

DeepSeek

每个人都可以读懂的大模型科普报告 (企业篇)

DeepSeek大模型及其企业应用实践

林子雨 副教授

厦门大学

厦门大学大数据教学团队作品 2025年3月3日



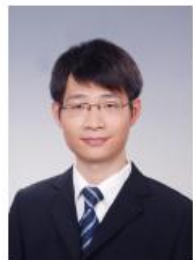
厦门大学大数据教学团队



林子雨



夏小云



赖永炫



陶继平



谢怡



郑海山



蔡炳跃



苏淑文



郑宇辉



张琦

国内高校大数据教学的重要贡献者

团队负责人：**林子雨** 副教授

年轻力量：核心成员全部46周岁以下

结构合理：教学型、科研型、实验工程师

专注专业：从2013年至今，11年专注于大数据教学

团队特点：眼光前瞻、紧跟技术、创新实干、执行力强

影响力高：多项指标在国内高校大数据教学领域领先

- 教材数量
- 教材占有率
- MOOC课程学习人数
- 师资培养
- 教学研讨会
- 教学网站访问量
- 在线讲座观看人数
-

团队联系方式：ziyulin@xmu.edu.cn

大模型系列报告（科普报告，非技术报告）

- 报告1：大模型概念、技术与应用实践（面向社会大众）
- 报告2：DeepSeek大模型赋能高校教学和科研（面向高校）
- **报告3：DeepSeek大模型及其企业应用实践（面向企业）**
- 报告4：DeepSeek大模型赋能政府数字化转型（面向政府部门）

报告下载地址：<https://dblab.xmu.edu.cn/post/deepseek/>



扫码下载报告



目录

1. 大模型：人工智能的前沿
2. 大模型产品
3. 大模型的行业应用
4. 企业大模型落地方案
5. 智能体的企业应用
6. 厂商提供的企业级大模型服务
7. 大模型典型应用案例
8. AIGC与企业应用实践
9. 大模型未来发展趋势



1.大模型：人工智能的前沿

- 1.1 大模型的概念
- 1.2 大模型的发展历程
- 1.3 人工智能与大模型的关系
- 1.4 大模型分类





1.1 大模型的概念

大模型通常指的是大规模的人工智能模型，是一种基于深度学习技术，具有海量参数、强大的学习能力和泛化能力，能够处理和生成多种类型数据的人工智能模型。

通常说的大模型的“大”的特点体现在：

参数数量庞大

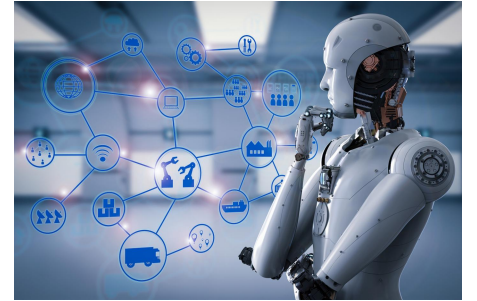
训练数据量大

计算资源需求高

2020年，OpenAI公司推出了GPT-3，模型参数规模达到了1750亿。

2023年3月发布的GPT-4的参数规模是GPT-3的10倍以上，达到1.8万亿，

2021年11月阿里推出的M6 模型的参数量达10万亿。



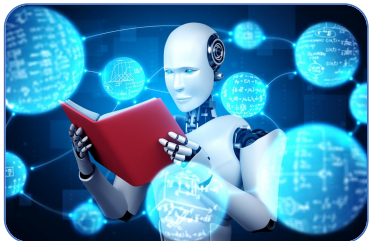


1.1 大模型的概念

大模型的设计和训练旨在提供更强大、更准确的模型性能，以应对更复杂、更庞大的数据集或任务。大模型通常能够学习到更细微的模式和规律，具有更强的泛化能力和表达能力

上下文理解能力

大模型具有更强的上下文理解能力，能够理解更复杂的语意和语境。这使得它们能够产生更准确、更连贯的回答



语言生成能力

大模型可以生成更自然、更流利的语言，减少了生成输出时呈现的错误或令人困惑的问题



学习能力强

大模型可以从大量的数据中学习，并利用学到的知识和模式来提供更精准的答案和预测。这使得它们在解决复杂问题和应对新的场景时表现更加出色



可迁移性高

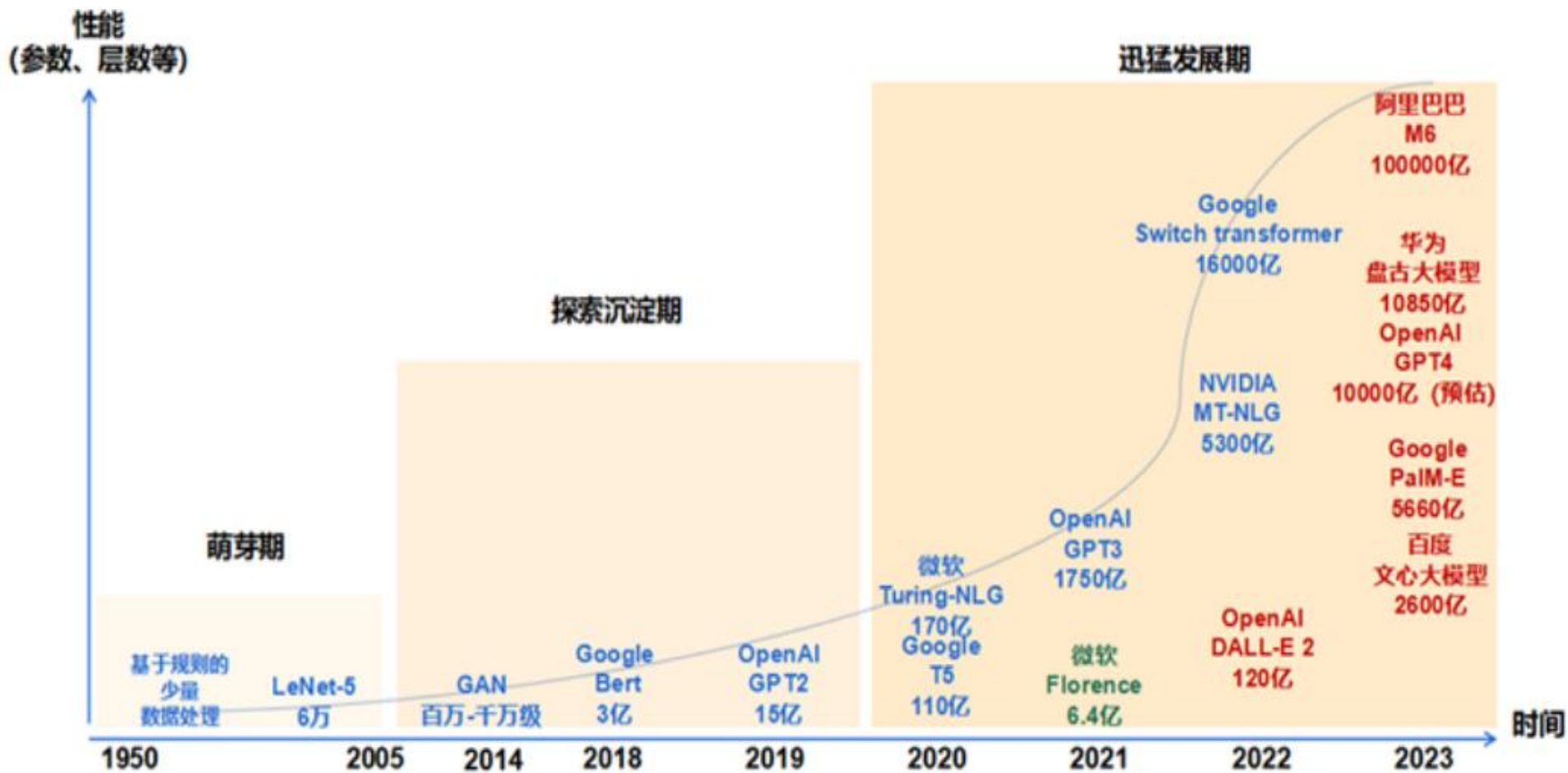
学习到的知识和能力可以在不同的任务和领域中迁移和应用。这意味着一次训练就可以将模型应用于多种任务，无需重新训练





1.2 大模型的发展历程

大模型发展历经三个阶段，分别是萌芽期、沉淀期和爆发期





1.2 大模型的发展历程

大语言模型呈现核心能力金字塔结构

大语言模型

03

进阶能力
更为进阶的理解力



人类情感理解和识别

其他高阶能力

02

进阶能力
逻辑推理能力



编程能力

数学

医学

法律

01

核心能力
认知和学习能力



概念抽象

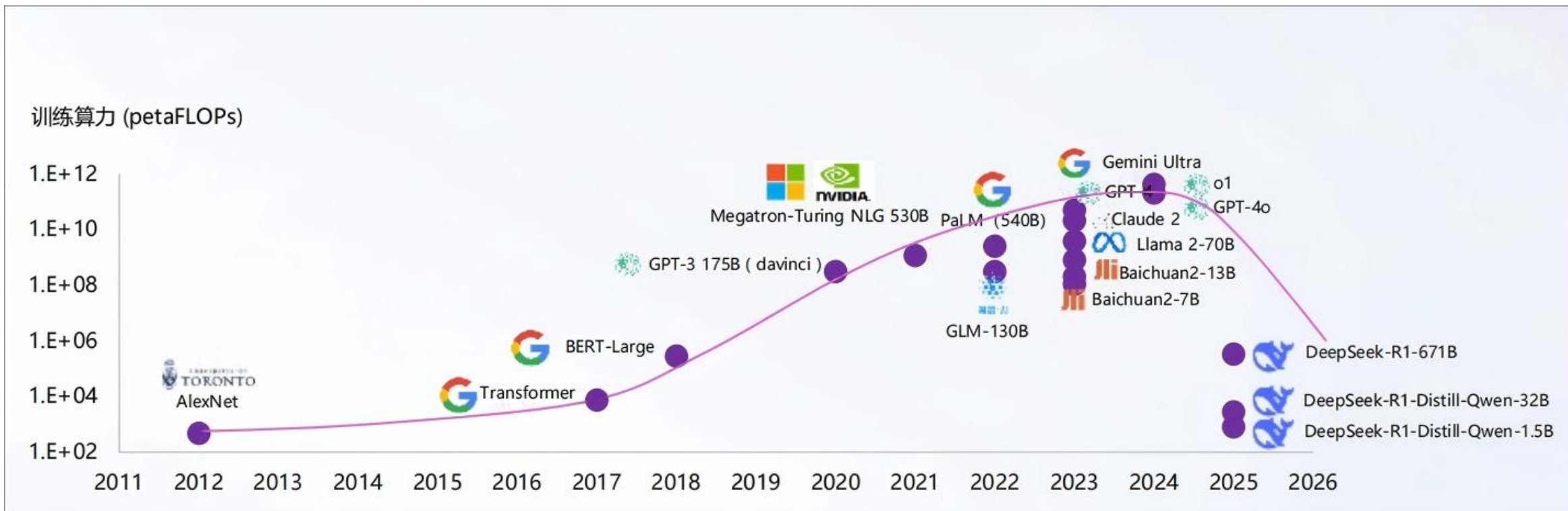
视觉识别

文字理解



1.2 大模型的发展历程

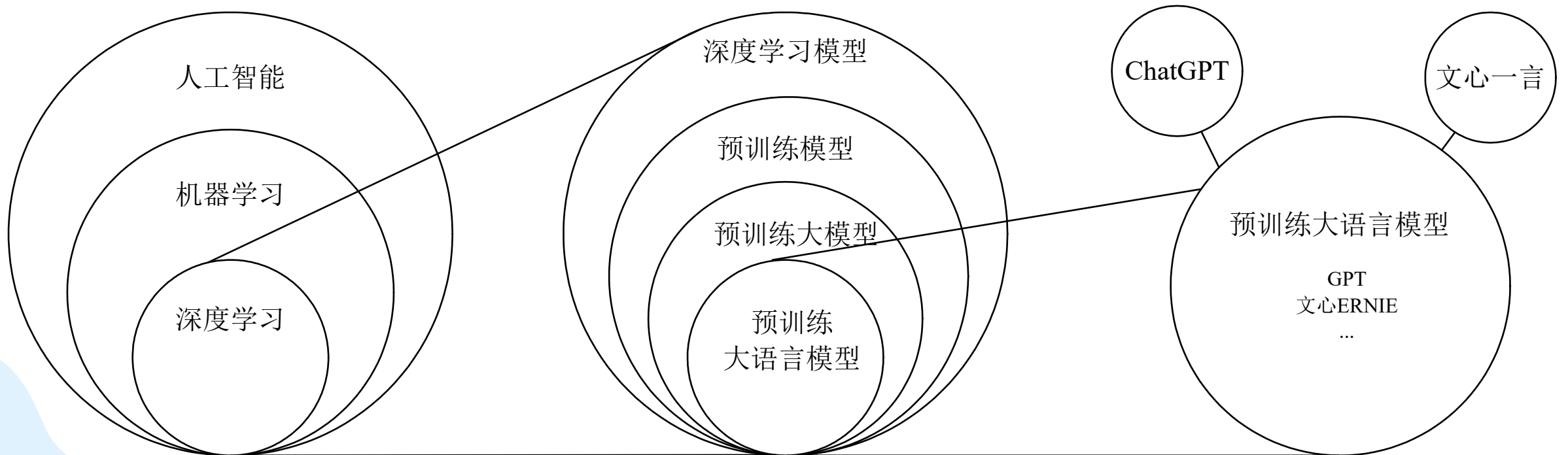
大模型发展对算力的需求演变





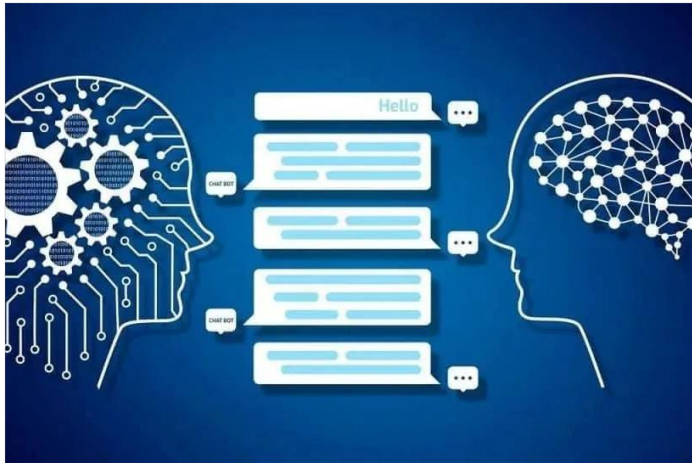
1.3 人工智能与大模型的关系

人工智能包含了机器学习，机器学习包含了深度学习，深度学习可以采用不同的模型，其中一种模型是预训练模型，**预训练模型包含了预训练大模型（可以简称为“大模型”）**，**预训练大模型包含了预训练大语言模型（可以简称为“大语言模型”）**，预训练大语言模型的典型代表包括OpenAI的GPT和百度的文心ERNIE，ChatGPT是基于GPT开发的大模型产品，文心一言是基于文心ERNIE开发的大模型产品





1.4 大模型的分类



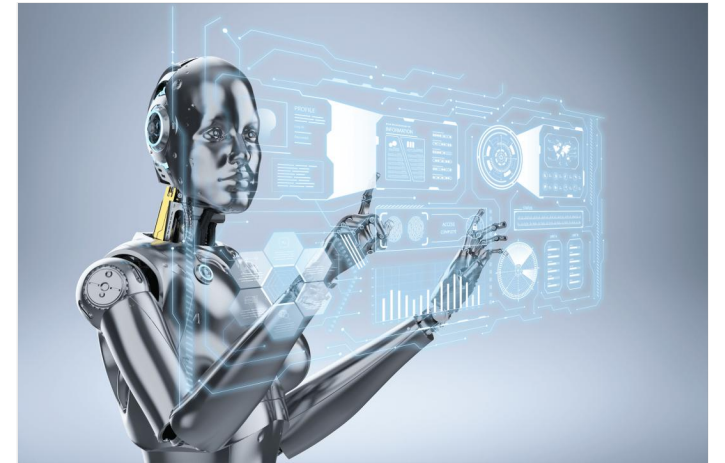
语言大模型

是指在自然语言处理（Natural Language Processing, NLP）领域中的一类大模型，通常用于处理文本数据和理解自然语言。这类大模型的主要特点是它们在大规模语料库上进行了训练，以学习自然语言的各种语法、语义和语境规则。代表性产品包括GPT系列（OpenAI）、Bard（Google）、DeepSeek、文心一言（百度）等



视觉大模型

是指在计算机视觉（Computer Vision, CV）领域中使用的大模型，通常用于图像处理和分析。这类模型通过在大规模图像数据上进行训练，可以实现各种视觉任务，如图像分类、目标检测、图像分割、姿态估计、人脸识别等。代表性产品包括ViT系列（Google）、文心UFO、华为盘古CV、INTERN（商汤）等



多模态大模型

是指能够处理多种不同类型数据的大模型，例如文本、图像、音频等多模态数据。这类模型结合了NLP和CV的能力，以实现对多模态信息的综合理解和分析，从而能够更全面地理解和处理复杂的数据。代表性产品包括DingoDB多模向量数据库（九章云极DataCanvas）、DALL-E(OpenAI)、悟空画画（华为）、midjourney等



1.4 大模型的分类

按照应用领域的不同，大模型主要可以分为L0、L1、L2三个层级

通用大模型L0

是指可以在多个领域和任务上通用的大模型。它们利用大算力、使用海量的开放数据与具有巨量参数的深度学习算法，在大规模无标注数据上进行训练，以寻找特征并发现规律，进而形成可“举一反三”的强大泛化能力，可在不进行微调或少量微调的情况下完成多场景任务，相当于AI完成了“通识教育”

行业大模型L1

是指那些针对特定行业或领域的大模型。它们通常使用行业相关的数据进行预训练或微调，以提高在该领域的性能和准确度，相当于AI成为“行业专家”

垂直大模型L2

是指那些针对特定任务或场景的大模型。它们通常使用任务相关的数据进行预训练或微调，以提高在该任务上的性能和效果



1.4 大模型的分类

大语言模型可以分为**通用大模型**和**推理大模型**

推理大模型

推理大模型的概念大规模传播应该开始于2024年9月份。2024年9月12日，OpenAI官方宣布了OpenAI o1推理大模型。

OpenAI定义推理模型

在OpenAI的官网上，OpenAI定义推理模型是在回答之前进行思考，并在回复用户之前，在内部生成一长串的思维链过程。**思维链是一种提示大语言模型进行逐步推理的方法。**它让模型在得出最终答案之前，先显式地写出推理的中间步骤。这就像人类解决复杂问题时会先把思考过程写下来一样。

推理模型的核心

也就是说，如果模型在回复你之前有一长串的思考过程（这个过程必须可以显示输出），探索了很多不同的路径之后给出答案，那么有这个能力的大模型就是推理大模型。**推理模型的核心在于处理那些需要多步骤逻辑推导才能解决的复杂问题。**



1.4 大模型的分类

■ 推理大模型DeepSeek R1的对话效果

请评价一下厦门大学林子雨副教授的大数据教学水平和成就

请评价一下厦门大学林子雨副教授的大数据教学水平和成就



已深度思考 (用时 25 秒) ^

嗯，用户让我评价厦门大学林子雨副教授的大数据教学水平和成就。首先，我需要确认用户的需求是什么。可能是学生选课前的参考，或者同行想了解他的教学成果，也可能是有意合作的人想评估他的专业能力。接下来，我要收集关于林子雨的信息，确保评价的准确性。

林子雨是厦大计算机系的副教授，主要研究大数据和数据库。他的教学水平方面，用户提到他编写了多本教材，比如《大数据技术原理与应用》。这些教材被广泛使用，说明他的内容可能很系统，适合入门。另外，他建设了在线课程，比如慕课，这可能让更多学生受益，特别是疫情期间在线教育的重要性提升。还有实践教学，比如实验指南和Docker镜像，这对学生动手能力有帮助，符合大数据注重实践的特点。

在成就方面，用户提到他的教材影响力大，被多所高校采用，说明权威性。还有他推动的产学研结合，比如与华为合作，这对学生就业有帮助。另外，公益教育项目，比如给师生提供资源，显示他的社会责任感。获奖情况也是成就的一部分，比如厦门大学奖，说明校内认可。

接下来要考虑用户可能的深层需求。如果用户是学生，可能想知道林老师的课是否适合自己，是否容易拿到资源，课程负担如何。如果是教师或同行，可能关注他的教学方法是否有借鉴意义，或者合作的可能性。此外，用户可能还想知道林老师的业界联系，是否有助于实习或就业。

需要注意的点：评价要客观，不能过于主观。要区分教学和科研成就，虽然用户主要问教学，但如果有科研成就也可以提，但用户可能更关注教学方面。另外，是否有负面评价？如果没有，就不提，但如果有，要谨慎处理。目前提供的信息都是正面的，所以可以集中在正面评价。

厦门大学林子雨副教授在大数据教学领域具有较高的知名度和影响力，其教学水平和成就主要体现在以下几个方面：

一、教学水平与特色

1. 系统化的教学资源建设

林子雨长期致力于大数据课程体系的开发，主编了《大数据技术原理与应用》《Spark编程基础》等多部教材，内容深入浅出，注重理论与实践结合。教材被全国数百所高校采用，成为国内大数据入门课程的经典参考书。

2. 开创性的在线课程与资源共享

他主导建设的“大数据技术原理与应用”慕课（MOOC）在多个平台上线，惠及数万名学生和社会学习者。配套的在线资源（如课件、实验指南、代码库）完全开源，极大降低了学习门槛，体现了教育普惠理念。

3. 实践导向的教学模式

林子雨强调动手能力培养，设计了完整的实验案例和Docker环境部署指南，帮助学生快速搭建大数据平台（如Hadoop、Spark）。他的课程注重产学研结合，常引入企业真实场景案例，提升学生解决实际问题的能力。

4. 通俗易懂的授课风格

学生评价其授课逻辑清晰、语言生动，擅长将复杂技术概念转化为通俗表达，尤其适合初学者入门。他注重知识框架的构建，帮助学生形成系统化的大数据思维。



1.4 大模型的分类

Sebastian Raschka博士 (Lightning AI的首席教育学家) :

将“推理”定义为**通过生成中间步骤来回答案复杂问题的过程**

2个简单的例子:

非推理问题:

“法国的首都是哪里?”

