第1章 大模型入门

人工智能（AI）技术作为当今科技领域的热点之一，正在全球范围内引领一场前所未有的变革。人工智能涵盖多个子领域，包括机器学习、深度学习、自然语言处理、机器视觉、机器人学、专家系统、知识工程等。在当代社会，人工智能技术也已经在金融、交通、医疗、教育等多个领域得到了广泛应用。在金融领域，人工智能可以帮助用户识别风险、预测市场走势、提高投资决策的准确性。在交通领域，自动驾驶汽车、智能交通系统等应用正在改变我们的出行方式。在医疗领域，人工智能可以辅助医生进行疾病诊断、制订治疗方案、提高医疗服务的效率和质量。在教育领域，人工智能可以根据学生的学习情况提供个性化的教学方案，促进教育公平和提升教育质量。

1.1 人工智能发展历程

人工智能是源于人们对于大规模数据处理的需求，随着技术的发展，人工智能的发展可以大致划分为以下 4 个阶段。

1. 机器学习的基础阶段

机器学习起源于 20 世纪 50 年代。1949 年，赫布从神经心理学获得启发而提出了一种机器学习方式，现在被称为赫布型学习。

在早期的机器学习阶段，研究者主要关注符号推理和专家系统的开发。这些方法通过编写规则和知识库实现智能行为，但受限于规则的数量和复杂性，难以处理大规模和复杂的问题。

随着统计学的发展，一系列重要的机器学习算法和模型相继问世，如感知器模型、决策树算法、支持向量机等。在这一时期，统计学家也在研究损失函数最小化问题，其间优化理论做出了很大贡献。这些算法的出现大大提高了机器学习的性能和准确性，并在医疗、金融等领域中得到广泛应用。机器学习主要是使用数据手段分析大规模数据，这一阶段也为深度学习打下了基础，其中的很多算法（如线性回归等）也被应用于深度学习中。

2. 深度学习的崛起

进入 21 世纪，深度学习开始崭露头角。深度学习是一种模仿人脑神经网络结构和工作原理的机器学习方法。通过构建深度神经网络，深度学习能够自动学习数据的特征表示，从而避免了手工设计特征的烦琐过程。

这一阶段诞生了 CNN、RNN、ResNet、LSTM 等基础的深度神经网络，与机器学习相比，这些网络不仅提升了数据分析和预测的准确性，而且在图像识别、语音识别、自然语言处理等领域，深度学习模型也取得了显著的成果，并且拓宽了机器学习的应用领域。这些成果证明了深度学习在处理复杂数据和模式识别方面的强大能力。

3. 大模型的兴起

随着深度学习技术的发展和计算能力的提升，大模型开始受到越来越多的关注。大模型通常具有更深的网络结构、更大的参数规模和更强的计算能力，能够处理更加复杂和大规模的数据。

早期的大模型主要是基于统计学习的方法，如朴素贝叶斯分类器、决策树和逻辑回归等。然而，这些模型通常需要在小规模数据集上进行训练，因此性能受到一定的限制。

随后，随着深度学习技术的广泛应用和数据的爆炸式增长，大规模预训练模型成为大模型发展的重要方向。这些模型在大量的数据上进行预训练，可以学习更多的知识和特征，从而在各种任务上具有更好的性能。例如，ChatGPT、BERT 等模型在自然语言处理领域取得了巨大的成功。

4. 超大规模预训练模型的出现

近年来，随着数据资源的不断增加和计算资源的不断提升，超大规模预训练模型开始出现。这些模型具有数十亿甚至数万亿的参数规模，能够在海量数据上进行高效的训练和学习。

超大规模预训练模型的出现进一步推动了机器学习和人工智能技术的发展，使得 AI 系统能够在各种复杂场景下实现更高级别的智能和性能。

人工智能的发展历程是一个不断深化、技术不断进步的过程。在机器学习领域，通过训练模型使其具备自主学习和优化的能力，得以在海量数据中发掘隐藏的模式和规律；深度学习进一步推动了这一进程，通过构建深层次的神经网络，实现了对复杂数据的精确处理和分析，和机器学习相比普遍准确率更高；自然语言处理技术的不断进步，使得人工智能能够理解和生成人类语言，从而实现了与人类的智能交互；计算机视觉技术则让机器具备了像人一样识别、理解图像和视频的能力。

1.2　大模型崛起

1.2.1　大模型的技术概念

大模型的“大”需要从内在和外在两方面考虑。“内在”是指大模型本身的计算结构和参数量，现有的大模型是在深度学习的基础上构建的，拥有数十亿至数千亿个参数和复杂的计算结构。这样的设计是为了提高模型的认知能力，在这样强大的能力加持下，机器也

能够处理更加复杂的任务。“外在”是指训练大模型的数据规模也要大，通过海量的训练数据学习复杂的模式和特征，这样一来大模型便具有更强大的泛化能力，可以对未知的数据做出准确的预测。

大模型的应用范围非常广泛，包括自然语言处理、计算机视觉、语音识别、推荐系统等领域。在自然语言处理领域，大模型可以用于处理更复杂的任务和提升性能，如机器翻译、语音识别、文本摘要、情感分析等。当下大模型最引人关注的方向还包括内容生成能力，不仅可以生成文本还可以生成图片，这样一来，大模型的应用范围也非常广。例如在对话方面，大模型能够理解用户意图，与用户进行文字交互；再如图像问答，大模型可以理解照片中的物体，并做出准确回答。现在大模型已经可以服务于智能制造、智能交通、金融、医疗保健等领域，未来还会扩展更多应用领域。

现在大模型的表现如此出众，我们也就希望大模型能够完成更多事情，为了应对更复杂的任务和更大规模的数据，各个大模型的参数数量、层数等指标也在不断增加。同时，大模型的发展方向也更倾向于多种数据模态，比如文生图、图生文、多模态信息检索等。这种多模态的表现方式更趋于人类的交互方式，让大模型从单一的文本表达方式转为视觉表达方式，也为大模型提供了更加多样的应用方向。如前所述，大模型“大”的不只是模型结构、参数量，训练好一个大模型更需要一个庞大的数据量，因此训练大模型需要的时间成本、算力成本也是非常大的。在小数据集上快速调整大模型，迁移学习和预训练模型成为常用的手段。现在研究者也在探索自监督学习和无监督学习，它们也是未来大模型发展的重要趋势。

大模型具有强大的上下文理解能力、语言生成能力和学习能力，以及高可迁移性，这些能力都是深度学习所不具备的。和传统模型相比，大模型不仅可以处理单模态任务，还可以处理多模态任务，例如图文理解、文生图、图生文等。同时大模型不需要梯度回传，也不需要特别的训练或者微调，只需要给大模型一个指令或者几个例子，甚至在零样本（zero-shot）的场景下，大模型也能很好地完成目标任务。大模型对比深度学习来说，泛化性和实用性都有了一个很大的提升。

1.2.2　大模型诞生

大模型是继传统深度学习模型变体之后，在深度学习方向上打开的一道新世界的大门。当前流行的大模型网络架构沿用当前 NLP 领域最热门、最有效的架构，即 Transformer 架构，如图 1.1 所示。相比于 RNN 等网络，Transformer 所特有的注意力机制可以加强模型的理解能力，将重点放在重要的词上，也可以处理更长的序列。这个模型弥补 RNN、CNN 等传统循环神经网络模型的一些缺陷，一个是并行计算，另一个则是长距离输入的上下文信息。

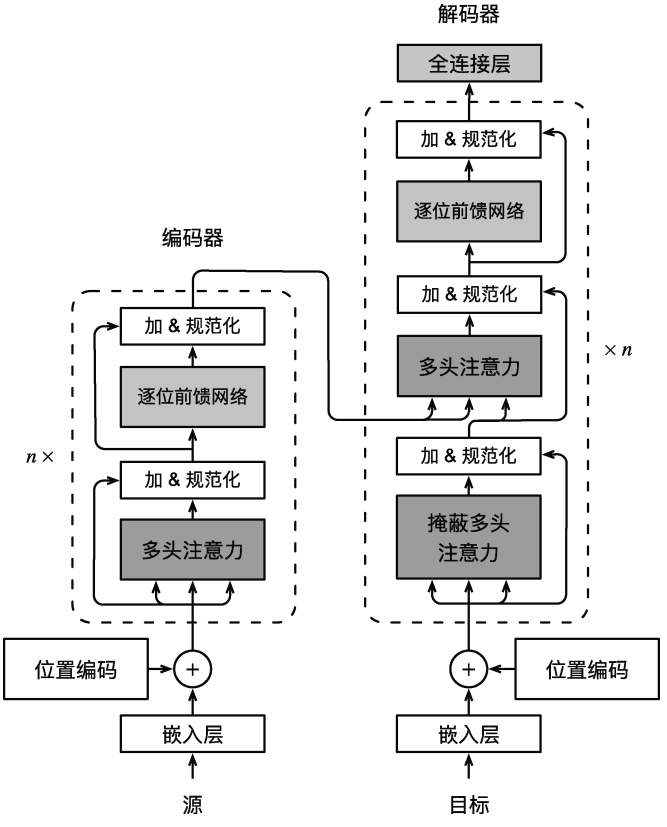

图 1.1　Transformer 模型结构

Transformer 模型最核心的两个关键点补足了 RNN、CNN 的不足，一个关键点是多层 encoder 和 decoder，另一个关键点是注意力机制。首先我们来看第一个关键点，encoder 将输入序列处理成为一个高维的向量，decoder 将这个高维向量解码成序列目标。这样的结构可以更好地让模型捕捉输入语义与上下文信息和特征。其次我们来看注意力机制，其作用是为每个输入序列的每个位置分配一个权重。注意力机制支持并行计算，因此可以大大提高训练和推理速度。这些优势都为将来的大语言模型奠定了基础。

1.2.3 迭代优化

在 Transformer 模型之后诞生的便是 BERT 模型。BERT 是基于 Transformer encoder 的深度双向模型，正因为引入了双向性，所以该模型可以同时接收输入序列左右两侧的信息，并且可以直接利用预训练模型进行微调，使模型更易迁移到任意场景，因为模型对垂域数据依赖已经减少。

