

# DeepSeek：回望AI三大主义与加强通识教育

浙江大学计算机科学与技术学院

吴 飞



# 提 纲

- 1 从达特茅斯启航的人工智能三大主义
- 2 从 ChatGPT 到 DeepSeek
- 3 人工智能通识教育



# 人工智能单词首次登上人类历史舞台

1955年8月 基本猜想：学习的所有特点以及大多数智能原则上都可被精确描述出来，从而用一台机器来模拟

## What I cannot create, I cannot understand

不可造也，未能知也

- 麦卡锡(John McCarthy)、明斯基 (Marvin Lee Minsky)、香农 (Claude Shannon) 和罗切斯特 (Nathaniel Rochester) 四位学者向美国洛克菲勒基金会递交了一份题为“关于举办达特茅斯人工智能夏季研讨会的提议 (a proposal for the Dartmouth summer research project on artificial intelligence)”的建议书，希望洛克菲勒基金会资助拟于1956年夏天在达特茅斯学院举办的人工智能研讨会。
- 评审意见：研究内容难以让人彻悟(difficult to grasp very clearly)，但是鉴于这一研究所具有长期挑战性特点，基金会愿意资助其申请经费的一半。希望你们不会觉得我们过于谨慎 (overcautious)，对思维的数学模型研究从长远来看非常具有挑战性，是一场适度的赌博，因此在现阶段冒任何大风险会令人犹豫重重。

## A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

August 31, 1955

John McCarthy, Marvin L. Minsky,  
Nathaniel Rochester,  
and Claude E. Shannon

■ The 1956 Dartmouth summer research project on artificial intelligence was financed by this August 31, 1955 proposal, authored by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. The original typescript consisted of 17 pages plus a title page. Copies of the typescript are housed in the archives at Dartmouth College and Stanford University. The first 5 papers state the proposal, and the remaining pages give qualifications and interests of the four who proposed the study. In the interest of brevity, this article reproduces only the proposal itself, along with the short autobiographical statements of the proposers.

guage, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

### 1. Automatic Computers

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack

# 人工智能诞生之初所提研究问题



达特茅斯会议合影  
(1956年6月18日至8月17日)

七大议题

40位专家

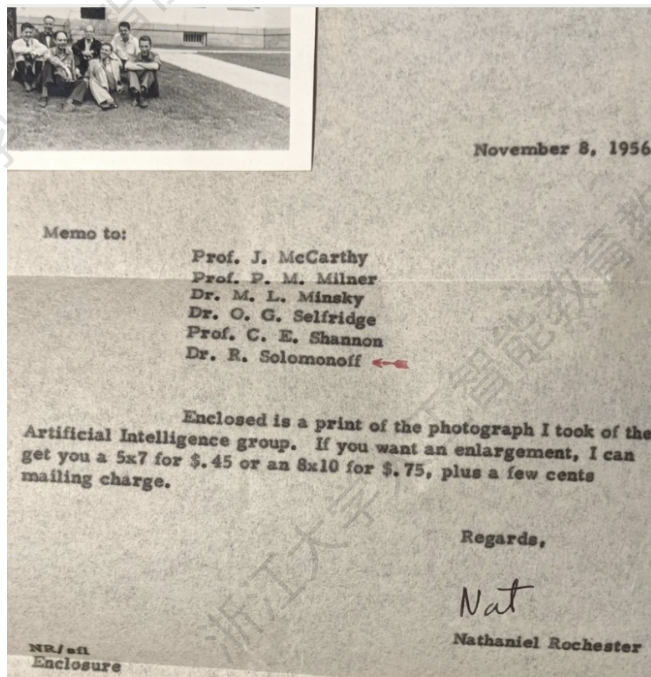
8周时间

议题	描述
自动计算机	让计算机完成特定人类工作
使用语言对计算机进行编程	通过语言对计算机编程
神经网络	通过神经网络让计算机形成抽象概念
计算复杂度	计算任务所耗费的时间和空间
智能算法的自我改进	算法能够苟日新、日日新
智能算法的抽象能力	大数据读薄、然后厚积薄发（归纳和演绎）
智能算法的随机性和创造力	智能算法具备学会学习能力

# 人工智能登上人类历史舞台的“舞者”

你们的名字已被知晓，你们的功绩永世长存

- 麦卡锡、明斯基和索洛莫洛夫（Ray Solomonoff）三位学者全程参与了会议。
- 参加会议的还包括1975年图灵奖得主纽厄尔（Allen Newell）、1975年图灵奖和1978年诺贝尔经济学奖得主西蒙（Herbert A. Simon）、1977年图灵奖得主巴克斯（John Backus）、机器学习一词的创立者塞缪尔（Arthur Samuel）和控制论之父维纳等等。