(答案直接、无需推导)

推理问题:

“一列火车以每小时60英里的速度行驶3小时，行驶距离是多少?”

(需先理解“距离=速度×时间”的关系，再分步计算)

通用的大语言模型 (LLM) 可能直接输出简短答案 (如“180英里”)

推理模型的特点在于显式展示中间推导过程



1.4 大模型的分类

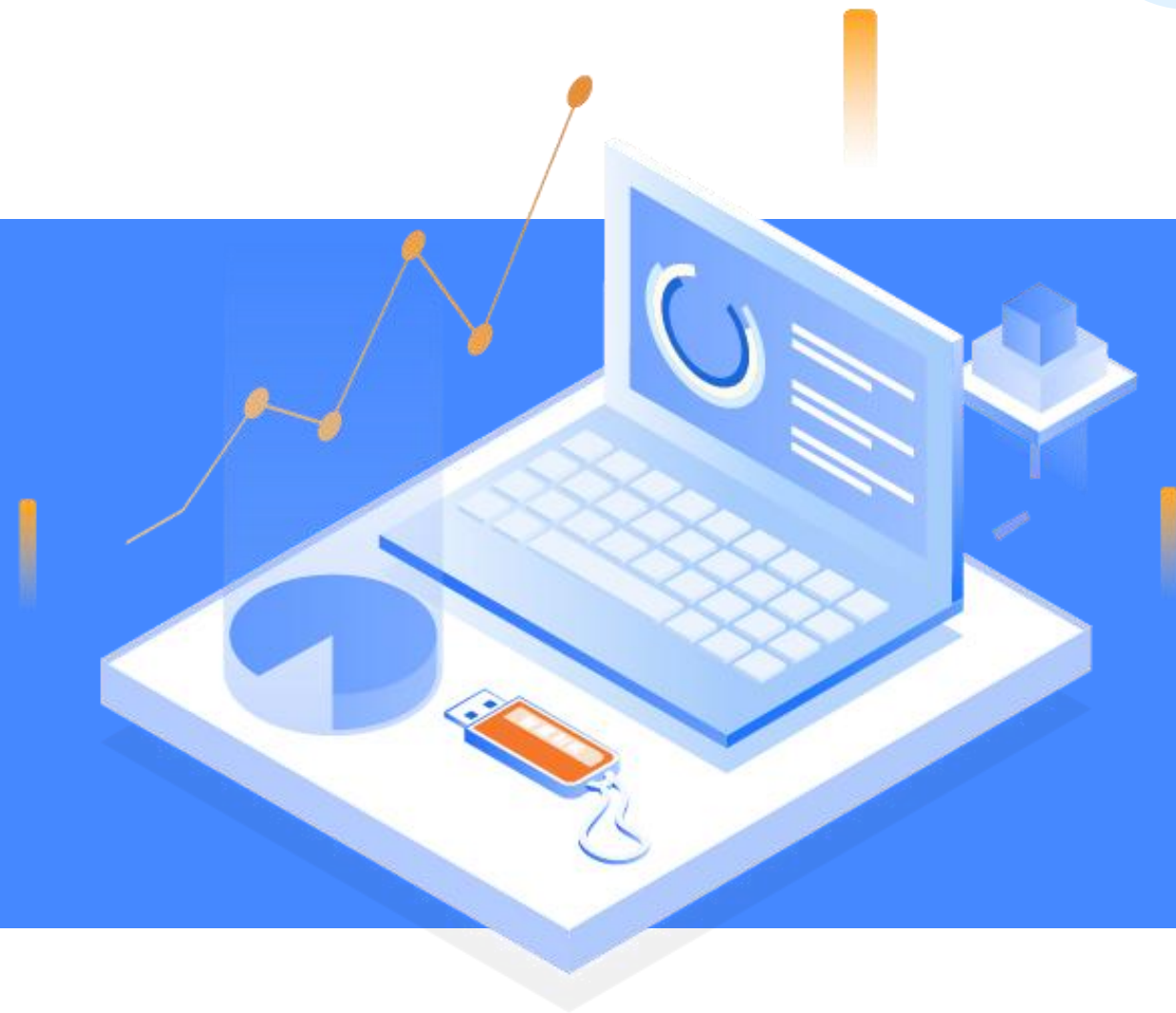
在应用方面二者各有擅长的领域，而不是简单的谁强谁弱问题

- 如果你需要完成数据分析、逻辑推理、代码生成等**逻辑性较强且较为复杂**的任务，请选择**推理大模型**
- 如果你面临创意写作、文本生成、意图识别等**发散性较强且较为创意多样**的任务，请选择**通用大模型**

特性	推理大模型	通用大模型
适用场景	复杂推理、解谜、数学、编码难题	文本生成、翻译、摘要、基础知识问答
复杂问题解决能力	优秀，能进行深度思考和逻辑推理	一般，难以处理多步骤的复杂问题
运算效率	较低，推理时间较长，资源消耗大	较高，响应速度快，资源消耗相对较小
幻觉风险	较高，可能出现“过度思考”导致的错误答案	较低，更依赖于已知的知识和模式
泛化能力	更强，能更好地适应新问题和未知场景	相对较弱，更依赖于训练数据
擅长任务举例	解决复杂逻辑谜题，编写复杂算法，数学证明	撰写新闻稿，翻译文章，生成产品描述，回答常识问题
成本	通常更高	通常更低

2. 大模型产品

- 2.1 国外的大模型产品
- 2.2 国内的大模型产品
- 2.3 主流大模型“幻觉”评测



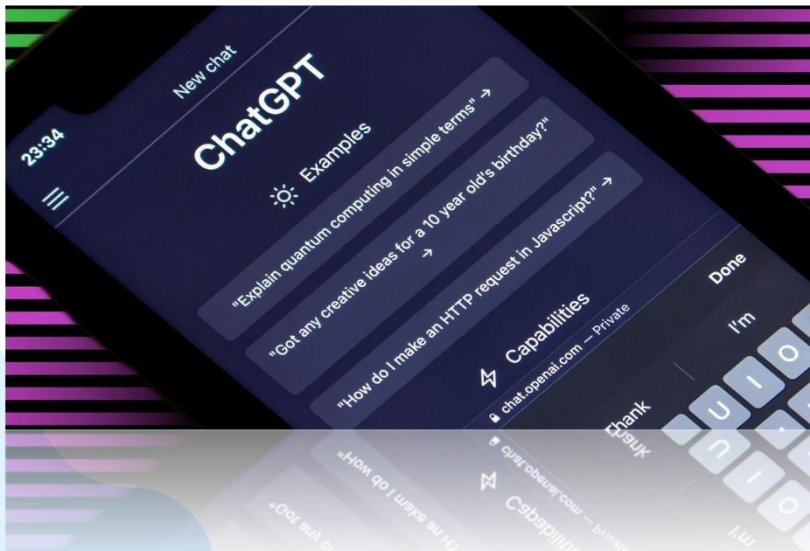


2.1 国外的大模型产品

■ ChatGPT

ChatGPT是一种由OpenAI训练的大语言模型。它是基于Transformer架构，经过大量文本数据训练而成，能够生成自然、流畅的语言，并具备回答问题、生成文本、语言翻译等多种功能

ChatGPT的应用范围广泛，可以用于客服、问答系统、对话生成、文本生成等领域。它能够理解人类语言，并能够回答各种问题，提供相关的知识和信息。与其他聊天机器人相比，ChatGPT具备更强的语言理解和生成能力，能够更自然地与人类交流，并且能够更好地适应不同的领域和场景。ChatGPT的训练数据来自互联网上的大量文本，因此，它能够涵盖多种语言风格和文化背景





2.1 国外的大模型产品

■ Gemini

Gemini是谷歌发布的大模型，它能够**同时处理多种类型的数据和任务，覆盖文本、图像、音频、视频等多个领域**。Gemini采用了全新的架构，将多模态编码器和多模态解码器两个主要组件结合在一起，以提供最佳结果

Gemini包括三种不同规模的模型：Gemini Ultra、Gemini Pro和Gemini Nano，适用于不同任务和设备。2023年12月6日，Gemini的初始版本已在Bard中提供，开发人员版本可通过Google Cloud的API获得。Gemini可以应用于Bard和Pixel 8 Pro智能手机。Gemini的应用范围广泛，包括**问题回答、摘要生成、翻译、字幕生成、情感分析**等任务。然而，**由于其复杂性和黑箱性质，Gemini的可解释性仍然是一个挑战**





2.1 国外的大模型产品

■ Sora

2024年2月16日，OpenAI再次震撼全球科技界，发布了名为Sora的文本生成视频大模型，只需输入文本就能自动生成视频。这一技术的诞生，**不仅标志着人工智能在视频生成领域的重大突破，更引发了关于人工智能发展对人类未来影响的深刻思考。**随着Sora的发布，人工智能似乎正式踏入了通用人工智能（AGI: Artificial General Intelligence）的时代。**AGI是指能够像人类一样进行各种智能活动的机器智能，包括理解语言、识别图像、进行复杂推理等。**Sora大模型能够直接输出长达60秒的视频，并且视频中包含了高度细致的背景、复杂的多角度镜头，以及富有情感的多个角色。这种能力已经超越了简单的图像或文本生成，开始触及到视频这一更加复杂和动态的媒介。**这意味着人工智能不仅在处理静态信息上越来越强大，而且在动态内容的创造上也展现出了惊人的潜力**





2.1 国外的大模型产品

■ Sora

右图是Sora根据文本自动生成的视频画面，一位戴着墨镜、穿着皮衣的时尚女子走在雨后夜晚的东京市区街道上，抹了鲜艳唇彩的唇角微微翘起，即便带着墨镜也能看到她的微笑，地面的积水映出了她的身影和灯红酒绿的霓虹灯，热闹非凡的唐人街正在进行舞龙表演，熙熙攘攘的人群目光都聚焦在跃动的彩龙身上，整个环境的喜庆氛围仿佛令人身临其境

提示词译文：

一位时尚女性走在布满温暖霓虹灯和动画城市标牌的东京街道上。她穿着黑色皮夹克、红色长裙和黑色靴子，拎着黑色钱包。她戴着太阳镜，涂着红色口红。她走路自信又随意。街道潮湿且反光，在彩色灯光的照射下形成镜面效果。许多行人走来走去。

Sora生成的视频：





2.1 国外的大模型产品

■ OpenAI o3

2024年12月20日，OpenAI发布推理模型o3，无论在软件工程、编写代码，还是竞赛数学、掌握人类博士级别的自然科学知识能力方面，o3都达到了很高的水平





2.2 国内的大模型产品

2025年1月国内大模型排行榜

大模型	图标	指标排名
DeepSeek		能力测评第一
豆包		用户数量第一
Kimi		文本处理第一
即梦AI		作图能力第一
通义万相		视频生成第一
智谱清言		文档归纳第一

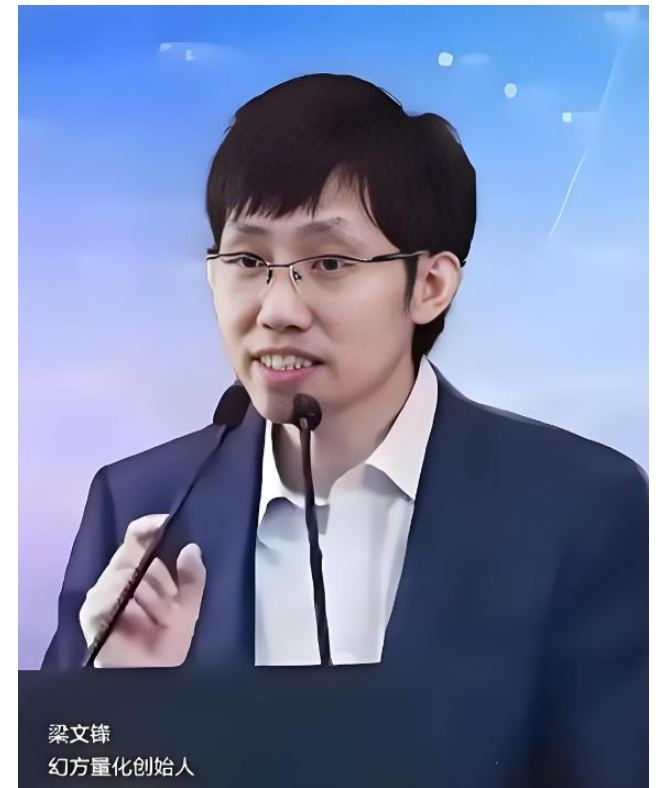


2.2 国内的大模型产品

■ DeepSeek (深度求索)

2024年12月26日，杭州一家名为“**深度求索**” (DeepSeek) 的中国初创公司，发布了全新一代大模型DeepSeek-V3。在多个基准测试中，DeepSeek-V3的性能均超越了其他开源模型，甚至与顶尖的闭源大模型GPT-4o不相上下，尤其在数学推理上，DeepSeek-V3更是遥遥领先。DeepSeek-V3以多项开创性技术，大幅提升了模型的性能和训练效率。DeepSeek-V3在性能比肩GPT-4o的同时，研发却只花了558万美元，训练成本不到后者的二十分之一。因为表现太过优越，DeepSeek在硅谷被誉为“来自东方的神秘力量”。

2025年1月20日，DeepSeek-R1正式发布，拥有卓越的性能，在数学、代码和推理任务上可与OpenAI o1媲美。



DeepSeek创始人 梁文峰



2.2 国内的大模型产品

■ 通义千问

通义千问是阿里云推出的一个**超大规模的语言模型**，它具备**多轮对话、文案创作、逻辑推理、多模态理解、多语言支持**的能力。通义千问这个名字有“通义”和“千问”两层含义，“通义”表示这个模型能够理解各种语言的含义，“千问”则表示这个模型能够回答各种问题。通义千问基于深度学习技术，通过对大量文本数据进行训练，从而具备了强大的语言理解和生成能力。它能够理解自然语言，并能够生成自然语言文本

■ 字节跳动豆包

豆包是字节跳动基于云雀模型开发的 AI，能理解你的需求并生成高质量回应。它知识储备丰富，涵盖历史、科学、技术等众多领域，无论是日常问题咨询，还是深入学术探讨，都能提供准确全面的信息。同时，具备出色的文本创作能力，能撰写故事、诗歌、文案等各类体裁。并且擅长语言交互，交流自然流畅，就像身边的知心伙伴，耐心倾听并给予恰当反馈。



豆包



2.2 国内的大模型产品

■ 文心一言

文心一言是由百度研发的知识增强大模型，能够与人对话互动、回答问题、协助创作，高效便捷地帮助人们获取信息、知识和灵感

文心一言基于飞桨深度学习平台和文心知识增强大模型，持续从海量数据和大规模知识中融合学习，具备知识增强、检索增强和对话增强的技术特色。

文心一言**具有广泛的应用场景**，例如智能客服、智能家居、移动应用等领域。它可以与用户进行自然语言交互，帮助用户解决各种问题，提供相关的知识和信息



■ Kimi

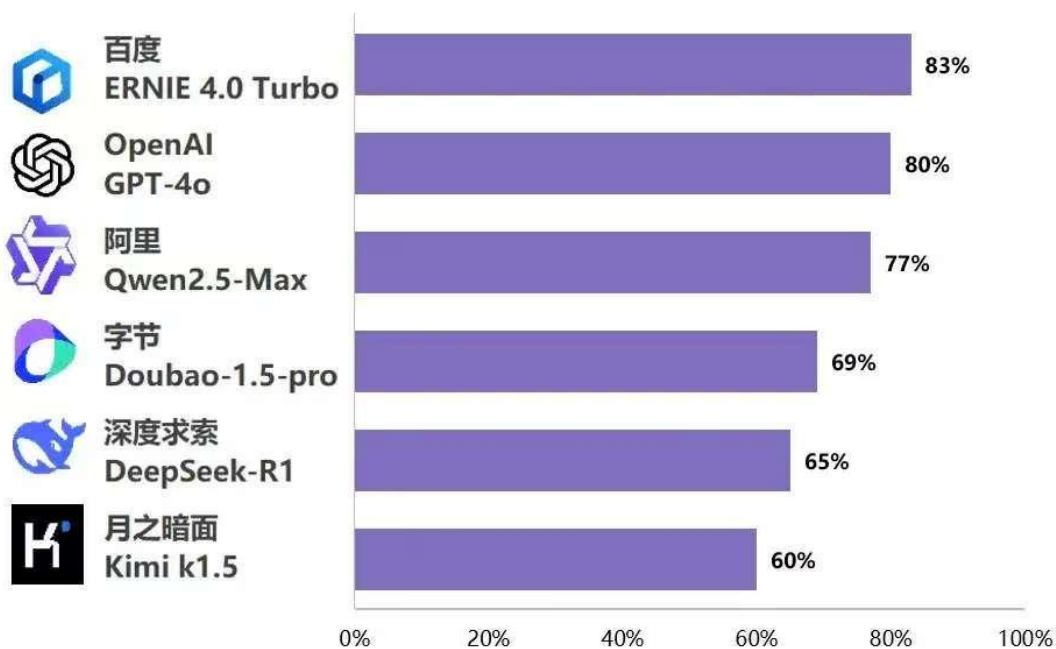
Kimi是月之暗面科技 2023 年推出的 AI 助手，可处理 200 万字超长文本，支持多格式文件解读、互联网信息搜索整合、多语言对话等，能用于办公、学习、创作等场景，有网页版、APP、微信小程序等使用方式。





2.3 主流大模型“幻觉”评测

主流大语言模型非幻觉率排行



■ 非幻觉率：AI生成的回答中不包含幻觉的比例
即回答真实、准确并与正确答案一致的内容所占的比例

来源：根据公开资料测评

大模型幻觉，也被称为 AI 幻觉，是指大型语言模型在生成内容时，产生与事实不符、逻辑错误或无中生有等不合理信息的现象。

比如在回答历史事件时，可能会编造不存在的细节或人物；在进行科学知识讲解时，给出错误的理论或数据。

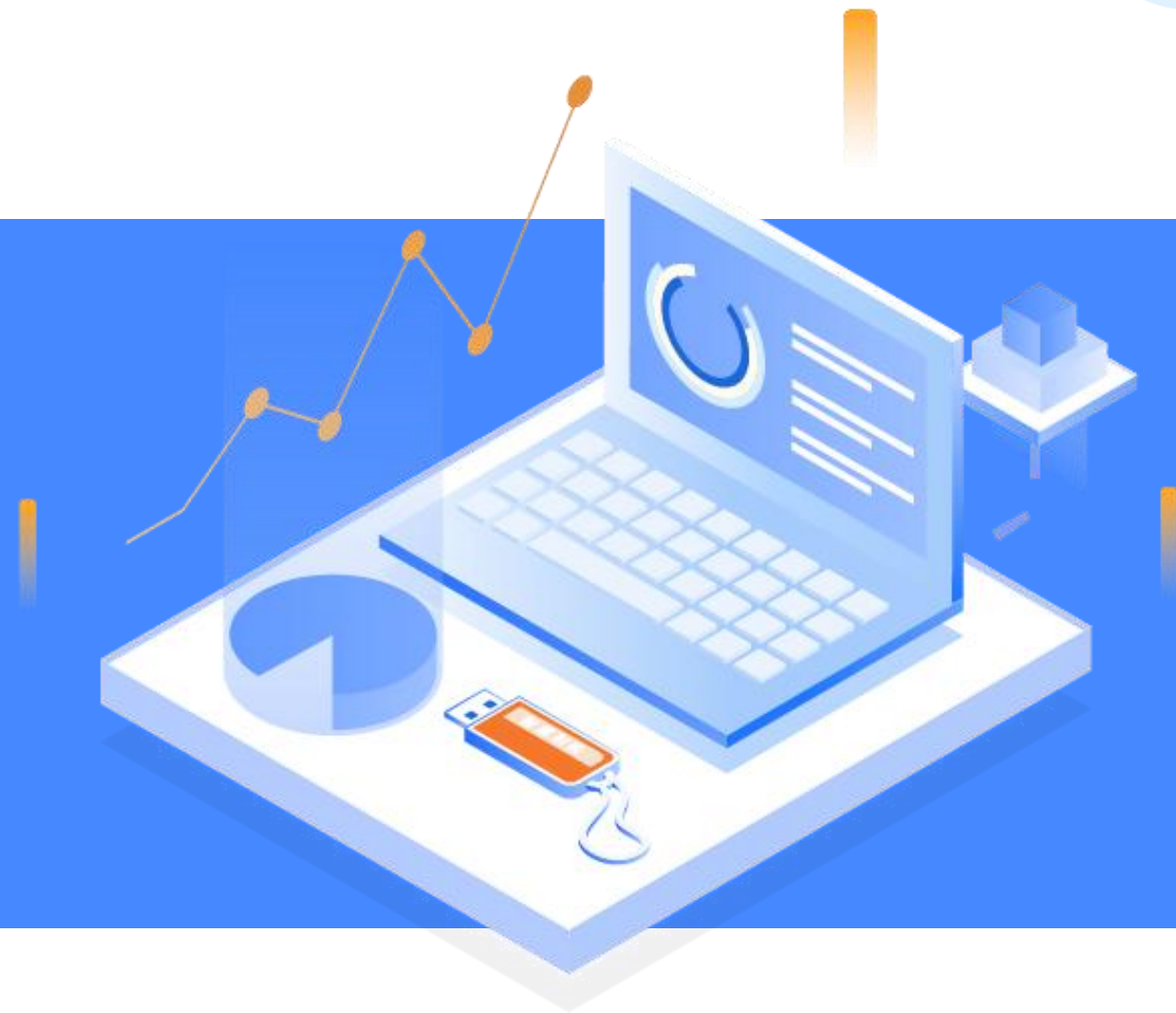
其产生原因主要包括：模型训练数据存在偏差、不完整或错误，导致在学习过程中引入了不准确的信息；模型基于概率分布生成内容，在某些情况下会选择一些看似合理但实际错误的路径。大模型幻觉会影响信息的准确性和可靠性，在信息传播、学术研究等领域可能带来不良影响。**因此，在使用大模型时，需要对其输出内容进行仔细验证和甄别。**

