AI营销培训 第一部分文案

# 一、内容纲领

## 1.1 关键词

原理

## 1.2 内容安排

1. 大模型发展背景 -> 基本概念原理概括

（简介）

1. 当前大模型发展现状，特别是国产模型状况以及其优势（ 智谱Qwen / 阿里Qwen / …… ）

（简介）

1. 语言大模型主流发展路线（ 注意力机制 / 循环神经网络 - RNNs ） -> 原理讲解

（略过）

1. 国产大模型介绍和 基本使用/本地部署 （针对不同基础，需求）讲解

（重点）

# 二、具体讲解内容

## 2.1 人工智能（大模型）沿革历史

人工智能（Artificial Intelligence，简称AI）是指由计算机系统所展现出来的智能行为。它是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。

学习、推理、自动规划、知识表示、自然语言处理、感知、机器学习、机器人技术

### 2.1.1 起步发展期 [1943年—20世纪60年代]

人工智能概念的提出后，发展出了符号主义、联结主义(神经网络)等五种流派

前后相继取得了一批令人瞩目的研究成果，如机器定理证明、跳棋程序、人机对话等，掀起人工智能发展的第一个高潮。

挑关键时间点来讲

1943年，美国神经科学家麦卡洛克（Warren McCulloch）和逻辑学家皮茨（Water Pitts）提出神经元的数学模型，这是现代人工智能学科的奠基石之一。

1950年，艾伦·麦席森·图灵（Alan Mathison Turing）提出“图灵测试”（测试机器是否能表现出与人无法区分的智能），让机器产生智能这一想法开始进入人们的视野。

图示

中度可信度描述已自动生成

1950年，克劳德·香农（Claude Shannon）提出计算机博弈。

1956年，达特茅斯学院人工智能夏季研讨会上正式使用了人工智能（artificial intelligence，AI）这一术语。这是人类历史上第一次人工智能研讨，标志着人工智能学科的诞生。

1957年,弗兰克·罗森布拉特（Frank Rosenblatt）在一台IBM-704计算机上模拟实现了一种他发明的叫做“感知机”（Perceptron）的神经网络模型。[感知机可以被视为一种最简单形式的前馈式人工神经网络，是一种二分类的线性分类判别模型，其输入为实例的特征向量想（x1,x2…），神经元的激活函数f为sign，输出为实例的类别（+1或者-1），模型的目标是要将输入实例通过超平面将正负二类分离。]

图示

描述已自动生成

1958年，David Cox提出了logistic regression[LR是类似于感知机结构的线性分类判别模型，主要不同在于神经元的激活函数f为sigmoid，模型的目标为(最大似然)极大化正确分类概率。]

图示

中度可信度描述已自动生成

1959年，Arthur Samuel给机器学习了一个明确概念：Field of study that gives computers the ability to learn without being explicitly programmed.（机器学习是研究如何让计算机不需要显式的程序也可以具备学习的能力）。

1961年，Leonard Merrick Uhr 和 Charles M Vossler发表了题目为A Pattern Recognition Program That Generates, Evaluates and Adjusts its Own Operators 的模式识别论文，该文章描述了一种利用机器学习或自组织过程设计的模式识别程序的尝试。

1965年，古德（I. J. Good）发表了一篇对人工智能未来可能对人类构成威胁的文章，可以算“AI威胁论”的先驱。他认为机器的超级智能和无法避免的智能爆炸最终将超出人类可控范畴。后来著名科学家霍金、发明家马斯克等人对人工智能的恐怖预言跟古德半个世界前的警告遥相呼应。

1966 年，麻省理工学院科学家Joseph Weizenbaum 在 ACM 上发表了题为《ELIZA-a computer program for the study of natural language communication between man and machine》文章描述了ELIZA 的程序如何使人与计算机在一定程度上进行自然语言对话成为可能，ELIZA 的实现技术是通过关键词匹配规则对输入进行分解，而后根据分解规则所对应的重组规则来生成回复。

1967年，Thomas等人提出K最近邻算法（The nearest neighbor algorithm）。[KNN的核心思想，即给定一个训练数据集，对新的输入实例Xu，在训练数据集中找到与该实例最邻近的K个实例，以这K个实例的最多数所属类别作为新实例Xu的类别。]

图示, 示意图

描述已自动生成

1968年，爱德华·费根鲍姆（Edward Feigenbaum）提出首个专家系统DENDRAL，并对知识库给出了初步的定义，这也孕育了后来的第二次人工智能浪潮。该系统具有非常丰富的化学知识，可根据质谱数据帮助化学家推断分子结构。

[专家系统（Expert Systems）是AI的一个重要分支

（！！！！！三阶段？？？）

，同自然语言理解，机器人学并列为AI的三大研究方向。它的定义是使用人类专家推理的计算机模型来处理现实世界中需要专家作出解释的复杂问题，并得出与专家相同的结论，可视作“知识库(knowledge base)”和“推理机(inference machine)” 的结合。]

图示

描述已自动生成

1969年，“符号主义”代表人物马文·明斯基（Marvin Minsky）的著作《感知器》提出对XOR线性不可分的问题：单层感知器无法划分XOR原数据，解决这问题需要引入更高维非线性网络（MLP, 至少需要两层），但多层网络并无有效的训练算法。这些论点给神经网络研究以沉重的打击，神经网络的研究走向长达10年的低潮时期。

图表

描述已自动生成

（！！！！奠定基础？？？？）

### 2.1.2 反思低谷期 [20世纪70年代]

人工智能发展初期的突破性进展大大提升了人们对人工智能的期望，人们开始尝试更具挑战性的任务，然而计算力及理论等的匮乏使得不切实际目标的落空，人工智能的发展走入低谷。

1974年，哈佛大学沃伯斯(Paul Werbos)博士论文里，首次提出了通过误差的反向传播(BP)来训练人工神经网络，但在该时期未引起重视。[BP算法的基本思想不是（如感知器那样）用误差本身去调整权重，而是用误差的导数（梯度）调整。通过误差的梯度做反向传播，更新模型权重, 以下降学习的误差，拟合学习目标，实现网络的万能近似功能的过程。]

图示

描述已自动生成

1975年，马文·明斯基(Marvin Minsky)在论文《知识表示的框架》(A Framework for Representing Knowledge)中提出用于人工智能中的知识表示学习框架理论。

1976年，兰德尔·戴维斯（Randall Davis）构建和维护的大规模的知识库，提出使用集成的面向对象模型可以提高知识库（KB）开发、维护和使用的完整性。

1976年，斯坦福大学的肖特利夫(Edward H. Shortliffe)等人完成了第一个用于血液感染病的诊断、治疗和咨询服务的医疗专家系统MYCIN。

1976年，斯坦福大学的博士勒纳特发表论文《数学中发现的人工智能方法——启发式搜索》，描述了一个名为“AM”的程序，在大量启发式规则的指导下开发新概念数学，最终重新发现了数百个常见的概念和定理。

1977年，海斯·罗思(Hayes. Roth)等人的基于逻辑的机器学习系统取得较大的进展，但只能学习单一概念，也未能投入实际应用。

1979年，汉斯·贝利纳（Hans Berliner）打造的计算机程序战胜双陆棋世界冠军成为标志性事件。(随后，基于行为的机器人学在罗德尼·布鲁克斯和萨顿等人的推动下快速发展，成为人工智能一个重要的发展分支。格瑞·特索罗等人打造的自我学习双陆棋程序又为后来的强化学习的发展奠定了基础。)

### 2.1.3 应用发展期 [20世纪80年代]

人工智能走入应用发展的新高潮。专家系统模拟人类专家的知识和经验解决特定领域的问题，实现了人工智能从理论研究走向实际应用、从一般推理策略探讨转向运用专门知识的重大突破。而机器学习(特别是神经网络)探索不同的学习策略和各种学习方法，在大量的实际应用中也开始慢慢复苏。

