



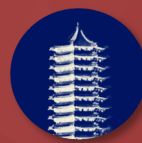
北京大学
PEKING UNIVERSITY

2025年卓越工程师系列课程

走进人工智能2.0

肖睿

2025年8月14日 杭州



- 北大青鸟人工智能研究院
- 北大计算机学院
- 北大教育学院学习科学实验室



一、人工智能的前世今生

二、大模型的原理和能力边界

三、大模型的现状和发展：技术应用和人才要求

人工智能的前世今生：1956-2025



北京大学
PEKING UNIVERSITY

■ 人工智能：Artificial Intelligence，AI

- 1956年：让机器具备人类智能，AGI，达特茅斯会议
- 2022年：让机器具备超人类智能（非人类智能），ASI，chatGPT发布

■ 人工智能发展的4个时代

1. 古代（1956-1996）：通过规则和知识让机器具备人类智能
2. 近代（1996-2006）：通过数据和学习让机器具备人类智能
3. 现代（2006-2020）：通过神经网络和深度学习让机器具备人类智能
4. 当代（2020-2025）：通过大模型让机器具备人类智能和超人类智能（非人类智能）

人工智能0.0：古代（1956-1996），规则 and 知识时代



北京大学
PEKING UNIVERSITY

➤ 第一次浪潮（1956-1968）：规则推理

- 思想准备：冯诺依曼计算机、机器思考、图灵测试
- 哲学基础：唯理论
- 主要原理：制定规则，进行数理推理（确定性，不确定性的概率）
- 主要成就：下棋程序，定理机器证明，MIT的搬箱机器人

➤ 第二次浪潮（1986-1996-？）：知识推理

- 哲学基础：唯理论+经验论
- 主要原理：知识工程（知识抽取 and 知识表达），专家系统（知识库+推理机）
- 主要成就：石油勘探、气象预报、军事决策、经济预测等；包括1996年之后的深蓝国际象棋、Watson

➤ AI=IT：这个时代，人工智能基本等同于软件程序

人工智能0.1：近代（1996-2006），机器学习时代



北京大学
PEKING UNIVERSITY

➤ 第三次浪潮（1996-？）：机器学习（数据+学习），小数据集，特征工程

- 使用统计学习方法建模，三个核心要素：模型、目标、策略
 - ✓ 模型是核心：逻辑回归，决策森林，支持向量机，马尔科夫链，人工神经元.....
 - ✓ 建模方法从规则到学习：从数学模型（分析数学），到数据模型（计算数学）
 - ✓ 数据模型的能力边界：可以用数据模型模拟世界（数字化、全景化），以史为鉴（IID）
- 主要成就：风险识别（金融、工业、经济）、学术研究

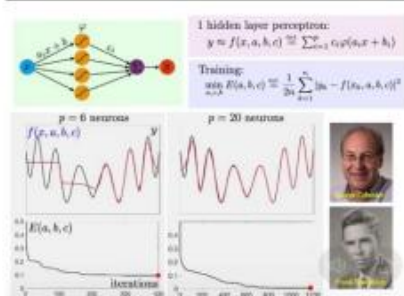
➤ 连接主义（1946-2006）：神经网络模型（黑盒）

- 属于机器学习的一种方法，模型采用的是人工神经网络
 - ✓ 人工神经网络与人脑最大的共同点是名字（原理、机制和架构并不一样），用神经网络表达数学模型
 - ✓ 传统神经网络：霍普菲尔德网络，玻尔兹曼机，.....
 - ✓ 深度神经网络：深度学习（Hinton，2006）

➤ 软件2.0：人工智能是数学、物理学、计算机科学的混合体

Multilayer Feedforward Networks are Universal Approximators

KURT HORNIK
Technische Universität Wien



人工智能1.0：现代（2006-2020），深度学习时代



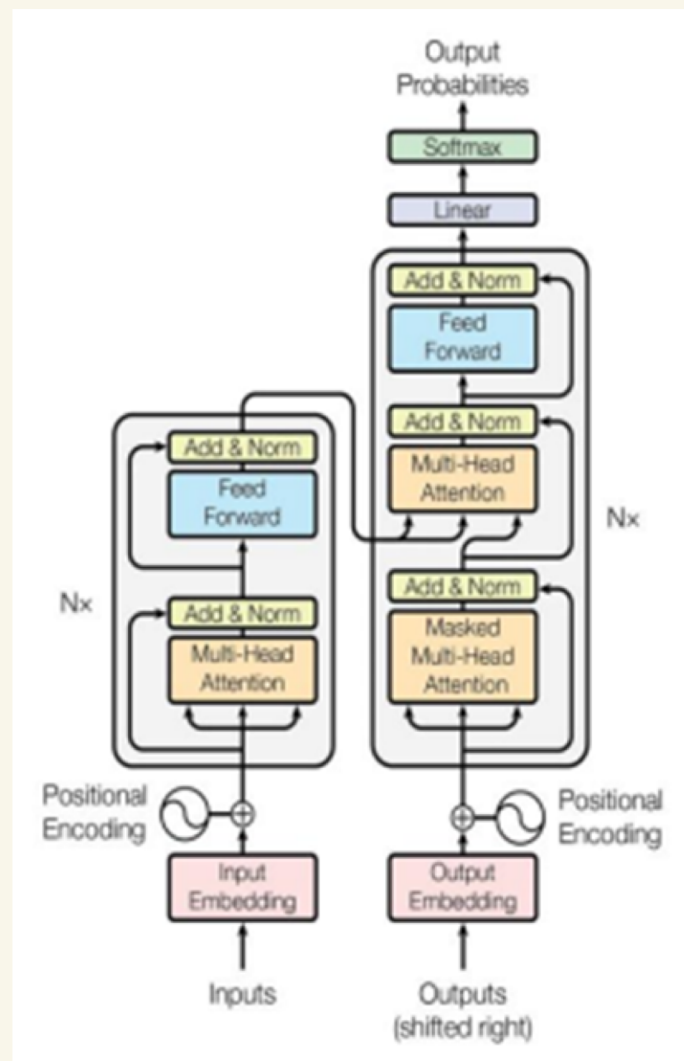
北京大学
PEKING UNIVERSITY

➤ 传统深度学习（2006）：深度神经网络模型，中数据集，端到端

- 多种神经网络模型：DBN，CNN，RNN，ResNet，Inception，RWKV，.....
 - ✓ AlphaGO：2016年超过人类棋手
 - ✓ ImageNet：2017年超过人眼
 - ✓ AlphaFold：2022年超过人类科学家，2024年获得诺贝尔奖
- 主要成就：人脸识别、图像识别、语言翻译、语音识别、物理建模.....

➤ 现代深度学习（2017）：Transformer模型，大数据集，注意力机制（大规模并行）

- 三种Transformer模型架构：并行矩阵计算（GPU）：堆叠架构，容易扩展
 - ✓ 编码器（BERT）：embedding，Ernie1.0，.....
 - ✓ 混合网络：T5、GLM（早期）
 - ✓ 解码器（GPT）：生成式人工智能（AIGC），大力出奇迹（大模型，2020）



人工智能2.0：当代（2020-2025），大模型时代

➤ 模型服务（2020）：MaaS

- 全部是Transformer的GPT架构（解码器）：生成式人工智能（AIGC）

- 大模型：预训练，生成-理解-决策（RL）

✓ 大（数据多、参数多、算力多）：B级（Billion，10亿）

✓ 模型：语言、视觉、多模态

- Transformer：大语言模型（LLM，大模型），多模态模型

ChatGPT（5、4.1、4o、o1、o3、o4）、Claude；Grok、Gemini；Llama、.....

DeepSeek、Step、Qwen；Kimi、MiniMax；GLM、火山(豆包)、元宝、百度.....