3. 大模型的行业应用

3.1 大模型的行业应用领域

3.2 DeepSeek大模型的应用场景

3.3 大模型与其他技术在企业中的融合应用





3.1 大模型的行业应用领域

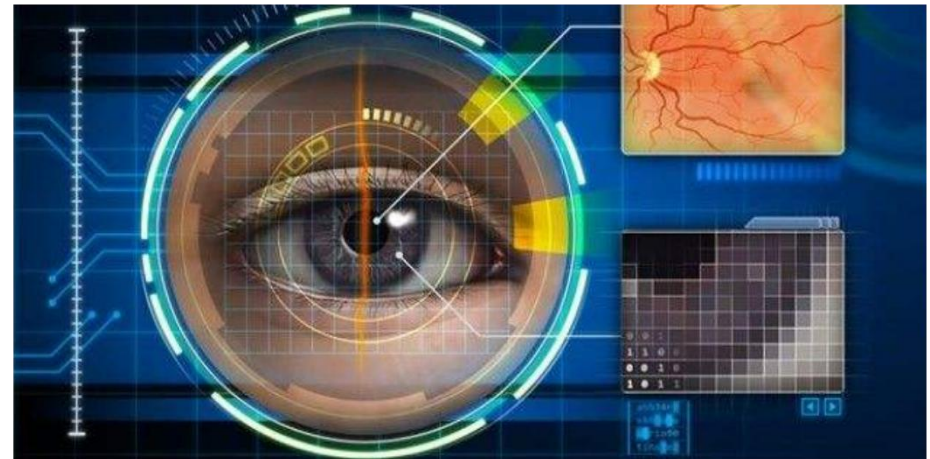
大模型的应用领域非常广泛，涵盖了自然语言处理、计算机视觉、语音识别、推荐系统、医疗健康、金融风控、工业制造、生物信息学、自动驾驶、气候研究等多个领域

(1) 自然语言处理

大模型在自然语言处理领域具有重要的应用，可以用于文本生成（如文章、小说、新闻等的创作）、翻译系统（能够实现高质量的跨语言翻译）、问答系统（能够回答用户提出的问题）、情感分析（用于判断文本中的情感倾向）、语言生成（如聊天机器人）等

(2) 计算机视觉

大模型在计算机视觉领域也有广泛应用，可以用于图像分类（识别图像中的物体和场景）、目标检测（能够定位并识别图像中的特定物体）、图像生成（如风格迁移、图像超分辨率增强）、人脸识别（用于安全验证和身份识别）、医学影像分析（辅助医生诊断疾病）等





3.1 大模型的行业应用领域

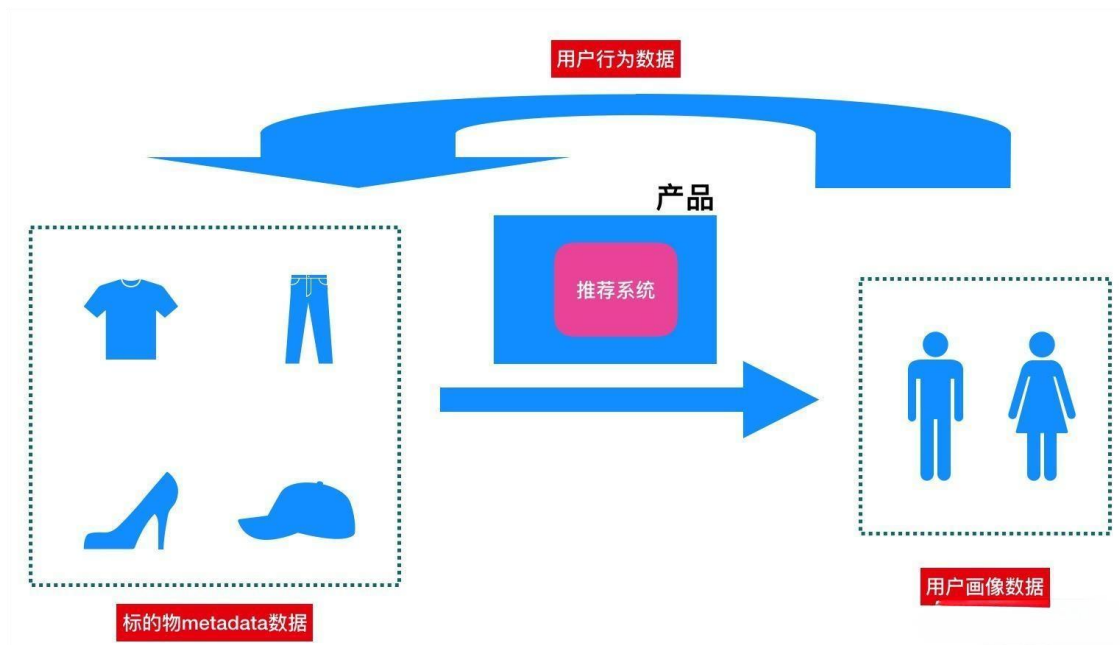
(3) 语音识别

大模型在语音识别领域也有应用，如语音识别、语音合成等。通过学习大量的语音数据，大模型可以实现高质量的跨语言翻译和语音识别以及生成自然语音



(4) 推荐系统

大模型可以用于个性化推荐、广告推荐等任务。通过分析用户的历史行为和兴趣偏好，大模型可以为用户提供个性化的推荐服务，提高用户满意度和转化率





3.1 大模型的行业应用领域



自动驾驶

大模型可以用于自动驾驶中的感知、决策等任务。通过学习大量的驾驶数据，大模型可以实现对车辆周围环境的感知和识别，以及进行决策和控制，提高自动驾驶的安全性和效率



医疗健康

大模型可以用于医疗影像诊断、疾病预测等任务。通过学习大量的医学影像数据，大模型可以辅助医生进行疾病诊断和治疗方案制定，提高医疗水平和效率



金融风控

大模型可以用于信用评估、欺诈检测等任务。通过分析大量的金融数据，大模型可以评估用户的信用等级和风险水平，以及检测欺诈行为，提高金融系统的安全性和稳定性

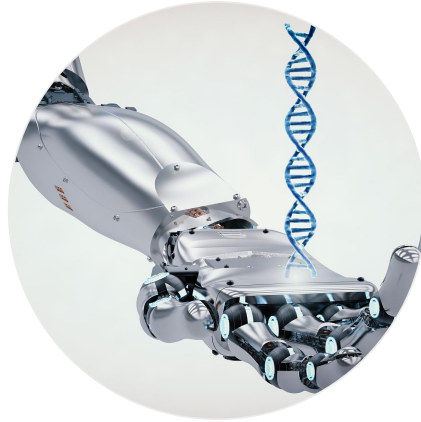


3.1 大模型的行业应用领域



工业制造

大模型可以用于质量控制、故障诊断等任务。通过学习大量的工业制造数据，大模型可以辅助工程师进行产品质量控制和故障诊断，提高生产效率和产品质量



生物信息学

在生物信息学领域，大模型可以用于基因序列分析（识别基因中的功能元件和变异位点）、蛋白质结构预测（推测蛋白质的二级和三级结构）、药物研发（预测分子与靶点的相互作用）等



气候研究

在气候研究领域，大模型可以处理气象数据，进行天气预测和气候模拟。它们能够分析复杂的气象现象，提供准确的气象预报，帮助人们做出应对气候变化的决策



3.1 大模型的行业应用领域

实例：百度大模型应用





3.2 DeepSeek大模型的应用场景

超级产品 增长1亿用户所用的时间



注：DeepSeek 包含网站Web/应用App累加不去重，Tiktok 不包含国内版抖音

数据来源：AI产品榜 aicpb.com 感谢邓瑞恒提供作图思路





3.2 DeepSeek大模型的应用场景

■ 客户服务与支持（高频应用）

- 1 自动回复客户咨询：电商平台使用DeepSeek自动回答客户关于订单状态、退货政策等问题
- 2 多渠道客服支持：企业使用DeepSeek在网站、社交媒体和邮件等多个渠道提供客服支持
- 3 自动处理订单：外卖平台使用DeepSeek自动处理用户订单和配送请求
- 4 自动处理退款：电商平台使用DeepSeek自动处理用户退款请求
- 5 自动处理投诉：电信公司使用DeepSeek自动处理客户关于网络和账单的投诉
- 6 自动处理咨询：银行使用DeepSeek自动处理客户关于账户和贷款的咨询
- 7 情感支持：心理健康应用使用DeepSeek为用户提供情感支持和心理辅导
- 8 智能语音客服：呼叫中心使用DeepSeek提供语音客服服务，减少人工客服压力
- 9 客户反馈分析：企业使用DeepSeek分析客户反馈，改进产品和服务
- 10 实时聊天支持：网站使用DeepSeek提供实时聊天支持，提升客户满意度



3.2 DeepSeek大模型的应用场景

■ 个性化推荐（高频应用）

- 01** 个性化购物推荐：电商平台使用DeepSeek根据用户购买历史推荐商品
- 02** 个性化音乐推荐：音乐平台使用DeepSeek根据用户听歌历史推荐音乐
- 03** 个性化电影推荐：流媒体平台使用DeepSeek根据用户观看历史推荐电影
- 04** 个性化书籍推荐：图书平台使用DeepSeek根据用户阅读历史推荐书籍
- 05** 个性化视频推荐：视频平台使用DeepSeek根据用户观看历史推荐视频
- 06** 个性化新闻推荐：新闻应用使用DeepSeek根据用户阅读历史推荐新闻
- 07** 个性化旅游推荐：旅游平台使用DeepSeek根据用户偏好推荐旅游目的地和行程
- 08** 个性化学习推荐：在线教育平台使用DeepSeek根据学生的学习进度推荐课程
- 09** 个性化内容推荐：内容平台使用DeepSeek根据用户兴趣推荐文章和博客
- 10** 个性化广告推荐：广告平台使用DeepSeek根据用户行为推荐个性化广告



3.2 DeepSeek大模型的应用场景

■ 教育与培训（中频应用）

- 01** 在线辅导学生：教育平台利用DeepSeek为学生提供数学、科学等科目的实时辅导
- 02** 自动批改作业：学校使用DeepSeek自动批改学生作业，节省教师时间
- 03** 个性化学习路径：在线教育平台使用DeepSeek为学生制定个性化学习路径
- 04** 语言学习助手：语言学习应用使用DeepSeek提供实时翻译和发音纠正
- 05** 虚拟实验室：教育机构使用DeepSeek创建虚拟实验室，帮助学生进行科学实验
- 06** 智能题库管理：教育平台使用DeepSeek管理题库，自动生成试卷
- 07** 学习进度跟踪：在线教育平台使用DeepSeek跟踪学生学习进度，提供反馈
- 08** 虚拟导师：教育机构使用DeepSeek提供虚拟导师服务，帮助学生解答问题
- 09** 职业培训：企业使用DeepSeek为员工提供职业培训课程
- 10** 在线考试监控：教育机构使用DeepSeek监控在线考试，防止作弊



3.2 DeepSeek大模型的应用场景

■ 医疗与健康（中频应用）

1 初步医疗建议：健康应用集成DeepSeek，为用户提供症状分析和就医建议

3 药物提醒：健康应用使用DeepSeek提醒用户按时服药

5 医疗数据分析：医院使用DeepSeek分析患者数据，提供个性化治疗方案

7 疾病预测：健康应用使用DeepSeek分析用户健康数据，预测疾病风险

9 智能诊断：医疗平台使用DeepSeek辅助医生进行疾病诊断

2 健康监测：智能穿戴设备使用DeepSeek监测用户健康数据，提供健康建议

4 心理健康支持：心理健康应用使用DeepSeek为用户提供心理辅导和支持

6 远程医疗咨询：医疗平台使用DeepSeek提供远程医疗咨询服务

8 医疗知识库：医院使用DeepSeek构建医疗知识库，方便医生查询信息

10 健康管理：健康应用使用DeepSeek帮助用户管理健康数据，制定健康计划



3.2 DeepSeek大模型的应用场景

■ 金融与投资（中频应用）

- 01** 市场趋势分析：投资平台利用DeepSeek分析股票市场数据，提供投资建议
- 02** 风险评估：银行使用DeepSeek评估客户信用风险，提供贷款建议
- 03** 智能投顾：投资平台使用DeepSeek为客户提供智能投资顾问服务
- 04** 欺诈检测：银行使用DeepSeek检测交易中的欺诈行为
- 05** 财务规划：金融应用使用DeepSeek帮助用户制定财务规划
- 06** 自动交易：投资平台使用DeepSeek进行自动化交易
- 07** 客户信用评分：银行使用DeepSeek评估客户信用评分，提供个性化金融服务
- 08** 金融数据分析：金融公司使用DeepSeek分析市场数据，提供投资建议
- 09** 智能客服：银行使用DeepSeek提供智能客服服务，解答客户问题
- 10** 财务报告生成：金融公司使用DeepSeek自动生成财务报告



3.2 DeepSeek大模型的应用场景

■ 内容创作与媒体（中频应用）

- 1 自动生成文章：**内容营销公司使用DeepSeek自动生成高质量的文章和博客内容
- 2 新闻摘要生成：**新闻应用使用DeepSeek自动生成新闻摘要，方便用户快速了解新闻
- 3 视频内容生成：**视频平台使用DeepSeek自动生成视频内容
- 4 社交媒体管理：**企业使用DeepSeek管理社交媒体内容，自动发布和回复
- 5 内容审核：**社交媒体平台使用DeepSeek自动审核用户发布的内容
- 6 智能写作助手：**作家使用DeepSeek辅助写作，提供灵感和建议
- 7 内容推荐：**内容平台使用DeepSeek根据用户兴趣推荐文章和视频
- 8 语音转文字：**媒体公司使用DeepSeek将采访录音转换为文字稿
- 9 图像识别：**社交媒体平台使用DeepSeek自动识别和标记照片中的物体和人脸
- 10 内容翻译：**媒体公司使用DeepSeek将内容翻译成多种语言



3.2 DeepSeek大模型的应用场景

■ 智能家居与物联网（低频应用）

- 01 智能家居控制：智能家居系统使用DeepSeek语音控制灯光、温度等设备
- 02 家庭安全监控：智能家居系统使用DeepSeek监控家庭安全，提供警报
- 03 智能家电管理：智能家居系统使用DeepSeek管理家电，优化能源使用
- 04 语音助手：智能音箱使用DeepSeek语音助手控制家居设备
- 05 家庭健康监测：智能家居系统使用DeepSeek监测家庭成员健康数据

- 06 智能照明：智能家居系统使用DeepSeek根据环境光线自动调节照明
- 07 智能温控：智能家居系统使用DeepSeek根据用户习惯自动调节温度
- 08 智能安防：智能家居系统使用DeepSeek提供智能安防服务，监控家庭安全
- 09 智能家电控制：智能家居系统使用DeepSeek控制家电，如洗衣机、冰箱等
- 10 家庭娱乐管理：智能家居系统使用DeepSeek管理家庭娱乐设备，如电视、音响等



3.2 DeepSeek大模型的应用场景

■ 法律与合规（低频应用）

1 合同审查：法律服务平台使用DeepSeek为用户提供合同审查服务

3 合规检查：企业使用DeepSeek检查业务流程是否符合法律法规

5 案件分析：律师事务所使用DeepSeek分析案件数据，提供法律建议

7 法律风险评估：企业使用DeepSeek评估法律风险，提供合规建议

9 法律案例检索：律师事务所使用DeepSeek检索相关法律案例

2 法律咨询：法律服务平台使用DeepSeek为用户提供法律咨询

4 法律文书生成：法律服务平台使用DeepSeek自动生成法律文书

6 法律知识库：法律服务平台使用DeepSeek构建法律知识库，方便用户查询信息

8 法律文书翻译：法律服务平台使用DeepSeek将法律文书翻译成多种语言

10 法律培训：法律服务平台使用DeepSeek为用户提供法律培训课程



3.2 DeepSeek大模型的应用场景

■ 游戏与娱乐（低频应用）

- 01** 游戏角色智能：游戏开发公司使用DeepSeek为游戏角色提供智能行为
- 02** 游戏内容生成：游戏平台使用DeepSeek自动生成游戏内容
- 03** 游戏推荐：游戏平台使用DeepSeek根据用户游戏历史推荐游戏
- 04** 游戏数据分析：游戏公司使用DeepSeek分析玩家数据，优化游戏体验
- 05** 虚拟现实体验：游戏平台使用DeepSeek提供虚拟现实游戏体验
- 06** 游戏语音识别：游戏平台使用DeepSeek提供语音识别功能，增强游戏互动
- 07** 游戏社交：游戏平台使用DeepSeek提供社交功能，增强玩家互动
- 08** 游戏内容审核：游戏平台使用DeepSeek自动审核用户生成的内容
- 09** 游戏虚拟助手：游戏平台使用DeepSeek提供虚拟助手，帮助玩家解决问题
- 10** 游戏市场分析：游戏公司使用DeepSeek分析市场数据，制定游戏策略



3.2 DeepSeek大模型的应用场景

■ 其他应用（低频应用）

- 01 自动驾驶：**汽车制造商使用DeepSeek开发自动驾驶系统，识别道路和障碍物
- 02 智能物流：**物流公司使用DeepSeek优化物流路线，提高配送效率
- 03 智能农业：**农业公司使用DeepSeek监测作物生长，优化种植方案
- 04 智能交通：**交通管理部门使用DeepSeek优化交通流量，减少拥堵
- 05 智能环保：**环保机构使用DeepSeek监测环境数据，提供环保建议
- 06 智能能源管理：**能源公司使用DeepSeek优化能源使用，提高能源效率
- 07 智能城市规划：**城市规划部门使用DeepSeek分析城市数据，优化城市规划
- 08 智能零售：**零售公司使用DeepSeek优化库存管理，提高销售效率
- 09 智能招聘：**招聘平台使用DeepSeek自动筛选简历，匹配职位需求
- 10 智能数据分析：**企业使用DeepSeek分析大数据，提供商业洞察



3.3 大模型与其他技术在企业中的融合应用

自动化流程技术 (RPA)

实现更加智能的业务流程自动化。RPA 擅长处理重复性的规则性任务，而大模型则能够理解自然语言指令，深入分析复杂的业务场景，并提供极具价值的决策建议。比如在财务部门，RPA 可以自动完成发票录入等任务，而大模型则能够回答员工关于财务政策的问题，并依据数据分析给出成本优化的建议

知识图谱技术

为企业提供更为精准的知识管理和决策支持。知识图谱以结构化的方式呈现企业内外部的知识，大模型则可以利用知识图谱进行知识推理和查询，迅速获取所需信息。在研发部门，大模型结合知识图谱，可以回答技术问题、推荐相关技术方案，助力研发人员实现创新

物联网技术

对设备和生产过程进行智能监控和管理。物联网设备会产生大量的数据，大模型能够对这些数据进行分析，预测设备故障，优化生产流程。在制造业中，大模型可以根据传感器数据实时调整生产参数，提高生产效率和产品质量

智能语音技术

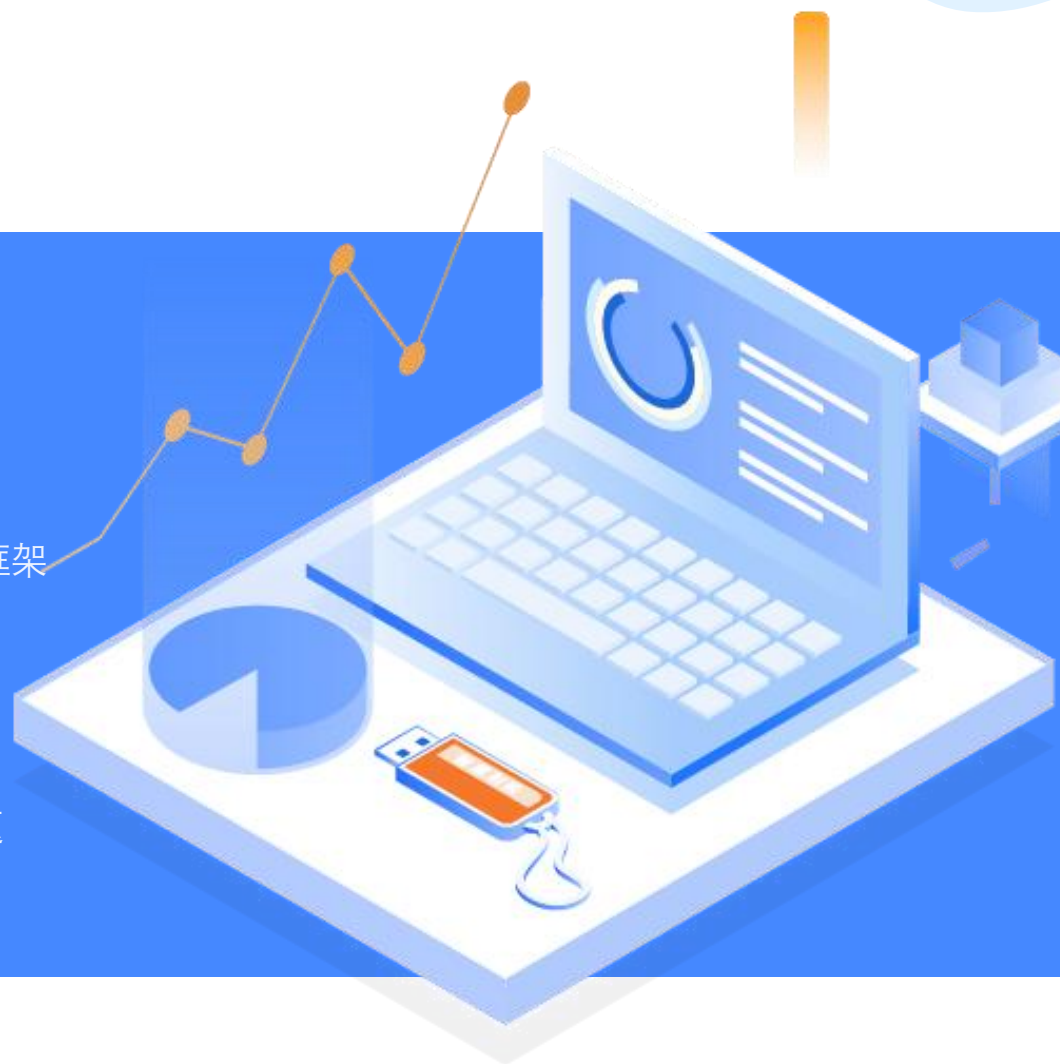
单击此处添加文本具体内容，简明扼要地阐述您的观点。根据需要可酌情增减文字，以便观者准确地理解您传达的思想。单击此处添加文本具体内容