1980年，在美国的卡内基梅隆大学(CMU)召开了第一届机器学习国际研讨会，标志着机器学习研究已在全世界兴起。

1980年，德鲁·麦狄蒙（Drew McDermott）和乔恩·多伊尔（Jon Doyle）提出非单调逻辑，以及后期的机器人系统。

1980年，卡耐基梅隆大学为DEC公司开发了一个名为XCON的专家系统，每年为公司节省四千万美元，取得巨大成功。

1981年，保罗（R.P.Paul）出版第一本机器人学课本，“Robot Manipulator：Mathematics，Programmings and Control”，标志着机器人学科走向成熟。

1982年，马尔（David Marr）发表代表作《视觉计算理论》提出计算机视觉（Computer Vision）的概念，并构建系统的视觉理论，对认知科学（CognitiveScience）也产生了很深远的影响。

1982年，约翰·霍普菲尔德（John Hopfield） 发明了霍普菲尔德网络，这是最早的RNN的雏形。霍普菲尔德神经网络模型是一种单层反馈神经网络（神经网络结构主要可分为前馈神经网络、反馈神经网络及图网络），从输出到输入有反馈连接。它的出现振奋了神经网络领域，在人工智能之机器学习、联想记忆、模式识别、优化计算、VLSI和光学设备的并行实现等方面有着广泛应用。

1983年，Terrence Sejnowski, Hinton等人发明了玻尔兹曼机（Boltzmann Machines），也称为随机霍普菲尔德网络，它本质是一种无监督模型，用于对输入数据进行重构以提取数据特征做预测分析。

图示

描述已自动生成

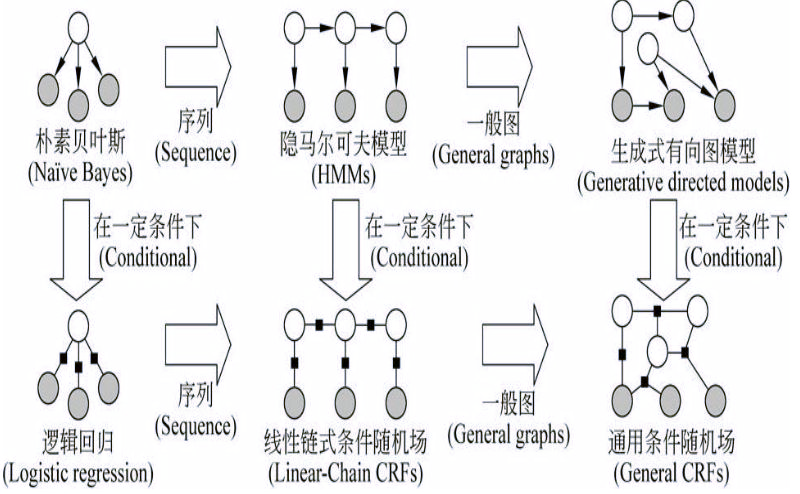
1985年，朱迪亚·珀尔提出贝叶斯网络(Bayesian network)，他以倡导人工智能的概率方法和发展贝叶斯网络而闻名，还因发展了一种基于结构模型的因果和反事实推理理论而受到赞誉。

[贝叶斯网络拓朴结构是一个有向无环图(DAG)，通过把某个研究系统中涉及的随机变量，根据是否条件独立绘制在一个有向图中，以描述随机变量之间的条件依赖，用圈表示随机变量(random variables)，用箭头表示条件依赖(conditional dependencies)就形成了贝叶斯网络。对于任意的随机变量，其联合概率可由各自的局部条件概率分布相乘而得出。如图中b依赖于a(即：a->b)，c依赖于a和b，a独立无依赖，根据贝叶斯定理有 P(a,b,c) = P(a)\*P(b|a)\*P(c|a,b)]

形状, 多边形

描述已自动生成

[贝叶斯网络是一种模拟人类推理过程中因果关系的不确定性处理模型，如常见的朴素贝叶斯分类算法就是贝叶斯网络最基本的应用。]



1986年，罗德尼·布鲁克斯(Brooks)发表论文《移动机器人鲁棒分层控制系统》，标志着基于行为的机器人学科的创立，机器人学界开始把注意力投向实际工程主题。

1986年，辛顿(Geoffrey Hinton)等人先后提出了多层感知器(MLP)与反向传播（BP）训练相结合的理念（该方法在当时计算力上还是有很多挑战，基本上都是和链式求导的梯度算法相关的），这也解决了单层感知器不能做非线性分类的问题，开启了神经网络新一轮的高潮。

图示

中度可信度描述已自动生成

1986年，昆兰（Ross Quinlan）提出ID3决策树算法。[决策树模型可视为多个规则(if, then)的组合，与神经网络黑盒模型截然不同是，它拥有良好的模型解释性。ID3算法核心的思想是通过自顶向下的贪心策略构建决策树：根据信息增益来选择特征进行划分（信息增益的含义是 引入属性A的信息后，数据D的不确定性减少程度。也就是信息增益越大，区分D的能力就越强)，依次递归地构建决策树。]

图示

描述已自动生成

1989年，George Cybenko证明了“万能近似定理”（universal approximation theorem）。简单来说，多层前馈网络可以近似任意函数，其表达力和图灵机等价。这就从根本上消除了Minsky对神经网络表达力的质疑。

[“万能近似定理”可视为神经网络的基本理论：⼀个前馈神经⽹络如果具有线性层和⾄少⼀层具有 “挤压” 性质的激活函数（如 sigmoid 等），给定⽹络⾜够数量的隐藏单元，它可以以任意精度来近似任何从⼀个有限维空间到另⼀个有限维空间的 borel 可测函数。]

图示, 多边形

描述已自动生成

1989年，LeCun (CNN之父) 结合反向传播算法与权值共享的卷积神经层发明了卷积神经网络（Convolutional Neural Network，CNN），并首次将卷积神经网络成功应用到美国邮局的手写字符识别系统中。

[卷积神经网络通常由输入层、卷积层、池化（Pooling）层和全连接层组成。卷积层负责提取图像中的局部特征，池化层用来大幅降低参数量级(降维)，全连接层类似传统神经网络的部分，用来输出想要的结果。]

图示

描述已自动生成

### 2.1.4 平稳发展期 [20世纪90年代—2010年]

由于互联网技术的迅速发展，加速了人工智能的创新研究，促使人工智能技术进一步走向实用化，人工智能相关的各个领域都取得长足进步。在2000年代初，由于专家系统的项目都需要编码太多的显式规则，这降低了效率并增加了成本，人工智能研究的重心从基于知识系统转向了机器学习方向。

1995年，Cortes和Vapnik提出联结主义经典的支持向量机(Support Vector Machine)，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。

[支持向量机（Support Vector Machine, SVM）可以视为在感知机基础上的改进，是建立在统计学习理论的VC维理论和结构风险最小原理基础上的广义线性分类器。与感知机主要差异在于：1、感知机目标是找到一个超平面将各样本尽可能分离正确(有无数个)，SVM目标是找到一个超平面不仅将各样本尽可能分离正确，还要使各样本离超平面距离最远(只有一个最大边距超平面)，SVM的泛化能力更强。2、对于线性不可分的问题，不同于感知机的增加非线性隐藏层，SVM利用核函数，本质上都是实现特征空间非线性变换，使可以被线性分类。]

感知器

图示

描述已自动生成

SVM

图表, 散点图

描述已自动生成

1995年， Freund和schapire提出了 AdaBoost (Adaptive Boosting)算法。AdaBoost采用的是Boosting集成学习方法——串行组合弱学习器以达到更好的泛化性能。另外一种重要集成方法是以随机森林为代表的Bagging并行组合的方式。以“偏差-方差分解”分析，Boosting方法主要优化偏差，Bagging主要优化方差

[Adaboost迭代算法基本思想主要是通过调节的每一轮各训练样本的权重(错误分类的样本权重更高)，串行训练出不同分类器。最终以各分类器的准确率作为其组合的权重，一起加权组合成强分类器。]

图示

描述已自动生成

1997年国际商业机器公司（简称IBM）深蓝超级计算机战胜了国际象棋世界冠军卡斯帕罗夫。深蓝是基于暴力穷举实现国际象棋领域的智能，通过生成所有可能的走法，然后执行尽可能深的搜索，并不断对局面进行评估，尝试找出最佳走法。

