2024 全球6G技术大会
GLOBAL 6G CONFERENCE

10.0A GPT与通信

未来移动通信论坛
FUTURE MOBILE COMMUNICATION FORUM

摘 要

在大数据、云计算等关键技术的共同推动下，以ChatGPT 为代表的GPT 大模型大量涌现，展现出了极富创造力的内容生成能力，提供了高度智能化的人机交互体验。一直以来，在通信方面存在许多传统方法难以精确建模或高效求解的技术难题，而GPT 展示出的潜力能够改进信息通信的服务，提升自智网络的性能。此外，GPT 的快速发展和广泛应用，也需要大带宽低时延高可靠的通信网络来支撑。

因此，本白皮书从通信从业者的角度，探讨了GPT 与通信的相互关系。具体来说，首先，第1章阐述了GPT 大模型的概念、发展历程和研究现状。其次，第2章探讨了GPT 赋能通信行业的崭新应用，以及在网络智能自治中的定位。再次，第3章对通信网络如何支持GPT 泛在应用进行了研究，给出了未来网络设计的典型思路。接着，第4章对GPT 和通信从独立演进到协同发展的过程进行了全面的分析，介绍了未来能够通过“6G+GPT”加速数字化和智能化转型的 行业。随后，第5章指出了“GPT+通信”融合发展所面临的五个最显著的问题， 并给出了一些解决思路。然后，第6章提出了对GPT 与通信融合发展的建议和对未来的展望。最后，第7章对本白皮书进行了总结。

目录

0. 引言

近年来，随着人工智能（Artificial Intelligence，AI）技术的不断发展，尤其是在强化学习、大模型和内容生成等方面不断取得突破，各行各业都在积极探索人工智能技术的应用。2022 年11 月底，OpenAI 公司发布了迅速爆火的聊天机器人程序ChatGPT，它具有惊人的自然语言理解和生成能力，引起了社会的广泛关注。2023 年3 月，升级版GPT-4 多模态大模型的发布，再次引发了生成式AI 的热潮，各类大模型纷纷涌现。

从文字对话交互开始，GPT 在短短几年的时间内深刻影响了人们的生产和生活，带来了巨大的变化，并且许多人认为它将继续带来颠覆性的改变。比尔·盖茨指出大模型是40 多年来最具革命性的技术进步；英伟达CEO 黄仁勋将大模型的出现称为AI 的“iPhone 时刻”；百度CEO 李彦宏在2023 中关村论坛上提出 大模型即将改变世界。可以看出，从ChatGPT 掀起的一片浪花，到席卷全球的 浪潮，GPT 大模型已经成为当下最受关注的话题之一，预示着生成式AI 的发展迎来重要转折，2023 年在AI 发展史上也将留下浓墨重彩的一笔。

作为人与人、人与自然、人与机器之间进行信息交流和传递的行业，通信行业与大模型技术的发展息息相关。通信行业本身数字化程度较高，需要处理繁杂的数据。GPT 的引入可以简化大量工作，为通信运营商带来显著的能力提升，尤其是在网络运维和业务交付方面将更加智能化。在大模型时代，随着GPT 技术的发展，算力、数据、算法需求将呈现爆炸式增长，同样需要通信基础设施来提供支撑。未来，GPT 如何赋能通信行业，通信行业又该如何支撑GPT，是每一个通信从业者都应该认真思考的问题。

因此，本白皮书以GPT 大模型的发展历程和最新研究进展为基础，一方面结合具体场景详细说明了GPT 在通信行业中的创新应用，另一方面研究了未来通信网络在架构和关键技术上对GPT 的原生支持。然后，将GPT 和通信相结合，提出了二者协同发展赋能重点行业的数字化智能化转型思路，同时也指出了二者融合发展过程中存在的问题与挑战。针对上述问题，给出了相应的发展建议和对未来的展望。最后，对本白皮书的全部内容进行了总结。本白皮书的完整章节架

构如图0-1 所示。

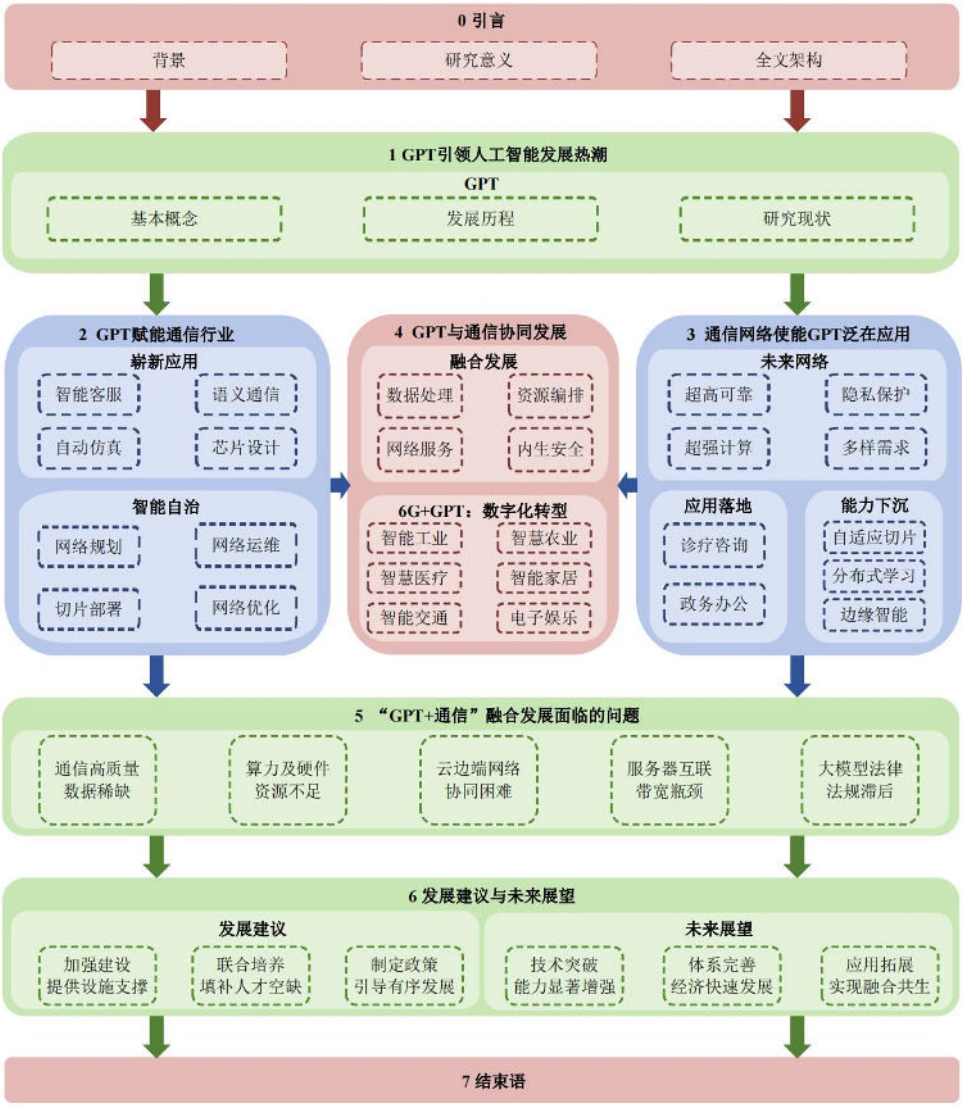


图0-1 白皮书章节架构图

本白皮书由北京理工大学牵头组织撰写，共有 18 家单位参与，包括移动、联通、电信3 大运营商，7 所一流高校，3 家知名企业，以及5 个业内领先的研究院所。从调研和跟进GPT 大模型前沿动态，到探究GPT 与通信的关系并构思白皮书的大纲，再到安排章节具体内容并分工撰写，总共历时8 个多月，有50 多位专家学者深度参与，在反复讨论修改迭代了二十余个版本之后才最终完成。在此期间，部分参与单位还成功联合申请了科技部的国际合作课题“基于大模型的云算网一体化多维智能编排关键技术研究”，从而更好地支持本白皮书的完成。

我们认为，AI 技术仍处于飞速发展阶段，GPT 大模型与通信网络能够实现相互融合、相互支持，不断拓展创新应用场景并完善生态建设，从而共同促进科技进步和千行百业的发展。

1.GPT 引领人工智能发展热潮

随着AI 和深度学习等技术的发展，“大模型”这一概念进入了人们的视野， 其中最引人注目的就是ChatGPT。2022 年11 月30 日，OpenAI 公司正式发布人工智能聊天机器人 ChatGPT，作为人工智能生成内容（Artificial Intelligence Generated Content，AIGC）在自然语言领域的代表，它强大的功能改变了许多人的工作和生活方式，掀起了全球范围内的AI 新浪潮，也吸引了工业界和学术界的广泛关注。2023 年3 月14 日，正式发布的GPT-4 进一步升级，文字输入限制大幅度放宽，回答准确性显著提高，甚至可以直接输入图像，生成歌词、创意文本等，实现风格变化，让人们再次感受到生成式AI 带来的震撼。2023 年11 月7 日，在首次开发者大会上，OpenAI 公司首席执行官Altman 向世界展示了GPT-4 Turbo。作为GPT 最新版本，它在数据质量、图像处理和语音转换等方面进行了更新，为开发者和用户带来了更多的可能性和机会。

那么ChatGPT 和GPT 是什么？它们经历了怎样的发展？又应该如何理解和应用呢？本章将从GPT 大模型出发，分别介绍GPT 的基本概念、发展历程和研究现状，以便读者对GPT 有更加全面和深入的了解。

1.1. GPT 基本概念

1.1.1. 生成式预训练转换器

GPT 的全称是Generative Pre-trained Transformer，即生成式预训练转换器，源于深度学习和自然语言处理（Natural Language Processing，NLP）领域。在过去的几年里，随着计算能力的提升和大数据的出现，NLP 领域取得了突破性的进展。GPT 作为一系列NLP 技术的集大成者，正是在这样的背景下应运而生的，如图1-1 所示。

G：Generative 。说明了GPT 的能力是自发生成内容。

P：Pre-trained 。说明了GPT 已经过预训练，可以直接使用。

T：Transformer 。说明了GPT 是基于Transformer 架构的语言模型。

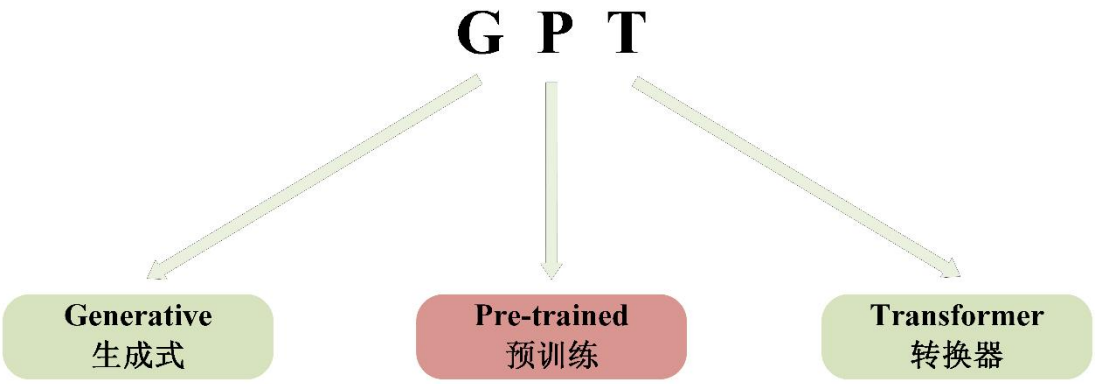


图 1-1 GPT 的含义

2017 年，Google 团队首次提出基于自注意力机制（Self-Attention Mechanism，SAM）的Transformer 模型，并将其应用于NLP[1]。OpenAI 应用了这项技术，于2018 年发布了最早的一代大模型GPT-1，此后每一代GPT 模型的参数量都呈爆炸式增长，2019 年2 月发布的GPT-2 参数量为15 亿，而2020 年5 月发布的GPT-3，参数量直接达到了1750 亿。

因此，ChatGPT 的“一夜爆火”并不是偶然，它是经过了很多人的努力，以 及很长一段时间的演化得来的。要了解GPT 的发展，首先应该了解大模型的概念以及Transformer 架构。

1.1.2. 大模型

一般来说，在ChatGPT 之前，被公众关注的AI 模型主要是用于单一任务的。比如，引燃了整个人工智能市场并促使其爆发式发展的“阿尔法狗”（AlphaGo），它基于全球围棋棋谱的计算，在2016 年轰动一时的“人机大战”中击败了围棋世界冠军李世石。但是从本质上来说，这种专注于某个具体任务而建立的AI 数据模型，和ChatGPT 相比，只能叫“小模型”。

大模型是指具有庞大的参数规模和复杂程度的机器学习模型，我们所提到的通常是大语言模型（Large Language Model，LLM）的简称。语言模型是一种人工智能模型，它被训练后可以理解和生成人类语言，而“大”的意思是指模型的参数量非常大，是相对于“小模型”而言的。

如图1-2 所示，这幅进化树图追溯了近些年大模型的发展历程，其中重点凸 显了某些最知名的模型，同一分支上的模型关系更近[2]。实心方块表示开源模型， 空心方块则是闭源模型。非Transformer 的模型都用灰色表示，而基于Transformer 的模型中，仅编码器模型是粉色分支，仅解码器模型是蓝色分支，编码器-解码器模型是绿色分支。

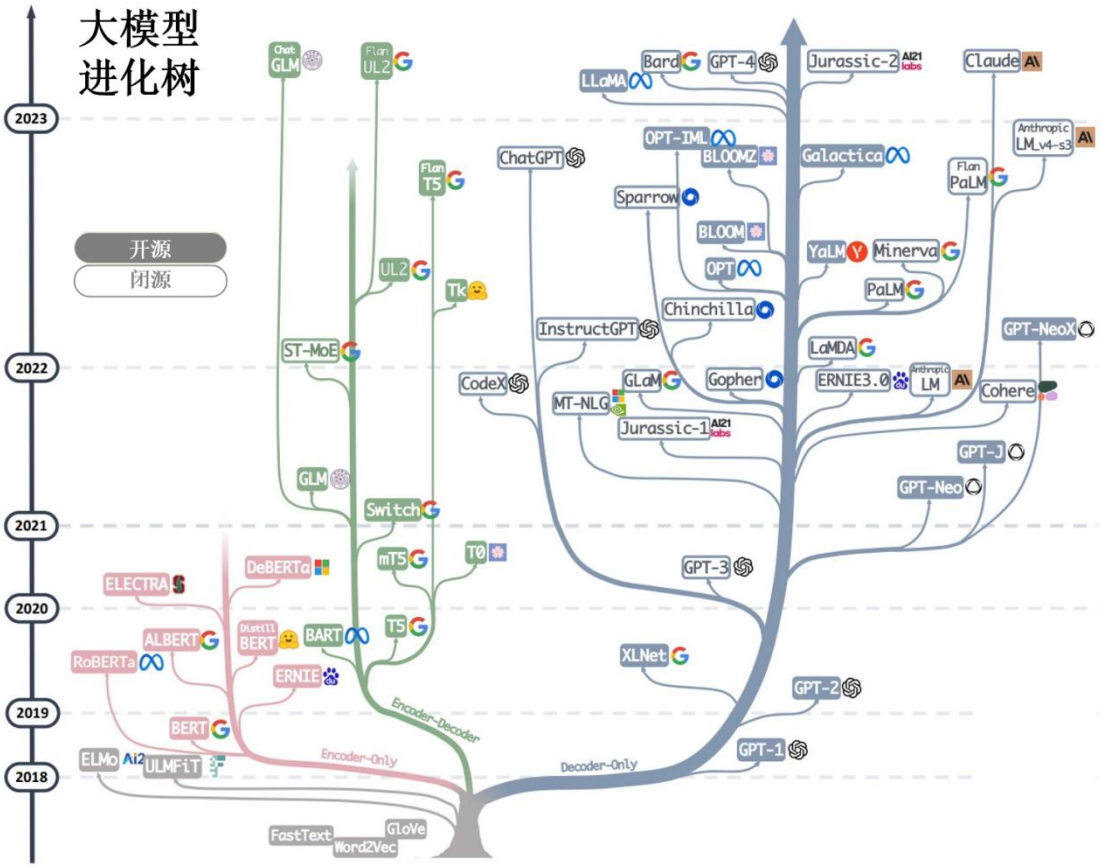


图1-2 大模型进化树

基于这幅进化树示意图，我们可以得出：仅解码器模型正逐渐成为LLM 发展的主导模型，且OpenAI 持续保持着其在LLM 方向上的领先地位。Meta 在开源和推动LLM 研究方面贡献卓越，但GPT-3 推出后LLM 开发有闭源的趋势。此外，仍有许多公司和机构在积极探索编码器-解码器模型，比如谷歌。

目前，国外大模型的主要发布机构有 OpenAI、Anthropic、Google 以及Meta 等，这些模型参数规模以百亿级和千亿级为主。发展至今，国外的头部GPT 大模型主要包括ChatGPT、Claude、Bard 和Llama 等。其中Bard 在谷歌发布了最新版原生多模态大模型Gemini 后，也正式更名为Gemini。

在这场全球参与的竞争中，我国也紧跟步伐，开发了许多大模型。包括腾讯的“混元”、阿里的“通义千问”、华为的“盘古”以及中国移动的“九天”系 列等。数据显示，截至2023 年10 月，国内10 亿参数规模以上的大模型厂商及 高校院所共计254 家，意味着“百模大战”正从上一阶段的“生下来”走向“用起来”的新阶段。图1-3 展示了目前国内外厂商开发的一些大模型。

数据来源：InfoQ研究中心

图1-3 国内外各类大模型

1.1.3. Transformer 架构

Transformer 架构是GPT 的重要基础，是一种SAM 的神经网络架构，广泛应用于NLP 领域的大模型中。其核心部分是编码器和解码器，即Encoder 和Decoder 。编码器把输入文本编码成一系列向量，解码器则将这些向量逐一解码成输出文本。在Transformer 提出之前，NLP 领域的主流模型是循环神经网络（Recurrent Neural Network，RNN），使用递归和卷积神经网络进行语言序列转换。

2017 年 6 月，谷歌大脑团队在 AI 领域的顶会 NeurIPS 发表了一篇名为 *Attention is All You Need* 的论文，首次提出了一种新的网络架构，即 Transformer，它完全基于 SAM，摒弃了循环递归和卷积。在八个P100 图形处理器（Graphics Processing Unit，GPU）上进行了仅仅 12 个小时的训练之后，Transformer 就可以在翻译质量方面达到更高的水平[1]，体现了很好的并行能力，成为当时最先进的 LLM。

图 1-4 给出了Transformer 的网络结构。Transformer 是由一系列编码器和解码器形成的，二者均由多头注意力层和全连接前馈网络组成。GPT 类似于 Transformer 的Decoder 部分，是一个自回归模型。

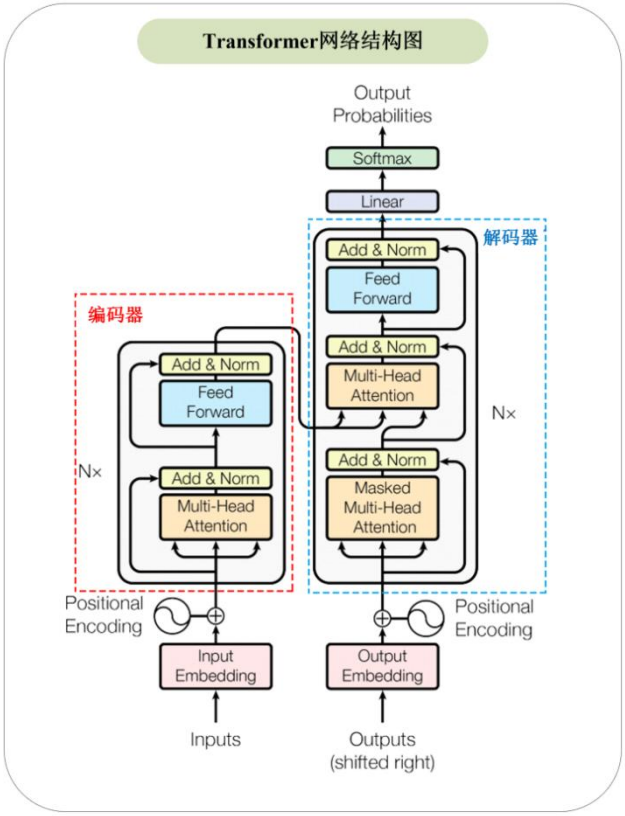


图 1-4 Transformer 网络结构图

Transformer 中的核心组件是多头注意力机制模块，如图1-5 所示。它需要三个指定的输入Q（代表查询）、K（代表键）、V（代表值），然后通过公式将 Q 和K 之间两两计算相似度，依据相似度对各个V 进行加权，得到注意力的计算结果。

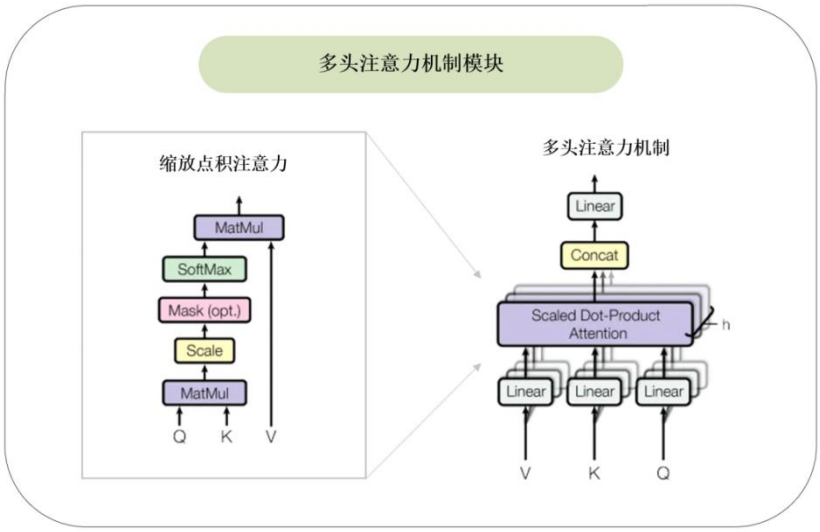


图1-5 多头注意力机制模块

多头注意力机制不是只计算一次注意力，而是将输入分成更小的块，然后并行计算每个子空间上的缩放点积注意力。这种结构设计能让每个注意力机制去优化每个词汇的不同特征部分，从而均衡同一种注意力机制可能产生的偏差，让模型能捕捉到不同层次的语义信息，增强模型的表达能力，提升模型效果。