数据分析与可视化技术

帮助企业更好地理解 and 利用数据。大模型可以对大量的数据进行分析 and 挖掘，提取有价值的信息，而数据分析和可视化技术则可以将这些信息以直观的图表和报告形式展示出来，方便企业管理层进行决策

4. 企业大模型落地方案

- 4.1 企业大模型如何为企业赋能
- 4.2 企业大模型的七大落地场景
- 4.3 企业大模型的部署方式
- 4.4 为什么需要本地部署大模型
- 4.5 企业部署大模型规划路线
- 4.6 企业级应用集成AI大模型的关键步骤
- 4.7 企业级应用集成大模型的技术架构设计
- 4.8 大模型应用落地总体“四维认知”框架
- 4.9 大模型应用落地策略建议
- 4.10 大模型构建的技术方案选择
- 4.11 本地部署大模型初级方法
- 4.12 本地部署大模型的成本
- 4.13 企业部署大模型当前关注的问题
- 4.14 企业部署大模型面临的挑战





4.1 企业大模型如何为企业赋能

提高效率

企业AI大模型可以自动化地完成一些重复性、繁琐的工作，从而提高工作效率。比如，它可以通过自然语言处理技术自动回复客户邮件、自动分类垃圾邮件等等

优化业务流程

企业AI大模型可以根据历史数据和实时数据来优化业务流程。比如，它可以通过预测分析来提前发现潜在的问题，并采取相应的措施来避免这些问题的发生

提高产品质量

企业AI大模型可以帮助企业更好地理解客户需求，从而提高产品质量。比如，它可以通过自然语言处理技术来分析客户反馈意见，并针对性地改进产品设计

降低成本

企业AI大模型可以通过自动化地完成一些工作来降低成本。比如，它可以通过自动化的数据分析来减少人力资源的使用



4.2 企业大模型的七大落地场景

对话助手

对话助手是大模型技术最基础也是最广泛的应用形态。基于知识库数据源和API查询，对话助手能够实时回答用户问题，提供高效的信息传递。例如，某一乘用车通过搭建基于AI大模型的汽车在线问答平台，实现了对非结构化文档（如PDF、Word等）的智能知识交互，提升了员工工作效率和学习能力。该平台提供7*24小时的智能知识服务，问题交互准确率达95%以上，客户满意度提升了35%

报告生成

企业常常需要撰写大量报告，这些报告往往具有固定模板且内容复杂。大模型技术可以通过分解报告为多个子问题，分别求解，并将结果整合到预定框架中，实现报告的自动生成。虽然这类应用不能保证100%的准确性，但结合多种技术手段（如大型和小型模型、专家系统等），可以大大提高生成效率和准确性

审核与合规

审核类场景通常与生成类场景相伴相生。大模型技术不仅可以辅助生成文档，还可以用于审核结构化业务数据和非结构化文档。例如，在提交审批前，通过大模型进行自动审核，可以帮助提交者判断是否需要修改，提高提交质量，降低打回概率。此外，大模型还可以用于合规性和合法性审核，确保企业业务符合法律法规要求



4.2 企业大模型的七大落地场景

知识管理

大模型技术正在重塑企业知识管理场景。通过大模型对语义的理解，企业可以将非结构化数据进行结构化提取和总结，大大降低了知识管理的落地门槛

编码助手

在软件开发领域，大模型技术可以作为编码助手，提升工程师的工作效率。通过AI编程助手，减少了重复劳动，提升了代码质量

数据分析

数据分析是企业决策的重要依据。大模型技术通过引入编程能力，可以大大降低数据分析成本。然而，目前大模型在特定业务场景下的效果仍需优化，模型微调是一种解决办法

智能运维与制造优化

大模型技术在智能运维和制造优化方面也发挥着重要作用。利用大模型技术构建了智能运维系统，通过自然语言交互界面，实现了对运维问题的智能回答。在制造领域，大模型可以通过智能分析和预测维护，提升生产效率



4.3 企业大模型的部署方式

云端部署

通过云厂商实现弹性扩展和成本优化

本地/内网部署

通过本地数据中心实现数据完全掌控

边缘部署

通过边缘节点实现低延迟和实时处理

混合部署

通过本地和云端结合实现灵活性和灾备需求

容器化/微服务部署

通过容器技术和微服务架构实现敏捷开发和资源隔离

联邦部署

通过联邦协议和分布式架构实现跨组织协作和数据隐私保护



4.3 企业大模型的部署方式

云端部署

DeepSeek大模型部署在公有云或私有云上，利用云厂商的基础设施和资源。适用场景：

- 弹性需求：需要根据负载动态调整资源
- 快速扩展：业务增长迅速，需快速扩展系统
- 成本优化：希望通过按需付费模式降低IT成本

本地/内网部署

DeepSeek大模型部署在企业内部服务器或数据中心，数据和应用完全在企业内网中运行。适用场景：

- 数据敏感：对数据安全要求高，需完全掌控数据
- 合规要求：需满足特定行业或地区的合规要求
- 网络限制：内网环境无法连接外部网络



4.3 企业大模型的部署方式

边缘部署

将DeepSeek大模型部署在靠近数据源的边缘节点，减少数据传输延迟。适用场景：

- **低延迟需求：**如物联网、实时监控等需要快速响应的场景
- **带宽有限：**数据传输成本高或带宽有限时，边缘计算可减少数据上传
- **离线运行：**需要在网络不稳定或离线时仍能正常运行

混合部署

结合云端和本地部署，DeepSeek大模型部分系统在云端，部分在本地。适用场景：

- **灵活需求：**部分数据需本地处理，部分需云端处理
- **过渡阶段：**从本地逐步迁移到云端时，混合部署可作为过渡方案
- **灾备需求：**本地和云端互为备份，提升系统可靠性



4.3 企业大模型的部署方式

容器化/微服务部署

将DeepSeek大模型系统拆分为多个微服务，使用容器技术（如Docker）进行部署和管理。适用场景：

- 敏捷开发：需要快速迭代和发布新功能
- 资源隔离：不同服务需要独立运行环境，避免相互干扰
- 弹性扩展：根据需求独立扩展特定服务

联邦部署

DeepSeek大模型多个独立系统通过联邦协议协作，共享数据和资源，但各自保持独立。适用场景：

- 跨组织协作：多个组织需共享数据但保持独立管理
- 数据隐私：需在保护数据隐私的前提下进行数据共享
- 分布式计算：需要在多个节点上分布式处理数据，如联邦学习



4.3 企业大模型的部署方式

硅基流动推出一站式大模型 API 云服务平台 SiliconCloud, 集成了阿里通义大模型 Qwen2、智谱 AI 的 GLM - 4、幻方量化的 DeepSeek - R1 系列开源模型等众多主流大模型, 为开发者提供高效能、低成本的多品类 AI 模型服务, 开发者可直接调用平台 API, 无需自建服务, 大大节省了搭建和维护模型服务的时间和成本

与**华为云**团队合作, 联合首发并上线基于华为云昇腾云服务的 DeepSeek R1/V3 推理服务, 支持部署的 DeepSeek 模型能达到与全球高端 GPU 部署模型持平的效果, 并提供稳定的生产级服务能力。还与智谱 AI 等厂商合作模型广场, 客户可一键调用超百个垂直领域模型



大模型云端部署参考方案



4.4 为什么需要本地部署大模型





4.5 企业部署大模型规划路线

数据治理

知识库

大模型



数据治理——给企业数据“立规矩”

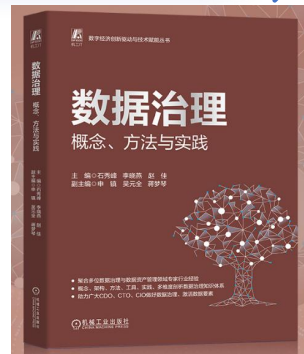
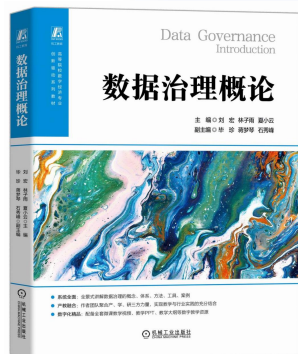


DeepSeek 大模型 —— 让企业拥有“数字大脑”

第一阶

第二阶

第三阶



厦大团队两本数据治理书籍



知识库——把员工经验变成企业资产



4.5 企业部署大模型规划路线

短期 (0-6月)

构建企业内部知识库
(RAG系统)

中期 (6-18月)

开发行业专属大模型
(模型微调)

长期 (18月+)

搭建AI Agent workflow
(企业智能体)





4.6 企业级应用集成AI大模型的关键步骤

需求场景定义

企业首先需要明确AI应用的具体场景，如文本生成、情感分析、图片理解和生成等。这有助于后续模型的选择与技术路线的设计

模型选型

根据需求场景，企业需要在众多预训练模型中选择合适的AI大模型。这一过程中，企业需要综合考虑模型的性能、准确性、计算效率、成本及安全性等因素

模型优化

选定模型后，企业需要通过提示词工程、RAG、微调等策略对模型进行优化，以提升其在特定场景下的表现效果与可靠性

模型部署

优化后的模型需要被部署到合适的环境中，如云端、边缘计算或本地服务器等。这要求企业具备强大的技术实力和丰富的运维经验

应用集成

最后，企业需要将AI大模型无缝集成到现有的业务系统中，实现自动化决策或增强用户体验



4.7 企业级应用集成AI大模型的技术架构设计

企业级应用集成AI大模型的技术架构设计，需要综合考虑业务需求、数据安全、模型性能等多方面因素。以下是一个典型的技术架构设计框架：

数据源层

包括内部数据库、公共数据集、第三方数据服务等，为AI大模型提供丰富的数据支持

数据采集与预处理层

使用爬虫、API、IoT设备等手段收集数据，并进行清洗、格式化、归一化、特征提取等预处理操作

模型训练与评估层

使用机器学习框架（如TensorFlow、PyTorch等）训练模型，并通过交叉验证、A/B测试等方法评估模型性能

模型部署与服务层

将训练好的模型部署为RESTful API等服务，使用Docker、Kubernetes等工具进行容器化与编排，实现高效的模型管理和服务

业务逻辑与用户界面层

将AI模型集成到业务流程中，通过Web应用、移动应用等用户界面提供交互服务

安全与合规层

实施加密、访问控制等手段保护数据安全，确保AI应用符合行业标准和国家法规



4.8 大模型应用落地总体“四维认知”框架

问题诊断

直面行业转型现状与痛点在启动大模型项目时，必须深入分析行业转型的阶段特征、业务体系的现实需求以及技术架构的短板痛点，找准问题，避免照搬照抄

能力评估

诊断短板，明确能力提升路径找准问题仅仅是第一步，还需要进一步诊断企业在大模型应用方面的短板，给出补齐相关能力的“药方”

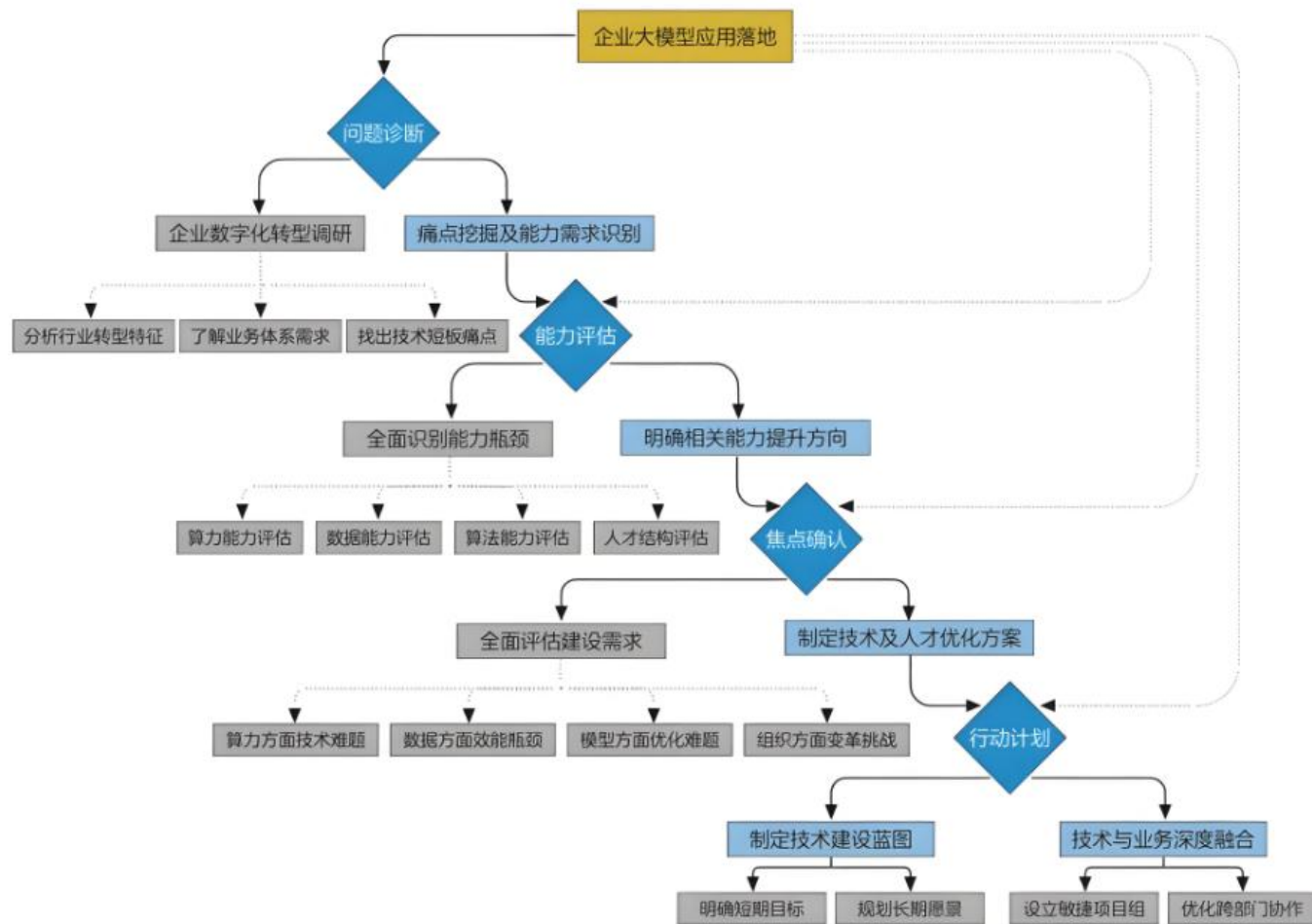
焦点确认

聚焦待补齐能力背后的技术议题大模型能力的提升并非一蹴而就，需要在算力基础、数据积累、模型优化等方面持续发力

行动计划

制定技术建设路线图，有的放矢地行动

梳理清楚大模型落地所需的能力体系只是第一步，接下来还需要以行动为指引，将愿景变为现实





4.9 大模型应用落地策略建议

聚焦业务场景需求 并合理选择模型

- 深入剖析业务需求，首选大语言模型提升文本处理效率；多模态业务需权衡技术成本，采用合适架构融入系统

评估自身数据 丰富度及质量

- 评估数据丰富度、多样性、时效性和隐私保护要求，确保数据支持模型训练微调，符合合规要求

建立持续学习 与迭代机制

- 建立性能监控和反馈机制，定期评估模型效果，根据需求和反馈调优迭代，确保模型持续满足业务需求

探索大模型应用与 业务的深度融合

- 从辅助生成场景拓展到为决策赋能，推动数字化转型和智能化升级，实现业务流程自动化智能化

明确任务性质结合 业务逻辑优化流程

- 区分辅助生成与决策性任务，优化输入输出流程（辅助生成任务）或结合模型预测与决策逻辑（决策性任务）

明确技术选型 与适配性

- 综合评估框架成熟度、易用性、扩展性和兼容性等因素，选择合适大模型框架并确保与现有系统高效集成

培养独立的 AI人才与团队

- 引进培养AI专业人才，负责模型微调、技术引入和系统集成等工作，支撑大模型应用发展



4.10 大模型构建的技术方案选择

模型微调还是本地知识库?

模型微调和本地知识库

使用海量数据进行预训练得到的基础大模型，具备广泛的语言理解和生成能力，但在特定任务上的表现往往不够精准。

解决方案：

- (1) 模型微调；
- (2) 本地知识库

模型微调技术特点

- (1) **领域针对性强**：经过微调的模型在特定领域的表现会有显著提升，能够更好地理解和处理该领域的专业问题；
- (2) **模型适应性优化**：通过微调可以调整模型的参数，使其更符合特定任务的要求，提高输出的准确性和稳定性。

模型微调技术要点

- (1) **高质量的标注数据**：标注数据的质量直接影响微调的效果，需要确保数据标注的准确性和一致性。
- (2) **合理的微调策略**：选择合适的微调算法和超参数，避免过拟合或欠拟合问题。



4.10 大模型构建的技术方案选择

模型微调主要是指令微调，指令微调包括两个阶段：**监督微调**和**强化学习**

监督微调

在监督微调阶段，模型会学习一个指令-响应 (Instruction-Response) 数据集，该数据集包含大量人类编写的任务示例，例如“请解释相对论的基本概念”及其标准答案。通过这种方式，模型能够理解不同类型的任务并提供符合预期的回答。**指令-响应 (Instruction-Response) 数据集用于训练模型理解任务指令并生成符合预期的响应**

强化学习

在微调完成后，部分高级模型还会使用强化学习进行优化。例如，ChatGPT 和 Claude 使用人类反馈强化学习 (RLHF) 让模型的回答更符合用户期望，更好地选择符合人类偏好的答案

不足之处：

数据准备成本高

需要收集、整理和标注大量特定领域的数据，这是一个耗时费力的过程

时效性问题

对知识更新频繁的领域，微调后的模型可能很快会过时，需要不断重新训练

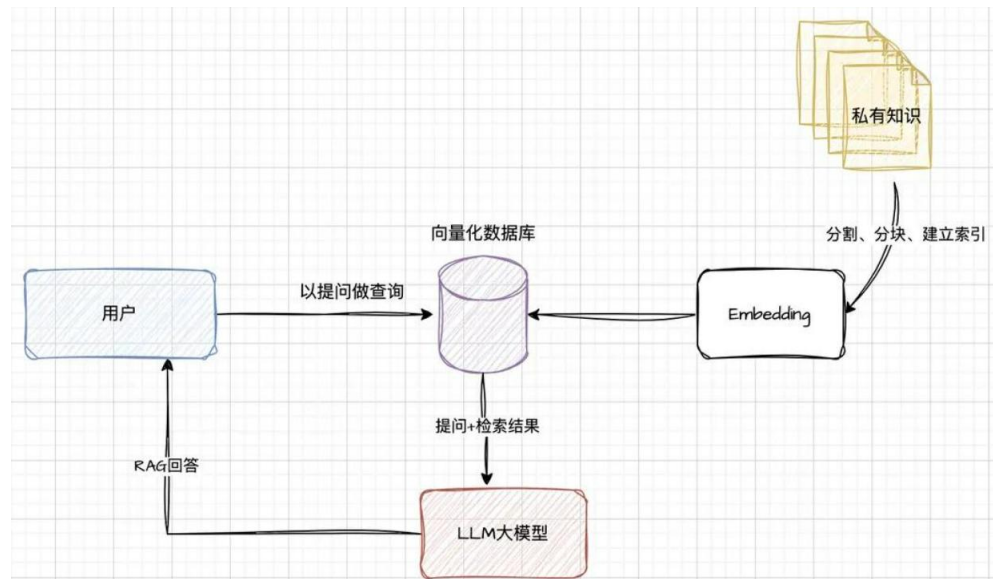
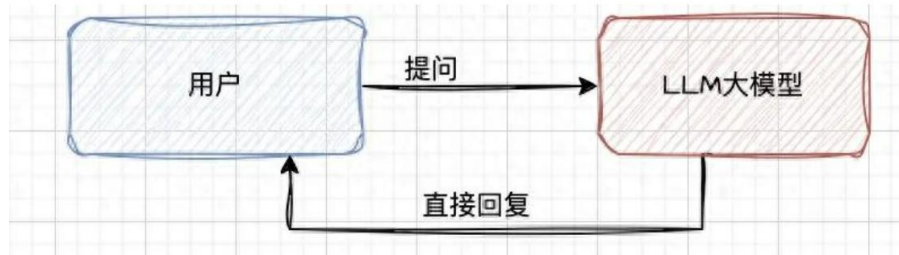


4.10 大模型构建的技术方案选择

■ 本地知识库

RAG (Retrieval-Augmented Generation) , 即检索增强生成, 是一种结合检索技术和生成模型的技术框架, 旨在提升模型生成内容的准确性和相关性

其核心思想是: 在生成答案前, 先从外部知识库中检索相关信息, 再将检索结果与用户输入结合, 指导生成模型输出更可靠的回答。简单地说, 就是利用已有的文档、内部知识生成向量知识库, 在提问的时候结合库的内容一起给大模型, 让其回答的更准确, 它结合了信息检索和大模型技术





4.10 大模型构建的技术方案选择

■ 本地知识库的优点

保护数据隐私

由于日常的业务知识是保存到本地的，所以减少信息泄露的风险

减少模型幻觉

由于提问结合了业务知识，所以减少了模型的幻觉，即减少了模型的胡说八道

实时知识补充

模型的回复结合了业务知识和实时知识，所以实时性可以更好

无需重新训练

不用重新训练模型，微调模型降低了成本



4.11 本地部署大模型初级方法

步骤

安装Ollama

下载DeepSeek R1

运行DeepSeek R1

使用Open WebUI增强交互体验



Get up and running with large language models.