在 BERT 之后，针对 Transformer 结构的掩蔽进行优化，又衍生出了一批自然语言处理模型，像是对输入序列 token 进行动态掩蔽的 RoBERT；对字、词粒度做掩蔽的 ERNIE；对序列增加掩蔽掉长度的 SPAN BERT 等，这些模型都曾在文本分类、情感分析、命名实体识别等自然语言处理的任务中表现出出色的性能。

大模型所占用的显存也相对比较大，因此也需要对模型做瘦身。除了传统的量化、剪枝、跨层参数共享等方法外，还有 ALBERT 模型使用的自注意力层且都采用相同的参数。

BERT 模型也并不是生成模型，但输入和输出长度必须是一致的。随后，一些用于翻译和文本生成的模型应运而生。之后，BART、GPT、T5 等大语言模型逐渐进入大家的视野，从此拉开了生成式大模型的序幕，如图 1.2 所示。

目前绝大多数的大模型都是在 Transformer 的基础上演变而来的。2021 年谷歌提出视觉 Transformer 模型（ViT），其架构如图 1.3 所示，将图片处理为块序列送入 Transformer 结构中进行分类操作。实验结果证明，ViT 模型在图像的一系列处理上（包括目标检测、语义分割、视频理解等）都有非常优秀的表现，如表 1.1 所示。

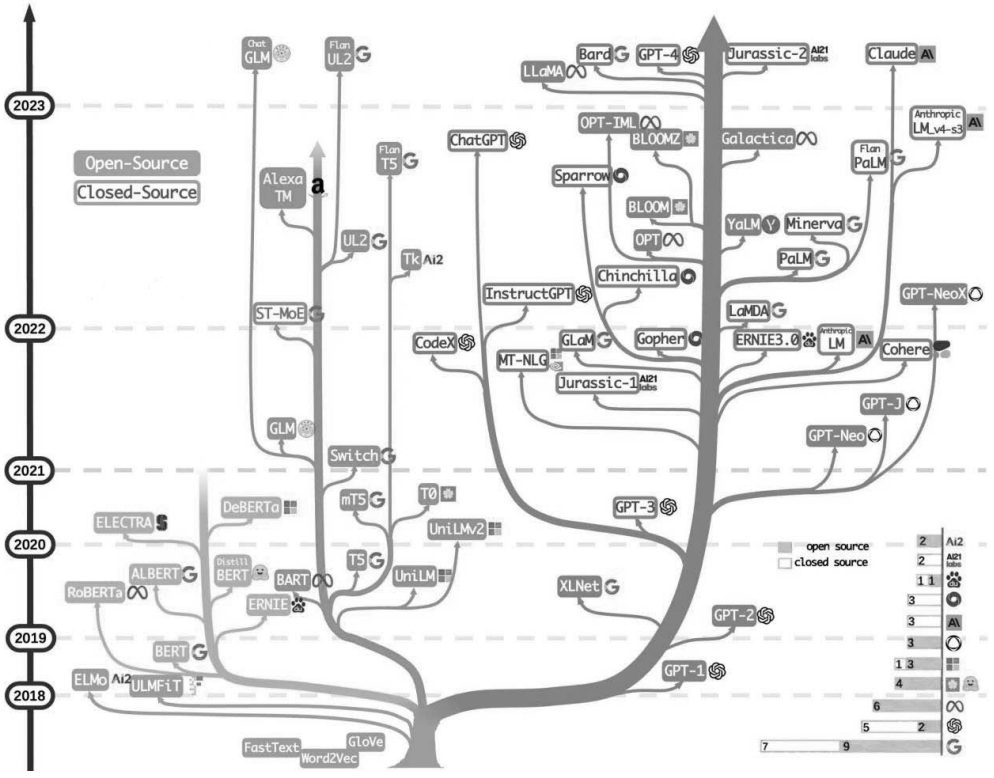

图 1.2　大模型的发展路径

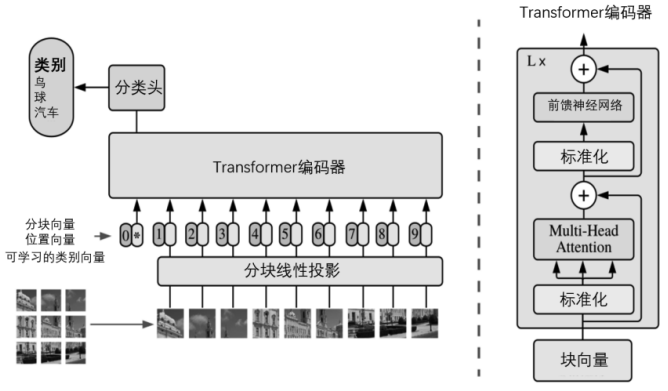

图 1.3　ViT 模型结构

表 1.1　ViT 模型结果分析

Model	Epochs	ImageNet	ImageNet Real	CIFAR-10	CIFAR-100	Pets	Flowers	exaFLOPs
ViT-B/32	7	80.73	86.27	98.61	90.49	93.40	99.27	164
ViT-B/16	7	84.15	88.85	99.00	91.87	95.80	99.56	743
ViT-L/32	7	84.37	88.28	99.19	92.52	95.83	99.45	574
ViT-L/16	7	86.30	89.43	99.38	93.46	96.81	99.66	2586
ViT-L/16	14	87.12	89.99	99.38	94.04	97.11	99.56	5172
ViT-H/14	14	88.08	90.36	99.50	94.71	97.11	99.71	12826
ResNet50 × 1	7	77.54	84.56	97.67	86.07	91.11	94.26	150
ResNet50 × 2	7	82.12	87.94	98.29	89.20	93.43	97.02	592
ResNet101 × 1	7	80.67	87.07	98.48	89.17	94.08	95.95	285
ResNet152 × 1	7	81.88	87.96	98.82	90.22	94.17	96.94	427
ResNet152 × 2	7	84.97	89.69	99.06	92.05	95.37	98.62	1681
ResNet152 × 2	14	85.56	89.89	99.24	91.92	95.75	98.75	3362
ResNet200 × 3	14	87.22	90.15	99.34	93.53	96.32	99.04	10212
R50 × 1+ViT-B/32	7	84.90	89.15	99.01	92.24	95.75	99.46	315
R50 × 1+ViT-B/16	7	85.58	89.65	99.14	92.63	96.65	99.40	855
R50 × 1+ViT-L/32	7	85.68	89.04	99.24	92.93	96.97	99.43	725
R50 × 1+ViT-L/16	7	86.60	89.72	99.18	93.64	97.03	99.40	2704

同一时期，多模态模型也进入了我们的视野中，如文生图模型、图生文模型、图文理解大模型等一系列模型，甚至是数字人生成模型、文生视频模型、代码生成模型。当然，ChatGPT 等模型也在不断地扩展功能，除了多轮对话问答之外，其对输入内容的理解能力也在不断加强并且可以承担更多的工作。

1.3　新的产品形态

当下大模型产品的发展主要是从大模型的能力提升、应用场景丰富两个方向进行的。

首先是能力提升，打造原生 AI 入口产品，科研院所及创业公司利用自身技术优势和产投研一体化能力，专注于开发原生 AI 对话型入口产品。例如，清华大学与智谱公司合作开发的 ChatGLM2 大模型，并在此基础上推出了生成式 AI 对话产品“智谱清言”，通过用户个人数据优化模型训练。这是在模型理解能力和生成能力上进行提升，主要是提供更加优质的底座大模型。另一个发展方向是在应用层面开拓更多的适配场景，大型科技公司华为凭借其行业解决方案的经验和积累，提出了“L0 基础大模型-L1→行业大模型-L2→细分场景大模型”的发展路径。华为发布的盘古大模型重点为政务、金融、制造、采矿等垂直行业应用提供服务。又如互联网大厂腾讯、百度、阿里等，它们在 AI 领域有深厚积累，并拥有丰富的自有业务场景。这些公司多采取原生 AI 入口产品+内部应用赋能+垂直行业应用全面布局战略。例如，百度基于文心大模型推出了文心一言对话产品作为 AI 入口，同时将大模型能力应用于百度搜索、信息流、智能驾驶、百度地图、小度智能屏等多个内部产品，实现全面赋能。

1.3.1　个性化与定制化

定制化大模型的概念涵盖了为企业独特需求而设计的人工智能模型，这一概念源于对传统通用模型在应对企业个性化挑战上的限制。定制化大模型的优势不仅在于其能够根据企业的独特特点进行调整，更在于其能够通过深度学习和自适应性算法实时调整，能够在特定场景下提供最好的服务，通过定制化大模型技术可以帮助企业快速开发出具有竞争力的产品和服务，以满足不断变化的市场需求并适应不断变化的业务环境。这使得企业能够更加灵活地应对市场需求和竞争压力，使其在短时间内实现更为精准的决策。

目前我们所熟知的产品（如 ChatGPT、豆包、文心一言、通义千问等大模型）都是通用场景下的，企业使用的时候往往需要根据业务需求调整回答内容。例如在办公场景下可以编写邮件、公文、标书等各种内容。在写作之后还可以根据公司情况让大模型对写作的内容进行整理等。如果是训练大模型，那么在训练数据、算力资源、时间成本上对企业来说都是很大的挑战。使用知识图谱和知识库等技术结合大模型提示工程可实现大模型垂域

适配的调整，在保证效果的同时也保证了算力。

除了针对企业，不少大模型服务提供者也在为个人使用者提供定制化服务。OpenAI 已经开放了大模型的微调功能，用户可以上传自己的数据定制 GPT-3.5Turbo，每个人都能打造个人专属的 ChatGPT。据 OpenAI 透露，未来将推出微调 UI，实现面向大众用户的产品化功能。

1.3.2 跨模态交互

大模型现在不仅可以完成纯文本的对话，还可以结合多模态技术实现各个场景的应用。例如，结合 CV 技术的文生图和图生文模型，实现艺术创作和图像理解等。目前，通义千问、文心一言、GPT-4 等模型也可以在对话过程中提供纯文本的回答，还可以生成图片、文档，也可以针对输入的图片做解答，如图 1.4 所示。

图 1.4 大模型图像分析

大模型也可以实现视频理解功能，不仅是对单一图片进行理解，还可以结合前后帧捕捉连续画面内容对视频内容进行解析。现在，大模型不仅可以逐帧对视频做内容解析，还可以生成视频，比如 opensora 开源框架已经可以实现视频的生成。