1997年，Sepp Hochreiter 和 Jürgen Schmidhuber提出了长短期记忆神经网络(LSTM)。

[LSTM是一种复杂结构的循环神经网络（RNN），结构上引入了遗忘门、输入门及输出门：输入门决定当前时刻网络的输入数据有多少需要保存到单元状态，遗忘门决定上一时刻的单元状态有多少需要保留到当前时刻，输出门控制当前单元状态有多少需要输出到当前的输出值。这样的结构设计可以解决长序列训练过程中的梯度消失问题。]

钟表的特写

低可信度描述已自动生成

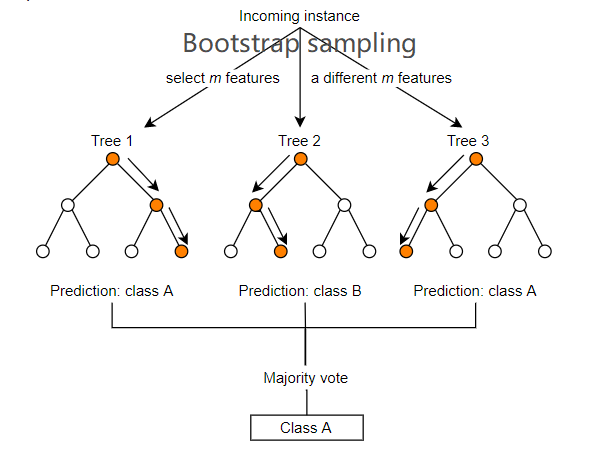
1998年，万维网联盟的蒂姆·伯纳斯·李(Tim Berners-Lee)提出语义网(Semantic Web)的概念。其核心思想是：通过给万维网上的文档（如HTML）添加能够被计算机所理解的语义(Meta data)，从而使整个互联网成为一个基于语义链接的通用信息交换媒介。换言之，就是构建一个能够实现人与电脑无障碍沟通的智能网络。

2001年，John Lafferty首次提出条件随机场模型（Conditional random field，CRF）。CRF是基于贝叶斯理论框架的判别式概率图模型，在给定条件随机场P ( Y ∣ X ) 和输入序列x，求条件概率最大的输出序列y \*。在许多自然语言处理任务中比如分词、命名实体识别等表现尤为出色。

2001年，布雷曼博士提出随机森林（Random Forest）。

随机森林是将多个有差异的弱学习器(决策树)Bagging并行组合，通过建立多个的拟合较好且有差异模型去组合决策，以优化泛化性能的一种集成学习方法。多样差异性可减少对某些特征噪声的依赖，降低方差（过拟合），组合决策可消除些学习器间的偏差。

[随机森林算法的基本思路是对于每一弱学习器(决策树)有放回的抽样构造其训练集，并随机抽取其可用特征子集，即以训练样本及特征空间的多样性训练出N个不同的弱学习器，最终结合N个弱学习器的预测（类别或者回归预测数值），取最多数类别或平均值作为最终结果。]



2003年，David Blei, Andrew Ng和 Michael I. Jordan于2003年提出LDA（Latent Dirichlet Allocation）。

[LDA是一种无监督方法，用来推测文档的主题分布，将文档集中每篇文档的主题以概率分布的形式给出，可以根据主题分布进行主题聚类或文本分类。]

图示

描述已自动生成

2003年，Google公布了3篇大数据奠基性论文，为大数据存储及分布式处理的核心问题提供了思路：非结构化文件分布式存储（GFS）、分布式计算（MapReduce）及结构化数据存储（BigTable），并奠定了现代大数据技术的理论基础。

图片包含 公司名称

描述已自动生成

2005 年，波士顿动力公司推出一款动力平衡四足机器狗，有较强的通用性，可适应较复杂的地形。

2006年，杰弗里·辛顿以及他的学生鲁斯兰·萨拉赫丁诺夫正式提出了深度学习的概念（Deeping Learning），开启了深度学习在学术界和工业界的浪潮。2006年也被称为深度学习元年，杰弗里·辛顿也因此被称为深度学习之父。

[深度学习的概念源于人工神经网络的研究，它的本质是使用多个隐藏层网络结构，通过大量的向量计算，学习数据内在信息的高阶表示。]

图表

描述已自动生成

2010年，Sinno Jialin Pan和 Qiang Yang发表文章《迁移学习的调查》。

[迁移学习(transfer learning)通俗来讲，就是运用已有的知识（如训练好的网络权重）来学习新的知识以适应特定目标任务，核心是找到已有知识和新知识之间的相似性。（知识蒸馏属于其中的特例）]

图示

描述已自动生成

### 2.1.5 蓬勃发展期 [2011年至今]

随着大数据、云计算、互联网、物联网等信息技术的发展，泛在感知数据和图形处理器等计算平台推动以深度神经网络为代表的人工智能技术飞速发展，大幅跨越了科学与应用之间的技术鸿沟，诸如图像分类、语音识别、知识问答、人机对弈、无人驾驶等人工智能技术实现了重大的技术突破，迎来爆发式增长的新高潮。

2011年，IBM Watson问答机器人参与Jeopardy回答测验比赛最终赢得了冠军。Waston是一个集自然语言处理、知识表示、自动推理及机器学习等技术实现的电脑问答（Q&A）系统。

2012年，Hinton和他的学生Alex Krizhevsky设计的AlexNet神经网络模型在ImageNet竞赛大获全胜，这是史上第一次有模型在 ImageNet 数据集表现如此出色，并引爆了神经网络的研究热情。

[AlexNet是一个经典的CNN模型，在数据、算法及算力层面均有较大改进，创新地应用了Data Augmentation、ReLU、Dropout和LRN等方法，并使用GPU加速网络训练。]

图示

描述已自动生成

2012年，谷歌正式发布谷歌知识图谱Google Knowledge Graph），它是Google的一个从多种信息来源汇集的知识库，通过Knowledge Graph来在普通的字串搜索上叠一层相互之间的关系，协助使用者更快找到所需的资料的同时，也可以知识为基础的搜索更近一步，以提高Google搜索的质量。

日程表

描述已自动生成

[知识图谱是结构化的语义知识库，是符号主义思想的代表方法，用于以符号形式描述物理世界中的概念及其相互关系。其通用的组成单位是RDF三元组(实体-关系-实体)，实体间通过关系相互联结，构成网状的知识结构。]

图片包含 图示

描述已自动生成

2013年，Durk Kingma和Max Welling在ICLR上以文章《Auto-Encoding Variational Bayes》提出变分自编码器（Variational Auto-Encoder，VAE）。

[VAE基本思路是将真实样本通过编码器网络变换成一个理想的数据分布，然后把数据分布再传递给解码器网络，构造出生成样本，模型训练学习的过程是使生成样本与真实样本足够接近。]

图示

描述已自动生成

2013年，Google的Tomas Mikolov 在《Efficient Estimation of Word Representation in Vector Space》提出经典的 Word2Vec模型用来学习单词分布式表示，因其简单高效引起了工业界和学术界极大的关注。

图示

描述已自动生成

[Word2Vec基本的思想是学习每个单词与邻近词的关系，从而将单词表示成低维稠密向量。通过这样的分布式表示可以学习到单词的语义信息，直观来看，语义相似的单词的距离相近。]

电脑合成图

中度可信度描述已自动生成

[Word2Vec网络结构是一个浅层神经网络（输入层-线性全连接隐藏层->输出层），按训练学习方式可分为CBOW模型(以一个词语作为输入，来预测它的邻近词)或Skip-gram模型 (以一个词语的邻近词作为输入，来预测这个词语)。]

图示

描述已自动生成

2014年，聊天程序“尤金·古斯特曼”（Eugene Goostman）在英国皇家学会举行的“2014图灵测试”大会上，首次“通过”了图灵测试。

2014年，Goodfellow及Bengio等人提出生成对抗网络（Generative Adversarial Network，GAN），被誉为近年来最酷炫的神经网络。

