性能同时会导致物理内存的大量浪费，如何科学合理的使用pin memory是ZeRO-Cache着重要解决的问题。

ZeRO-Cache本着极致化去冗余的理念引入了chunk对内存和显存进行管理，保证所有模型状态只存储一份通常模型会存储在内存或显存上，ZeRO-Cache提出异构统一存储，采用内存和显存共同作为存储空间，击破了异构存储的壁垒，极大扩充了模型存储可用空间。



▲ 异构统一存储框架

为了更加低成本的扩展模型参数，ZeRO-Cache进一步引入了SSD作为三级存储，针对GPU高计算吞吐、高通信带宽和SSD低PCIE带宽之间的差异，ZeRO-Cache将所有fp16参数和梯度移动到内存中，让forward和backward的计算不受SSD低带宽影响，同时通过对优化器状态做半精度压缩来缓解SSD读写对性能的影响。



▲ ZeRO-Cache SSD三级存储框架

大模型训练过程中有大量的计算和通信，包括GPU计算、H2D和D2H单机通信、NCCL多机通信等，涉及的硬件有GPU、CPU、PCIE等。ZeRO-Cache为了最大化的利用硬件，多流异步化GPU计算、H2D和D2H单机通信、NCCL多机通信，参数预取采用用时同步机制，梯度后处理采用多buffer机制，优化器状态拷贝采用多流机制。

3.2 云服务商打造三大软件合作路线

根据电信运营商的优势能力、业务发展情况以及具体应用场景，云服务商可提供的支持可总结为三大合作路线：

▲ 云服务商三大软件合作路线

3.2.1 行业智算云+标准化应用：合力推广开箱即用的大模型软件

标准软件落地路线是指企业直接采购基于生成式AI模型能力增强的成熟应用软件。该路线下，生成式AI内嵌于应用之中，可以开箱即用。

该场景需求呈现出应用成熟、通用程度高、安全要求高等特点，应用多搭载于公有云/行业云之上。此路线以模型服务商为主导，大模型应用服务商提供开箱即用的应用，全权负责从底层基础设施到上层应用建设的全部职责。

电信运营商拥有遍布全国各地的行业云资源，可以提供底层智算基础设施及行业云平台，与企业转型、企业宽带结合提供就近服务；云服务商具有标准化产品的开发及运维优势，可以将成熟的AI Apps部署到运营商的属地行业云平台。双方强强联合，构成端到端的应用方案，共同满足属地企业客户的需要。

代码生成、会议纪要等通用程度高的场景适合采用该路线。标准软件通常具有较好的稳定性和可靠性，无需进行定制化开发，支持开箱即用快速落地，可以帮助电信运营商高性价比快速落地生成式AI应用。

3.2.2 私有云集成+标准化组件：合力承建私域化的知识增强型应用

标准模型能力增强路线是指通过提示词设计、检索增强生成等提示词工程，增强基础模型输出的准确性、知识实时性。检索增强生成可以通过加载组织/垂直知识数据，改善模型输入的提示词准确性。该路线实施时，会将生成式AI模型能力封装为API服务接口，企业将API嵌入自己的应用软件增强智能化水平，或是基于API创建定制化的全新智能应用。该落地路线的场景需求具有私域知识引入、研发资源与预算投入适中、算力资源投入低、数据安全要求高等特点，有一定的私有化部署及

定制化开发需求。大模型应用服务商需要配合完成提示工程与检索增强生成、应用定制开发等工作，并帮助企业完成基础模型构建、模型精调训练等工作。

电信运营商具有丰富的DICT项目集成经验，可以承担整体项目的私有云建设、定制开发等工作；云服务商可根据项目需要提供标准化组件（如机器学习平台、向量数据库等），并提供提示工程与检索增强生成、基础模型构建、模型精调训练等方面的技术支持。

知识助手类场景适合采用该路线。借助生成式AI的知识助手，可提升各个岗位的人效，通过问答形式为每位员工提供赋能，员工的个人经验页可以持续沉淀在知识助手上，进而提升公司的整体经验。

3.2.3 项目总集成+智算技术底座：合力支持按需定制的客户大模型

定制化模型精调训练落地路线，是指基于基础模型，企业通过私域数据集进行模型训练，以实现企业专属大模型的建设。企业专属大模型改变了基础模型参数，并且在特定任务上表现出来的性能和在企业知识的深度理解两方面得以增强。

该落地路线的场景需求呈现出高度定制化、资源和预算投入高等特点，对专属训练数据集、算法工程师团队、大量的智算资源、高性能的智算底座都有要求。若企业缺乏主导项目落地的能力，则需要一个强大的实施集成商来帮助企业完成全流程的项目建设。

电信运营商的DICT集成能力可在此类项目得到最大程度的发挥，一方面帮助企业构建起项目所需的、包括硬件、网络、软件在内的私有云/行业云智算平台；另一方面配合企业完成数据集准备、模型精调、数据检索、提示工程、应用建设等一系列工作。云服务商则可以基于自身经验按需输出一系列技术能力，例如帮助构建智算中心的高性能网络和高性能存储、帮助构建大模型训练环境的预制化行业大模型和训练平台等，从而大幅提升企业专属大模型的训练效率。

在面对检索增强的模型输出依然无法达到企业准确性和实时性要求，或者企业需要将生成式AI应用于业务逻辑极其复杂的场景、而基础模型完全无法理解等情景时，企业需要考虑采用定制化模型精调训练落地路线，定制企业专属大模型。

3.3 电信运营商和云服务商的融合共建价值

3.3.1 同质化的硬件堆叠难以保证竞争中的优势

在当今快速发展的人工智能领域，智算集群作为支撑大模型运行的关键基础设施，其建设与优化显得尤为重要。然而，目前市场上智算集群的同质化现象严重，这不仅限制了技术的进步，也难以在激烈的市场竞争中保持优势。

为了突破这一瓶颈，我们需要提高智算集群的资源利用率和可用性是保证竞争优势的关键。这要求我们在集群建管用上进行创新，采用高性能的网络底座、存储底座，高效的训练框架，以及更先进的运营运维体系，实现资源的高效分配和充分利用。同时，还需要加强集群的容错能力和自愈能力，确保在面对各种挑战时，集群能够稳定运行，提供可靠的服务。

3.2 云服务商打造三大软件合作路线

根据电信运营商的优势能力、业务发展情况以及具体应用场景，云服务商可提供的支持可总结为三大合作路线：

▲ 云服务商三大软件合作路线

3.2.1 行业智算云+标准化应用：合力推广开箱即用的大模型软件

标准软件落地路线是指企业直接采购基于生成式AI模型能力增强的成熟应用软件。该路线下，生成式AI内嵌于应用之中，可以开箱即用。

该场景需求呈现出应用成熟、通用程度高、安全要求高等特点，应用多搭载于公有云/行业云之上。此路线以模型服务商为主导，大模型应用服务商提供开箱即用的应用，全权负责从底层基础设施到上层应用建设的全部职责。

电信运营商拥有遍布全国各地的行业云资源，可以提供底层智算基础设施及行业云平台，与企业转型、企业宽带结合提供就近服务；云服务商具有标准化产品的开发及运维优势，可以将成熟的AI Apps部署到运营商的属地行业云平台。双方强强联合，构成端到端的应用方案，共同满足属地企业客户的需要。

代码生成、会议纪要等通用程度高的场景适合采用该路线。标准软件通常具有较好的稳定性和可靠性，无需进行定制化开发，支持开箱即用快速落地，可以帮助电信运营商高性价比快速落地生成式AI应用。

3.2.2 私有云集成+标准化组件：合力承建私域化的知识增强型应用

标准模型能力增强路线是指通过提示词设计、检索增强生成等提示词工程，增强基础模型输出的准确性、知识实时性。检索增强生成可以通过加载组织/垂直知识数据，改善模型输入的提示词准确性。该路线实施时，会将生成式AI模型能力封装为API服务接口，企业将API嵌入自己的应用软件增强智能化水平，或是基于API创建定制化的全新智能应用。该落地路线的场景需求具有私域知识引入、研发资源与预算投入适中、算力资源投入低、数据安全要求高等特点，有一定的私有化部署及

定制化开发需求。大模型应用服务商需要配合完成提示工程与检索增强生成、应用定制开发等工作，并帮助企业完成基础模型构建、模型精调训练等工作。

电信运营商具有丰富的DICT项目集成经验，可以承担整体项目的私有云建设、定制开发等工作；云服务商可根据项目需要提供标准化组件（如机器学习平台、向量数据库等），并提供提示工程与检索增强生成、基础模型构建、模型精调训练等方面的技术支持。

知识助手类场景适合采用该路线。借助生成式AI的知识助手，可提升各个岗位的人效，通过问答形式为每位员工提供赋能，员工的个人经验页可以持续沉淀在知识助手上，进而提升公司的整体经验。

3.2.3 项目总集成+智算技术底座：合力支持按需定制的客户大模型

定制化模型精调训练落地路线，是指基于基础模型，企业通过私域数据集进行模型训练，以实现企业专属大模型的建设。企业专属大模型改变了基础模型参数，并且在特定任务上表现出来的性能和在企业知识的深度理解两方面得以增强。

该落地路线的场景需求呈现出高度定制化、资源和预算投入高等特点，对专属训练数据集、算法工程师团队、大量的智算资源、高性能的智算底座都有要求。若企业缺乏主导项目落地的能力，则需要一个强大的实施集成商来帮助企业完成全流程的项目建设。

电信运营商的DICT集成能力可在此类项目得到最大程度的发挥，一方面帮助企业构建起项目所需的、包括硬件、网络、软件在内的私有云/行业云智算平台；另一方面配合企业完成数据集准备、模型精调、数据检索、提示工程、应用建设等一系列工作。云服务商则可以基于自身经验按需输出一系列技术能力，例如帮助构建智算中心的高性能网络和高性能存储、帮助构建大模型训练环境的预制化行业大模型和训练平台等，从而大幅提升企业专属大模型的训练效率。

在面对检索增强的模型输出依然无法达到企业准确性和实时性要求，或者企业需要将生成式AI应用于业务逻辑极其复杂的场景、而基础模型完全无法理解等情景时，企业需要考虑采用定制化模型精调训练落地路线，定制企业专属大模型。

3.3 电信运营商和云服务商的融合共建价值

3.3.1 同质化的硬件堆叠难以保证竞争中的优势

在当今快速发展的人工智能领域，智算集群作为支撑大模型运行的关键基础设施，其建设与优化显得尤为重要。然而，目前市场上智算集群的同质化现象严重，这不仅限制了技术的进步，也难以在激烈的市场竞争中保持优势。

为了突破这一瓶颈，我们需要提高智算集群的资源利用率和可用性是保证竞争优势的关键。这要求我们在集群建管用上进行创新，采用高性能的网络底座、存储底座，高效的训练框架，以及更先进的运营运维体系，实现资源的高效分配和充分利用。同时，还需要加强集群的容错能力和自愈能力，确保在面对各种挑战时，集群能够稳定运行，提供可靠的服务。

3.3.2 电信运营商优质资源和云服务商最佳实践的结合

在集群的建设和优化需要考虑竞争力、性价比以及可持续发展的问题。随着人工智能技术的不断进步，对资源投入、能源的需求也在不断增加。因此，我们需要在集群设计时考虑到资源利用率、能效比，采用更高效节能的算力集群建设技术，减少能源消耗，实现绿色计算。

在建设有竞争力的智算集群方面，电信运营商和云服务商各有所长。运营商通常拥有丰富的硬件资源，如GPU服务和IDC资源，这些资源是构建高效能智算集群的基础。而云服务商则在集群建设和业务应用上具有优势，拥有大规模算力集群的建设经验和丰富的业务应用经验。两者的结合，可以实现资源共享和优势互补，推动智算集群的商业化成功。

综上所述，要建设有竞争力的算力集群，需要发挥电信运营商和云服务商的双方优势，从资源差异化、集群建设能力、业务应用能力、可持续发展等多个方面进行创新和改进。通过这些措施，可以构建出更高效、更可靠、更绿色的智算集群，为人工智能的发展提供强有力的支持。

04

N个场景，云服务商支持电信运营商构建AI大模型场景化解决方案

云服务商支持运营商发展大模型应用的策略涉及多个层面，旨在提供全面的技术和服务支持。首先，云服务商拥有强大的 AI 算力资源，可以补齐运营商在算力分布、算力种类、算力合规方面的需求，以支撑大模型的训练和推理需求。此外，云服务商积极促进大模型生态建设，通过开源社区和合作伙伴网络，推动技术的共享与创新。提供易用的开发工具和工作流，帮助运营商降低技术门槛，加速大模型的开发和部署。同时，云服务商注重数据安全和隐私保护，确保平台符合法规要求，保护数据安全。

在商业模式创新方面，云服务商与电信运营商共同探索新的服务模式，如基于订阅的服务或按使用量计费，以实现商业价值的最大化。支持多云和混合云策略，为运营商提供灵活的云服务选项，适应不同部署和管理需求。

云服务商支持电信运营商大模型应用落地，参与多个场景的共建中，包括：企业知识应用、视联网内容分析、增值内容创作、客服场景、DICT 合作等场景。

3.3.2 电信运营商优质资源和云服务商最佳实践的结合

在集群的建设和优化需要考虑竞争力、性价比以及可持续发展的问题。随着人工智能技术的不断进步，对资源投入、能源的需求也在不断增加。因此，我们需要在集群设计时考虑到资源利用率、能效比，采用更高效节能的算力集群建设技术，减少能源消耗，实现绿色计算。

在建设有竞争力的智算集群方面，电信运营商和云服务商各有所长。运营商通常拥有丰富的硬件资源，如GPU服务和IDC资源，这些资源是构建高效能智算集群的基础。而云服务商则在集群建设和业务应用上具有优势，拥有大规模算力集群的建设经验和丰富的业务应用经验。两者的结合，可以实现资源共享和优势互补，推动智算集群的商业化成功。

综上所述，要建设有竞争力的算力集群，需要发挥电信运营商和云服务商的双方优势，从资源差异化、集群建设能力、业务应用能力、可持续发展等多个方面进行创新和改进。通过这些措施，可以构建出更高效、更可靠、更绿色的智算集群，为人工智能的发展提供强有力的支持。

04

N个场景，云服务商支持电信运营商构建AI大模型场景化解决方案

云服务商支持运营商发展大模型应用的策略涉及多个层面，旨在提供全面的技术和服务支持。首先，云服务商拥有强大的 AI 算力资源，可以补齐运营商在算力分布、算力种类、算力合规方面的需求，以支撑大模型的训练和推理需求。此外，云服务商积极促进大模型生态建设，通过开源社区和合作伙伴网络，推动技术的共享与创新。提供易用的开发工具和工作流，帮助运营商降低技术门槛，加速大模型的开发和部署。同时，云服务商注重数据安全和隐私保护，确保平台符合法规要求，保护数据安全。

在商业模式创新方面，云服务商与电信运营商共同探索新的服务模式，如基于订阅的服务或按使用量计费，以实现商业价值的最大化。支持多云和混合云策略，为运营商提供灵活的云服务选项，适应不同部署和管理需求。

云服务商支持电信运营商大模型应用落地，参与多个场景的共建中，包括：企业知识应用、视联网内容分析、增值内容创作、客服场景、DICT 合作等场景。

4.1 企业知识应用场景

01 场景理解

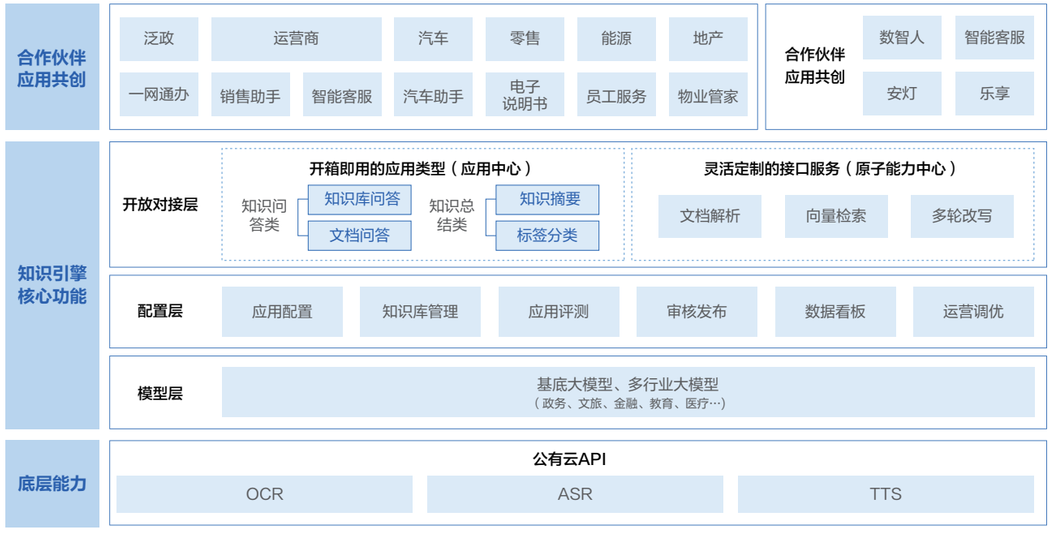
企业知识库是企业知识管理和信息化建设的重要一环，它通过集中存储、管理和利用组织的关键知识资产，支持企业在多个方面提高效率和效能。它用于文档管理，确保重要文件和记录的安全访问。作为经验传承的工具，保存企业的专业技能和教训。对新员工进行培训，帮助他们快速融入公司，并为员工解决问题提供参考和指导。

02 解决方案

企业知识库作为信息和知识管理的中心枢纽，尽管具有显著的优势，但在实际使用过程中也可能面临一系列挑战。技术障碍和复杂的操作界面可能阻碍用户有效使用知识库，而搜索工具的不精确可能影响信息检索的效率。员工对于向知识库贡献内容的参与度不足，以及缺乏定期的用户培训，都可能限制知识库功能的充分发挥。兼容性问题可能影响知识库与其他企业系统的协同工作，而维护知识库的持续成本对于一些组织可能是一个负担。

解决上述的问题，可以在企业知识库场景中引入基于行业大模型的知识引擎产品，这至关重要，因为它能显著提升知识管理的效率和效果。知识引擎具备自动化知识检索能力，通过智能推荐系统根据用户行为提供个性化内容，从而增加知识库的实用性。它的自然语言处理功能使用户能够以自然语言与知识库互动，提高检索的准确性和用户体验。

腾讯、阿里、百度陆续推出知识引擎产品，都旨在通过先进的人工智能技术，提高企业的知处理能力和应用开发效率。例如腾讯云知识引擎具有强大的知识获取、处理与推荐能力，能为企业提供精准、全面且实用的知识服务。通过智能分析用户需求，它助力企业提升员工技能，优化决策流程，并加速创新。知识引擎还具备出色的跨平台能力，能满足不同场景下的知识需求，有效降低知识获取成本，从而全面提升企业的竞争力和可持续发展能力。



▲ 知识引擎产品定位是助力客户重构企业级知识服务

03 关键技术

基于大模型的知识引擎集成了一系列尖端技术，以实现高级的知识管理和智能决策支持。关键技术包括预训练大模型、迁移学习、微调技术等。其中预训练大模型这些具备出色的语言理解和生成能力，迁移学习和微调技术使模型能够适应特定的领域或任务，而多模态处理能力让引擎能够同时处理文本、图像和声音等多类数据。

知识引擎还采用上下文理解技术，以确保对用户查询提供准确和相关的响应。长短期记忆网络（LSTM）和注意力机制赋予模型处理序列数据和集中注意力于关键信息的能力。Transformer架构，基于自注意力机制，是实现语言理解和生成任务的广泛应用模型。