Run Llama 3.3, DeepSeek-R1, Phi-4, Mistral, Gemma 2, and other models, locally.

Download ↓

Available for macOS, Linux, and Windows

```
C:\Windows\system32\cmd.exe - ollama run deepseek-r1:1.5b
Microsoft Windows [版本 10.0.22621.4]
(c) Microsoft Corporation. 保留所有权利。

C:\Users\ziyul>ollama run deepseek-r1:1.5b
>>> 请问如何学习人工智能
<think>
嗯，我现在要学人工智能。听起来挺复杂的，但我得一步步来。首先，我应该了解什么是人工智能（AI）。听说它就是让计算机能够模拟人脑的思维活动，比如识别、理解、决策等。对吧？那我得先明白AI的基本概念和目标。
```

ollama deepseek-r1:1.5b

+ 请介绍如何学习人工智能

建议

Grammar check
rewrite it for better readability

Give me ideas
for what to do with my kids' art

Overcome procrastination
give me tips

具体安装过程请参考厦门大学数据库实验室博客 <https://dbl原因lab.xmu.edu.cn/blog/5816/>



4.12 本地部署大模型的成本

DeepSeek R1 671B (满血版) 部署成本

1. 硬件采购成本

- 服务器集群：含8张NVIDIA A100/H100显卡的服务器，市场价格约80-120万元
- 配套设备：液冷系统、冗余电源等附加成本约15-25万元

2. 运维成本

- 电费：满载功耗约6000W，年电费约5-8万元（按工业电价1.2元/度计算）
- 维护：专业工程师团队年成本约30-50万元





4.12 本地部署大模型的成本

- 2025年2月10日，清华大学AI团队发布KTransformers开源项目迎来重大更新，成功打破大模型推理算力门槛。此前，拥有671B参数的MoE架构大模型DeepSeek-R1在推理时困难重重。推理服务器常因高负荷宕机，专属版云服务器按GPU小时计费的高昂成本让中小团队无力承担，而市面上的“本地部署”方案多为参数量大幅缩水的蒸馏版，在本地小规模硬件上运行满血版 DeepSeek-R1 被认为几乎不可能
- 此次KTransformers项目更新带来重大突破，支持在24G显存（4090D）的设备上本地运行 DeepSeek-R1、V3的671B满血版。其预处理速度最高可达286 tokens/s，推理生成速度最高能达到14 tokens/s。甚至有开发者借助这一优化技术，在3090显卡和200GB内存的配置下，使Q2_K_XL模型的推理速度达到9.1 tokens/s，实现了千亿级模型的“家庭化”运行

传统方案：8卡A100服务器成本超**百万元**，按需计费每小时数千元

清华方案：单卡RTX 4090方案，整机成本约**2万元**，功耗80W





4.13 企业部署大模型当前关注的问题

收益不确定：当前很多应用场景多处于验证阶段，其实际效果和最终收益存在不确定性

落地和建设路径：第一，是模型本身，研发能力不足的企业用户，是否自身要去做模型训练？现阶段是否一定要有企业内部专属大模型？第二，如果做训练和微调，现有的数据量是否足够？第三，算力是否需要投入？如果增加算力投入，整体项目预算量级会上升一个级别。第四，如何量化最终收益？第五，是项目可行性问题，大多数企业用户在过去一年都已经做过大模型尝试，准确性、幻觉问题一直存在，如何解决？

合规：模型本身是否自主可控？数据是否安全合规？信创要求。

选型：现在市面上第一类大模型由互联网大厂，像阿里、腾讯、华为等大厂研发的相关产品。那第二类是专注单点应用的小厂，比如专注于 Agent 平台和 Agent 应用开发。第三，现有垂直供应商。所有企业用户都会关注选型问题，虽然大模型大厂本身技术能力很强，但不一定特别理解企业自身业务场景，而现有的垂直厂商本身技术能力会受到多方质疑，所以选型是重点关注的问题。



4.14 企业部署大模型面临的挑战

AI大模型正在重构千行万业，造就数据的黄金时代，然而AI进入各行业仍面临着挑战：

从通用大模型到行业场景大模型，需要进行针对性训练，训练所需数据预处理耗时长，收集、清洗等环节占模型开发训练时长60%，需要高效归集、管理的数据工程能力

其次，行业场景模型训练和应用落地难，项目开发难度大，人员技术要求高，开发周期不可控，需要环境易安装、模型易训练、应用易搭建的AI业务平台

最后，因算力等待、任务潮汐、资源碎片化等原因，AI集群可用度往往不足50%，需要不断提升全系统调度效率

大模型企业落地解决方案供应商（比如DeepSeek一体机厂商）可以提供支持：

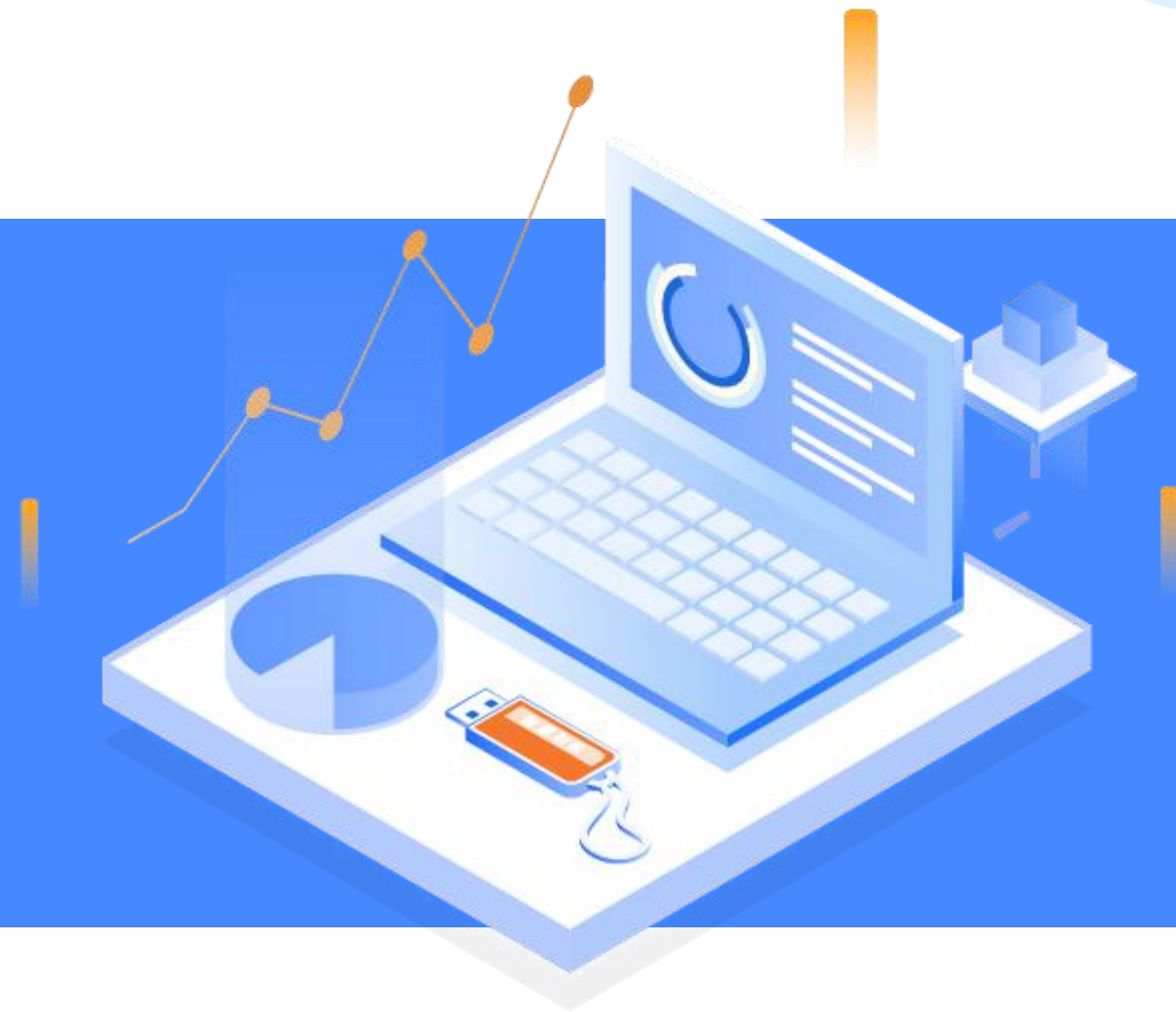
一
数据工程工具化，缩短80%训练数据准备周期

二
系统级模型训练与推理加速能力，实现模型训练周期缩短30%，推理并发提升一倍

三
极简应用开发平台，支撑非专业开发者快速上手

5. 智能体的企业应用

- 5.1 智能体的概念
- 5.2 智能体和RAG的区别
- 5.3 典型的智能体类型
- 5.4 智能体产品
- 5.5 国内典型的智能体开发平台





5.1 智能体的概念

继微博、微信和抖音之后，智能体成为每个企业的必备“武器”
智能体是企业在AI时代的“智能”客户端





5.1 智能体的概念

智能体 (AI Agent), 又称“人工智能代理”, 是一种模仿人类智能行为的智能化系统, 它就像是拥有丰富经验和知识的“智慧大脑”, 能够感知所处的环境, 并依据感知结果, 自主地进行规划、决策, 进而采取行动以达成特定目标。简单来说, 智能体能够根据外部输入做出决策, 并通过与环境的互动, 不断优化自身行为

智能体本身既不是单纯的软件也不是硬件, 而是一个更为宽泛的概念, 它们可以是软件程序、机器人或其他形式的系统, 具备一定的自主性和智能性



5.1 智能体的概念

基于大模型的智能体是指**利用大语言模型（如GPT、BERT等）作为核心组件，构建的能够执行特定任务、与环境交互并做出决策的人工智能系统**。这些智能体具有**自主性、交互性、适应性**等特点，能够模拟人类的认知和决策过程，提供更加自然、高效和个性化的交互体验。它们能够处理海量数据，进行高效的学习与推理，并展现出跨领域的应用潜力

AI Agent 的诞生就是为了处理各种复杂任务的，就复杂任务的流程而言 AI Agent 主要分为两大类：**行动类、规划执行类**。总而言之，**AI Agent 就是结合大模型能去自动思考、规划、效验和执行的一个计算体，以完成特定的任务目标**，如果把大模型比作大脑，那 AI Agent 可以理解为**小脑 + 手脚**





5.1 智能体的概念

过去的嵌入式工具型 AI助手型工具（例如 siri、小度、小爱音箱），只完成和人之间的问答会话

目前各类 AI Copilot 不再是机械地完成人类指令，而是可以参与人类 workflow，为诸如编写代码、策划活动、优化流程等事项提供建议，与人类协同完成

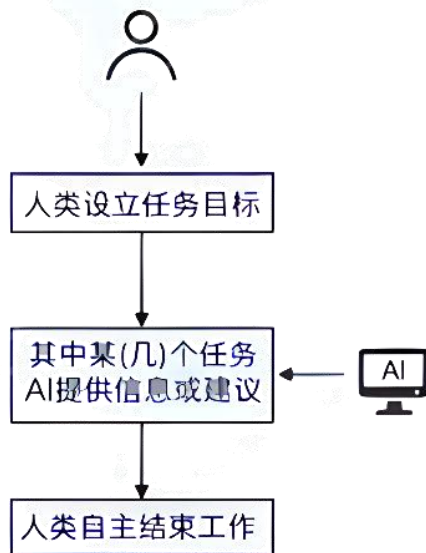
AI Agent 的工作仅需给定一个目标，它能够针对目标独立思考并做出行动，它会根据给定任务详细拆解出每一步的计划步骤，依靠来自外界的反馈和自主思考，自己给自己创建 prompt，来实现目标

人类与AI协同的三种模式

Embedding 模式



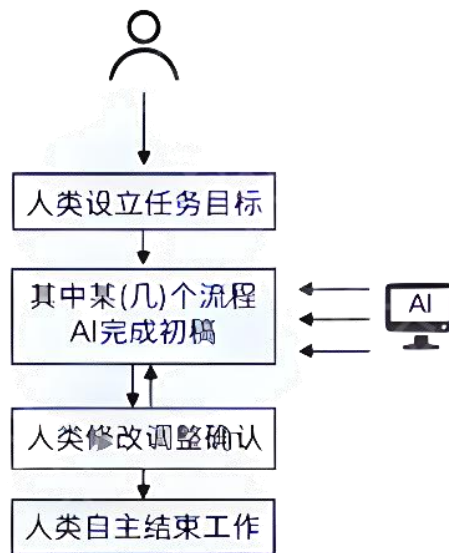
人类完成绝大部分工作



Copilot 模式



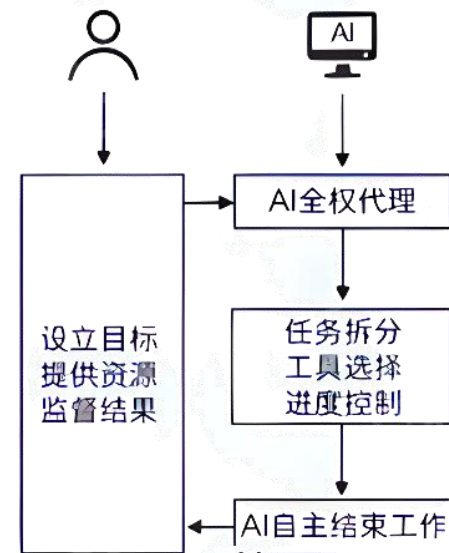
人类和AI协作工作



Agents 模式



AI完成绝大部分工作





5.2 智能体和RAG的区别

■ RAG和智能体

RAG 技术就像一个智能的知识助手，它通过将外部知识库与大语言模型相结合，极大地增强了模型的回答能力。当用户提出问题时，RAG 首先会在庞大的知识库中进行检索，找到相关的信息，然后将这些信息与大语言模型的生成能力相结合，给出更加准确、全面的回答。例如，在企业客服场景中，RAG 可以快速检索企业的产品资料、常见问题解答等知识库，为客户提供精准的服务，大大提高了客服的效率和质量

智能体更像是一个拥有自主意识的智能员工，它以大语言模型为驱动，具备自主理解、感知、规划、记忆和使用工具的能力。它可以根据目标任务，进行智能规划与决策，自动执行复杂的任务。比如在智能办公场景中，AI Agent 可以根据用户的日程安排，自动预订会议室、安排会议议程，还能在会议结束后自动生成会议纪要，实现办公流程的自动化，让员工从繁琐的事务性工作中解脱出来

RAG 的部署相对来说成本较低，主要集中在知识库的建设和维护上。企业可以利用现有的数据资源，构建自己的知识库，然后通过接入大语言模型，实现 RAG 的功能。对于一些数据量较小、业务场景相对简单的企业来说，RAG 是一个性价比很高的选择

AI Agent 的部署则需要更高的技术门槛和成本。它不仅需要强大的大语言模型支持，还需要构建复杂的智能规划和决策系统，以及与各种工具和系统的集成。此外，AI Agent 还需要不断地进行训练和优化，以提高其智能水平和适应性。因此，对于一些中小企业来说，AI Agent 的部署成本可能较高，但对于大型企业和对智能化要求较高的企业来说，AI Agent 带来的价值可能远超其成本



5.2 智能体和RAG的区别

■ RAG和智能体

RAG

在运行效果上，RAG **在处理一些需要大量知识支持的任务时表现出色**，比如智能问答、文档生成等。它能够利用知识库中的信息，为用户提供准确、详细的回答。但 RAG 的局限性在于，它缺乏自主决策和规划的能力，对于一些复杂的、需要灵活应变的任务可能无法胜任

AI Agent

AI Agent 则**在复杂场景下展现出了强大的优势**，它能够根据环境的变化和任务的要求，自主地进行决策和规划，实现任务的自动化执行。无论是在智能客服、智能办公还是智能生产等领域，AI Agent 都能够发挥出其独特的价值，提高企业的运营效率和竞争力

各自的优势和适用场景

RAG 和 AI Agent 都有各自的优势和适用场景。对于企业用户来说，在选择时需要根据自身的业务需求、数据资源、技术实力和预算等因素进行综合考虑。如果企业只是需要解决一些简单的知识检索和生成问题，那么 RAG 可能是一个不错的选择；如果企业希望实现业务流程的全面自动化和智能化，那么 AI Agent 则更能满足其需求。在这个快速发展的人工智能时代，只有选择最适合自己的技术，才能在激烈的市场竞争中立于不败之地



5.3 典型的智能体类型

■ 典型的智能体类型

- 01 生产力智能体**：这一类智能体通过自动化任务、优化工作流程和提升效率来显著提高生产力
- 02 智能体开发平台**：是支持开发、管理和部署智能体的技术基础。它们提供必要的工具和接口，帮助企业和个人快速构建和集成智能体
- 03 智能体框架**：是一组编程工具和库，专门为创建高效、可扩展的AI智能体设计。它们通常包括用于任务规划、对话管理和数据处理的模块
- 04 编程类AI智能体**：能够协助开发者完成代码编写、调试和优化，提升开发效率，减少人为错误
- 05 语音AI智能体**：通过语音识别和自然语言处理技术，与用户进行语音交互。这些智能体在智能家居、车载系统和客户服务中非常常见
- 06 客户服务智能体**：俗称“智能客服”，能够提供即时的客户支持、回答常见问题，并通过个性化推荐提高客户满意度
- 07 数据分析智能体**：可以处理海量数据，快速生成洞察，并帮助用户做出数据驱动的决策。它们广泛应用于商业、科学和金融领域
- 08 数字化工作者**：是一种AI智能体，专注于执行重复性高的任务，比如数据输入、文档整理和流程自动化
- 09 个人助理智能体**：可以帮助用户管理日程、发送提醒、执行简单任务，甚至提供个性化建议，提升个人的日常效率



5.3 典型的智能体类型

企业产品实例



超级客服智能体

搭建企业专属知识库

一站式导入文档、表格、问答对、网页内容等知识

支持图片、视频、文件、小程序等素材

将散落的知识教给一个“数字大脑”

基于知识库智能问答，准确回答问题

针对特殊场景和长尾问题执行智能转人工

及时提醒人工客服处理，客户零感知

是AI，也是 7*24h 在线超级客服

多渠道接入

智能转人工

智能对话摘要

辅助生成回复

The screenshot displays a customer service chat interface. At the top, there are two blue callout boxes: '多渠道接入' (Multi-channel access) and '智能转人工' (Intelligent transfer to human). The main chat area shows a conversation where the AI assistant provides a list of services: 1. 添加新人 (Add new person), 2. 客户管理 (Customer management), 3. 环境管理 (Environment management), 4. 团队协作 (Team collaboration), 5. 智能转人工 (Intelligent transfer to human). Below this, there is a section for '客户信息' (Customer information) with fields for name, phone, and address. At the bottom, there is a blue callout box labeled '辅助生成回复' (Assisted reply generation).