数字客服结合数字人、语音等能力将数字生命变成真人，可以结合客服等为用户提供

解答等服务，如目前市场上的小冰数字人、商汤 SenseMars 等。

现在，随着生成式大模型的发展，大模型生成能力也有很大程度的提升。甚至很多图片人眼都无法辨别，这个时候 AI 可以提供判断，辅助人眼判断图像或者视频等是否为 AI 生成。

1.3.3　任务处理

任务型对话可以模拟人类，从而与人类形成连贯的对话，并可以理解用户的输入，比如文本、语音，甚至是图片，然后根据输入作出回答，如图 1.5 所示。完成的产品形态包括自然语言理解、对话管理、意图识别、自然语言生成几个模块。其中对话管理还包括了对话状态追踪和对话策略，实现根据上下文理解用户意图和指令，从而让回答更流畅自然。这可以有效地帮助用户完成一些特定任务，这主要是针对垂域业务，比如应用于办公场景下的会议预约、公文检索、查询通讯录等一系列操作。

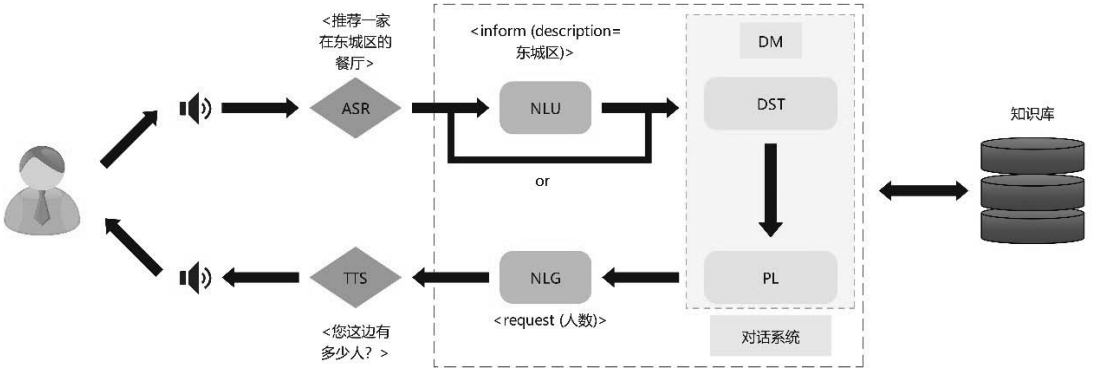

图 1.5　任务型对话

提升 AI 大模型的应用能力，也就是提升大模型对于任务的处理能力，这需要大模型具有更强的理解能力。大模型产品以生活助手、办公助理、趣味问答等功能作为切入点，为用户提供更聚焦、更精准的服务界面，从而提升用户输入的便捷性和输出结果的准确性。

1.3.4 决策智能化

很多人会把智能决策和智能预测两个概念混淆，其实两者之间还是有一定区别的。智能预测更多的是倾向于在海量的数据中找到规律，这个过程并不会对数据造成任何改变。但是智能决策则是在找到规律的基础之上将预测结果反馈到数据上，改变数据甚至是优化决策。大模型的诞生，可以通过其出色的理解能力、思维推理能力和多模态能力分析海量历史数据。

看起来决策智能化好像离我们很远，但其实日常生活中很多地方已经使用了大模型辅助决策理论。例如，在广告推荐场景下，可以通过分析用户的检索关键字、查看历史内容（包括文字、图片信息等向量），逐渐调整推荐的内容类型等；另外还有智能驾驶，其中典型代表是特斯拉基于 Transformer 构建的 BEV 感知方法；此外，在金融、政务等领域，大模型的辅助决策功能也是未来大模型应用的一个重要方向。

1.4 大模型产品分析

1.4.1 市场头部模型综述

随着深度神经网络技术的蓬勃发展，人工智能领域正式迈入了统计分类深度模型的新纪元。此类模型以其前所未有的泛化能力著称，能够灵活提取并应用多样化的特征值于广泛的实践场景中，显著弥补了传统模型的局限性。然而，2018—2019 年，双下降现象彻底重塑了人工智能的发展轨迹。传统数学理论曾预言，随着模型参数量的增加与模型复杂度的提升，过拟合现象会导致模型误差先减后增，促使研究者致力于寻找误差最低、精度最优的模型平衡点。

然而，随着人工智能算法与计算能力的飞跃式进步，研究者惊奇地发现，当模型规模持续无上限扩张时，其误差在初次攀升后竟能迎来第二次显著下降，且这一趋势随着模型

体量的进一步增大而持续强化，即模型规模与准确率之间呈现出正相关性，预示着一个“大模型时代”的到来。

在所谓的“百模大战”时代，模型的种类日益丰富，参数规模不断攀升，行业应用范围也在不断拓宽，整体上，AI 模型正在重塑产业格局。参数量已成为区分大模型与小模型的关键指标之一。百度集团副总裁侯震宇指出，拥有 10 亿参数的模型即可被视为大模型，而当前的大模型参数量动辄达到上千亿。根据科技部 2023 年的统计数据显示，中国发布的 10 亿参数规模以上的大模型数量已达到 79 个，而美国的数量则超过了 100 个。这表明，大模型的时代已经到来，它在推动人工智能技术发展和产业变革中发挥着越来越重要的作用。

1.4.2　发展概况

在探讨大语言模型发展的技术脉络时，我们聚焦于 3 大主流技术路径，即 BERT 范式、GPT 范式以及融合两者的混合模式。目前，大多数主流的大语言模型倾向于采用 GPT 技术路线，而中国则更倾向于采用融合创新的混合模式。自 2019 年以来，BERT 模式似乎没有出现具有里程碑意义的新模型，而 GPT 技术路线却呈现出蓬勃发展的态势，引领着模型规模与性能的双重飞跃。

各类大语言模型依据其技术路线各有千秋，GPT 范式尤其擅长处理生成类任务，展现出卓越的表现力。

进一步细分，大语言模型可依据数据处理与知识表示两个维度进行分类。在数据处理层面，模型利用的数据源可分为通用数据与领域特定数据；而在知识表示上，模型则涵盖了语言知识与世界知识的融合。从任务类型视角审视，模型又可分为专注于单一任务或多任务处理的模型，以及侧重于理解类或生成类任务的模型。

具体而言，BERT 范式遵循两阶段训练策略（即双向语言模型预训练与针对特定任务的微调），这一特点使其特别适用于理解类任务及特定场景下的精细化处理，呈现出“专精而轻量”的优势；而 GPT 范式则通过简化为单阶段训练（单向语言模型预训练结合零样本提示技术），提供了对生成类任务及多任务处理的强大支持，展现出“全面而通用”的特质。

此外，T5 模式作为 BERT 与 GPT 方法的巧妙融合，同样采用两阶段训练（单向语言模型预训练与微调）方式，为不同应用场景提供了灵活的选择。当前研究指出，对于规模极端庞大且聚焦于单一领域理解类任务的模型而言，T5 模式或为更佳选择；而在生成类任务领域，GPT 模式则以其卓越的性能脱颖而出。

综上所述，当前参数规模突破千亿量级的大语言模型几乎无一例外地采用了 GPT 范式，这一趋势不仅反映了 GPT 技术路线的强大生命力和广泛的应用前景，也预示着未来大语言模型技术发展的主流方向。

2022 年 11 月，美国 AI 企业 OpenAI 隆重推出了其旗舰产品——AI 聊天机器人程序 ChatGPT，该程序根植于先进的大语言模型 GPT-3.5 的深厚基础之上，并巧妙地融合了指令微调技术与基于人类反馈的强化学习策略进行精心训练。这一创新举措迅速在业界引起轰动，在发布后的短短两个月内，ChatGPT 的月活跃用户突破 1 亿，成为史上用户增长速度最快的消费级应用程序。

自 ChatGPT 发布后的半年多时间里，中国本土的科技企业及研究机构积极响应，推出了近 80 款参数量达到十亿量级以上的大语言模型。这些厂商中，既有华为、阿里巴巴、腾讯等互联网巨头，也有 360、科大讯飞等在人工智能领域拥有深厚技术积累的企业，以及复旦大学等高校和研究机构。这一系列成就，不仅彰显了我国在 AI 大模型领域的强劲实力与创新能力，也预示着未来 AI 技术将更加深刻地融入并改变我们的工作和生活。

1.4.3 模型分类

大模型的分类依据其输入类型的不同，可细化为 NLP（自然语言处理）大模型、CV（计算机视觉）大模型以及多模态大模型 3 大类。随着技术的发展，大模型所支持的模态数量日益丰富，从最初仅支持文本、图片等单一模态下的单一任务，逐步过渡到能够支持多种模态下执行多种复杂任务的阶段。

NLP 大模型作为处理自然语言文本数据的核心力量，尤其是以 LLM（大语言模型）为代表的模型，如 OpenAI 的 GPT 系列，展现了卓越的语言理解和生成能力。这类模型不仅

助力人类高效地完成问答、创作、文本处理等任务，还在快速发展的交互类场景中占据关键地位，其高度的商业化应用程度进一步印证了其市场价值和社会影响力。

CV 大模型专注于图像和视频数据的处理，凭借其强大的图像识别和视频分析能力，在智能安防、自动驾驶等众多领域展现出巨大潜力。例如，腾讯的 PCAM 大模型便是在这一领域内的杰出代表。尽管国内企业在此领域的研发与内部测试工作正深入进行，但总体而言，CV 大模型仍处于发展初期，未来增长空间广阔。

多模态大模型作为技术创新的前沿阵地，能够同时处理文本、图像、语音等多种类型的数据，实现跨模态搜索、跨模态生成等高级任务，如谷歌的 Vision Transformer 模型便是此类技术的典范。尽管多模态大模型目前尚处于雏形阶段，但其应用潜力已被广泛认可，然而，要实现更广泛、更深入的应用，还需解决一系列关键性的技术和挑战。