另外，针对电信运营商的垂直场景，我们还需要通过数据的导入，训练出专业场景的大模型，例如：网络运维类大模型、营销服务类大模型。能提高知识引擎的鲁棒性与泛化能力，确保模型在不同数据和任务面前保持稳定表现。

04 方案效果

知识引擎在企业知识库中的应用显著提升了组织运作的多个方面。它通过快速检索和高度准确的搜索结果，极大提高了员工获取信息的效率。个性化的知识推荐系统根据员工的行为和偏好定制内容，进一步提升了用户体验。此外，知识引擎促进了知识共享和团队协作，降低了新员工培训成本，并加速了知识的更新和传播。通过自动化和智能化的工具，知识引擎还有助于避免知识流失，提高员工生产力，确保客户服务的快速响应，并增强知识的可发现性。

05 方案价值

知识引擎为企业带来的价值是多维度的。它不仅提升了决策的质量和效率，还通过确保内容的合规性和及时风险管理，为企业的稳健运营提供了保障。知识引擎的引入，加强了企业对市场变化的适应能力，培养了持续创新的环境，从而增强了企业的竞争优势。此外，它通过促进协作和团队工作，帮助企业构建了学习型组织，不断探索和实践新方法和解决方案。知识引擎的使用，使企业能够更好地保护和利用其知识资产，为企业长期发展和创新奠定了坚实的基础。

4.2 视联网内容分析场景

01 场景理解

运营商视联网是由运营商构建的新型视频服务基础设施，它利用5G、AI、云计算等新技术，集成了视频监控、AI分析、云存储等多种服务能力。视联网不仅是一个技术平台，更是运营商服务经济社会发展、推动数字化转型的重要举措。

中国电信的天翼视联网是一个典型的例子，它被定义为继移动网、宽带网、物联网、卫星网之后的“第五张网”，致力于构建全国统一的新型视频服务基础设施。中国移动也发布了自己的视联网服务，通过视频加速网和视频大模型提供服务，实现全国9200万路摄像头云端互联，并在家庭和政企领域提供智慧化解决方案，推动了视联网技术的创新和应用拓展。视联网的建设和应用对于运营商来说具有战略意义，它不仅能够提升运营商的服务能力，还能够促进数字经济的发展，推动社会治理现代化，丰富人民的数字生活体验。

4.1 企业知识应用场景

01 场景理解

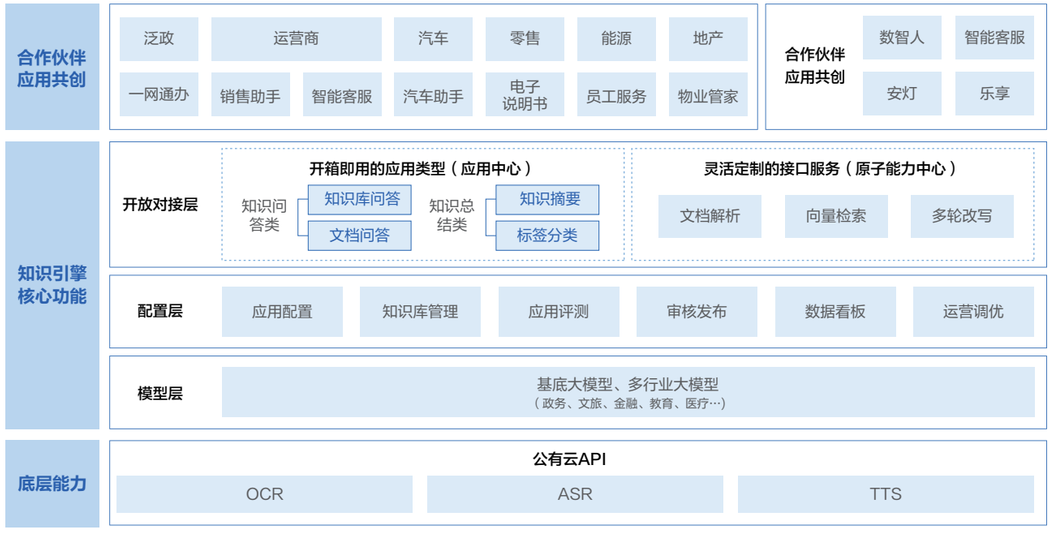
企业知识库是企业知识管理和信息化建设的重要一环，它通过集中存储、管理和利用组织的关键知识资产，支持企业在多个方面提高效率和效能。它用于文档管理，确保重要文件和记录的安全访问。作为经验传承的工具，保存企业的专业技能和教训。对新员工进行培训，帮助他们快速融入公司，并为员工解决问题提供参考和指导。

02 解决方案

企业知识库作为信息和知识管理的中心枢纽，尽管具有显著的优势，但在实际使用过程中也可能面临一系列挑战。技术障碍和复杂的操作界面可能阻碍用户有效使用知识库，而搜索工具的不精确可能影响信息检索的效率。员工对于向知识库贡献内容的参与度不足，以及缺乏定期的用户培训，都可能限制知识库功能的充分发挥。兼容性问题可能影响知识库与其他企业系统的协同工作，而维护知识库的持续成本对于一些组织可能是一个负担。

解决上述的问题，可以在企业知识库场景中引入基于行业大模型的知识引擎产品，这至关重要，因为它能显著提升知识管理的效率和效果。知识引擎具备自动化知识检索能力，通过智能推荐系统根据用户行为提供个性化内容，从而增加知识库的实用性。它的自然语言处理功能使用户能够以自然语言与知识库互动，提高检索的准确性和用户体验。

腾讯、阿里、百度陆续推出知识引擎产品，都旨在通过先进的人工智能技术，提高企业的知处理能力和应用开发效率。例如腾讯云知识引擎具有强大的知识获取、处理与推荐能力，能为企业提供精准、全面且实用的知识服务。通过智能分析用户需求，它助力企业提升员工技能，优化决策流程，并加速创新。知识引擎还具备出色的跨平台能力，能满足不同场景下的知识需求，有效降低知识获取成本，从而全面提升企业的竞争力和可持续发展能力。



▲ 知识引擎产品定位是助力客户重构企业级知识服务

03 关键技术

基于大模型的知识引擎集成了一系列尖端技术，以实现高级的知识管理和智能决策支持。关键技术包括预训练大模型、迁移学习、微调技术等。其中预训练大模型这些具备出色的语言理解和生成能力，迁移学习和微调技术使模型能够适应特定的领域或任务，而多模态处理能力让引擎能够同时处理文本、图像和声音等多类数据。

知识引擎还采用上下文理解技术，以确保对用户查询提供准确和相关的响应。长短期记忆网络（LSTM）和注意力机制赋予模型处理序列数据和集中注意力于关键信息的能力。Transformer架构，基于自注意力机制，是实现语言理解和生成任务的广泛应用模型。

另外，针对电信运营商的垂直场景，我们还需要通过数据的导入，训练出专业场景的大模型，例如：网络运维类大模型、营销服务类大模型。能提高知识引擎的鲁棒性与泛化能力，确保模型在不同数据和任务面前保持稳定表现。

04 方案效果

知识引擎在企业知识库中的应用显著提升了组织运作的多个方面。它通过快速检索和高度准确的搜索结果，极大提高了员工获取信息的效率。个性化的知识推荐系统根据员工的行为和偏好定制内容，进一步提升了用户体验。此外，知识引擎促进了知识共享和团队协作，降低了新员工培训成本，并加速了知识的更新和传播。通过自动化和智能化的工具，知识引擎还有助于避免知识流失，提高员工生产力，确保客户服务的快速响应，并增强知识的可发现性。

05 方案价值

知识引擎为企业带来的价值是多维度的。它不仅提升了决策的质量和效率，还通过确保内容的合规性和及时风险管理，为企业的稳健运营提供了保障。知识引擎的引入，加强了企业对市场变化的适应能力，培养了持续创新的环境，从而增强了企业的竞争优势。此外，它通过促进协作和团队工作，帮助企业构建了学习型组织，不断探索和实践新方法和解决方案。知识引擎的使用，使企业能够更好地保护和利用其知识资产，为企业长期发展和创新奠定了坚实的基础。

4.2 视联网内容分析场景

01 场景理解

运营商视联网是由运营商构建的新型视频服务基础设施，它利用5G、AI、云计算等新技术，集成了视频监控、AI分析、云存储等多种服务能力。视联网不仅是一个技术平台，更是运营商服务经济社会发展、推动数字化转型的重要举措。

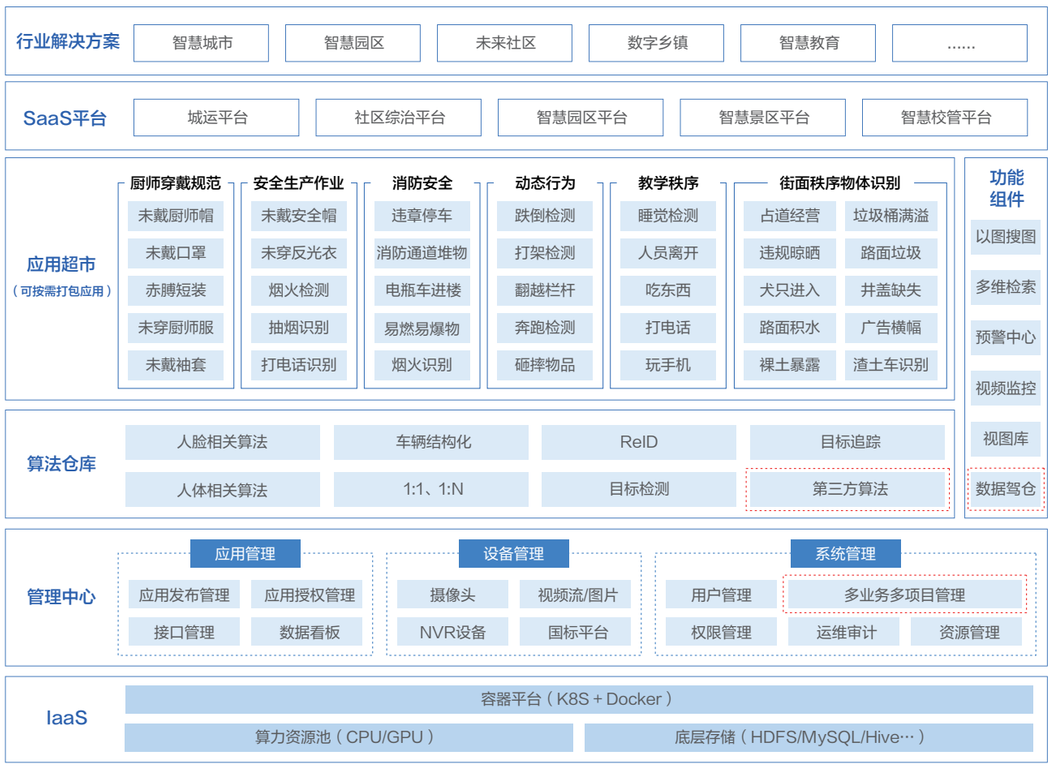
中国电信的天翼视联网是一个典型的例子，它被定义为继移动网、宽带网、物联网、卫星网之后的“第五张网”，致力于构建全国统一的新型视频服务基础设施。中国移动也发布了自己的视联网服务，通过视频加速网和视频大模型提供服务，实现全国9200万路摄像头云端互联，并在家庭和政企领域提供智慧化解决方案，推动了视联网技术的创新和应用拓展。视联网的建设和应用对于运营商来说具有战略意义，它不仅能够提升运营商的服务能力，还能够促进数字经济的发展，推动社会治理现代化，丰富人民的数字生活体验。

02 解决方案

运营商视联网的内容分析面临一系列挑战和问题。技术层面，视联网涉及体系架构、终端技术、编码技术、网络传输、CV大模型、算网融合、安全防护和体验评估等多个方面的挑战。随着业务规模的增长，算力和存储资源的消耗急剧增加，导致成本上升，因此需要研究新型数据表征与编码技术以节约资源。同时，实现低时延和高可靠的多媒体传输是提升用户体验的关键。CV大模型虽然在视联网中具有广泛应用前景，但其部署成本成为普及应用的障碍，需要探索算力资源的高效复用。安全防护方面，视联网业务的广泛性和终端的多样化对端到端业务安全提出了挑战，需要建立统一的分类分级的业务安全方案。最后，随着视联网的广泛应用，如何处理和保护海量视频数据中的个人隐私和数据合规性成为一个重要问题。解决这些问题对于推动运营商视联网内容分析的发展至关重要。

面对视联网内容分析中的挑战，CV大模型提供了一系列的解决方案来优化视频处理。这些模型通过自动化视频分析，减少人工审核需求，同时从视频中提取多维度特征以提高识别和分类的准确性。它们能够实时处理视频流，为监控和安全应用提供快速响应，并运用智能压缩技术减少存储和传输需求，降低成本。

腾讯云混元大模型，通过引入多模态能力，扩展大模型的应用范围，可以使其在更广泛的场景中表现出色。在识别和交互上，多模态大模型通过结合视觉与语言理解等能力，突破了此前的局限，实现了更精准的语义分析和全面的根因分析。而阿里云通义千问模型则以其高灵活性和开源开放的战略吸引了大量开发者的关注。企业可以根据自身的需求和场景，选择最适合的CV大模型产品。



▲ 视频智能分析平台架构

03 关键技术

CV大模型集成了一系列先进技术，以实现高效的图像和视频分析。深度学习作为核心，通过卷积神经网络（CNNs）自动提取图像特征，而循环神经网络（RNNs）及其变体长短期记忆网络（LSTM）则处理时间序列数据，捕捉视频中的时间动态特征。Transformer架构利用自注意力机制处理序列数据，增强了特征提取的能力。

目标检测技术如区域建议网络（R-CNN）和YOLO提高了对象识别和定位的准确性。生成对抗网络（GANs）用于生成逼真的图像或视频内容，而自编码器则用于数据压缩和特征学习。注意力机制使模型集中处理数据中的关键部分，提升效率和准确性。

多任务学习和迁移学习允许模型执行多项任务或利用预训练的知识适应新任务。模型微调和模型压缩技术，如剪枝和量化，使模型更适用于资源受限的环境。数据增强技术提高模型的泛化能力，而端到端学习则让模型直接从输入学习到输出，无需手动特征提取。

最后，多模态学习结合不同数据源的信息，提升模型的理解和预测能力。这些技术的结合为CV大模型提供了强大的功能，使其在各种复杂场景中都能执行高效的视频分析和图像识别任务。

04 方案效果

CV大模型在运营商视联网场景中的应用带来了显著的正面效果。它们极大地提升了视频分析能力，通过深度学习技术自动识别和分类视频中的对象。增强的目标检测和识别功能对安全监控至关重要，同时CV大模型还能够根据文本描述生成新的图像和视频内容，为运营商的增值业务提供支持。

在视频监控和安全领域，CV大模型通过自动化分析减少了人工审核的需求，有效降低了运营成本。此外，多模态数据处理能力让模型能够更全面地理解场景，而对敏感内容的识别和过滤增强了数据隐私和系统的安全性。CV大模型还推动了技术创新和应用拓展，运营商利用它们发布了视联网相关的白皮书和大模型，指导技术发展方向。

构建的行业大模型基于庞大的用户数据和业务经验，更好地理解用户需求，提供实际应用价值。总体而言，CV大模型的应用不仅提高了视频处理的智能化水平，还为运营商带来了成本效益和更丰富的用户体验，同时确保了数据安全和隐私保护。

05 方案价值

CV大模型在运营商视联网场景中的应用极大地增强了运营商的服务能力，带来了深远的价值。首先，它通过自动化视频分析提升了服务效率并降低了运营成本，同时通过个性化服务增强了用户体验。

此外，CV大模型优化了资源分配，提升了资源的利用效率，并从海量视频数据中提取有价值的信息，增强了业务洞察力。它还提升了服务质量，特别是在客户服务和故障排查方面，通过精准的目标检测和识别提高了服务的准确性和可靠性。技术层面上，CV大模型推动了运营商在人工智能和机器学习领域的技术进步。

在市场竞争中，CV大模型的应用使运营商能够提供差异化服务，增强了其市场竞争力。它还帮助运营商更好地管理和保护用户数据，确保了法规遵从。

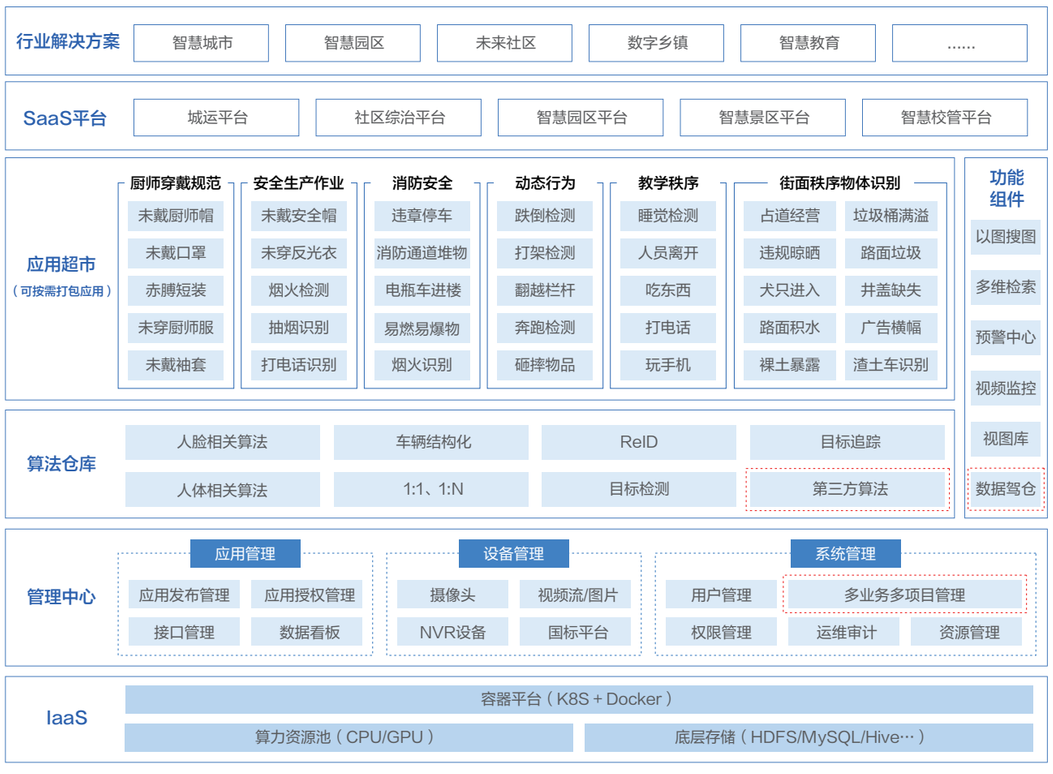
最后，CV大模型的应用促进了与技术供应商、内容提供商等合作伙伴的生态合作，共同推动了视联网生态的发展。总体而言，CV大模型为运营商带来了技术升级和市场竞争力的双重提升。

02 解决方案

运营商视联网的内容分析面临一系列挑战和问题。技术层面，视联网涉及体系架构、终端技术、编码技术、网络传输、CV大模型、算网融合、安全防护和体验评估等多个方面的挑战。随着业务规模的增长，算力和存储资源的消耗急剧增加，导致成本上升，因此需要研究新型数据表征与编码技术以节约资源。同时，实现低时延和高可靠的多媒体传输是提升用户体验的关键。CV大模型虽然在视联网中具有广泛应用前景，但其部署成本成为普及应用的障碍，需要探索算力资源的高效复用。安全防护方面，视联网业务的广泛性和终端的多样化对端到端业务安全提出了挑战，需要建立统一的分类分级的业务安全方案。最后，随着视联网的广泛应用，如何处理和保护海量视频数据中的个人隐私和数据合规性成为一个重要问题。解决这些问题对于推动运营商视联网内容分析的发展至关重要。