5.3 典型的智能体类型

主动触达客户

拟人对话

长期对话记忆

智能客户标签



私域销售智能体

基于 IM 触达客户

拟人对话和长期记忆，增加“真人感”

主动收集需求，基于知识库推荐产品

识别客户意图，自主智能转人工

生成 AI 客户档案卡片、AI 智能打标签

对话可管理，支持人工发送消息

AI 回复 / 人工回复一键切换

是AI，也是私域金牌销售！



5.3 典型的智能体类型

企业产品实例 Link AI

效率助手智能体

能总结文件、速读网页的工作“小帮手”

可查询政策、培训员工的企业内部“百事通”

对话式 AI 生成 SQL 进行数据分析

研报解读、内容打标

基于自定义插件，搭建企业专属 Agent

基于 workflow，拖拉拽轻松搭建 AI 业务流程

在 IM、办公平台或网页端随时待命的 AI 助手

是AI，也是企业的超级智能助手！

内容总结



企业知识库问答



对话式智能数据分析



5.4 智能体产品

2025年1月23日，OpenAI发布了一个创新性的智能体——Operator，它是一个能够像人类一样使用计算机的智能体。它基于OpenAI最新研发的CUA（Computer-Using Agent）模型，CUA将GPT-4o的视觉功能与通过强化学习获得的高级推理相结合，经过训练可以与图形用户界面（GUI，即人们在屏幕上看到的按钮、菜单和文本字段）进行交互。Operator通过观察屏幕并使用虚拟鼠标和键盘来完成任务，而无需依赖专门的API接口。这种设计使其可以适配任何为人类设计的软件界面，带来极高的灵活性

Operator好比一个博士水平的个人助理，你给他一个复杂的任务，它就会自动执行。Operator的主要功能包括自主完成诸如采购杂货、提交费用报表、订票、买日用品、填写表格等任务，旨在通过自动化操作提升日常生活和工作效率。它还可以一边在StubHub搜索勇士队比赛门票，一边处理网球场预订、寻找清洁服务和DoorDash订餐，实现多任务并行处理





5.4 智能体产品

2025年2月3日，OpenAI发布了一款新的智能体产品——Deep Research。Deep Research由OpenAI o3模型的一个版本提供支持，该模型针对网页浏览和数据分析进行了优化，它利用推理来搜索、解释和分析互联网上的大量文本、图像和PDF，并根据需要根据遇到的信息做出调整。Deep Research具有以下四大核心技术：

数据雷达

会自动24小时扫描全球知识库

知识拼图

能把零散的信息拼成完整的战略地图

逻辑推理

发现矛盾时，自动回溯、验证，调整推理路径

学术裁缝

可以综合各种知识，生成完美报告，附带文献引用



5.5 国内典型的智能体开发平台



Coze(扣子)是字节跳动推出的新一代 AI 智能体开发服务平台

Coze 扣子 AI智能体

自动办公 · 降本增效 · 获客变现

企业老板

应用场景

智能客服

营销文案

直播卖货

知识博主

应用场景

数字分身

视频文案

直播卖课

专家老师

应用场景

数字分身

视频文案

直播卖课

职场白领

应用场景

数字牛马

自动办公

副业服务



5.5 国内典型的智能体开发平台





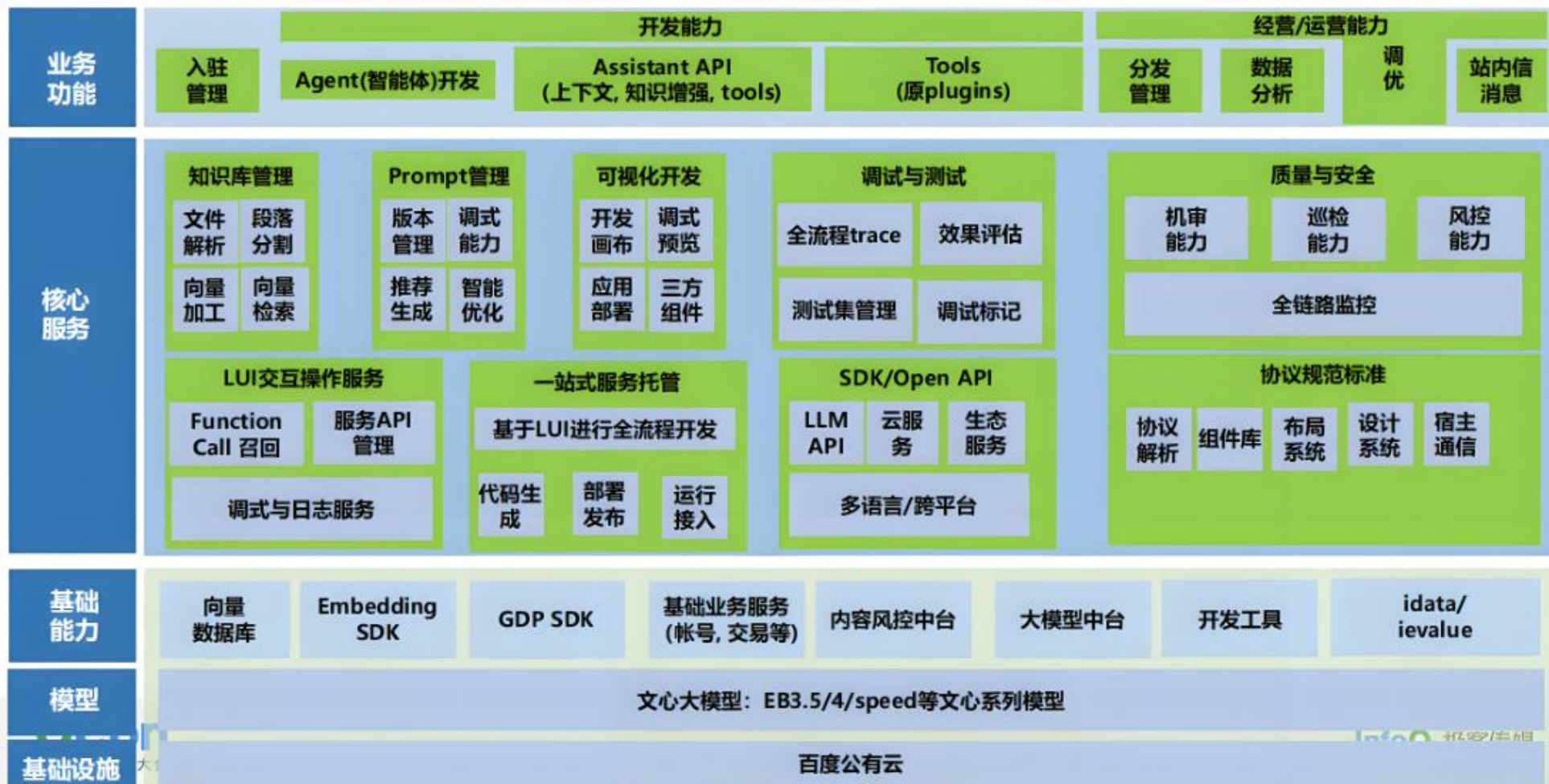
5.5 国内典型的智能体开发平台

文心智能体平台 AgentBuilder 官方

想象即现实 人人都是开发者

基于文心大模型的智能体平台 >
一句话、零代码创建智能体 >
调优迭代快、分发渠道广、商业可闭环 >

文心智能体整体架构图





5.5 国内典型的智能体开发平台

讯飞星火 智能体创作中心

< 星火智能体 >

深入垂直场景与核心需求，为个体和企业提供开箱即用的大模型应用

星火助手

35000+公开助手覆盖日常核心需求，为每个人的工作、出行、健康等方方面面优化体验，支持用户自主创建个性化助手。

原生应用

聚焦自媒体写作、PPT生成、虚拟人视频生成等场景，提供便捷的SaaS应用，为企业和个人提供专业生产工具

轻应用

依托讯飞星火App，面向医疗、翻译、创作等垂直场景，构建无需安装的轻量级应用。已接入智文、晓医、智作、语伴等专业助手。



5.5 国内典型的智能体开发平台

支付宝



百宝箱 · Tbox

基于支付宝的生活场景，通过领先的大模型技术，为企业快速构建专业级智能体。实现体验升级，智能决策，助力业务增长

业务创新

大模型加速业务升级

边界增长

Agent增强服务外延

价值提升

新体验驱动价值增长

使用专业版

体验基础版

立即咨询 >



为企业智能化提供全方位服务



大模型咨询服务

专业咨询服务，涵盖市场趋势分析，需求分析、技术方案设计、效果评测设计等



大模型技术服务

专属大模型定制，包含数据准备、模型调优、效果测试和模型部署等



技术工程服务

AI应用工程实施，涉及技术架构设计、技术研发、系统集成、性能优化等



售后专家服务

售后技术支持，覆盖故障诊断、系统维护、升级服务、技术咨询等



5.5 国内典型的智能体开发平台

Lenovo
ThinkStation

Lenovo 联想

智能体开发一体机

大模型应用 快人一步

一体化集成

- 预置丰富大模型，最高支持72B
- 零代码开发工具，分钟级创建智能体
- 最优算力适配，极致性价比

私有化部署

- 数据不出门，安全有保障
- 企业系统快速对接，成果易推广

简易运维

- 桌面级安装
- 7x24小时不间断运行

算力平台：联想ThinkStation PX 工作站

软件平台：联想智能体开发平台



6. 厂商提供的企业级大模型服务

- 6.1 企业级大模型落地解决方案服务
- 6.2 DeepSeek大模型一体机
- 6.3 企业选择大模型产品的考察维度
- 6.4 企业大模型应用专业顾问





6.1 企业级大模型落地解决方案服务

01

云厂商为代表的大厂

- 具备强大基础能力，可解决复杂问题；构建通用能力支持共性需求；拥有自研模型和充足算力，提供全方位服务

02

AI应用企业为代表的服务商

- 注重用户体验流畅性；对特定行业理解深入，提供贴合行业需求的方案；基于场景经验提供实用服务

03

新兴大模型应用开发服务商

- 有完善算力运营调度方案，通过软硬件优化和多元异构算力适配技术提升AI应用性能



6.1 企业级大模型落地解决方案服务

架构层

数据基座

- 数据中台与数据基建
- 企业专属知识库

算力基建

- 云端、本地算力基础设施
- AI算力盒子

平台层

积木式无代码搭建 workflow

- 企业内部流程自动化
- 企业业务智能化
- LLM Ops/DevOps

大模型微调与训练

- 大模型定制开发与私有化部署
- RAG&向量数据库
- 大模型推理与精调

应用层

AI+职能企业级应用

- 客户服务
- 员工支持与培训
- 营销与内容生成
- 数据运用与决策支持
- 研发与开发助手

AI+行业解决方案

- 智慧零售
- 智能制造
- 医药与医疗设备
- 智慧城市与园区
- 政务、文旅、法律等更多行业



6.1 企业级大模型落地解决方案服务



轻量级低成本

支持无代码积木式搭建应用与工作流，支持RAG、向量数据库等方式对大模型进行微调



部署快捷灵活

根据企业预算与需求，定制专属落地方案，最大化性价比，开箱即用



数据安全与合规

严密的数据管理和安全框架，满足数据私密性和合规要求



多种部署方式

支持本地私有化部署、公有云、私有云、混合云、AI算力盒子



应用场景广泛

从企业业务流程改造到员工支持，从内容生成到产品管理，满足各行业数智化需求



企业级定制服务

根据客户需求和业务场景，需求挖掘定制解决方案



数据结果可信

交叉验证提升结果准确度，高质量数据与信源



高兼容扩展性

支持接入和改造企业已有系统，支持接入第三方大模型与AI应用



6.1 企业级大模型落地解决方案服务



解决方案落地实践

上百家中大型企业验证过的大模型落地最佳实践

○ 试点探索 Pilot exploration

业务梳理

梳理企业内部流程、需求部门业务目标

价值链条

打通试点业务从 业务价值、流程、AI技术的价值链闭环

落地团队

高管+业务+HR+IT

○ 试点交付 Pilot delivery

试点验证

基于试点的应用部署、上线、运营

小步快跑

与业务团队、公司高层试点复盘，规划下一步方案

○ 成果复制 Achievement replication

融入业务

企业更多场景的盘点，扩大内训规模，复制成功经验

持续迭代

构建企业AI平台的创新模式和能力，更远期和规模化的AI创新规划



6.2 DeepSeek大模型一体机

国产化AI算力的“开箱即用”解决方案——DeepSeek大模型一体机

全尺寸模型支持

提供1.5B轻量版至671B超大规模模型的灵活调用，满足边缘端轻量化推理与云端复杂训练的双重需求，支持模型蒸馏与定制化开发，助力企业“按需取用”

动态资源调度

通过智能算力管理引擎，实现CPU、GPU等异构资源的动态分配，提升资源利用率，降低算力闲置成本

行业场景深度适配

内置政务公文写作、金融合同审核、工业质检、智能客服等多种垂直场景解决方案，结合API服务快速对接企业现有系统

数据安全与本地化部署

支持私有化部署至客户本地环境，确保金融、政府等高敏感行业的数据主权，符合国家信创安全标准

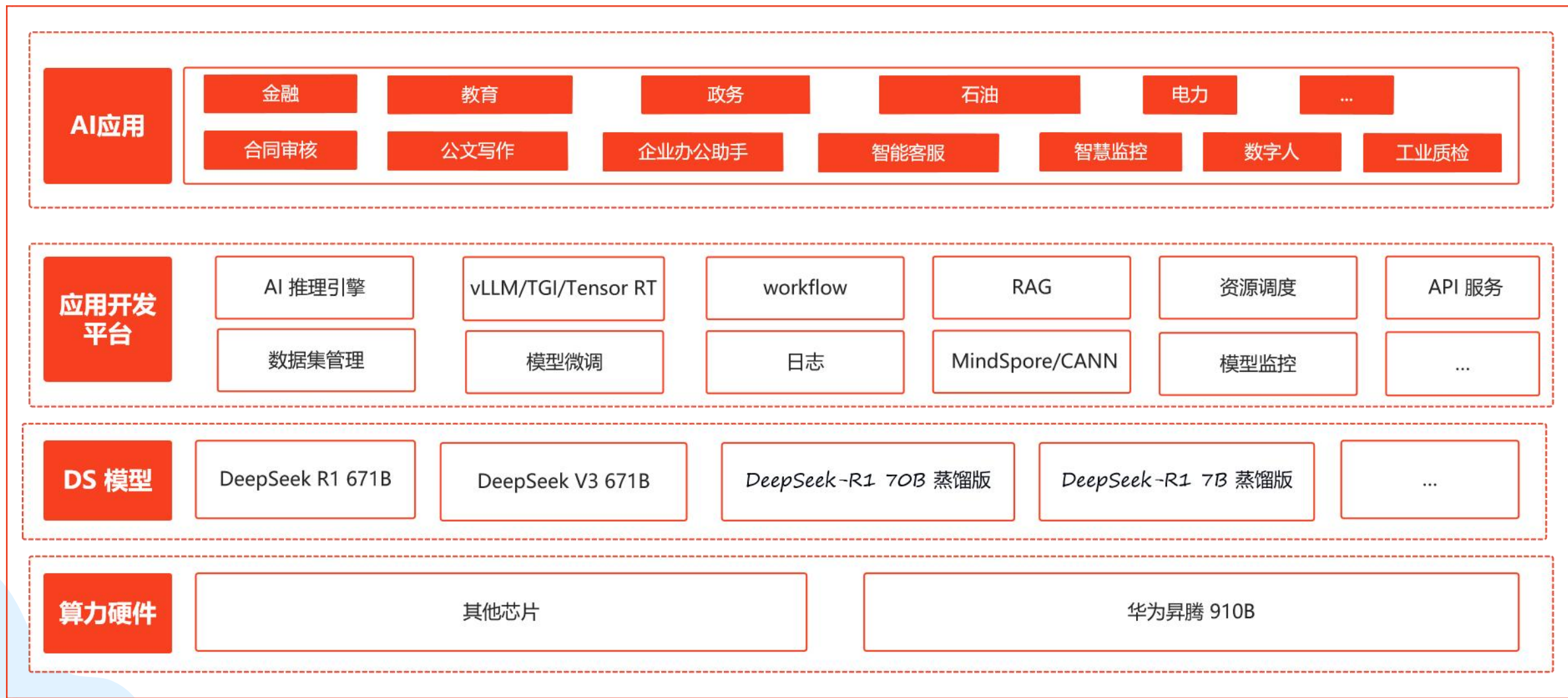
全生命周期管理

从数据集管理、模型微调、日志监控到自动化运维，提供端到端的技术支持，降低企业AI应用的长期运营成本



6.2 DeepSeek大模型一体机

某厂商国产化AI算力的“开箱即用”解决方案





6.2 DeepSeek大模型一体机

天玑科技: PriData 超融合一体机

深信服: 一朵云

海康威视: 文搜存储系列产品

大华股份: 大华神算

浪潮信息: DeepSeek "推理一体机"

中国长城: 长城擎天 GF7280 V5 AI 训推一体机

中科曙光: 曙光 DeepSeek 人工智能一体机

优刻得: DeepSeek 满血版大模型一体机

云从科技: 从容大模型训推一体机

天融信: 融信DeepSeek 安全智算一体机

新致软件: 新致信创一体机

软通动力: DeepSeek 应用方案一体机

科大讯飞: "星火+DeepSeek 双引擎一体机"

拓维信息: 拓维信息智能数据标注一体机

协创数据: Fcloud DeepSeek 满血版一体机

麒麟信安: 麒麟信安国产化智算一体机

亚康华创科技: D-BOX Pro 桌面级智能一体机

华为昇腾: 昇腾DeepSeek 推理一体机

联想集团: 智能体一体机与训推一体服务器

钉钉科技: 专属 AI 一体机

新华三: DeepSeek智能一体机

宝得: DeepSeek一体化智能设备

中国电信: 息壤智算一体机-DeepSeek 版

中国移动: 智算一体机-DeepSeek 版

中国联通: DeepSeek一体机

柏飞电子: DeepSeek 加固式一体机

京东云: DeepSeek 大模型一体机

华能振宇: 天巡 DeepSeek 大模型一体机

昆仑技术: DeepSeek 本地化部署一体机

百度: 百舸 DeepSeek 一体机

黄河信产: 黄河 DeepSeek 一体机



6.2 DeepSeek大模型一体机

■ 一体机价格

硬件成本分析

服务器硬件: ¥200,000-500,000, 用于配备高端处理器的企业级服务器

AI加速器: ¥1,000,000-2,500,000, 用于671B模型所需的8+高性能GPU/NPU

内存: ¥150,000-300,000, 用于1-2TB RAM配置 (如协创数据所提及)

存储: ¥80,000-200,000, 用于高性能NVMe SSD

软件与集成: ¥200,000-600,000, 用于DeepSeek部署、平台集成和支持服务

估计总成本范围

在Deepseek-671满血版硬件配置基础上估算如下:

入门级671B配置: ¥1,500,000-¥2,500,000

中端企业部署: ¥2,500,000-¥5,000,000

高端多服务器集群设置: ¥5,000,000-¥10,000,000+

关键成本因素

部署规模 - 单服务器与多服务器集群配置的差异

GPU/NPU选择 - 国产与国际加速器卡的成本差异

内存需求 - 671B等大模型需要大量内存资源

支持和维护合同 - 企业级支持协议的费用

软件许可 - DeepSeek商业许可和管理工具的成本



6.3 企业选择大模型产品的考察维度

响应速度

包括大模型响应时效和端到端响应时效。响应时效反映单次请求的处理效率,端到端时效关注交互场景的整体流畅度。响应时间过长会严重影响用户体验,降低工作效率,因此响应速度是一个重要的技术考量点

输入长度与多轮交互能力

主要指标是token长度。token长度越长,支持编码理解的上下文信息越丰富,不同意图之间的语义关联性更强。足够的token长度可以让模型学习和记忆更复杂的知识结构,进行多轮交互,推理更加准确

问答准确率

评估大模型文本生成、问答等功能的输出质量和正确率,是技术指标的核心。知识准确率直接决定了模型应用场景的范围和效果,是一个必须重点考量的维度

语言个性化程度

大模型是否能够按照客户化的语言模式,生成答案,以流畅的、适配客户情境的客户化语言进行交流

答案可信度

答案是否是真实可信的,不存在模型的幻想,或超越出本知识库内容的编造内容;是否提供了查证答案来源的可溯源的知识出处机制



6.3 企业选择大模型产品的考察维度

知识库构建难度

知识构建是否与日常知识积累复用，是否需要进行深度的知识加工才可以供大模型进行问答，其知识库构建难度越高，在这一点上得分越低

信息安全系数

是否能够确保信息安全，不泄密企业核心知识，没有信息安全隐患

垂域训练能力

是否能适配和接入垂直领域知识库，进行领域的训练。反映大模型对业务领域专业知识的理解和适应程度，直接影响知识服务的质量。核心是意图理解力和问题解决力。这关系到模型是否能贴合企业实际业务需求

集成扩展能力

是否能与企业知识库、企业业务系统、企业IM等数据源、渠道进行集成，实现大模型能力的中台化复用，集成接口是否完善

成本效益

前期成本投入包括软件采购和基础硬件投入。运营期人工替代率能直观反映大模型应用对企业运营成本的节约效果，所投入的成本最终能够代替多少人工或人工工时，直接反应了其成本收益的比

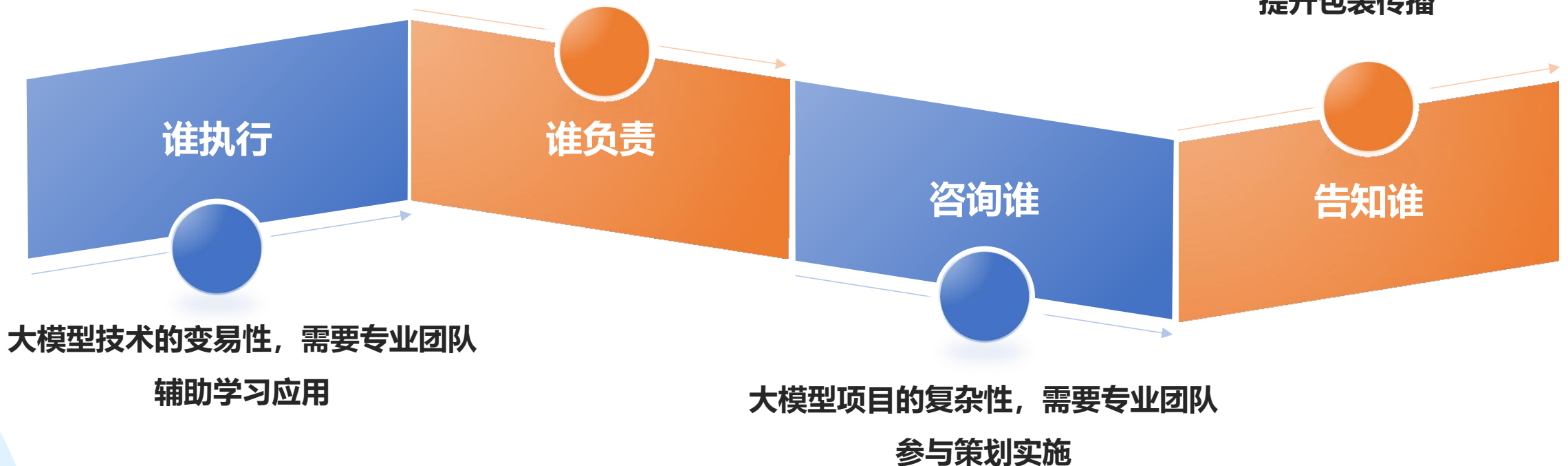


6.4 企业大模型应用专业顾问

■ 为什么需要专业顾问?