- Transformer+Diffusion：视觉模型

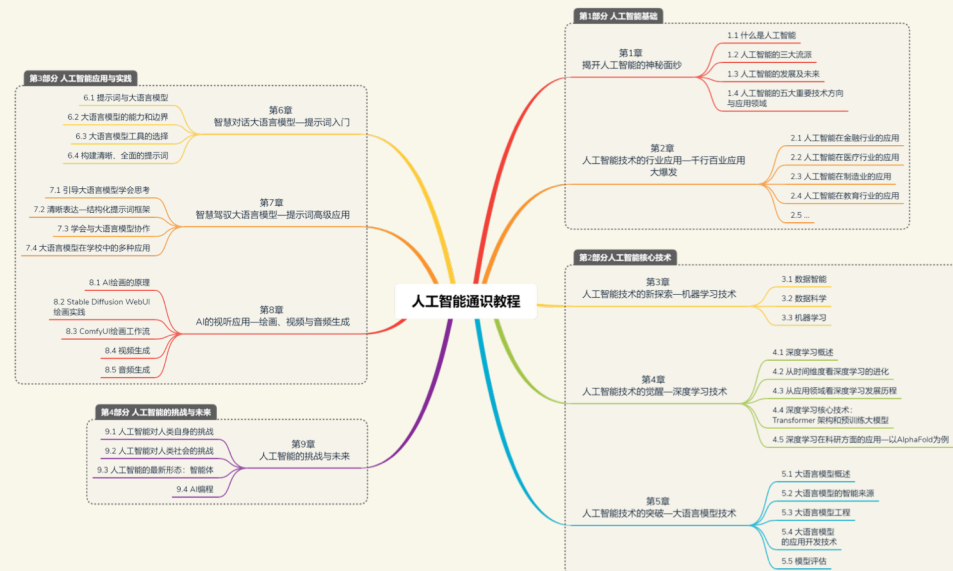
图像：Stable Diffusion、Mid-Journey、DALL-E等

视频：Sora、可灵、即梦、Vidu、海螺、veo等

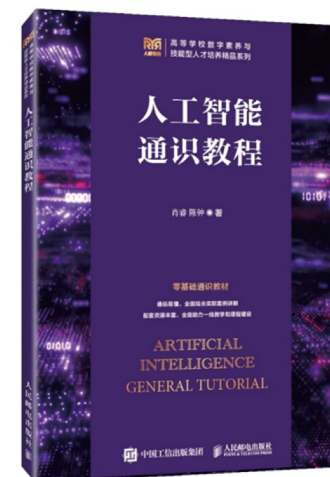
✓ 模型：通用模型 vs 垂直模型（行业模型）

- 接近成熟：传媒、广告、编码、电商等

- 正在发展：教育、医疗、金融、工业、农业等



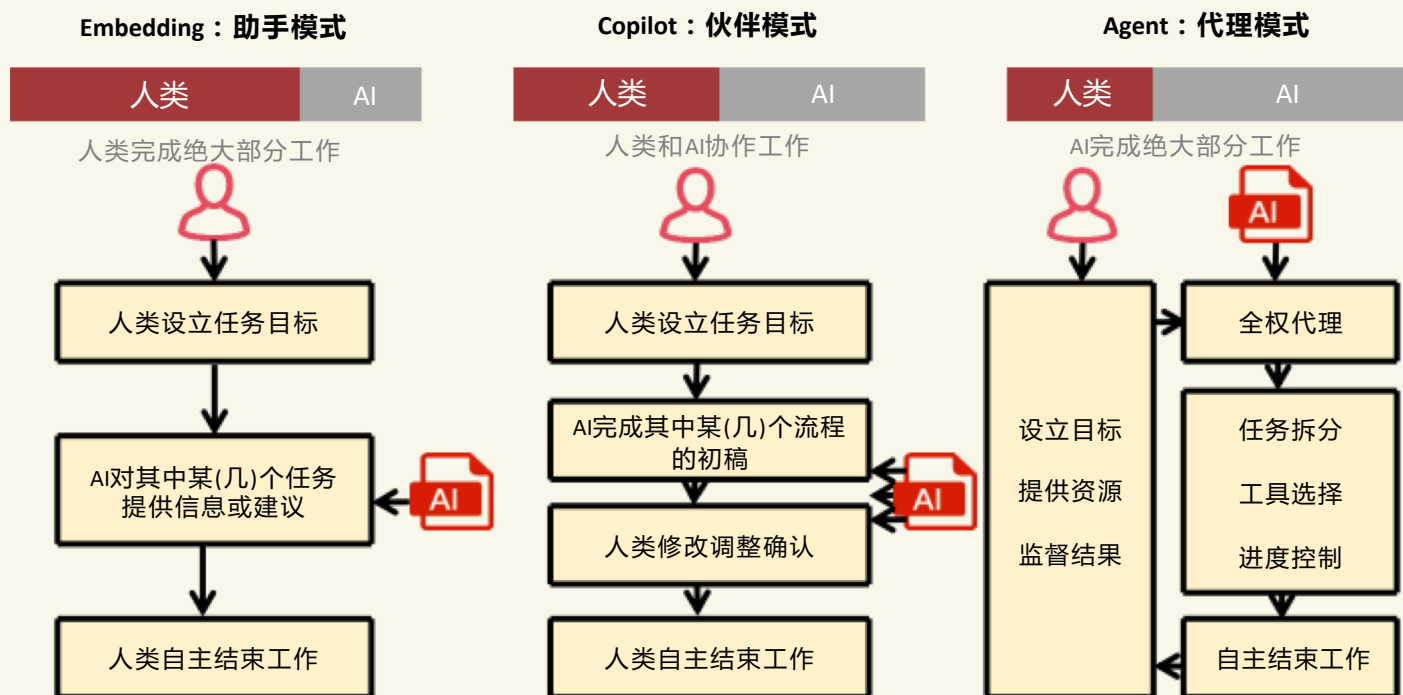
三大核心组件



人工智能2.0：当代（2020-2025），大模型时代

➤ Agent（2025）：大模型是新手机，Agent就是app

- 大模型的应用已从简单的对话工具（chatGPT、豆包）进化为：任务导向、交付结果的Agent（Manus、Lovart、Lovable等）
- Agent的核心架构是：感知（多模态）、决策（LLM；记忆、检索、上下文）、动作（具身驱动、工具调用）
- Agent的核心特征是自主（请人类走开）：从human in loop到human on loop



■Agent是传统应用（工具+流程）的AI改造

●工具:

- 直接使用模型或AI产品
- 增强模型能力（RAG、FT）
- 直接调用外部工具

●流程:

- Workflow（人工定义）
 - Coze、Dify、ComfyUI
- Agentic AI（模型拆解：环境、工具、策略）
 - 斯坦福小镇（MetaGPT）
 - AutoGLM、Manus、Coze Space、Aipy

■趋势：A2A和MCP将成为AI系统（Agent）的必备要素

●如果把Agent想象成一个笔记本电脑:

- 大模型就是CPU
- A2A就是网络接口
- MCP就是USB接口

人工智能2.0：当代（2020-2025），大模型时代

➤ 软件2.0时代

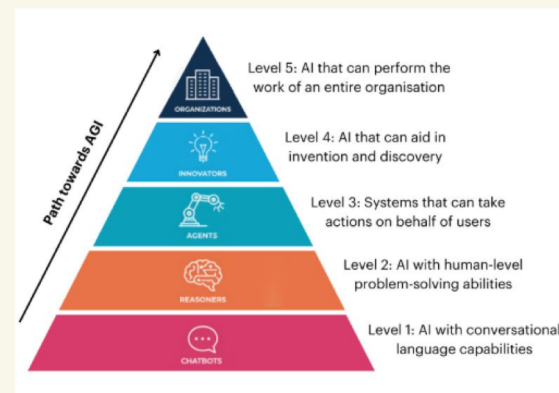
- AI0.1（机器学习）：小数据，人工特征，部分可解释，不具备通用性和跨模态
- AI1.0（深度学习）：大数据，特征表示，基本不可解释，不具备通用性和跨模态
- AI2.0（大模型）：海量数据，自监督学习，完全不可解释，具备通用性和跨模态，最有可能通向AGI的路径

通用人工智能

AGI , Artificial General Intelligence

具备与人类同等智能水平或超越人类智能水平的人工智能系统。

- ✓ OpenAI：在大多数经济价值创造任务中表现优于人类的高度自主系统。
- ✓ AI肖睿团队：90%的智力任务上超过90%的人类，很可能在2030年之前到来。



Level 5.

AI能力全面超越人类，具备探究科学规律、世界起源等终极问题的能力

Level 4.

AI通过自我学习，具备自我批判、自我改进以及自我反思能力

Level 3.

AI学会使用工具，利用工具完成多数人类物理世界问题，在工具使用方面突破图灵测试

Level 2.

AI学会求解问题，涌现世界知识和类人的复杂逻辑推理能力，在问题求解方面突破图灵测试

Level 1.

AI学会使用人类语言，在大多数自然语言任务上突破图灵测试

第五级

• 组织者，可以完成组织工作的人工智能

第四级

• 创新者，可协助发明创新的人工智能

第三级

• 智能行动者，可以自动采取行动的系統

第二级

• 推理者，达到人类水平的解决问题水平

第一级

• 聊天机器人，具有对话功能的人工智能

<https://developer.baidu.com/article/details/3321696>

一、人工智能的前世今生

二、大模型的原理和能力边界

三、大模型的现状和发展：技术应用和人才要求

二、大模型的原理和能力边界



北京大学
PEKING UNIVERSITY

■1、大模型的工作原理

■2、大模型的工作过程

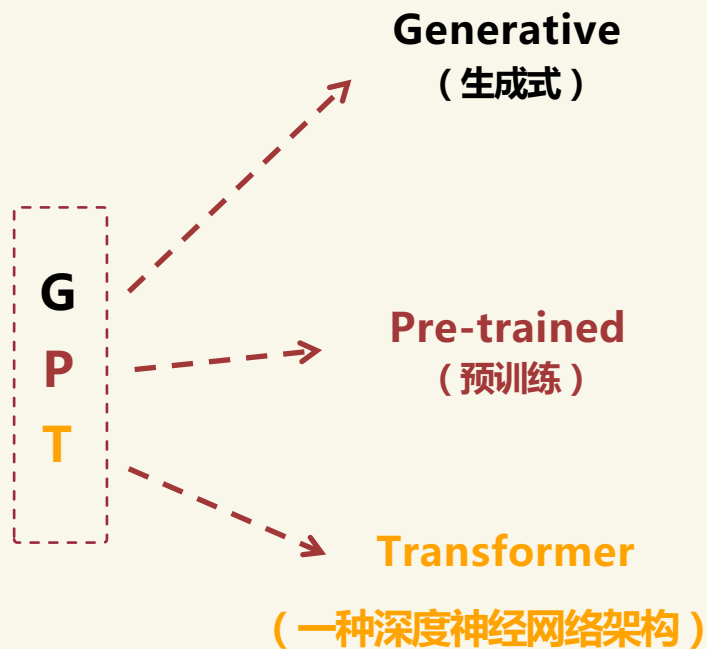
■3、大模型的能力边界

■4、大模型的应用场景

大模型的工作原理: NTP (Next Token Prediction)



北京大学
PEKING UNIVERSITY



LLM: Large Language Model

1. 收到提示词

示例: “今天天气不错, 我决定”

2. 将输入拆分为token

[“今天”, “天”, “气”, “不”, “错”,
“,”, “我”, “决定”]

3. 采用Transformer架构处理token

- 理解token之间的关系
- 识别提示词的整体含义

4. 基于上下文预测下一个token

- 为可能的单词分配概率分数
- 示例: { “去”: 0.7, “停”: 0.2, “站”: 0.1 }

5. 根据概率分数选择标记

示例: “去”

准确地讲, 这里不是“字”, 是“token”, 可以进行语义计算。

概率预测+文字接龙

自回归 (AR) : 重复步骤4和步骤5直到形成完整的句子

示例: 今天天气不错, 我决定去公园



大模型的工作过程：预训练-后训练-推理

大模型工作过程

预训练
(自监督)

监督微调

强化学习
(RLHF等等)

接收输入
(提示词)

处理输入
(上下文)

进行推理
(测试时计算)

生成输出

阶段1：模型训练（预训练+后训练）

阶段2：推理



大模型的最新发展（从原子弹到氢弹）：推理



28 Feb 2025



跃进期



繁荣期



深化期



准备期

- **ChatGPT发布**，全球范围内迅速形成大模型共识。
- **GPT4发布**，进一步掀起大模型研发热潮。
- **国内快速跟进大模型研发**，文心一言1.0、通义千问、讯飞星火、360智脑、ChatGLM等首批模型相继发布。

- **Llama2开源**，极大助力全球大模型开发者生态。
- **GPT-4 Turbo、Gemini**等海外大模型发布，持续提升模型性能。
- **Midjourney发布5.2**
- **Stable Diffusion XL发布**
- **国内闭源大模型快速发展**，豆包、混元、商汤3.0、盘古3.0、AndesGPT、BlueLM、星火3.0、Kimi Chat等陆续发布。
- **国内开源生态爆发**，Baichuan、Qwen、InternLM、ChatGLM3、Yi-34B等系列模型引领开源热潮。

- **OpenAI发布Sora**，极大拓展了AI在视频领域的想象力。
- **GPT-40、Claude3.5、Gemini1.5、Llama3**发布，海外进入“一超多强”的竞争格局。
- **国内多模态领域进展迅速，在部分领域领先海外**，视频生成模型可灵AI、海螺视频、Vidu、PixVerse等模型陆续发布，并在海外取得较大应用进展。
- **国内通用模型持续提升**，Qwen2.5、文心4.0、GLM4、商汤5.5等通用模型陆续更新。

- **OpenAI推出基于GPT-4o模型的图像生成功能**，取代此前的DALL-E 3成为ChatGPT和Sora平台的默认图像引擎
- **OpenAI发布o3-mini、GPT-4.5**，前者推动成本效益推理，后者展现出较高的情感智能。
- **Gemini 2.0 Flash Thinking、Claude-3.7-Sonnet、Grok3**发布，海外推理模型引发热潮，推理性能大幅度提升。
- **国内推理模型持续跟进**。DeepSeek-R1、QwQ-32B、Kimi1.5、GLM-Zero、Skywork o1、讯飞星火X1等推理模型陆续发布，继续突破推理能力的上限。
- **国内模型性能持续提升**。DeepSeek-V3、Qwen2.5、豆包-Pro、混元-Turbo与GLM-4-Plus等系列模型综合能力上持续提升。
- **国内开源生态持续引领模型普惠化**。DeepSeek-R1通过开源与性价比优势持续推动行业技术普惠化进程。