面对视联网内容分析中的挑战，CV大模型提供了一系列的解决方案来优化视频处理。这些模型通过自动化视频分析，减少人工审核需求，同时从视频中提取多维度特征以提高识别和分类的准确性。它们能够实时处理视频流，为监控和安全应用提供快速响应，并运用智能压缩技术减少存储和传输需求，降低成本。

腾讯云混元大模型，通过引入多模态能力，扩展大模型的应用范围，可以使其在更广泛的场景中表现出色。在识别和交互上，多模态大模型通过结合视觉与语言理解等能力，突破了此前的局限，实现了更精准的语义分析和全面的根因分析。而阿里云通义千问模型则以其高灵活性和开源开放的战略吸引了大量开发者的关注。企业可以根据自身的需求和场景，选择最适合的CV大模型产品。



▲ 视频智能分析平台架构

03 关键技术

CV大模型集成了一系列先进技术，以实现高效的图像和视频分析。深度学习作为核心，通过卷积神经网络（CNNs）自动提取图像特征，而循环神经网络（RNNs）及其变体长短期记忆网络（LSTM）则处理时间序列数据，捕捉视频中的时间动态特征。Transformer架构利用自注意力机制处理序列数据，增强了特征提取的能力。

目标检测技术如区域建议网络（R-CNN）和YOLO提高了对象识别和定位的准确性。生成对抗网络（GANs）用于生成逼真的图像或视频内容，而自编码器则用于数据压缩和特征学习。注意力机制使模型集中处理数据中的关键部分，提升效率和准确性。

多任务学习和迁移学习允许模型执行多项任务或利用预训练的知识适应新任务。模型微调和模型压缩技术，如剪枝和量化，使模型更适用于资源受限的环境。数据增强技术提高模型的泛化能力，而端到端学习则让模型直接从输入学习到输出，无需手动特征提取。

最后，多模态学习结合不同数据源的信息，提升模型的理解和预测能力。这些技术的结合为CV大模型提供了强大的功能，使其在各种复杂场景中都能执行高效的视频分析和图像识别任务。

04 方案效果

CV大模型在运营商视联网场景中的应用带来了显著的正面效果。它们极大地提升了视频分析能力，通过深度学习技术自动识别和分类视频中的对象。增强的目标检测和识别功能对安全监控至关重要，同时CV大模型还能够根据文本描述生成新的图像和视频内容，为运营商的增值业务提供支持。

在视频监控和安全领域，CV大模型通过自动化分析减少了人工审核的需求，有效降低了运营成本。此外，多模态数据处理能力让模型能够更全面地理解场景，而对敏感内容的识别和过滤增强了数据隐私和系统的安全性。CV大模型还推动了技术创新和应用拓展，运营商利用它们发布了视联网相关的白皮书和大模型，指导技术发展方向。

构建的行业大模型基于庞大的用户数据和业务经验，更好地理解用户需求，提供实际应用价值。总体而言，CV大模型的应用不仅提高了视频处理的智能化水平，还为运营商带来了成本效益和更丰富的用户体验，同时确保了数据安全和隐私保护。

05 方案价值

CV大模型在运营商视联网场景中的应用极大地增强了运营商的服务能力，带来了深远的价值。首先，它通过自动化视频分析提升了服务效率并降低了运营成本，同时通过个性化服务增强了用户体验。

此外，CV大模型优化了资源分配，提升了资源的利用效率，并从海量视频数据中提取有价值的信息，增强了业务洞察力。它还提升了服务质量，特别是在客户服务和故障排查方面，通过精准的目标检测和识别提高了服务的准确性和可靠性。技术层面上，CV大模型推动了运营商在人工智能和机器学习领域的技术进步。

在市场竞争中，CV大模型的应用使运营商能够提供差异化服务，增强了其市场竞争力。它还帮助运营商更好地管理和保护用户数据，确保了法规遵从。

最后，CV大模型的应用促进了与技术供应商、内容提供商等合作伙伴的生态合作，共同推动了视联网生态的发展。总体而言，CV大模型为运营商带来了技术升级和市场竞争力的双重提升。

4.3 增值内容创作场景

01 场景理解

运营商提供的增值业务种类丰富，涵盖了从基础通信到高端信息化服务的多个层面。增值业务不仅提升了用户的通信体验，也为电信运营商带来了新的收入来源和市场机会。随着技术的不断发展，如5G和大模型的应用，这些服务的应用场景将更加广泛和深入，进一步推动运营商业务的创新和转型。其中，大模型图像创作和大模型视频创作是非常重要的两种内容生成能力，已经逐步应用在运营商增值业务中。有别于传统的由运营商单向提供给用户，这类内容生成能力，能让用户参与增值产品的制作中，提高对用户的黏度和单价。

02 解决方案

大模型的图像创作引擎为运营商的增值业务提供了多样化的应用场景，从而丰富了用户体验并开拓了新的商业机会。在个性化内容生成方面，该引擎能够定制5G视频彩铃、音频彩铃，满足用户对个性化服务的需求。

社交媒体服务通过提供内容创作工具，使用户能够创作并分享个性化图像和视频，增加用户粘性。5G视频彩铃服务利用该引擎生成个性化视频内容，而企业宣传则可以依靠定制化的图像和视频制作服务来增强品牌形象。教育和培训材料通过动画和图解变得更加生动，内容推荐系统则通过个性化推荐图像或视频内容来提升用户体验。

此外，图像创作引擎在游戏和娱乐领域生成视觉元素，智能客服和虚拟助手的形象设计也因它而更加友好。数据可视化通过图表和信息图帮助用户快速理解复杂数据，版权内容保护则通过独特的视觉标记加强。这些应用场景展示了大模型图像创作引擎在推动运营商增值业务发展和创新方面的巨大潜力。

图像创作引擎、视频创作引擎等产品或服务通过生成高质量的图像、视频内容，为电信运营商发展提供了创新动力。它能够根据文本描述快速创建视觉素材，大幅降低传统图像制作的时间和成本。图像创作引擎、视频创作引擎还支持个性化定制，提升消费者参与度，孵化出5G视频彩铃等产品。

行业	运营商	广告	游戏	电商	文旅	传媒
应用	内容创作			内容编辑		
原子能力	图像生成	视频生成	文本生成	语音生成		
	AI绘画	广告植入	广告文案生成	声音工坊		
	动漫创作	虚偶制作	写作训练	音色定制		
	人脸融合	人像驱动	春联创作	AI作曲		
	人像变换	图文转视频	字幕生成	唱演		
训练平台						
原子能力	NLP	多模态	CV			
	文本生成通用	图生图	图片感知通用			
	文本理解通用	文生图	图片分类			
	摘要生成	视频理解	物体检测			
	序列标注	跨模态检索	.....			
	.....	.....				

▲ AIGC为增值业务提供制作能力

03 关键技术

大模型图像创作引擎依赖于一系列先进的技术来实现其功能。这些关键技术包括深度学习技术，特别是卷积神经网络（CNNs），它们在图像识别和特征提取方面至关重要。生成对抗网络（GANs）通过对抗训练生成逼真图像，而变分自编码器（VAEs）则用于学习数据的潜在分布，以生成新的图像样本。自注意力机制，常见于Transformer模型中，增强了模型处理图像区域关系的能力。

大模型音频创作引擎集成了一系列关键技术，使其在音频生成、处理和理解方面表现出色。深度学习技术是其核心，利用循环神经网络（RNNs）和长短期记忆网络（LSTMs）处理音频序列数据，捕捉时间动态特征。生成对抗网络（GANs）通过对抗训练生成高质量、逼真的音频样本。波形建模技术如WaveNet直接生成原始音频波形，而频谱建模技术则将音频信号转换为频谱表示，便于进一步处理。

音频效果处理技术如混响、压缩、均衡等，用于改善音频质量和创造特定的听觉效果。数据增强技术通过增加训练数据的多样性，提高模型的鲁棒性。模型压缩和优化技术如剪枝、量化减小模型大小，提高在资源受限设备上的运行效率。实时处理能力通过算法优化和硬件加速，确保音频创作引擎能够快速响应。这些技术的集合为音频创作引擎提供了强大的功能，使其在音乐制作、语音合成、声音效果设计等多种应用场景中都能发挥出色的音频创作和编辑能力。

04 方案效果

将大模型技术应用到电信运营商增值业务中，可以带来显著的商业和运营效益。首先，它能够通过个性化推荐和智能客服提升用户体验，从而增加用户满意度和忠诚度。同时，自动化和智能化处理可以显著降低人工成本，提高服务效率。此外，大模型技术有助于开发新的增值服务，如智能分析和预测服务，为运营商创造额外的收入来源。

此外，大模型技术的应用促进了与技术供应商、内容提供商等的生态合作，共同开发增值服务。它还推动了运营商在AI和大数据领域的技术进步，并增强了运营商的市场适应性，使其能够快速响应市场变化，及时调整业务策略。总体而言，大模型技术为运营商提供了一个强大的工具，以实现更高效、更智能的运营和更丰富的服务，从而在竞争激烈的市场中保持领先地位。

05 方案价值

将大模型技术应用到电信运营商增值业务场景中，可以带来显著的商业和社会效益。首先，它能够通过个性化服务和智能推荐显著提升客户体验，增加用户满意度和忠诚度。同时，自动化和智能化处理可以简化工作流程，减少人力成本和时间成本，从而提高运营效率。大模型技术还有助于创新服务产品，如智能分析和预测服务，拓宽服务范围并创造新的收入渠道，提高盈利能力。

此外，利用大数据分析提供更准确的市场洞察，可以支持更明智的业务决策，优化决策制定。展示运营商的技术先进性和创新能力，能够增强品牌形象，吸引更多用户。加强数据安全，利用AI进行安全监控和异常检测，提高网络安全防护能力。与技术供应商、内容提供商等建立合作关系，共同开发和推广增值服务，促进生态合作。

4.3 增值内容创作场景

01 场景理解

运营商提供的增值业务种类丰富，涵盖了从基础通信到高端信息化服务的多个层面。增值业务不仅提升了用户的通信体验，也为电信运营商带来了新的收入来源和市场机会。随着技术的不断发展，如5G和大模型的应用，这些服务的应用场景将更加广泛和深入，进一步推动运营商业务的创新和转型。其中，大模型图像创作和大模型视频创作是非常重要的两种内容生成能力，已经逐步应用在运营商增值业务中。有别于传统的由运营商单向提供给用户，这类内容生成能力，能让用户参与增值产品的制作中，提高对用户的黏度和单价。

02 解决方案

大模型的图像创作引擎为运营商的增值业务提供了多样化的应用场景，从而丰富了用户体验并开拓了新的商业机会。在个性化内容生成方面，该引擎能够定制5G视频彩铃、音频彩铃，满足用户对个性化服务的需求。

社交媒体服务通过提供内容创作工具，使用户能够创作并分享个性化图像和视频，增加用户粘性。5G视频彩铃服务利用该引擎生成个性化视频内容，而企业宣传则可以依靠定制化的图像和视频制作服务来增强品牌形象。教育和培训材料通过动画和图解变得更加生动，内容推荐系统则通过个性化推荐图像或视频内容来提升用户体验。

此外，图像创作引擎在游戏和娱乐领域生成视觉元素，智能客服和虚拟助手的形象设计也因它而更加友好。数据可视化通过图表和信息图帮助用户快速理解复杂数据，版权内容保护则通过独特的视觉标记加强。这些应用场景展示了大模型图像创作引擎在推动运营商增值业务发展和创新方面的巨大潜力。

图像创作引擎、视频创作引擎等产品或服务通过生成高质量的图像、视频内容，为电信运营商发展提供了创新动力。它能够根据文本描述快速创建视觉素材，大幅降低传统图像制作的时间和成本。图像创作引擎、视频创作引擎还支持个性化定制，提升消费者参与度，孵化出5G视频彩铃等产品。

行业	运营商	广告	游戏	电商	文旅	传媒
应用	内容创作			内容编辑		
原子能力	图像生成	视频生成	文本生成	语音生成		
	AI绘画	广告植入	广告文案生成	声音工坊		
	动漫创作	虚偶制作	写作训练	音色定制		
	人脸融合	人像驱动	春联创作	AI作曲		
	人像变换	图文转视频	字幕生成	唱演		
训练平台						
原子能力	NLP	多模态	CV			
	文本生成通用	图生图	图片感知通用			
	文本理解通用	文生图	图片分类			
	摘要生成	视频理解	物体检测			
	序列标注	跨模态检索	.....			
	.....	.....				

▲ AIGC为增值业务提供制作能力

03 关键技术

大模型图像创作引擎依赖于一系列先进的技术来实现其功能。这些关键技术包括深度学习技术，特别是卷积神经网络（CNNs），它们在图像识别和特征提取方面至关重要。生成对抗网络（GANs）通过对抗训练生成逼真图像，而变分自编码器（VAEs）则用于学习数据的潜在分布，以生成新的图像样本。自注意力机制，常见于Transformer模型中，增强了模型处理图像区域关系的能力。

大模型音频创作引擎集成了一系列关键技术，使其在音频生成、处理和理解方面表现出色。深度学习技术是其核心，利用循环神经网络（RNNs）和长短期记忆网络（LSTMs）处理音频序列数据，捕捉时间动态特征。生成对抗网络（GANs）通过对抗训练生成高质量、逼真的音频样本。波形建模技术如WaveNet直接生成原始音频波形，而频谱建模技术则将音频信号转换为频谱表示，便于进一步处理。

音频效果处理技术如混响、压缩、均衡等，用于改善音频质量和创造特定的听觉效果。数据增强技术通过增加训练数据的多样性，提高模型的鲁棒性。模型压缩和优化技术如剪枝、量化减小模型大小，提高在资源受限设备上的运行效率。实时处理能力通过算法优化和硬件加速，确保音频创作引擎能够快速响应。这些技术的集合为音频创作引擎提供了强大的功能，使其在音乐制作、语音合成、声音效果设计等多种应用场景中都能发挥出色的音频创作和编辑能力。

04 方案效果

将大模型技术应用到电信运营商增值业务中，可以带来显著的商业和运营效益。首先，它能够通过个性化推荐和智能客服提升用户体验，从而增加用户满意度和忠诚度。同时，自动化和智能化处理可以显著降低人工成本，提高服务效率。此外，大模型技术有助于开发新的增值服务，如智能分析和预测服务，为运营商创造额外的收入来源。

此外，大模型技术的应用促进了与技术供应商、内容提供商等的生态合作，共同开发增值服务。它还推动了运营商在AI和大数据领域的技术进步，并增强了运营商的市场适应性，使其能够快速响应市场变化，及时调整业务策略。总体而言，大模型技术为运营商提供了一个强大的工具，以实现更高效、更智能的运营和更丰富的服务，从而在竞争激烈的市场中保持领先地位。

05 方案价值

将大模型技术应用到电信运营商增值业务场景中，可以带来显著的商业和社会效益。首先，它能够通过个性化服务和智能推荐显著提升客户体验，增加用户满意度和忠诚度。同时，自动化和智能化处理可以简化工作流程，减少人力成本和时间成本，从而提高运营效率。大模型技术还有助于创新服务产品，如智能分析和预测服务，拓宽服务范围并创造新的收入渠道，提高盈利能力。

此外，利用大数据分析提供更准确的市场洞察，可以支持更明智的业务决策，优化决策制定。展示运营商的技术先进性和创新能力，能够增强品牌形象，吸引更多用户。加强数据安全，利用AI进行安全监控和异常检测，提高网络安全防护能力。与技术供应商、内容提供商等建立合作关系，共同开发和推广增值服务，促进生态合作。

4.4 客户服务场景

01 场景理解

随着人工智能技术的快速发展，大模型技术在各个领域的应用越来越广泛。特别是在运营商客服场景中，大模型的应用不仅可以提高客服效率，还能提升用户体验。本文将从运营商客服场景理解、大模型解决方案、关键技术、应用效果和价值五个方面进行详细分析。

随着用户需求的多样化和个性化，传统的客服系统已经难以满足用户的需求。因此，运营商需要通过技术创新来提升客服系统的智能化水平，满足用户的个性化需求。**客服场景的多样性：**运营商客服场景包括但不限于咨询、投诉、故障报修、业务办理等。每个场景都有其特定的需求和挑战。例如，在咨询场景中，用户可能需要获取关于套餐、资费、网络覆盖等信息；在投诉场景中，用户可能需要解决网络质量、服务中断等问题。**用户需求的个性化：**随着用户需求的多样化，传统的“一刀切”服务模式已经难以满足用户的需求。用户希望得到更加个性化、精准的服务。这就要求运营商客服系统能够根据用户的行为、偏好等信息，提供个性化的服务。**客服系统的挑战：**传统的客服系统主要依赖人工服务，存在服务效率低、成本高、服务质量不稳定等问题。此外，随着用户数量的增加，客服系统的工作压力也在不断增加。因此，运营商需要通过技术创新来提升客服系统的智能化水平，提高服务效率，降低成本，提升服务质量。

02 解决方案

为了解决传统客服系统存在的问题，运营商可以利用大模型技术来构建智能化的客服系统。大模型技术可以通过机器学习、自然语言处理等技术，实现对用户需求的精准识别和响应。

智能客服机器人：利用大模型技术，电信运营商可以构建智能客服机器人，实现对用户咨询的自动响应。智能客服机器人可以通过自然语言处理技术，理解用户的咨询内容，并提供精准的答案。

语音识别和语音合成：在语音客服场景中，运营商可以利用大模型技术实现语音识别和语音合成。通过语音识别技术，智能客服系统可以将用户的语音转化为文本，然后通过大模型进行处理；通过语音合成技术，智能客服系统可以将处理结果转化为语音，提供给用户。

多轮对话管理：在复杂的客服场景中，用户可能需要与客服系统进行多轮对话。利用大模型技术，运营商可以构建多轮对话管理系统，实现对用户需求的连续跟踪和响应。

例如企点客服产品，目前已形成综合性的在线客服解决方案，接入混元大模型使得智能客服机器更聪明、更拟人，可提供全天候自动回复服务，同时支持实时沟通，并能以文字、图片等形式与客户互动。该产品还包括客户信息管理、工单系统、数据统计分析等模块，帮助企业管理客户数据并优化服务策略。

▲ 客户服务需要构建一体化的全渠道服务运营中心

01 关键技术

大模型技术在运营商客服场景中的应用，涉及到多个关键技术，包括自然语言处理、机器学习、深度学习等。

自然语言处理：自然语言处理是大模型技术的核心，它使得机器能够理解和处理人类语言。在客服场景中，自然语言处理技术可以帮助智能客服机器人理解用户的咨询内容，并提供精准的答案。

机器学习：机器学习是大模型技术的基础，它使得机器能够从数据中学习规律，并做出预测。在客服场景中，机器学习技术可以帮助智能客服系统根据用户的行为、偏好等信息，提供个性化的服务。

深度学习：深度学习是大模型技术的重要组成部分，它通过模拟人脑的神经网络结构，实现对复杂数据的高效处理。在客服场景中，深度学习技术可以帮助智能客服系统处理大量的用户数据，提供更加精准的服务。