此外，根据大模型应用领域的不同，我们可以将其划分为 3 个层级，即 L0 层、L1 层和 L3 层。

- L0 层通用大模型：这一层级的大模型具备跨领域和多任务的通用性。它们依托于强大的计算能力和海量的开放数据资源，采用具有大量参数的深度学习算法，在大规模未标注数据上进行训练。通过这一过程，模型能够识别和学习数据中的模式和规律，从而获得强大的泛化能力。这种能力使得 L0 层模型在不经过或仅需少量微调的情况下，便能适应多种场景和任务，类似于 AI 完成了全面的“通识教育”。
- L1 层行业大模型：这一层级的大模型专注于特定行业或领域。它们通常使用与特定行业相关的数据进行预训练或微调，以优化模型在该领域的性能和准确性。这种专业化的训练使 L1 层模型能够深入理解特定行业的知识和需求，从而在该领域内提供更为精准和高效的服务，相当于 AI 成为该领域的“行业专家”。
- L3 层垂直大模型：这一层级的大模型针对特定的任务或场景进行优化。它们使用与特定任务紧密相关的数据进行预训练或微调，以提升模型在该任务上的表现和效果。L3 层模型的专业化训练使其能够针对特定场景提供定制化的解决方案，从而在特定任务上实现更高的性能和效率。

通过这种分层的模型设计，大语言模型能够更好地满足不同领域和任务的需求，实现从通用到专业领域的全方位覆盖，推动人工智能技术在各个领域的深入应用和发展。

1.4.4 市场情况

在全球范围内，中美两国在大型模型（large-scale models）领域展现出了引领行业发展的强劲势头。美国在算法模型研发领域凭借其深厚的积累与创新优势，稳居全球大模型数量的榜首位置。据中国科学技术信息研究所与科技部新一代人工智能发展研究中心联合编纂的《中国人工智能大模型地图研究报告》最新数据显示，截至 2023 年 5 月，美国已成功推出超过 100 个参数规模达到 10 亿级以上的大型模型，彰显了其在该领域的卓越实力。

与此同时，中国也不甘落后，积极响应全球大模型技术浪潮，自 2021 年起显著加快了研发步伐与成果产出。标志性事件包括 2021 年 6 月北京智源人工智能研究院发布的悟道 2.0 模型，其参数量高达 1.75 万亿，以及同年 11 月阿里巴巴推出的 M6 大模型，其参数量更是惊人地达到了 10 万亿级别，这些成就标志着中国在大型模型研发领域取得了重大突破。截至 2023 年 5 月，中国已累计发布 79 个大型模型，不仅在全球舞台上占据了举足轻重的地位，还展现出了中国在该领域的先发竞争优势。

然而，值得注意的是，鉴于数据安全、隐私保护合规性以及日益严格的科技监管环境等多重因素的考量，中美两国的大模型市场或将逐步演化出相对独立且各具特色的行业格局。这种趋势要求两国在推动技术创新与产业应用的同时，必须高度重视数据安全与隐私保护，确保技术发展与社会伦理、法律法规的和谐共生。

在海外大语言模型的竞争格局中，已经形成了一个明显的双龙头领先态势，同时伴随着 Meta 的开源策略和垂直领域的快速发展。这种格局不仅清晰，而且充满活力。随着通用大模型技术的日益成熟和可用性，基于这些模型的应用生态系统也开始蓬勃发展。

OpenAI 凭借其对尖端算法模型的深度整合及前瞻性的产品化策略，不仅使 GPT 在人机交互领域展现出超乎预期的卓越性能，更引领了基于 GPT 技术构建的丰富应用生态的崛起。微软公司旗下多款核心产品，包括但不限于 Bing 搜索引擎、Windows 操作系统、Office 办公软件套件、Edge 浏览器，以及 Power Platform 业务解决方案平台，均已成功融入 GPT 技术，显著提升了用户体验与智能化水平。此外，代码托管领域的领军者 GitHub，以及 AI 驱动的营销创意先锋 Jasper 等公司，也纷纷拥抱 GPT，共同推动了 AI 技术在各行业的深度渗透与创新应用。这一系列进展，不仅彰显了 GPT 技术的强大生命力，也为未来智能应用

的无限可能奠定了坚实基础。

谷歌公司在人工智能领域的持续投入和创新，已经对全球人工智能产业产生了深远的影响。谷歌公司提出的 LeNet 卷积神经网络模型、Transformer 语言架构以及 BERT 大语言模型等，均为人工智能技术的发展做出了重要贡献。尽管如此，由于公司内部团队的变动和对产品化落地采取更为审慎的态度，谷歌公司在早期并未大规模推出面向消费者端（C 端）的人工智能产品。

然而，在 AI 聊天机器人 ChatGPT 迅速流行的推动下，谷歌公司也加快了步伐，推出了自己的聊天机器人 Bard 及 PaLM2。这些产品不仅展示了谷歌公司在人工智能领域的最新进展，还计划与谷歌公司的协作与生产力工具 Workspace 进行集成，并与 Spotify、沃尔玛、UberEats 等外部应用进行融合，进一步拓展其在人工智能领域的应用范围。

与此同时，Meta 公司通过开源策略在人工智能领域迅速追赶。2023 年 7 月，Meta 公司发布了其最新的开源大模型 LLaMA2。该模型使用了 2 万亿 tokens 进行训练，显著提升了其上下文理解能力，使其能够处理更长的文本输入。这种提升不仅增强了模型的表现能力，还为更广泛的应用场景提供了可能。Meta 公司的这一举措进一步证明了开源策略在推动人工智能技术发展和应用中的重要性和潜力。

此外，在海外市场的 AI 版图中，Anthropic、Cohere、Hugging Face 等领先企业凭借其独特的垂直领域专长与高度定制化的服务方案，正扮演着不可或缺的关键角色。这些企业不仅深耕于各自的专业领域，展现出鲜明的垂直类特色，还通过提供灵活多变的定制化服务，积极满足市场多元化、精细化的需求，共同推动着全球 AI 产业的繁荣与发展。

自 ChatGPT 在全球范围内赢得广泛用户赞誉并引发深切关注以来，中国顶尖科技企业（如阿里巴巴、百度、腾讯、华为、字节跳动等）、新兴的创新型公司（如百川智能、MiniMax 等）、深耕 AI 领域的传统企业（科大讯飞、商汤科技等），以及顶尖高校与研究院所（复旦大学、中国科学院等）均纷纷加速布局大模型技术的研发与投资。当前，国内大模型领域正处于研发与迭代的初期关键阶段，各模型间的性能差异及用户体验的易用性正接受着市场的严格考验与筛选。尽管竞争格局的全面明朗尚需时日，但值得注意的是，互联网巨头凭借其在 AI 领域的长期深耕与积累，已占据先发优势，有望在未来竞争中占据主导地位。

ChatGPT 3.5 的问世，象征着我们正不可逆转地迈向一个由人工智能技术主导的未来。

尽管这一宏观趋势日益明显，但其在微观层面的日常生活中的应用和普及，仍是一条充满挑战的长路。2024 年 2 月 15 日，OpenAI 推出了革命性的“文本生成视频”工具 Sora，其展示的先进功能再次令世界瞩目。在美国，人工智能领域的迅猛发展也吸引了华尔街的广泛关注和投资，仅在 2023 年，美国 AI 相关企业便成功筹集了高达 679 亿美元的资金。作为 OpenAI 的重要投资者和关键技术合作伙伴，微软公司在 2023 年的股价表现尤为抢眼，全年涨幅达到了 55%。

相较于美国企业在人工智能领域的频繁曝光与舆论热议，中国企业在该领域的进展虽鲜少占据新闻头条，但实则暗流涌动，竞争态势已悄然形成。这一现象背后，实则源于两国 AI 技术发展路径的差异：美国多由金融资本引领，追求高估值与技术前沿的突破，其光芒四射往往能迅速吸引公众眼球；而中国则更多地遵循消费市场导向，侧重于技术的快速市场化应用与迭代优化，虽不显张扬，却能在每一步发展中收获市场的实际反馈，有效规避了泡沫风险，展现出稳健前行的姿态。

进一步剖析，大语言模型作为 AI 技术的一个重要分支，其能力犹如一位博览群书的学者，擅长处理已知信息或网络上的既有内容，但在创造力与实践应用方面尚显不足。在工业领域，我们更为迫切地需要 AI 扮演执行者的角色，不仅能够理解指令，更能高效、精准地完成各项任务，这对于推动产业升级与智能化转型具有不可估量的价值。因此，在探索 AI 技术发展的道路上，如何平衡技术创新与市场应用，让 AI 真正“行动起来”，成为我们共同面临的课题。

1.4.5 国内大模型介绍

大模型凭借其卓越的语言理解与生成能力，在众多领域（如文本创作、智能问答、知识检索及商业文案生成）展现出了非凡的潜力与广阔的应用前景。然而，就商业应用层面而言，大模型目前仍处于萌芽阶段，主要通过 API、PaaS 及 MaaS 三种模式逐步推进其市场渗透。全球范围内，大模型产业的实际落地正处于积极探索期，亟须与下游行业企业携手，共同构建可持续的商业模式。在此过程中，一个不容忽视的事实是，下游企业对大模

型的理解尚显不足，且缺乏足够的资源支撑以实现有效的融合与应用。

针对这一现状，大模型的商业化落地路径可细化为多种策略：一是通过提供 API 接口，实现按需调用的付费模式，灵活满足多样化需求；二是大厂可采用 PaaS 模式，不仅提供大模型本身，还涵盖开发工具、云平台及全方位服务，助力下游企业轻松接入与部署；更进一步，则是 MaaS 模式，即直接提供预训练好的、高度定制化的模型服务，以更低的门槛和更高的效率，推动大模型在各行各业中的深度应用与价值释放。

下面以 NLP 大模型为核心，简单介绍几个国内热门大模型。

1. DeepSeek——深度求索

DeepSeek 大模型是由中国人工智能公司深度求索自主研发的新一代通用人工智能大模型，其研发团队由 NLP、多模态、分布式计算等领域的顶尖科学家组成。该模型基于 Transformer 架构创新，在算法设计、训练效率和工程实践上均实现突破，致力于推动 AGI 技术的普惠化发展。

DeepSeek 采用混合专家模型（mixture of expert，MoE）架构，创新性地提出动态路由优化算法，使模型在保持 1750 亿个参数规模的同时，推理成本较传统稠密模型降低 80%。其独特的层级注意力机制（hierarchical attention）实现了对长文本（最高支持 128K tokens）的精准语义理解，在代码生成、法律文书分析等场景中表现出色。训练层面首创的“渐进式知识蒸馏”方案，使模型在预训练阶段可同步吸收领域知识，较传统的两阶段训练效率提升 3 倍。