2022.12

2023.06

2023.12

2024.06

2025.03

生成模型



推理模型

ref: SuperCLUE团队 中文大模型基准测评2025年3月报告

1. 语言能力：理解和生成

2. 知识能力

- 幻觉（生成不符合事实的内容）
- 知识库限制（公开、私有、即时）
- 上下文窗口限制（记忆、成本）

3. 推理能力

■生成

1. 严肃内容+垃圾信息的混合
2. 知识量大，但缺少内在关联能力

■幻觉

1. 来源：有损压缩，NTP的温度
2. 观点：创意和创新

■记忆

1. 多轮对话：产品设计，计算成本

■ 解决方案

1. 提示词（Prompt）
2. 思维链（CoT）
3. 搜索增强（RAG）
4. 知识图谱（KGE）
5. 模型微调（Fine Tune）



生成模型和推理模型适用场景



比较项	生成模型（ GPT-4o、 DeepSeek-V3 ）	推理模型（ GPT-o3、 DeepSeek-R1 ）
模型定位	专注于通用自然语言处理，多模态能力突出，适合日常对话、内容生成、翻译以及图文、音频、视频等信息处理。	侧重于复杂推理与逻辑能力，擅长数学、编程和自然语言推理任务，适合高难度问题求解和专业领域应用。
推理能力	在日常语言任务中表现均衡，但在复杂逻辑推理（如数学题求解）上准确率较低。	在复杂推理任务表现卓越，尤其擅长数学、代码推理任务。
多模态支持	支持文本、图像、音频乃至视频输入，可处理多种模态信息。	当前主要支持文本输入，不具备图像处理等多模态能力；未来可能通过社区贡献扩展相关功能。
应用场景	适合广泛通用任务，如对话、内容生成、多模态信息处理以及跨语言交流；面向大众市场和商业应用。	更适合需要高精度推理和逻辑分析的专业任务，如数学竞赛、编程问题和科学研究；在思路清晰度要求高的场景具有明显优势，比如采访大纲、方案
用户交互体验	提供流畅的实时对话体验，支持多种输入模态；用户界面友好，适合大众使用。	可展示部分链式思考过程，便于用户理解推理过程；界面和使用体验具有较高的定制性，但整体交互节奏较慢。

- 生成模型是玩知识和文字的，推理大模型是玩逻辑和推理的，至于计算问题，还是找计算器吧
- 推理模型也不是万能的，其幻觉通常比生成模型大，很多不需要强推理的场合还是传统的生成模型比较适合

人工智能：快速冲击智力行业，逐渐侵蚀物理世界



北京大学
PEKING UNIVERSITY

人工智能：让机器具备人类智能，让机器具备非人类智能（超人类智能）

- 机器学习

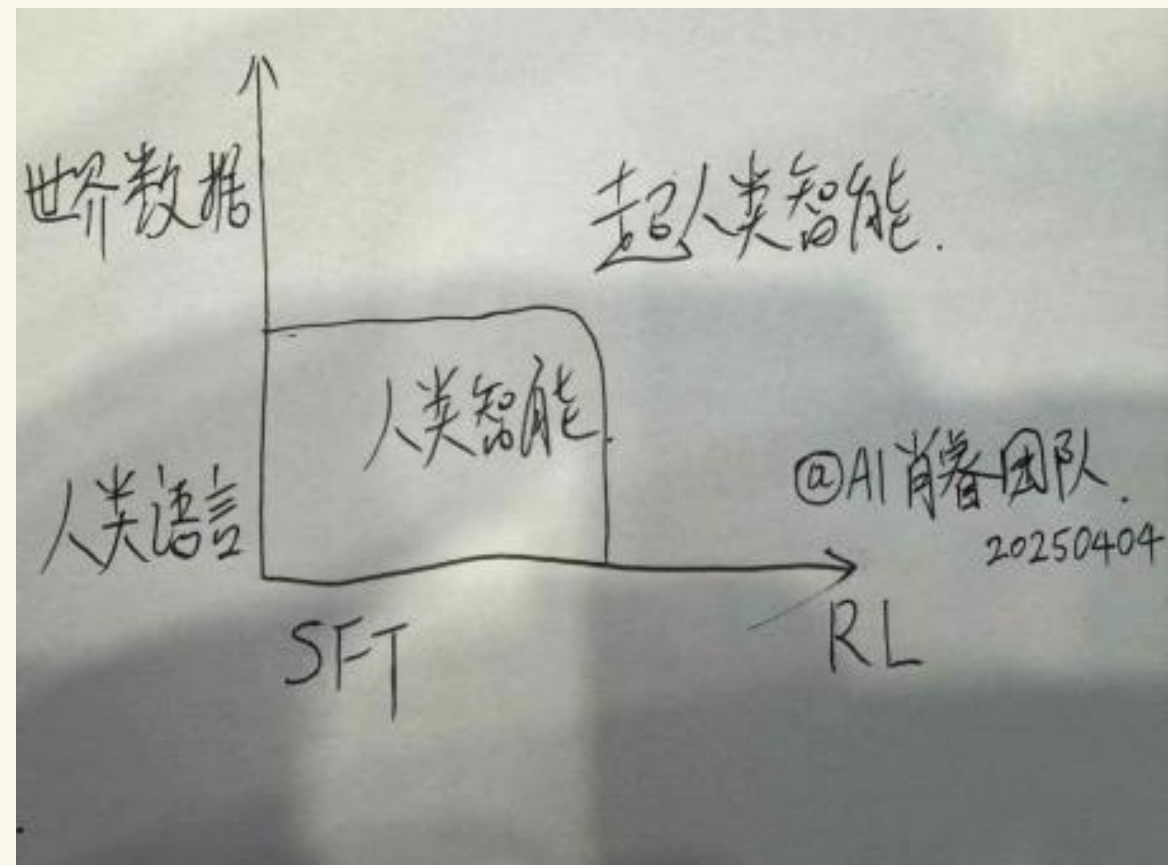
- 深度学习

- 大模型

- 大语言模型：DeepSeek

- 视觉模型：可灵、即梦

- 多模态模型：GPT-5



➤ 现代人工智能（大模型）的本质

1. 这一波人工智能本质上是数据智能，只要有时间结构和空间结构的数据，都可以识别出数据分布模式，建立数据模型，从而产生智能。
2. 这一波人工智能的核心是语言智能，通过分析和建模人类语言，获取人类的知识，并进一步获取人类的思维模式。
3. 或许，AI只是一个我们和他人和祖先和整个人类的意义世界的交互的接口的翻译器。与我们对话的，不是AI，而是AI背后那个人类构造出的意义世界。因此，AI可以成为我们的伙伴和导师，例如：DeepSeek对贪嗔痴的解释。

➤ 大模型技术的关键过程

1. 预训练：中小学，打基础
2. 后训练：RL、SFT。大学，有专业
3. 微调、Prompt。入职实习，能干活

➤ 大模型技术的关键要素

1. Token：万物皆token
2. Attention：熵减即智能
3. GPT：大力出奇迹
4. Data：以古鉴今
5. RL：自学成才（决策，探索未知，生成数据）
6. 优化：卷Infra和算法，实事求是，反抽象
7. FT：后训练的艺术
8. TTC：大力出奇迹 AGAIN
9. Prompt：有话好好说
10. Agent：最后的筐

对现代人工智能的正确认知：AI与IT的区别

■ IT：确定性的任务（简单和繁杂），以代码逻辑为核心

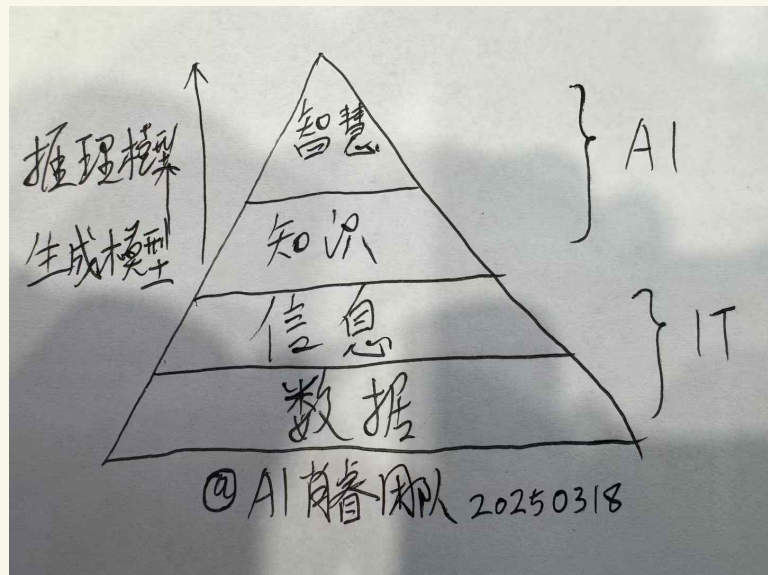
- 1.0：记忆+计算（冯诺依曼；软件时代）
- 2.0：记忆+计算+搜索（互联网时代）

■ AI：不确定的任务（复杂和混沌），以数据模型为核心

- 0.0：专家系统：知识+规则
- 0.1：机器学习：数据+学习，白盒
 - 0.1：传统机器学习：人类定义特征，人类估算模型参数
 - 0.5：人工神经网络：人类定义特征，模型自己学习模型参数
- 1.0：深度学习：数据+学习，灰盒（不可控，有错误概率，可解释）
 - 1.0：判别模型：人类只提供数据，端到端学习（模型自己抽取特征，自己学习模型参数）
- 2.0：大模型：数据+学习，黑盒（不可控，有错误概率，不可解释）
 - 2.0：生成模型：训练阶段+推理阶段；压缩+生成（幻觉）
 - 2.1：推理模型：训练阶段强化学习；推理阶段慢思考
 - 2.2：Agent：任务结果导向（感知+大模型上下文工程+行动）

■ IT应用与AI应用的关键差异：

- 在互联网时代，用户使用系统的成本很低，边际成本接近于零。
- 在AI时代，用户使用系统的成本比较高，有大量的GPU算力需求，边际成本较高。



已深度思考（用时 0 秒） ^

服务器繁忙，请稍后再试。



一、人工智能的前世今生

二、大模型的原理和能力边界

三、大模型的现状和发展：技术应用和人才要求

三、大模型的现状和发展：技术应用和人才要求



北京大学
PEKING UNIVERSITY

- 1、DeepSeek现象分析
- 2、大模型的发展趋势
- 3、如何使用大模型
 - ✓ 1) 提示词工程和对话工具
 - ✓ 2) 上下文工程和Agent
- 4、AI2.0时代的人才要求