02 方案效果

大模型技术在运营商客服场景中的应用，已经取得了显著的效果。通过实际应用，运营商可以明显感受到大模型技术带来的改变。

提升服务效率：利用大模型技术，运营商可以构建智能客服系统，实现对用户咨询的自动响应。这不仅提高了服务效率，还降低了人工成本。

4.4 客户服务场景

01 场景理解

随着人工智能技术的快速发展，大模型技术在各个领域的应用越来越广泛。特别是在运营商客服场景中，大模型的应用不仅可以提高客服效率，还能提升用户体验。本文将从运营商客服场景理解、大模型解决方案、关键技术、应用效果和价值五个方面进行详细分析。

随着用户需求的多样化和个性化，传统的客服系统已经难以满足用户的需求。因此，运营商需要通过技术创新来提升客服系统的智能化水平，满足用户的个性化需求。**客服场景的多样性：**运营商客服场景包括但不限于咨询、投诉、故障报修、业务办理等。每个场景都有其特定的需求和挑战。例如，在咨询场景中，用户可能需要获取关于套餐、资费、网络覆盖等信息；在投诉场景中，用户可能需要解决网络质量、服务中断等问题。**用户需求的个性化：**随着用户需求的多样化，传统的“一刀切”服务模式已经难以满足用户的需求。用户希望得到更加个性化、精准的服务。这就要求运营商客服系统能够根据用户的行为、偏好等信息，提供个性化的服务。**客服系统的挑战：**传统的客服系统主要依赖人工服务，存在服务效率低、成本高、服务质量不稳定等问题。此外，随着用户数量的增加，客服系统的工作压力也在不断增加。因此，运营商需要通过技术创新来提升客服系统的智能化水平，提高服务效率，降低成本，提升服务质量。

02 解决方案

为了解决传统客服系统存在的问题，运营商可以利用大模型技术来构建智能化的客服系统。大模型技术可以通过机器学习、自然语言处理等技术，实现对用户需求的精准识别和响应。

智能客服机器人：利用大模型技术，电信运营商可以构建智能客服机器人，实现对用户咨询的自动响应。智能客服机器人可以通过自然语言处理技术，理解用户的咨询内容，并提供精准的答案。

语音识别和语音合成：在语音客服场景中，运营商可以利用大模型技术实现语音识别和语音合成。通过语音识别技术，智能客服系统可以将用户的语音转化为文本，然后通过大模型进行处理；通过语音合成技术，智能客服系统可以将处理结果转化为语音，提供给用户。

多轮对话管理：在复杂的客服场景中，用户可能需要与客服系统进行多轮对话。利用大模型技术，运营商可以构建多轮对话管理系统，实现对用户需求的连续跟踪和响应。

例如企点客服产品，目前已形成综合性的在线客服解决方案，接入混元大模型使得智能客服机器更聪明、更拟人，可提供全天候自动回复服务，同时支持实时沟通，并能以文字、图片等形式与客户互动。该产品还包括客户信息管理、工单系统、数据统计分析等模块，帮助企业管理客户数据并优化服务策略。

▲ 客户服务需要构建一体化的全渠道服务运营中心

01 关键技术

大模型技术在运营商客服场景中的应用，涉及到多个关键技术，包括自然语言处理、机器学习、深度学习等。

自然语言处理：自然语言处理是大模型技术的核心，它使得机器能够理解和处理人类语言。在客服场景中，自然语言处理技术可以帮助智能客服机器人理解用户的咨询内容，并提供精准的答案。

机器学习：机器学习是大模型技术的基础，它使得机器能够从数据中学习规律，并做出预测。在客服场景中，机器学习技术可以帮助智能客服系统根据用户的行为、偏好等信息，提供个性化的服务。

深度学习：深度学习是大模型技术的重要组成部分，它通过模拟人脑的神经网络结构，实现对复杂数据的高效处理。在客服场景中，深度学习技术可以帮助智能客服系统处理大量的用户数据，提供更加精准的服务。

02 方案效果

大模型技术在运营商客服场景中的应用，已经取得了显著的效果。通过实际应用，运营商可以明显感受到大模型技术带来的改变。

提升服务效率：利用大模型技术，运营商可以构建智能客服系统，实现对用户咨询的自动响应。这不仅提高了服务效率，还降低了人工成本。

提高服务质量：通过自然语言处理和机器学习技术，智能客服系统可以更加精准地理解用户的需求，并提供个性化的服务。这不仅提高了用户的满意度，还提升了企业的品牌形象。

降低运营成本：传统的客服系统主要依赖人工服务，成本较高。利用大模型技术，运营商可以减少对人工客服的依赖，降低运营成本。

提升用户体验：智能客服系统可以提供24/7的服务，不受时间和地点的限制。此外，智能客服系统还可以根据用户的行为、偏好等信息，提供个性化的服务，提升用户体验。

03 方案价值

大模型技术在运营商客服场景中的应用，不仅能够提升服务效率和质量，还能够为企业带来更大的价值。

增强企业竞争力：通过大模型技术，运营商可以提供更加智能化、个性化的服务，增强企业的竞争力。在激烈的市场竞争中，这无疑是一个重要的优势。提升品牌形象：高质量的服务是提升品牌形象的关键。通过大模型技术，运营商可以提供更加精准、个性化的服务，提升用户的满意度，从而提升企业的品牌形象。促进业务创新：大模型技术的应用，为运营商提供了新的机会。运营商可以利用大模型技术，开发新的服务和产品，推动业务创新。提高决策效率：通过大数据分析，运营商可以获得更加准确的市场洞察，支持更加明智的业务决策。这不仅可以提高决策效率，还可以降低决策风险。

大模型技术在运营商客服场景中的应用，已经展现出巨大的潜力和价值。通过大模型技术，运营商可以提供更加智能化、个性化的服务，提升服务效率和质量，降低运营成本，增强企业竞争力。同时，大模型技术的应用还可以推动业务创新，提升品牌形象，促进社会进步。未来，随着技术的不断发展，大模型技术在运营商客服场景中的应用将会更加广泛和深入。

4.5 DICT 合作场景

01 场景理解

行业大模型是一种专为特定行业或领域设计的人工智能模型，按照行业分，当前比较成熟的行业大模型有：医疗行业大模型、金融行业大模型、政务大模型等。行业大模型需要通过大量数据训练和优化，具备高度专业化和智能化的特点。这些模型专注于处理特定行业的专业数据和问题，依赖大量行业数据进行训练，能够辅助或自动化决策过程，提供预测、分析和建议。它们具有多模态处理能力，能够处理和分析文本、图像、声音等多种类型的数据，同时具备持续学习与优化的能力，适应不断变化的行业需求。行业大模型在设计和应用时需考虑数据的安全性和隐私保护，符合相关的数据保护法规和标准。此外，它们需要具备良好的扩展性、跨平台和设备兼容性，以及用户友好的交互界面，确保非技术用户也能方便地使用。在开发和应用过程中，还需考虑其对社会的影响，确保符合伦理标准和社会责任。

以医疗大模型为例，它可以帮助医生进行诊断和提供治疗建议。这个模型需要基于广泛且高质量的医学数据进行训练。该模型需经过大量医学文本数据训练，具有强大的理解和生成医学文本的能力，能够准确理解和回答医疗相关的问题。

02 解决方案

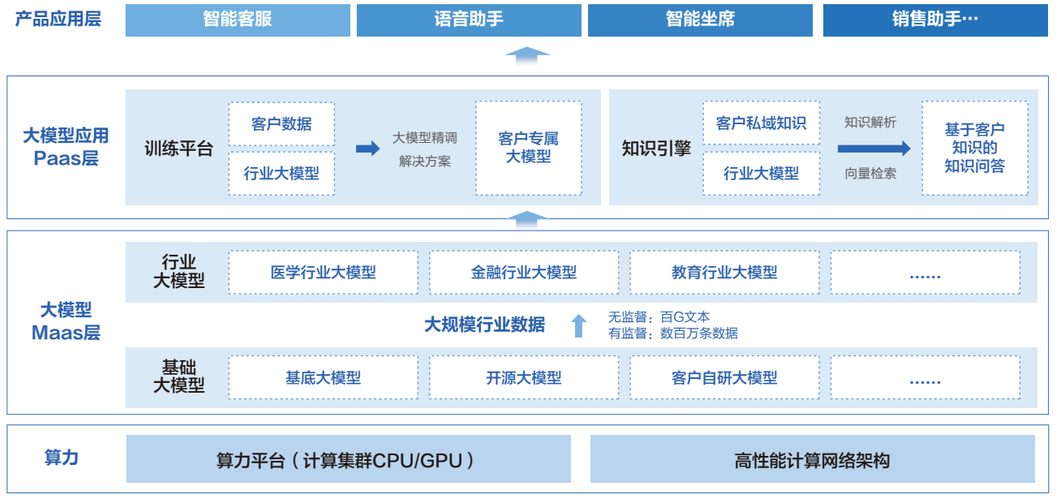
以医疗大模型产品为例，需要满足：医典问答、导诊、院务问答、用药咨询、疾病诊断、医学术语标准化、预问诊、病例结构化等方面的需求。

以医典问答为例，该任务的目标是让系统能够准确回答疾病、症状、药品等医学相关的问题。医典问答定位是一个集合了大量医学知识和信息的平台，内容涵盖用户的全医疗周期，包括百科词条、图文、视频等多种形式的医学科普知识。这些内容

通常由公立三甲名医提供，确保信息的专业性和准确性。医典问答类产品在医学大模型研究与测试中具有极高的价值。

首先，基于医典构造的问答测试集的内容丰富多样，可以提供全面的医学知识测试。医学是一个涵盖了众多领域和专业的学科，包括内科、外科、妇产科、儿科、眼科、耳鼻喉科等等。医典测试集包含了这些领域的知识，可以全面测试医学大模型在各个领域的表现，帮助我们了解模型在处理不同类型医学问题时的能力。

其次，该测试集的内容具有很高的专业性和准确性。这些内容由公立三甲名医提供，确保了信息的权威性。通过这些专业、准确的内容，可以测试医学大模型的准确性，看看它是否能够准确理解和处理专业的医学知识。此外，医典测试集的内容覆盖了用户的全医疗周期，包括预防、诊断、治疗和康复等各个阶段，可以使用医典测试集来测试医学大模型在处理不同阶段医疗问题时的能力，以验证其应用效果。



▲ 行业大模型框架

01 关键技术

医疗行业大模型的关键技术是多方面的，涵盖了从数据处理到患者护理的各个环节。它们包括机器学习和深度学习算法，用于分析和筛选庞大的健康数据集；自然语言处理技术，以改善医患沟通并提供具体治疗方案；计算机视觉技术，用于医学影像分析，自动识别病变区域；以及强大的大数据处理能力，整合和分析结构化和非结构化数据。智能手术机器人规划导航技术与手术机器人结合，提供复杂的手术辅助。个性化治疗和患者管理技术，根据患者具体情况定制治疗方案。药物研发领域中，AI大模型通过预测蛋白质复合物和加速药物筛选来优化临床试验设计。医疗质控方面，AI大模型能够生成规范医疗文书并检测缺陷，提高医疗效率和质量。跨模态数据处理能力使医疗大模型能够综合多种数据类型进行分析和预测。安全性和隐私保护技术确保在处理敏感数据时的数据安全和隐私。技术与临床的结合需要对底层算法进行优化，以适应医疗场景需求。此外，随着AI大模型在医疗领域的应用，标准化和监管机制的建立也变得至关重要，以确保技术的规范化应用和患者安全。这些技术的融合和发展，正推动医疗行业实现更高效、更智能化的服务，同时也带来了新的挑战和机遇。

02 方案效果

医疗大模型的应用在医疗健康领域产生了显著的积极效果。它们极大地提升了诊断的效率和准确性，通过快速分析医学影像和临床数据辅助医生做出更准确的判断。此外，医疗大模型优化了治疗方案，为患者提供个性化的治疗建议，同时加速了新药的研发流程，降低了成本并提高了药物研发的成功率。在改善患者体验方面，医疗大模型通过智能导诊和症状分析服务，使患者能够便捷地获取医疗信息。它们还提高了医疗服务的整体质量，通过自动化的质量控制减少了人为错误，并且帮助医院和医疗机构更有效地规划和使用医疗资源。

提高服务质量：通过自然语言处理和机器学习技术，智能客服系统可以更加精准地理解用户的需求，并提供个性化的服务。这不仅提高了用户的满意度，还提升了企业的品牌形象。

降低运营成本：传统的客服系统主要依赖人工服务，成本较高。利用大模型技术，运营商可以减少对人工客服的依赖，降低运营成本。

提升用户体验：智能客服系统可以提供24/7的服务，不受时间和地点的限制。此外，智能客服系统还可以根据用户的行为、偏好等信息，提供个性化的服务，提升用户体验。

03 方案价值

大模型技术在运营商客服场景中的应用，不仅能够提升服务效率和质量，还能够为企业带来更大的价值。

增强企业竞争力：通过大模型技术，运营商可以提供更加智能化、个性化的服务，增强企业的竞争力。在激烈的市场竞争中，这无疑是一个重要的优势。提升品牌形象：高质量的服务是提升品牌形象的关键。通过大模型技术，运营商可以提供更加精准、个性化的服务，提升用户的满意度，从而提升企业的品牌形象。促进业务创新：大模型技术的应用，为运营商提供了新的机会。运营商可以利用大模型技术，开发新的服务和产品，推动业务创新。提高决策效率：通过大数据分析，运营商可以获得更加准确的市场洞察，支持更加明智的业务决策。这不仅可以提高决策效率，还可以降低决策风险。

大模型技术在运营商客服场景中的应用，已经展现出巨大的潜力和价值。通过大模型技术，运营商可以提供更加智能化、个性化的服务，提升服务效率和质量，降低运营成本，增强企业竞争力。同时，大模型技术的应用还可以推动业务创新，提升品牌形象，促进社会进步。未来，随着技术的不断发展，大模型技术在运营商客服场景中的应用将会更加广泛和深入。

4.5 DICT 合作场景

01 场景理解

行业大模型是一种专为特定行业或领域设计的人工智能模型，按照行业分，当前比较成熟的行业大模型有：医疗行业大模型、金融行业大模型、政务大模型等。行业大模型需要通过大量数据训练和优化，具备高度专业化和智能化的特点。这些模型专注于处理特定行业的专业数据和问题，依赖大量行业数据进行训练，能够辅助或自动化决策过程，提供预测、分析和建议。它们具有多模态处理能力，能够处理和分析文本、图像、声音等多种类型的数据，同时具备持续学习与优化的能力，适应不断变化的行业需求。行业大模型在设计和应用时需考虑数据的安全性和隐私保护，符合相关的数据保护法规和标准。此外，它们需要具备良好的扩展性、跨平台和设备兼容性，以及用户友好的交互界面，确保非技术用户也能方便地使用。在开发和应用过程中，还需考虑其对社会的影响，确保符合伦理标准和社会责任。

以医疗大模型为例，它可以帮助医生进行诊断和提供治疗建议。这个模型需要基于广泛且高质量的医学数据进行训练。该模型需经过大量医学文本数据训练，具有强大的理解和生成医学文本的能力，能够准确理解和回答医疗相关的问题。

02 解决方案

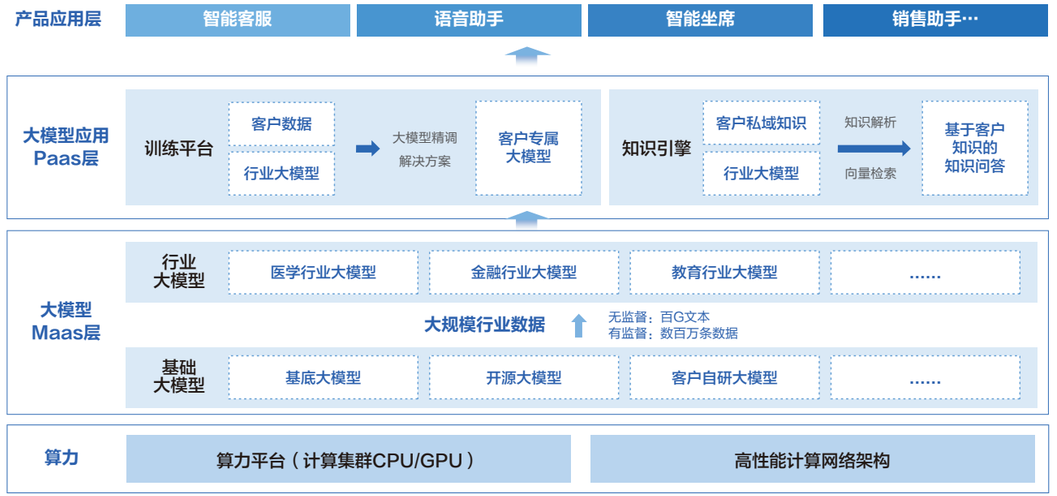
以医疗大模型产品为例，需要满足：医典问答、导诊、院务问答、用药咨询、疾病诊断、医学术语标准化、预问诊、病例结构化等方面的需求。

以医典问答为例，该任务的目标是让系统能够准确回答疾病、症状、药品等医学相关的问题。医典问答定位是一个集合了大量医学知识和信息的平台，内容涵盖用户的全医疗周期，包括百科词条、图文、视频等多种形式的医学科普知识。这些内容

通常由公立三甲名医提供，确保信息的专业性和准确性。医典问答类产品在医学大模型研究与测试中具有极高的价值。

首先，基于医典构造的问答测试集的内容丰富多样，可以提供全面的医学知识测试。医学是一个涵盖了众多领域和专业的学科，包括内科、外科、妇产科、儿科、眼科、耳鼻喉科等等。医典测试集包含了这些领域的知识，可以全面测试医学大模型在各个领域的表现，帮助我们了解模型在处理不同类型医学问题时的能力。

其次，该测试集的内容具有很高的专业性和准确性。这些内容由公立三甲名医提供，确保了信息的权威性。通过这些专业、准确的内容，可以测试医学大模型的准确性，看看它是否能够准确理解和处理专业的医学知识。此外，医典测试集的内容覆盖了用户的全医疗周期，包括预防、诊断、治疗和康复等各个阶段，可以使用医典测试集来测试医学大模型在处理不同阶段医疗问题时的能力，以验证其应用效果。



▲ 行业大模型框架

01 关键技术

医疗行业大模型的关键技术是多方面的，涵盖了从数据处理到患者护理的各个环节。它们包括机器学习和深度学习算法，用于分析和筛选庞大的健康数据集；自然语言处理技术，以改善医患沟通并提供具体治疗方案；计算机视觉技术，用于医学影像分析，自动识别病变区域；以及强大的大数据处理能力，整合和分析结构化和非结构化数据。智能手术机器人规划导航技术与手术机器人结合，提供复杂的手术辅助。个性化治疗和患者管理技术，根据患者具体情况定制治疗方案。药物研发领域中，AI大模型通过预测蛋白质复合物和加速药物筛选来优化临床试验设计。医疗质控方面，AI大模型能够生成规范医疗文书并检测缺陷，提高医疗效率和质量。跨模态数据处理能力使医疗大模型能够综合多种数据类型进行分析和预测。安全性和隐私保护技术确保在处理敏感数据时的数据安全和隐私。技术与临床的结合需要对底层算法进行优化，以适应医疗场景需求。此外，随着AI大模型在医疗领域的应用，标准化和监管机制的建立也变得至关重要，以确保技术的规范化应用和患者安全。这些技术的融合和发展，正推动医疗行业实现更高效、更智能化的服务，同时也带来了新的挑战和机遇。