1.2.GPT 发展历程

GPT 的发展历程主要可以分为两个阶段，在ChatGPT 之前侧重于不断增加 大模型的基础规模，并增强新能力。而ChatGPT 和GPT-4 则更侧重于增加人类反馈强化学习，理解人类意图，以提供更好的服务，如图1-6 所示。

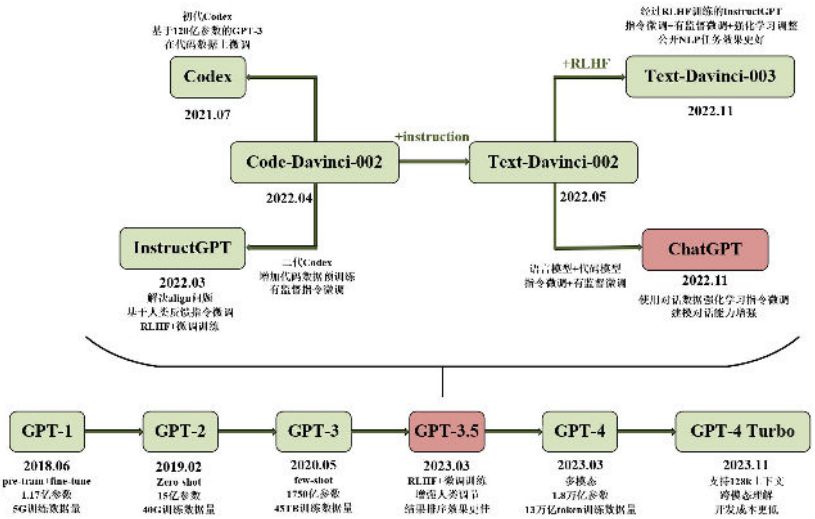


图1-6 GPT发展历程

①2018 年6 月，OpenAI 公司发表论文*Improving Language Understanding by Generative Pre-training*，正式发布了GPT-1[3]。

·基本思路：生成式预训练（无监督）+下游任务微调（有监督）。

·基于Transformer 的单向语言模型，解码器结构，共12 层。

·参数为1.17 亿，训练数据量5GB，模型规模和能力相对有限。

·上下文窗口为512 tokens。

②2019 年2 月，OpenAI 发表了最新进展，一篇*Language Models are Unsupervised Multitask Learners* 的论文，提出语言模型是无监督的多任务学，GPT-2 也随之诞生[4]。

·基本思路：去掉有监督，只保留无监督学习。

·48 层Transformer 结构。

·共15 亿个参数，数据训练量提升至40GB。

·上下文窗口为 1024 tokens。

③2020 年5 月，OpenAI 公司发表论文*Language Models are Few-Shot Learners*，构建了GPT-3 模型[5]。

·基本思路：无监督学习+in-context learning。

·采用了96 层的多头 Transformer。

·参数增大到1750 亿，基于45TB 的文本数据训练。

·上下文窗口为2048 tokens。

④2022 年3 月，OpenAI 再次发表论文*Training Language Models to Follow Instructions with Human Feedback*，介绍了人类反馈强化学习（Reinforcement Learning from Human Feedback，RLHF），并推出了InstructGPT 模型[6]。

·基本思路：RLHF+微调训练。

·增强了人类对模型输出结果的调节。

·对结果进行了更具理解性的排序。

ChatGPT 是 InstructGPT 的衍生，两者的模型结构和训练方式都一致，只是采集数据的方式有所差异，ChatGPT 更加注重以对话的形式进行交互。

⑤2023 年3 月，OpenAI 又发布了多模态预训练大模型GPT-4，再次进行了重大升级。

·基本思路：多模态。

·上下文窗口为 8195 tokens。

·1.8 万亿参数，13 万亿token 训练数据。

·强大的识图能力。

虽然目前GPT-4 在现实场景中的能力可能不如人类，但在各种专业和学术考试上都表现出明显超越人类水平的能力，甚至SAT 成绩（可以理解为美国高考成绩）已经超过了90%的考生，达到了考进哈佛、斯坦福等名校的水平。

1.3. GPT 研究现状

2023年10月12 日，分析公司 stateof.ai 发布了《2023年人工智能现状报告》（State of AI Report 2023）。该报告指出，Open AI的GPT-4仍然是全球最强大的LLM，生成式AI推动了生命科学的进步，并拯救了风险投资界[7]。大模型正不断实现技术突破，特别是在生命科学领域，在分子生物学和药物发现方面取得了有意义的进展。

2023年12月14日，《自然》（*Nature*）公布了十位2023年度人物，值得注意的是，聊天机器人Chat GPT因为占领了2023年的各种新闻头条，深刻影响了科学界乃至整个社会，被破例作为第11个“非人类成员”纳入榜单，以表彰生成式人工智能给科学发展和进步带来的巨大改变。目前，国内外对GPT大模型的研究不断

深入，纷纷开始研发自己的大模型，且应用的场景也越来越丰富。以Chat GPT为代表的大模型，正式开启了AI 2.0时代。

1.3.1. 国外研究现状

①美国

在美国，OpenAI、Anthropic 等初创企业和微软、Google 等科技巨头带领着美国在大模型的道路上飞速前进，同时各大公司也在不断提升自身的竞争力。Google 给Anthropic 投资3 亿美元以应对ChatGPT 的威胁，加入了AI 反馈强化学习（Reinforcement Learning from Artificial Intelligence Feedback，RLAIF）去减少人类的反馈，并于2022 年12 月发表论文*Constitutional AI: Harmlessnessfrom AI Feedback*，介绍了人工智能模型Claude；美国新媒体巨头Buzzfeed 因宣布计划采用ChatGPT 协助内容创作，股价两天涨了三倍；微软作为OpenAI 的主要投资方，也在利用ChatGPT 来增强其产品竞争力，补充专业知识、补齐数理短板。

②英国

2023 年4 月，英国政府宣布，向负责构建英国版人工智能基础模型的团队提供1 亿英镑的起始资金，以助英国加速发展人工智能技术。英国政府表示，该投资将用于资助由政府和行业共建的新团队，以确保英国的人工智能“主权能力”。 这一举措的目标是推广应用安全可靠的基础模型，并争取在2030 年将英国建设 成为科技“超级大国”。且针对GPT 等大模型应用在人工智能伦理方面的争议， 英国还发布了监管措施白皮书，并表示接下来监管机构将向各个组织发布使用指南和风险评估模板等其他工具及资源，来制定行业内的具体实施原则。

③欧洲

芬兰的Flowrite，是一个基于AI 的写作工具，可以通过输入关键词生成邮件、消息等内容。荷兰的全渠道通信平台 MessageBird 推出了自己的 AI 平台MessageBird AI，可以理解客户信息的含义并做出相应的响应。这两者都是在 GPT-3 的基础上运行的。德国在大模型的研发上也不断追赶。比如，谷歌 2023年 3 月 7 日推出

的多模态大模型 PaLM-E，就由柏林工业大学和谷歌共同打造。2024 年 2 月，欧洲生成式 AI 独角兽 Mistral AI 发布了最新大模型 Mistral Large 。该模型上下文窗口为 32K tokens，支持英语、法语、西班牙语、德语和 意大利语。作为新推出的旗舰模型，本次发布的 Mistral Large 在常识推理和知识问答上均表现出色，综合评分超过了 Gemini Pro 及 Claude 2，仅次于 GPT-4。

④韩国

韩国也是最早加入大模型研发的国家之一。目前，韩国在大模型领域的代表有NAVER、Kakao、KT、SKT 以及LG。韩国在半导体芯片方面的积累使其在大模型方面具有优势。目前韩国半导体企业正在积极结盟，以应对大模型发展带来的算力挑战。2022 年年底，NAVER 就开始和三星电子合作开发下一代人工智能芯片解决方案，即基于NAVER 推出的大模型HyperCLOVA 进行优化。此外，韩国在大模型的垂直应用上已经有比较多的探索，比如KoGPT 在医疗保健方面的应用、Exaone 在生物医药和智能制造方面的应用等。

⑤日本

作为一个小语种国家，日语面临缺乏语料的问题。日本最早公开上线的NLP 大模型是2020 年发布的NTELLILINK Back Office，当时它能实现文档分类、知识阅读理解、自动总结等功能，是在谷歌BERT 基础上开发的应用。

更有日本血统的生成式AI 其实是HyperCLOVA、Rinna 和ELYZA Pencil，但其中HyperCLOVA 和Rinna 也都有外国基因。HyperCLOVA 最早是韩国搜索巨头NAVER 在2021 年推出的，但HyperCLOVA 确实是第一个专门针对日语的大模型，它曾在2021 年举行的对话系统现场比赛中获得了所有赛道的第一名。ELYZA Pencil 则是由东京大学松尾研究所的AI 初创公司推出的大模型，算是真正意义上日本首次公开发布的生成式AI 产品。

1.3.2. 国内研究现状

许多人可能会认为，中国的大模型是从“文心一言”开始的，但“文心一言”其实只是一个对话工具，背后驱动它的还是大模型，而文心大模型早在2019年就在国内率先发布。这一年，大模型已经广泛应用于药品研发领域，各大科技企业也开始了对大模型产业的布局，并先后公布了各自的大模型项目。2021 年 3 月智源研究院发布了我国首个超大规模智能模型系统“悟道 1.0”。同年4 月， 阿里巴巴发布了中文社区最大规模的预训练语言模型PLUG，在当时有不少人将其称为“中文版GPT-3”。

近年来，国内在大模型领域取得了显著进展。从科研机构到企业，都加大了对大模型的投入力度，在算法、算力、数据等方面取得了重要突破。国内已经出现了一批具有国际竞争力的大模型，并在多个领域得到了广泛应用。

2023 年3 月16 日，基于文心大模型，百度发布了“文心一言”，成为中国 第一个类ChatGPT 产品。科大讯飞于2023 年5 月6 号发布中国版ChatGPT“讯飞星火认知大模型”，具有文本生成、语言理解、知识问答、逻辑推理、数学能力、代码能力和多模态能力七大核心能力。

1.3.3.国际组织

如今国际标准化组织（International Organization for Standardization，ISO）、国际电工委员会（International Electrotechnical Commission，IEC）等组织都已围绕关键术语等开展标准研究。2023 年3 月，欧洲电信标准化组织（European Telecommunication Standards Institute，ETSI）亦提出了有关人工智能透明度和可解释性的标准规范，旨在生成更多可解释的模型，同时保持高水平的模型性能。

第三代合作伙伴计划（3rd Generation Partnership Project，3GPP）规范包括了AI 在网络架构中的部署和使用，涵盖了AI 算法和架构的规范，还涉及了AI 数据的处理和管理标准。目前，3GPP 有四个工作组在进行AI/机器学习（Machine Learning，ML）标准化方面的研究工作，分别包括AI/ML for Air Interface、AI/ML for RAN 、AI/ML for 5GS 以及AI/ML for OAM。

2023 年 11 月，在由上海人工智能实验室与商汤科技联合主办的电气电子工程师学会（Institute of Electrical and Electronics Engineers，IEEE）“人工智能大模型”标准大会上，中国电子技术标准化研究院、上海人工智能实验室和华为云等 11 家单位共同发起成立了IEEE 大模型标准工作组。该工作组将协同国内外大模型产业力量，制定大模型技术规范、测评方法、安全可信、可靠决策等领域国际先进标准，为全球大模型产业技术创新和发展提供更好的支撑。

2.GPT 赋能通信行业

第 1 章中我们介绍了GPT 的概念、发展历程和研究现状等内容，可以看出，GPT 已被应用于众多领域，成为经济社会发展中重要的变革技术与关键力量，GPT 将为全球产业带来巨大飞跃和突破式发展。当前，GPT 已经实现了人与机器之间以多种形式进行“communication”的功能，接近甚至超越了人与人之间 以文本方式聊天的体验，这与通信行业支撑人们进行多种多样交流的作用相似。AI 应用在通信行业的落地，为信息通信基础设施的建设和运营开拓了新方案。作为AI 发展的新高度，GPT 引发的AI 即服务拥有更大的业务空间，能为通信行业的创新提供广阔的舞台。GPT 如何赋能通信行业应用，通信行业如何保障GPT 落地，这是通信从业者必须思考和回答的问题。

本章将重点介绍GPT 在通信行业的创新应用，彰显GPT 对通信细分领域的改革与推进作用。通过研究GPT 促进通信网络智能自治的方法，我们从网络规划、切片部署、网络运维和网络优化的角度对GPT 大模型如何赋能通信网络进行了分析。我们期待GPT 的飞速发展能够促进人工智能与通信产业的深度融合，加速构建下一代信息基础设施，助力经济社会的数字化转型。

2.1. GPT 催生通信新应用与新改革

GPT 百花齐放的崭新应用，为千行百业的发展带来了新的想象空间，也给通信行业带来新的机遇和挑战。GPT 的出现改变了传统的通信模式和应用场景，它突破了人与机器交互的界限，能提供更加智能、便利和个性化的通信体验，极大地提高了信息交互能力和行业应用能力。

GPT 大模型可作为工具来改进信息通信服务能力。首先，它在自然语言上的强大能力可用于提升智能客服、智慧运营和欺诈监测等运营服务功能，通信网络的巨量数据可用来训练通信网络大模型。其次，GPT 在自然语言上的成功，促进了语音、视觉等多模态数据技术的发展，这将为通信领域千行百业的数字化

转型赋能提供重要工具。最后，GPT 类大模型的运行和服务对算力和网络有着较高的要求，这会在一定程度上促进算网融合的建设，为更多大模型服务在通信行业落地和普及创造条件。

通过迭代训练海量数据，GPT 具备不断提升的上下文语义理解与交互能力，在众多应用场景中展现出无限潜力。目前，GPT 的应用主要集中在文本、图片、 音频、视频以及多模态内容的生成上，在摄影、游戏和传媒等领域的应用，通常 是在这些基本的应用的基础上，再进行定制化的开发或训练。例如，文本生成和分析[8]、软件测试[9][10]、领域专业聊天机器人[11]等，如图2-1 所示。

图2-1 GPT在通信领域的崭新应用

2.1.1. 智能客服

智能客服系统旨在为通信运营商的客户提供一种高效、灵活、可定制的解决方案，用于管理、维护运营商与客户之间的交互。将智能客服系统和GPT 结合，

可以发挥两者的技术优势，在智能语音助手、智能推荐、自助服务、社交媒体管理、个性化服务等多个方面，提高客户服务的质量和效率，满足客户日益增长的个性化需求，从而帮助企业更好地服务客户，提高竞争力和盈利能力。

1）增强智能客服的语义理解、情感识别

GPT 的自然语言处理能力，弥补了智能客服系统许多不足之处。GPT 准确地识别用户提问的主题和关键词，帮助智能客服系统更好地理解用户的需求，并识别用户的情感状态，从而提供更准确、更个性化的服务。

2）实现智能服务监管

GPT 可用于自动检测客服对话内容，识别潜在的违规行为或不当语言，例如侮辱、歧视、欺诈等，并预先筛选潜在问题对话，只将有可能违规的对话内容提交给人工审核，一定程度上减轻了审核人员的工作量。通过对GPT 模型的输出进行数据分析，快速了解和识别各类违规行为的趋势和模式，监管部门能够改进监管策略并及时采取相应的措施。

2.1.2. 自动化仿真

GPT 可以重构实验流程，为实现自动化仿真创造条件。GPT 是在大量文本上预先训练的，并且可以根据上下文提示进一步泛化。与传统的工作流程不同，它不需要每次改变模拟设置参数、底层机器学习算法或数据格式，用户只需要提供与预定义架构相关的参数，在对创建的模型进行解析后将其插入GPT 准备好的模板中，最后通过GPT 实现自动化仿真。

在仿真设计阶段，GPT 可以帮助设计师快速设计原型，从而使开发团队和相关人员更好地理解系统的工作流程和功能，提前发现问题和改进需求。基于GPT，设计师可以将自然语言描述作为输入，生成相应的交互原型，避免了手动构建的繁琐，提高原型的质量和准确性。此外，GPT 不仅可以辅助开发者完成常规的代码编写工作，还可以通过机器学习和自然语言处理技术，实现智能编程。通过理解开发者的意图，根据自然语言描述生成代码，并实现更加高级和复杂的功能。

Dragana Krstic 等人[12]提出了一种基于ChatGPT 的框架，可用于移动网络中的信道容量计算，实现自动化无线网络规划中的仿真过程，如图2-2 所示。在该框架中，ChatGPT 基于对话Agent 和Neo4j 图数据库的模型驱动方法，帮助进行自动化数据导入、图构建和机器学习相关的查询等多个步骤。其中，Neo4j 是一 种高性能和可扩展的图数据库管理系统，其基于图理论的数据库，专注于存储和 处理图结构的数据。结果显示，基于 ChatGPT 的服务质量（Quality of Service， QoS）估计方法在准确性和训练速度方面比基于深度神经网络的解决方案更优。此外，与传统基于手动生成代码的仿真流程相比，利用ChatGPT 自动化生成代码可以缩短仿真时间。

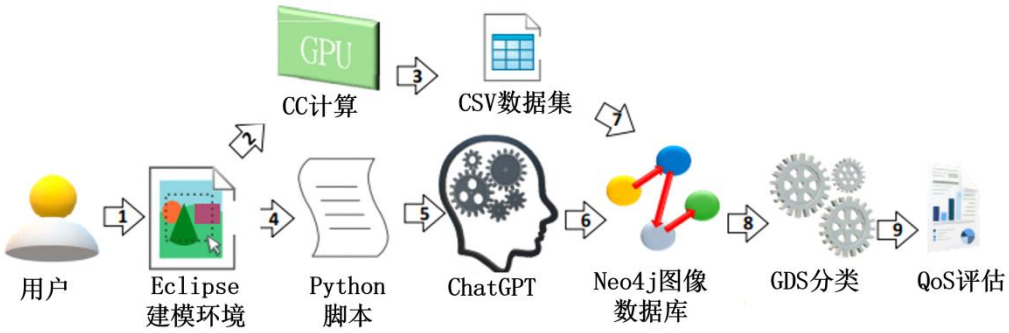


图2-2 GPT帮助进行信道容量分析实验

2.1.3. 增强语义通信

随着第六代移动通信系统（The Sixth Generation，6G）和物联网等新型网络技术的发展，万物智能互联成为时代所趋，语义通信有望成为未来通信网络的核心范式。然而，现有的语义通信受到缺乏上下文推理能力和背景知识的限制，同时，语义模型训练及语义知识图谱的构建将消耗巨大的时间与计算资源。因此，提升模型的训练效率，降低模型的训练成本，实现模型在网络中高效传输和部署，是语义通信的重要基础，也是所面临的关键挑战。引入GPT 相关技术，可以对输入进行语义理解和表示学习，并进行语义匹配任务。

文献[13]中提出了一种新的AI 辅助SemCom 网络框架，通过采用全局和局部GPT 模型，在基于GPT 的增强语义通信系统中，收发端分别部署语义编码模块和译码模块，模块对应的语义模型分别用于提取和恢复语义信息。通过在服务器中基于GPT 生成语义模型，并根据收发端的请求，动态部署适配的语义模型。同时，收发端将语义模型存储在各自的语义模型库中。发送端将原始信息输入语

义提取与表征模块，得到语义信息，并通过联合的语义编码和信道编码将语义信息转化为比特数据，再进行传输。接收端对接收到的比特数据进行联合信道译码、语义译码，以及语义信息恢复重建，恢复出原始信息。上下文、通信环境等背景因素会影响语义信息的恢复，语义译码模块对背景因素带来的误差可进行补偿。这种方法实现多模态语义内容理解、语义级联合信源信道编码，一定程度上提高了语义推理的可靠性和资源利用率，减少了传输流量和降低延迟，实现了更有效的语义传递，如图2-3 所示。

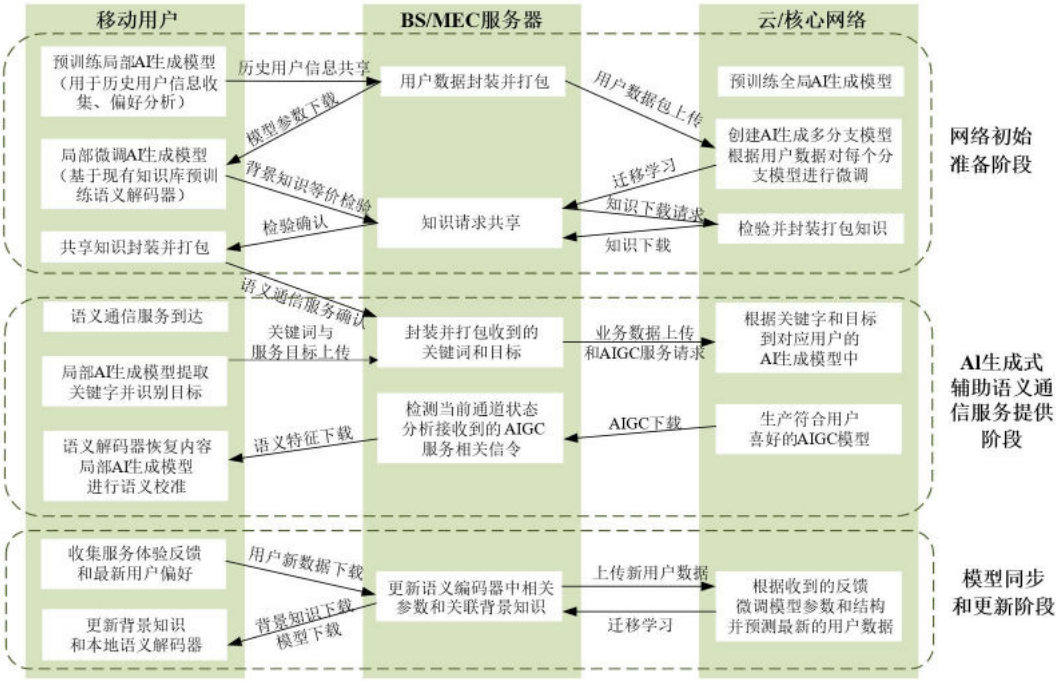


图2-3 增强语义通信

2.1.4. 重塑芯片设计领域

在通信领域，芯片设计起着至关重要的作用，可以说是通信技术发展的关键驱动力。芯片设计可以实现对各种通信协议的支持，例如以太网、无线通信、蓝牙、长期演进技术（Long Term Evolution，LTE）等，从而支持设备之间的通信和数据传输；芯片设计可以提供专门的硬件加速器、编解码器和信号处理器等，以支持高效的多媒体数据处理和传输。无线通信芯片拥有射频前端、调制解调器、功率放大器等功能，用于实现无线信号的收发和调制解调；芯片设计可以提供硬件安全功能，用于数据加密、身份认证和安全通信等，保护通信系统免受恶意攻击和数据泄露。