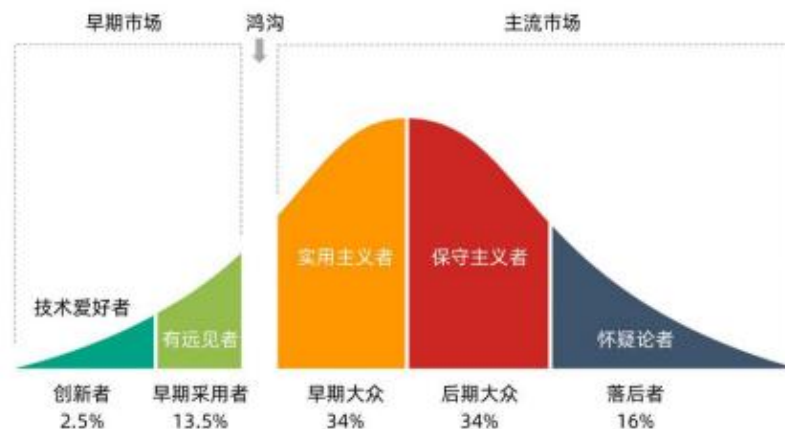
现象：DeepSeek快速出圈，全民硬控

- 2024年12月26日，DeepSeek推出对标OpenAI GPT-4o的 语言模型DeepSeek V3，随后在美国AI行业内部引起轰动。
- 2025年1月20日，DeepSeek发布对标OpenAI o1 的DeepSeek R1大语言模型，并于1月24日引起美国投资界KOL关注。
- 2025年1月26日，关于DeepSeek颠覆了大模型的商业模式（堆算力、拼资本），引发英伟达股价大跌，DeepSeek首先在美国出圈，引发国际社会讨论。
- 2025年1月底（春节前后），DeepSeek在中国出圈，并上升到中美竞争高度，同时纷纷接入DeepSeek，DeepSeek成为AI和大模型的代名词。

■ DeepSeek让AI跨越了鸿沟。



2025年1月20日下午，中共中央政治局常委、国务院总理李强主持召开专家、企业家和教科文卫体等领域代表座谈会，听取对《政府工作报告（征求意见稿）》的意见建议。DeepSeek公司创始人梁文峰作为企业家代表之一参加了此次座谈会。



到底谁是DeepSeek？公司、模型、产品



北京大学
PEKING UNIVERSITY

系统组成

部署方案

SaaS

应用服务（网页、APP、桌面软件、设备软件）

- 生产办公系统：目前多数为定制开发（RAG、Agent等）
- 对话机器人：目前多数为标准产品，使用云服务
 - 免费：**DeepSeek**、豆包、Kimi、元宝、Monica、秘塔搜索、问小白、Molly R1等
 - 收费（月费；充值）：文小言、通义、可灵、即梦、Vidu、Liblib、ChatGPT等

MaaS

模型服务（数据工程、推理加速、训练框架、API调用）

- 开源模型：**DeepSeek**、Qwen、Llama、SD、微调模型等
 - 向上提供：私有化部署，公有云服务（DeepSeek官网、硅基流动、火山引擎等）
- 闭源模型：OpenAI、Claude、Gemini、Kimi、百度等
 - 向上提供：公有云服务（OpenAI官网）

PaaS

平台服务（存储、计算、数据、安全、中间件）

IaaS

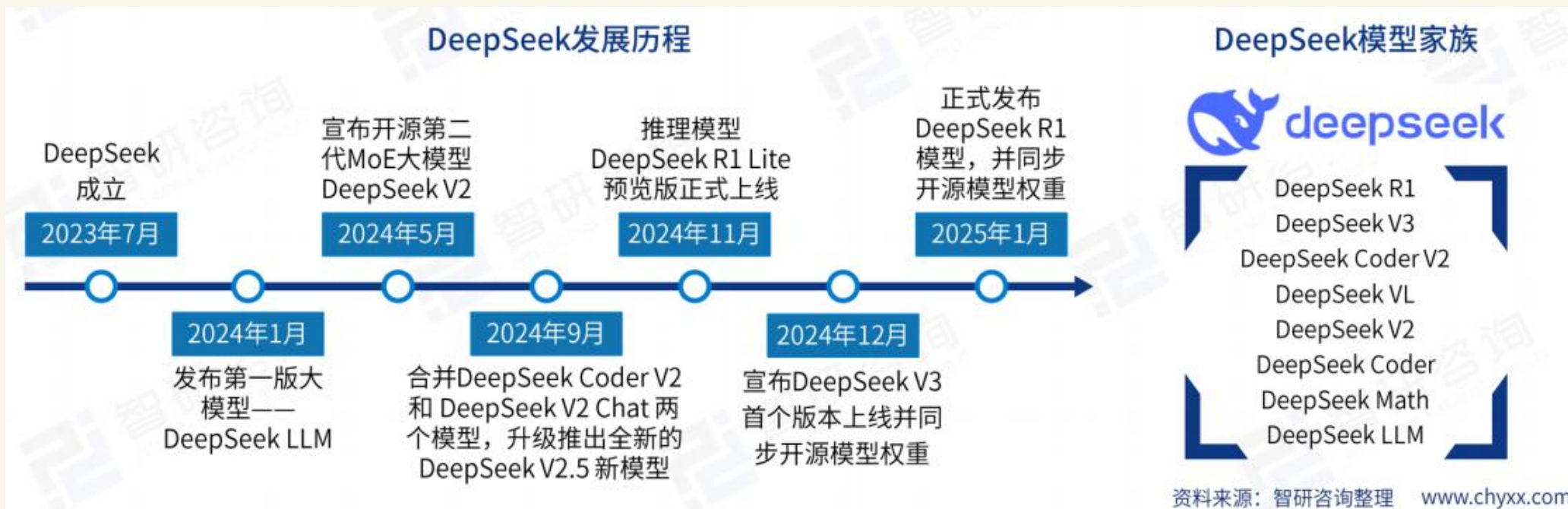
硬件服务（设备/电脑/服务器/GPU、网络、操作系统）

- 私有化部署
 - 本地：个人设备或电脑、企业服务器
 - 数据中心（IDC）：企业服务器、服务器集群（私有云）
- 公有云服务：阿里云、火山云、腾讯云、金山云、华为云等；AWS、Azure等

到底谁是DeepSeek？公司、模型、产品



北京大学
PEKING UNIVERSITY



■ DeepSeek于2023年5月成立北京公司，2023年7月成立杭州公司，是幻方量化孵化出的一家大模型研究机构（目标AGI），160人的团队分布在杭州和北京，是中国AI2.0（大模型）时代的七小龙之一。

• 注：AI1.0（深度学习）时代的四小龙：商汤、旷世、云从、依图

■ 除了DeepSeek公司之外，其它六家也被投资界称为中国大模型企业六小虎（智谱AI、百川智能、月之暗面、阶跃星辰、MiniMax、零一万物）。

到底谁是DeepSeek？公司、模型、产品



Model	Base Model	
DeepSeek-R1-Distill-Qwen-1.5B	Qwen2.5-Math-1.5B	蒸馏模型，能力稍弱 ■ 实际上是增加了推理能力的Qwen模型和Llama模型，不能称为DeepSeek模型。 ■ 市场上有误解，厂商有误导，甚至Ollama工具的模型选项中也有误导。
DeepSeek-R1-Distill-Qwen-7B	Qwen2.5-Math-7B	
DeepSeek-R1-Distill-Llama-8B	Llama-3.1-8B	
DeepSeek-R1-Distill-Qwen-14B	Qwen2.5-14B	
DeepSeek-R1-Distill-Qwen-32B	Qwen2.5-32B	
DeepSeek-R1-Distill-Llama-70B	Llama-3.3-70B-Instruct	满血版，能力最强
DeepSeek-R1-671B	DeepSeek-V3-Base	

到底谁是DeepSeek？公司、模型、产品



北京大学
PEKING UNIVERSITY

公司、模型、产品

- 引起中美AI行业内部关注的是：模型
- 引起美国关注的是：模型和公司
- 春节前后在中国出圈的是：产品

产品优势（用户可感受到）

- 思考过程展示
- 中文好有情商
- 容易获得，使用简单
- 联网和来源引用
- 速度快
- 免费



能力突破

开源、低成本、国产自主

DeepSeek以“推理能力+第一梯队性能”为核心基础，叠加：开源开放、超低成本、国产自主研发三大优势，不仅实现技术代际跨越，更推动AI技术普惠化与国产化生态繁荣，成为全球大模型赛道的重要领跑者。

混合专家
MOE

多头潜注意力
MLA

多Token预测
MTP

强化学习
GRPO

测试时计算
TTC

混合精度训练
FP8

通讯优化
DualPipe

并行训练框架
HAI

直接硬件编程
PTX

■ 基础能力：进入推理模型阶段，并跻身全球第一梯队

1. **推理能力跃升**：DeepSeek大模型核心技术突破，实现复杂推理任务的精准处理与高效执行，覆盖多模态场景应用。
2. **国际竞争力对标**：模型综合性能跃居全球第一梯队，技术指标与国际顶尖水平（如GPT系列、Claude等）直接对标，奠定国产大模型的行业标杆地位。

■ 核心加分项：开源、低成本、国产自主

1. 开源：技术共享，生态共建

- 全量开源训练代码、数据清洗工具及微调框架
- 开发者可快速构建教育、金融、医疗等垂直领域应用，推动协同创新。

2. 低成本：普惠企业级AI应用

- 针对H系列芯片做了大量的模型架构优化和系统工程优化。
- 最后一次训练成本仅\$557w：显著低于行业同类模型，打破高价壁垒。
- 推理成本降低83%：千亿参数模型适配中小企业需求，加速商业化落地。

3. 国产自主：技术自主，缩短差距

- 将国产模型与美国的代际差距从1-2年缩短至3-5个月，突破“卡脖子”技术瓶颈。
- 构建多行业专属模型矩阵，全面支持国内产业智能化升级。

DeepSeek V3/R1模型的创新

一、基础架构：

1. **混合专家模型 (MoE)**：DeepSeek采用MoE架构，通过动态选择最适合输入数据的专家模块进行处理，提升推理能力和效率。
2. **无辅助损失的专家负载均衡策略 (EP)**：该策略使DeepSeekMoE在不对优化目标产生干扰的前提下，实现各个专家的负载均衡，避免了某些专家可能会被过度使用，而其他专家则被闲置的现象。
3. **多头潜在注意力机制 (MLA)**：MLA通过低秩压缩减少Key-Value缓存，显著提升推理效率。
4. **强化学习 (RL)**：DeepSeek-R1在训练中大规模应用强化学习(让模型自我探索和训练)，将传统的PPO替换为GRPO训练算法，显著提升推理能力。
5. **多Token预测 (MTP)**：通过多Token预测， Deepseek不仅提高了推理速度，还降低了训练成本。