02 方案效果

医疗大模型的应用在医疗健康领域产生了显著的积极效果。它们极大地提升了诊断的效率和准确性，通过快速分析医学影像和临床数据辅助医生做出更准确的判断。此外，医疗大模型优化了治疗方案，为患者提供个性化的治疗建议，同时加速了新药的研发流程，降低了成本并提高了药物研发的成功率。在改善患者体验方面，医疗大模型通过智能导诊和症状分析服务，使患者能够便捷地获取医疗信息。它们还提高了医疗服务的整体质量，通过自动化的质量控制减少了人为错误，并且帮助医院和医疗机构更有效地规划和使用医疗资源。

医疗大模型还支持医学教育和科研，为医学教育和科研提供了丰富的数据资源和分析工具，推动了医学知识的传播和创新。在公共卫生管理方面，医疗大模型通过疫情监控和疾病预防提高了公共卫生事件的应对能力。它们还促进了医学与数据科学、生物信息学等领域的交叉融合，推动了医疗健康领域的全面创新。随着医疗大模型的广泛应用，医疗伦理和监管机制也得到了加强，确保了技术应用的安全性和合规性。

03 方案价值

医疗大模型通过其强大的数据处理能力和智能化分析，在医疗健康领域创造了显著的价值。它们提高了诊断的准确性，辅助医生快速识别疾病；个性化治疗计划根据患者具体情况定制，提高疗效；加速药物研发流程，降低成本；优化患者体验，提供便捷人性化服务；合理配置医疗资源，提高运营效率；确保医疗质量控制，减少差错；支持医学教育和科研，构建知识图谱；在公共卫生管理中，通过数据分析预测疾病趋势；促进医学与其他学科的交叉融合，推动创新；同时也促进了医疗伦理和监管机制的完善。医疗大模型的应用不仅提升了医疗服务的质量和效率，还为医疗行业的数字化转型和智能化升级提供了动力，具有深远的社会和经济意义。

05

电信运营商大模型应用案例

医疗大模型还支持医学教育和科研，为医学教育和科研提供了丰富的数据资源和分析工具，推动了医学知识的传播和创新。在公共卫生管理方面，医疗大模型通过疫情监控和疾病预防提高了公共卫生事件的应对能力。它们还促进了医学与数据科学、生物信息学等领域的交叉融合，推动了医疗健康领域的全面创新。随着医疗大模型的广泛应用，医疗伦理和监管机制也得到了加强，确保了技术应用的安全性和合规性。

03 方案价值

医疗大模型通过其强大的数据处理能力和智能化分析，在医疗健康领域创造了显著的价值。它们提高了诊断的准确性，辅助医生快速识别疾病；个性化治疗计划根据患者具体情况定制，提高疗效；加速药物研发流程，降低成本；优化患者体验，提供便捷人性化服务；合理配置医疗资源，提高运营效率；确保医疗质量控制，减少差错；支持医学教育和科研，构建知识图谱；在公共卫生管理中，通过数据分析预测疾病趋势；促进医学与其他学科的交叉融合，推动创新；同时也促进了医疗伦理和监管机制的完善。医疗大模型的应用不仅提升了医疗服务的质量和效率，还为医疗行业的数字化转型和智能化升级提供了动力，具有深远的社会和经济意义。

05

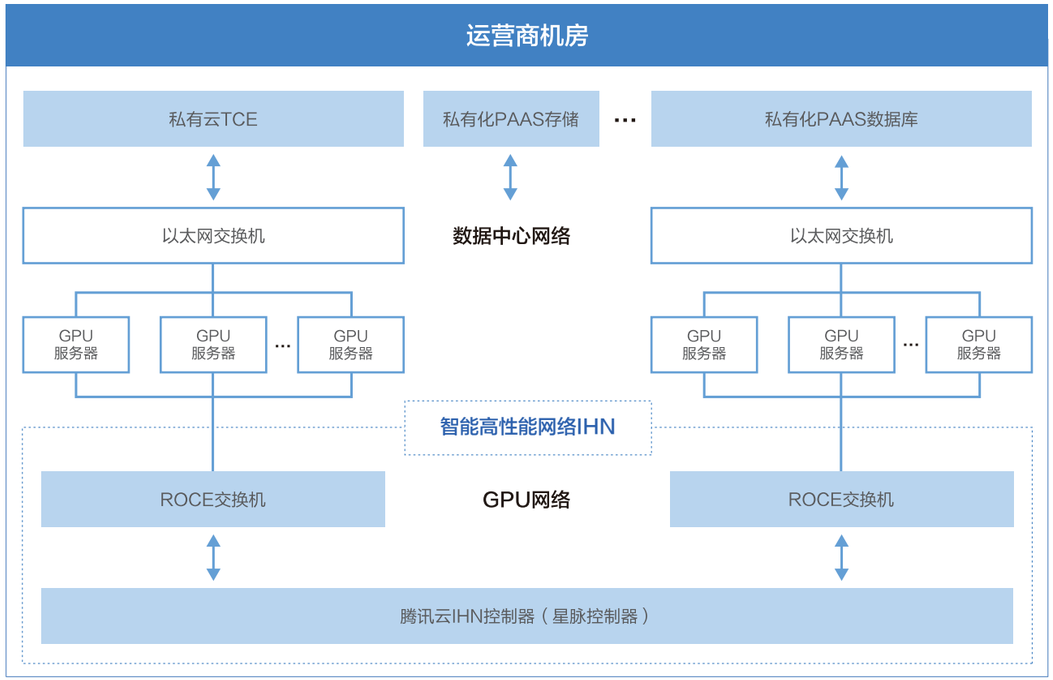
电信运营商大模型应用案例

5.1 强强联合共建大模型算力集群

大模型算力集群建设，是应对当前AI大模型训练需求爆炸式增长的关键举措。随着AIGC技术的迅猛发展，大模型在处理复杂任务、提供智能化服务方面展现出巨大潜力。参数的指数级增长，对大模型的训练和推理过程的算力集群规模和能力提出了极高的要求，单集群规模从千卡发展到万卡，甚至往10万卡级别递进。

为了应对AI大模型算力需求的爆炸式增长，运营商作为算力基础设施建设的主力军，大力推进构建高效的大模型算力集群。这一解决方案涉及多个层面，包括高性能计算、高性能网络以及高性能存储，以及各节点间高效协作的生产框架。

以某运营商智算中心项目为例，其采用私有云集成+标准化组件合作路线，与腾讯云联合搭建智算集群、运营运维平台，满足数据本地化要求。在基础设施层面提供强有力的保证，支持灵活的扩展能力，在当前项目基础上，进一步做万卡规模集群的规划。

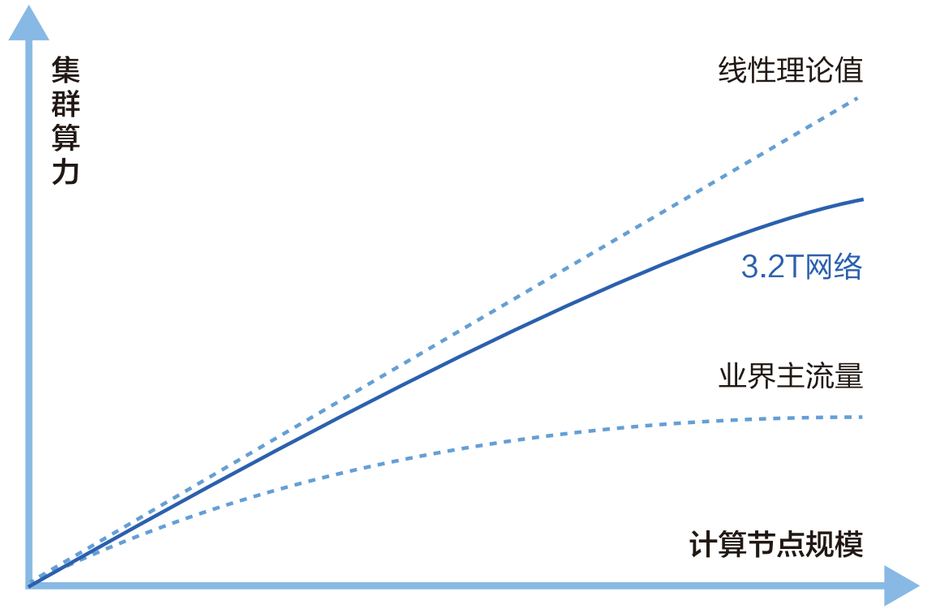


▲ 同运营商合建算力集群

该集群的建设包含了高性能计算、高性能网络、高性能存储、高效的模型生产框架、高效资源调度与管理技术、智能化运维等。

通过对单机算力、网络架构和存储性能进行协同优化，能够为大模型训练提供高性能、高带宽、低延迟的智算能力支撑。

网络层面，计算节点间存在海量的数据交互需求，随着集群规模扩大，通信性能会直接影响训练效率。此项目中采用了星脉网络，为新一代集群带来了业界最高的3.2T的超高通信带宽。节点内外统一的AllReduce通信带宽，实现网络和算力的最大协同。实测结果显示，搭载同样的 GPU，最新的3.2T星脉网络相较1.6T网络，能让集群整体算力提升20%。



▲ 星脉网络带宽示意

基于多轨道聚合的无阻塞网络架构、主动拥塞控制和定制加速通信库，云服务商可提供强大的集群构建能力，支持单集群高达十万卡级别的组网规模。在超大集群场景下，仍然能保持优秀的通信开销比和吞吐性能，满足大模型训练以及推理业务的横向扩展。

同时，高性能集合通信库TCCL的应用，基于星脉网络硬件平台深度优化，在全局路径规划、拓扑感知亲和性调度、网络故障实时告警/自愈等方面融入了定制设计的解决方案。相对业界开源集合通信库，为大模型训练优化40%负载性能，消除多个网络原因导致训练中断问题。

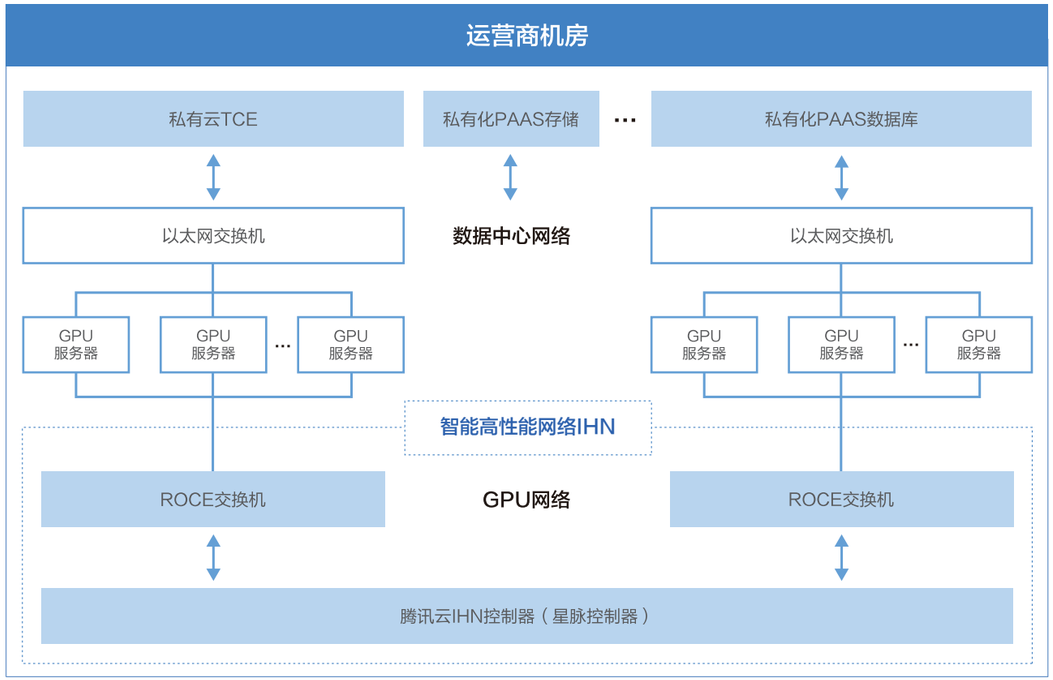
存储层面，训练场景下，几千台计算节点会同时读取一批数据集，需要尽可能缩短数据集的加载时长。对象存储COS+数据加速器GooseFS的存储方案可以提供多层缓存加速，大幅提升端到端的数据读取性能；将公开数据集、训练数据、模型结果统一存储到对象存储中，实现数据统一存储和高效流转。同时，GooseFS按需将热数据缓存到 GPU 内存和本地盘中，利用数据本地性提供高性能访问。

5.1 强强联合共建大模型算力集群

大模型算力集群建设，是应对当前AI大模型训练需求爆炸式增长的关键举措。随着AIGC技术的迅猛发展，大模型在处理复杂任务、提供智能化服务方面展现出巨大潜力。参数的指数级增长，对大模型的训练和推理过程的算力集群规模和能力提出了极高的要求，单集群规模从千卡发展到万卡，甚至往10万卡级别递进。

为了应对AI大模型算力需求的爆炸式增长，运营商作为算力基础设施建设的主力军，大力推进构建高效的大模型算力集群。这一解决方案涉及多个层面，包括高性能计算、高性能网络以及高性能存储，以及各节点间高效协作的生产框架。

以某运营商智算中心项目为例，其采用私有云集成+标准化组件合作路线，与腾讯云联合搭建智算集群、运营运维平台，满足数据本地化要求。在基础设施层面提供强有力的保证，支持灵活的扩展能力，在当前项目基础上，进一步做万卡规模集群的规划。

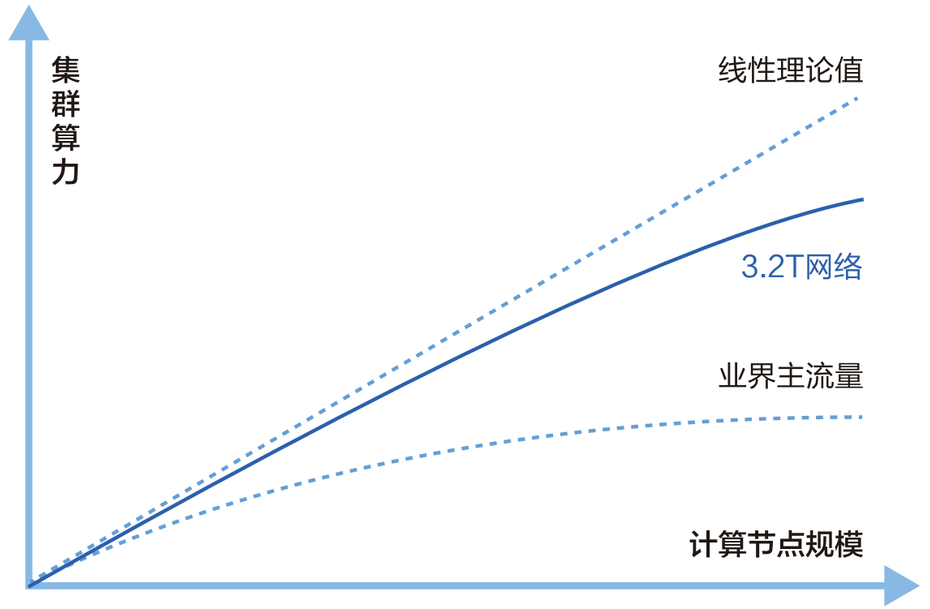


▲ 同运营商合建算力集群

该集群的建设包含了高性能计算、高性能网络、高性能存储、高效的模型生产框架、高效资源调度与管理技术、智能化运维等。

通过对单机算力、网络架构和存储性能进行协同优化，能够为大模型训练提供高性能、高带宽、低延迟的智算能力支撑。

网络层面，计算节点间存在海量的数据交互需求，随着集群规模扩大，通信性能会直接影响训练效率。此项目中采用了星脉网络，为新一代集群带来了业界最高的3.2T的超高通信带宽。节点内外统一的AllReduce通信带宽，实现网络和算力的最大协同。实测结果显示，搭载同样的 GPU，最新的3.2T星脉网络相较1.6T网络，能让集群整体算力提升20%。

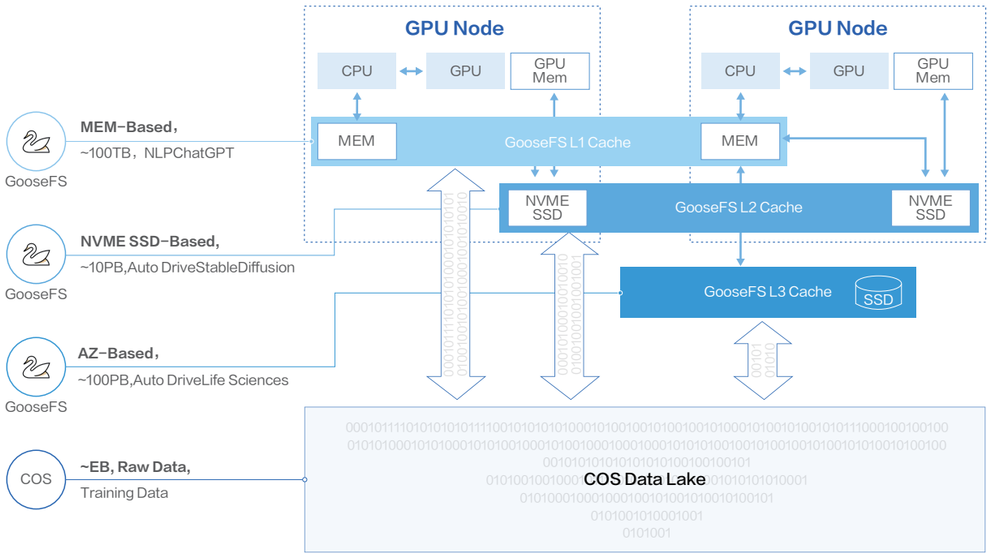


▲ 星脉网络带宽示意

基于多轨道聚合的无阻塞网络架构、主动拥塞控制和定制加速通信库，云服务商可提供强大的集群构建能力，支持单集群高达十万卡级别的组网规模。在超大集群场景下，仍然能保持优秀的通信开销比和吞吐性能，满足大模型训练以及推理业务的横向扩展。

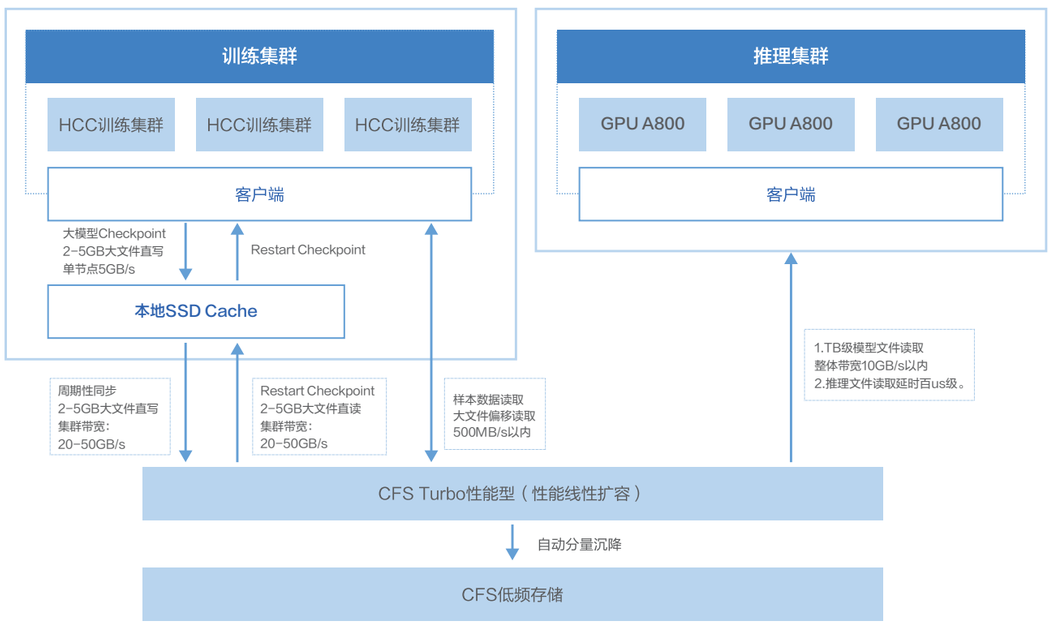
同时，高性能集合通信库TCCL的应用，基于星脉网络硬件平台深度优化，在全局路径规划、拓扑感知亲和性调度、网络故障实时告警/自愈等方面融入了定制设计的解决方案。相对业界开源集合通信库，为大模型训练优化40%负载性能，消除多个网络原因导致训练中断问题。