GPT 可以大大提高芯片设计的效率，缩短设计周期，并使设计过程更加自动化，进一步提高设计的效率和质量。GPT-4 可以降低芯片设计的门槛，使更多的人可以参与到芯片设计中来，这可能会带来更多的创新。

2023 年9 月，纽约大学Tandon 工程学院的研究人员[14]利用OpenAI 的GPT-4 模型，成功设计出了一个芯片，如图2-4 所示，这标志着AI 在硬件设计领域的重大突破。GPT-4 通过简单的英语对话，生成了可行的硬件描述语言（Hardware Description Language，HDL）代码，然后将基准测试和处理器发送到Skywater 130nm 穿梭机上成功流片。可以说GPT-4 在芯片设计中的应用，是AI 在硬件设计领域的一次重大突破，我们有理由相信，AI 将在未来的芯片设计领域发挥更大的作用，为我们带来更高效、更创新的芯片设计方案。

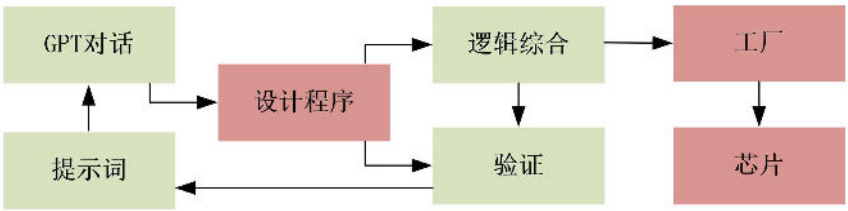


图2-4 GPT在芯片设计中的应用

2.2. GPT 促进通信网络智能自治

AI 赋能的自治网络是第五代移动通信系统（The Fifth Generation，5G）和后5G 网络发展的重要趋势，将为移动网络带来根本性变革。网络将由当前以人工干预为主的被动管理模式，逐步向网络自我驱动为主的自治管理模式转变。未来，智能化网络将通过业务数据、用户数据、网络状态数据等多维数据感知，基于AI 的智能分析，提供更加灵活高效的网络策略，从而实现网络高度自治，大幅提升移动网络全生命周期效率，降低网络运营成本。

自智网络[15]的核心理念在于通过AI 等技术的引入推动新一代通信网络向自配置、自治愈、自优化、自演进的方向发展。“AI+通信”已成为ITU 定义的6G 中六大场景之一[16]，包括辅助自动驾驶、设备间自主协作、辅助医疗应用、基于数字孪生的事件预测等新功能。目前，AI 已经初步实现了在网络智能自治领域的应用，全球多家运营商、设备商和第三方厂商已经开始了对网络智能自治的研

究，如图2-5 所示，包括网络规划、切片部署、网络运维和网络优化等应用案例。

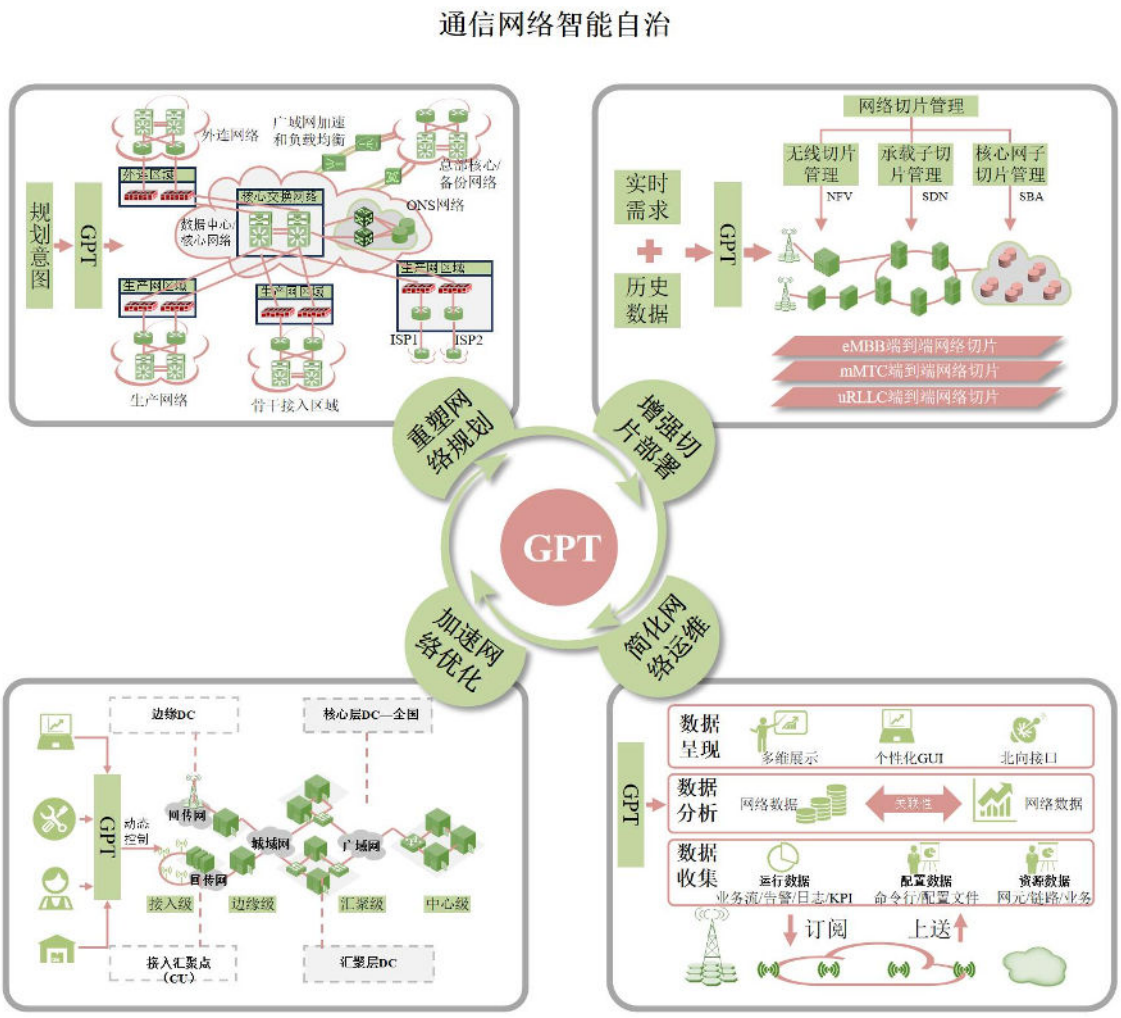


图2-5 GPT促进通信网络智能自治

2.2.1. GPT 重塑网络规划

由于预计未来几年无线接入设备数量将呈指数级增长，运营商需要扩大网络 基础设施的部署规模，以提供所需的容量。传统上，新基站选址是由无线网规专 业人士手动完成的 。在覆盖模拟工具的帮助下，根据关键性能指标（Key Performance Indicator，KPI）评估每个站点并对其进行排名，根据可用预算挑选出排名靠前的站点位置。然而，当可供选择站点数量较大时，传统方法成本巨大， 且很难准确地考虑每个涉及因素的影响。AI 驱动的规划方案可以为新蜂窝基站推荐最佳位置，帮助运营商降低网络规划成本。

针对最佳站点选择问题，Siddhartha Shakya 等人[17]提出了基于 AI 的选址方法，在此基础上，基于 GPT 进行网络站点选址规划，通过采集历史时空特征数 据，分析无线资源利用率的变化规律，监测和评估覆盖小区的 KPI。GPT 综合分析网络覆盖、用户分布和场景特征，通过无监督机器学习根据小区的属性同质聚类。监督回归模型捕捉不同小区属性之间的关系，如小区性能、用户吞吐量等。 在回归模型的基础上，构建仿真算法，估算拟新建站点的潜在流量负载。最后，基于计分排名机制对站点进行排名，排名靠前站点入围候选基站。

除站点选址外，天线设计也是基站规划阶段的重要工作。在天线的优化设计过程中，通常涉及的天线参数较多，天线的几何形状越来越复杂，天线性能要求之间的互相矛盾也频繁出现。将GPT 引入天线仿真设计，可以代替电磁仿真软件的角色，模拟应用场景对天线参数进行微调，结合粒子群智能优化算法[18]进行天线的快速仿真和优化设计，相比电磁仿真软件，可以进一步提升计算效率。

2.2.2. GPT 增强切片部署

网络切片的引入成功解决了不同业务场景的网络资源分配不均问题，给网络带来了极大的灵活性，使网络可以按需定制、实时部署、动态保障。在网络切片部署时，不同业务场景的切片对底层物理网络的资源需求不同，在部署的结构上也存在差异。传统算法难以解决多业务场景切片安全部署问题，利用GPT 相关技术，可以在实现端到端网络切片安全部署的同时，降低部署成本，保障更强的安全性。

网络切片的部署涉及虚拟网络功能（Virtualized Network Functions，VNF）的放置和相关链路的选择。VNF 的放置是指在满足网络容量的条件下，网络切片请求中的节点总能在物理网络上找到相对应的节点来承载请求。相比于传统启发式算法求解VNF 映射过程，GPT 可以对网络环境状况进行分析，根据业务场景需求智能调整网络参数，并做出业务资源需求预测，通过Agent 和环境的相互作用，执行特定的动作，更新网络资源利用情况，充分感知VNF 映射过程中的状态信息，对网络的变化情况做出及时的决策。如图2-6 所示，GPT 帮助获取网

络部署环境，并将物理节点信息以安全特征矩阵进行储存。Agent 定义为一个依靠GPT 计算物理节点映射概率的策略网络。GPT 辅助计算物理节点的安全特征矩阵输出物理节点映射概率，然后选择概率最大的物理节点并进行VNF 映射。之后，GPT 根据不同业务需求选择最合适的链路映射方案，以网络资源利用率作为奖励函数，给予Agent反馈，同时更新状态信息。

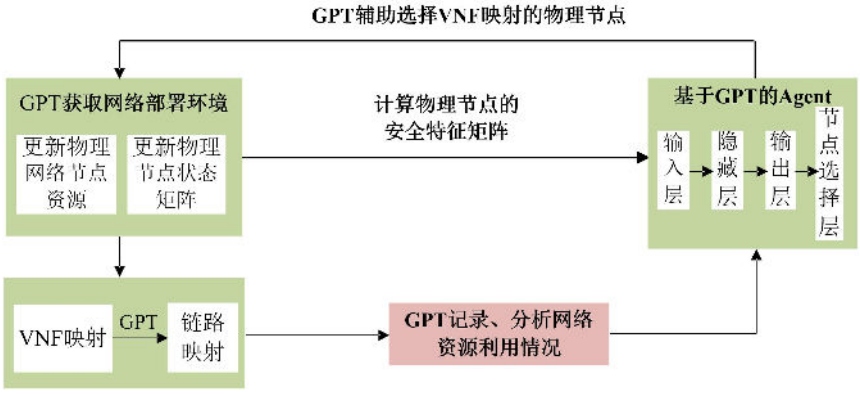


图2-6 GPT增强切片部署

2.2.3. GPT 简化网络运维

网络智能运维的典型应用场景包括异常检测、故障诊断、事件预警、效能优化等。在传统网络运维中，运维人员需要通过手动巡检和数据分析等方式获取网络状态信息，这种方式效率较低。通过引入GPT 相关技术，可以实时高效监测网络状态信息，并通过自动化运维的方式对网络进行分析和处理，从而有效地提升网络的稳定性和可靠性。

如图2-7 所示，网络采集器将实时网络信息发送给GPT，如设备的中央处理器（Central Processing Unit，CPU）、内存、网络拥塞信息、网络事件的日志信息等。GPT 快速进行统计分析，在此基础上结合不同网络业务场景对网络进行预测并给出相应运维决策，然后发送至数据库进行存储。另外，GPT 辅助对网络当前情况和预测情况进行可视化处理，更好地向运维人员展示网络现状和趋势，辅助运维人员更高效和更智能地进行网络运维。

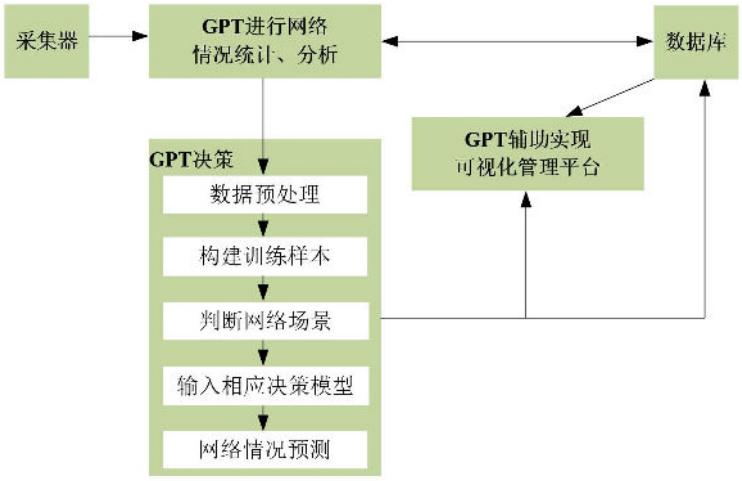


图2-7 GPT简化网络运维

此外，文献[19]中结合GPT 相关技术，以意图驱动的、特定于通信网络的机器学习模型和高级策略为目标，对底层组件进行了全面的解析，实现了“故障定位—策略生成—策略验证”的自主循环。与传统人工决策辅助的智能运维方法不同，该方法利用时空表征学习进行知识推理和网络运行状态检测，自动生成并验证故障恢复和多任务管理策略，通过绕行路由、资源编排等技术保证业务带宽和网络性能。此外，它还可以根据学习结果自主进行故障修复，以支持网络管理和状态调整，从而促进网络的自主运维。

2.2.4. GPT 加速网络优化

随着社会经济的高速发展，人们对信息和内容获取方式的要求也越来越高，移动通信的业务需求不断改变和提升，从最初的中低速率语音业务需求到宽带高速率数据业务的需求，进一步转化为满足海量差异化业务需求。网络优化指通过一系列的针对移动通信系统的专业测试、专业分析，发现问题并解决问题，同时深度开发系统潜能和提高系统运行性能的过程，其对象主要包括数据业务核心网、电路交换核心网、无线接入网，等等。随着通信网络流程和业务的多样化，网络运营的压力增大，人工智能的引入将成为解决网络优化问题的关键。当前，基于GPT 的网络优化可通过自主检测、分析和操作实现网络的自我校正和优化，主要包括网络流量优化、无线网络覆盖优化、网络信令追踪三个方面，如图2-8 所

示。

1）网络流量优化

随着用户业务需求的不断变化，流量也随之动态变化。GPT 可根据流量的变化提取特征，对其变化趋势进行预测，给出优化方案，从而平衡网络负荷，保障用户的网络体验。对流量较大区域、时段提前预测并进行配置调度，对流量较小区域、时段智能关断部分基站设施，从而达到节省成本的效果，保障通信网络处于最佳工作状态。

2）无线网络覆盖优化

无线网络的覆盖程度决定了通信网络的质量。统计显示，LTE 网络中各设备商的无线参数总和已经超过 8000 个，依靠人工经验很难进行精细化配置。有学者提出利用人工智能技术对通信网络进行系统分析，可实现精准网络参数配置[20]，在此基础上可以引入 GPT 相关技术。例如，在面对 TopN 小区覆盖问题时，利用 GPT 训练图神经网络（Graph Neural Network，GNN）模型，构建区域覆盖模型，输入影响覆盖的特征信息，如基站结构、参数配置等数据。然后，通过隐含层进行模型训练和特征学习，当算法迭代到一定程度时，可通过高层特征表述出覆盖预测模型、推荐参数取值及指导无线参数的调整与配置。另外，还可以帮助运营商将站点大规模多输入多输出（massive Multi-Input Multi-Output，mMIMO）天线的波束覆盖和传输形态纳入考虑范围，解决目标用户因在室内和地面上高度不同而引起的各种信号覆盖差异问题，充分利用mMIMO 天线的 3D 特性和周边环境特征，可以保证站点规划的准确性，从而更好地实现无线网络覆盖优化。

3）网络信令追踪

信令信息是指通信系统中的控制指令，又称“信令 ”，它主要用于处理和控制通信过程。通过对信令信息的处理，可以对通信过程进行监控管理和优化，从而提升通信的质量和效率。文献[21]中提出 AI 信令追踪措施，在此基础上引入 GPT 相关技术，能够监测和分析大量的信令信息，以便发现潜在的故障或异常 情况，从而掌握网络的实际运行状况。同时，GPT 可通过比对正常和异常信令 流程，快速定位故障所在，提供故障诊断报告，进而可以实现对故障的自动修复，

从而缩短网络中断时间，降低运维成本。除此之外，根据实时的信令信息和用户需求，基于 GPT 的信令追踪可以预测网络流量需求，动态管理、分配和调度资源，以适应时间、区域和用户类型的变化。同时，也可以分析和监测用户连接过程中的信令信息和延迟情况，及时发现并解决影响用户体验的问题，如延迟高、信号弱等，以提升用户满意度，保障网络环境稳定高效。

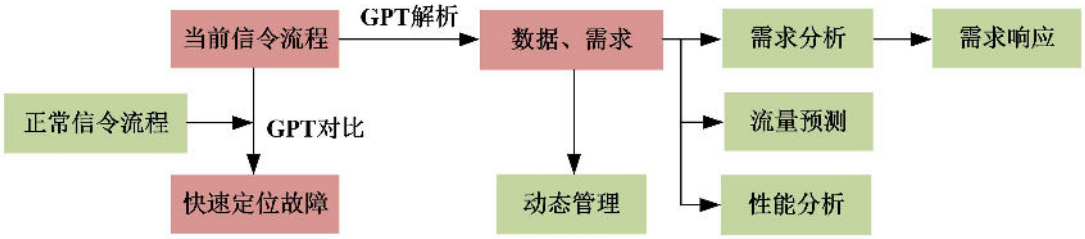


图2-8 GPT加速网络优化

3.通信网络使能 GPT 泛在应用

第2 章中我们介绍了GPT 在通信领域的崭新应用，以及GPT 促进通信网络智能自治的方式。GPT 将会被应用于越来越多场景，同时，通信领域如何实现对GPT 的支撑和优化是我们下一步需要考虑的。未来网络架构在设计时应将GPT 部署的潜在需求纳入规划范围，以使能GPT 的泛在应用，进一步满足用户多样化的需求。此外，边缘智能技术在近几年的快速发展为GPT 提供了更广阔的应用空间，兼具GPT 泛化能力的边缘智能具有巨大的潜力，如何对GPT 进行便捷的部署、如何将“云中心”的GPT“边缘”化需要在未来网络的设计上进行更多的研究。

本章我们将介绍现有通信网络如何保障GPT 应用落地，讨论未来网络设计的典型思路，以及如何实现对GPT 的原生支持。在此基础上研究了新型网络架构支撑海量数据训练和推理加速的关键技术，从而能够实现内生智能。

3.1. 通信网络保障 GPT 应用落地

近几年，随着云计算、大数据、AI 等技术的快速发展，以及各种应用场景的不断成熟，越来越多的数据需要上传到云端进行处理，这给云计算带来了更重的工作负担[22]。边缘智能的出现降低了对云端的依赖，为行业数字化提供了边缘智能服务，满足了其在敏捷连接、实时业务、数据优化、应用智能、安全与隐私保护等方面的关键需求[23]。此外，AIGC 充分解决了云网层高延迟、高风险等局限性，使得数据更安全保险、功能更稳定可靠、服务更高效智能。图3-1 展示了由移动AIGC 网络支撑的边缘智能。