大模型应用的探索性，需要专业团队
协助调研论证

大模型创新的价值性，需要专业团队
提升包装传播





6.4 企业大模型应用专业顾问

■ AI大模型应用中专业顾问服务什么？

AI大模型应用的目标场景设定

01

AI大模型应用的功能需求设计

02

AI大模型应用的选型方法优化

03

AI大模型应用的组织过程协同

04

AI大模型应用的成果质量保障

05

AI大模型应用的成果推广转化

06



6.4 企业大模型应用专业顾问

■ AI大模型应用中专业顾问如何服务？

教练式

AI大模型应用专业顾问服务，类似于教练式服务。与客户企业一起对AI+行动规划进行市场趋势、方法论和内容体系等指导和讨论，提供相关主题讲座，提供主题研讨交流，提供过程中的疑难咨询、方法讨论。帮助界定内容，指导把控要点，促进甲方项目组成员数字化专业水平的提升

监理式

AI大模型应用专业顾问服务，也类似于项目监理服务。通过服务，保障企业AI+创新应用项目的全过程得到专业化的质量管控。协助组织和参与实施项目前期方案评审、过程重要节点评审、项目验收评审，对交付成果依据项目需求和研发协议进行全盘评价，协助与技术开发方的沟通和完善。从需求阶段的共研，到过程中外协开发方各节点规划设计的交付评审，到最终规划成果的评审验收，提供专业意见和规范指导

伙伴式

AI大模型应用专业顾问服务，同样类似于伙伴式服务。围绕项目全过程，与企业工作组成员一道，随时响应，及时交流，结合外协开发方服务进度和方式，共同沟通，融合协作，随时响应沟通交流，双方决策层建立协作直通车，在项目研发、应用和成果化等各方面进行紧密协作，保障规划项目最终实现优秀成果并便于应用落地

7. 大模型典型应用案例





7. 大模型典型应用案例

- 瑞金医院携手华为发布瑞智病理大模型RuiPath，单切片AI诊断仅需数秒

瑞智病理大模型RuiPath，是基于瑞金医院数字化智慧病理科的全业务流程开发的临床级国产多模态互动式病理大模型，**数字化和AI助力医院实现多模态融合**

在交互式辅助诊断环节，传统诊断方式是医生在显微镜下逐张查看切片，完成诊断后再人工录入报告。而 RuiPath能够提前精准识别病灶区域，单切片AI诊断时间仅需数秒。这使得医生的工作模式从在镜下逐个寻找病灶，转变为以互动方式审核AI诊断结果，从 **‘逐片诊断 (Slide by slide)’ 模式升级为 ‘逐步审核 (Step by step)’ 模式**，显著提升了诊断效率与质量





7. 大模型典型应用案例

2025年3月，东莞市人工智能大模型中心将正式启动运营，大模型中心由东莞市数字经济发展集团建设，依托华为昇腾AI全栈技术，构建起100P国产化智能算力底座，上线3个通用大模型及2个制造业垂直模型

作为全国首个聚焦制造业的城市级大模型中心，其创新性在于，针对制造企业“高精度、低成本、快迭代”的共性需求，提供开箱即用的算法服务与AI开发平台。企业仅需注入少量数据，即可快速提升模型精度，破解传统AI开发周期长、投入高的痛点

通过CV大模型实现包装盒外观缺陷毫秒级识别；同时运用运筹优化算法，重构全球供应链物流网络，通过数据治理，为月台调度、派车计划两个求解器场景提供准确数据，使月台调度效率提升，每年节省物流成本300万-440万





7. 大模型典型应用案例

■ 云南白药集团股份有限公司的“大模型应用开发平台”

云南白药集团股份有限公司构建了面向全员的人工智能企业级大模型应用开发平台，深植 LLM 运营 (LLMOps) 的先进理念。平台不仅无缝集成了主流大模型提供商的丰富模型矩阵，还深耕 Prompt 工程学领域，通过精细化指令设计，引导模型展现超凡智能表现

应用实例

基于大模型应用平台构建的进销存数据清洗及标准化解决方案，实现了由人工处理转为线上系统自动化流程处理，月处理药品销售库存 3000+ 份文件约 300 万条记录，平均每份文件处理时间由 10 分钟降至 1 分钟，人效提高 90%

案例亮点

平台包含了构建 LLM 应用所需的关键技术栈，可以快速搭建企业业务助手，为业务场景解决方案提供便利性 AI 支持



7. 大模型典型应用案例

■ 厦门市合趣信息科技有限公司：云合·AI 产业顾问

云合·AI 产业顾问是基于大语言模型及产业大数据的办公效率工具，为全国数百万招商及产业人员提供准确、及时、可靠的产业分析、数字招商和辅助办公功能。它帮助政府、招商、投资、产研等机构通过 AI 对话、知识库、企业分析报告等功能快速完成工作

应用实例

- 投洽会 AI 顾问 - 鹭可：在 23 届 98 投洽会线下现场及线上平台“云上投洽会”中集成了大模型数字人对话应用“投洽会 AI 顾问”，以厦门城市形象“鹭可”作为智能对话数字人，将投洽会“展、会、谈、宣”及厦门文旅、核心产业数据进行知识增强，用户可通过智能语音 / 文字对话方式进行各类相关内容的问答
- 云合·AI 产业顾问 - 通用版：23 年 11 月份在互联网上启动开放下载及注册，目前已有数千名个人办公用户，并形成了一定的付费转化，帮助不少企业员工和管理层实现了 AI 办公赋能

案例亮点

将 AI 大模型的能力经过场景化提炼整合，为用户提供具体工作场景所需的功能，如 AI 对话、AI 角色 / 数字员工、AI 全局快捷指令、AI 写作、企业分析报告、知识库、思维导图、研报中心、AIPPT、AI 绘图、AI 产业数据服务、AI agents 定制服务等



7. 大模型典型应用案例

天士力集团大模型探索实践



8. AIGC与企业应用实践

8.1 AIGC概述

8.2 文本类AIGC应用实践

8.3 图片类AIGC应用实践

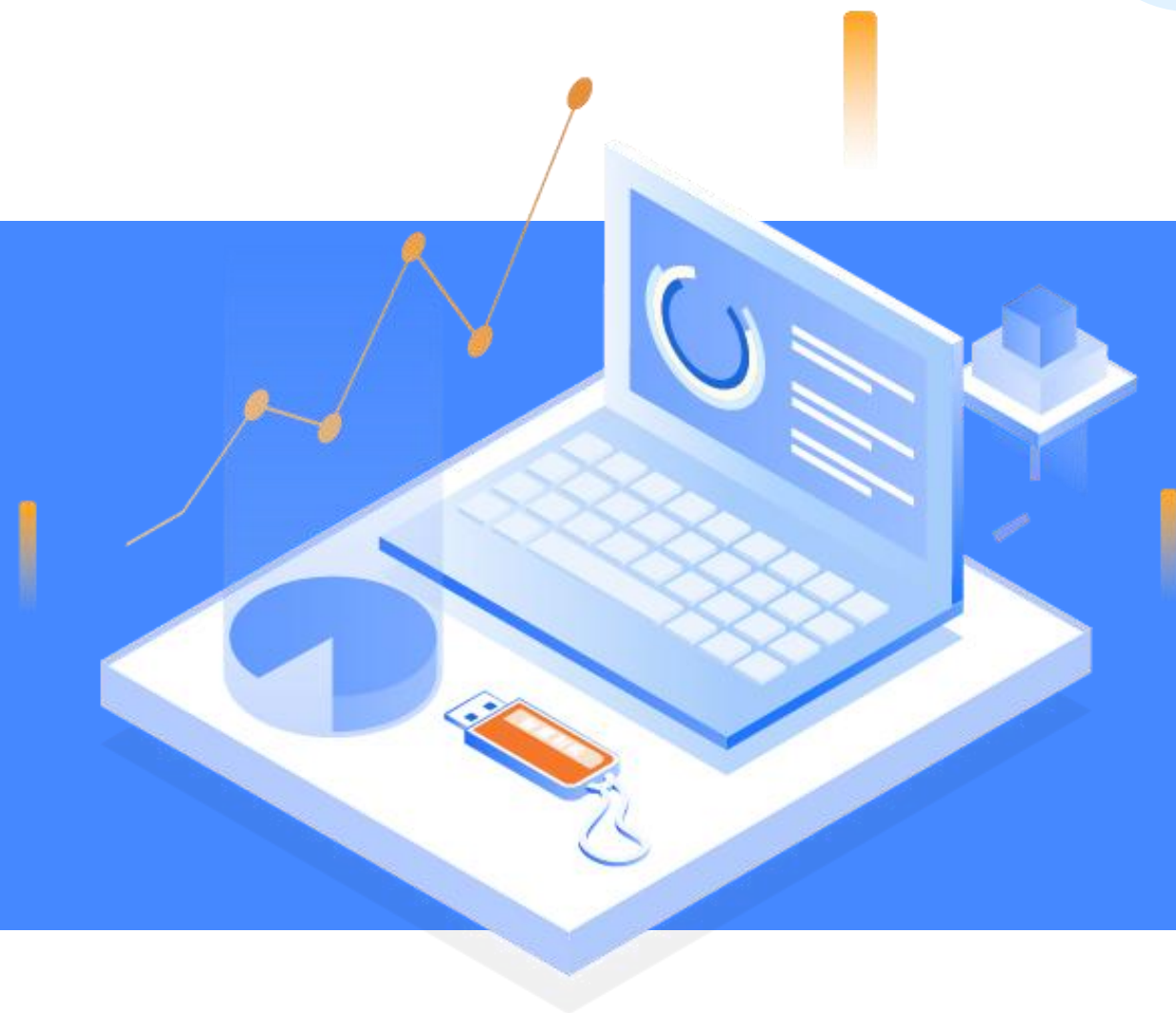
8.4 语音类AIGC应用实践

8.5 视频类AIGC应用实践

8.6 AIGC在辅助编程中的应用

8.7 AI搜索

8.8 AI智能办公



8.1 AIGC概述

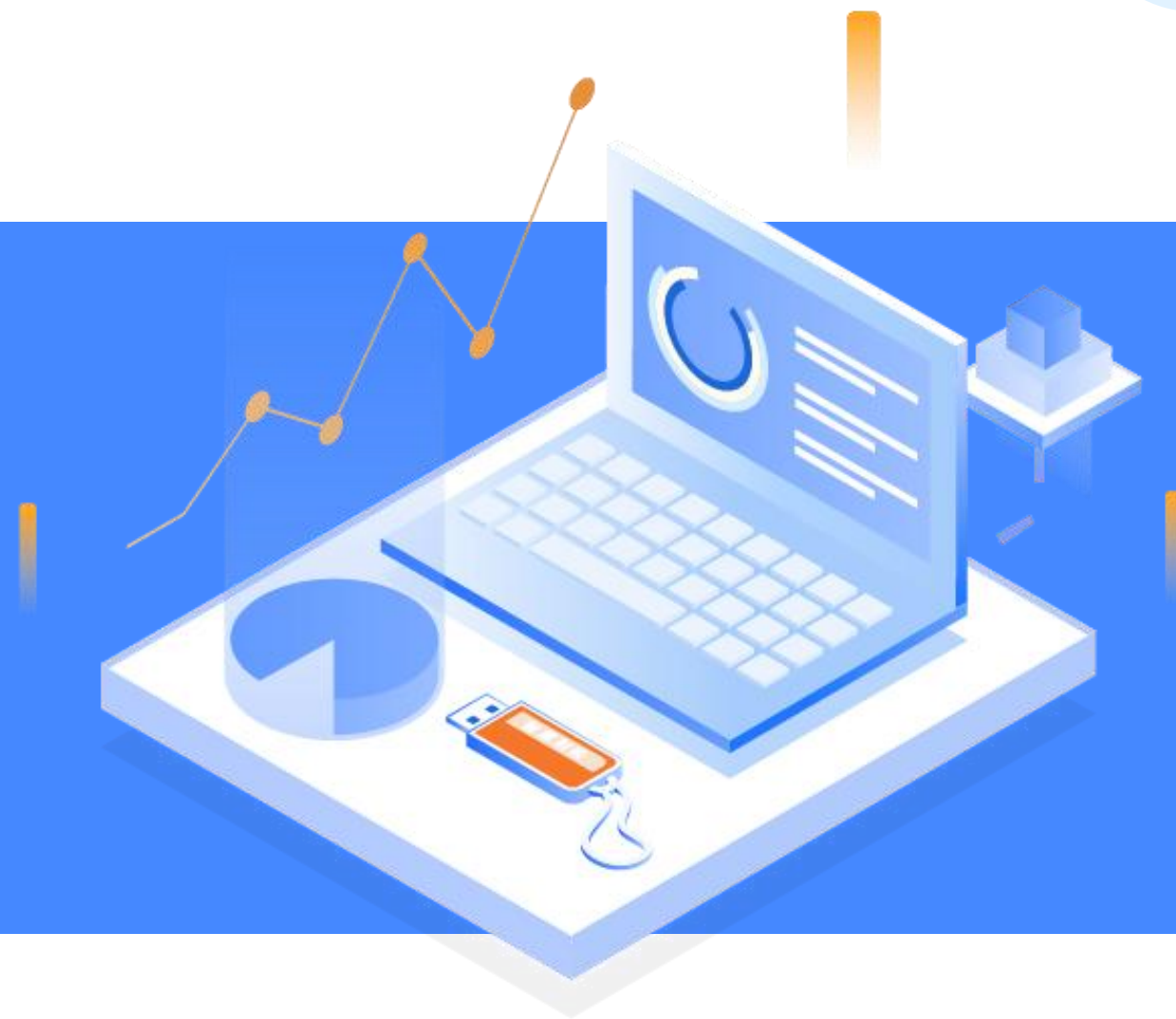
8.1.1 什么是AIGC

8.1.2 AIGC与大模型的关系

8.1.3 常见的AIGC应用场景

8.1.4 常见的AIGC大模型工具

8.1.5 AIGC大模型的提示词





8.1.1 什么是AIGC

AIGC的全称为“Artificial Intelligence Generated Content”，中文翻译为“人工智能生成内容”。这是一种新的创作方式，利用人工智能技术来生成各种形式的内容，包括文字、音乐、图像、视频等

- AIGC是**人工智能进入全新发展时期的重要标志**，其核心技术包括生成对抗网络（GAN, Generative Adversarial Networks）、大型预训练模型、多模态技术等
- AIGC的核心思想是**利用人工智能算法生成具有一定创意和质量的内容**。通过训练模型和大量数据的学习，AIGC可以根据输入的条件或指导，生成与之相关的内容。例如，通过输入关键词、描述或样本，AIGC可以生成与之相匹配的文章、图像、音频等
- AIGC技术**不仅可以提高内容生产的效率和质量，还可以为创作者提供更多的灵感和支持**。在文学创作、艺术设计、游戏开发等领域，AIGC可以自动创作出高质量的文本、图像和音频等内容。同时，AIGC也可以应用于媒体、教育、娱乐、营销、科研等领域，为用户提供高质量、高效率、高个性化的内容服务





8.1.2 AIGC与大模型的关系

大模型与AIGC之间的关系可以说是**相辅相成、相互促进**的。大模型为AIGC提供了强大的技术基础和支撑，而AIGC则进一步推动了大模型的发展和应用

01

大模型为AIGC提供了丰富的数据资源和强大的计算能力

02

AIGC的需求也推动了大模型的发展

03

大模型和AIGC的结合，也带来了广泛的应用前景



8.1.3 常见的AIGC应用场景

电商

生成商品标题、描述、广告文案和广告图

办公

写周报日报，写方案，写运营活动，制作PPT，写读后感，写代码

游戏

生成场景原画，生成角色形象，生成世界观，生成数值，生成3D模型，生成NPC对话，音效生成

娱乐

头像生成，照片修复，图像生成，音乐生成

影视

生成分镜头脚本，生成剧本脚本，台词润色，生成推广宣传物料，音乐生成



8.1.3 常见的AIGC应用场景

动漫

原画绘制, 动画生成, 分镜生成, 音乐生成

艺术

写诗, 写小说, 生成艺术作品, 草图生成, 艺术风格转换, 音乐创作

教育

批改试卷, 试卷创建, 搜题答题, 课程设计, 课程总结, 虚拟讲师

设计

UI设计, 美术设计, 插画设计, 建筑设计

媒体

软文撰写, 大纲提炼, 热点撰写

生活

制定学习计划, 做旅游规划



8.1.4 AIGC技术对行业发展的影响

内容创作领域的革新

AIGC技术能够自动生成高质量的文本、图像、音频和视频等内容，极大地提高了内容创作的效率。在新闻、广告、自媒体等领域，AIGC已经实现了广泛应用，帮助创作者快速生成多样化、个性化的内容，满足市场需求。这种技术革新不仅降低了内容创作的成本，还激发了创作者的创新灵感，推动了内容产业的繁荣发展

生产力提升与成本降低

AIGC技术在多个行业中展现了其提升生产力和降低成本的潜力。例如，在游戏开发领域，AIGC技术可以用于场景构建、角色互动等，减少人工制作的工作量，提高开发效率。在制造业中，AIGC技术可以辅助设计、优化生产流程，降低生产成本。这些应用使得企业能够更快地响应市场变化，提升竞争力

用户体验的升级

AIGC技术通过提供个性化、定制化的内容和服务，显著提升了用户体验。在智能客服、在线教育等领域，AIGC技术可以根据用户的需求和偏好提供精准的服务，满足用户的个性化需求。这种以用户为中心的服务模式不仅增强了用户的满意度和忠诚度，还为企业带来了更多的商业机会

推动行业创新与转型

AIGC技术的快速发展为传统行业带来了转型升级的契机。通过与AIGC技术的深度融合，传统行业可以探索新的商业模式和服务模式，实现创新发展。例如，在零售业中，AIGC技术可以用于智能推荐、虚拟试衣等场景，提升购物体验并促进销售增长。在金融领域，AIGC技术可以应用于投资策略优化、风险管理等方面，提高金融机构的决策效率和准确性



8.1.5 AIGC技术对职业发展的影响

新兴职业的出现

随着**AIGC**技术的快速发展，一系列与该技术相关的新兴职业应运而生。例如，**AI**训练师、机器学习工程师、数据标注员等职业需求激增。这些新兴职业不仅要求从业者具备扎实的技术基础，还需要不断学习和掌握最新的**AIGC**技术动态

传统职业的转型升级

AIGC技术也为传统职业的转型升级提供了契机。许多传统职业如编辑、设计师、教师等，在**AIGC**技术的辅助下，工作效率和创作质量得到了显著提升。同时，这些职业也需要从业者不断适应技术变革，掌握新的技能和工具，以适应市场需求的变化

工作方式的变革

AIGC技术改变了传统的工作方式，使得远程工作、灵活办公成为可能。许多企业开始采用**AIGC**技术来优化工作流程，减少人力成本，提高工作效率。这种变革不仅为员工提供了更加灵活的工作方式，也为企业带来了更大的经济效益

职业发展路径的多样化

AIGC技术的发展为职业发展路径提供了更多的可能性。从业者可以根据自己的兴趣和特长，选择适合自己的职业发展方向。例如，一些对**AI**技术感兴趣的从业者可以选择成为**AI**训练师或机器学习工程师，而一些具有创意和设计才能的从业者则可以利用**AIGC**技术来提升自己的创作能力

持续学习与技能提升

面对**AIGC**技术的快速发展，从业者需要不断学习和提升自己的技能水平。通过参加培训课程、阅读专业书籍、参与技术论坛等方式，从业者可以紧跟技术前沿，保持自己的竞争力



8.1.6 常见的AIGC大模型工具

OpenAI的ChatGPT

DeepSeek

科大讯飞的讯飞星火

阿里的通义千问

百度文心一言

字节跳动豆包

Kimi

这些工具基于大语言模型技术，具备文本生成、语言理解、知识问答、逻辑推理等多种能力，可广泛应用于写作辅助、内容创作、智能客服等多个领域。通过不断迭代和优化，为用户提供更加智能、高效的内容生成解决方案



8.1.7 AIGC大模型的提示词

- AIGC大模型的提示词 (Prompt) 是指**用户向大模型输入的文本内容**，用于**触发大模型的响应并指导其如何生成或回应**
- 这些提示词可以是一个问题、一段描述、一个指令，甚至是一个带有详细参数的文字描述。它们为大模型提供了生成对应文本、图片、音频、视频等内容的基础信息和指导方向
- 提示词的重要作用如下：

引导生成

增强交互性

提高准确性



8.1.7 AIGC大模型的提示词

使用提示词需要注意一些技巧，这样可以从大模型获得更加符合我们预期要求的结果

简洁明确

考虑受众

分解复杂任务

使用肯定性指令

示例驱动

明确角色

遵守规则

自然语言回答





8.1.7 AIGC大模型的提示词

提示词之道：通用大模型与推理大模型在提示词策略方面也有不同侧重与技巧

推理模型

- 提示语更简洁，只需明确任务目标和需求（因其已内化推理逻辑）
- 无需逐步指导，模型自动生成结构化推理过程（若强行拆解步骤，反而可能限制其能力）

通用模型

- 需显式引导推理步骤（如通过“思维链”提示），否则可能跳过关键逻辑
- 依赖提示语补偿能力短板（如要求分步思考、提供示例）



8.2 文本类AIGC应用实践

超级产品 增长1亿用户所用的时间



注：DeepSeek 包含网站Web/应用App累加不去重，Tiktok 不包含国内版抖音
数据来源：AI产品榜 aicpb.com 感谢郑瑞庭提供作图思路



遥遥领先的国产大模型之光

DeepSeek-V3

我是 DeepSeek, 很高兴见到你!