存储层面，训练场景下，几千台计算节点会同时读取一批数据集，需要尽可能缩短数据集的加载时长。对象存储COS+数据加速器GooseFS的存储方案可以提供多层缓存加速，大幅提升端到端的数据读取性能；将公开数据集、训练数据、模型结果统一存储到对象存储中，实现数据统一存储和高效流转。同时，GooseFS按需将热数据缓存到 GPU 内存和本地盘中，利用数据本地性提供高性能访问。



▲ COS+GooseFS缓存加速方案

高性能并行文件存储CFS Turbo方案，采取多级缓存加速，基于全分布式架构，提供 100GB/s 带宽、1000 万 IOPS 的极致性能。并通过持久化客户端缓存技术，将裸金属服务器本地 NVMe SSD 和 Turbo 文件系统构成统一命名空间，实现微秒级延时，解决大模型场景大数据量、高带宽、低延时的诉求。同时，通过智能分层技术，自动对冷热数据分层，节省80%的存储成本，提供极致的性价比。



▲ CFS Turbo大模型解决方案

底层架构之上，针对大模型训练场景，新一代集群集成了训练加速引擎TACO Train，对网络协议、通信策略、AI 框架、模型编译进行大量系统级优化，大幅节约训练调优和算力成本。

此案例通过运营商优质的计算和IDC资源结合腾讯侧智算集群建设方案，建设高性能的智算集群，结合丰富的智算套件能力，提高算力集群的服务化水平，更好的发挥集群的利用率、生产效率、运维效率的水平，建成面向头部市场有竞争力的大模型算力集群。

5.2 助力电信运营商提高视频分析能力

中国电信已经初步建成其第五张基础网络——天翼视联网，并成立了天翼视联科技有限公司，致力于构建全国统一的新型视频服务基础设施，提供高质量产品和平台服务，构建开放合作的视联生态，打造服务经济社会发展的国家级数字化平台。截至2023年10月底，天翼视联网的接入设备数量为6000万，原子能力调用超过3800万，输出视频路数超过1700万，AI日均调用超过4.2亿次。

中国移动也在积极布局视联网，并于2024年5月25日正式发布了“中国移动视联网”。已经实现了全国9200万路摄像头云端互联，服务个人家庭用户超9400万，服务政府企业客户超1000万。

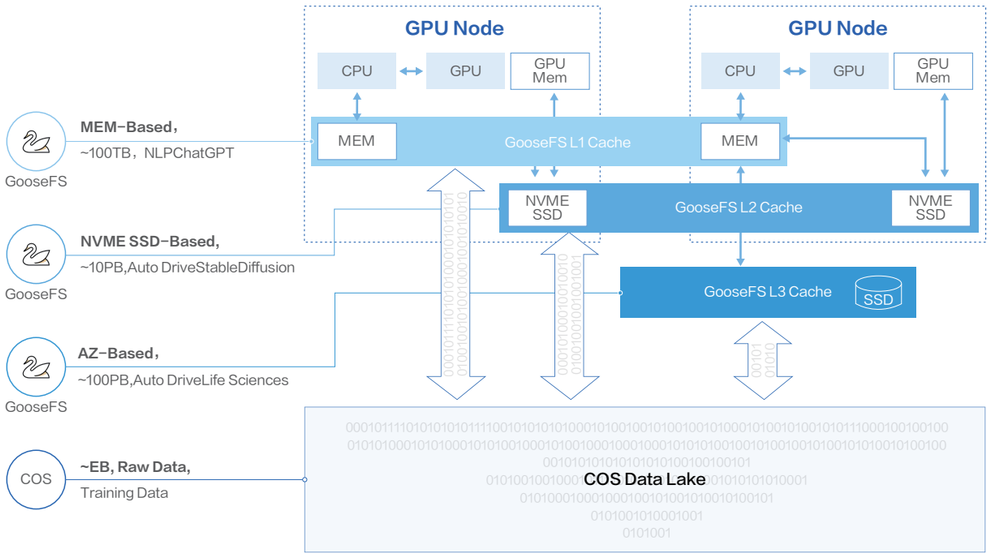
中国联通在视联网领域的布局也取得了显著进展。早在2019年，中国联通便打造了“1+4+X”智慧家庭新应用生态布局，视频监控是其中的重要应用之一。2024年，中国联通智慧家庭业务面向AI时代进行了全新升级，形成“AI+4+X”创新产品策略，其中“联通看家”将面向视联网升级。目前，中国联通视联网平台已经推出了文旅慢直播、扶贫电商直播、宠物喂养直播、店铺宣传直播、AI球场赛事直播等场景，以及“视联中国”APP。

视联网作为一个融合了视频、音频和文本等多种数据源的平台，其内容的多样性和复杂性要求一种能够全面理解和分析这些信息的技术。多模态大模型以其能够同时处理多种数据类型的能力，成为分析视联网内容的理想工具。它通过整合不同模态的数据，不仅提高了内容理解的准确性，还增强了对上下文信息的捕捉能力，从而为视联网内容的智能管理和自动化处理提供了强有力的技术支持。

01 实现方案

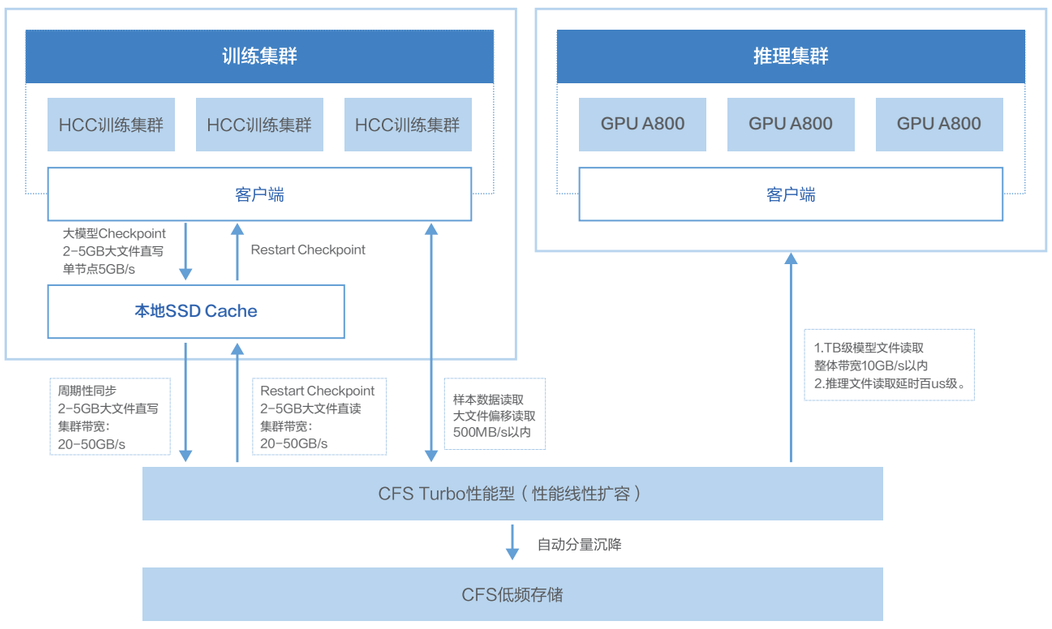
视觉感知大模型（Large Pretrained Model，简称LPM）作为人工智能领域的一项关键技术，在视频分析中的应用前景非常广阔。它们不仅能够提升视频分析的自动化水平和效率，还能够增强分析的准确性和深度。随着技术的不断进步，LPM将在安全监控、交通管理、医疗影像分析和内容创作等多个领域发挥关键作用，推动社会向更加智能化和自动化的方向发展。

LPM基于LLM同源范式，融合文本和位置的提示指令，统一基于分类、检测和分割的输出范式。权衡模型掌握的语义概念广度和目标定位精度，打造业界效果最好的感知模型。如下图所示，LPM处于“Recognition”区域的右侧，标示为红色矩形。LPM代表了具有高准确性和广泛范围（Scope）的模型，能够在识别任务中表现出色。它与ML-Decoder、Tag2Text、CLIP和BLIP等其他模型一起，位于“Recognition”区域内，表明它们都具备较高的准确性和广泛的应用范围。



▲ COS+GooseFS缓存加速方案

高性能并行文件存储CFS Turbo方案，采取多级缓存加速，基于全分布式架构，提供 100GB/s 带宽、1000 万 IOPS 的极致性能。并通过持久化客户端缓存技术，将裸金属服务器本地 NVMe SSD 和 Turbo 文件系统构成统一命名空间，实现微秒级延时，解决大模型场景大数据量、高带宽、低延时的诉求。同时，通过智能分层技术，自动对冷热数据分层，节省80%的存储成本，提供极致的性价比。



▲ CFS Turbo大模型解决方案

底层架构之上，针对大模型训练场景，新一代集群集成了训练加速引擎TACO Train，对网络协议、通信策略、AI 框架、模型编译进行大量系统级优化，大幅节约训练调优和算力成本。

此案例通过运营商优质的计算和IDC资源结合腾讯侧智算集群建设方案，建设高性能的智算集群，结合丰富的智算套件能力，提高算力集群的服务化水平，更好的发挥集群的利用率、生产效率、运维效率的水平，建成面向头部市场有竞争力的大模型算力集群。

5.2 助力电信运营商提高视频分析能力

中国电信已经初步建成其第五张基础网络——天翼视联网，并成立了天翼视联科技有限公司，致力于构建全国统一的新型视频服务基础设施，提供高质量产品和平台服务，构建开放合作的视联生态，打造服务经济社会发展的国家级数字化平台。截至2023年10月底，天翼视联网的接入设备数量为6000万，原子能力调用超过3800万，输出视频路数超过1700万，AI日均调用超过4.2亿次。

中国移动也在积极布局视联网，并于2024年5月25日正式发布了“中国移动视联网”。已经实现了全国9200万路摄像头云端互联，服务个人家庭用户超9400万，服务政府企业客户超1000万。

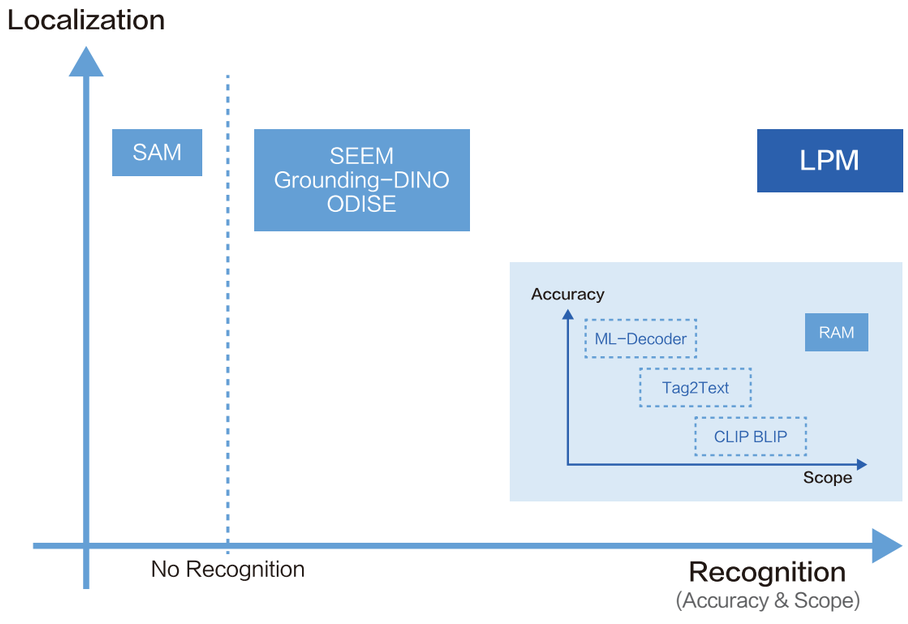
中国联通在视联网领域的布局也取得了显著进展。早在2019年，中国联通便打造了“1+4+X”智慧家庭新应用生态布局，视频监控是其中的重要应用之一。2024年，中国联通智慧家庭业务面向AI时代进行了全新升级，形成“AI+4+X”创新产品策略，其中“联通看家”将面向视联网升级。目前，中国联通视联网平台已经推出了文旅慢直播、扶贫电商直播、宠物喂养直播、店铺宣传直播、AI球场赛事直播等场景，以及“视联中国”APP。

视联网作为一个融合了视频、音频和文本等多种数据源的平台，其内容的多样性和复杂性要求一种能够全面理解和分析这些信息的技术。多模态大模型以其能够同时处理多种数据类型的能力，成为分析视联网内容的理想工具。它通过整合不同模态的数据，不仅提高了内容理解的准确性，还增强了对上下文信息的捕捉能力，从而为视联网内容的智能管理和自动化处理提供了强有力的技术支持。

01 实现方案

视觉感知大模型（Large Pretrained Model，简称LPM）作为人工智能领域的一项关键技术，在视频分析中的应用前景非常广阔。它们不仅能够提升视频分析的自动化水平和效率，还能够增强分析的准确性和深度。随着技术的不断进步，LPM将在安全监控、交通管理、医疗影像分析和内容创作等多个领域发挥关键作用，推动社会向更加智能化和自动化的方向发展。

LPM基于LLM同源范式，融合文本和位置的提示指令，统一基于分类、检测和分割的输出范式。权衡模型掌握的语义概念广度和目标定位精度，打造业界效果最好的感知模型。如下图所示，LPM处于“Recognition”区域的右侧，标示为红色矩形。LPM代表了具有高准确性和广泛范围（Scope）的模型，能够在识别任务中表现出色。它与ML-Decoder、Tag2Text、CLIP和BLIP等其他模型一起，位于“Recognition”区域内，表明它们都具备较高的准确性和广泛的应用范围。



▲ 不同模型在自然语言处理和计算机视觉任务中的定位和识别能力

LPM在视频分析中展现出显著的优势。首先，它们能够实现自动化和高效的视频分析流程，减少人工干预，提高处理速度。其次，这些模型经过大量数据训练，具有高准确率和鲁棒性，能够在不同环境和条件下稳定工作。此外，它们支持实时视频流分析，为需要快速响应的应用场景提供支持。视觉感知大模型还能够同时执行多种视频分析任务，如目标检测、行为识别和场景分类，提供丰富的上下文感知和数据驱动的决策支持。

02 案例价值

LPM通过深度学习技术，为运营商视联网提供了强大的视频分析能力，从而显著提升了用户体验和管理效率，为运营商视联网带来显著的价值：

- 提升视频分析能力：视觉感知大模型能够自动识别视频中的对象、行为和场景，提供更准确的视频内容分析。
- 增强用户体验：通过智能视频推荐和内容过滤，提升用户在视联网平台上的观看体验。
- 优化资源管理：自动化的视频监控和分析有助于运营商更有效地管理网络资源，提高服务质量。
- 促进业务创新：大模型的应用为运营商提供了开发新服务和应用的机会，如智能安防、智慧城市。

在电信运营商视联网中应用LPM，实现了实时视频监控、智能内容管理、安全防护增强以及生态合作的深化。这些模型通过高效的图像语义识别和场景识别，提高了视频监控的实时性和准确性。同时，它们优化了视频内容的存储和检索，提升了内容管理效率，并识别潜在安全威胁，加强了网络安全防护。此外，大模型的应用还拓宽了运营商与内容提供商、设备制造商等合作伙伴的合作空间，促进了视联网生态的共同繁荣。

5.3 为 5G 视频彩铃提供内容制作能力

5G视频彩铃是一种依托5G网络的先进通信服务，它允许用户在通话等待期间享受短视频内容，从而提升通话体验。这种服务利用5G网络的大带宽和低时延特性，提供更丰富的视觉和互动体验。用户可以自定义3D形象作为视频彩铃，展示个性化社交名片，或通过IMS DC实时互动视频彩铃功能，在观看视频时进行点赞和一键设置同款彩铃，增强互动性。此外，超高清自适应视频彩铃能够根据用户设备的能力自动选择播放相应清晰度的视频，提供沉浸式观看体验。随着5G技术的不断发展，5G视频彩铃正在推动短视频社交平台的创新，为用户提供更便捷、有趣的互动体验，并有望在全球范围内得到更广泛的应用。AIGC技术已经逐渐成熟，能低成本的服务于不同的内容创作场景中，能为5G视频彩铃提供了个性化和多样化的内容生成能力，极大地丰富了用户的通话体验。

01 实现方案

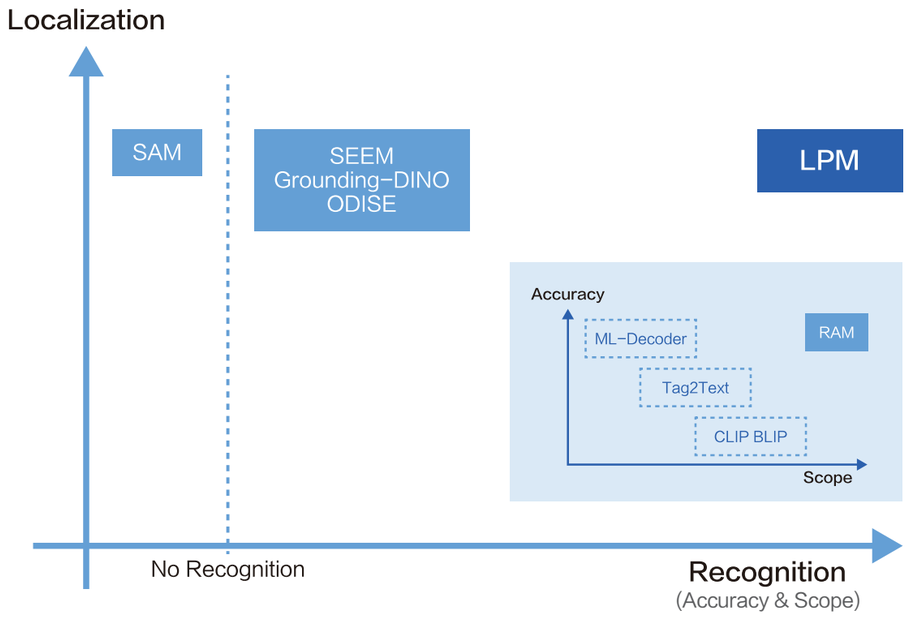
通过AIGC技术完成5G视频彩铃的制作是一个高效且用户友好的过程。用户首先确定彩铃的主题，然后输入相关的文本提示。AIGC系统根据这些提示，自动从大量素材中选择或生成匹配的图像和视频片段，并配上相应的背景音乐。用户可以对AIGC生成的草稿进行个性化编辑，如更换图片、调整视频长度或修改音乐风格，以确保最终作品符合自己的创意和情感表达。