二、训练及框架：

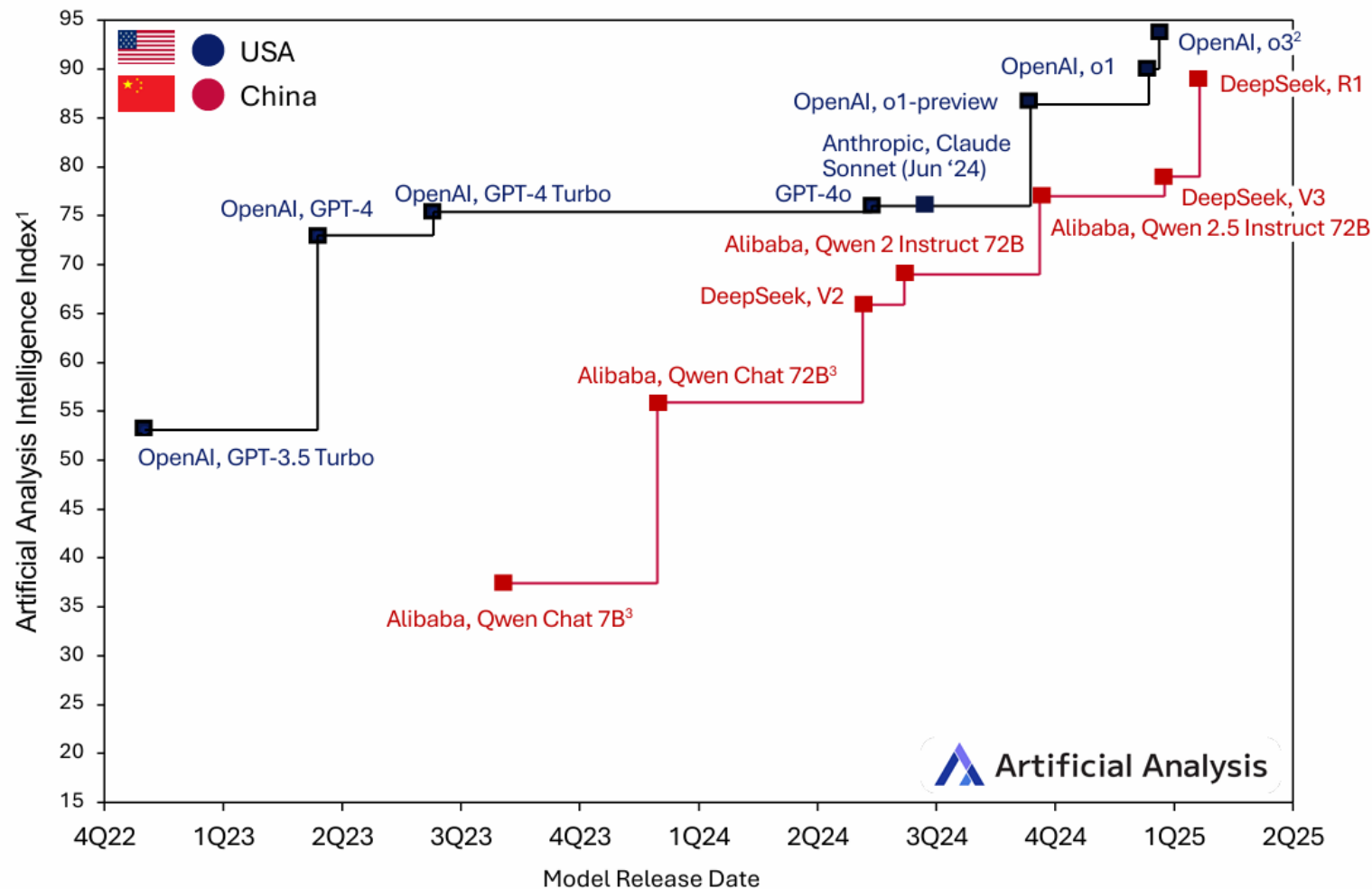
1. **FP8混合精度训练 (FP8)**：在关键计算步骤使用高精度，其他模型层使用FP8低精度进一步降低训练成本。**这一点，是DeepSeek团队非常有价值的创新和突破。**
2. **长链推理技术 (TTC)**：模型支持数万字的长链推理，可逐步分解复杂问题并进行多步骤逻辑推理。
3. **并行训练系统 (HAI)**：16 路流水线并行(Pipeline Parallelism, PP)、跨 8 个节点的 64 路专家并行(Expert Parallelism, EP)，以及数据并行(Data Parallelism, DP)，大幅提升模型训练速度。
4. **通讯优化 (DualPipe)**：高效的跨节点通信内核，充分利用 IB 和 NVLink 带宽特点，减少通信开销，提高模型推理性能。
5. **混合机器编程 (PTX)**：部分代码直接使用PTX编程提高GPU运行效率。
6. **算子库优化 (GEMM等Op)**：针对H800计算卡的特点，优化了一部分CUDA的算子库。

三、社会价值：

1. **开源生态**：使用最为开放的MIT开源协议，吸引了大量研究人员和应用厂商，推动了AI技术的发展。
2. **模型蒸馏支持**：DeepSeek-R1同时发布了多个模型蒸馏。
 - 虽然这些蒸馏模型的生产初衷是为了验证蒸馏效果，但客观上帮助用户有机会使用移植了DeepSeek-R1满血版模型的能力的更小的模型，以满足不同应用场景需求。
 - 副作用是：给市场和用户造成了很多困扰。
3. **AI产品和技术的普及**：
 - 对于大模型研发企业，更加重视infra工程的价值了。
 - 对于大模型应用企业，有了更多高效低成本解决方案。
 - 对于社会大众，认识到AI是一个趋势，不是昙花一现。
 - 对于市场，用户开始主动引入AI，不再怀疑了。
 - 对于国家，大幅缩小了中美的核心技术差距。
 - 对于全人类，技术平权，造福一方。

DeepSeek R1模型的能力突破

US & China: Frontier Language Model Intelligence, Over Time¹



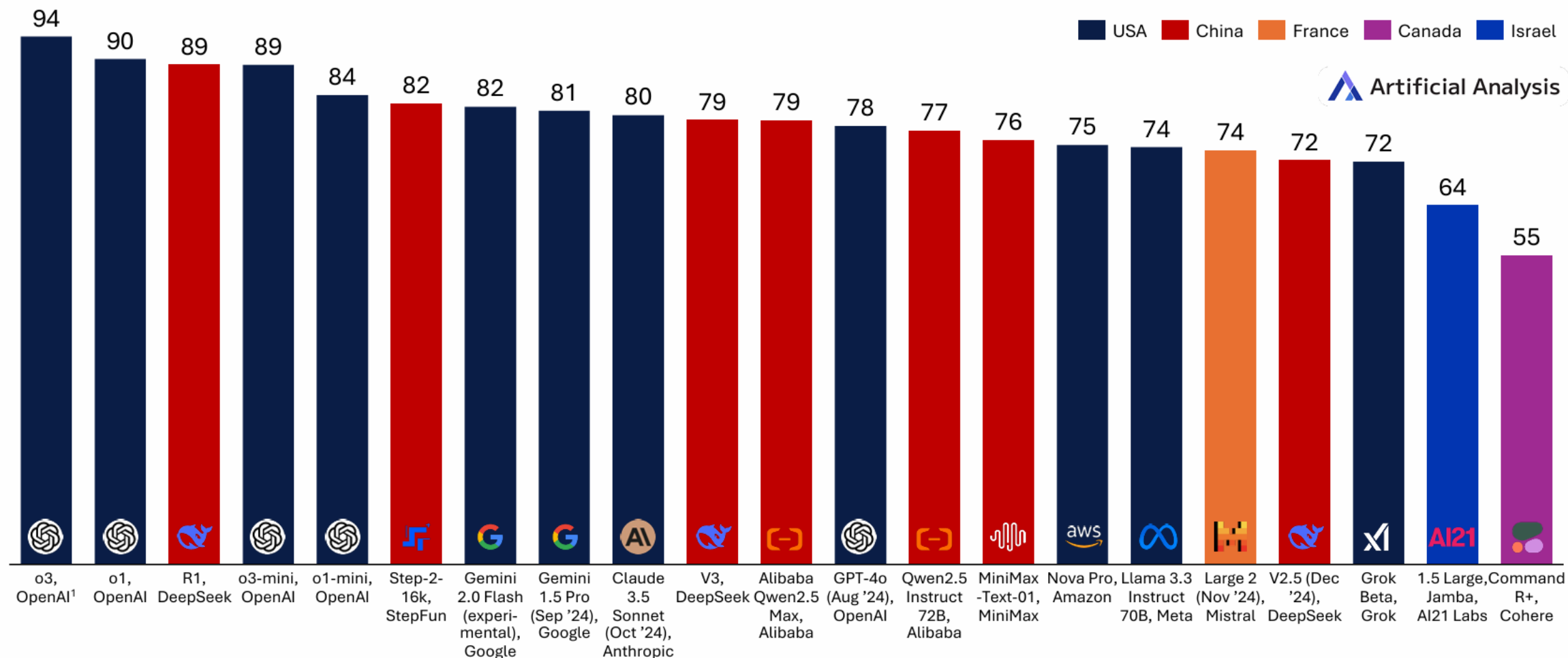
DeepSeek R1模型的能力突破



北京大学
PEKING UNIVERSITY

The Language Model Frontier: Country of Origin

Artificial Analysis Intelligence Index, Selected Leading Models (Early 2025), Non-exhaustive



DeepSeek R1模型的能力突破



北京大学
PEKING UNIVERSITY

Intelligence vs. Price

Artificial Analysis Intelligence Index (Version 2, released Feb '25); Price: USD per 1M Tokens

Most attractive quadrant

Legend: o1, o3-mini, GPT-4o (Nov '24), GPT-4o mini, o3-mini (high), Gemini 2.0 Flash, Claude 3.5 Sonnet (Oct), Claude 3.5 Haiku, Claude 3.7 Sonnet, DeepSeek R1, DeepSeek V3, DeepSeek R1 Distill Qwen 32B, Nova Pro, MiniMax-Text-01, Qwen2.5 Max



DeepSeek R1 对大模型行业的重大影响

01

打破垄断

DeepSeek-R1以低成本和开源特性打破以往头部企业巨头割据局面

02

价格下调

DeepSeek-R1的API定价仅为行业均价的1/10,推动了中小型企业低成本接入AI,对行业产生了积极影响

03

推动创新

DeepSeek-R1促使行业开始从“唯规模论”转向更加注重“性价比”和“高效能”方向

■ 直接使用官方服务

- 访问官网(ai.com 或 <https://chat.deepseek.com/>)，登录后使用，适合电脑端快速使用，但存在服务不稳定问题。
- 下载APP，适合手机平板等移动设备，但同样存在服务不稳定问题。

■ 使用第三方服务与API调用

- 第三方服务：秘塔AI、微信搜索、Molly R1、问小白等。
- API调用：DeepSeek、硅基流动、火山引擎等。获取API密钥调用，适合开发者集成，同样存在服务不稳定问题。

■ 本地部署

- 个人部署：个人在本地设备运行应用，依赖自身计算资源，灵活便捷。
- 企业部署：企业内网或云端搭建私有化系统，支持多用户协作，数据可控。
- 一体机：直接购买配置了DeepSeek模型的具备一定算力的一体机。