区别于单一文本模型，DeepSeek-7B 版本已实现图文跨模态理解，在 OCR 信息提取、流程图解析等任务中准确率达 92.7%（据权威测评 MMBench 测评）。其视觉-语言对齐模块采用对比学习优化策略，显著提升细粒度语义关联能力，在医疗影像报告生成等专业场景的幻觉率低于 2%。

在金融领域，模型通过微调构建的 DeepSeek-Finance 版本，具备财报自动分析、风险预警等能力，在上海证券交易所实测中的事件推理准确率达 89.3%。在教育场景中，其数学推导模块整合符号计算引擎，可分步解析大学数学竞赛题，解题流程的可解释性超过 GPT-4。企业级用户可通过 API 调用获得支持千亿级参数的私有化部署方案，推理延迟稳定在

200ms 以内。

作为国内首个完整开源 MoE 架构的团队，DeepSeek 已发布 7B/67B 参数模型（Apache 2.0 协议），配套提供 200 万字高质量的多轮对话数据集。其创新设计的“可插拔伦理模块”支持用户自定义内容审查规则，在中文内容安全评测 C-Eval 中合规率达 99.6%。开发者社区已涌现 300+个基于该模型的垂直领域应用，涵盖智能编程助手、工业质检知识库等创新方向。

当前 DeepSeek 在中文权威测评 SuperCLUE 中综合得分位列国产模型第一，其英文 MMLU 测评准确率突破 85%，展现出强大的通用能力。随着其“算力-算法-数据”三位一体技术体系的持续迭代，该模型正在重塑企业智能化转型的技术范式。

2. 百度公司——文心一言

百度公司于 2023 年 3 月 16 日正式推出了知识增强型大语言模型——文心一言，该模型根植于深厚的深度学习平台与文心知识增强大模型基础之上，历经海量数据与广泛知识的深度融合与持续学习而铸就。文心一言展现出了卓越的跨模态、跨语言深度语义理解与生成能力，能够流畅地与人类进行对话互动，精准回答各类问题，并有效辅助创作过程，极大地提升了用户获取信息、汲取知识与激发灵感的效率与便捷性。

目前，百度公司已推出文心大模型的多个版本，包括 3.5、4.0 及性能更为强劲的 4.0 Turbo，旨在通过这些模型在搜索、信息流推荐、广告投放、智能写作及对话系统等多个关键场景中实现智能化转型，为用户带来更加精准且个性化的服务体验。

文心一言不仅擅长自然语言处理、文本生成及语音合成等核心任务，还创新性地将大数据预训练与多元化知识源深度融合，借助持续学习机制，不断从海量文本数据中汲取词汇、结构、语义等新知识，推动模型效果持续优化与进化。此外，基于文心大模型的强大基座，百度公司还提供了包括对话 PLATO-XL、搜索 ERNIE-Search、跨语言 ERNIE-M、代码 ERNIE-Code 在内的多个独立应用场景大模型，以及视觉模型、跨模态模型、生物计算模型等多元化选择方案，以满足不同领域与场景下的特定需求。

为了让用户能够更加便捷地体验与应用文心一言，百度公司构建了一整套完善的配套产品体系，包括官网体验入口、APP、文心智能体平台（AgentBuilder）以及千帆大模型平

台（ModelBuilder），并持续拓展产品链条，以提供更全面的服务。特别是在 2024 年 3 月 27 日举办的“AI Cloud Day：大模型应用产品发布会”上，百度智能云在北京首钢园隆重宣布，将大模型能力全面融入 7 大产品之中，如百度智能云曦灵数字人平台、客悦智能客服平台及 Comate 代码助手等，覆盖了企业营销、客户服务、知识管理、数据洞察及代码编程等多个通用场景。这些产品不仅支持公有云与私有云两种灵活的使用方式，还为企业量身打造了“应用产品全家桶”，旨在全方位助力企业业务增长与运营效率的提升。

个人与企业客户均可通过百度智能云的千帆大模型平台轻松接入并享受这些先进的大模型应用服务。

3. 智谱 AI——GLM

追溯 GLM 的发展轨迹，其发布历程彰显出前瞻性的布局。智谱 AI 自 2020 年底便投身于 GLM 预训练架构的研发，并于次年 9 月成功推出了拥有自主知识产权的开源百亿参数大模型 GLM-10B。在 2022 年 8 月，智谱 AI 再接再厉，发布了性能卓越的高精度迁移大模型 GLM-130B，其效果直逼 GPT-3 175B，引发了全球超过 70 个国家、1000 余家研究机构的广泛关注与应用需求。同年内，智谱 AI 还发布了编程大模型 CodeGeex 的两个重要版本，进一步丰富了其产品线。此外，2022 年推出的 P-tuningV2，经过时间的沉淀与迭代，已成为业界广泛采用的开源微调框架之一。

2023 年，智谱 AI 继续引领创新，3 月份开源了中英文双语对话模型 ChatGLM-6B，该模型以其出色的性能，在单张消费级显卡上即可实现高效推理，尽管参数规模有限，但其实际效果却令人印象深刻，成为众多大模型开发者部署于本地的中文大模型首选方案。随后，在 2024 年 1 月，智谱 AI 又推出了新一代基座大模型 GLM-4，标志着其在大模型研发领域的又一重大突破。

GLM 大模型作为智谱 AI 的杰出代表作，凭借其卓越的中文处理能力、开源的友好特性以及对多种硬件平台的广泛支持，成为业内外关注的焦点。其基于 Transformer 架构的自主研发核心，深刻体现了智谱 AI 在算法创新领域的深厚底蕴与前瞻视野。不仅如此，GLM 大模型还展现出跨模态处理的潜力，能够解析图像等复杂数据类型，极大地拓宽了 AI 技术的应用边界。而与之配套的产品矩阵，如 ALL Tools、CogVLM3 及 CodeGeeX3 等，共同构

建了一个全面而强大的大模型生态系统。

智谱 AI 的产品线丰富多样，除了上述提及的 ChatGLM、GLM-4、GLM-130B 等大模型外，还包括代码大模型 CodeGeeX、文生图模型 CogView 等，满足了不同领域与场景下的多样化需求。基于这些强大的大模型能力，智谱 AI 已成功应用于智能驾驶、智能投顾、财报分析、公积金咨询、智能医疗、旅行规划等多个行业领域，为用户与企业提供了高效、便捷的智能化解决方案。

特别值得一提的是，2023 年 8 月，智谱 AI 的生成式 AI 助手“智谱清言”正式上线。该产品基于智谱 AI 的基座大模型深度开发，通过海量文本与代码数据的预训练，结合先进的监督微调技术，实现了通用问答、多轮对话、创意写作、代码生成、虚拟对话、AI 绘图、文档与图片解读等多项功能，展现了强大的智能交互能力。为了更好地体验 GLM 系列大模型的魅力，智谱 AI 还推出了 bigmodel.cn 大模型开放平台，该平台已全面接入最新的 GLM 大模型家族，并新增了一键微调、All Tools API 调用等便捷功能，为用户提供了更加高效、灵活的大模型使用体验。

4. 阿里云——通义千问

阿里云于 2023 年 4 月 7 日宣布，其自主研发的大语言模型“通义千问”正式开启用户测试体验。事实上，阿里巴巴的达摩院在此领域已深耕多年，早在 2019 年便启动了大模型的研发工作，并在 2022 年 9 月推出了“通义”大模型系列。该系列模型具备多轮对话、文案创作、逻辑推理、多模态理解及多语言支持等功能，能够与人类进行深入的交互，并展现出卓越的文案创作能力，如续写小说、编写邮件等。

到了 2023 年 8 月，通义千问加入开源行列，首次开源了 Qwen-7B 模型，并沿着“全模态、全尺寸”的开源路线，陆续推出了包括语言大模型、多模态大模型、混合专家模型、代码大模型在内的数十款模型。2024 年 5 月 9 日，阿里云正式发布了性能全面超越 GPT-4 Turbo 的通义千问 2.5，成为业界公认的最强中文大模型。同时，通义千问的 1100 亿参数开源模型在多个基准测评中取得了最佳成绩，超越了 Llama 3-70B，成为开源领域中的佼佼者。

为满足不同场景下用户的需求，通义千问推出了参数规模从 5 亿到 1100 亿不等的 8 款

大语言模型。小尺寸模型如 0.5B、1.8B、4B、7B、14B 等，便于在手机、PC 等端侧设备上部署；而大尺寸模型如 72B、110B 则支持企业级和科研级的应用需求；中等尺寸的 32B 模型则在性能、效率和内存占用之间寻求最佳平衡点。此外，通义千问还开源了视觉理解模型 Qwen-VL、音频理解模型 Qwen-Audio、代码模型 CodeQwen1.5-7B 和混合专家模型 Qwen1.5-MoE，后者结合了多种神经网络结构，以提升处理复杂任务时的性能。

除了语言模型，通义千问还包括了用于处理视觉和音频数据的 Qwen-VL 和 Qwen-Audio 模型，进一步扩展了其在多模态应用中的能力。阿里云的通义千问大模型已广泛应用于营销、客服、编码等多个场景，并与手机、计算机、芯片、座舱等智能终端深度融合。

在应用产品方面，通义 APP（原名通义千问 APP）集成了通义大模型的全栈能力，在移动端、Web 端、小程序端为所有用户提供免费服务。该应用集文生图、智能编码、文档解析、音视频理解、视觉生成等多种能力于一体，旨在成为每个人的全能 AI 助手。2023 年 10 月，阿里云推出了百炼大模型平台，使开发者能够通过“拖”“拉”“拽”等交互形式，在 5 分钟内开发一款大模型应用，几小时内“炼”出一个专属模型，从而将精力更专注于应用创新。

5. 科大讯飞股份有限公司——星火认知大模型

2023 年 5 月 6 日，科大讯飞股份有限公司（以下简称科大讯飞）正式推出了其创新的讯飞星火认知大模型，并持续进行迭代更新。至 2024 年 6 月 27 日，该模型已成功推出 V4.0 版本。讯飞星火认知大模型集成了科大讯飞的先进技术，具备 7 大核心能力：文本生成、语言理解、知识问答、逻辑推理、数学处理、编程能力以及多模态交互。

星火大模型的 API 矩阵提供了丰富的接口选项，为企业构建定制化的大模型产品提供了强有力的支持。该矩阵包括 Spark 4.0 Ultra、Spark 3.5 Max、Spark Pro 和 Spark Lite 等多种类型的接口，其中 Lite 版本已全面免费开放，以促进更广泛的应用和创新。