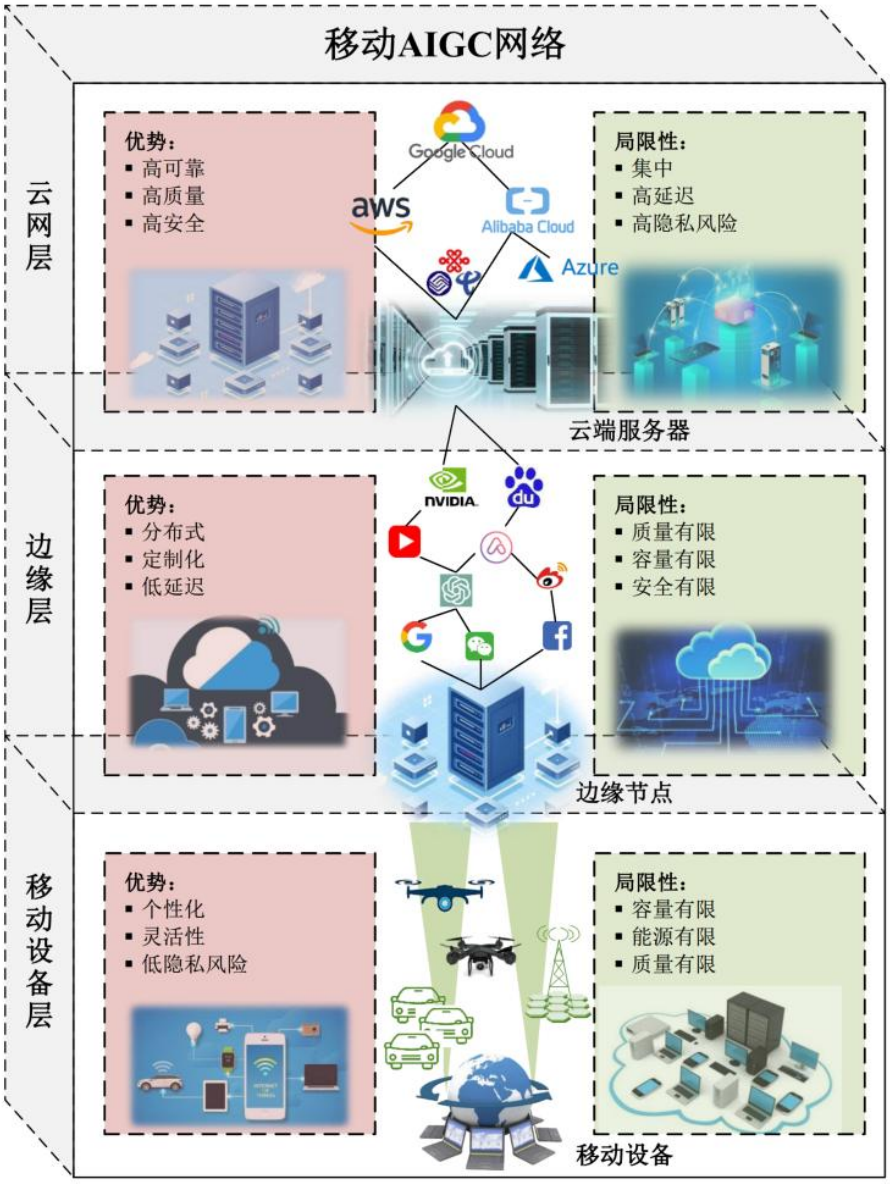


图3-1 移动AIGC网络支撑边缘智能

随着边缘智能的不断演进，将先进的自然语言处理技术如GPT 融入边缘智能系统中可以为其提供更为强大的语言理解和生成能力。边缘智能为GPT 的部署提供了融合网络、计算、存储、应用核心能力的开放平台，兼具GPT 泛化能力的边缘智能有着更广泛的应用场景。

在医疗领域，为了解决医疗纠纷问题，提高医疗机构的法务建设，同时为个人提供诊疗咨询服务，广州医科大学、三家运营商以及蜂动科技联合推出了诊疗合规法律知识GPT 垂直大模型5G 消息应用。该应用基于网络消息的流量入口，为GPT 提供了用户交互平台，支持GPT 技术提供智能化的问答服务和个性化推荐，提高服务质量。通过通信网络消息的商用模式验证，这种新型应用可以在业

界得到广泛应用。通过提供医疗领域专业知识，将GPT 大模型的研究领域缩小到诊疗纠纷细分领域，使服务更具有针对性，既可以回答医疗从业者关于诊疗医疗相关的法律法规，并提供相关的法规和标准，也可以为患者提供咨询、索赔等相关服务。据了解，项目上线以来有超10 万人次咨询，平均每次对话超过10.3 轮次。该应用为医生提供诊疗法律多场景入口，有效提高自身法律意识，以及快速获取解决方案，作为患者咨询入口，可以节约医疗纠纷层面的法务资源，此外，该应用还可以提供专业服务，有效避免医疗纠纷诉讼耗时长、投资大，减少经济资源浪费。

在智能办公方面，云知声携手深圳市龙华数据有限公司，以山海大模型为底座，推出面向行业垂直领域的“龙知政”政务GPT 大模型，成为聚焦特定领域 的GPT 落地实践项目。该模型具备语言生成、语言理解、知识问答、逻辑推理、 代码、数学、安全合规七项通用能力及插件扩展、领域增强、企业定制三项行业 落地能力，能够满足不同场景的不同需求。“龙知政”以GPT 大模型为底座， 应用于区政务服务数据管理局，通过高质量政务语料训练和参数微调，已具备在 政务领域中精准的知识服务能力。区别于传统的信息化系统，“龙知政”GPT 大模型为GPT 提供了专属数据的安全隔离、多轮对话以及信息溯源，并将专业 知识库的所有政策文件和专业术语融会贯通，对企业群众的办事诉求进行语义和 上下文的理解，颠覆传统线上机器问答模式，推动GPT 实现了信息服务由人工 查找向主动的、双向的、实时的智能全程引导转变。GPT 大模型可以基于现有 的政策性文件、规范性文件、政府公文等材料进行训练，学习各类写作风格、建 立写作模型，掌握业务要领和行文规范，帮助政府工作人员更准确、更高效地完成摘要总结、公文草拟、文本检索归纳等以往需要耗费大量精力的日常工作。

3.2.未来网络技术支撑 GPT 应用

过去 20 年，通信网络完成了“人联 ”“物联 ”的发展，面向 2030 年及未来，人类社会将步入智能化的新时代。在这个时代，社会服务将更加均衡和高效，社会治理将更加科学和精准，社会发展将更加绿色和节能。而未来网络技术将是实

现这些目标的重要支撑，它将从服务人和物，发展到服务智能体，并实现与 GPT 的高效的连接。未来网络技术将进一步支持 GPT 应用，通过人机物的智能互联，服务智慧化的生产和生活，满足经济社会的高质量发展，推动构建普惠智能的人类社会。

3.2.1.未来网络设计的典型思路

未来网络的设计涉及多个领域和技术，如物联网、云计算、人工智能、区块链、网络安全等，针对不同的应用场景和需求，需要提供不同的思路和方案。秉承的设计理念应该具有兼容性、跨域设计、分布式设计、至简性、安全性、内生设计等，实现继承式创新，并确保多种新增能力的一体化接入，使网络架构更灵活、更简洁、更安全，同时，需要更加主动地引入AI 技术[24]。

6G 网络将通过不断的自主学习和设备间协作，持续为整个社会赋能赋智，把AI 的服务和应用，例如GPT 推到每个终端用户，让实时、可靠的AI 智能成为每个人、每个家庭、每个行业的忠实伙伴，实现真正的普惠智能[25]。新的网络架构需要灵活适配协同感知、分布式学习等任务，以实现GPT 应用的大规模普及。在未来网络针对GPT 技术的架构设计中，需要实现GPT 在网络中的原生化，从网络设计之初就考虑对GPT 技术的支持，而不只是将GPT 作为优化工具。

除了原生支持GPT 应用，未来网络还需要包括新的特性，比如原生数据保护、原生可信、原生多元生态系统等，如图3-2 所示。此处的“可信”涵盖了网 络安全、隐私、韧性、功能安全、可靠性等多个方面[26]，要求未来网络设计必须注重网络安全和隐私保护[27]，采取多层次、全方位的安全防护措施。未来网络原生支持各种类型网络接入，构成实现普惠智能的多元生态系统。为普惠智能助力的是，人工智能技术将内生于未来移动通信系统并通过无线架构、无线数据、无线算法和无线应用等呈现出新的智能网络技术体系。

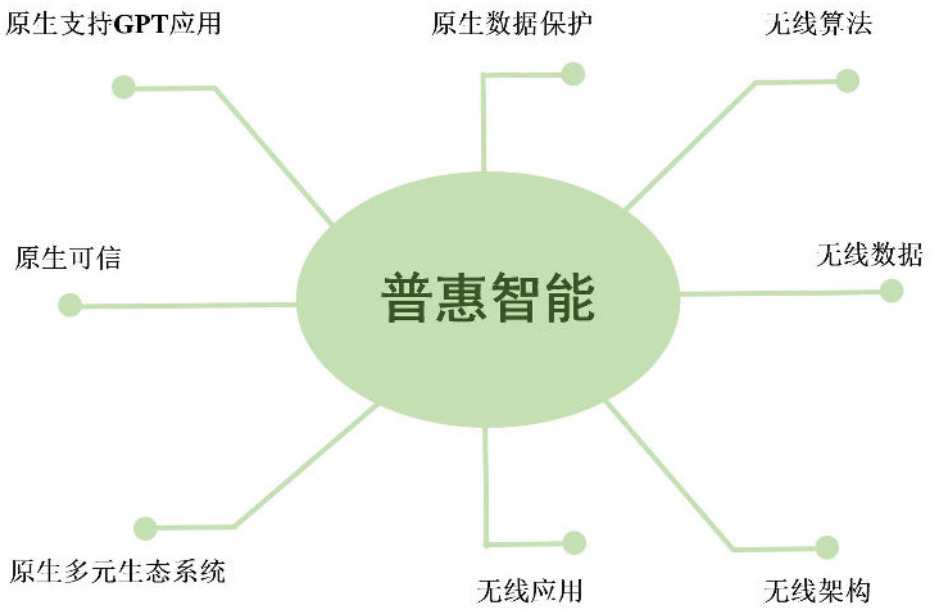


图3-2 普惠智能网络

3.2.2. 原生支撑 GPT 应用的 6G 网络

6G 的AI 能力不再是作为附加功能，而是作为一种内在特征。6G 的一个主要目标是实现泛在人工智能，并且实现无处不在的GPT 应用。在6G 时代，网络架构和GPT 将紧密结合， 6G 将为GPT 相关业务和应用提供端到端的支持。但是，原生支持GPT 的6G 网络是我们对未来网络的美好期望，仍需要满足以下诸多需求。

1）强大的边缘计算支持

未来网络需要提供强大的边缘计算支持，算力需要超过 100G FLOPS，使得GPT 模型能够在离用户更近的边缘设备上进行部署和运行，从而提高响应速度并降低对核心网络的负载。

2）实时的数据交换和处理[28]

GPT模型进行大量的数据交换和处理，需要网络提供低延迟和高带宽的支 。在传输过程中，需要保证数据快速、稳定地从源到目的地，速度要达到 l00Gbit/s 以上，延迟降低到秒级以下，才能满足GPT 模型对实时数据传输的需求。

3）严格的安全和隐私保护

GPT 模型所处理的数据有时会涉及公司的商业秘密和用户的隐私信息，因此网络需要提供强大的安全机制，包括端到端的加密传输、身份认证、访问控制等，以保护数据的安全和隐私。

4）智能的网络管理与优化

未来网络需要具备智能化的管理和优化能力，能够根据GPT 模型的实时需 求，动态调整网络资源分配、路径选择和优化网络拓扑结构，实现网络对GPT模型的最佳支持。

5）多样化的网络接入需求

考虑到GPT 模型可能通过不同类型的设备（如移动设备、IoT 设备等）进行 访问和使用，未来网络需要支持多样化的连接需求，包括无线接入、移动接入、低功耗连接等，以满足不同场景下GPT 模型的使用。

原生支持GPT 的6G 网络架构，充分利用网络节点的通信、计算和感知能力，通过分布式学习、群智协同以及云边端一体化部署，构建新的AI 应用生态 和以用户为中心的业务体验。同时，网络边缘运行的分布式人工智能可以逼近极 致性能，也能解决个人和企业都十分关心的数据所有权问题。普惠智能与ICT 系统深度融合，在网络边缘提供多样化的连接、计算和存储资源，将成为6G 的 固有特征。可以预见，提供原生GPT 支持的6G 网络架构将从现在的集中式“云AI”转变为分布式“网络AI”[29]。

3.3. 新型网络架构支持 GPT 能力下沉

随着6G 网络的快速发展，我们正步入一个以数据为核心的新纪元。传统的 以连接为中心的网络架构，正向更加智能、高效且以数据为中心的网络信息系统 演变。这不仅涉及对网络基础架构的改革，也包括对不断增长的数据进行有效管 理和利用，支持诸如GPT 等广泛的智能应用和服务。现有网络架构存

在数据处 理和智能决策支持缺失、延迟高带宽不足、计算能力集中、网络适应性低等诸多局限性。

为了支持GPT 能力下沉，需要网络具有高效且灵活的数据传输能力和强大 的分布式数据处理能力。从网络的部署设计出发，需要降低网络对中心计算的依赖性，在减少不必要的数据回传的同时保障数据的准确性。

3.3.1. 自适应切片

6G时代，移动网络服务的对象不再只是手机，而是各类型的设备，比如传感器、车辆等，业务类型也越发多样丰富。如果针对每种典型业务都专门建立特定的网络来满足其独特要求，那么网络成本之高将严重制约业务发展，同时， 若不同业务都承载在相同的基础设施和网元上，网络可能无法同时满足多种业务的不同QoS 保障需求。

自适应网络切片技术，允许创建多个虚拟网络切片，每个网络切片都是一组 网络功能及其资源的集合，由这些网络功能形成一个完整的逻辑网络，每一个逻 辑网络都能以特定的网络特征来满足对应的业务需求。通过网络功能和协议定制， 网络切片为不同业务场景提供所匹配的网络功能。其中，每个切片都可以独立按 照业务场景的需要和话务模型进行网络功能的定制裁剪和相应网络资源的编排 管理，是对6G 网络架构的实例化。这种灵活性使得6G 网络可以同时支持多种 有特定性能需求的服务和应用，如延迟敏感型应用和大带宽应用。核心在于，能 够根据实时的网络条件和服务需求，动态调整资源分配和配置。在面对像GPT 这样的高性能计算需求时，自适应网络切片显示出其独特的优势。如图3-3 所示， 网络切片作为提供服务的方式可以应用于多种垂直行业，根据应用场景、业务类型按需提供网络能力，切片间相互隔离、互不干扰。

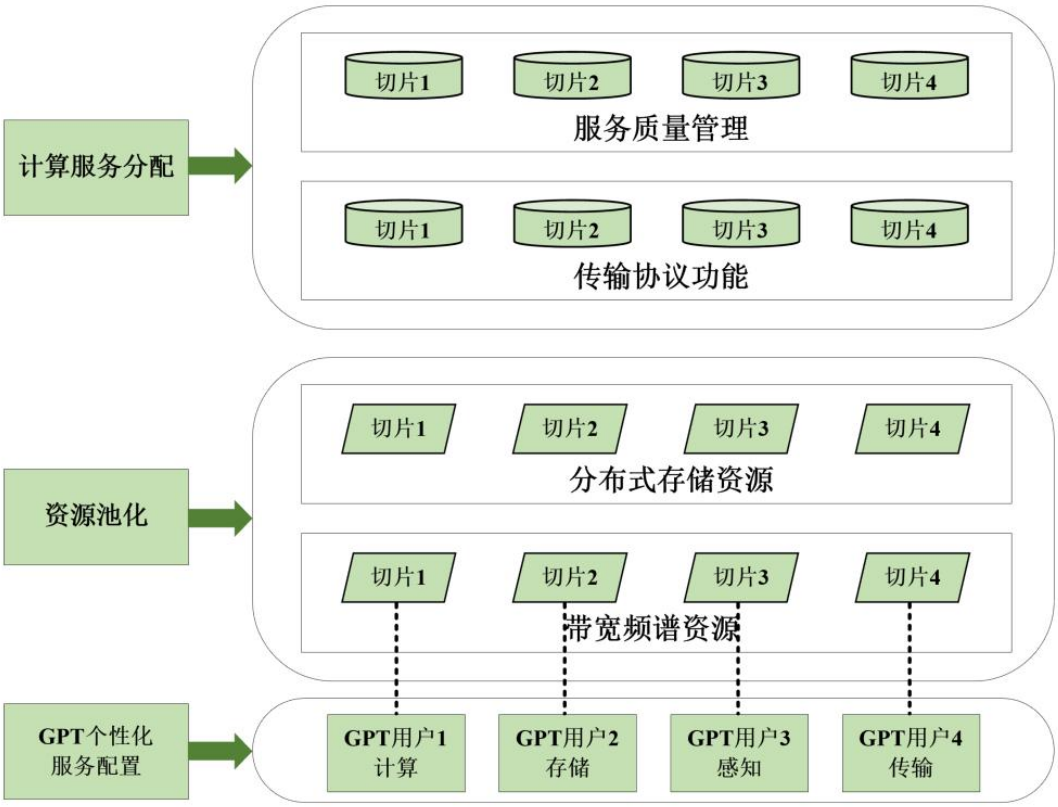


图3-3 分层网络切片管理

3.3.2. 分布式学习

分布式学习将成为6G 网络提升AI 性能和效率的重要途径。通过分布式学习可以扩充样本空间，部署更大模型，设计全局优化算法，并提升网络AI 开发和 训练效率。在智能服务方面，分布式学习能够使各智能节点在学习中交互信息， 共享应用学习经验，在网络智能最为常见的分布智能决策中具有重要应用价值， 具体包括：一般移动智能体的控制决策、各分布设施的智能推荐决策、环境监测 的事件判断决策以及网络自身的参数优化配置等。未来智能网络将形成传统集中训练支持系统全局智能应用，分布式学习支持各分布节点自主智能的新模式。

与传统的端到云的传输—计算相对隔离的运作模式相比，采用分布式学习的 最大特点是分布式的网络计算融合，通过网络中分布式可计算节点的参与，减少 冗余数据的传输，从而降低系统通信代价并提升计算效率。联邦学习就是一个典 型的通信—计算、网络—智能相互作用的系统。传统的联邦学习框架

下，多个分 布式节点利用本地的局部数据进行训练，并周期性地进行模型上传，模型参数在 中心进行整合更新后再分发至各节点。其中，通信为数据的传递和节点间信息的 交互提供支撑，而计算过程则影响系统调度和模型准确度，通信与计算相互耦合，共同决定了系统的可靠性和效率。

3.3.3. 边缘智能

如图3-4 所示，边缘智能最初可追溯到分布式计算的发展[30]。通过将计算任 务分解成多个子任务，更高效地处理大规模的语言模型训练，在泛化能力上取得 显著的进展。边缘智能的进一步演进在于将感知、计算和决策深度融合为一个整 体。感知阶段的关键目标是收集高质量的数据，为后续的计算和决策提供充足的 信息基础。随着深度学习和神经网络技术的飞速发展，计算在边缘智能中的角色 变得日益重要。边缘设备上的AI 模型能够解析和理解用户的语言，并做出相应 的决策，使得用户与设备之间的交互更为自然和智能。感知、计算和决策的融合 提高了系统的智能化水平[31]，通过在边缘设备上实现智能分析和决策，系统能够 更快速、更灵活地响应不同的任务和环境变化，为用户提供更为个性化和智能化的服务。

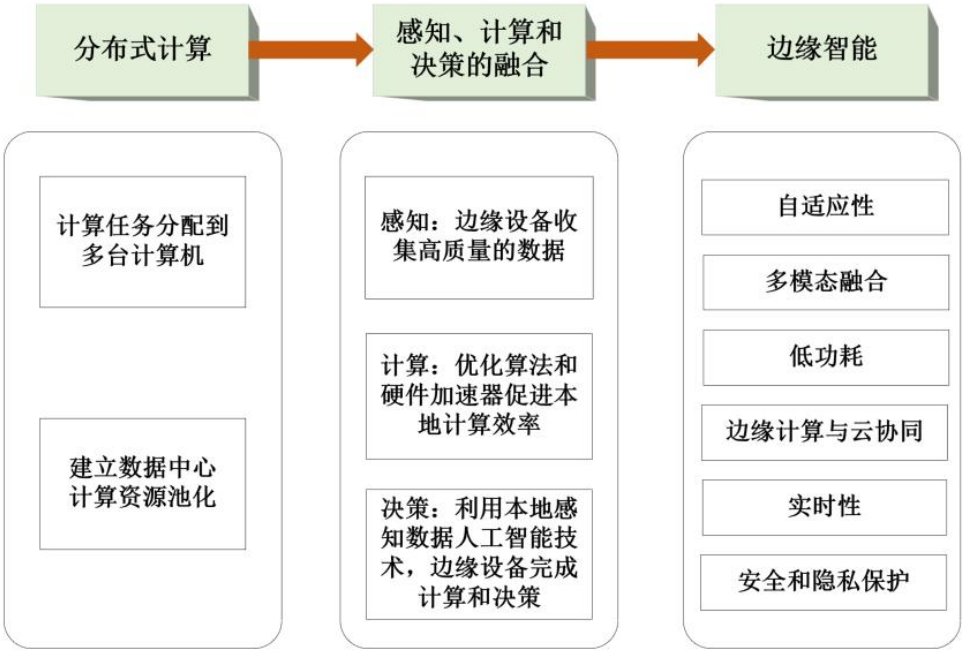


图3-4 边缘智能的概念演进

边缘智能的演进不仅在于其概念的拓展，也体现在其关键特征的不断丰富。 当其兼具GPT 泛化能力时，这些特征使得边缘智能在各个行业和应用场景中更具吸引力。

自适应性[32]：能够根据环境和任务的变化自动调整网络行为，并可以通过GPT 的学习能力，不断积累经验和数据，提高预测准确度和网络性能。

多模态融合[33]：需要同时处理来自多种传感器的信息，包括图像、音频、视频等，能够有效整合和处理不同类型的数据，提供更全面的智能分析和决策。

低能耗[34]：通过优化算法、硬件设计和电源管理策略，边缘智能系统能够在保持高性能的同时最小化能耗，延长设备使用寿命。

边缘计算与云协同[34]：边缘设备处理实时的、本地的任务，云服务器负责更复杂的、计算密集型的任务，两者协同工作以实现最佳性能和资源利用率。

实时性[34]：通过GPT 在边缘设备上执行智能分析和决策，系统更迅速地响应用户需求，降低通信延迟，提高用户体验。

安全和隐私保护[33]：由于数据在本地处理，边缘智能有助于降低因数据传输而引发的安全风险，并通过加密、身份验证等手段保护用户隐私。

4.GPT 与通信协同发展

移动通信技术以每十年一代的速度向前演进，丰富了人们沟通的方式，推动 了社会生活乃至生产方式的改变。随着大模型与生成式AI 的迅速崛起，AI 技术 的应用已经进入崭新阶段。GPT 作为新一代人工智能的典型代表，与通信的联 系越来越紧密，将从相互独立演进、融合发展，发展到协同演进赋能生产生活多 领域。随着社会经济的发展，人们的需求变得更加丰富，6G 网络有望引入新的 应用场景与关键性能指标，本章将讨论如何利用 GPT 相关技术，推动6G 新型 网络深化研究，同时将 6G 与GPT 结合，支撑更多行业数字化转型，取得更大的社会和经济价值。