[GAN是基于强化学习(RL)思路设计的，由生成网络(Generator, G)和判别网络(Discriminator, D)两部分组成， 生成网络构成一个映射函数G: Z→X（输入噪声z, 输出生成的伪造数据x）, 判别网络判别输入是来自真实数据还是生成网络生成的数据。在这样训练的博弈过程中，提高两个模型的生成能力和判别能力。]

图示, 示意图

描述已自动生成

2015年，为纪念人工智能概念提出60周年，深度学习三巨头LeCun、Bengio和Hinton(他们于2018年共同获得了图灵奖)推出了深度学习的联合综述《Deep learning》。

[《Deep learning》文中指出深度学习就是一种特征学习方法，把原始数据通过一些简单的但是非线性的模型转变成为更高层次及抽象的表达，能够强化输入数据的区分能力。通过足够多的转换的组合，非常复杂的函数也可以被学习。]

图示

描述已自动生成

2015年，Microsoft Research的Kaiming He等人提出的残差网络（ResNet）在ImageNet大规模视觉识别竞赛中获得了图像分类和物体识别的优胜。

[残差网络的主要贡献是发现了网络不恒等变换导致的“退化现象（Degradation）”，并针对退化现象引入了 “快捷连接（Shortcut connection）”，缓解了在深度神经网络中增加深度带来的梯度消失问题。]

图示

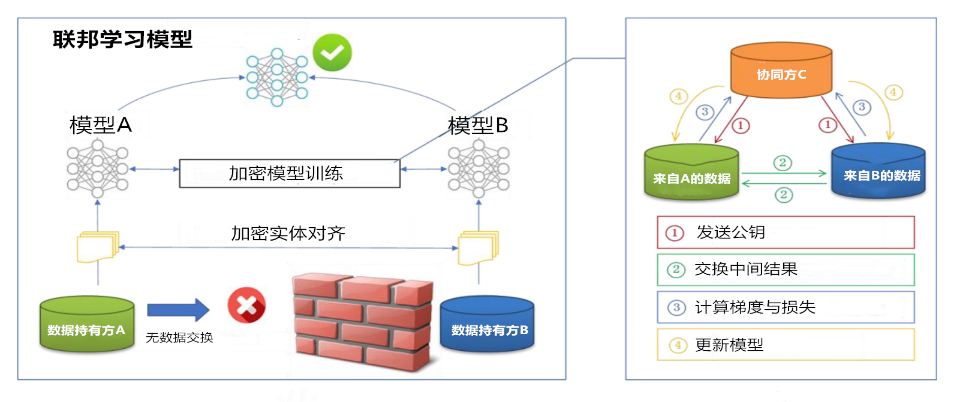
描述已自动生成

2015年，谷歌开源TensorFlow框架。它是一个基于数据流编程（dataflow programming）的符号数学系统，被广泛应用于各类机器学习（machine learning）算法的编程实现，其前身是谷歌的神经网络算法库DistBelief。

2015年，马斯克等人共同创建OpenAI。它是一个非营利的研究组织，使命是确保通用人工智能 (即一种高度自主且在大多数具有经济价值的工作上超越人类的系统）将为全人类带来福祉。其发布热门产品的如：OpenAI Gym，GPT等。

2016年，谷歌提出联邦学习方法，它在多个持有本地数据样本的分散式边缘设备或服务器上训练算法，而不交换其数据样本。

[联邦学习保护隐私方面最重要的三大技术分别是： 差分隐私 ( Differential Privacy )、同态加密 ( Homomorphic Encryption )和 隐私保护集合交集 ( Private Set Intersection )，能够使多个参与者在不共享数据的情况下建立一个共同的、强大的机器学习模型，从而解决数据隐私、数据安全、数据访问权限和异构数据的访问等关键问题。]



### 2.1.6 国产模型

见2.4介绍

## 2.2 大模型基本原理

人工智能研究目的

人工智能（Artificial Intelligence，AI）研究目的是通过探索智慧的实质，扩展人类智能——促使智能主体会听（语音识别、机器翻译等）、会看（图像识别、文字识别等）、会说（语音合成、人机对话等）、会思考（人机对弈、专家系统等）、会学习（知识表示，机器学习等）、会行动（机器人、自动驾驶汽车等）。一个经典的AI定义是：“ 智能主体可以理解数据及从中学习，并利用知识实现特定目标和任务的能力。

人工智能学派

主要思想

代表方法

联结主义

利用数学模型来研究人类认知的方法，用神经元的连接机制实现人工智能

神经网络、SVM等

符号主义

认知就是通过对有意义的表示符号进行推导计算，并将学习视为逆向演绎，主张用显式的公理和逻辑体系搭建人工智能系统

专家系统，知识图谱，决策树等

演化主义

对生物进化进行模拟，使用遗传算法和遗传编程

遗传算法等

贝叶斯主义

使用概率规则及其依赖关系进行推理

朴素贝叶斯等

行为主义

以控制论及感知-动作型控制系统原理模拟行为以复现人类智能

强化学习等

其中，**符号主义及联结主义**为主要的两大派系：

“符号主义”（Symbolicism），又称逻辑主义、计算机学派，认为认知就是通过对有意义的表示符号进行推导计算，并将学习视为逆向演绎，主张用显式的公理和逻辑体系搭建人工智能系统。如用决策树模型输入业务特征预测天气：

图示

描述已自动生成

“联结主义”（Connectionism），又叫仿生学派，笃信大脑的逆向工程，主张是利用数学模型来研究人类认知的方法，用神经元的连接机制实现人工智能。如用神经网络模型输入雷达图像数据预测天气：

图表

描述已自动生成

模态（输入）、参数 / 提示词、输出（结果）、黑盒

第一阶段：基于规则的方法（符号主义）

第二阶段：传统机器学习

第三阶段：深度学习与大模型

下面是基于这种划分方式的详细说明：

第一阶段：基于规则的方法（符号主义）

时间范围：大约从1950年代到1970年代末

核心特征：这一阶段的人工智能研究侧重于逻辑和符号推理，试图通过明确编程的方式来模仿人类的推理过程。

里程碑事件：

1956年的达特茅斯会议标志着人工智能作为一个研究领域的正式开始4。

专家系统：如MYCIN和DENDRAL等，这些系统使用规则和逻辑表达式来解决特定领域的问题1。

语言处理系统：例如SHRDLU，它是一个能够理解简单英语命令并执行相应操作的玩具世界系统1。

第二阶段：传统机器学习

时间范围：大约从1980年代初到2000年代末

核心特征：这一阶段开始转向数据驱动的方法，利用统计学习理论和算法来构建预测模型。

里程碑事件：

神经网络的复兴：1980年代末至1990年代初，神经网络的研究重新受到重视

支持向量机（SVMs）：在1990年代中期，SVMs因其在小样本数据上的出色表现而得到广泛应用11。

决策树和集成方法：如随机森林等，这些方法在分类和回归任务中表现出色

第三阶段：深度学习与大模型

时间范围：大约从2010年代初至今

核心特征\*\*\*基础\*\*\*：这一阶段的特点是深度学习技术的兴起，特别是卷积神经网络（CNNs）、循环神经网络（RNNs）和最近的Transformer架构。