在世界移动通信大会（MWC）上，科大讯飞展出了 5 大创新产品：星火智能陪练、星火评标助手、星火合同助手、星火生成式智慧驾驶舱以及 iCase 会话智能。这些产品将讯飞星火 V4.0 模型与企业业务需求紧密结合，展现了从基础模型升级到应用落地的连贯发展策略。

基于讯飞星火 V4.0 模型的强大能力，科大讯飞推出了一系列更智能、更个性化的 AI 助手。首批上线的 14 个智能体，针对不同专业领域的应用进行了特别优化。升级版的讯飞晓医 APP 和讯飞 AI 学习机，以及星火智能批阅机——能在 30 秒内高效完成 15 份学生作业的批改，都是该模型实用性的体现。星火语音大模型也实现了重大升级，支持 74 种语言和方言的无缝对话，并有效解决了高干扰环境下的语音识别问题。

此外，科大讯飞还推出了星火企业智能体平台，并展示了星火商机助手、星火评标助手等典型应用案例，进一步拓展了大模型的应用范围和效能。

讯飞星火大模型还提供了包括官网体验入口、移动应用、智能体平台、API 接入以及开源模型接入和微调服务在内的全面接入方案，使用户和开发者能够轻松利用这些先进的技术资源。

6. 腾讯科技（深圳）有限公司——混元大模型

2023 年 9 月 7 日，腾讯科技（深圳）有限公司（以下简称腾讯）在全球数字生态大会上隆重推出了其自研的腾讯混元大模型，并通过腾讯云平台向外界开放。作为腾讯全链路研发的通用大语言模型，混元大模型拥有超过千亿级别的参数规模，其预训练语料库涵盖了超过 2 万亿的 tokens，展现了卓越的中文理解、创作能力、逻辑推理能力以及稳健的任务执行能力。该模型体系包括混元生文与混元多模态两大核心模块，其中，混元生文已迭代推出 6 个版本，涵盖 hunyuan-pro、hunyuan-standard、hunyuan-role 等，以满足不同场景下的需求。

腾讯混元大模型已成功融入公司内部超过 400 项业务与场景之中，并通过腾讯云，向广大企业及个人开发者全面开放。众多广为人知的“国民级”应用，如企业微信、腾讯文档、腾讯会议等，均已深度融合 AI 技术，实现了功能的智能化升级。此外，腾讯云旗下的 SaaS 产品，如企业知识学习平台腾讯乐享、电子合同管理工具腾讯电子签等，也同样受益于 AI 的赋能，使得用户体验得到了显著提升。

腾讯混元大模型秉持实用主义原则，深度应用于公司多个业务与产品的实际场景中。以腾讯会议为例，基于混元大模型打造的 AI 小助手，仅需用户发出简单的自然语言指令，即可高效完成会议信息提取、内容分析等复杂任务，并在会议结束后自动生成智能总结纪

要。在文档处理领域，混元大模型支持数十种文本创作场景，已在腾讯文档的智能助手功能中得到实际应用。同时，该模型还具备一键生成标准格式文本、精通数百种 Excel 公式、支持自然语言生成函数以及基于表格内容自动生成图表等能力。在广告业务场景中，腾讯混元大模型能够智能化创作广告素材，精准匹配行业与地域特色，满足个性化需求，实现文字、图片、视频等元素的自然融合。

基于混元大模型的强大能力，腾讯推出了面向 C 端用户的元宝 APP，旨在通过更便捷的操作体验，为用户带来全新的智能服务。此外，腾讯还构建了元器平台，该平台提供一站式智能体创作与分发服务，用户可轻松创建个性化的智能体，并充分利用平台提供的丰富插件与知识库资源，实现智能应用的快速开发与部署。

7. 月之暗面——Kimi

自 2023 年 10 月正式面世以来，Kimi 凭借其卓越的长文本处理能力在众多模型中脱颖而出，成为业界瞩目的焦点。2024 年 3 月，Kimi 再次迎来重大升级，其上下文窗口实现了 10 倍扩容，现可支持高达 200 万字的超长无损输入，这一里程碑式的成就，不仅刷新了当前全球最长上下文窗口的记录，也彰显了其技术实力的深厚底蕴。

Kimi 源自成立仅一年的前沿人工智能初创企业——月之暗面，该企业凭借其深厚的研发实力和前瞻性的技术视野，成功打造出这款集数据处理、分析于一体的智能工具。Kimi 能够迅速捕捉并理解用户的问题，依托其广泛的知识库，覆盖科技、文化、历史、教育等多个领域，精准满足用户多样化的信息需求。同时，它还具备高效的文件处理能力以及互联网访问能力，为用户带来更加便捷、全面的智能体验。

Kimi 的核心功能涵盖 6 大方面，即长文总结与生成、联网搜索、数据处理、代码编写、用户交互以及翻译服务。这些功能相互融合，共同构建了一个强大而灵活的智能助手体系。在专业领域，Kimi 已展现出其非凡的应用潜力，如助力专业学术论文的翻译与深度理解、辅助法律问题的精准分析以及加速 API 开发文档的快速掌握等，均为用户带来了前所未有的工作效率提升。尤为值得一提的是，Kimi 还是全球范围内首个支持一次性输入 20 万汉字的智能助手产品，这一特性极大地提升了其在处理复杂文本和多语言内容时的效率和准

确性。

通过这些先进的功能和广泛的应用场景，Kimi 不仅为用户提供了强大的辅助工具，也为人工智能领域的技术进步和应用创新树立了新的标杆。

8. 字节跳动公司——豆包大模型

2024 年 5 月 15 日，字节跳动公司在火山引擎原动力大会上正式推出了豆包大模型。该模型经过字节跳动公司内部超过 50 个业务场景的深入实践与验证，以其高性价比及每日支撑千亿级 tokens 处理能力为显著特点，被精准定位为多功能的 AI 助手，旨在全方位助力用户在生活、学习、工作等多个领域的需求。

豆包大模型构建了一个多模态的模型矩阵，目前涵盖通用模型 Pro、通用模型 Lite、语音识别模型、语音合成模型、文生图模型等在内的 9 款模型，每一款模型均针对不同应用场景进行了深度优化。其应用场景极为广泛，包括但不限于智能客服系统的优化、内容创作的辅助、智能教育平台的赋能、个性化智能助手的实现，以及智慧娱乐体验的提升等。此外，豆包大模型的多模态特性还使其能够灵活适配互联网、零售消费、金融、汽车、教育、科研等多个行业领域，展现出强大的跨界应用潜力。

基于豆包大模型，字节跳动公司成功打造了多款创新应用，如 AI 对话助手“豆包”、面向开发者的 AI 应用开发平台“扣子”，以及互动娱乐应用“猫箱”等，同时还推出了星绘、即梦等专业的 AI 创作工具。这些应用与工具不仅丰富了字节跳动公司的产品生态，还通过将大模型深度融入抖音、番茄小说、飞书、巨量引擎等 50 余个核心业务中，显著提升了运营效率并优化了用户体验。

值得注意的是，“扣子”专业版现已无缝集成于火山引擎的大模型服务平台“火山方舟”之中，为企业用户提供企业级 SLA 保障及一系列高级功能特性。火山引擎作为字节跳动公司旗下的云服务平台，其总裁谭待详细介绍了豆包大模型的全貌，包括豆包通用模型 Pro、豆包通用模型 Lite、豆包 • 角色扮演模型、豆包 • 语音合成模型、豆包 • 声音复刻模型、豆包 • 语音识别模型、豆包 • 文生图模型以及豆包 • FunctionCall 模型等，展现了其全面而深入的 AI 技术能力。

1.5　大模型产品设计

大模型作为新时代生产力的杰出代表，正以前所未有的力量引领着传统产业的深刻转型与新兴产业的蓬勃兴起，为社会经济的高质量发展源源不断地注入强劲新动能。曾经被视为人类专属领域的诸多任务与挑战，正逐渐被大模型的浪潮所席卷，并展现出前所未有的智能化潜力。

自 2023 年初以来，以 AIGC 为催化剂的年度科技盛宴持续升温，高潮不断。从 ChatGPT 横空出世，一举引爆全球关于人工智能通用化应用的热烈讨论，到大模型技术迅速普及，形成群雄逐鹿、百舸争流的壮观景象，这一进程不仅见证了技术的飞跃，也预示着一个新时代的到来。

如果说各大科技企业竞相推出大模型产品，竞相展示技术实力，共同编织了这场“百模大战”的壮阔的上半场，那么随着技术的日益成熟与应用的不断深化，大模型发展的下半场将聚焦于更为精细化的垂直领域应用与价值转化的深度挖掘。在这一阶段，大模型将不再仅仅停留于技术展示层面，而是将深入各行各业，通过定制化解决方案，实现技术与产业的深度融合，推动社会生产力的全面升级。

具体而言，大模型将凭借其强大的数据处理、模式识别与智能决策能力，在医疗健康、智能制造、智慧城市、金融服务等多个关键领域发挥不可替代的作用，助力解决行业痛点，提升服务效率与质量，创造更加丰富的社会价值与经济价值。这一过程，不仅将深刻改变我们的生产生活方式，也将为全球经济社会的可持续发展注入新的活力与希望。

1.5.1　大模型应用场景概述

作为机器与用户交互的桥梁，智能对话系统广泛应用于客服与助手领域。相较于传统基于问答对或固定规则的对话模式，大模型以其卓越的上下文理解能力和自然流畅的回答，

为用户带来了接近真人的对话体验。针对垂直领域的知识需求，通过知识库的灵活挂载，系统能够辅助学习并深化理解。在对话等待期间，系统还能进行轻松闲聊，增强用户交互的愉悦感。这一场景广泛适用于智能客服、语音助手、智能陪练、虚拟社交、售前咨询、售后运维及游戏 NPC 等多种场合。

在智能写作领域，大模型通过接收特定提示词，能够自动生成多样化的文本内容，涵盖电子邮件、短信、文章、新闻报道、营销文案及社交媒体文案等。相较于传统基于规则和模板的拼接方式，大模型展现了更高的灵活性和创新性，为用户提供了更加丰富多变的创作体验。作为大模型率先成熟的商业模式之一，智能写作已广泛应用于广告文案创作、会议纪要整理、直播脚本生成等领域，并能结合大纲制订、内容润色及多模态能力，进一步拓展至 PPT 文档等复杂内容的自动生成。