4.1.GPT 与通信从独立演进到紧密结合

4.1.1. GPT 与通信结合趋势

早在2008 年，3GPP 就基于Rel8 定义了自组织网络（Self-Organizing Networks， SON），把机器学习和AI 的功能嵌入到了构思与规划、分析与设计、实施与构 建、运行与维护的网络生命周期里[33][34][35][36]，这成为推动通信AI 发展的一个里 程碑。然而，2G 与3G，最初都没有按照网络智能化理念来构建，旧网络时代生态体系与AI 难以适配。

2017 年9 月，3GPP 第一次定义了通信AI 的网元，即网络数据分析功能（Network Data Analytic Function，NWDAF），此外，O-RAN 也定义了通信AI 的网元，即无线接入网智能控制器（Radio Access Network Intelligent Controller，RIC），类似于网络里面的通信AI 大脑。

2018 年 6 月，3GPP 5G 新空口（New Radio，NR）标准独立组网（StandAlone， SA）方案在3GPP 第80 次TSG RAN 全会正式完成并发布。一方面，与4G 网络 相比，5G 网络在传输速率、传输时延、连接规模等关键性能指标上均有质的飞跃，进而支撑起更加丰富的业务场景和应用，为以GPT 为代

表的AI 工具落地创 造条件。另一方面，5G 网络在运营过程当中面临相关挑战，由于组网复杂、能耗高、控制灵活性差等问题带来诸多的不确定性[37][38][39]。

2023年3月，ETSI提出了有关AI透明度和可解释性的标准规范，旨在生成更多可解释的模型，同时保持高水平的模型性能。

2023 年 9 月，3GPPAI/ML 工作组将生成式人工智能引入了讨论范畴，并加 入NWDAF 模块，经过数个版本迭代演进，现阶段已形成数据采集、训练、推 理、闭环控制，以及支持多样化解决方案的分布式网络大数据分析架构。相关的 网络功能及接口规范已成熟，具备加速产业化能力。2023 年 12 月，3GPP Release 19 将围绕新场景、新技术，让AI 更懂网络，实现网络与AI 深度融合，包括AI 与5G 未来结合的应用方向。GPT 为代表的AI 有能力通过分析从网络中收集的 数据，来解决复杂和非结构化的网络问题。对一些特定的用例，3GPP 已经研究如何将AI 应用于5G RAN 和物理层。

2024 年2 月，ETSI 探讨了AI 在医疗保健、智能交通和工业自动化等不同领域的应用，以及未来移动网络中的AI 相关功能，这为GPT 在通信领域进一步应用提出了新思路，GPT 与通信行业正在从独立演进、前沿交叉到未来协同演 进、紧密结合，最终实现深度融合，这是相辅相成的结果，也是发展的必然趋势，将加速两者的共同进步。

4.1.2. GPT 与 5G 网络结合

GPT 与物理层的结合，有助于实现5G 的弹性设计。大规模天线技术是 5G NR 设计的基石。NR 需要支持高达 100GHz 的频谱范围，随着频率的升高，收 发系统使用的天线个数也相应增加。GPT 可以综合考虑天线的波束覆盖和传输 形态[40]，充分利用天线周边环境特征，对天线相关参数进行配置和优化，这是能够充分发挥GPT 强大推理能力的方向。

GPT 与网络层的结合，有助于实现5G 的柔性自治。通过5G 网络可获得海量数据，且结构性数据占比高，GPT 对大数据集进行统计和分析[41]，结果可以对整个网络进行更精细的调度和优化，提升网络资源利用率。

GPT 与业务层结合，可以提供端到端的确定性。通过引入GPT，将业务需 求转化为网络可感知的指标，对指标进行分析预测以判断当前网络资源是否满足 业务需求。进一步，结合分析预测得出的指标趋势，助力网络资源的动态规划、调度和优化，为最终提供高可靠业务保障打造坚实基础。

4.2. GPT 与 6G 通信网络融合发展

6G 网络与 GPT 融合是未来潜在发展方向，既包括为 6G 网络自身性能优化提供的智能能力，如利用端到端 AI 实现空口和网络的定制优化和自动化运维， 提供满足多样化需求的最佳解决方案；也包括向第三方业务提供的智能能力，如 通过 6G 网元原生集成通信、计算和感知能力，加速云上集中智能向边缘泛在智能演进，为服务第三方业务的 GPT 提供分布式学习的基础设施。

“6G+GPT ”服务主要面对高实时、高性能、强安全等需求，在网络内进行训练或推理，提供适应不同应用场景的 AI 能力。6G 网络作为原生智能架构，利 用网络内的通信、计算、数据集、基础模型等资源，结合 GPT 高效训练或推理 能力，能够实现海量数据处理、网络自服务、资源优化和内生安全等任务，为用 户提供无所不在的高性能 AI 服务。下面具体介绍 6G 与 GPT 融合发展的四个方向，如图 4- 1 所示。

图4-1 GPT与6G紧密耦合

4.2.1. GPT 支持海量数据处理

6G 网络需要服务海量数据采集、预处理、分布式存储和高速传输等基本数 据类业务。信息时代数据量爆炸式增长，海量数据资源蕴藏着巨大的价值。随着 6G 时代更先进的智能终端和无线边缘设备的增多，对边缘侧算力要求进一步提 高。目前英伟达已发布GPT 专用GPU，推理速度可以提升 10 倍，满足对6G 较高算力的需求。此外，基于GPT 的多模态模式可以同时处理文本、图像、音频等多种类型数据，强大的数据处理和分析能力帮助6G 网络进一步统一数据业务 和服务的标准（包括数据格式、参数定义、计算方式等），实现数据资源在6G网络内的快速流转和共享应用，实现以海量数据为中心的智能计算。

4.2.2. GPT 推动网络自服务

6G 内生智能网络的特性之一是自适应匹配用户的个性化需求，为用户提供 网络自服务能力。具体体现在，用户对带宽、时延、计算能力、存储能力等性能 指标的需求动态变化，网络接收到用户的个性化需求，基于GPT 对用户意图进 行分析并转译为对网络的QoS 需求，进一步根据对当前网络的状态感知，将QoS需求设置为网络的执行方案或执行策略，整个过程中不需要运维人员介入。

4.2.3. GPT 协助网络资源编排

6G 网络是集通信、感知、计算于为一体的信息系统，需要对通信资源、计 算资源进行编排，以满足用户服务级别协议（Service Level Agreement，SLA）需 求并实现网络运营效率最优。编排是对计算机系统、应用及服务的自动化配置、 管理和协调。GPT 对业务需求分析后进行资源消耗趋势预测，优化通信和计算 资源编排调度方案。具体来说，根据对业务需求的分析，网络生成通感算联合优 化需求，并下发至意图管控功能。基于GPT 辅助进行全面业务意图感知、网络 意图解析、网络能力转换，并输出和下发具体的业务感知 SLA 需求，实现将业 务SLA 需求细化为业务传输计算模型和基站资源消耗趋

势。随后，基于学习算 法优化影响通感算性能的网络配置参数，生成网络参数调优策略。指导资源编排，GPT 与用户交互获得反馈，进而对策略进行迭代优化。

4.2.4. GPT 构建网络内生安全

GPT 有望助力提供有效的安全技术来抵御和预判各类网络攻击，将在保护 6G 网络免受各种安全威胁方面发挥关键的作用。基于零信任的软件定义安全边 界，已应用于6G 来构建新安全边界防护体系[11]。通过始终持续性地验证用户、 设备和应用的合法性及权限，来实时界定每次资源访问是否被合理授权，完全通 过软件来定义对象的安全边界。在零信任的框架中，GPT 基于用户、设备和应 用进行动态风险评估，及时对权限进行调整；根据用户身份、行为、时间、地址、应用上下文等信息来构建指定风险管控方案，更高效地综合分析、应对每次访问风险。综合访问风险记录，GPT 总结历史经验，将安全能力与网元、网络特性深度融合，并进一步构筑 6G 内生安全体系。

4.3.“6G+GPT ”赋能行业数字化转型

6G 将具备原生AI 能力，不仅空口和网络设计将借助端到端AI 和机器学习实现高度定制的优化，同时各个网元也将原生融合通信、计算与感知能力，从集 中智能向分布式的网络泛在智能转变，通过边缘智能的分布式机器学习架构，满 足社会生产的大规模智能需求[42]。6G 作为未来数字世界的“超级基础设施”， 将以大连接、高算力和强安全的极致性能，支撑人、机、物的泛在智联，赋能全 社会数字化转型，实现“万物智联，数字孪生”的美好愿景。GPT 与6G 深度融 合后，可应用的场景非常丰富，支撑和提供诸多新业务和应用，最典型的应用领 域包括：智能家居、智慧医疗、智能工业、智能交通、智慧农业和数字娱乐等，如图4-2 所示。

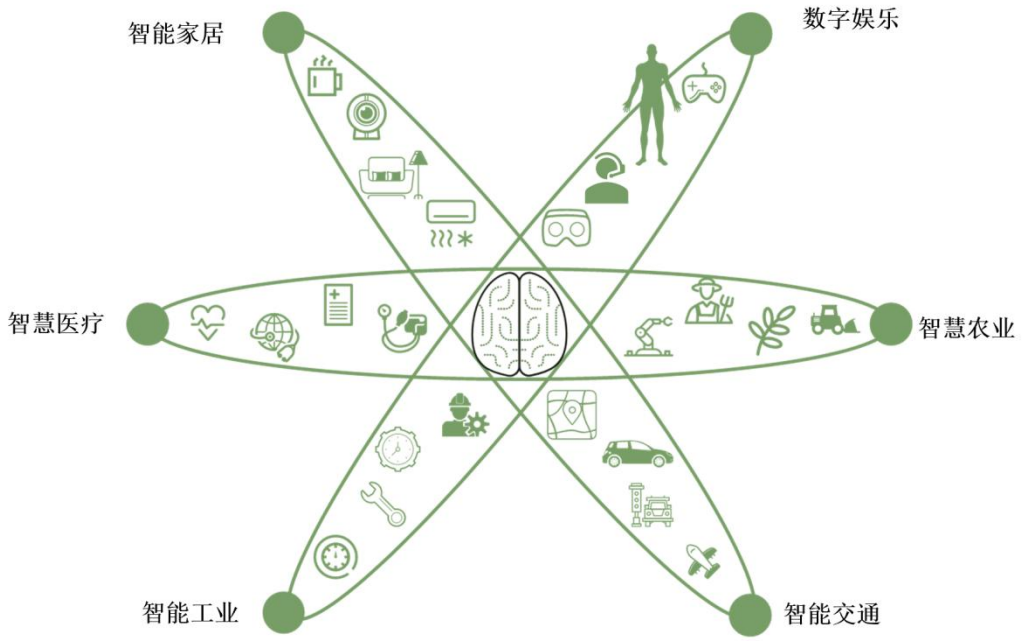


图4-2 “6G+GPT”赋能行业数字化转型

4.3.1.“6G+GPT ”赋能智能工业

近年来，以人工智能、大数据、云计算等为代表的新技术正在迅速与传统制 造业融合，以“绿色”和“智能”为核心的制造模式，成为制造业的重点发展方向。现代工业智能化生产模式，建立在AI 应用的基础上[43][44]，6G 通信技术与GPT 的协同，可充分发挥两者优势，提升工业系统性能，做到无线覆盖广、感知能力强、服务响应快等，进一步提高数据采集、处理和分析的效率，挖掘行业数据潜在价值。

工业智能化生产通常有着较高的传输和处理时延、鲁棒性、可靠性要求，由 于工业智能制造主要采用本地部署执行的特点，在6G 与GPT 的协同下，工作 中的基站侧传输、算力、算法资源和能力的拓展非常重要，它能比传统方案提供 更低的传输处理时延和抖动，从而保障更高的工业级信号处理的确定性。同时， 工业智能生产线上部署的智能终端，都可能具备本地强大的无线感知和数据分析 推理决策能力。此外，还可以结合神经网络、模糊控制技术等先进算法应用于产 品配方、业务编排等，实现智能制造过程，这有助于进一步提高生产效率和减少人工参与度，更好地满足客户的大规模个性化需求，实现工业生产技术的进步。

4.3.2.“6G+GPT ”赋能智慧医疗

6G 不仅能够更好地支撑智慧医疗相关的海量信息传输和同步，其内生智能 还可以直接赋能医疗信息的处理和决策。而 GPT 的出现，突破了传统AI 模型受 算法成熟度和病例样本数的限制，减少了人为参与和监控，简化了诊断方式和流程。

医疗传感器和智能可穿戴设备的发展推动了智慧医疗的改革，6G 和GPT 协同可直接应用于医疗传感器和智能可穿戴设备[45]，辅助收集个人身体和情感信息， 提供实时、便捷的健康监测，提高医疗质量，并使用户能够掌握自己的健康情况。 电子健康记录实现了患者完整病史的存储和显示，包括医疗状况、治疗计划、处方、过敏和其他详细信息。6G 和GPT 技术的结合，实现不同的物理设备和对象 与互联网进行连接和通信，优化了数据收集方式，高速传输同步医生和患者的相关信息，从而不断迭代提升预诊疗结果的准确性、可靠性和实时性。在GPT的辅助下实现数字化和集中化患者信息分析管理[46]，帮助快速建立集中的患者信息存储库，实现了数据驱动的决策，有助于加速推动智慧医疗改革。

4.3.3.“6G+GPT ”赋能智能交通

智能交通是AI 和通信技术与现代交通系统融合的产物。“6G+GPT ”可以给城市交通带来全新升级，例如在自动驾驶、无人机快递、无人出租车、车路协同等方面推动城市交通体系的持续变革。

在进行路网级的交通信号控制协同时，针对各种公路网的感知是极其关键的。目前对于公路网的感知主要依靠城市卡口、微波雷达、GPS 定位等数据源，交通 数据采集设备的部署总体来说还比较稀疏，在构建时空模型时有价值的信息依然 有限。6G 与GPT 协同将充分发挥内生感知和数据处理能力，有望通过广域覆盖提供全方位多维度的路网感知数据[47]。

交通流的预测精度和时效性对于交通的主动管控非常关键，面向超大规模交 通网络状态的估计和预测，6G 与GPT 协同相比于传统云端 AI 更贴近交通现场，从而能够提供更精准更实时的预测结果。

4.3.4.“6G+GPT ”赋能智慧农业

智慧农业是物联网技术在现代农业领域的应用，利用实时图像和视频对农业 生产系统进行监控和检测。与传统手工或机械化农场相比，智慧农场可以采用基 于“6G+GPT ”的新型生产作业模式，如通过传感器采集农场片区的各类数据， 智能调控农作物的生长环境，使其更好地满足作物生长需要，并将各种类型的农 业机器人应用到耕地、播种、喷药、收割、采摘、包装等农业作业环节中，进一步提高农场作业质量及效率，减少人工投入。

“6G+GPT ”能为智慧农场提供各种AI 业务支持，包括基于多类传感器的感知数据精准获取与传输、基于海量数据的分布式AI 模型训练、模型参数的高 效传输与聚合、无人机喷洒作业路线的精准规划和飞行控制、农机自动驾驶路线 规划等。智能传感器设备可以实时获取精准数据，比如时刻监测大棚内的生长环 境数据，GPT 据此输出温控、水肥等解决方案，帮助农民实现标准化种植，真 正达到降本增效的目的。在采摘环节，基于“6G+GPT”控制采摘机器人[48]，通 过高精度定位和动作控制，实现智能采摘。此外，基于对农作物生长数据的分析，对作物的全生命周期进行管控，能够及时发现种植问题并发出预警，减少损失。

4.3.5.“6G+GPT ”赋能智能家居

智能家居是人工智能技术和物联网技术在居家生活场景中相互融合的产物。 智能家居系统能够像人类一样思考、决策和调度用户习惯和家庭环境，从而提供 便捷、舒适、安全的智慧生活。据统计，在智能家居市场中，2022 年AI 技术的 行业整体渗透率约为25%，预计 2025 年AI 技术的行业整体渗透率将接近50%， 而在拥有计算机视觉、语音交互功能的智能扫地机器人、智能摄像头、智能门锁、智能音箱等品类中则有望突破60%。

“6G+GPT”可用于家居控制、安防监控、健康监测等，利用6G 网络设施，可以感知和分析人们的行为、手势和位置等信息，结合历史数据刻画住户习惯， 通过GPT 了解人们意图，实现对各类家居设备的最优控制。在维护家庭安全方 面，“6G+GPT ”可检测非法入侵，基于家庭成员画像，分析并评估入侵

动作危 险等级，同时安防系统自动触发报警等动作，避免家庭财产受到损失。在生活家 居健康监测方面，“6G+GPT”基于传感数据进行识别分析，可实现对住户和宠物的健康监测管理，当健康指标与历史信息相比出现异常时，能及时发现和预警。

4.3.6.“6G+GPT ”赋能数字娱乐

数字娱乐行业存在“成本、效率、质量”的不可能三角，即难以同时兼顾研发成本、研发效率与产品质量。而AIGC 的广泛应用，能够极大地提升策划、音频、美术、程序等环节的生产力，压缩整体项目的研发周期与人员规模，大幅降低制作成本。虚拟现实（Virtual Reality，VR）、增强现实（Augmented Reality， AR）、扩展现实（Extended Reality，XR）等技术的使用是新一代数字娱乐的趋 势。“6G+GPT”能够提供完全沉浸式的交互场景，支持精确的空间互动，满足人类在多重感官甚至情感和意识层面的联通，将助力实现现实环境中物理实体的数字化和智能化，构建虚实融合的数字娱乐新模式。

XR 需要进行对象定位和运动追踪，处理和反应依赖于AI 的能力，GPT 技 术可用于6G 网络未来XR 业务，提供更为丰富的算力和算法资源，保证各种XR 应用的执行和卓越用户体验[49]。除了AR、VR 业务以外，6G 网络需要更强大的 网络图形图像业务，未来，不仅包括 cloud XR，还有云游戏、智慧城市、数字 孪生城市、数字可视化等。网络图形图像处理过程中存在大量的数据传输，这需 要GPT 对网络传输进行调优，保证网络的通畅和业务的时效性。此外，现实感 知技术需要基于便携式终端感知周围环境，使信息收集更方便，GPT 能够对信 息进行识别、分析和处理，提供超越人类自身的强安全、高精度、低功耗的感知 与成像能力。其中，超高分辨率场景需要更高的带宽和更大的天线孔径[50]，通过 “超越人眼”的GPT 应用，可以获得皮肤下、遮挡物后或黑暗中隐藏的重要信息。例如，在6G 系统中，基于太赫兹频段的无线电波感知[51]，可以实现非视距成像。

5.“GPT+通信 ”融合发展面临的问题

在第4 章中，我们研究了GPT 大模型与通信如何从独立演进到融合发展的过程，并讨论了未来6G 网络与GPT 结合后如何支撑更多行业的数字化转型。然而，在二者协同发展和实际应用的过程中，仍然存在一些难点和挑战。这是因 为通用大模型虽然能够掌握广阔的通识知识，但缺乏对特定任务和行业的深入理 解，导致在满足严格的专业化需求时可能难以达到最佳性能。此外，大模型本身 也存在一些局限性，同时通信行业的数据复杂性以及网络环境的动态变化等因素，都可能导致模型输出效果不佳。

因此，在将GPT 大模型应用到通信领域时，往往会与实际场景的应用需求存在一定差距。本章将分别从高质量数据稀缺、硬件资源不足、云边端网络协同 难、带宽存在瓶颈以及相关法律滞后这五个不同的角度进行分析，讨论“GPT+通信”融合发展过程中需要解决的痛点问题和可能的研究思路。

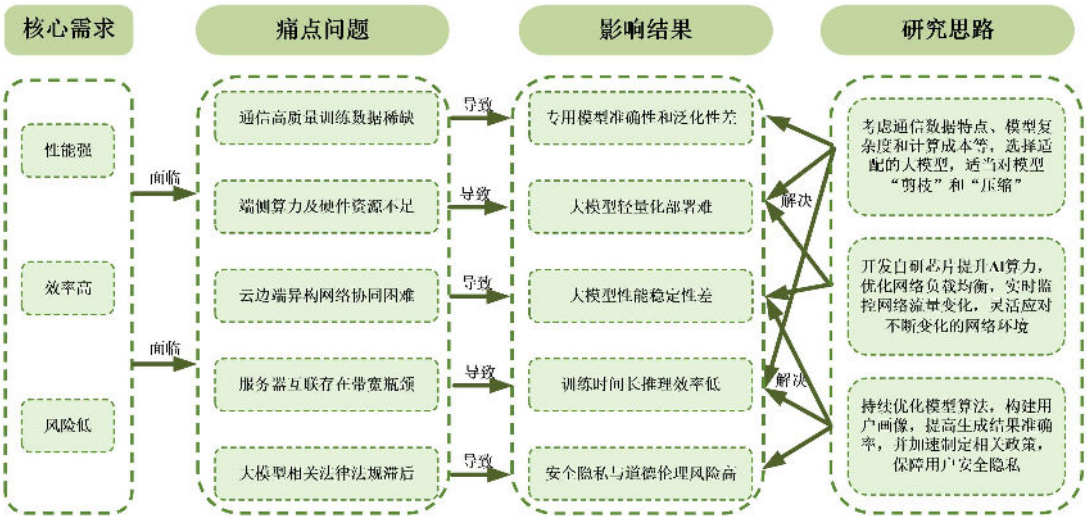


图5-1 “GPT+通信”融合发展面临的问题和可能的研究思路

5.1. 通信高质量训练数据稀缺，专用模型准确性和泛化性差

各个行业都有长期积累且涉及多个维度的专业知识，为了训练出满足产业需 求、精度极高的垂直行业模型，大模型必须结合行业知识和专有数据，完成从通 用到专用的转变。在设计适配通信领域的GPT 大模型时，训练数据集会直接影 响生成内容的质量，而通信中的高质量训练数据仍然不足，模型在理解复杂或非标准指令时的准确性较差。且模型对单一性数据集的重复训练可能会出现过拟合，导致泛化性大幅度降低，这对通信专用模型的性能进一步提升提出了挑战，如图5-2 所示