按训练方法分类\*\*\*基础\*\*\*：

无监督学习大模型：这些模型通过自我监督的方式在未标注数据上学习。

半监督学习大模型：结合少量标注数据和大量未标注数据进行训练。

监督学习大模型：仅在标注数据上进行训练。

按应用场景分类\*\*\*基础\*\*\*：（文生图！！！ NLP + CV ）

自然语言处理（NLP）大模型：例如用于文本生成、翻译、问答等任务的模型。

计算机视觉（CV）大模型：例如用于图像分类、目标检测、语义分割等任务的模型。

语音识别大模型：例如用于语音转文字、音频分类等任务的模型

里程碑事件：

深度学习的突破：2012年，AlexNet在ImageNet竞赛中取得胜利，标志着深度学习时代的开始。

自然语言处理的进展：2018年，BERT模型的发布标志着预训练模型在自然语言处理领域的巨大突破。

大模型的出现：例如GPT-3、PaLM等，这些模型在大量数据上进行预训练，能够在多种任务上展现出强大的性能。

（语言大模型LLM、多模态？）

二、人工智能的主要分支

（一）机器学习

介绍监督学习、非监督学习和强化学习。

常见算法和应用场景。

（二）自然语言处理（NLP）

语言模型、聊天机器人和情感分析的基础。

（三）计算机视觉

图像识别、物体检测和图像生成技术。

（四）机器人技术

自动化与智能化的机器人应用实例。

（五）生成式人工智能

（六）通用人工智能

（七）具身人工智能与机器人技术

## 2.3 大模型发展路线

### 2.3.1 要素

专家系统（Expert Systems）是AI的一个重要分支，同自然语言理解，机器人学并列为AI的三大研究方向

人工智能有三个要素：数据、算力及算法，数据即是知识原料

图示

描述已自动生成

#### 2.3.1.1 数据

数据是现实世界映射构建虚拟世界的基本要素，随着数据量以指数形式增长，开拓的虚拟世界的疆土也不断扩张。不同于AI算法开源，关键数据往往是不开放的，数据隐私化、私域化是一种趋势，数据之于AI应用，如同流量是互联网的护城河，有核心数据才有关键的AI能力。

图表

描述已自动生成

#### 2.3.1.2 算力

计算是AI的关键，自2010年代以来的深度学习浪潮，很大程度上归功于计算能力的进步。

量子计算发展

在计算芯片按摩尔定律发展越发失效的今天，计算能力进步的放慢会限制未来的AI技，量子计算提供了一条新量级的增强计算能力的思路。随着量子计算机的量子比特数量以指数形式增长，而它的计算能力是量子比特数量的指数级，这个增长速度将远远大于数据量的增长，为数据爆发时代的人工智能带来了强大的硬件基础。

边缘计算发展

边缘计算作为云计算的一种补充和优化，一部分的人工智能正在加快速度从云端走向边缘，进入到越来越小的物联网设备中。而这些物联网设备往往体积很小，为此轻量机器学习（TinyML）受到青睐，以满足功耗、延时以及精度等问题。

类脑计算发展

以类脑计算芯片为核心的各种类脑计算系统，在处理某些智能问题以及低功耗智能计算方面正逐步展露出优势。类脑计算芯片设计将从现有处理器的设计方法论及其发展历史中汲取灵感，在计算完备性理论基础上结合应用需求实现完备的硬件功能。同时类脑计算基础软件将整合已有类脑计算编程语言与框架，实现类脑计算系统从“专用”向“通用”的逐步演进。

人工智能计算中心成为智能化时代的关键基础设施

人工智能计算中心基于最新人工智能理论，采用领先的人工智能计算架构，是融合公共算力服务、数据开放共享、智能生态建设、产业创新聚集的“四位一体”综合平台，可提供算力、数据和算法等人工智能全栈能力，是人工智能快速发展和应用所依托的新型算力基础设施。未来，随着智能化社会的不断发展，人工智能计算中心将成为关键的信息基础设施，推动数字经济与传统产业深度融合，加速产业转型升级，促进经济高质量发展。

当前主流……（显卡NVIDIA）

#### 2.3.1.3 算法

机器学习自动化(AutoML)发展

自动化机器学习（AutoML）解决的核心问题是：在给定数据集上使用哪种机器学习算法、是否以及如何预处理其特征以及如何设置所有超参数。随着机器学习在许多应用领域取得了长足的进步，这促成了对机器学习系统的不断增长的需求，并希望机器学习应用可以自动化构建并使用。借助AutoMl、MLOps技术，将大大减少机器学习人工训练及部署过程，技术人员可以专注于核心解决方案。

图片包含 日程表

描述已自动生成

向分布式隐私保护方向演进

当前全球多个国家和地区已出台数据监管法规，如HIPAA（美国健康保险便利和责任法案）、GDPR（欧盟通用数据保护条例）等，通过严格的法规限制多机构间隐私数据的交互。分布式隐私保护机器学习(联邦学习)通过加密、分布式存储等方式保护机器学习模型训练的输入数据，是打破数据孤岛、完成多机构联合训练建模的可行方案。

数据和机理融合

AI模型的发展是符合简单而美的定律的。从数据出发的建模从数据中总结规律，追求在实践中的应用效果。从机理出发的建模以基本物理规律为出发点进行演绎，追求简洁与美的表达。一个好的、主流的的模型，通常是高度总结了数据规律并切合机理的，是“优雅”的，因为它触及了问题的本质。就和科学理论一样，往往简洁的，没有太多补丁，而这同时解决了收敛速度问题和泛化问题。

神经网络模型结构发展

神经网络的演进一直沿着模块化+层次化的方向，不断把多个承担相对简单任务的模块组合起来。神经网络结构通过较低层级模块侦测基本的特征，并在较高层级侦测更高阶的特征，无论是多层前馈网络，还是卷积神经网络，都体现了这种模块性(近年Hinton提出的“胶囊”（capsule）网络就是进一步模块化发展)。因为我们处理的问题（图像、语音、文字）往往都有天然的模块性，学习网络的模块性若匹配了问题本身内在的模块性，就能取得较好的效果。层次化并不仅仅是网络的拓扑叠加，更重要的是学习算法的升级，仅仅简单地加深层次可能会导致BP网络的梯度消失等问题。

多学派方法融合发展

通过多学派方法交融发展，得以互补算法之间的优势和弱点。如 1）贝叶斯派与神经网络融合，Neil Lawrence组的Deep Gaussian process, 用简单的概率分布替换神经网络层。2）符号主义、集成学习与神经网络的融合，周志华老师的深度随机森林。3) 符号主义与神经网络的融合：将知识库(KG)融入进神经网络，如GNN、知识图谱表示学习。4) 神经网络与强化学习的融合，如谷歌基于DNN+强化学习实现的Alpha Go 让AI的复杂任务表现逼近人类。

基于大规模无(自)监督预训练发展

“如果智能是一块蛋糕，那么蛋糕的大部分是无监督学习，蛋糕上的糖衣是监督学习，而蛋糕上的樱桃是强化学习（RL）——Yann Lecun”监督学习需要足够的带标签数据，然而人工标注大量数据既耗时又费力，在一些领域(如医学领域)上几乎不太可能获得足量的标注数据。通过大规模无(自)监督预训练方法利用现实中大量的无标签数据是一个研究的热点，如GPT-3的出现激发了对大规模自监督预训练方法继续开展探索和研究。未来，基于大规模图像、语音、视频等多模态数据的跨语言的自监督预训练模型将进一步发展，并不断提升模型的认知、推理能力。

基于因果学习方法发展

当前人工智能模型大多关注于数据特征间相关性，而相关性与更为本源的因果关系并不等价，可能导致预测结果的偏差，对抗攻击的能力不佳，且模型往往缺乏可解释性。另外，模型需要独立同分布(i.i.d.)假设(现实很多情况，i.i.d.的假设是不成立的)，若测试数据与训练数据来自不同的分布，统计学习模型往往效果不佳，而因果推断所研究的正是这样的情形：如何学习一个可以在不同分布下工作、蕴含因果机制的因果模型(Causal Model)，并使用因果模型进行干预或反事实推断。

可解释性AI (XAI)发展

可解释的人工智能有可能成为未来机器学习的核心，随着模型变得越来越复杂，确定简单的、可解释的规则就会变得越来越困难。一个可以解释的AI（Explainable AI, 简称XAI）意味着AI运作的透明，便于人类对于对AI监督及接纳，以保证算法的公平性、安全性及隐私性。