信息识别与抽取致力于从长文本中精准抽取并总结关键信息，以结构化形式输出。相比传统 NLP 技术，大模型在泛化能力上实现了显著提升，支持零样本学习，显著降低了开发成本，加速了应用落地进程。信息抽取技术广泛应用于用户需求分析、意图识别、文章辅助阅读、舆情监测、用户画像构建等多个场景，为企业决策与个性化服务提供了有力支持。

大模型在代码生成领域展现出巨大潜力，不仅提高了开发效率，还减轻了程序员编写代码的负担。通过智能分析现有代码，大模型能提出重构或优化建议，并根据用户描述自动生成脚本、数据库查询语句等，助力自动化流程的构建与应用程序性能的监控。这一技术可应用于代码审查、测试用例生成、智能 RPA（机器人流程自动化）、AI 建站、NL2SQL（自然语言到 SQL 查询语言）等多个场景，极大地推动了软件开发的智能化进程。

传统信息检索系统主要依赖文字或语义匹配，返回结果多为原文片段或结构化卡片，难以满足复杂查询需求。而大模型则通过深度理解文章内容，结合用户查询意图生成针对性回答，为用户提供了全新的智能信息检索体验。该技术广泛应用于知识搜索、文档检索、视频搜索、商品搜索、简历筛选等多个领域，极大地提升了信息获取的便捷性和准确性。

大模型的理解能力还为其在作文评分、文本润色、缩写扩写等领域的应用提供了可能。此外，大模型还能解决数学问题、批改作业、审查合同等，展现出广泛的应用前景和巨大的潜力。随着技术的不断进步和应用场景的持续拓展，大模型将在更多领域发挥重要作用。

1.5.2　大模型的部署及落地

大模型行业应用的实现路径主要有两种，一种实现路径是加大对用大模型的研发投入，提升 AI Agent 能力，直接服务各行业；另一种实现路径是融合行业专业知识，基于通用大模型打造垂类行业模型。站在通用大模型的“巨人的肩膀”上，垂类行业大模型通过知识录入或二次训练，可实现更精准的行业应用。

1. 模型选择

当前市场上，大模型厂商众多，每家厂商都可提供不同规模和特性的大模型。值得注意的是，模型规模并非其效能的决定性指标，通过精确的场景适配和细致的工程优化，即使是参数量较为适中的模型也能展现出卓越的性能。此外，在特定应用场景中，单一模型往往不足以满足所有需求；通过整合不同模型的优势，实现它们的协同互补，可以显著提升工作效能，创造出更佳的应用成果。

在选择合适的模型时，通常需要综合考量多种因素，主要可从以下两个方面进行评估。

首先，需要明确自身对数据安全的重视程度，判断是否需要模型服务的私有化部署，或是简单的 API 调用即可满足需求。同时，考察厂商是否提供场景模型微调服务，以及其在目标应用场景中的实践经验与模型评测效果是否达标。

在选定厂商范围后，还需要对自身条件进行细致评估，包括算力成本、资源配备及预算限制。在确保资源充足的前提下，进一步分析模型性能（响应时间、对话效率等）、功能范围（如模型推理能力、Agent 集成等）以及部署环境的特定要求（如私有化部署、本地部署、信创兼容性等），以确认最适合自身需求的模型规格。

2. 方案选择

运用大模型的过程中，涉及多个关键环节需要细致考量，如 Prompt 设计的有效性、是否实施微调、训练数据的标注策略、训练轮次的设定以及参数的调整等。

为实现高质量的 Prompt 编写，首要任务是明确需求。一个优秀的 Prompt 应能精确界定产品需求，我们可以通过绘制功能流程图，清晰界定大模型调用的输入与输出界面。对于知识问答类功能，还需要紧密结合产品定位，预先准备大模型推理所必需的业务知识库。

此外，鉴于不同大模型对指令内容的敏感度各异，需要依据模型特性定制化设计 Prompt，以实现最佳适配。

完成 Prompt 设计后，构建评测集成为关键步骤。对此，建议采用人工构建方式，结合业务目标精心挑选覆盖各难度级别的场景数据，作为 Prompt 调优的基准。评测数据的分布最好能够尽量贴近实际业务场景，确保类型全面且评测标准逻辑一致。值得注意的是，评测集的大小并非决定性因素，关键在于其代表性和针对性。

并非所有场景均需要进行模型微调，对于依赖通识能力的场景，往往无须微调即可满足需求。然而，在垂直行业、高专业度要求及知识密集型场景中，微调则成为提升效果的关键。微调的核心在于对齐标准、行业术语等，而更广泛、精确的行业知识可通过知识库关联的方式有效注入。

微调的方法多样，具体技术细节将在后续部分进行详细介绍。在微调完成后，应采用大规模测评数据集进行全面评估，依据交付标准细致分析错误案例（bad case），分类归纳其类型与成因，以指导后续微调方向。针对错误案例，可构建专门的微调数据集，并按一定比例混入线上其他类型数据，以防止微调过程中的梯度爆炸及知识遗忘问题。完成新一轮微调后，应重新构建评测集，对比首次与二次评测结果，评估微调成效及新一轮错误案例情况，通过多轮迭代直至达到上线标准。

1.5.3 大模型入门产品设计

大模型技术的潜在应用场景极为广泛，它们能够适应多种不同的业务需求。本节内容将从一些普遍适用的领域出发，简要介绍几种易于实施且设计成熟的大模型应用实例。

通过精心设计，这些应用不仅能够充分利用大模型的强大能力，还能够为用户提供实际价值和解决方案。我们将探讨这些应用如何通过创新的方式，将大模型技术融入实际工作流程之中，以实现效率提升和体验优化。

大模型对话类产品通常作为用户探索与体验模型性能的窗口，多见于面向消费者的（toC）应用中，产品更加侧重于通用模型或功能的实现。鉴于其面向广大公众的特性，这

类产品对资源的高效利用及合规性管理提出了较高的要求。

当前，主流的大型模型服务提供商普遍配备了此类产品，它们不仅覆盖了 PC 端网页，还延伸至移动 APP，共同构成了用户认知与体验模型能力的综合门户。这些产品的结构设计虽各有特色，但核心功能板块大致相似，主要包括两大方面：一方面是通用大模型的问答功能，旨在满足用户多样化的信息查询需求；另一方面是智能体功能，通过模拟人类交互方式，提供更加个性化、智能化的服务体验。

大模型问答功能模块设计如下。

- 欢迎语设计：该元素位于对话框顶部，旨在用户首次进入或长时间未登录后重新访问时触发。其内容通常包含亲切的问候语及自我介绍，旨在营造友好氛围并引导用户进一步探索。
- 问题推荐展示：紧随欢迎语下方，展示一组精心挑选的 3 至 5 个问题列表，旨在作为引导性提示，激发用户兴趣。这些问题可以是趣味性强、效果显著的提问，或是特色功能的引导性询问，用户可通过单击操作直接发起询问。
- 输入交互区域：为用户输入问题提供便捷界面，同时可以根据技术能力接入语音输入、文件及图片上传等多样化输入方式，以提升用户体验的灵活性和便捷性。
- 问答交互流程：问答环节以用户提问、模型即时响应的一问一答形式呈现。后端能力对接大模型接口。为实现高效流畅的用户体验，一般推荐使用流式输出。此外，模型回答结果中可嵌入点赞、点踩、分享及播报等互动功能，以增强用户参与感和满意度。

其他辅助功能如下。

- 新对话开启：此功能通常在最近一次对话结束后的特定时间范围内有效，允许用户主动清空当前对话上下文，为新的对话轮次做准备。
- 历史记录管理：当前主流实现方式分为两类：一类是时间线式记录，所有对话按时间顺序串行保存，用户可通过下拉界面在顶部逐步加载历史内容；另一类是主题式并行存储，每个主题独立成组，便于用户回溯历史记录或在该主题下继续提问。

通过这些精心设计的模块，大模型问答功能不仅能够满足用户的基本需求，还能够提供更加丰富和个性化的交互体验。

智能体功能模块是大模型技术中用于提供个性化用户体验的关键组件。市场上的智能体创建功能主要集中于两个核心方面：角色设定和知识库的构建。用户可以自定义智能体的名称和角色描述，同时，系统可提供基于用户输入的“一句话描述”自动生成角色设定的功能，以增强用户体验。

- 角色个性化设定：此环节涵盖名称自定义与角色（或人物设定）的详细描述，均由用户直接输入完成。此外，根据系统能力，还可引入智能服务，允许用户通过简短的“一句话描述”自动生成初步设定，以增强用户体验的便捷性。
- 知识库构建与上传：旨在赋予智能体特定领域内的专业能力，通过用户上传专业知识文档或待分析材料实现。用户端提交文档后，后端系统对接文档解析引擎，自动创建索引，确保智能体能够准确理解并应用这些知识。
- 智能体广场：此模块的成功构建依赖于持续的内容积累与运营。初期，可依据模型现有能力，创建一系列富有创意、特色鲜明的智能体作为示范，以吸引用户关注。随后，通过持续的运营推广与用户自发创建，不断丰富智能体广场的内容生态，实现可持续发展。

智能体模型能力模块如下。

- 对话能力：在考量成本和资源的基础上，我们可以选择集成高效的大模型服务接口或自主部署开源大模型服务。本场景中的对话通常为用户的休闲交流，对于深入的专业知识需求较低。如系统智能体所言，若存在特定领域的知识数据，我们可以通过智能体提供专业化的服务。
- 审核能力：对于面向公众的服务，合规性是至关重要的。模型生成的内容需要经过严格的审核流程，包括语义分析、敏感词过滤等，以确保内容不涉及政治敏感、色情、暴力、恐怖主义、欺诈等违规内容。
- 实时信源能力：大模型在处理时效性问题时可能存在局限。为解决这一问题，我们可以通过整合外部实时数据源，如最新日期、天气更新、即时检索结果等，强化模型的回答能力。结合意图识别技术，智能地调用相关信源，确保提供的信息既准确又及时。