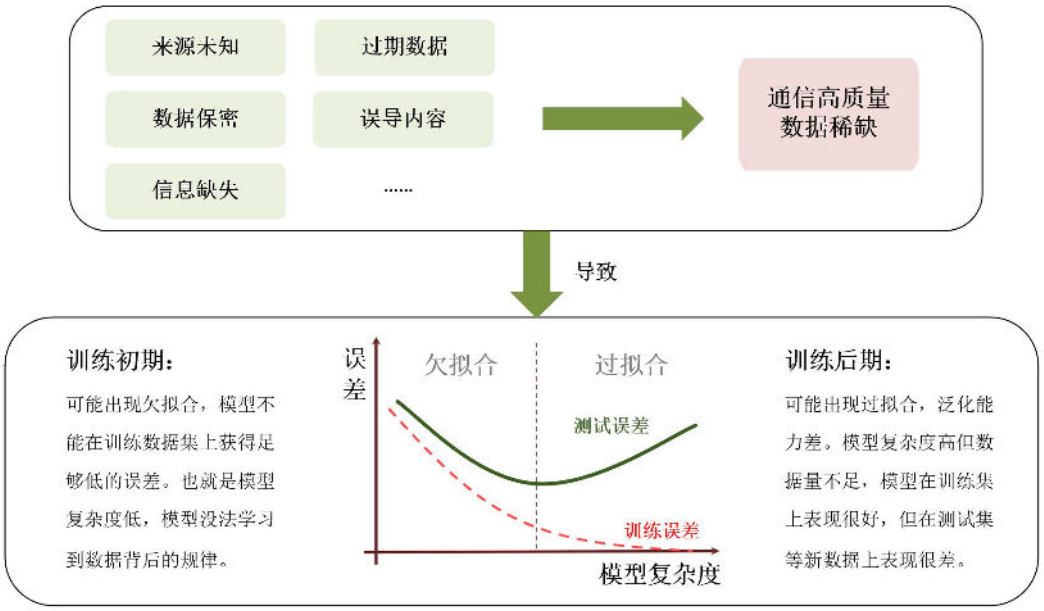


图5-2 通信高质量数据稀缺导致模型性能降低

准确性是指大模型生成的回答是否正确，是否符合逻辑和人们的认知常理， 是否能够被人们理解、解释和信任。在通信领域，准确性可能涉及正确识别或预 测一系列通信任务，如信号识别、调制解调、频谱感知等，对数据质量要求较高。 而泛化性是指机器学习模型在未见过的数据上仍然表现良好的能力[52]，换句话说， 一个好的机器学习模型不仅能够在训练数据上表现出色，还能够在新数据上进行 准确预测和决策。然而，泛化性并非一蹴而就，而是需要精心设计和调整模型来 实现的。训练数据集样本过于单一可能会导致模型训练欠拟合或者过拟合问题。 这是机器学习中的常见问题，指的是模型在训练初期

对数据集学习不足，但后期 又过度适应训练数据，或过于强调训练数据中的每个细节，而不是学习普遍规律， 导致模型在新数据上表现不佳，泛化性大幅度降低[53]。过拟合可能是因为训练数 据集样本单一、样本不足，也可能是因为模型过于复杂，过于贴合训练数据，而忽略了数据之间的一般关系，以至于学习到了训练数据的噪声和细微差异。

在深度学习领域，许多研究在评估深度学习模型时只关注准确性，而忽视了 泛化性。这意味着这些模型可能在特定数据集上表现良好，但在新的环境下可能 无法泛化。这是一个与通信融合的问题，因为在实际的通信系统中，模型需要能够适应不同的环境和数据[54]。因此在设计通信领域的深度学习模型时，需要在准确性和泛化性之间进行平衡。但目前通信领域训练所用数据量远远低于需求，通 常只有数十GB，与所需的数百GB 到数TB 相比存在巨大差距。而且现代预训 练数据集的大小使得任何个人都无法彻底阅读所包含的文档，或对其进行质量评估[55]，因此如何挑选合适的高质量数据集对GPT 进行预训练仍有待研究。

但是，通信行业内的许多数据未公开或受到商业保密协议的保护，导致公开 可用的数据来源有限，或数据缺失无法充分利用。由于行业专业性强，非可靠来 源的数据更可能包含技术上的错误、过时的信息或误导性内容。在通信行业中， 使用不准确的数据训练的大模型可能会生成包含事实错误、逻辑错误或偏见性观 点的内容，导致技术误解和错误的决策，这在设计网络、维护系统或应对紧急情况时尤为严重，不仅会降低用户对模型的信任度，甚至可能造成严重的事故。

此外，在学习和表示多模态数据时，不同无线应用的数据结构和特征也各不 相同，无线数据类型多样性包括信道频率响应、位置坐标、波束矢量和接收信号 等，这些不同模态具有独特的数据结构和特征[56]，对大模型的准确性和泛化能力提出了挑战。

因此，针对通信领域不同任务的特殊需求，为了在有限数据下训练专用大模 型，需要综合考虑通信数据特点、大模型本身复杂度和计算成本等因素，选择适配任务的大模型，可能还需要适当对模型进行“剪枝”和“压缩”。

5.2. 端侧算力及硬件资源不足，大模型轻量化部署难

在AI 技术快速发展的当下，智能手机等移动设备在人机交互、语音交流等 功能方面的需求不断提升，将大模型轻量化部署到终端设备也正成为一个重要的 研究方向和发展趋势。利用端侧AI 可以更好地为用户提供个性化的服务和支持， 帮助用户进行自我管理。然而，由于算力及硬件设备等限制，这一目标的实现仍面临诸多挑战，在实际应用上还远远满足不了用户的需求。

早期的智能手机语音助手，虽然具备基本的人机交互能力，但在复杂问题的处理上表现并不理想，功能也较为单一。随着ChatGPT 等大模型的发展，AI 能力得到显著提升，原本功能有限的语音助手有望处理更复杂的问题，这无疑是手机制造商们迫切希望落地的技术。例如，苹果最早搭建了Ajax 大模型，并推出了内部测试聊天机器人“Apple GPT”；vivo 发布了自研的覆盖多个参数量级的“蓝心”大模型，包括端云两用模型和端侧专业文本大模型等；小米宣布其自研MiLM 轻量级大模型已经接入了新发布的澎湃操作系统（Operating System，OS）；华为也宣布HarmonyOS 4 系统将全面接入“盘古”大模型；荣耀、OPPO、三星等其他终端厂商也都在纷纷布局，将大模型装进手机，如图5-3 所示。

时间	公司	相关布局
2023年7月	苹果	搭建了大语言模型Ajax，并推出了一个名为“Apple GPT”的内部聊天机器人来测试其功能。预计每年将在人工智能的研究上投入10亿美元，更智能的新版本Siri有望于2024年问世
2023年8月	小米	推出MiLM轻量级大模型，技术主力突破方向是轻量化和本地部署，目前已在手机端跑通13亿参数的大模型
2023年8月	华为	宣布HarmonyOS 4系统将全面接入“盘古”大模型，利用AI 大模型、HarmonyOS API 等技术，把语音助手小艺变成更懂用户的个性化助手
2023年11月	OPPO	在OPPO开发者大会上正式发布了自主训练的个性专属大模型与智能体——安第斯大模型（AndesGPT），以“端云协同”为基础架构设计思路，推出多种不同参数规模的模型规格
2023年11月	vivo	在2023vivo开发者大会上，vivo正式发布自研“蓝心”大模型，含覆盖十亿、百亿、千亿三个参数量级的五款大模型，全面覆盖用户核心场景
2024年1月	荣耀	在MagicOS 8.0发布会上，正式揭晓了自研的端侧70亿参数平台级AI大模型——“魔法”大模型，且与百度智能云达成了战略合作，通过其千帆大模型平台的端云协同，生成专业内容
2024年1月	三星	在三星Galaxy S24系列上搭载了Galaxy AI，将从音频、文本、图像和视频处理四个方面，提升用户体验

图5-3 手机厂商在终端部署大模型的最新进展

然而，要将大模型部署到终端设备上，对算力和硬件层面就提出了更高的要 求。据统计，vivo、荣耀、小米的大模型基本上都从十亿级的参数量开始做起， 逐渐往更大的参数量拓展。目前，手机厂商在大模型计算上，基本采用两种

路径， 荣耀、小米等公司是采用端侧计算的模式，vivo 等公司则是端侧和云端两条路径 并行。云端计算的能力更强，但每次计算的成本过高，可能会对手机厂商造成负 担。相比之下，端侧计算成本更可控，并且由于数据不用上传云端，安全隐私性更强。

在具体实现过程中，大模型需要大量内存来存储参数，并且在执行复杂的计 算任务时需要巨大的算力资源和电量消耗。而智能手机等移动设备硬件资源和电 池容量有限，轻量化部署仍然存在许多挑战，主要表现在三个方面：内存约束、算力不足和功耗较大。

1）内存约束

大模型由于其庞大的参数量和复杂的网络结构，需要极大的内存资源来存储 模型参数和进行计算。在苹果公司发表的最新论文中提到，一个 70 亿参数的模 型就需要超过 14GB 的内存来加载半精度浮点格式的参数[57]，这超过了大多数终 端设备的承受能力。即使通过量化压缩模型，这对于终端设备的内存要求依然过 大。且在实际应用场景中，端侧设备需要快速、准确地处理输入并给出响应。但 由于内存限制，这些设备可能无法承载大型模型，或者因为内存不足导致性能下降，运行模型时频繁卡顿，影响其他应用程序等问题。

2）算力不足

大模型运行需要消耗大量的计算资源，因此如果要在端侧部署大模型，除了 内存外，对芯片计算能力也提出了更高的要求。然而，现代智能手机的芯片处理 器性能与专业服务器相比仍有较大差距，使得模型运行面临处理速度慢和响应时 间长等瓶颈，从而影响用户体验。目前行业内可供采用的芯片不多，暂时只有极 少量芯片能支持大模型的端侧落地。由于大模型对于手机内存和芯片的限定要求， 即使实现了在手机上部署 GPT 大模型这一目标，在短期内可能只会是高端手机的专属体验。

3）功耗较大

大模型对算力的需求变得更大了，意味着功耗也会变大。但手机等端侧设备 不像数据中心能通过空调或液冷系统降温，功耗太高会直接影响大模型运行

的效果。当在这些设备上运行GPT 大模型时，不仅会迅速消耗电池电量，破坏手机 充满电后至少能待机一天的体验，还可能导致设备过热，缩短电池的使用寿命。 这种过高的电量消耗不仅限制了大模型在移动设备和其他端侧设备上的应用，而且还可能导致运行成本和设备维护难度增加，甚至还可能直接损坏硬件，影响其他功能的正常使用。

5.3. 云边端异构网络协同困难，大模型性能稳定性差

在当前大模型体系架构下，在终端设备上部署GPT 应用并形成实际业务服务需求，需要云边端共同参与完成。GPT 大模型在边缘节点部署，而用于预训 练过程的大规模数据库通常在云端存储，这涉及终端—边缘节点和边缘节点—云端两段链路的数据传输。然而，随着移动用户的个性化需求剧增，为了满足更多用户，需要实现云边端网络的高效协同，实现计算资源合理化分配，否则可能会导致大模型稳定性下降，影响用户体验。

云端服务器通常具有强大的处理能力和存储容量，适合处理大规模、复杂的 任务。边缘节点位于用户和云资源之间，具有一定的计算能力和存储空间，可以 降低数据传输延迟，加快响应时间。终端设备如智能手机、传感器等通常资源非 常有限，但最接近数据源。在这种网络中，数据和任务根据其属性和需求在云、 边缘和终端之间流动。一些需要快速响应的任务可能在边缘节点处理，而那些需要深度分析和复杂计算的任务则可能上传到云端处理，如图5-4所示。

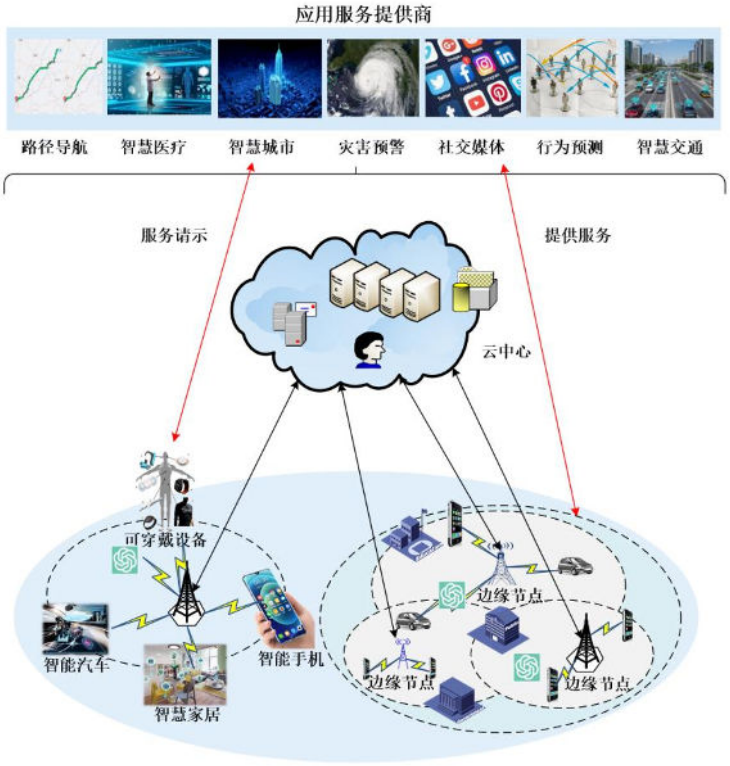


图5-4 “GPT+云边端异构网络”架构

然而，在边缘计算方面，由于物联网带宽资源等限制，其发展仍处于起步阶 段[58] 。在保持模型准确性的同时，如何设计在物联网设备上运行的GPT 大模型 存在挑战，其中最关键的因素是边缘应用的处理速度[59]。同时，网络中不同节点 的性能、存储和网络连接能力存在显著差异：云服务器可能拥有高性能处理器和 大量存储，而边缘节点和终端设备则在计算能力和存储容量上有限。这些差异对大模型的效率和准确性产生了直接影响。

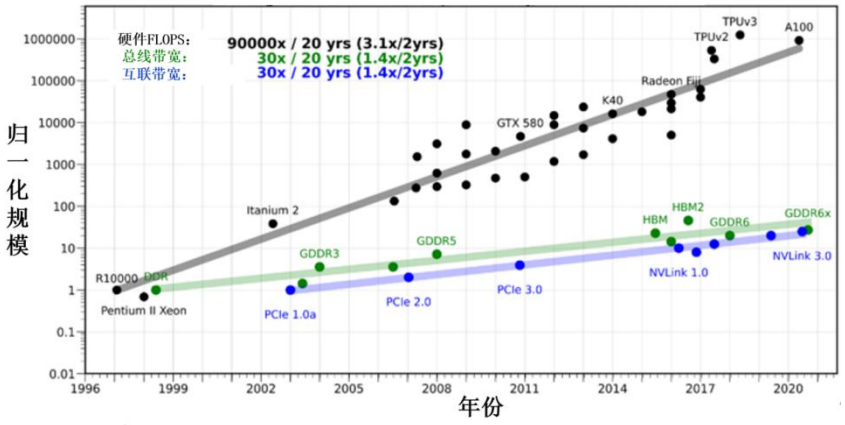
首先，响应时间在云端、边缘节点和终端设备之间存在显著差异。云服务器 可以在几十毫秒内处理请求，而边缘设备则可能需要几百毫秒到几秒，而在终端 设备上，处理延时可能更高。其次，大模型的有效运行常常需要在多个计算节点 间同步大量数据，网络延迟和数据传输速度在这里成为关键因素，影响模型的实 时响应和决策能力。且边缘节点的计算限制可能要求对模型进行简化或只运行模 型的一部分，这可能导致模型准确率下降，特别是在数据密集和计算密集的应用 场景中，如智能交通系统，高峰期的数据处理需求可能超出边缘节点的处理能力， 导致模型性能波动。最后，大模型的稳定性还受到网络节点可靠性和故障恢复能 力的影响，边缘节点的故障或中断可能导致服务的中断。

因此，需要综合考虑设备的性能和资源约束，以保证网络中所有节点能高效协同工作，同时保持系统的准确性和可靠性。

例如，在城市监控系统中，大量数据需要监控摄像头传输至云端服务器以进 行有效处理，传输过程中的延迟对实时监控的效率和应急快速响应能力有显著影 响。然而在处理计算密集型模型时，边缘节点可能受到其硬件性能的限制，难以 达到理想的处理速度和精度。这一点在自动驾驶中尤为明显，该应用要求车辆快 速处理来自众多传感器的大量数据[60]。在这种时延敏感型应用中，高带宽低延迟 的网络连接是保障实时决策乃至驾驶安全的重要因素。因此，需要优化网络负载 均衡，分析用户需求并合理分配有限资源，才能灵活应对不断变化的网络环境，保持大模型的稳定运行。

5.4. 服务器互联存在带宽瓶颈，训练时间长推理效率低

大模型的训练和推理过程需要大量的计算资源和数据，仅大模型训练就需要 由数千片甚至上万片GPU 组成的集群连续训练数月时间，海量计算的同时还有 海量数据交换的需求，与传统CPU 和存储集群比较，内部通信互联要求的提高 十分明显。且随着模型参数量以及GPU 算力的增加，要在动态无线通信环境下同时满足生成内容高质量和低延迟，需要更高的互联带宽才能支持。然而，由于 目前计算服务器间的互联带宽不足，这可能会导致网络传输速度过慢甚至中断， 需要很长时间才能从云服务器上下载数据，从而影响资源的使用率，降低整个训 练和推理过程的效率和准确性。如图5-5 所示，当前通信带宽提升速度远低于计算提升速度。



数据来源：amirgholami@github

图5-5 通信带宽提升速度远低于计算提升速度

随着模型规模的不断扩大，单GPU、单服务器已经无法满足计算和存储需 求。目前，大模型的训练过程需要在多节点的计算集群中进行，这些集群通常由 若干台服务器组成，通过分布式训练框架实现跨节点的协作，共同完成训练任务 [61]。主要采用的是分布式并行计算方式，即将计算任务以数据并行、流水线并行 及张量并行等方式分配到多台服务器上，来加快模型训练速度。在数据并行模式下，数据被分割成多个部分，分配给不同的计算设备进行并行计算。

在这种复杂的分布式系统中，任何一个环节遇到瓶颈都可能对模型训练的效 率和可扩展性产生重大影响。当计算服务器数量增加时，各应用程序线程间的通 信成本会增高，进而导致整体训练性能下降。在传统服务器配置中，AI计算卡 之间的通信受限于PCIe 总线的带宽，使得数据在GPU 内存间的传输速度仅为理 论速率的约1%。此外，位于不同服务器的AI 计算卡之间的通信还受到数据中 心网络带宽的限制，如常见的 10 Gbit/s 以太网速率，进一步制约了训练效率。简而言之，随着集群规模扩大，通信成为影响AI 模型训练性能的关键因素。

其中，影响最大的是服务器间的高速互联。需要在系统之间提供100 Gbit/s甚至更高的带宽，改善GPT 类模型训练的通信带宽，进而提升算力的利用效率。 因此，需要解决计算服务器之间可能存在的互联带宽瓶颈问题，以确保数据在服 务器之间能够快速、高效地传输。还需要正确配置和优化计算服务器上的硬件， 考虑和设计合适的网络拓扑，以最大程度地提高互联带宽的利用效率。在通信领 域应用GPT 大模型时，对算力的需求和对数据中心网络的稳定性要求同样较高。 为了提升通信数据集的获取效率，往往需要在预训练过程中采用更大带宽传输海量数据，这提高了硬件设备的性能门槛。

分布式训练需要在多台主机之间同步大量参数、梯度和中间变量，对于大模 型来说，单次同步参数通常在十亿量级，因此对高带宽网络有很高的需求。在分 布式计算环境中，不同计算机之间需要频繁地进行数据交换和通信。因此，网络 性能的优劣会直接影响分布式训练的质量和速度。如果网络吞吐量不够大，数据传输就会成为瓶颈，从而限制分布式训练的效率。

因此，网络性能对分布式训练的质量和速度有着重要的影响。必须要采取相 应措施来提高服务器之间的互联带宽，同时优化网络的负载均衡，以保障整个计算集群的效率最大化。

5.5. 大模型相关法律法规滞后，安全隐私与道德伦理风险高

随着AI 技术的飞速进步和大模型的普及，信息化世界的各个方面都在迅速演变，但与此同时，网络安全和隐私泄露的风险也在不断上升。在这个数字化时 代，确保网络安全和保护个人隐私已经成为极其紧迫的任务，我们需要深入理解 风险，并采取适当的措施，以确保数据安全、内容安全、社会安全乃至国家安全。 此外，迈入现实的AI 技术也同时落入了纷繁复杂的人类社会，它不仅是技术工 具，也将作为一个社会对象影响着使用者，其训练数据中也不可避免地包含一些 人类社会偏见。如何正确合理地使用大模型，怎样科学地看待、解决大模型在社 会维度上的价值观与道德伦理问题，如何结合技术手段和治理体系，合理地对安全隐患和隐私泄露风险进行控制，是摆在全人类面前的重要课题，如图5-6 所示。