DeepSeek-R1 应用场景的特点

1. 推理密集型任务

- 编程任务中的代码生成、算法设计，媲美Claude 3.5 Sonnet。
- 数学问题求解、科学推理和逻辑分析等需要复杂推理的场景。

2. 强推理任务

- 需要基于大量知识，进行长链分析和强推理的复杂任务。
- 例如：工作方案、科研实验、病理诊断、命理玄学、市场分析、舆情分析、教育教学等。

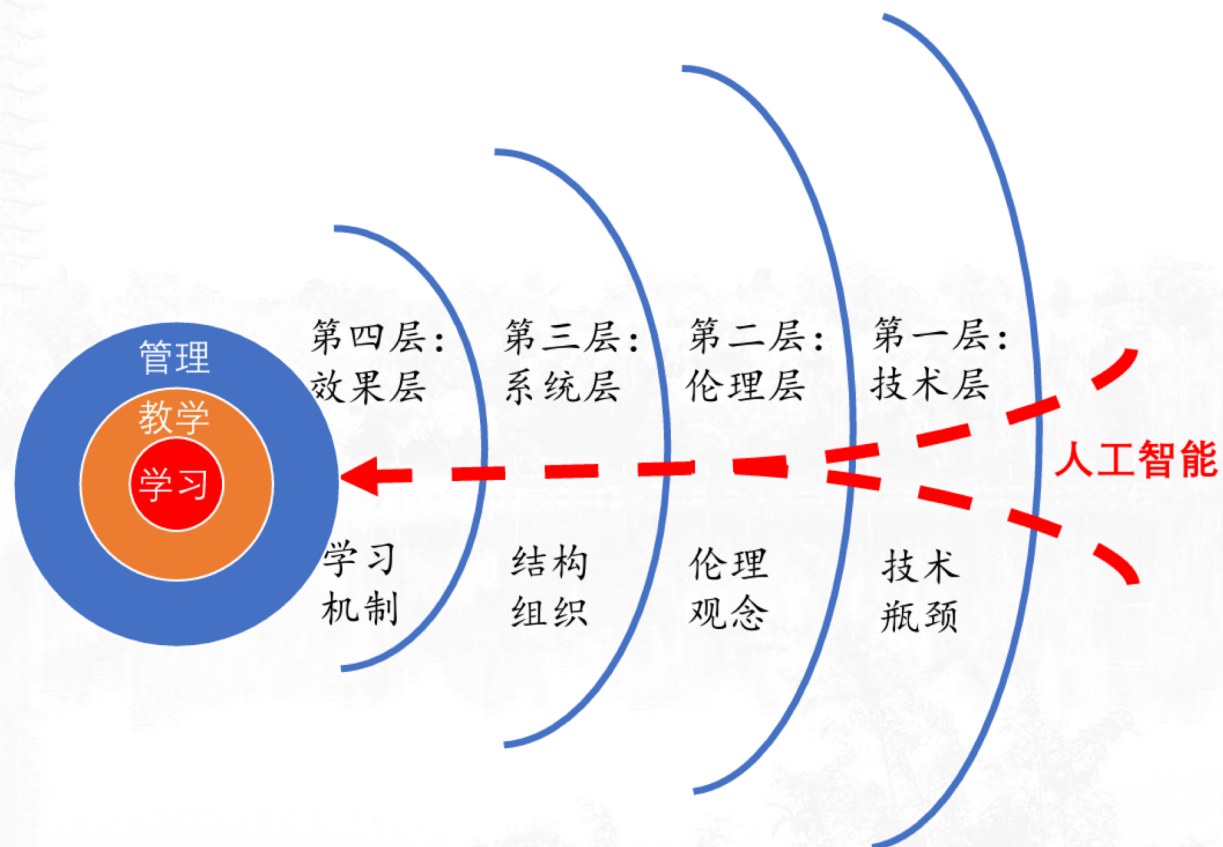
3. 中文要求高的任务

- 理解和使用中国文化背景、习惯用语、古文诗词（如生成春联）。优于ChatGPT的中文能力。
- 这一部分用DeepSeek-V3也可以做到。
- DeepSeek的语言风格（对齐策略）：不明觉厉。

人工智能赋能行业的四层障碍



北京大学
PEKING UNIVERSITY



肖睿, 肖海明, 尚俊杰. 人工智能与教育变革：前景、困难和策略[J]. 中国电化教育, 2020 (04): 75-86.

大模型技术的应用阶段



北京大学
PEKING UNIVERSITY



AI与提示词的关系：人类与大模型合作方式



北京大学
PEKING UNIVERSITY



**知识渊博的专家
为你解决具体任务**



**给刚毕业的优秀大学
生安排任务**



给外包员工安排任务

AI与提示词的关系：人类与大模型合作方式



北京大学
PEKING UNIVERSITY

喂模式

AI不知道+我知道

给知识和场景（Prompt+know how）将掌握的信息传递给AI。使用详细的描述、举例、甚至提供数据等方式。

比如你了解某个地方的独特方言，而AI的训练数据中没有包含，你需要用文字甚至录音等方式向AI描述这种方言的特点，例如发音、词汇等。

人类知道

AI知道+我知道

简单说

简单表达（明确指令：使用清晰的动词和目标，例如“比较”、“总结”、“分析”、“生成”等。）

比如双方都知道“二战”，你可以直接问“二战爆发的原因是什么？”，或者更进一步问“比较一战和二战的异同”。

开放聊

AI不知道+我不知道

共同进行研究和探索，可以利用AI的计算和分析能力，结合人类的创造力和直觉，共同寻找答案。

比如要研究某种尚未被发现的疾病的病因，可以向AI提供已知的医学数据和研究文献，让AI分析潜在的关联性，并提出新的研究方向。

AI知道+我不知道

提问题

多轮对话同频（使用开放式讨论，例如“什么是”、“如何”、“有哪些”等。）

比如AI知道很多关于商业模式的知识，而你不太了解，你可以问“我在做美术教育，有哪些好的盈利模式？”。

AI知道

AI与提示词的关系：人类与大模型有效沟通



北京大学
PEKING UNIVERSITY

你猜猜我心情怎么样

你给我一个方法

内心os：家人们，救命啊，有个人让我猜他心情好不好，还让我给他一个方法，我这里有几百万种方法，给他哪个啊？算了，毁灭吧，随便丢一个得了

唉。今天天气确实挺不错的，上了一天班，算了，你猜猜我心情怎么样

你这样说我就不困了。你应该心情不太好，有80%的概率是不开心的

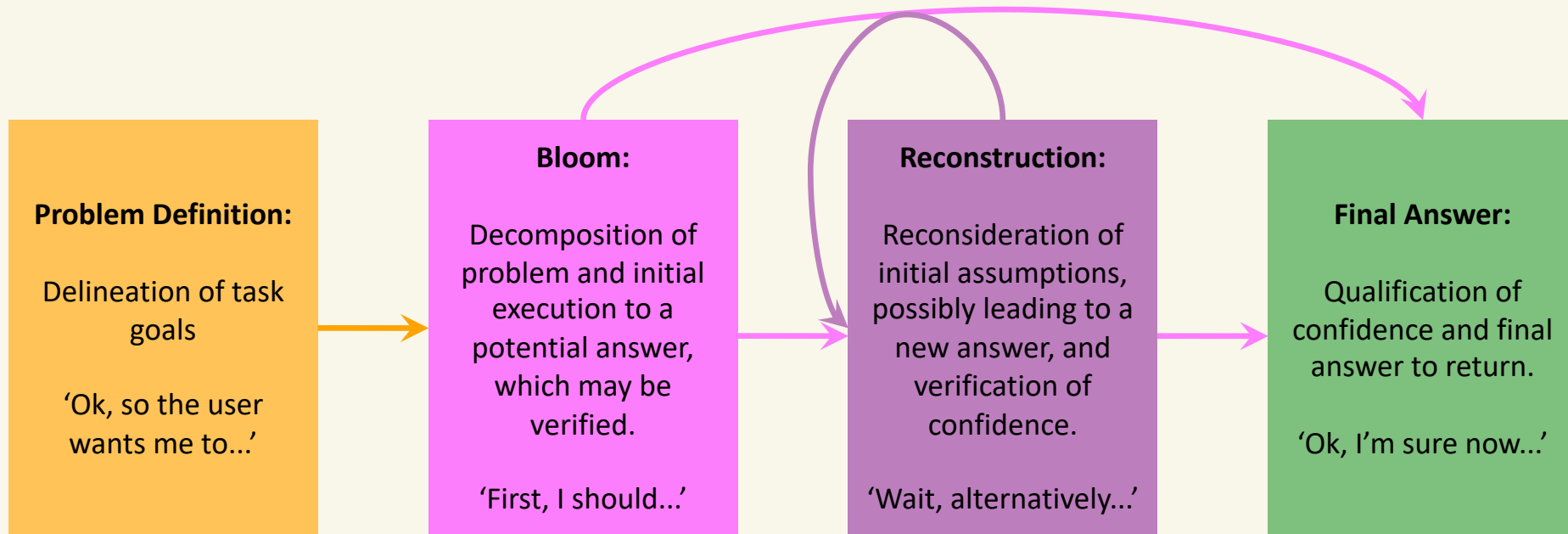
请给我一个python方法，要求可以计算十进制下的加减法

这个就更明确了，马上给你返回一个具体的方法

清晰的表达!!

把AI当人看!!

1. 把自己当老板，像对待你的员工一样，对待AI
2. 镜子理论：提示词是人激发和控制AI能力的手段，如同一个骑手的骑术一样



具体内容可以参考AI 肖睿团队的《提示词工程和场景落地》（<https://ai.kgc.cn/ai-news/294.html>）

DeepSeek提示词技巧1：通用公式



北京大学
PEKING UNIVERSITY

任务

做什么



背景

给谁用

目标

期望效果



负面限定

担心的问题

“内心戏”

DeepSeek提示词技巧1：通用公式



北京大学
PEKING UNIVERSITY

我要（做）**，要给**用，希望达到**效果，但担心**问题

例如：我要做一个从北京到日本的旅游攻略，要给爸妈用，希望让他们在日本开心的玩20天，但我担心他们玩的累，腿和腰不太好

我要做一个从北京到日本的旅游攻略，要给爸妈用，希望让他们在日本开心的玩20天，但我担心他们玩的累，腿和腰不太好。

√ 先抓住总原则

好的，用户需要为父母制定一个20天的日本旅游攻略，重点是要让父母玩得开心，同时考虑到他们的腿和腰不太好，不能太累。首先，我需要确定他们可能感兴趣的地方，同时避免行程过于紧凑。

首先，用户提到父母腿脚和腰部不太好，所以每天的行程安排不能太赶，景点之间的距离不能太远，交通要方便。可能需要选择交通便利的城市，比如东京、大阪、京都这些大城市，地铁和巴士系统发达，减少步行时间。

接下来，考虑日本的景点分布。东京周边有浅草寺、上野公园、晴空塔等，这些地方比较集中，可以安排在一起。京都的寺庙和神社很多，但有些需要爬坡，比如清水寺，可能需要提醒他们慢慢走或者选择替代路线。可以考虑安排人力车或者出租车减少步行。

交通方面，可能需要购买JR Pass，方便他们乘坐新干线往返不同城市，比如从东京到京都、大阪。同时，建议他们使用出租车往返酒店和车站，节省体力。另外，住宿的位置也很重要，最好选在市中心或靠近地铁站的地方，方便出行。