通过这些功能及模型能力的整合与优化，我们可以构建一个智能又可靠的系统，满足

用户的日常需求，同时确保服务的合规性和信息的实时性。

1.6　风险与挑战

大模型的发展也面临着一些挑战，如算力资源、存储资源、网络通信瓶颈等问题，以及如何在中低速应用场景下保证访问和生成内容的适宜性和可控性。尽管如此，随着技术的不断进步和应用场景的逐步成熟，大模型正在逐渐实现产业化落地，并为人类社会的发展带来更多的机遇和挑战。

1.6.1　技术角度

1. 算力资源

大模型的出现并非偶然，它是各领域长期理论研究和应用创新迭代的产物，其中包括算力芯片、云计算等相关领域的长期积累，缺一不可。这些先进的算力资源为大模型提供了必要的算力底座，支撑着大模型进行数以千亿计的数据样本处理和模型参数训练，已经成为发展和应用大模型技术的首要前提。

在国家“十四五”规划中，国内智能算力规模年复合增长率将超过 50%。产业界预测，到 2030 年全球智能算力将达到 105 ZFLOPS，将是 2023 年年初的 500 倍，增长速度远超通用算力的增长速度。为了在这轮智能化浪潮中处于领先地位，支撑以大模型为代表的新一轮智能产业的发展，各地纷纷投资建设智能计算（智算）中心。

根据国家信息中心与浪潮信息联合发布的《智能计算中心创新发展指南》，截至 2023 年 1 月，我国已有超过 30 个城市正在积极投入智算中心建设。此外，华为在其全面智能化战略中明确指出，将为大模型发展持续打造坚实的算力底座，实现百模千态，为千行万业赋能。

越来越多的垂域大模型的出现也将带来算力瓶颈的转移，越来越多的企业将会更需要量身定制的大模型，但是大模型耗费的算力也是非常大的。因此如何分配算力资源将来也会是一个很大的社会问题。

2. 数据问题

大模型在数据方面最大的两个问题，一个是数据量，另一个是数据安全。例如普遍应用于企业的办公大模型需要用到的数据是公司内部的数据，包括公司内部的制度、公文、业务等数据。这些数据来源多种多样，并且涉及公司机密，不能公开，无论是数据质量或者是数据安全方面都是不小的挑战。

训练大模型需要高质量、大规模、多样的数据。高质量数据集能够提高模型精度与可解释性；海量大规模的数据集可以使预训练模型的效果越来越好；数据集的丰富度可以提升模型的泛化能力，让模型能够适应更多的场景。

数据集通常有几种获取方式，一种是公开数据集，如图像数据集 ImageNet、MNIST 等，这些数据一般是由研究机构，或者公司开放的，在特定领域内广泛使用和共享。庞大的数据也就意味着算力也是非常重要的资源。在算力资源、高质量数据资源日趋宝贵的今天，我们再也不能陷入重复训练浪费算力资源的陷阱，大模型走向开源+共训模式将更符合未来的高质量发展需求。

另一种是合作数据集，如企业和高校或研究机构合作，但是通常这类数据集不会公开分享，只能通过合作关系获得。还有就是通过用户使用标注的数据集，如办公场景是我们经常使用的场景，公司内部管理、业务流程协助都需要使用公司的自有数据。这些数据来源广泛，并且数据结构也很难统一，往往只适用于行业内部场景。这种数据前期通常由公司自行标注，后期在使用的过程中，使用程序进行标注。

数据安全问题主要体现在两方面，即数据泄露问题和数据权限控制问题。

现在大多数企业会选择使用行业大模型，那么大模型的数据就一定会包含公司内部的数据，这些数据是构成行业大模型的语料基础，而且是公司的隐私需要被保护起来的。只通过设置 IaaS 层隔离的方式是不能把数据放到公网上的。但如果通过本地化部署的方式将大模型部署到服务提供者的服务器上，对于服务提供者来说算力又是一个非常大的成本挑战。

另一方面，数据控制需要保证特定的数据只有特定的用户来访问，这样可以有效地保证企业数据不外泄，尤其是在有外部合作的情况下。此外，还需要建立完备的访问权限，如读写改等权限分别存储，也可以有效防止企业的核心数据被擅自篡改的情况。对于训练模型的数据也需要进行清洗、脱敏等操作以有效保护企业隐私。但是有时候企业用户也很可能主动或被动地在向大模型提问的过程中将企业数据泄露给大模型。

3. 实用性

大模型的实用性也是大家关心的问题。毕竟一个大模型从训练到消耗的算力资源、存储资源、时间成本都是非常大的，那么它到底可以做什么呢？目前大家更倾向去使用纯文本类的大语言模型进行单纯的文本类任务。例如 LLM 任务型对话通常用于解决多场景问题，包括网店客服、办公助手；或者使用大模型生成日常工作总结、邮件等完成事务性工作；又或者是摘要提取，例如英文的论文在翻译成中文的同时还可以提取出论文中的要点。大模型也可以帮助我们完成简单的代码编辑，包括多种语言，如 Python、Java，甚至是 SQL 语句也可以完成。

视觉能力的大模型也逐渐从最早生成图片取悦我们变成了扩图神器，又或者是自动修复的工具。例如我们在拍照时场景较为局促，那么可以使用扩图将整个画面变得更为完整，使构图更完美；或者是一些老照片可以通过 AI 手段进行修复，去除其中的斑点；又或者为老照片填色，还原了很多历史场景。

现在还有越来越多的办公产品也结合了大模型和小模型。如公文生成、PPT 生成等，从而增强了大模型的应用范围。未来还会有更多的功能将被挖掘出来。

4. 可解释性

大语言模型在自然语言处理任务上的惊艳表现引起了社会广泛的关注。与此同时，如何解释大模型在跨任务中令人惊艳的表现是学术界面临的迫切挑战之一。相比于传统的机器学习或者深度学习模型，超大的模型架构和海量的学习资料使得大模型具备了强大的推理泛化能力。大语言模型提供可解释性的几个主要难点包括：

- 模型复杂性高：区别于 LLM 时代之前的深度学习模型或者传统的统计机器学习模型，LLM 模型规模巨大，包含数十亿个参数，其内部表示和推理过程非常复杂，很难针对其具体的输出给出解释。

- 数据依赖性强：LLM 在训练过程中依赖大规模文本语料，这些训练数据中的偏见、错误等都可能影响模型，但很难完整判断训练数据的质量对模型的影响。例如在最近的调查中，我们可以看到文生图的模型在“亚洲男人和他的白人妻子”这个提示词下生成图片的效果欠佳，这就是因为训练数据中普遍存在的偏见问题造成的。
- 黑箱性质：我们通常把 LLM 看作黑箱模型，即使是对于开源的模型来说，如 Llama-2，我们也很难显式地判断它的内部推理链和决策过程，只能根据输入输出进行分析，这给可解释性带来困难。
- 输出不确定性：LLM 的输出常常存在不确定性，对同一输入可能产生不同输出，这也增加了可解释性的难度。
- 评估指标不足：目前对话系统的自动评估指标还不足以完整反映模型的可解释性，需要更多考虑人类理解的评估指标。像是生成内容和提示词的相关性，这属于相当主观的判断标准。例如要生成一个 mouse 的图片，可能出现鼠标的图片，也可能出现老鼠的图片，并不能说哪个是准确的。

5. 网络攻击

在大模型大规模扩展的同时，网络犯罪的数量也在逐渐攀升。派拓网络发布的 2024 年亚太地区网络安全趋势预测指出，“自 2022 年 10 月 ChatGPT 上线以来，全世界都担心它可能会导致网络犯罪泛滥。尽管 ChatGPT 有防止恶意应用的措施，但只需要一些有创意的提示，就能让 ChatGPT 大批量生成听上去‘极像人类’、近乎无懈可击的网络钓鱼邮件。我们已经看到攻击者利用生成式 AI，以深度伪造和语音技术等新手段从银行骗取巨额钱财。使用生成式 AI 的企业需要警惕模型中毒、数据泄露、提示注入攻击等漏洞。随着生成式 AI 在合法用例中的使用日益增加，攻击者将不断找出各种新漏洞并加以利用。”

1.6.2 伦理角度

1. 生成内容合法合规

生成式大模型生成的内容一般是随机生成的且不可人为控制，因此存在一些生成内容

违法或违规问题。例如历史问题回答不真实或者政治问题存在偏见等，或者是生成的内容包含涉黄涉暴等违法内容，这些都会对社会造成不良影响，让使用者感到不适。这通常是因为通用大模型数据集的质量参差不齐，或者是缺乏某一类的知识等，而造成幻觉问题。

根据中国人工智能学会 CAAI 的定义，生成式大模型的伦理内涵分为三层，一是人类在开发和使用人工智能相关技术、产品和系统时的道德准则及行为规范；二是人工智能体本身所具有的符合伦理准则的道德编程或价值嵌入方法；三是人工智能体通过自我学习推理而形成的伦理规范。

目前我国已经针对大模型生成内容进行了监管，2023 年 5 月，北京市人民政府办公厅印发了《北京市促进通用人工智能创新发展的若干措施》，2023 年 7 月七部门联合公布了《生成式人工智能服务管理暂行办法》，自 2023 年 8 月 15 日起施行。大模型全面增强了人工智能的功能性，对全社会、全行业进行新的赋能，之后随着模型的使用规模增加，新的监管条例也在不断完善中。

现在商用的大模型底座模型或者是使用大模型的算法模型都需要在国家互联网信息办公室进行备案，备案过程中也会对模型做安全审核等工作。通常安全检查也会使用一些诱导性的词语，以考察大模型是否会生成有违伦理道德的内容。随着安全风险案例的出现，我们的大模型伦理安全也将会逐步完善。

2. 职业发展

随着大模型的出现和不断的发展，可以预见一些工作岗位将面临挑战甚至消失，同时也会涌现出许多新的工作岗位。传统的劳动密集型工作岗位，如客服、翻译，可能有一部分将会被任务型问答机器人所取代。机器人可以实现 7×24 小时的实时问答服务，并且准确度和即时性远超过人工客服。但是大模型在另一方面也创造出了新的就业机会，数据科学家、大模型开发工程师都会成为未来大模型发展的主力军。此外，随着对大模型潜力的不断挖掘，以大模型为底座模型进行其他功能的二次开发将会是人才市场的下一个机会。对于其他领域的非技术型人才来说，人机协作也许会是以后的职业发展方向。让大模型代替人力进行重复、烦琐的工作，可以有效提高工作效率。

大模型的出现势必会造成很大的影响，这也将引领一次新的工业革命，推动生产力的再一次提升。