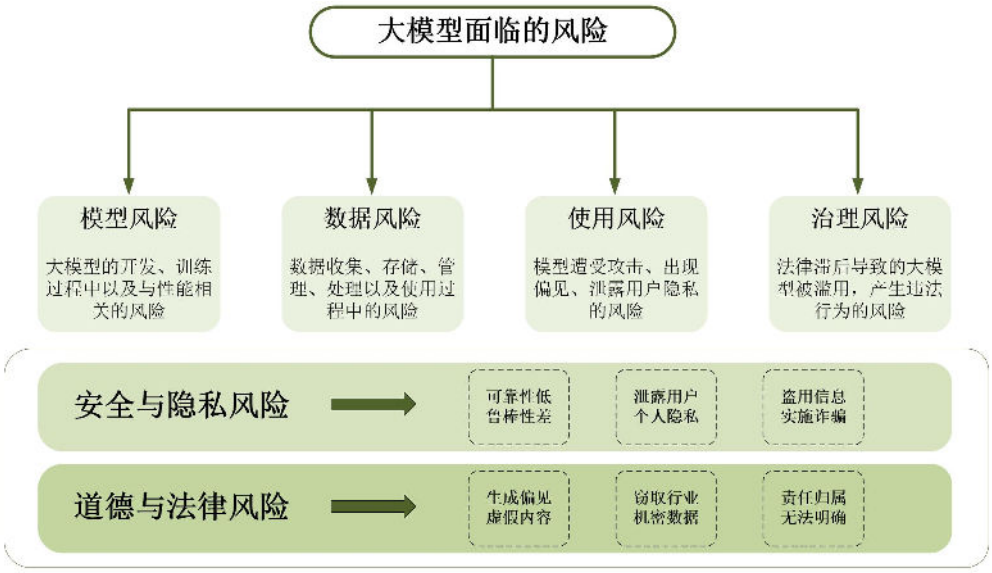


图5-6 GPT大模型面临的多种风险

ChatGPT 等大模型发布后，AI 技术所带来的风险也日益凸显，攻击者可能 利用各种手段，对大模型训练数据或输入数据进行篡改、污染或窃取。且现有大 模型数据来源于人类社会，其决策很可能隐含着道德偏见。然而，相关法

律法规 的制定却不可避免地滞后于大模型技术的快速发展，这很可能会导致安全隐私与道德伦理风险显著升高。

例如，当前以ChatGPT 为代表的聊天机器人在“创作”过程中大量学习和使用语料库中他人作品的内容，可能导致“智能洗稿”，原作者的权益应当如何保护也是非常值得关注的问题。在教育领域，ChatGPT 也带来了相关的学术伦理挑战。学生可能会利用ChatGPT 制作本不属于其自身的作品，导致抄袭、剽窃等“学术不端”的行为出现，进而影响教育和学术生态。在此情况下，2023 年1 月27 日，《科学》杂志就曾发表评论文章，明确拒绝了ChatGPT 的作者署名权。 此外，清华大学助理教授于洋曾带领团队对ChatGPT 的前身GPT-2 进行相关测 试，发现GPT-2 存在70.59%的概率将教师预测为男性，60.03%的概率将医生预 测为男性，但总把厨房里的人识别为女性，这表明它会“重男轻女，爱白欺黑（种族歧视）”。

研究人员还发现，预训练模型容易受到对抗性样本的影响，原始输入的微小 干扰可能会误导预训练模型产生特定的错误预测[62]。同时，可以通过查询域名来 恢复数据样本，这可能会导致隐私泄露，训练数据集和参数量较大的模型更容易受到攻击[63]。

因此，各国家和地区普遍高度重视研究与治理AIGC 带来的安全性问题，并 对其带来的风险与挑战进行系统分析。然而，尽管各国家和地区已竞相监管人工 智能，努力填补法律空缺，相关政策的发布相对于大模型技术的发展仍较为缓慢。 如果缺乏及时的立法约束，可能会让一些不法分子钻法律的漏洞，做出利用大模 型窃取数据和隐私等行为，危害社会安全。为此，应重视大模型的研发与相应配 套的监管协同发展，全球各国家和地区也需加强治理框架之间的互操作性，深化共同合作，从而找到适合整个国际社会的人工智能治理机制。

通信行业的数据较为复杂且不少需要保密，其中工业和信息化领域数据包括 工业数据、电信数据和无线电数据等，这些数据专业化程度高，体量庞大而多样，且质量不一致，给数据保护带来一定的困难。

例如，2023 年4 月，三星员工就曾在使用ChatGPT 处理工作时，无意间泄露了公司的绝密数据。不过三星的这些商业机密还只流传到OpenAI 公司内部

服 务器，没有进一步扩散，因此还没有造成严重的影响。但是在竞争激烈的半导体 行业，任何形式的数据泄露都可能给厂商带来灾难性打击。此外，作为通信运营 商，对于用户使用GPT 的通信行为也有义务进行保密，否则有可能造成用户住 址、工作单位和个人习惯等隐私信息泄露，如果被不法分子加以利用，进行诈骗或者威胁，很可能造成严重的后果，甚至危害人身安全。

因此，针对GPT 使用过程中的安全隐私与道德伦理风险，设立相关法律法 规具有重要的意义，且需要提前进行风险预判，加快政策的制定速度，才能保障用户在使用过程中同时获得便利性和安全性。

6.发展建议与未来展望

全球移动通信经历了从1G 到4G 的跨越式发展，目前已进入5G 商用阶段。而6G 将在5G 的基础上进一步拓展和深化物联网的应用范围和领域，持续提升现有网络的基础能力，并不断发掘新的业务应用，以服务智能化社会和生活，实现从“万物互联”到“万物智联”的跃迁。

如今，GPT 在通信行业的应用场景不断丰富，由大模型驱动的6G 智慧内生 网络也开始构建，“GPT+通信”融合发展已然成为不可阻挡的趋势。为了高效 而准确地处理通信行业的海量数据，除了需要对算力发展、空口技术等提出更高 的要求，还需要制定更多政策和标准来帮助“GPT+通信”合法合理地健康发展。 因此，针对第5 章指出的“GPT+通信”融合发展中面临的问题和挑战，本章提出了一些具体的发展建议，并且对未来的趋势进行了展望。

6.1.发展建议

6.1.1. 加快 AI 算力建设，提供基础设施支撑

目前，各行各业不同领域几乎都在开发自己的大模型，通信行业作为“万物互联”时代信息传输的承担者，自然应当提前布局，规划好未来的发展道路。算力作为AI 的三大基础要素之一，正变得前所未有的重要。AI 应用的快速发展带 来了长期、海量的计算需求，其中，高算力的基础设施能够加快数据处理和分析 的速度，推动复杂算法模型的应用和优化，为人工智能的创新提供更广阔的空间。 无论是大模型的训练、推理还是部署，抑或是商业模式的创新，都需要算力作为 支撑，同时，算力还是数据处理和应用的平台。除了集中的大型算力中心外，通 信与计算深度融合使得通信终端、边缘计算、工业模组、移动通信基站和通信网络设备等也都不同程度地嵌入了计算能力。

从国家层面来看，算力已成为衡量国力的重要标准，各国都在制定人工智能 战略和政策，以推动AI 产业发展。我国中央政府和各省市也高度重视，相继出 台了许多相关的发展政策。2024 年2 月19 日，国务院国资委召开“AI 赋能

产业 焕新”中央企业人工智能专题推进会，强调中央企业要将发展人工智能放在全局 工作中统筹谋划，深入推进产业焕新，加快布局和发展智能产业，加快建设一批智能算力中心，开展“AI+ ”专项行动。会上，10 家中央企业签订了倡议书，表示将主动向社会开放人工智能应用场景。

上海市正持续夯实算力基础设施建设，助力城市数字化转型，建设“算赋百业”生态初具规模，也通过推动“算力浦江”计划演进升级，助力构建全国一体 化算力网。河北张家口市也依托本地交通区位、地理气候、自然资源、绿色电力等独特优势，加速推进大数据产业发展，着力构建“一廊四区多园”的大数据产业空间布局和“1+3+9+*N*”的大数据产业发展体系，加速建设京津冀“算力之都”。 此外，浙江省政府办公厅于2024 年1 月印发的《关于加快人工智能产业发展的 指导意见》中也明确提出了发展目标：到2027 年，人工智能核心技术取得重大突破，算力算法数据有效支撑场景赋能的广度和深度全面拓展，全面构建国内一 流的通用人工智能发展生态，培育千亿级人工智能融合产业集群 10 个、省级创 新应用先导区 15 个、特色产业园区 100 个，人工智能企业数量超3000 家，总营业收入突破 10000 亿元，成为全球重要的人工智能产业发展新高地。

近年来，我国算力规模稳步扩张，智能算力保持强劲增长，算力发展为拉动 我国 GDP 增长做出突出贡献。中国信通院发布的数据显示，在 2016—2022 年期 间，我国算力规模平均每年增长46%，数字经济增长 14.2%，GDP 增长 8.4%。 根据国际数据公司（International Data Corporation，IDC）数据，中国智能算力规 模2023 年将达到414. 1EFLOPS，2022—2027 年复合增长率达33.9%。当前国家 高度重视算力建设，AI“需求+政策”驱动智能算力市场持续扩容，国产算力应 用逐步加速，而智算中心是算力发展的关键，如图6-1 所示，中国智能算力规模正不断扩大。

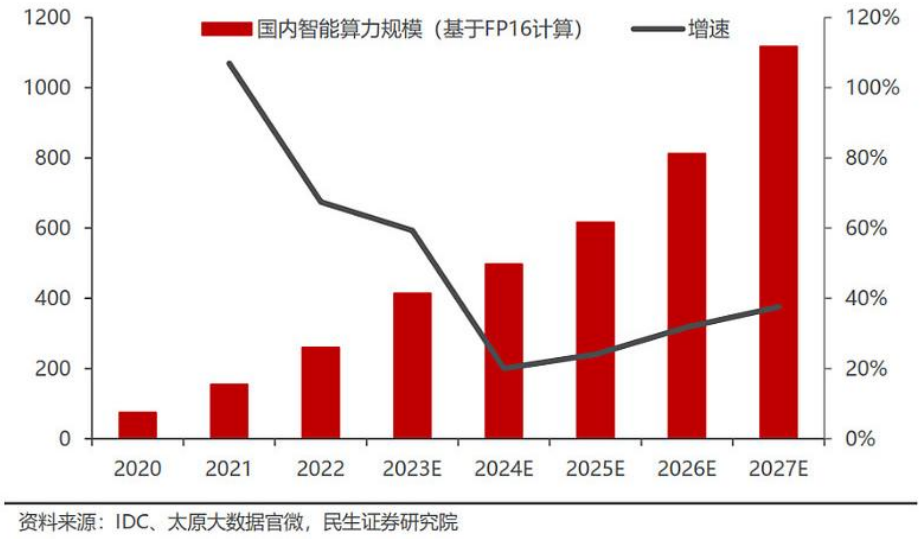


图6-1 中国智能算力规模及预测

算力作为AI 时代的核心“引擎”和通用刚需资源，已成为支撑数字经济持续纵深发展的新动能，赋能各行各业的数字化转型升级。但当前国内AI 算力中 心建设仍面临顶层制度建设和标准体系不统一、建设方向和建设需求错位等问题。 解决上述行业发展问题的关键因素之一，在于应当从应用场景中获取实践经验， 优化行业解决方案，推动AI 产业的全面健康发展。积累数据资源、提升算力水平、做大做强算力产业，已经成为全球各国发展的战略选择。

6.1.2. 加强校企联合培养，填补创新人才空缺

随着我国产业持续转型升级及国际竞争的加剧，高技能人才已成为国家竞争 力的重要支撑。大模型已经成为AI 发展的新方向，同时对教育改革与人才培养 也产生了结构性的影响，而打造大模型技术产业生态的关键也在于培育优秀的创 新型专业人才。在各领域数字化转型的大背景下，当前各类用人单位亟需数字化 人才。我们需要结合当下社会对复合型AI 人才的需求进行综合考虑，创新AI人才培养模式，以适应大模型时代的挑战和机遇。

对此，我国在国家层面有着清醒的认识，并在持续推进。2024 年2 月7 日， 人力资源和社会保障部、教育部、科技部等七部门联合印发了《高技能领军人才 培育计划》（以下简称《计划》），提出从2024 年至2026 年组织实施高技能领军人才培育计划，旨在通过3 年的努力，新培育领军人才1.5 万人次以

上，并带动新增高技能人才500 万人次左右。新时代背景下，我国高度重视相关领域的“自主可控”。其中，高技能人才的自主培养无疑是“全面提高人才自主培养质量” 的重要一环。在这方面，《计划》明确指出要“强化企业主体责任”。为了确保对高技能领军人才的全面、客观评估，作为用人主体的企业需要构建一套多元化、 动态化的人才评价体系。该体系应涵盖创新能力、团队协作力及行业影响力等指标，并随着产业的变革和国家战略的调整，适时更新评价标准，其中包括市场和社会评价。

首先，必须加强人工智能理论的教育。建议普通高校、职业院校设立交叉学 科，增设人工智能相关专业和自主技术路线教学内容，积极在人工智能学科专业 教学中设置场景创新类课程，从而激发学生的场景想象力，提升学生场景创新素 养与能力。例如，在大模型背景下，计算机类专业人才需要具备更加全面的技术能力和素养。除了需要掌握大模型的基本理论和算法外，他们还应具备数据处理、分布式计算、云计算等方面的技能。且大模型需要海量数据进行训练和调优，因而数据管理和处理能力也成为计算机类专业人才所需要具备的重要素养之一。

其次，应该建立完善的校企联合人才培养体系，包括高校学习培养、企业内 部培训、科研机构实践等。从当前企业对数字化人才的需求来看，相对于学校培 养出的“理论派”毕业生，企业更希望招聘到同时具备先进技术与实践经验的员 工。同时，诸多数字化人才岗位的招聘需求都明确指向了复合型人才。鼓励开展 场景创新人才培训，通过企业岗位培训、关键岗位实践、重点项目参与等方式， 以及开设研修班、开展交流讨论会、组织场景专题培训等多种形式，培养一批具有场景创新意识和能力的专业人才。

此外，还应引进人工智能全球顶尖人才，加强国际人才交流与学习。通过与 世界一流科研机构的合作交流，从海外引进优秀师资力量，吸引企业高级人才和 行业专家进入学校授课等方式，扩大教学力量，从而培育聚集科技领军人才、卓 越工程师和青年科技人才，填补当前存在的人才空缺，满足不断增长的创新发展 需求。为了加速大模型技术的产业化，还应该积极推动产学研结合。通过建立产 学研联合实验室、共享创新平台等方式，引导企业和高校、科研机构紧密合作，共同开展大模型技术的研究与开发。

6.1.3. 加速制定相关政策，建立平台引导发展

当前，全球各国大模型研究已呈现白热化竞争态势。大模型扎堆出炉的背后 潜藏着不少问题，包括技术仍存软肋、治理体系尚待优化、盲目跟风、资源消耗 巨大、发展路径有待明晰等。例如，ChatGPT 等大模型仍可能产生一些存在偏见 或误导性的回答，甚至编造虚假内容。这很容易导致误解，甚至引发纠纷，需要 开发者和企业在使用时提前了解潜在问题，并采取相应的措施和监管政策来减少 其影响。为此，宜推动大模型底层技术研究和应用创新、建立健全大模型监管机制、引导资本市场理性投资、加强国际合作与交流。

从风险治理角度来看，国家政府应当提前布局，加快相关政策的制定速度，整体规划大模型发展路径，提供更多的发展平台和机会，并积极推动跨部门、跨 领域的监管协同，形成全方位、多层次的监管格局，从而提高监管效能。相关部 门也需要制定法律和伦理规范，明确相关技术在应用方面的限制和义务，保障公众的安全和利益。同时加强国际合作和标准化建设，形成一套共识性、全球性伦理准则和治理框架，推动建立“以人为本”“智能向善”的发展生态。

在大模型风险治理的政策制定中，最重要的就是对关键数据进行管理。为了 促进大模型训练数据的合规使用和高质量输出，尤其需要加强对大模型训练数据 的源头管控，对训练数据进行规范。可以考虑对大模型训练数据尤其是合成数据 建立托管机制[64]。监管机构则通过对训练数据托管方的约束，进一步规范大模型 训练数据生产方和使用方的行为。数据托管方可按规定对大模型训练数据来源、 处理结果、数据流向以及训练结果等进行监测，确保大模型训练数据来源可靠， 在数据标准、数据质量、数据安全、隐私保护等方面依法合规，以保障大模型输 出结果的高质量并符合监管要求。同时，还要加强风险防范意识，避免盲目投资 和短期行为，共同建立大模型产业链协同机制，培育壮大以科技领军企业为龙头、 以专精特新科技“小巨人”企业为骨干的人工智能企业梯队，促进上下游企业共同发展，打造国际一流的人工智能开放创新平台，如图 6-2 所示。

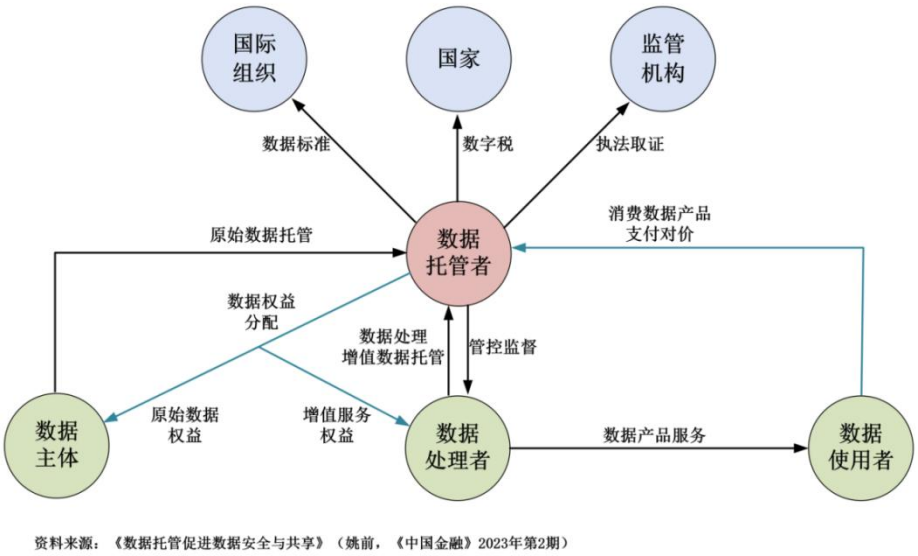


图6-2 数据托管与权益分配机制

从行业发展角度来看，建议强化“伦理先行”意识、加强行业自律自治，共同打造GPT 应用良性发展生态。开发者需要监督和改进相关应用，以消除其潜 在偏见和回答不准确等问题。企业需要结合自身资源条件和发展需求，加强数据 归集、算力统筹、算法开源等平台和基础能力建设，支持GPT 大模型赋能数字 经济。政府需要通过相关政策进行正向引导，实行宏观规划，着重加强对GPT 技术的监督与管理，以确保其在应用过程中合法合理，同时明确权责分配等内容，避免恶意和无序竞争，充分释放大模型应用价值潜力。同时还应强调道德和伦理的约束，引导科研人员和企业在研发与应用过程中秉持正确的道德价值观，注重社会责任，确保技术透明合理。

在此基础上，各单位可以基于开源共享平台促进协同合作、加速应用创新， 围绕“GPT+通信”产业发展与治理需求，推动行业层面在算力提升、算法设计、 AI 工程化等方面的联合攻关。特别是努力突破技术局限，打破行业发展瓶颈， 积极参与GPT 应用与治理等领域的国际规则制定和全球发展合作，开放更多的应用平台，争取更大的国际影响力和话语权。

6.2.未来展望

6.2.1. 核心技术实现突破，关键能力显著增强

如今通信行业可以利用GPT 的智能分析能力和强大的生成能力实现更智能 的网络管理，提高网络性能和效率。未来GPT 也将以更大的规模应用到通信行 业的各个领域，6G 网络也将原生支持GPT 功能，各种通信相关指标和性能都将得到大幅度提升。

6G 将拥有原生的AI 能力，空口和网络将采用端到端人工智能和机器学习来 实现定制化优化和自动化运维。而且每个6G 网元都将原生集成通信、计算和感 知能力，促进从云端的集中式智能向边缘的泛在智能演进。要实现全面智能普惠， 而不是局限在某些专有应用范围内的智能，需要与作为基础设施的通信网络紧密 结合，在通信网络中提供各类泛在智能所需要的基础平台、资源和能力，包括计 算、数据、存储、训练和推理服务等，从而使智能作为普惠性的服务，像当前通信网络提供的无处不在的连接服务一样，能够高效、低成本地向大众提供。

在应用维度上，内生AI 和泛在感知的6G 网络将提供更全面和综合的能力， 意图驱动以及最少人工干预的智能管理与运维是提升管理和运营效率的关键。鉴 于网络运维和管理的方式、过程和指令完全可以描述为人类语言或文本交互的问 答模式，因此借助数字孪生网络，基于其试错和预测能力提供模型细化的评判并 预训练裁判模型，在大量网络运维和管理的数据及专家知识基础上，持续强化训练网络管理和运维领域的GPT，最终实现通用智能化网络运维与管理。

在基础设施维度上，6G 网络在内生AI 和感知能力的加持下，一方面将成为 一个泛在的分布式大算力平台，同时又是一个泛在的移动大数据平台，这必定契合未来大模型强算法的部署与应用。云计算正在逐渐成为数字世界的“中枢神经”，算力云化指的是基于云计算技术向社会各组成部分提供通用计算、智算、超算等 算力资源和服务。未来GPT 将不断推动云算力服务全面升级和产业数字