02 使用AIGC技术，完成图片风格化

图片风格化是一种将艺术风格或特定的视觉效果应用到图片上的过程，它通过调整颜色、纹理和光影效果，模拟不同材料或艺术手法，使图片呈现出独特的艺术表现。现代图片风格化技术常利用深度学习模型，尤其是卷积神经网络，学习并应用不同风格的特征。风格迁移是其中一种流行技术，它将内容图片与风格图片结合，创造出具有新颖视觉效果的新图片。用户还可以通过交互式界面选择和调整风格效果，增加个性化体验。图片风格化技术广泛应用于艺术创作、广告设计、影视制作和游戏开发等领域，为视觉内容增添吸引力。随着人工智能技术的发展，图片风格化方法不断进步，能够实现更加精细和多样化的风格效果，拓展了创意的边界，使得用户能够轻松尝试和实现各种独特的视觉风格。

以下图为例，选择油画风格，输入Prompt：“亚麻色头发美丽的少女在森林里抚摸一只鹿”，调用算法生成图片。

▲ 图片风格化



▲ 不同模型在自然语言处理和计算机视觉任务中的定位和识别能力

LPM在视频分析中展现出显著的优势。首先，它们能够实现自动化和高效的视频分析流程，减少人工干预，提高处理速度。其次，这些模型经过大量数据训练，具有高准确率和鲁棒性，能够在不同环境和条件下稳定工作。此外，它们支持实时视频流分析，为需要快速响应的应用场景提供支持。视觉感知大模型还能够同时执行多种视频分析任务，如目标检测、行为识别和场景分类，提供丰富的上下文感知和数据驱动的决策支持。

02 案例价值

LPM通过深度学习技术，为运营商视联网提供了强大的视频分析能力，从而显著提升了用户体验和管理效率，为运营商视联网带来显著的价值：

- 提升视频分析能力：视觉感知大模型能够自动识别视频中的对象、行为和场景，提供更准确的视频内容分析。
- 增强用户体验：通过智能视频推荐和内容过滤，提升用户在视联网平台上的观看体验。
- 优化资源管理：自动化的视频监控和分析有助于运营商更有效地管理网络资源，提高服务质量。
- 促进业务创新：大模型的应用为运营商提供了开发新服务和应用的机会，如智能安防、智慧城市。

在电信运营商视联网中应用LPM，实现了实时视频监控、智能内容管理、安全防护增强以及生态合作的深化。这些模型通过高效的图像语义识别和场景识别，提高了视频监控的实时性和准确性。同时，它们优化了视频内容的存储和检索，提升了内容管理效率，并识别潜在安全威胁，加强了网络安全防护。此外，大模型的应用还拓宽了运营商与内容提供商、设备制造商等合作伙伴的合作空间，促进了视联网生态的共同繁荣。

5.3 为 5G 视频彩铃提供内容制作能力

5G视频彩铃是一种依托5G网络的先进通信服务，它允许用户在通话等待期间享受短视频内容，从而提升通话体验。这种服务利用5G网络的大带宽和低时延特性，提供更丰富的视觉和互动体验。用户可以自定义3D形象作为视频彩铃，展示个性化社交名片，或通过IMS DC实时互动视频彩铃功能，在观看视频时进行点赞和一键设置同款彩铃，增强互动性。此外，超高清自适应视频彩铃能够根据用户设备的能力自动选择播放相应清晰度的视频，提供沉浸式观看体验。随着5G技术的不断发展，5G视频彩铃正在推动短视频社交平台的创新，为用户提供更便捷、有趣的互动体验，并有望在全球范围内得到更广泛的应用。AIGC技术已经逐渐成熟，能低成本的服务于不同的内容创作场景中，能为5G视频彩铃提供了个性化和多样化的内容生成能力，极大地丰富了用户的通话体验。

01 实现方案

通过AIGC技术完成5G视频彩铃的制作是一个高效且用户友好的过程。用户首先确定彩铃的主题，然后输入相关的文本提示。AIGC系统根据这些提示，自动从大量素材中选择或生成匹配的图像和视频片段，并配上相应的背景音乐。用户可以对AIGC生成的草稿进行个性化编辑，如更换图片、调整视频长度或修改音乐风格，以确保最终作品符合自己的创意和情感表达。

02 使用AIGC技术，完成图片风格化

图片风格化是一种将艺术风格或特定的视觉效果应用到图片上的过程，它通过调整颜色、纹理和光影效果，模拟不同材料或艺术手法，使图片呈现出独特的艺术表现。现代图片风格化技术常利用深度学习模型，尤其是卷积神经网络，学习并应用不同风格的特征。风格迁移是其中一种流行技术，它将内容图片与风格图片结合，创造出具有新颖视觉效果的新图片。用户还可以通过交互式界面选择和调整风格效果，增加个性化体验。图片风格化技术广泛应用于艺术创作、广告设计、影视制作和游戏开发等领域，为视觉内容增添吸引力。随着人工智能技术的发展，图片风格化方法不断进步，能够实现更加精细和多样化的风格效果，拓展了创意的边界，使得用户能够轻松尝试和实现各种独特的视觉风格。

以下图为例，选择油画风格，输入Prompt：“亚麻色头发美丽的少女在森林里抚摸一只鹿”，调用算法生成图片。

▲ 图片风格化

03 使用AIGC技术，完成以图生图

以图生图是一种先进的人工智能图像处理技术，它通过深度学习模型分析原始图片并根据用户指令生成新图片。这一过程包括图像内容分析、用户指令识别、应用风格迁移或内容生成，以及图像合成等多个步骤。利用如生成对抗网络（GANs）或卷积神经网络（CNNs）等深度学习模型，AI能够理解并实现从风格转换到内容修改等多样化的图像处理需求。此外，一些工具还提供交互式编辑功能，使用户能够对生成的图片进行进一步的微调。以图生图技术在艺术创作、设计、广告和娱乐等多个领域有广泛应用，为用户带来了丰富的创意空间和个性化体验。随着人工智能技术的持续进步，以图生图技术在生成质量和多样性上不断提升，使得用户即使没有专业技能也能轻松创作出专业级别的图像作品。

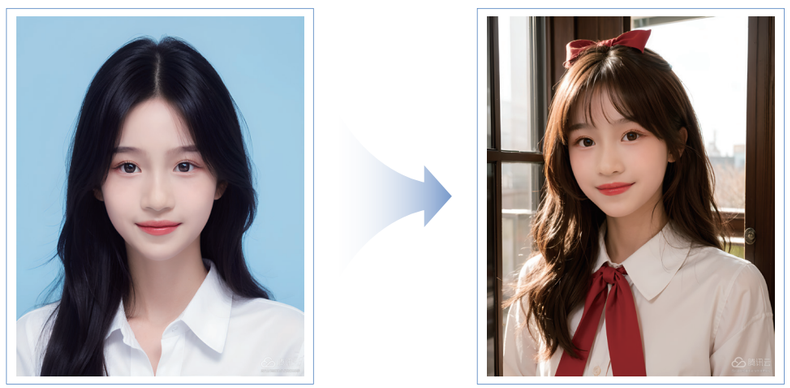
以下图为例，上传一张照片，选择选择“最终幻想”风格，调用算法，生成图片。

▲ 以图生图

04 使用AIGC技术，完成AI写真

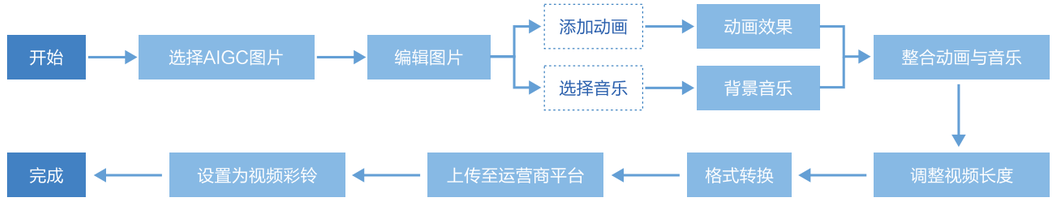
AI写真是一种利用人工智能技术生成和编辑肖像或场景图片的过程，它通过深度学习模型如生成对抗网络（GANs）来模拟复杂的图像特征，实现逼真或具有特定艺术效果的图片创作。这种技术能够进行人脸合成、风格化渲染、背景移除与替换、图像修复、色彩校正与增强等多种操作。用户可以通过交互界面输入指令或进行微调，以获得所需的视觉效果。AI写真的应用场景广泛，包括电影制作、游戏开发、虚拟现实、广告摄影和社交媒体等。然而，随着技术的发展，AI写真也引发了关于肖像权、版权和深度伪造的伦理和隐私问题，需要在创新应用的同时考虑相应的法律法规和道德标准。总体而言，AI写真技术为用户提供了强大的图像创作和编辑工具，使得创建高质量、个性化的图片变得更加容易和快捷。

以下图为例，上传一张照片，选择选择各类写真风格，调用算法生产图片。



▲ AI写真

通过AIGC生成的图片可以通过一系列步骤转化为5G视频彩铃，首先从AIGC创作的图片中选择适合的素材，并使用视频编辑软件添加动画效果，如淡入淡出、平移或缩放，以赋予图片动态感。接着，配上合适的背景音乐，调整视频长度以符合运营商的时间限制，并进行格式转换以适应5G视频彩铃的标准格式。完成编辑和渲染后，将视频上传至运营商的平台，并设置为个人彩铃。这样，每当有人呼叫时，他们就能看到这段个性化的5G视频彩铃。整个过程不仅提供了一种新颖的自我表达方式，还随着AI技术的发展，变得更加自动化和便捷，使用户能够轻松地将静态图像转化为引人注目的动态视频内容。



▲ 使用AIGC技术的5G视频彩铃制作流程

▲ 应用AI写真技术的5G视频彩铃

05 案例价值

通过AIGC，用户可以享受到根据个人兴趣和偏好定制的实时视频彩铃，降低了视频创作门槛，让每个人都能轻松创作出高质量的个性化视频内容。此外，AIGC增强了视频彩铃的互动性，通过5G网络的高速率和低延迟特性，实现了实时点赞、评论和分享等功能。AIGC技术还通过智能推荐系统提升了用户体验，根据用户的观看习惯和反馈提供定制化内容。同时，它帮助运营商和内容提供商扩展了内容生态，吸引了更多的合作伙伴和创作者。商业品牌可以利用AIGC技术定制视频彩铃，

03 使用AIGC技术，完成以图生图

以图生图是一种先进的人工智能图像处理技术，它通过深度学习模型分析原始图片并根据用户指令生成新图片。这一过程包括图像内容分析、用户指令识别、应用风格迁移或内容生成，以及图像合成等多个步骤。利用如生成对抗网络（GANs）或卷积神经网络（CNNs）等深度学习模型，AI能够理解并实现从风格转换到内容修改等多样化的图像处理需求。此外，一些工具还提供交互式编辑功能，使用户能够对生成的图片进行进一步的微调。以图生图技术在艺术创作、设计、广告和娱乐等多个领域有广泛应用，为用户带来了丰富的创意空间和个性化体验。随着人工智能技术的持续进步，以图生图技术在生成质量和多样性上不断提升，使得用户即使没有专业技能也能轻松创作出专业级别的图像作品。

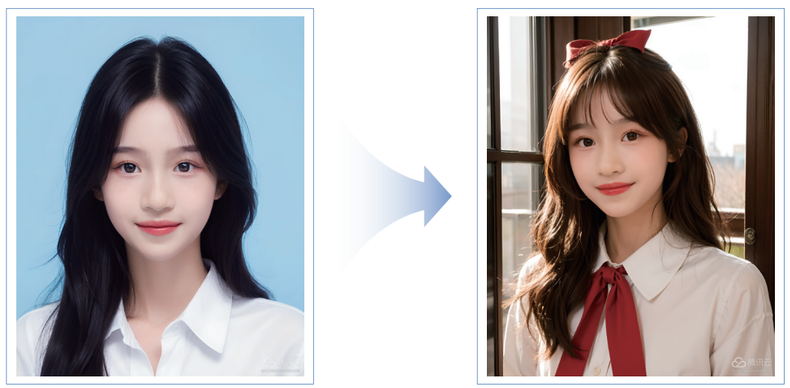
以下图为例，上传一张照片，选择选择“最终幻想”风格，调用算法，生成图片。

▲ 以图生图

04 使用AIGC技术，完成AI写真

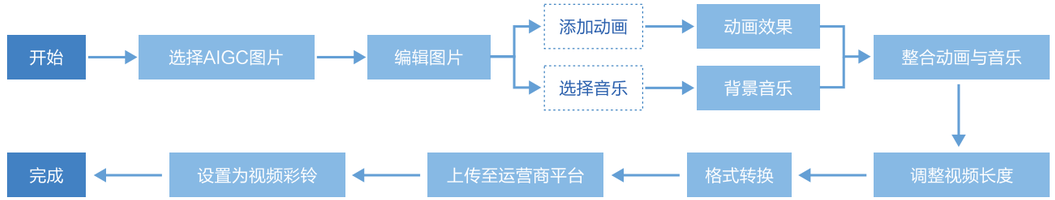
AI写真是一种利用人工智能技术生成和编辑肖像或场景图片的过程，它通过深度学习模型如生成对抗网络（GANs）来模拟复杂的图像特征，实现逼真或具有特定艺术效果的图片创作。这种技术能够进行人脸合成、风格化渲染、背景移除与替换、图像修复、色彩校正与增强等多种操作。用户可以通过交互界面输入指令或进行微调，以获得所需的视觉效果。AI写真的应用场景广泛，包括电影制作、游戏开发、虚拟现实、广告摄影和社交媒体等。然而，随着技术的发展，AI写真也引发了关于肖像权、版权和深度伪造的伦理和隐私问题，需要在创新应用的同时考虑相应的法律法规和道德标准。总体而言，AI写真技术为用户提供了强大的图像创作和编辑工具，使得创建高质量、个性化的图片变得更加容易和快捷。

以下图为例，上传一张照片，选择选择各类写真风格，调用算法生产图片。



▲ AI写真

通过AIGC生成的图片可以通过一系列步骤转化为5G视频彩铃，首先从AIGC创作的图片中选择适合的素材，并使用视频编辑软件添加动画效果，如淡入淡出、平移或缩放，以赋予图片动态感。接着，配上合适的背景音乐，调整视频长度以符合运营商的时间限制，并进行格式转换以适应5G视频彩铃的标准格式。完成编辑和渲染后，将视频上传至运营商的平台，并设置为个人彩铃。这样，每当有人呼叫时，他们就能看到这段个性化的5G视频彩铃。整个过程不仅提供了一种新颖的自我表达方式，还随着AI技术的发展，变得更加自动化和便捷，使用户能够轻松地将静态图像转化为引人注目的动态视频内容。



▲ 使用AIGC技术的5G视频彩铃制作流程

▲ 应用AI写真技术的5G视频彩铃

05 案例价值

通过AIGC，用户可以享受到根据个人兴趣和偏好定制的实时视频彩铃，降低了视频创作门槛，让每个人都能轻松创作出高质量的个性化视频内容。此外，AIGC增强了视频彩铃的互动性，通过5G网络的高速率和低延迟特性，实现了实时点赞、评论和分享等功能。AIGC技术还通过智能推荐系统提升了用户体验，根据用户的观看习惯和反馈提供定制化内容。同时，它帮助运营商和内容提供商扩展了内容生态，吸引了更多的合作伙伴和创作者。商业品牌可以利用AIGC技术定制视频彩铃，

开展新型的商业宣传，而公益组织也可以通过它来传播重要的社会信息。随着技术的不断进步，AIGC与5G视频彩铃的结合预示着更加沉浸和互动的通信体验，展现出广阔的应用前景和商业潜力。

AIGC技术在5G视频彩铃中的应用场景多样，包括个性化5G视频彩铃创作、AI视频模仿秀、商务彩铃等。这些应用不仅提升了用户的创作自由度，还增强了社交分享的活跃度。对企业而言，AIGC提供了一种新型的品牌传播方式，帮助提高宣传效率。同时，这种技术的应用也极大地丰富了用户的产品体验，让通话等待时间变得更加有趣和个性化。

5.4 AI 代码助手助力运营商研效提升

随着科技的快速发展和数字化转型的推进，软件开发已经成为当今世界的重要驱动力之一。企业和开发者面临着越来越多的挑战，如市场竞争加剧、技术更新换代、开发周期缩短以及对高质量代码的需求不断增长。为了应对这些挑战，人工智能（AI）技术逐渐应用于软件开发领域，以提高开发效率、降低开发成本并提升代码质量。

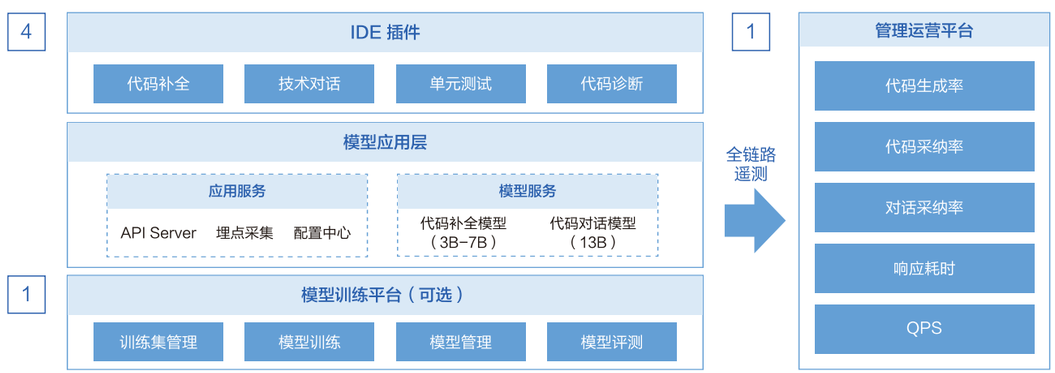
近年来，基于大型语言模型（LLM）的 AI 技术取得了显著的进步，同时，腾讯、阿里、智谱等公司也推出了以大模型为代表的大模型产品。这些技术的发展为软件开发领域提供了新的可能性，使得基于 LLM 的代码辅助工具成为现实，促使研效提升。

在此背景下，AI代码助手作为一款基于 LLM 大模型技术的代码辅助工具应运而生。通过结合先进的自然语言处理（NLP）技术和大量编程知识库，腾讯、阿里、智普等公司的AI代码助手能够理解开发者的需求，自动生成代码片段，提供智能建议，并帮助开发者在各种编程语言中进行代码上下文识别、补全和重构。同时，还具有单元测试生成和代码修复功能，可以对开发者编写的代码进行实时分析，自动生成相关的单元测试代码，提醒开发者注意潜在的问题和改进点，促进软件质量提升。

01 实现方案

AI代码助手模型采用通用大模型，它可能使用了基于Transformer的模型架构。Transformer是一种利用自注意力机制（Self-AttentionMechanism）的深度学习模型，可以捕捉输入数据中的长距离依赖关系。

在AI代码助手中，训练了两个大型编程语言模型，分别为代码补全模型和技术对话模型。其中，针对代码补全场景，训练了一个数十亿规模的大型编程语言模型，用于理解和生成代码，可以高效的支撑代码补全功能；针对技术对话、单元测试、代码诊断等场景，训练了一个百亿规模的大型编程语言模型，支持更丰富的自然语言理解、代码上下文理解并回答编程问题。