图形用户界面, 应用程序

描述已自动生成

### 2.3.2 发展方向

多模态（多模型混合）：

多模态大模型能够处理多种类型的数据，例如文本、图像、音频等，以实现更全面的理解和生成能力。

端侧智能：

针对移动设备和边缘计算的优化，使大模型能够在低功耗设备上运行，提供即时响应的服务。

Agents（智能体）：

大模型与智能体技术的结合，使模型能够执行更复杂的任务，例如规划、决策和交互。

网络架构创新：

探索新的网络架构，以克服现有Transformer架构的局限性，例如降低计算资源消耗和提高效率。

自主学习能力：

发展模型的自我进化能力，让模型能够自动创建数据进行自我训练，减少对外部数据的依赖。

模型压缩与优化：

通过模型压缩、蒸馏等技术减少模型大小，提高推理速度，降低部署成本。

跨学科融合：

将计算机科学与其他学科如心理学、生物学等结合，以促进更深入的理解和更广泛的应用。

通用人工智能：

朝着构建能够处理广泛任务的通用人工智能方向发展，这要求模型具备更强的适应性和自我规划能力。

物理形态模拟：

模型不仅限于数字形态，还可能探索物理形态的模拟，如机器人技术中的应用。

自我规划与多步规划：

让模型能够规划复杂的任务序列，进行自我管理并完成一系列任务。

## 2.4 国产大模型介绍

### 阿里系 -> 通义千问

通义千问是由阿里云自主研发的大语言模型，它用于理解和分析用户输入的自然语言，在不同领域和任务为用户提供服务和帮助。通义千问于2023年9月13日正式向公众开放，并且随着不断的更新，它已经成为了一个多功能、多模态的大模型，不仅支持文本生成、对话模拟等基本功能，还能处理图像和视频等多媒体数据。

### 技术特点与应用

通义千问采用了先进的自然语言处理技术，拥有庞大的参数量，比如在2024年的某个版本中达到了18亿参数。这使得它能够在处理复杂的语言任务时表现优异，包括但不限于多轮对话、实时交互、异步通信等。

除了文本处理之外，通义千问还推出了专门针对视觉语言处理的版本——通义千问VL，这是阿里云研发的大规模视觉语言模型，能够以图像、文本、检测框（多模态）作为输入，并以文本和检测框作为输出，支持中文多模态对话及多图对话。

### 开源与应用

通义千问不仅是商业服务的一部分，同时也积极参与到了开源社区中。阿里云表示，通义是大模型开源最积极的推动者和实践者。通过开源，通义千问促进了AI技术的发展，并且为开发者提供了一个学习和改进AI模型的平台。

### 应用场景

通义千问的应用场景十分广泛，它不仅可以帮助解答问题、文档阅读、联网搜索并写作总结，还能用于文字创作，例如撰写故事、公文、邮件、剧本和诗歌等。此外，它也被用于智能编码助手——通义灵码，以及与联发科天玑9300芯片合作，实现了移动端AI的重大突破。

### 总结

通义千问代表了阿里云在人工智能技术方面的重大进步，它不仅展示了公司在语言模型研究上的深厚积累，也为用户带来了实际的价值。无论是个人用户还是企业客户，都能从中受益，享受到更智能、更便捷的服务。随着技术的不断进步和应用场景的拓展，通义千问有望在未来继续引领AI技术的发展方向。

模型系列介绍…………

1. 文字创作与文本处理

通义千问能够基于用户的指令和需求，生成符合要求的文本内容，如创作故事、撰写公文或邮件等。其强大的文本生成能力也为文艺创作者提供了便利，如辅助创作剧本、诗歌等。此外，通义千问还可以进行文本润色和摘要提取，帮助用户优化文本表达和提高阅读效率。

2. 编程辅助

通义千问能够理解编程语言和逻辑，辅助程序员编写和优化代码，提高开发效率。这一功能对于软件开发人员来说尤为重要，可以大大减轻他们的工作负担，提高编程质量和速度。

3. 翻译服务

通义千问支持多种语言的实时翻译服务，为跨国交流提供了便利。无论是商务沟通还是日常交流，用户都可以通过通义千问实现即时翻译，打破语言障碍。

4. 对话模拟与智能客服

通义千问能够扮演不同角色进行交互式对话，为用户提供模拟对话体验。在电商行业，通义千问可以作为智能客服使用，理解消费者的问题并给予准确解答，提高客户满意度。此外，通义千问还可以应用于教育领域，作为智能问答系统使用，为学生提供实时的学习辅导和解答疑问。

5. 数据可视化与智能推荐

通义千问能够根据用户提供的数据和信息，自动生成各类图表，帮助用户更好地展示和分析数据。同时，基于大数据处理能力，通义千问可以更准确地挖掘用户喜好和需求，为用户推荐更为精准的商品和内容。这一功能在电商行业中尤为重要，可以帮助商家提高转化率和销售额。

配套平台 --- 魔搭社区（ModelScope）

阿里巴巴魔搭社区（ModelScope）是一个由阿里云发起的人工智能模型社区平台，旨在降低人工智能技术的应用门槛，并促进AI模型的共享与交流。以下是关于魔搭社区的一些关键信息：

1. \*\*创立背景\*\*：2022年，阿里云在业界率先提出了“模型即服务”（MaaS）的理念，并以此为基础创建了魔搭ModelScope社区。该社区致力于提供一个全面的AI模型开发与应用的生态系统，包括但不限于模型的选择、数据集的获取、算力的支持等。

2. \*\*模型资源\*\*：魔搭社区首批上架了超过300个AI模型，涵盖视觉、语音、自然语言处理（NLP）、多模态等多个领域，其中中文模型占了相当大的比例。这些模型包括了150多个最先进模型（SOTA）和十多个大型模型，均全面开源并开放使用。

3. \*\*一站式服务\*\*：为了方便开发者，魔搭社区提供了一站式的服务，包括模型探索、推理、训练、部署和应用等环节。同时，社区还推出了SwingDeploy服务，支持将魔搭上的模型一键部署到用户指定的云资源上。

4. \*\*合作与贡献\*\*：魔搭社区的合作方包括了澜舟科技、智谱AI、深势科技等多家科研机构。这些合作机构共同为魔搭社区贡献了丰富的AI模型资源。

5. \*\*社区特色\*\*：魔搭社区的一大特色是提供了详尽的模型卡片描述，帮助用户更好地理解模型的工作原理和使用方法。此外，魔搭社区还关注AI模型的定制化需求，提供必要的工具和服务以帮助开发者根据具体应用进行模型调整。

6. \*\*开源计划\*\*：2024年全球开发者先锋大会上，阿里云魔搭社区启动了ModelScope-Sora开源计划，聚焦于中国多模态大模型的研究。这项计划旨在推动AI技术的发展，提高模型处理效率与质量。

7. \*\*技术支撑\*\*：魔搭社区背后有着强大的技术支持，包括数据处理系统Data-Juicer等工具链，这些工具帮助开发者高效处理多模态数据。

8. \*\*愿景与使命\*\*：魔搭社区的愿景是成为下一代开源的模型即服务共享平台，为AI开发者提供灵活、易用、低成本的一站式模型服务产品。通过降低AI应用的成本和技术门槛，魔搭希望加速AI技术在各个行业的普及应用。

魔搭社区作为一个综合性的AI模型社区，不仅为专业开发者提供了丰富的资源和服务，同时也为AI技术的推广和应用创造了条件。随着更多机构和个人的加入，魔搭社区将继续成长壮大，进一步推动AI领域的创新与发展。

魔搭社区（ModelScope）与阿里通义系列模型之间存在着紧密的联系。以下是二者之间的具体关系：

1. \*\*模型发布与开源\*\*：阿里通义系列模型，作为阿里巴巴达摩院研发的大规模预训练模型，通过魔搭社区进行发布和开源。例如，通义千问系列模型（Qwen）就在魔搭社区上进行了开源。这使得开发者能够直接访问、下载和使用这些模型，以便于在自己的项目中进行集成和微调。

2. \*\*模型即服务（MaaS）\*\*：魔搭社区旨在打造一个“模型即服务”的平台，而通义系列模型正是这种服务模式的重要组成部分。通过魔搭社区，用户可以方便地获取通义系列模型提供的服务，例如自然语言处理、图像识别等，而无需从头开始训练模型。

3. \*\*技术资源共享\*\*：魔搭社区不仅提供模型本身，还提供了一系列工具和服务，如模型训练推理工具箱，支持LLaMA、千问、ChatGLM、BaiChuan等多种模型及LoRA等多种训练方式。这些工具和服务可以帮助用户更有效地利用通义系列模型。