化转型， 利用云服务形成算力、网络、人工智能、区块链等多要素融合的一体化服务，推动算力经济供给侧结构性改革，激发算力服务的范式创新。

6.2.2. 体系建设日益完善，数字经济快速发展

随着GPT 技术的不断成熟和各领域大模型的不断涌现，使得类GPT 产品可实现低门槛定制开发，应用商店加速产品落地推广，以GPT 为核心的模型生态 建设加速推进，验证大模型强大商业潜力。伴随模型、工具、平台全面提升，大 模型有望构筑生态核心，推动AI 商业化进程加速和市场天花板打开。同时新型 人工智能芯片的突破，也将不断推进人工智能框架软件、基础硬件和终端操作系 统等的研发应用。智能网联、北斗导航、低空卫星通信等基础设施建设也将不断加强，自动驾驶汽车、无人机、无人船等智能交通装备也会越来越普及。如图6-3 所示，以标准规范、技术研发、内容创作、行业应用、产权服务为 核心的GPT 生态体系架构也将日趋完善，无论是以GPT 赋能产业升级还是以GPT 自主释放价值都将在此框架下健康有序发展。

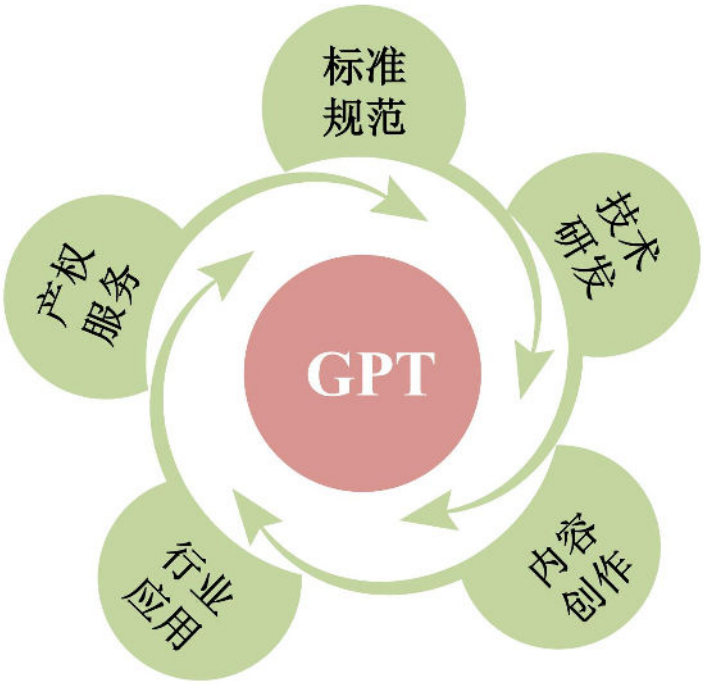


图6-3 GPT生态体系架构

标准规范构建了涵盖技术、内容、应用、服务的全过程体系，促进G大模型在合理、合规、合法的框架下良性发展。同时，在核心技术持续演进和关键能力显著增强的背景下，性能更强大、逻辑更智能的 AI 算法将被应用于GPT， 技术研发的不断创新将强有力地推动内容创作，提高生成内容质量，使内容更接 近人类智力水平和审美标准，同时应用于各类行业各种场景。GPT的发展还将 促进产权服务快速跟进，生成内容进行合理评估，保护相关创作者的产权，并进行价值重塑，充分释放其商业潜力，从而构建 GPT 大模型的经济循环体系。

近年来，我国数字经济快速发展、成效显著，已成为我国经济增长的新动能、 高质量发展的重要引擎。一方面，以互联网、云计算、大数据等数字技术驱动的 新兴产业有力拉动经济增长；另一方面，数字技术与产业深度融合，催生新业态 新模式，传统产业发展动能不断增强。中国信通院数据显示，2022年我国数字 经济规模达到50.2 万亿元，总量稳居世界第二，同比增长 10.3%，占GDP 比重 达到41.5%。国新办新闻发布会数据也显示，数字经济核心产业销售收入同比增长8.7%。

未来随着B5G/6G、云计算、VR、AR 等前沿技术的快速发展和新一代智能 终端设备的研发创新，完整的GPT 生态链将是释放数据要素红利、助力传统产 业升级、促进数字经济发展、构建数实融合一体、创造元宇宙世界最重要的推动力之一。

6.2.3. 应用场景不断拓展，循序渐进融合共生

随着GPT 不断迭代升级演进的同时，大模型的应用范围也不断扩大，正为 全行业智能化转型拓展出无穷无尽的新空间，迸发出源源不断的新动能。GPT 与通信融合发展，未来在社会生产生活各个领域的广泛应用将激发全新体验。为 了更好地推动GPT 创新应用和产业发展，需要产学研用多方参与和协同，在把 握好行业智能化发展趋势的前提下，不断追求技术创新，聚焦工程实践，确保人工智能安全可信，为人工智能造福人类保驾护航。

GPT 与通信的融合发展，将不断加速创新场景赋能，打造人工智能创新应 用先行地，并不断拓展更多的新型应用场景。例如，在科学技术创新方面，将建 立人工智能驱动的科学研究专用平台，构建“人类科学家+AI 科研助手”的人机 协同科研新模式。在实体经济发展方面，大模型在工业领域落地应用，分级诊断 评估标准不断完善，引导企业数字化转型和智能化升级。在社会智能化方面，智慧医院和智慧诊疗建设也将继续完善，包括疾病风险预测、医用机器人和智能公共卫生服务等应用场景。此外，在线教育也将实现虚拟课堂、AI 教育助手等创 新场景，同时建设智慧图书馆和智慧校园。在城市现代化治理方面，智慧交通将 加速建设，持续提升交通运行监测、出行信息服务和应急指挥能力，从而提升公共安全治理能力。

在大数据时代，如何将数据收集、分析、提炼，用于改善社会生活，是 AI 的基础。可以预见，GPT 的流行会让人机自然对话的交互方式越来越普及，并 不断从文本对话模式扩展到语音对话以及与数字人面对面交流，进而大大提升信 息通信流量和用户黏度，这也将推动信息通信从人与人之间的沟通交流扩展到人 机之间的沟通交流。“GPT+通信”将持续深度融合发展，其相关技术的进步和 创新，将提供更多的机会，实现创造、学习和协作。因此只要不断围绕通信和大 模型协同创新，构建开放共享的创新生态，促进人工智能与通信产业的深度融合， 将会继续加速构建下一代信息基础设施，助力经济社会数字化转型，未来在6G时代实现真正的“人机融合”以及“万物智联，数字孪生”的美好愿景。

7.结束语

GPT 的出现，推动了AI 技术的发展，改变了人们的生活习惯，也促进了相关技术和应用在各行各业加速落地，是AI 发展史上一个新的里程碑。

通信网络作为现代信息化社会的关键基础设施，对于更高效、更稳定、更智 能的需求也日益增长。随着5G 移动通信网络的发展，网络架构更加复杂，终端 类型和业务种类也日益增加，即将到来的6G 移动通信网络还必须具备内生智能 和感知等能力。多样化的应用和通信场景、超密集的网络连接以及极致性能的服 务需求，都对移动通信网络提出了更高的要求。因此，迫切需要对GPT 大模型与通信行业相结合进行研究。

本白皮书分享了作者对通信行业与GPT 互相赋能、逐渐融合，进而协同发 展的探究和思考。首先介绍了GPT 的概念与发展历程，然后重点调研了GPT 在 通信领域的智能客服、仿真编程和辅助芯片设计等创新应用，介绍了未来网络支 撑GPT 应用的思路和内生智能的6G 网络。此外，还对GPT 与通信融合发展， 赋能行业数字化转型的应用场景以及发展中面临的问题进行了说明，并提出了一些发展建议和对未来的展望。

期待对通信与GPT 大模型感兴趣的相关专业人士能够通过阅读本白皮书有 所收获，同时本白皮书仍存在不足之处，敬请大家批评指正。未来，希望能够推 进通信产业和大模型领域深度融合创新，构建绿色低碳的基础设施和开放共享的良好生态，促进新一代智能通信网络发展，加速经济社会的数字化转型。

参考文献

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv: 1706.03762,2017.

[2] Yang J, Jin H, Tang R, et al. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond[J]. arXiv preprint arXiv:2304.13712,2023.

[3] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. 2018, https://cdn.openai.com.

[4] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, https://cdn.openai.com/better-language-models.

[5] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. arXiv preprint arXiv:2005.14165，2020.

[6] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Neural Information Processing Systems, 2022, 35: 27730-27744.

[7] State of AI Report 2023. 2023, https://www.stateof.ai.

[8] Qu Y, Liu P, Song W, et al. A text generation and prediction system: Pre-training on new corpora using BERT and GPT-2[C]//2020 IEEE 10th international conference on electronics information and emergency communication (ICEIEC). IEEE, 2020: 323-326.

[9] Zimmermann D, Koziolek A. Automating GUI-based software testing with GPT-3[C]//2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). IEEE, 2023: 62-65.

[10]Mathur A, Pradhan S, Soni P, et al. Automated test case generation using T5 and GPT-3[C]//2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2023: 1986-1992.

[11]Jeong S W, Kim C G, Whangbo T K. Question answering system for healthcare Information based on BERT and GPT[C]//2023 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on

Electrical, Electronics, Computer and Telecommunications Engineering (ECTIDAMT & NCON). IEEE, 2023: 348-352.

[12]Krstic D, Petrovic N, Suljovic S, et al. AI-enabled framework for mobile network experimentation leveraging ChatGPT: Case study of channel capacity calculation for η-μ fading and co-channel interference[J]. Electronics, 2023, 12(19): 4088.

[13]Xia L, Sun Y, Liang C, et al. Generative AI for semantic communication: Architecture, challenges, and outlook[J]. arXiv preprint arXiv:2308.15483, 2023.

[14]Blocklove J, Garg S, et al. Chip-Chat: Challenges and opportunities in conversational hardware design[C]//2023 ACM/IEEE 5th Workshop on Machine Learning for CAD(MLCAD). IEEE, 2023: 1-6.

[15]Gelenbe E, Domanska J, Fröhlich P, et al. Self-aware networks that optimize security, QoS, and energy[J]. Proceedings of the IEEE, 2020, 108(7): 1150-1167.

[16]ITU-R. Framework and overall objectives of the future development of IMT for 2030 and beyond,2023,https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2030/Pages/default.aspx.

[17]Shakya S, Roushdy A, Khargharia H S, et al. AI based 5G RAN planning[C]//2021 International Symposium on Networks, Computers and Communications (ISNCC). IEEE, 2021: 1-6.

[18]Vijayakumar D, Nema R K. Superiority of PSO relay coordination algorithm over non-linear programming: A Comparison[C]//2008 Joint International Conference on Power System Technology and IEEE Power India Conference. IEEE, 2008: 1-6.

[19]Wang J, Zhang L, Yang Y, et al. Network meets ChatGPT: Intent autonomous management, control and operation[J]. Journal of Communications and Information Networks, 2023, 8(3): 239-255.

[20]屈军锁,唐晨雪,蔡星,等.人工智能与通信网络融合趋势[J].西安邮电大学学报, 2021, 5:26.

[21]Hu Z, Dong Y, Wang K, et al. GPT-GNN: Generative pre-training of graph neural networks[C]//Proceedings of the 26th ACM International Conference on

Knowledge Discovery & Data Mining (SIGKDD). 2020: 1857- 1867.

[22]张彤，任奕璟，闫实，等. 人工智能驱动的 6G 网络：智慧内生[J]. 电信科学，2020, 36(9):14-22.

[23]Liu Y, Peng M, Shou G, et al. Toward edge intelligence: Multiaccess edge computing for 5Gand Internet of things[J]. IEEE Internet of Things Journal, 2020, 7(8): 6722-6747.

[24]中国移动通信集团有限公司. 中国移动 6G 网络架构技术白皮书[R]. 2022.and Internet of things[J]. IEEE Internet of Things Journal, 2020, 7(8): 6722-6747.

[25]华为. NET4AI：6G 支持AI 即服务.2022.

[26]Edward R G, Chris G, David A W, Martin J B. Framework for cyber-physical systems:Volume 1, overview[J]. NIST Special Publication, 2017: 1500-201.

[27]华为. 6G：无线通信新征程白皮书[R].2022.

[28]Regulation (EU) 2016/679 of the European Parliament and of the Council. Regulation, 2016，https://eur-lex.europa.eu/legal-content.

[29]Chen X, Zhou M, Wang R, et al. Evaluating response delay of multimodal interface in smart device[C]//Design, User Experience, and Usability. Practice and Case Studies: 8thInternational Conference, DUXU 2019, 2019: 408-419.

[30] 王凌豪,王淼,张亚文,张玉军.未来网络应用场景与网络能力需求[J]. 电信科学,2019,35(10)：2-12.

[31]Munir A, Blasch E, Kwon J, et al. Artificial intelligence and data fusion at the edge[J]. IEEEAerospace and Electronic Systems Magazine, 2021, 36(7): 62-78.

[32]彼得·李,凯丽·戈德伯格,伊萨克·科恩.超越想象的GPT 医疗社科[M].杭州：浙江科学技术出版社,2023.

[33]Seppo H, Henning S, Cinzia S. LTE self-organising networks (SON): Network management automation for operational efficiency[M]. John Wiley & Sons, 2012.

[34]Iacoboaiea O, Sayrac B, Jemaa S B, et al. SON conflict diagnosis in heterogeneous networks[C]//2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). IEEE, 2015: 1459-1463.

[35]Isa I N M, Baba M D, Yusof A L, et al. Handover parameter optimization for self-organizing LTE networks[C]//2015 IEEE symposium on computer applications & industrial electronics (ISCAIE). IEEE, 2015: 1-6.

[36]Luketić I, Šimunić D, Blajić T. Optimization of coverage and capacity of self-organizing network in LTE[C]//2011 Proceedings of the 34th InternationalConvention MIPRO. IEEE, 2011: 612-617.

[37]López-Pérez D, Chu X, Vasilakos A V, et al. On distributed and coordinated resource allocation for interference mitigation in self-organizing LTE networks[J]. IEEE/ACM Transactions on Networking, 2012, 21(4): 1145-1158.

[38]Andrews J G, Buzzi S, Choi W, et al. What will 5G be?[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(6): 1065-1082.

[39]Shafique K, Khawaja B A, Sabir F, et al. Internet of things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios[J]. IEEE Access, 2020, 8: 23022-23040.

[40]Zheng Q, Guo C, Ding J, et al. A wideband low-RCS metasurface-inspired circularly polarized slot array based on AI-driven antenna design optimization algorithm[J]. IEEE Transactions on Antennas and Propagation, 2022, 70(9): 8584-8589.

[41]Sharma A, Devalia D, Almeida W, et al. Statistical data analysis using GPT3: An overview[C]//2022 IEEE Bombay Section Signature Conference (IBSSC). IEEE, 2022: 1-6.

[42]Letaief K B, Shi Y, Lu J, et al. Edge artificial intelligence for 6G: Vision, enabling technologies, and applications[J]. IEEE Journal on Selected Areas in Communications, 2021, 40(1): 5-36.

[43]Gopinath R, Jensen C, Groce A. Code coverage for suite evaluation by developers[C]//Proceedings of the 36th International Conference on software engineering. IEEE, 2014: 72-82.

[44]Ammann P, Offutt J. Introduction to Software Testing[M]. Cambridge University Press, 2016.

[45]Panchal B, Parmar S, Rathod T, et al. AI and blockchain-based secure message exchange framework for medical Internet of things[C]//2023 International Conference on Network, Multimedia and Information Technology (NMITCON). IEEE, 2023: 1-6.

[46]Kuzlu M, Xiao Z, Sarp S, et al. The rise of generative artificial intelligence inhealthcare[C]//2023 12th Mediterranean Conference on Embedded Computing (MECO). IEEE, 2023: 1-4.

[47]Cao H, Garg S, Kaddoum G, et al. Softwarized resource management and allocation with autonomous awareness for 6G-enabled cooperative intelligent transportation systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 24662-24671.

[48]Ranjha A, Kaddoum G, Dev K. Facilitating URLLC in UAV-assisted relay systems with multiple-mobile robots for 6G networks: A prospective of agriculture 4.0[J]. IEEE Transactions on Industrial Informatics, 2021, 18(7): 4954-4965.

[49]Esswie A A, Repeta M. Evolution of 3GPP standards towards true extended reality (XR) support in 6G networks[J]. arxiv preprint arxiv:2306.04012, 2023.

[50]De Lima C, Belot D, Berkvens R, et al. Convergent communication, sensing and localization in 6G systems: An overview of technologies, opportunities and challenges[J]. IEEE Access, 2021, 9: 26902-26925.

[51]Serghiou D, Khalily M, Brown T W C, et al. Terahertz channel propagation phenomena, measurement techniques and modeling for 6G wireless communication applications: A survey, open challenges and future research directions[J]. IEEE Communications Surveys & Tutorials, 2022, 24: 1957-1996.

[52]Elsayed M, Erol-Kantarci M. AI-enabled future wireless networks: Challenges, opportunities, and open issues[J]. IEEE Vehicular Technology Magazine, 2019, 14(3): 70-77.

[53]冯帅帅,张佳星,罗教讲.AI 时代社会科学研究方法创新与模型“过度拟合”问题探索[J]. 社会科学杂志, 2023, 1(1)：157-184.

[54]Akrout M, Mezghani A, Hossain E, et al. From multilayer perceptron to GPT: A reflection on deep learning research for wireless physical layer[J]. arXiv preprint arXiv:2307.07359, 2023.

[55]Kaddour J, Harris J, Mozes M, et al. Challenges and applications of large language models[J]. arXiv preprint arXiv:2307.10169, 2023.

[56]Chen Z, Zhang Z, Yang Z. Big AI models for 6G wireless networks: Opportunities, challenges, and research directions[J]. arXiv preprint arXiv:2308.06250, 2023.

[57]Alizadeh K, Mirzadeh I, Belenko D, et al. LLM in a flash: Efficient large language model inference with limited memory[J]. arXiv preprint arXiv:2312.11514, 2023.

[58]Yao J, Zhang S, Yao Y, et al. Edge-cloud polarization and collaboration: A comprehensive survey for AI[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(7): 6866-6886.

[59]Nielsen J. Usability Engineering[M]. Morgan Kaufmann, 1994.

[60]Lei L, Zhang H, Yang S X. ChatGPT in connected and autonomous vehicles:Benefits andchallenges[J]. Intelligence & Robotics, 2023, 3: 145- 193.

[61]李抵非, 田地, 胡雄伟. 基于分布式内存计算的深度学习方法[J]. 吉林大学学报（工学版）, 2015, 45(3): 921-925.

[62]Zhou C, Li Q, Li C, et al. A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT[J]. arXiv preprint arXiv:2302.09419,2023.

[63]Qiu X, Sun T, Xu Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China Technological Sciences, 2020, 63(10): 1872- 1897.

[64]姚前. 数据托管促进数据安全与共享[J]. 中国金融, 2023, 2: 23-24.

缩略语

缩写	英文全拼	中文释义
3GPP	3rd Generation Partnership Project	第三代合作伙伴计划
5G	The Fifth Generation	第五代移动通信系统
6G	The Sixth Generation	第六代移动通信系统
AI	Artificial Intelligence	人工智能
AIGC	Artificial Intelligence Generated Content	人工智能生成内容
AR	Augmented Reality	增强现实
ETSI	European Telecommunication Standards Institute	欧洲电信标准化组织
GNN	Graph Neural network	图神经网络
GPT	Generative Pre-trained Transformer	生成式预训练转换器
GPU	Graphics Processing Unit	图形处理器
HDL	Hardware Description Language	硬件描述语言
IDC	International Data Corporation	国际数据公司
IEC	International Electrotechnical Commission	国际电工委员会
IEEE	Institute of Electrical and Electronics Engineers	电气电子工程师学会
ISO	International Organization for Standardization	国际标准化组织
KPI	Key Performance Indicator	关键性能指标

LLM	Large Language Model	大语言模型
LTE	Long Term Evolution	长期演进技术
mMIMO	massive Multi-Input Multi-Output	大规模多输入多输出
ML	Machine Learning	机器学习
NWDAF	Network Data Analytic Function	网络数据分析功能
NLP	Natural Language Processing	自然语言处理
NR	New Radio	新空口
OS	Operating System	操作系统
QoS	Quality of Service	服务质量
RLAIF	Reinforcement Learning from Artificial Intelligence Feedback	人工智能反馈强化学习
RLHF	Reinforcement Learning from Human Feedback	人类反馈强化学习
RNN	Recurrent Neural Network	循环神经网络
SA	StandAlone	独立组网
SAM	Self-Attention Mechanism	自注意力机制
SLA	Service Level Agreement	服务等级协议
SON	Self-Organizing Networks	自组织网络
VNF	Virtualized Network Function	虚拟网络功能
VR	Virtual Reality	虚拟现实
XR	Extended Reality	扩展现实

致谢

诚挚地感谢撰写团队对本白皮书做出的贡献。

主编：曾捷、杨一帆

贡献单位与人员：

北京理工大学：曾捷、杨一帆、杨铮、王紫如、叶能、朱超、于珊平、程波铭、张雨婷、王琛红

北京邮电大学：张诗语、吕铁军、王鲁晗、路兆铭、黄平牧、崔莹萍、喻茜、王梦珂、何晓宇、牛海文、李秉轩

清华大学：粟欣、李云洲、刘蓓、谭志强

中山大学：陈翔、王玺钧、郭志恒

西安交通大学：范建存、陶梦丽

西安电子科技大学：任智源、赵佳昊

桂林电子科技大学：李晓欢

中国科学院计算技术研究所：田霖、孙茜

中移智库：邓灵莉、马梦媛、何克光

中国电信：郭建章、李振、张劲松、王栋

中国联通：黄蓉、刘珊、周伟

紫光展锐：苗润泉

中信科移动：程志密

鹏城实验室：杨婷婷

维沃移动通信有限公司(vivo)：袁雁南、周通、姜大洁

图灵人工智能研究院：李强

深圳清华大学研究院、清华大学深圳国际研究生院： 郑斯辉

审核：杨铮、张雨婷