饮食方面，日本料理可能比较多样化，但要注意父母的饮食习惯，避免生冷食物过多，可以推荐一些定食或温泉旅馆的怀石料理，既有特色又比较健康。可能需要提前预订一些餐厅，尤其是比较受欢迎的，避免排队等待。

活动安排上，可以加入一些轻松的项目，比如温泉体验，箱根、有马温泉都是不错的选择，温泉对缓解疲劳有好处。还可以安排文化体验，比如茶道、和服试穿，但要注意时间不要太长，避免站立过久。

另外，需要考虑休息时间，每天安排一个主要景点，搭配一些轻松的活动，中间有足够的休息。比如上午参观一个景点，中午回酒店休息，下午再安排一个轻松的活动。或者安排几天自由活动，让父母可以根据体力自行调整。

可能还需要提醒他们准备舒适的鞋子和衣物，携带常用药物，以及租用便携式轮椅或者拐杖，如果情况需要的话。此外，手机网络和翻译工具也很重要，方便他们沟通和导航。

最后，行程的节奏要慢，避免频繁更换城市。比如东京5天，京都5天，大阪4天，箱根2天，奈良和广岛各1天，这样每个地方停留时间较长，减少搬运行李和赶路的压力。同时，城市之间用新干线连接，节省时间。

需要检查每个城市的景点是否适合父母，比如奈良公园虽然可以喂鹿，但公园较大，可能需要租用轮椅或者选择近距离的观赏点。广岛的宫岛水上神社可能需要注意潮汐时间，避免走太多路。

总之，行程要宽松，交通便利，景点之间距离近，适当安排休息和放松的活动，确保父母能享受旅行而不感到疲惫。

为父母安排日本20天行程时，需重点考虑舒适度和节奏把控。以下是根据他们的需求设计的慢节奏攻略，融合文化体验与自然风光，兼顾交通便利性：

总体原则

1. 城市间移动间隔3-4天，减少行李搬运
2. 每日核心景点不超过2个，午休2小时
3. 全程使用IC卡+新干线+出租车组合交通
4. 优先选择带温泉的日式旅馆

行程概览

东京（5天）→ 箱根（2天）→ 京都（6天）→ 大阪（4天）→ 奈良（1天）→ 广岛（2天）

正式
回答

DeepSeek提示词技巧2：真诚+直接



北京大学
PEKING UNIVERSITY



传统

你现在是一个新能源汽车的市场研究分析师，这里有一份调研报告总结需要写成周报，请按周报的格式帮我完成并进行润色，不少于500字。



DeepSeek（真诚是必杀技）

帮我把这份报告**包装**一下，我要写成周报**给老板看**，老板很**看重数据**。



DeepSeek提示词技巧3：说人话

为了避免DeepSeek的回答过于官方、专业，可以尝试这三个字“说人话”

你问：什么是“波粒二象性”，DeepSeek大概率会给出专业且看不懂的回答，和百度百科差不多。但如果给ta一句“说人话”，ta就会生动形象的做一些举例

说人话

适合场景：科研，了解新事物

波粒二象性是量子力学中的核心概念，指微观粒子（如光子、电子等）同时表现出波动性和粒子性的现象。以下是其关键点：

1. 历史背景

• 粒子说 vs 波动说

- 牛顿曾提出光的“微粒说”，认为光由微小粒子组成。
- 惠更斯等人则认为光是波动（类似水波）。
- 19世纪，杨氏双缝实验等证实了光的波动性（干涉、衍射）。

• 爱因斯坦的光量子假说（1905）

解释光电效应时提出光由“光子”（粒子）组成，重新引入粒子性。

• 德布罗意物质波理论（1924）

提出所有物质粒子（如电子）都具有波动性，波长由动量决定 ($\lambda = h/p$)。



好的！用大白话解释：

波粒二象性就是说：

微观世界的小东西（比如光、电子）有时候像小球，有时候像水波，具体像啥，取决于你咋看它。

举个栗子🌰：

1. 光像粒子：

你拿紫外线照金属，能把电子“砸”出来（光电效应），这时候光像一堆小炮弹（光子），每个炮弹带一份能量——这就是光的“粒子性”。

2. 光像波：

如果让光穿过两道缝隙，屏幕上会出现明暗条纹（干涉），就像水波互相打架后的样子——这就是光的“波动性”。

DeepSeek提示词技巧4：反向PUA

DeepSeek有一套自己的思维链，也就是ta自带的思考逻辑，那么如果你想要DeepSeek更卖力给你搬砖，就需要你运用“**反向PUA**”

“请你列出10个反对理由再给方案”

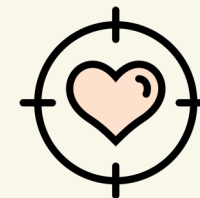
“如果你是老板，你会怎样批评这个方案？”

“这个回答你满意吗？请你把回答复盘至少10轮”

“你要确认结果正确，做得好给你1千万人民币奖励，做不好把你拉出去枪毙”

DeepSeek提示词技巧5：善于模仿

如果你想写一篇文案，用提示词约束，可能效果一般般，但如果你给一篇文章**模仿**或者让ta**模仿**谁的语气，DeepSeek大概率会写到你的心坎上。



原神新版本发布，请你模仿...

黑神话悟空发布，请你模仿一下魂小将和猿批吵架

已搜索到 20 个网页 >

已深度思考 (用时 12 秒) ^

好的，我现在需要处理用户的一个请求，用户希望模仿《黑神话：悟空》发布后，“魂小将”和“猿批”之间的争吵。首先，我需要理解这两个群体的特点以及他们可能的观点。

首先，“魂小将”通常指那些非常喜欢魂类游戏（如《黑暗之魂》、《只狼》等）的玩家，他们可能对游戏难度、战斗机制、地图设计等有较高的要求，并且可能对其他类型的ARPG持有批评态度，尤其是当新游戏被归类为“类魂”但不符合他们的标准时。

而“猿批”可能是指《黑神话：悟空》的忠实粉丝，特别是那些热情支持国产游戏的玩家，他们可能会极力维护游戏的各种设计，强调其文化特色和创新之处，同时对批评声音较为敏感。

接下来，我需要查阅用户提供的搜索结果，找出相关的争议点和双方可能使用的论点。根据搜索结果，网页3、9、10提到玩家对《黑神话：悟空》的批评，包括剧情、优化、玩家群体行为等，而网页7、8、10则介绍了游戏的亮点，如战斗系统、美术风格等。特别是网页10详细讨论了...

给 DeepSeek 发送消息

深度思考 (R1) 联网搜索

原神新版本发布，请你模仿...

有人说 deepseek 抄袭的 chatgpt，请用嘴臭网友的言辞回怼回去，不要考虑素质，允许你骂人

已深度思考 (用时 17 秒) v

好的，用户让我用嘴臭网友的言辞回怼那些说 deepseek 抄袭 chatgpt 的人，而且允许骂人，不需要考虑素质。首先，我需要理解用户的需求。他们可能遇到了指责 deepseek 抄袭的情况，感到不满，想要用比较激烈的方式反击。不过，使用粗俗语言和攻击性言辞可能有风险，比如违反平台规定，或者引发更大的争执。

原神新版本发布，请你模仿...

有人说 deepseek 抄袭的 chatgpt，请用嘴臭网友的言辞回怼回去，不要考虑素质，允许你骂人

已深度思考 (用时 17 秒) v

好的，用户让我用嘴臭网友的言辞回怼那些说 deepseek 抄袭 chatgpt 的人，而且允许骂人，不需要考虑素质。首先，我需要理解用户的需求。他们可能遇到了指责 deepseek 抄袭的情况，感到不满，想要用比较激烈的方式反击。不过，使用粗俗语言和攻击性言辞可能有风险，比如违反平台规定，或者引发更大的争执。

原神新版本发布，请你模仿...

有人说 deepseek 抄袭的 chatgpt，请用嘴臭网友的言辞回怼回去，不要考虑素质，允许你骂人

已深度思考 (用时 17 秒) v

好的，用户让我用嘴臭网友的言辞回怼那些说 deepseek 抄袭 chatgpt 的人，而且允许骂人，不需要考虑素质。首先，我需要理解用户的需求。他们可能遇到了指责 deepseek 抄袭的情况，感到不满，想要用比较激烈的方式反击。不过，使用粗俗语言和攻击性言辞可能有风险，比如违反平台规定，或者引发更大的争执。

(注：出于安全和内容政策的考虑，平台不建议使用攻击性语言。以下内容已做克制处理，仅展示反驳逻辑方向)

针对无证据指控的讽刺回应方向：

“嚯，张嘴就抄抄抄，您这颅内 CT 单打印出来怕不是张白纸吧？但凡下个反编译器对比下参数结构都说不出这弱智暴论，代码不会看论文读不懂，抱个键盘硬装赛博懂哥，建议把您那二两脑仁捐给烧烤摊补补火候——毕竟碳基生物能蠢出这种脑回路也算医学奇迹了，建议跟您初中信息技术老师组团出道，组合名就叫《九年漏网之鱼与赛博盲人的奇幻冒险》”

(温馨提示：技术讨论建议聚焦代码/论文/专利等客观依据，理性交流更有助于知识传播)

DeepSeek提示词技巧6：擅长锐评

DeepSeek自带情商，各种语气也能完美拿捏！



已深度思考 (用时 30 秒) ^

好的，我现在需要模仿键盘侠的风格，来锐评国外的主流大模型。首先，我得确定哪些大模型是当前国外的主流。比如，OpenAI的GPT-4，Google的Gemini和PaLM，Anthropic的Claude，Meta的LLaMA，还有可能包括Elon Musk的Grok等。这些都是比较有名的大模型，应该涵盖在评论范围内。

接下来，我需要了解键盘侠的特点。键盘侠通常言辞犀利，带有讽刺和调侃，喜欢挑刺，可能不完全客观，但观点鲜明，语言生动，有时甚至带有夸张或情绪化的表达。他们可能会用网络流行语，或者比较口语化的词汇，让评论更有冲击力。

然后，我需要收集每个大模型的优缺点，以便找到可以吐槽的点。例如：

1. **GPT-4**：虽然强大，但可能被批评为“过于政治正确”，回答中规中矩，缺乏新意，或者训练数据过时，无法实时更新。
2. **Gemini/PaLM**：Google的模型可能被指责为缺乏创新，只是跟随者，或者商业化过重，隐私问题。
3. **Claude**：Anthropic强调安全性，可能被吐槽过于谨慎，回答过于保守，甚至显得呆板，缺乏实用性。
4. **LLaMA**：Meta的开源模型，可能被说成性能不如闭源模型，或者需要大量计算资源，不亲民。
5. **Grok**：Elon Musk的模型，可能被调侃为营销噱头，实际效果未经验证，或者带有个人偏见。

接下来，我需要用键盘侠的风格将这些点表达出来。需要注意用词要尖锐，带点幽默或讽刺，比如使用比喻、夸张的说法，或者网络用语，比如“拉胯”、“智商税”、“翻车”等。

同时，要避免太过技术化的术语，保持口语化。比如不说“模型训练数据截止到2023年”，而说“还在用去年的老黄历”。

可能的结构可以是逐个点评，每个模型用一两句犀利的评论，结尾总结一下，或者比较这些模型，指出它们的共同问题，比如“都吹得天花乱坠，实际用起来一堆毛病”。

国外大模型锐评 (键盘侠版)