4. \*\*模型应用案例\*\*：魔搭社区展示了通义系列模型在不同场景下的应用案例。例如，中华联合财产保险股份有限公司就利用阿里云的通义灵码大幅提升了编码效率。这样的案例有助于其他企业了解如何在实际业务中应用通义系列模型。

5. \*\*社区互动与支持\*\*：魔搭社区为通义系列模型的用户和开发者提供了一个互动平台，用户可以在社区内分享使用经验和技巧，也可以获得来自阿里云及其他社区成员的技术支持和指导。

6. \*\*模型丰富性与多样性\*\*：魔搭社区覆盖了视觉、语音、自然语言处理、多模态等AI主要领域，而通义系列模型作为其中的重要组成部分，为这些领域提供了强有力的支持。这意味着无论是研究人员还是商业用户，都能在魔搭社区找到适合自己的通义系列模型。

通过上述几点可以看出，魔搭社区与阿里通义系列模型之间存在着一种共生关系，魔搭社区为通义系列模型提供了一个展示和使用的平台，而通义系列模型则丰富了魔搭社区的内容，两者共同促进了AI技术的发展和应用。

### 月之暗面（阿里系？） -> Kimi

Kimi模型是一款由国内人工智能初创公司“月之暗面”开发的大规模语言模型，以其卓越的长文本处理能力和实时信息搜索功能著称。以下是关于Kimi模型的主要技术和特点的详细介绍：

核心技术

解决长上下文序列问题的新机制：Kimi模型采用了一种新机制来解决长上下文序列的问题，这使得它能够处理高达200万字的长文本输入，这是一个非常显著的技术突破。

无损压缩技术：Kimi模型还使用了无损压缩技术来优化存储和处理大量文本数据的方式，这对于处理超长文本尤为重要。

深度学习和自然语言处理：Kimi模型基于深度学习和自然语言处理技术，设计灵感来源于人类大脑的工作原理，通过模拟神经网络的方式来实现智能决策和学习能力。

个性化定制：Kimi模型能够根据用户的个性化需求和偏好来进行定制化服务。

实时信息搜索：Kimi模型具备快速整合实时信息的能力，可以为用户提供详尽的答案并给出信息来源，无论是日常知识查询、专业领域研究还是最新的新闻追踪。

数据来源

Kimi模型的数据来源非常广泛，不仅限于知乎，还包括国内外的其他网站。但是由于国内的运营限制，其主要依赖知乎等专业网站的数据。

技术来源

Moonshot AI的技术来源主要包括以下几个方面：

核心团队的经验：Moonshot AI的核心团队成员在谷歌等国际领先的人工智能研究机构积累了丰富的经验。他们在开发大规模语言模型方面的背景，为Moonshot AI提供了坚实的技术基础。

学术研究成果：创始人杨植麟在自然语言处理领域的学术研究为公司带来了前沿的技术洞见和技术积累。特别是他在Transformer-XL和XLNet等项目中的工作，对公司在长文本处理方面的能力产生了直接影响。

技术创新：Moonshot AI在处理超长文本方面进行了创新，例如开发了能够处理200万字长文本的技术，这在业界是非常先进的。

合作伙伴与投资：与阿里巴巴等大型科技公司的合作也为Moonshot AI带来了资金和技术资源的支持。这些合作伙伴关系有助于Moonshot AI加速技术研发和市场推广。

开源社区：虽然没有直接提到Moonshot AI是否利用了开源社区的技术成果，但通常来说，开源软件和模型对于推动人工智能领域的发展至关重要，Moonshot AI也很可能受益于开源社区的技术进步。

### 清华系 -> 智谱清言

智谱AI是一家源自清华大学的技术成果转化公司，在人工智能领域有着深厚的技术积累和创新成果。以下是关于智谱AI的一些关键信息，包括其股东资本组成和技术来源。

股东资本组成

智谱AI在成立和发展过程中经历了多次融资，吸引了多家知名投资机构和个人投资者的关注和支持。以下是一些重要的投资节点和股东信息：

早期融资：智谱AI成立于2019年6月，初始注册资本约为2084.6万元，随后在2024年8月，注册资本增加到了约2511.5万元。

B轮融资：2022年9月，智谱AI完成了数亿元人民币的B轮融资，由君联资本和启明创投联合领投。

后续融资：2023年，智谱AI进行了新一轮融资，估值突破了200亿人民币。

股东变动：2023年，中科创星作为创始股东，曾连续出售其持有的部分智谱AI股份。

投资方：智谱AI的投资方包括但不限于：君联资本、启明创投、中科创星、美团、蚂蚁集团、阿里巴巴、腾讯、小米、金山、顺为资本、Boss直聘等。

中东石油巨头沙特阿美（Aramco）旗下风险投资部门管理的基金Prosperity7

主要技术来源

智谱AI的技术来源可以追溯到清华大学计算机系的KEG（知识工程）实验室，该实验室拥有超过27年的人工智能技术积累。具体来说：

成立背景：智谱AI成立于2019年6月，源自清华大学KEG实验室的技术成果转化。

团队构成：智谱AI的团队主要由来自清华大学的专家和学者组成，包括CEO张鹏、总裁王绍兰、董事长刘德兵等，他们都曾在清华大学计算机系工作或学习。

研发模式：智谱AI采用了一种被称为“产学研用”紧密压缩的闭环的研发模式，这有助于快速推进技术创新和产品迭代。

技术成果：智谱AI专注于开发新一代的认知智能大模型，致力于打造中国创新的大模型技术。其技术突破主要集中在大规模语言模型的开发上，其中包括GLM-130B，这是一个双语千亿级超大规模预训练模型。

技术特色：智谱AI的大模型产品能够处理更长的上下文，使对话更加连贯和自然，并且支持图像、文本等多种模态的数据处理。

### 百度系 -> 文心一言 / 飞桨平台

自研

百度文心一言（ERNIE Bot）是百度基于其在自然语言处理（NLP）领域的深厚积累推出的一款人工智能大模型。以下是关于百度文心一言系列模型的主要技术来源的详细介绍：

技术基础

ERNIE系列模型：文心一言的基础是ERNIE系列模型，这是百度在自然语言处理领域长期研究和技术积累的成果。ERNIE系列模型通过大规模的预训练和特定任务的微调，能够在多种NLP任务上表现出色。

飞桨平台：文心一言依托于百度自主研发的飞桨（PaddlePaddle）开源深度学习平台。飞桨是一个全面的深度学习平台，包含了深度学习核心框架、基础模型库、端到端开发套件、工具组件和服务平台。

知识增强：文心一言采用了知识增强的技术路径，这意味着模型在训练过程中不仅使用了大量的文本数据，还整合了大规模的知识图谱，以此增强模型的理解能力和生成质量。

训练数据

大规模数据集：文心一言的训练数据包括万亿级网页数据、数十亿的搜索数据和图片数据、百亿级的语音日均调用数据以及5500亿事实的知识图谱等。

多模态数据融合：文心一言不仅限于文本数据的处理，还能处理图像、音频等多种模态的数据，这使得它能够更好地理解和生成多模态内容。

关键技术

监督精调：在预训练基础上，通过监督学习的方法对模型进行进一步的精调，以提高特定任务的表现。

人类反馈的强化学习：利用人类的反馈来指导模型的学习过程，使模型能够更好地适应人类的需求和偏好。

提示技术：通过特定的提示（prompt）引导模型生成符合预期的内容，这有助于提高生成内容的质量和多样性。

检索增强：在生成过程中结合外部信息检索，以增强生成内容的真实性和准确性。

对话增强：通过对话历史信息的考虑，使模型能够进行连贯且有意义的对话。

### 字节跳动 -> 豆包

主要技术来源（自研）

Transformer架构：豆包模型的核心技术之一是基于Transformer架构，这种架构通过自注意力机制和位置编码来处理序列数据，从而能够有效地捕捉文本中的长距离依赖关系。