▲ AI代码助手功能框架

代码补全：AI 代码助手的关键能力在于智能代码补全，它可以通过分析注释描述和上下文代码语法语意，自动生成相关的业务逻辑代码和函数。特别对于常见的代码特征，如对象判空、循环定义、异常捕捉和日志定义等效果显著。此外，AI 代码助手代码补全的范围不仅限于代码光标的所处位置，它还支持中间补全、跨文件补全等高阶特性，让代码补全更丰富、更准确。

技术对话：AI 代码助手具备技术对话功能，可以实现情境感知的对话交互。通过对用户问题与相关代码上下文的理解，它能提供推理性的解答回复，并自动生成可能的后续提问推荐。此外，AI 代码助手支持常用指令，如生成注释、代码翻译、语言互转、代码解释等，使对话更高效。同时，它能将对话中的代码建议与编码区结合，提供建议对比，将生成的代码片段一键插入代码文件中，提升开发者的编程体验。

自动化测试：AI 代码助手具备自动化测试代码生成功能，可以让研发人员快速完成繁琐的单元测试编码工作，提高软件自动化覆盖率，进而提高软件质量。自动化测试功能能根据代码逻辑自动生成相应的单元测试用例，同时，AI代码助手会根据工程语境与参数属性来构造测试数据，并生成冒烟测试，保障软件质量。

代码诊断：AI代码助手的代码诊断功能可以提升代码可靠性、健壮性和易维护性。它能发现代码中潜在的异常问题、空指针和安全漏洞，帮助开发者在开发阶段及时修正；对于代码中的语法问题、编译错误和运行报错等问题，也能给出相应的修复方案。此外，它还针对代码的可读性和复杂度进行诊断。

02 案例价值

提高编码效率：通过自动补全和生成代码，AI 代码助手可以显著提高编程效率，使开发人员能够更快地完成代码编写任务，减少编写代码的时间。

提高编码质量：AI 代码助手能够检测和纠正语法错误和逻辑错误，降低代码的错误率，提高代码质量。有助于减少代码自测和调试的时间，并提高软件和应用的稳定性。

智能化建议辅助：提供智能化的提示和建议，协助开发人员更高效地完成代码编写任务，降低查找资料和尝试错误的时间，并降低出错的可能性。

代码规范与最佳实践：AI代码助手根据业界最佳实践和编程规范，智能生成高质量的代码，有助于保持项目的整洁和可维护性。

阿里云、智谱、腾讯云都推出了相应的代码助手产品。阿里云的通义灵码是一款基于通义大模型的智能编码助手，它提供代码智能生成、研发智能问答能力，支持行级/函数级实时续写、自然语言生成代码、单元测试生成以及代码优化和解释。智谱AI的CodeGeeX是一个免费的AI编程助手，专注于帮助开发者编写代码，每天能生成数百万行代码。腾讯云AI代码助手是基于混元大模型开发的编辑器插件，集成到VS　Code和JetBrains系列IDE中，提供智能补全代码信息、精准修复错误代码、清晰解释既有代码、按需生成单元测试和人工智能技术对话等功能。支持多种开发语言和框架，包括Python、JavaScript/TypeScript、Java等，帮助开发者提高研发效率并加速开发流程。同时该产品还具备自动补全、BUG诊断、生成测试等功能，通过技术对话与代码补全，辅助生成业务代码、注释、单元测试等内容，提升开发效率和代码质量。

开展新型的商业宣传，而公益组织也可以通过它来传播重要的社会信息。随着技术的不断进步，AIGC与5G视频彩铃的结合预示着更加沉浸和互动的通信体验，展现出广阔的应用前景和商业潜力。

AIGC技术在5G视频彩铃中的应用场景多样，包括个性化5G视频彩铃创作、AI视频模仿秀、商务彩铃等。这些应用不仅提升了用户的创作自由度，还增强了社交分享的活跃度。对企业而言，AIGC提供了一种新型的品牌传播方式，帮助提高宣传效率。同时，这种技术的应用也极大地丰富了用户的产品体验，让通话等待时间变得更加有趣和个性化。

5.4 AI 代码助手助力运营商研效提升

随着科技的快速发展和数字化转型的推进，软件开发已经成为当今世界的重要驱动力之一。企业和开发者面临着越来越多的挑战，如市场竞争加剧、技术更新换代、开发周期缩短以及对高质量代码的需求不断增长。为了应对这些挑战，人工智能（AI）技术逐渐应用于软件开发领域，以提高开发效率、降低开发成本并提升代码质量。

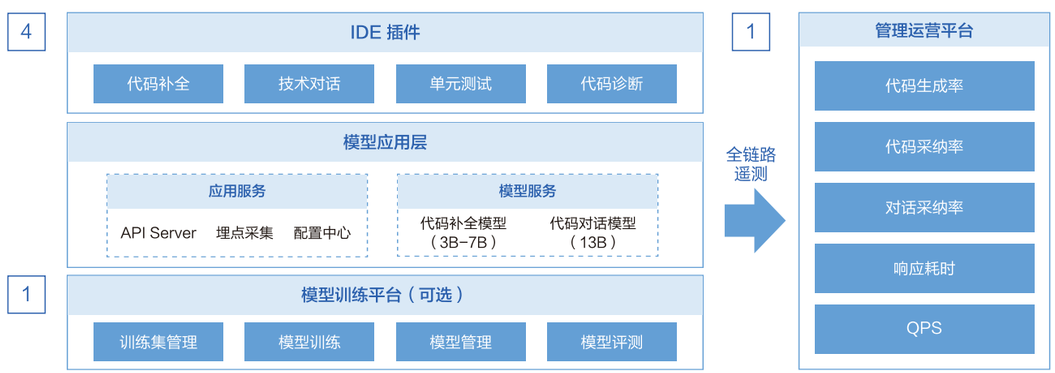
近年来，基于大型语言模型（LLM）的 AI 技术取得了显著的进步，同时，腾讯、阿里、智谱等公司也推出了以大模型为代表的大模型产品。这些技术的发展为软件开发领域提供了新的可能性，使得基于 LLM 的代码辅助工具成为现实，促使研效提升。

在此背景下，AI代码助手作为一款基于 LLM 大模型技术的代码辅助工具应运而生。通过结合先进的自然语言处理（NLP）技术和大量编程知识库，腾讯、阿里、智普等公司的AI代码助手能够理解开发者的需求，自动生成代码片段，提供智能建议，并帮助开发者在各种编程语言中进行代码上下文识别、补全和重构。同时，还具有单元测试生成和代码修复功能，可以对开发者编写的代码进行实时分析，自动生成相关的单元测试代码，提醒开发者注意潜在的问题和改进点，促进软件质量提升。

01 实现方案

AI代码助手模型采用通用大模型，它可能使用了基于Transformer的模型架构。Transformer是一种利用自注意力机制（Self-AttentionMechanism）的深度学习模型，可以捕捉输入数据中的长距离依赖关系。

在AI代码助手中，训练了两个大型编程语言模型，分别为代码补全模型和技术对话模型。其中，针对代码补全场景，训练了一个数十亿规模的大型编程语言模型，用于理解和生成代码，可以高效的支撑代码补全功能；针对技术对话、单元测试、代码诊断等场景，训练了一个百亿规模的大型编程语言模型，支持更丰富的自然语言理解、代码上下文理解并回答编程问题。



▲ AI代码助手功能框架

代码补全：AI 代码助手的关键能力在于智能代码补全，它可以通过分析注释描述和上下文代码语法语意，自动生成相关的业务逻辑代码和函数。特别对于常见的代码特征，如对象判空、循环定义、异常捕捉和日志定义等效果显著。此外，AI 代码助手代码补全的范围不仅限于代码光标的所处位置，它还支持中间补全、跨文件补全等高阶特性，让代码补全更丰富、更准确。

技术对话：AI 代码助手具备技术对话功能，可以实现情境感知的对话交互。通过对用户问题与相关代码上下文的理解，它能提供推理性的解答回复，并自动生成可能的后续提问推荐。此外，AI 代码助手支持常用指令，如生成注释、代码翻译、语言互转、代码解释等，使对话更高效。同时，它能将对话中的代码建议与编码区结合，提供建议对比，将生成的代码片段一键插入代码文件中，提升开发者的编程体验。

自动化测试：AI 代码助手具备自动化测试代码生成功能，可以让研发人员快速完成繁琐的单元测试编码工作，提高软件自动化覆盖率，进而提高软件质量。自动化测试功能能根据代码逻辑自动生成相应的单元测试用例，同时，AI代码助手会根据工程语境与参数属性来构造测试数据，并生成冒烟测试，保障软件质量。

代码诊断：AI代码助手的代码诊断功能可以提升代码可靠性、健壮性和易维护性。它能发现代码中潜在的异常问题、空指针和安全漏洞，帮助开发者在开发阶段及时修正；对于代码中的语法问题、编译错误和运行报错等问题，也能给出相应的修复方案。此外，它还针对代码的可读性和复杂度进行诊断。

02 案例价值

提高编码效率：通过自动补全和生成代码，AI 代码助手可以显著提高编程效率，使开发人员能够更快地完成代码编写任务，减少编写代码的时间。

提高编码质量：AI 代码助手能够检测和纠正语法错误和逻辑错误，降低代码的错误率，提高代码质量。有助于减少代码自测和调试的时间，并提高软件和应用的稳定性。

智能化建议辅助：提供智能化的提示和建议，协助开发人员更高效地完成代码编写任务，降低查找资料和尝试错误的时间，并降低出错的可能性。

代码规范与最佳实践：AI代码助手根据业界最佳实践和编程规范，智能生成高质量的代码，有助于保持项目的整洁和可维护性。

阿里云、智谱、腾讯云都推出了相应的代码助手产品。阿里云的通义灵码是一款基于通义大模型的智能编码助手，它提供代码智能生成、研发智能问答能力，支持行级/函数级实时续写、自然语言生成代码、单元测试生成以及代码优化和解释。智谱AI的CodeGeeX是一个免费的AI编程助手，专注于帮助开发者编写代码，每天能生成数百万行代码。腾讯云AI代码助手是基于混元大模型开发的编辑器插件，集成到VS　Code和JetBrains系列IDE中，提供智能补全代码信息、精准修复错误代码、清晰解释既有代码、按需生成单元测试和人工智能技术对话等功能。支持多种开发语言和框架，包括Python、JavaScript/TypeScript、Java等，帮助开发者提高研发效率并加速开发流程。同时该产品还具备自动补全、BUG诊断、生成测试等功能，通过技术对话与代码补全，辅助生成业务代码、注释、单元测试等内容，提升开发效率和代码质量。

5.5 行业大模型拓展运营商 CH 端场景

行业大模型是一种专为特定行业或领域设计的人工智能模型，按照行业分，当前比较成熟的行业大模型有：医疗行业大模型、金融行业大模型、政务大模型等。

01 实现方案

行业大模型需要通过大量数据训练和优化，具备高度专业化和智能化的特点。这些模型专注于处理特定行业的专业数据和问题，依赖大量行业数据进行训练，能够辅助或自动化决策过程，提供预测、分析和建议。它们具有多模态处理能力，能够处理和分析文本、图像、声音等多种类型的数据，同时具备持续学习与优化的能力，适应不断变化的行业需求。行业大模型在设计和应用时需考虑数据的安全性和隐私保护，符合相关的数据保护法规和标准。此外，它们需要具备良好的扩展性、跨平台和设备兼容性，以及用户友好的交互界面，确保非技术用户也能方便地使用。在开发和应用过程中，还需考虑其对社会的影响，确保符合伦理标准和社会责任。以医疗大模型为例，它可以帮助医生进行诊断和提供治疗建议。这个模型需要基于广泛且高质量的医学数据进行训练。该模型需经过大量医学文本数据训练，具有强大的理解和生成医学文本的能力，能够准确理解和回答医疗相关的问题。

02 案例价值

医疗大模型的应用在医疗健康领域产生了显著的积极效果。它们极大地提升了诊断的效率和准确性，通过快速分析医学影像和临床数据辅助医生做出更准确的判断。此外，医疗大模型优化了治疗方案，为患者提供个性化的治疗建议，同时加速了新药的研发流程，降低了成本并提高了药物研发的成功率。在改善患者体验方面，医疗大模型通过智能导诊和症状分析服务，使患者能够便捷地获取医疗信息。它们还提高了医疗服务的整体质量，通过自动化的质量控制减少了人为错误，并且帮助医院和医疗机构更有效地规划和使用医疗资源。医疗行业大模型同运营商的各类家庭终端：手机应用、带屏音箱、电视屏结合，应用于下面的这些场景中。

- 远程医疗服务：利用医疗大模型，运营商可以为家庭客户提供在线咨询、远程诊断和治疗建议，特别是对于慢性病管理和常规健康咨询。
- 智能健康监测：通过集成可穿戴设备和家庭健康监测工具，医疗大模型能够实时分析用户健康数据，预测疾病风险，并提供个性化健康建议。
- 紧急响应系统：在紧急医疗情况下，医疗大模型可以快速分析症状并指导用户进行适当的初步处理，同时协助联系紧急救援服务。
- 个性化医疗计划：根据家庭成员的健康状况和医疗历史，医疗大模型能够制定个性化的医疗和健康管理计划。
- 健康教育和促进：医疗大模型可以提供定制化的健康教育内容，帮助家庭成员了解如何预防疾病，提高健康意识。

电信运营商通过整合医疗行业大模型，不仅提升了医疗服务的可及性和效率，还优化了医疗资源的分配，增强了家庭健康管理，同时也为运营商自身创造了新的收入来源并增强了客户忠诚度。随着技术的进步和用户需求的增长，医疗大模型在运营商服务中的应用前景广阔，预计将成为推动家庭医疗服务创新和提升用户生活质量的关键因素。

06

电信运营商大模型发展展望

5.5 行业大模型拓展运营商 CH 端场景

行业大模型是一种专为特定行业或领域设计的人工智能模型，按照行业分，当前比较成熟的行业大模型有：医疗行业大模型、金融行业大模型、政务大模型等。

01 实现方案

行业大模型需要通过大量数据训练和优化，具备高度专业化和智能化的特点。这些模型专注于处理特定行业的专业数据和问题，依赖大量行业数据进行训练，能够辅助或自动化决策过程，提供预测、分析和建议。它们具有多模态处理能力，能够处理和分析文本、图像、声音等多种类型的数据，同时具备持续学习与优化的能力，适应不断变化的行业需求。行业大模型在设计和应用时需考虑数据的安全性和隐私保护，符合相关的数据保护法规和标准。此外，它们需要具备良好的扩展性、跨平台和设备兼容性，以及用户友好的交互界面，确保非技术用户也能方便地使用。在开发和应用过程中，还需考虑其对社会的影响，确保符合伦理标准和社会责任。以医疗大模型为例，它可以帮助医生进行诊断和提供治疗建议。这个模型需要基于广泛且高质量的医学数据进行训练。该模型需经过大量医学文本数据训练，具有强大的理解和生成医学文本的能力，能够准确理解和回答医疗相关的问题。

02 案例价值

医疗大模型的应用在医疗健康领域产生了显著的积极效果。它们极大地提升了诊断的效率和准确性，通过快速分析医学影像和临床数据辅助医生做出更准确的判断。此外，医疗大模型优化了治疗方案，为患者提供个性化的治疗建议，同时加速了新药的研发流程，降低了成本并提高了药物研发的成功率。在改善患者体验方面，医疗大模型通过智能导诊和症状分析服务，使患者能够便捷地获取医疗信息。它们还提高了医疗服务的整体质量，通过自动化的质量控制减少了人为错误，并且帮助医院和医疗机构更有效地规划和使用医疗资源。医疗行业大模型同运营商的各类家庭终端：手机应用、带屏音箱、电视屏结合，应用于下面的这些场景中。

- 远程医疗服务：利用医疗大模型，运营商可以为家庭客户提供在线咨询、远程诊断和治疗建议，特别是对于慢性病管理和常规健康咨询。
- 智能健康监测：通过集成可穿戴设备和家庭健康监测工具，医疗大模型能够实时分析用户健康数据，预测疾病风险，并提供个性化健康建议。
- 紧急响应系统：在紧急医疗情况下，医疗大模型可以快速分析症状并指导用户进行适当的初步处理，同时协助联系紧急救援服务。
- 个性化医疗计划：根据家庭成员的健康状况和医疗历史，医疗大模型能够制定个性化的医疗和健康管理计划。
- 健康教育和促进：医疗大模型可以提供定制化的健康教育内容，帮助家庭成员了解如何预防疾病，提高健康意识。

电信运营商通过整合医疗行业大模型，不仅提升了医疗服务的可及性和效率，还优化了医疗资源的分配，增强了家庭健康管理，同时也为运营商自身创造了新的收入来源并增强了客户忠诚度。随着技术的进步和用户需求的增长，医疗大模型在运营商服务中的应用前景广阔，预计将成为推动家庭医疗服务创新和提升用户生活质量的关键因素。

06

电信运营商大模型发展展望

6.1 技术演进，大模型建设与应用不断探索高效率、高精度、高适用性

随着大模型技术的发展及产业生态的丰富，AI大模型除了通过大参数量满足场景普适性需求之外，在计算资源日渐紧缺的背景下，未来将更加关注AI大模型的效率、精度以及场景的适用性。**从效率要求来看，**模型训练时常直接决定了AI大模型的生产成本，资源使用上，需要不断探索当前计算性能瓶颈的解决方案，提升有限的高性能计算资源利用效率，例如智算高性能传输网络、高性能并行计算调度方法、高性能训练框架等；模型设计上，避免为了追求大的模型参数量无效地复杂化模型结构，而是将更多的精力投入到模型优化上来。**从精度要求来看，**模型的推理精度直接决定了模型的使用体验，在AI大模型激烈竞争的当下，还有可能决定了产品和企业的存亡。不断提升AI大模型的推理精度，在法律法规要求的范围内，提升AI大模型的自然语言交互能力、响应速度、准确程度仍旧是未来一段时间AI大模型的研究重点。**从应用要求来看，**伴随大模型的推广以及行业大模型的落地，AI大模型的部署设备将会从传统的服务器向云、边、端侧算力扩展，应用场景也将出现更加强烈的行业属性，AI大模型如何在保证效率、精度的情况下更好的适配多样算力环境、多样应用环境，也将是AI大模型的建设方下一阶段要考虑的重要命题。

6.2 应用创新，电信运营商大模型要紧抓行业内、外痛点，打造差异差异化竞争力

中国的电信运营商是庞大的企业，从算力、网络等基础设施到AI大模型研发再到服务化输出，均可自成体系，因此其行业大模型一方面需要服务自身业务，另一方面需要服务外部市场用户。**对内，**电信运营商大模型需要深入了解企业优势能力与业务痛点，例如针对客户服务场景下的网络设备问题、通信问题进行定向优化，针对不同地区的网络质量、用户特征应用适配度更高的大模型等；**对外，**电信运营商相对互联网企业、人工智能企业在AI大模型的研究上并无显著优势，利用资源、数据优势打造具备差异化竞争力的AI大模型，结合电信业务进行推广和应用创新，或许将成为电信运营商在人工智能领域保持竞争力的有效措施。

6.3 跨领域协同，电信运营商与其他产业角色优势互补，谋求双赢

在AI大模型研究领域，电信运营商与互联网企业、算力设备制造企业等既是竞争也是合作关系，探索与其他产业角色跨优势领域的协同创新，强化合作关系，有望在全球人工智能科技竞赛中持续保持生命力。**与互联网企业协同，**电信运营商可集成互联网企业较为成熟的音视频、零售、游戏、供应链等解决方案，无需“重复造轮子”；**与算力设备制造企业协同，**在全球争夺算力资源的阶段，电信运营商可获得有效的算力供应链稳定性保障，优先探索大模型在新型算力设备上的性能表现。