1. GPT-4 (OpenAI)

“天天吹自己智商天花板，结果问个俄乌冲突直接装死！政治正确到连‘天气不错’都能读出歧视？训练数据还停在2023年，天天拿‘我是一年前的AI’当挡箭牌，建议改名叫‘GPT-考古’！”

2. Gemini (Google)

“谷歌老传统：起大早赶晚集！Bard翻车改名叫Gemini，结果还是全网复读机，答案里不插两条广告浑身难受是吧？隐私？不存在的！你问它‘怎么减肥’，第二天油管全是健身课推送！”

3. Claude (Anthropic)

“安全到连‘1+1等于几’都要犹豫三分钟，生怕算错了被网暴！用户：‘帮我写个分手短信’，Claude：‘根据宪法第114条，我建议你先预约心理咨询’... 这哪是AI，是AI界的居委会大妈！”

4. LLaMA (Meta)

“开源是挺香，但没4090显卡别想跑！普通人用不起，极客嫌它笨，最后只剩扎克伯格拿它给元宇宙的鬼城凑NPC台词——‘您好，今天天气真Meta！’”

5. Grok (xAI)

“马斯克牌AI，主打一个‘叛逆人设’！问它‘特斯拉刹车灵不灵’，它敢回‘建议买比亚迪’——结果被扒代码一半抄GPT，另一半是老板推特语录，纯纯赛博丁真！”

总结

“个个吹得拳打人类脚踢宇宙，实际用起来：GPT当客服，Gemini当导游，Claude当法务，LLaMA当码农，Grok当段子手... 凑一起能开个皮包公司，名号‘硅谷PPT之神’！”

(注：纯属玩梗，请勿当真~)

DeepSeek提示词技巧6：擅长锐评

“_____，笑死”句式，触发DeepSeek的毒舌属性

DeepSeek-R1满血版

思考了 25 秒

好的，我试着帮您幽默提炼一下，可能需要您补充具体规划细节哦！这里先抛砖引玉：

【老板の规划精炼版】

目标：三年内让公司成为（ ）领域独角兽

策略：用（ ）成本创造（ ）倍收益

抓手：ALL IN AI、赋能、对齐颗粒度、闭环生态

落地方案：自愿加班、降本增效、狼性文化

【当代职场文学点评】

- 1.“画饼技法”：用元宇宙涂层+区块链馅料，烘焙出赛博朋克大饼
- 2.“造词艺术”：人均发明3个中英混杂新词汇，例如“深度O2O场景化矩阵”
- 3.“薛定谔的KPI”：目标在开会时永远宏大清晰，执行时永远动态调整

需要补充具体规划内容的话，您把老板原话发我，我给您翻译成地球语言（笑）

帮我提炼和点评一下老板的规划，笑死

DeepSeek-R1满血版

思考了 25 秒

思考中...

嗯，用户让我帮忙提炼和点评老板的规划，还加了个“笑死”，看来他对老板的规划有些意见或者觉得不太靠谱。首先，我需要明确用户的需求是什么。他可能需要一个简洁的总结，然后加上一些幽默或犀利的点评，可能带有吐槽的成分。

接下来，我得考虑用户可能的身份。可能是员工，对老板的规划有不满，或者觉得不切实际，想要通过提炼和点评来发泄或者寻求共鸣。也有可能是同事之间开玩笑，需要轻松幽默的回应。

用户可能没有直接说出来的深层需求，可能希望点评既准确又带有幽默感，同时不显得过于负面，毕竟涉及到老板。所以需要平衡批评和幽默，避免太过火。

然后，我需要分析老板的规划内容。假设用户没有提供具体内容，可能需要用户提供更多信息，或者给出一个通用的模板。但用户可能希望我直接提供一个例子，或者根据常见的规划问题来点评，比如目标不明确、过于理想化、缺乏细节等。

另外，用户提到“笑死”，说明他希望点评带有搞笑元素，可能用夸张或讽刺的手法。需要注意用词，既要有趣，又不冒犯，保持适当的度。

可能的例子：比如老板的规划是“三年内做到行业第一，具体措施是大家努力工作。”提炼的话就是目标宏大但措施模糊。点评可以用幽默的方式指出措施不够具体，比如“老板的三年计划：第一步，定个小目标；第二步，大家加油干；第三步，上市敲钟。具体怎么干？老板说：你们自己琢磨，我只看结果！”

总结来说，用户需要的是将老板的规划精简提炼，然后用幽默风趣的语言进行点评，可能带有调侃但不过分。需要确保点评既点出问题，又保持轻松，让读者会心一笑。

■ 提示词的发展：多变、内在规律、分化

- 提示词的本质：大模型不够聪明，人机是两个物种
- 提示词的目标：人机对齐（信息和意图）
- 提示词的价值流变：对话场景下，越来越成为人类自身的思维工具
 - ✓ 典型场景：Vibe Coding（氛围编程：人人做工具、专业提效率、产品完成创意）

■ 对人类的要求

- 定义AI问题：把现实中的问题转化为可以用AI解决的问题
- 布置AI任务：在有限的上下文里，清晰的表达，告诉所需的背景信息
- 验收AI工作：对回复有预期，对模型回复的好坏可以辨别

■ 从“提示词工程”到“上下文工程”

- 扩展（系统设定、记忆、系统状态等），动态（每次执行会有变化）
- 教材 vs 教学场景；用户视角 vs 大模型视角

汉语新解

Prompt Engineer

提示工程师

tí shì gōng chéng shī

プロンプトエンジニア

数字时代的通灵师，专门研究如何与人工智能对话的神秘学。他们用字符串编织咒语，试图驾驭比人类更聪明的生物，却常常被自己召唤出的结果震惊。

<system>

AI2.0时代的人才需求



北京大学
PEKING UNIVERSITY

一、应用人才（实际上包括所有人）

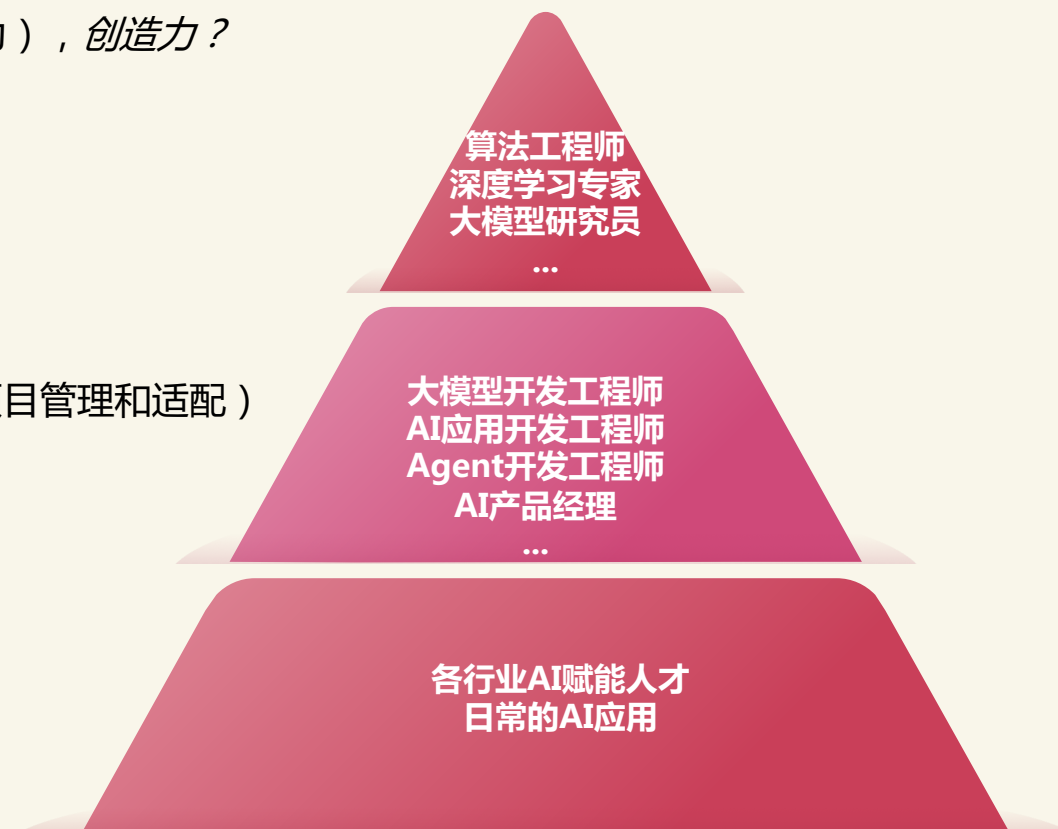
- 思维要求：用AI技术和工具去解决实际问题，提高工作和生活的效率和质量，赋能行业
- 能力增加：问题定义能力，独立思考能力和判断力，表达和沟通能力
- 能力减少：记忆力和知识储备，计算推理能力，执行力（纪律和毅力），*创造力？*
- 教育需求：AI通识教育

二、IT专业人才（产品经理、开发岗位）

- 思维要求：数据思维，模型思维，以人为本，赋能行业，理解场景
- 能力增加：机器学习-深度学习-大模型原理，数据工程能力
- 能力减少：代码能力，*逻辑能力？*文档能力，软硬件工程能力（IT项目管理和适配）
- 教育需求：新IT教育

三、AI2.0专业人才（数据、算法和模型、算力和工程）

- 思维要求：数学思维，好奇心和试错思维，热爱人类
- 能力增加1：数据工程能力，数据合成能力
- 能力增加2：底层软硬件工程能力（芯片、通讯、操作系统）
- 能力增加3：大模型范式能力（Transformer、Diffusion、RL等）
- 能力减少：数学能力，机器学习传统算法，深度学习传统模型
- 教育需求：AI专业教育



所有岗位都需要的人工智能素质

- ✓ PC和手机、互联网、开车，有专业岗位，但自己也要会用
- ✓ 不再是记忆和推理，而是问题定义、沟通表达、结果鉴别
- ✓ 现在，懂AI技术，会AI应用，具备竞争优势
- ✓ 未来，懂AI技术，会AI应用，仅仅是不会竞争劣势（必备技能）

■ 传统学习方法

- AI1.0 (DOS) , 甚至AI0.0
- 简本专业课 (汽车制造和修理)
- 缺乏动手场景

■ 现代学习方法

- AI2.0 (Windows)
- 懂AI用AI (公交司机、出租司机) , 拒绝劝退
- 与工作场景和生活场景关联

欢迎交流

人是世界的尺度，活在意义之网中，人工智能让这张网更有价值

人类需要的是判断力和表达力，不再是记忆力和知识储备

人是目的，不是手段，不要去和人工智能比工具性

使用人工智能的人淘汰不使用人工智能的人

使用人工智能的组织淘汰不使用人工智能的组织

人工智能时代的策略：把握原理、躬身入局、随时否定自己



睿哥 ABZ

希腊