云雀模型：豆包模型是基于云雀语言模型开发制作的，云雀模型也是字节跳动公司基于Transformer架构研发的语言模型。这意味着豆包模型继承了云雀模型的诸多优点，比如在大规模数据上的训练能力和对各种自然语言处理任务的支持。

自注意力机制：自注意力机制允许模型关注到输入序列中的每个位置，这对于捕捉文本中的长距离依赖关系至关重要。

目标函数：豆包模型使用最大似然估计等优化算法来最小化预测序列与真实序列之间的差异，以此来优化模型参数。

训练数据：豆包模型是在大量的文本数据上进行训练的，这意味着它能够学习到丰富的语言模式和表达方式。

火山引擎云服务：豆包模型的应用中可能会使用火山引擎提供的云服务，例如用于文档检索和信息整合的OpenSearch服务。

云雀模型（Skylark）是字节跳动开发的一个大型语言模型

山引擎是字节跳动旗下的一家云服务平台，它将字节跳动在快速发展过程中积累的增长方法、技术能力和应用工具开放给外部企业。

### 腾讯系(企鹅系) -> 混元

腾讯混元大模型是腾讯研发的一款通用大语言模型，具备强大的中文创作能力、复杂语境下的逻辑推理能力以及可靠的任务执行能力。以下是对腾讯混元大模型的详细介绍，特别是关注其主要技术来源：

技术来源与特点

全链路自研：

腾讯混元大模型是由腾讯全链路自研的通用大语言模型，这意味着从模型架构设计、训练到部署，都是由腾讯自主完成的。

参数规模：

混元大模型的参数规模超过千亿，预训练语料超过2万亿tokens10，这样的规模为模型提供了强大的表达能力。

训练框架：

混元大模型采用了腾讯太极机器学习平台自研的训练框架AngelPTM21，这一框架相比于业界主流解决方案有着更高的模型容纳能力和更快的训练速度。

混合专家模型架构：

腾讯混元大模型的技术架构已升级为混合专家模型（MoE）架构，这种架构能够根据输入的不同部分选择不同的专家模块进行处理，从而提高模型的效率和灵活性。

代码处理能力：

混元大模型在代码处理方面表现突出，处理水平提升超过20%，在HumanEval的公开测试集指标中超过了业界头部开源代码大模型。

降低幻觉：

为了提升模型的可靠性和成熟度，混元大模型在降低幻觉方面进行了专门的技术自研。腾讯采用了包括搜索、知识图谱增强在内的多种方法来辅助降低模型的幻觉现象。

多模态能力：

混元大模型具备多模态理解能力，能够处理文本、图像等多种类型的数据。

多轮问答流畅性：

在腾讯云智能客服体系中，基于混元大模型的垂直领域精调模型大幅度提升了智能对话的意图理解准确性和多轮问答的流畅性。

应用场景

内部应用：

腾讯混元大模型已经整合到了腾讯的多个业务中，如微信读书、腾讯会议、腾讯文档等。

对外服务：

通过腾讯云对外开放，供企业接入使用。

AI助手APP：

推出了基于混元大模型的AI助手APP腾讯元宝，提供AI搜索对话框等功能。

行业解决方案：

腾讯云联合生态伙伴，将大模型技术与20多个行业结合，提供了超过50个行业大模型解决方案

### 科大讯飞 -> 星火

科大讯飞星火模型是科大讯飞研发的一款认知智能大模型，具有多项核心能力。以下是关于科大讯飞星火模型的详细介绍，特别关注其主要技术来源：

主要技术来源

国产算力训练：

科大讯飞星火大模型是基于首个全国产算力训练的。这意味着星火模型的训练是在国产的计算平台上完成的，这不仅保证了技术的安全可控，也体现了科大讯飞在国产化技术上的领先地位。

大规模数据和知识学习：

星火模型从海量数据和大规模知识中持续学习进化，这为其提供了广泛的知识基础和语言理解能力。

跨领域的知识和语言理解：

星火模型具备跨领域的知识和语言理解能力，能够基于自然对话方式理解与执行任务，这是通过综合运用多种自然语言处理技术实现的。

多模态输入和表达能力：

星火模型支持多模态输入和表达，这表明模型能够处理不仅仅是文本数据，还包括图像、视频等其他形式的数据。

自主研发的模型架构：

科大讯飞星火模型的所有技术模块均为完全自主研发，这表明科大讯飞在人工智能领域的技术积累和技术独立性。

核心能力

文本生成：

星火模型可以创作新闻稿、撰写文章、润色群聊通知等，并能够根据不同场景进行衍生。

语言理解：

星火模型能够准确理解不同语境下的语义，包括复杂的中文语言特性。

知识问答：

星火模型能够基于广泛的知识库进行问答。

逻辑推理：

星火模型能够进行情景式思维链逻辑推理。

数学能力：

星火模型具有多题型步骤级数学能力。

代码能力：

星火模型具备多功能多语言代码能力。

多模态能力：

星火模型支持多模态输入和表达，包括处理图像、音频等非文本数据。

技术特点

多模态能力提升：

星火模型在多模态处理能力方面进行了显著的提升，特别是在V4.0版本中。

联网搜索能力：

星火模型新增了联网搜索功能，能够直接获取实时信息，增强答案的准确性和可信度。

工具集成：

星火模型集成了多款常用工具，如联网搜索、代码解释器、文生图、天气查询等，提高了处理复杂任务的效率。

数学和语言理解能力超越GPT-4 Turbo：

在V3.5版本中，星火模型在数学和语言理解方面超越了GPT-4 Turbo5。

整体超越GPT-4 Turbo：

在V4.0版本中，星火模型整体超越了GPT-4 Turbo8，并在多个国际权威测试集上取得了领先的成绩。

### 华为系 -> 盘古（专用 / 商用）

华为盘古模型是华为公司在自然语言处理（NLP）、计算机视觉（CV）和其他人工智能（AI）领域开发的一系列大型预训练模型。以下是关于华为盘古模型的详细介绍，特别关注其主要技术来源：

主要技术来源

深度学习技术：

华为盘古模型的核心技术之一是深度学习，特别是基于Transformer架构的大规模语言模型。这种架构非常适合处理自然语言处理任务，因为它能够在处理长序列数据时保持良好的性能。

大规模预训练：

华为盘古模型通过在大量未标注文本数据上进行无监督学习来进行预训练，从而获得了丰富的语言知识和上下文理解能力。这种预训练方法是现代自然语言处理模型的基础。

自然语言处理技术：

华为盘古模型在自然语言处理领域取得了显著成就，例如在文本生成、语音生成、视频生成和翻译等方面的应用。

计算机视觉技术：

华为盘古模型也利用了计算机视觉技术，例如在气象预测中使用的盘古气象模型，它将AI技术应用于气象学领域，提升了天气预报的准确性和时效性。

多模态融合：

华为盘古模型支持多模态数据处理，即能够处理图像、视频和文本等多种类型的数据，这使得模型在多媒体内容理解和生成方面表现出色。

行业应用：

华为盘古模型针对特定行业进行了定制化开发，比如在中医药领域，模拟中医的诊断过程，虽然不能替代真实医生的工作，但可以辅助医生进行初步诊断。

自主知识产权：

华为盘古模型的技术模块都是自主研发的，这体现了华为在人工智能技术领域的自主创新能力。

### 360×

调用，噱头

## 2.5 基本使用 / 本地部署

环境配置，基本技能要求

GitHub（Fast GitHub/Watt-原steam++）

GitLab（中国代理）

Gitee（GitHub中国山寨兼镜像）

Gitcode -> ×

通义千问

（https://tongyi.aliyun.com/qianwen/）

[通义千问大语言模型介绍\_大模型服务平台百炼(Model Studio)-阿里云帮助中心 (aliyun.com)](https://help.aliyun.com/zh/model-studio/developer-reference/what-is-qwen-llm)

智谱清言

（[Swiper demo (jykjrj.cn)](https://zp.jykjrj.cn/)）

文心一言（√）

（[文心一言 (baidu.com)](https://yiyan.baidu.com/)）

（[文心一言文档 (baidu.com)](https://yiyan.baidu.com/learn)）

使用

智能体

部署