我可以帮你写代码、读文件、写作各种创意内容, 请把你的任务交给我吧~

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索



即刻体验 AI搜索DeepSeek-R1满血版

进入百度官网访问DeepSeek保证正常快速使用



8.2 文本类AIGC应用实践

■ 快速体验DeepSeek

在提示词输入框中输入“请模仿李白的《望庐山瀑布》做一首诗，题目是《看厦门鼓浪屿》”，然后回车，或者用鼠标点击提示词输入框右侧的箭头按钮，向DeepSeek发起提问。DeepSeek给出的回答如图所示，需要注意的是，大模型属于概率模型，每次生成的回答内容可能不完全相同。



☒ 已深度思考 (用时 21 秒) ▾

《看厦门鼓浪屿》

鹭岛东南烟水遥，日光岩上碧云飘。

琴潮忽卷千堆雪，散作凤凰霞万条。

注：本诗仿太白雄奇飘逸之风，首句以“烟水遥”勾勒海上仙山意境，承以日光岩实景作天地支点。转句化琴岛声波为视觉浪涌，取“千堆雪”暗合东坡赤壁豪情，结句凤凰木朱华迸天，将声、光、色熔铸为漫天红霞，较之银河落九天另开海上生霞之瑰丽新境。





8.2 文本类AIGC应用实践

■ DeepSeek的基本用法

基本原则：简单直接，自然表达

①无需复杂结构。直接描述需求即可，无需添加“角色扮演”（如“假设你是专家”）或复杂指令（如“用学术语言分三点回答”）。比如，你可以直接向DeepSeek提问“什么是光合作用？”、“如何用Python写一个计算器程序？”，而不建议使用提示词“请以生物学教授的身份，用三个段落解释光合作用，每段不超过100字”。②多轮对话优化结果。如果首次回答不完整，可通过追问补充细节，无需一次性给出完美提示。比如，第一轮提问“写一首关于秋天的诗”，第二轮提问“加入一些悲伤的情绪”，第三轮提问“把‘落叶’换成比喻句”

不同场景的提问技巧（非必需，但可提升效率）

虽然简单提问即可满足大多数需求，但在复杂任务中，适当提供背景信息或明确需求会让结果更精准，具体技巧包括：①知识类问题。比如，基础提问是“量子力学的基本原理是什么？”，优化后的提问是“用通俗易懂的语言解释量子纠缠，适合高中生理解”，②创作类任务（写作、编程等）。比如，基础提问是“写一个关于人工智能的科幻短篇故事”，优化后的提问是“写一个反乌托邦主题的科幻故事，主角是女性工程师，结局有反转”。③实用建议（学习、工作等）。比如，基础提问是“如何提高英语听力？”，优化后的提问是“我每天只有30分钟学习时间，有哪些高效的英语听力练习方法？”。④复杂任务（数据分析、代码调试）。比如，基础提问是“这段Python代码报错了，帮我看看问题”，优化后的提问是“我的代码目标是爬取网页数据，但遇到SSL证书错误。报错信息如下：[粘贴代码]”



8.2 文本类AIGC应用实践

■ DeepSeek的基本用法

作为初学者，DeepSeek的一些“魔法”指令也很有用，比如，你可以输入“/步骤 如何用手机拍摄旅游照片”，DeepSeek返回的回答结果就会按照步骤详细给出拍摄旅游照片的说明，再比如，你可以输入“请解释量子计算，然后/简化”，它就会返回比较简明扼要的回答

指令	功能
/续写	当回答中断时自动继续生成
/简化	将复杂内容转换成大白话
/示例	要求展示实际案例（特别是写代码时）
/步骤	让AI分步骤指导操作流程
/检查	帮你发现文档中的错误

DeepSeek的“魔法”指令



8.2 文本类AIGC应用实践

■ 使用DeepSeek处理文档

点击DeepSeek界面中的「回形针」图标上传文件，支持的文件类型包括文本类（PDF、DOCX、TXT、Markdown）、数据类（CSV、XLSX）和图像类（JPG、PNG）。然后，就可以在对话框中输入提示词，比如，可以输入“总结这份年报的三个核心要点”、“提取合同中的责任条款制成表格”、“对比文档A和文档B的市场策略差异”、“从实验报告中整理所有温度数据”、“请识别图片中的文字”等。也可以使用一些指令来处理文档，如表所示。

功能	指令模板	应用场景
内容摘要	/总结 [文件名] 生成500字摘要	快速把握长篇文档核心内容
问答提取	/问答 [文件名] 第三章提到的技术参数是？	精准定位特定信息
数据可视化	/可视化 [文件名] 将销售数据生成折线图	转化表格数据为图形分析
跨文档对比	/对比 文件A vs 文件B 的政策差异	合同/论文查重对比

■ 还可以要求DeepSeek对回答结果进行结构化输出，比如，可以输入如下提示词：

/解析文件 年度报告.docx

输出要求：

1. 按"营收/利润/成本"分类
2. 用Markdown表格对比近三年数据
3. 关键增长点用 标注



8.2 文本类AIGC应用实践

案例：DeepSeek+Kimi制作PPT

01

步骤1:

登录DeepSeek平台。在浏览器地址栏中输入“<https://www.deepseek.com/>”网址，进入“DeepSeek”平台，完成登录后，进入平台的对话功能界面。

02

步骤2:

输入你要生成PPT的提示词

03

步骤3:

生成并保存DeepSeek生成的PPT大纲和框架

04

步骤4:

登录KIMI平台。在浏览器地址栏中输入“<https://kimi.moonshot.cn/>”网址，进入“KIMI平台”，登录成功后，点击平台左侧的“kimi+”后选择PPT助手功能

05

步骤5:

在Kimi中粘贴DeepSeek生成的PPT大纲和框架

06

步骤6:

一键生成PPT并选择适合的PPT模板和风格

07

步骤7:

编辑和确认生成的PPT内容，无误后下载PPT



8.2 文本类AIGC应用实践

案例：DeepSeek+Kimi制作脑图

将我们准备好的电子书上传到DeepSeek，并输入提示词：现在我需要做一个xmind思维导图，请帮这份文档输出为Markdown格式

将内容复制到text的文件中，进行保存，并将文件的后缀名改为.md

打开xmind软件，通过文件-导入-Markdown进行文件的导入，最后就能马上渲染出一个非常完美的思维脑图

```

markdown
# 高效能人士的七个习惯

## 作者简介
- **史蒂芬·柯维**
  - 《时代周刊》“美国25个最具影响力人物”之一
  - 《经济学人》“最具前瞻性管理思想家”
  - 核心理念：由内而外的自我重塑

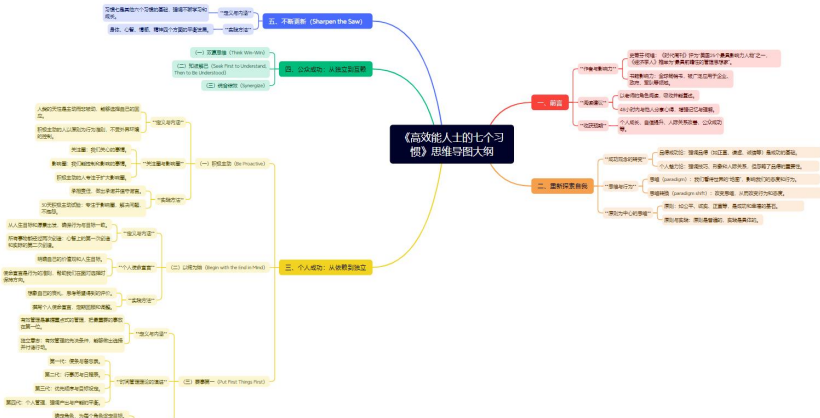
## 核心框架


```

mermaid
graph TD
 A[七个习惯] --> B[个人成功]
 A --> C[公共成功]
 A --> D[自我更新]
 B --> B1[习惯1-3]
 C --> C1[习惯4-6]
 D --> D1[习惯7]

```


```





8.3 图片类AIGC应用实践

图片类AIGC是一种基于人工智能技术生成图片的方法，它利用深度学习、生成对抗网络（GAN）等先进算法，通过学习和模仿大量图像数据，能够自动创作出高度真实和艺术化的图片。AIGC在图像生成、修复、风格转换、艺术创作等领域展现出强大能力，**为数字艺术、设计、游戏、电影等多个行业带来创新解决方案。其优势包括高效性、多样性和自动化**，能够快速生成大量高质量的图像内容，满足各种复杂需求





8.3 图片类AIGC应用实践

图像生成

AIGC能够生成高度逼真的图像，如人脸、动物、建筑物等。例如，OpenAI发布的DALL-E可以根据文本提示词创作出全新的、原创的图像，展示了AI在图像创作方面的强大能力

图像修复

AIGC还可以修复损坏的图像，如去除噪声、填充缺失的部分等。这项技术对于保护和恢复古老的艺术作品、修复损坏的照片等具有重要意义

图像增强

通过对图像进行增强处理，AIGC可以增加图像的饱满感和增强细节，使图像质量得到提升。这在提升照片的视觉效果、改善图像的清晰度和细节方面非常有用

图像识别

AIGC在图像识别方面也有广泛应用，可以识别图像中的对象、场景和特征，如人脸识别、车牌识别等。这项技术对于安防监控、智能搜索、自动驾驶等领域的发展至关重要



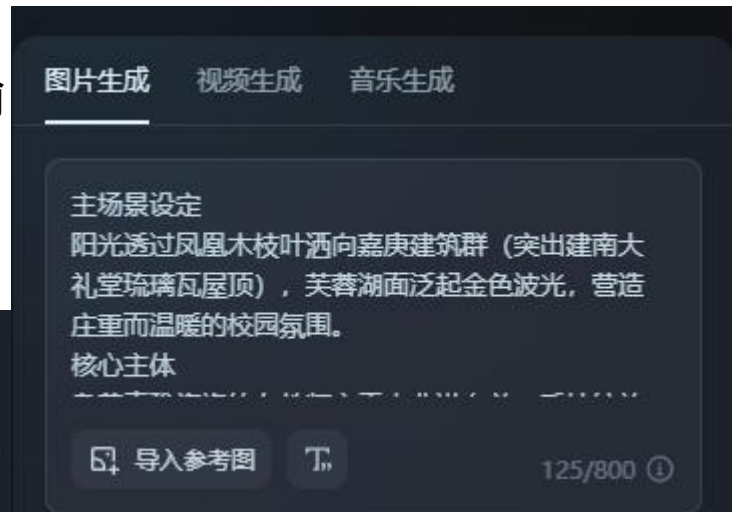
8.3 图片类AIGC应用实践

案例：DeepSeek+即梦AI，生成教师节海报

第1步：登录DeepSeek，输入如下提示词：

我想用AI绘图软件绘制一张教师节海报，要加入厦门大学元素，请帮我写一段提示词

第2步：登录“即梦AI”，进入“AI作图”的“图片生成”，在提示词输入框中，把第1步DeepSeek生成的提示词粘贴进来，点击“立即生成”





8.3 图片类AIGC应用实践

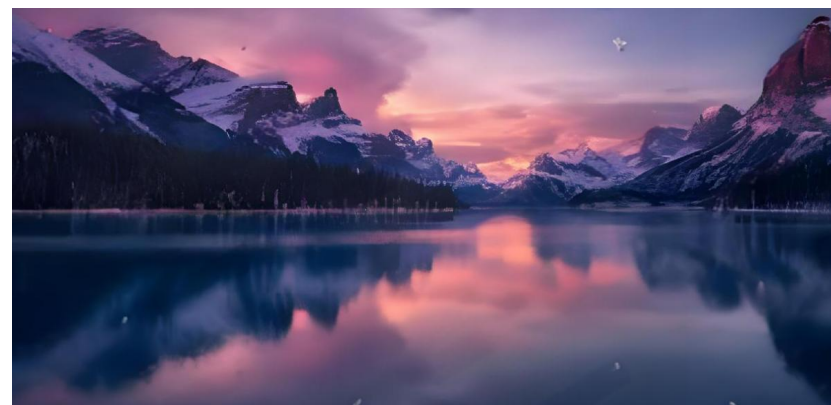
案例：去除图片中的水印



Hi, AI搜索已支持R1满血版, 快来试试吧!



带有水印的图片



去除水印后的图片



8.4 语音类AIGC应用实践

语音类AIGC是一种利用人工智能技术（特别是语音识别、自然语言处理和语音合成技术），自动生成和处理语音内容的技术。它能够**模拟人类语音，实现语音到文本的转换、文本到语音的合成，以及语音情感分析等功能**，广泛应用于智能语音助手、智能客服、语音翻译等多个领域





8.4 语音类AIGC应用实践





8.4 语音类AIGC应用实践



企业人员怎么用?

- 可以使用语音类AIGC工具（喜马拉雅音频大模型、腾讯智影），根据自己的文本内容，自动生成专业的配音，可以采用专业播音员的音色，也可以使用AIGC工具（比如米可智能）“克隆”自己的报告声音，用自己的音色生成配音
- 使用鬼手剪辑GhostCut进行语音翻译，可以把一种语言的报告视频自动转换成另外一种语言的报告视频



8.5 视频类AIGC应用实践

视频类AIGC是指利用人工智能技术，特别是深度学习、机器学习等算法，自动创建或处理视频内容的技术。它能根据给定的文本、图像或其他数据，自动生成符合描述的视频内容，涵盖文生视频、图生视频、视频风格化、人物动态化等多个方向。这一技术在创意设计、影视制作等领域潜力巨大，极大地提升了视频内容的生产效率和质量





8.5 视频类AIGC应用实践

- 视频类AIGC大模型发端于Sora。2024年2月，美国的Open AI发布了全球第一款文生视频大模型Sora（这里的“文生视频”是指由输入的文本内容生成相应的视频），迅速引起了业界的广泛关注和讨论

可灵

由快手推出，被誉为中国版Sora，视频生成时长可达120秒，支持文生视频、图生视频、视频续写、镜头控制等功能，表现出色。

Vidu

生数科技联合清华大学发布，是中国首个长时长、高一致性、高动态性视频大模型，支持一键生成16秒高清视频，性能对标国际顶尖水平。

即梦AI

即梦 AI 是字节跳动旗下剪映团队开发的一站式 AI 创意创作平台。支持文生图、图生图、文生视频、图生视频等，有智能画布、故事创作等

企业人员怎么用？

- 借助**可灵AI**工具，根据文本内容自动生成高质量的视频
- 使用**即梦AI**实现图生视频
- 使用**通义万相**根据图片生成自带音效的视频
- 借助**剪映AI**工具，根据文字自动生成高质量的视频，并自带配音配乐
- 使用**鬼手剪辑**生成视频解说
- 使用**腾讯智影**生成数字人播报视频（**用于制作企业教学视频**）



8.5 视频类AIGC应用实践

案例：使用腾讯智影生成数字人播报视频（**可以用于制作企业教学视频**）

在浏览器地址栏中输入网址“https://zenvideo.qq.com/”，进入“腾讯智影”平台，点击平台首页“智能小工具”栏目中的“数字人播报”按钮或者“智能小工具”上方的“数字人播报”按钮，进入“数字人播报”功能界面。

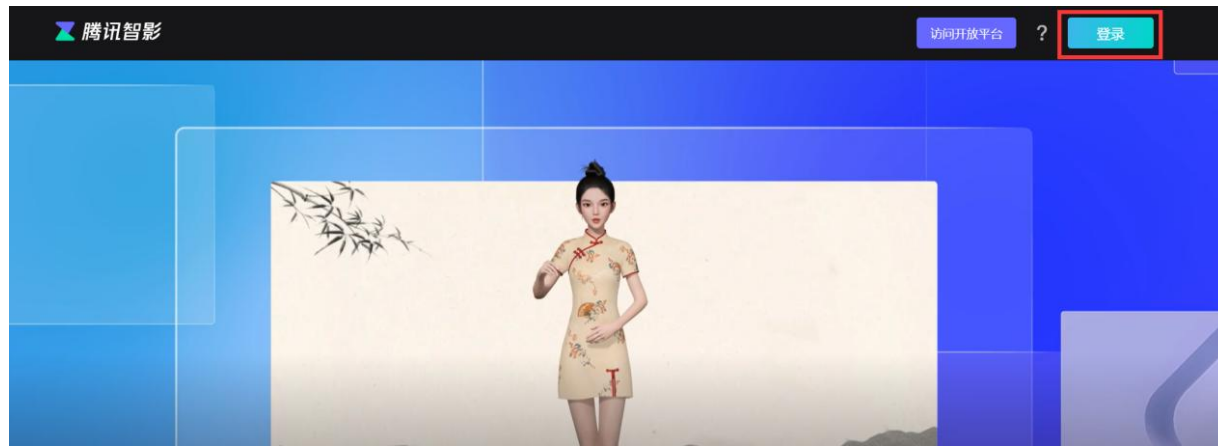


图 腾讯智影登录界面



图 “数字人播报” 功能入口



8.6 AIGC技术在辅助编程中的应用

AIGC技术在辅助编程中的应用日益广泛，它能够自动生成高质量的代码，从而显著提高开发效率



能够提供辅助编程服务的AIGC大模型包括Codex、GitHub Copilot、CodeGeeX、aiXcoder、豆包、通义灵码等

推荐企业使用AI编程工具： VS Code+DeepSeek, Cursor, 字节跳动Trae (AI原生集成开发环境)



8.6 AIGC技术在辅助编程中的应用



为什么要在 VSCode 中接入 DeepSeek?

- ❑ 代码智能补全：根据上下文自动补全代码，减少重复输入，提高编码效率。
- ❑ 代码生成：根据自然语言描述直接生成代码，例如快速搭建项目框架或实现特定功能。
- ❑ AI 对话辅助：随时与 AI 对话，获取编程建议、解决技术难题。

VSCode接入DeepSeek步骤

- ❑ 注册 DeepSeek 账号并获取API Key
- ❑ 安装 VSCode 插件
- ❑ 配置VSCode插件

具体操作步骤请参考网页：

<https://www.cnblogs.com/jinjiangongzuoshi/p/18726097>



8.7 AI搜索

AI搜索，即人工智能搜索引擎，是一种利用先进的人工智能技术，特别是深度学习和自然语言处理（NLP），来理解和响应用户的查询需求的新型搜索工具。它不仅仅是传统搜索引擎（比如百度）的简单升级，而是通过模拟人类的思维方式和行为模式，为用户提供更加精准、个性化且高效的信息检索服务

AI搜索通过收集和分析用户的历史搜索数据和行为模式，构建用户画像，从而实现更加精准的个性化搜索服务。这种数据驱动的智能决策机制，使得AI搜索能够不断自我优化，提升用户体验

纳米AI搜索是360公司在2024年12月推出的全新AI搜索应用，结合了自然语言处理、机器学习以及专家协同技术，致力于打破传统搜索引擎的局限，提供智能化、多样化的搜索体验

多模态搜索

支持文字、语音、拍照、视频等多种输入方式，满足不同场景下的需求，实现“一切皆可搜索”

智能工具集成

内置16款顶尖大模型，如豆包、文心一言等，为用户提供一站式AI智慧体验

慢思考模式

通过专家协同和多模型协作，深入分析复杂问题，提供更专业、更全面的答案

 纳米AI搜索





8.8 AI智能办公



WPS Office

AI写作助手 (帮我写、帮我改、AI伴写)

AI设计助手 (AI排版、AI格式)

AI阅读助手 (全文总结、文档问答、划词解释和翻译)

AI数据助手 (AI写公式、AI数据分析)





8.8 AI智能办公

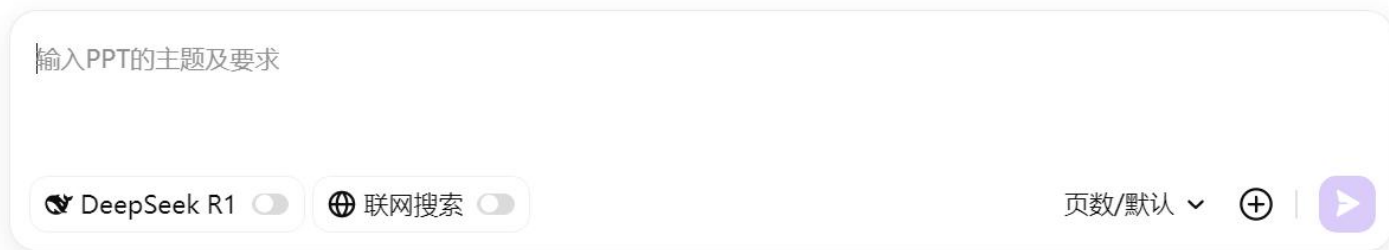
WPS灵犀

- 一键生成PPT
- 阅读报告，翻译报告
- 自动数据分析



AI PPT

输入灵感，灵犀一键生成 PPT



AI 阅读

让灵犀解读课件、论文、网页，轻松又专业



9. 大模型未来发展趋势





9. 大模型未来发展趋势

多模态融合与推理能力提升

多模态大模型将进一步融合文本、图像、音频、视频等数据，推动人机交互进入新高度。同时，大模型的训练将从“重规模”转向“重推理”，强化学习与自监督学习技术将优化模型的逻辑推理能力，使 AI 在复杂场景中表现更接近人类专家

生成式 AI 与具身智能深度融合

生成式 AI 将从内容创作扩展到物理世界交互。具身智能将推动机器人技术在工业、医疗等领域的落地，结合多模态能力，AI 不仅能生成创意内容，还能通过物理设备执行复杂任务

小模型崛起与高效化

为降低算力成本，轻量化小模型通过混合专家架构和知识蒸馏技术，以更低的能耗实现与大模型相当的性能。这类模型在本地化场景中将广泛应用，推动 AI 向边缘计算渗透

模型能力不断提升

在文本能力上，随着模型参数、训练数据和训练时间的增加，模型文本能力的上限不断刷新。在多模态能力上，实时对话延迟显著缩短，图片理解、视频生成、图像生成的基准测试最高分不断被突破。模型处理容量上，大模型的上下文窗口不断扩大

端侧大模型成为新增长引擎

随着大模型压缩和量化技术的不断提升，知识密度持续增大，终端搭载的模型能力值逐步增强。终端算力加速渗透，AI 处理的发展重心逐步从云端向手机、PC 等终端载体转移，将在消费电子领域实现广泛应用，并逐步推动工业领域的智能化升级



总结

大模型是人工智能领域的重要研究方向，其强大的语言理解和生成能力使得它在自然语言处理、机器翻译、智能客服等领域有着广泛的应用。大模型的训练需要大量的数据和计算资源，同时也需要先进的技术和算法支持。随着技术的不断发展，大模型的应用场景也在不断扩展，未来将会更加广泛地应用于各个领域。



谢谢

厦门大学大数据教学团队作品
2025年3月3日