姓名	情况描述
W. Ross Ashby	神经科学和控制论研究先行者
John Warner Backus	在 Trenchard More 提供的与会名单中被提及，1977 年图灵奖获得者
Alex Bernstein	数学家、IBM 供类人工智能程序研究者
Julian Bigelow	普林斯顿高等研究院工作，美国计算机工程先行者
Woodrow Wilson Bledsoe	在 Trenchard More 提供的与会名单中被提及，美国数学家和模式识别研究先行者
Wesley A. Clark	MIT 物理学家，第一台现代个人电脑发明者
R. Culver	Trenchard More 在回忆中提及并于 7 月 24 日参与了讨论
Thomas Etter	英文散文自动生成程序 Racter 的研制者
B. G. Farley	MIT 研究者
Frederic Brenton Fitch	在 Herbert Simon 一个回忆排播中提及参加，美国逻辑学家
Stanley Frankel	在 Trenchard More 提供的与会名单中被提及，美国计算机科学家，曼哈顿计划顾问
Herbert Gelernter	在 Ronald Kline 的文献“The Cybernetics Moment”中提及，IBM 研究者
David W. Hagebarger	在 Trenchard More 提供的与会名单中被提及，贝尔实验室研究员
John L. Holland	在 Trenchard More 提供的与会名单中被提及，美国心理学家
Donald MacCrimmon MacKay	在被邀请与会名单中，英国物理学家，但因夫人怀孕而未参加
John McCarthy	1971 年图灵奖得主
Warren Sturgis McCulloch	神经科学家和控制论研究先驱者，与 Walter Pitts 首次提出了 Nervous Net 概念来描述神经网络
Peter Milner	在 Trenchard More 提供的与会名单中被提及，加拿大神经科学家
Marvin Minsky	1969 年图灵奖得主，是第一位获此殊荣的人工智能学者
Gloria Minsky	1969 年图灵奖获得者 Marvin Lee Minsky 的妻子
Edward F. Moore	在 Trenchard More 提供的与会名单中被提及，美国计算机科学家、摩尔有限状态机提出者
Trenchard More	曾经在 MIT 和 Yale 等大学任职教授
John Nash	1994 年诺贝尔经济学奖获得者，提出了纳什均衡等博弈论概念
Allen Newell	1975 年图灵奖得主
Anatol Rapoport	在 Trenchard More 提供的与会名单中被提及，美国教学心理学家
Abraham Robinson	德国数学家、美国加州大学洛杉矶分校教授
Nat Rochester	IBM 第一台商用计算机 701 总设计师
Arthur Samuel	机器学习研究先行者，第一个棋类人工智能程序开发者
David Sayre	FORTAN 语言编译器开发者、晶体学研究 Sayre 方程提出者
Oliver Selfridge	机器感知之父
Claude Shannon	信息论创始人、信息熵提出者
Norman Zalmon Shapiro	在 Trenchard More 提供的与会名单中被提及，美国数学家
Kenneth R. Shoulders	实验物理学家、集成电路研究先行者
Bill Shutz	Ray Solomonoff 在回忆中提及并于 7 月 10 日参与了归纳推理问题的讨论
Herbert Simon	1975 年图灵奖得主和 1978 年诺贝尔经济学奖得主
Ray Solomonoff	算法概率论创始人，大多数时间工作于自办公司 Oxbridge Research
Albert Uttley	在 Trenchard More 提供的与会名单中被提及，英国计算机科学和神经生理学家
Pitts, Walter	在 Trenchard More 提供的与会名单中被提及，与 McCulloch 首次提出了 Nervous Net 概念来描述神经网络
Bernard Widrow	美国斯坦福大学教授、最小均方差法提出者（LMS）
Norbert Wiener	在 Trenchard More 提供的与会名单中被提及，控制论创始人

# 人工智能三剑客之一：符号主义人工智能的逻辑推理

“逻辑”指进行正确推理和充分论证的研究（the study of correct reasoning and good arguments），其关心的是从一个或若干前提出发，是否存在一个有效的论证或推理来支持所得到的结论，也就是说在前提和结论之间架构逻辑结构的桥梁。

## 苏格拉底三段论 (syllogism)

大前提：所有人都是要死的

小前提：苏格拉底是人

结论：苏格拉底是要死的

---

人晓人语

$$\forall x (Person(x) \rightarrow Mortal(x))$$
$$Person(Socrates)$$
$$Person(Socrates) \rightarrow Mortal(Socrates)$$

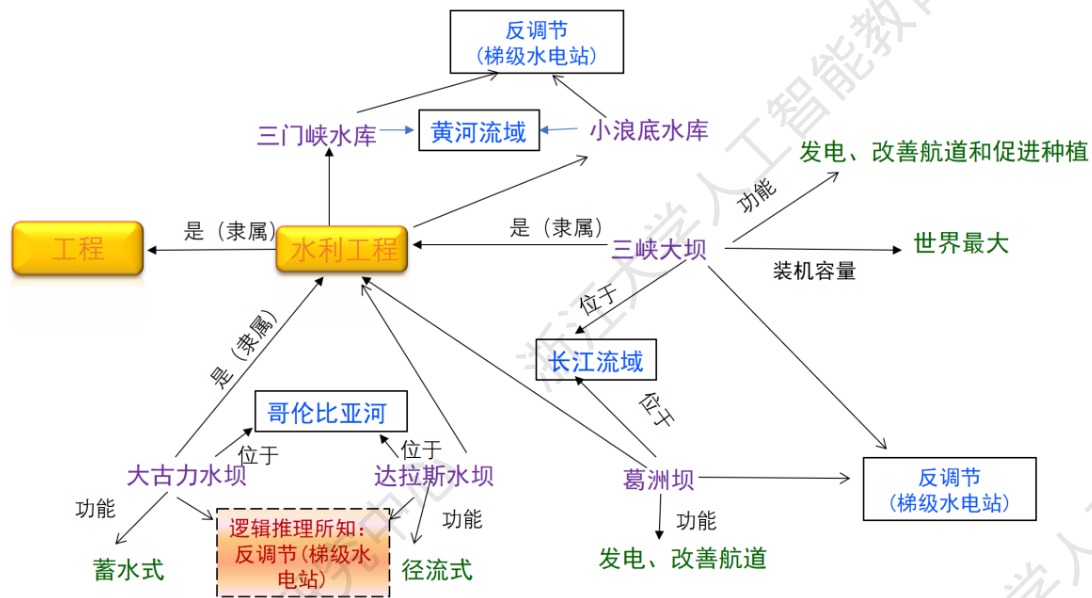
---

机懂机言

路德维希·维特根斯坦《逻辑哲学论》：**语言的边界就是思想的边界**

The limits of my language are the limits of my mind. All I know is what I have words for. (Ludwig Wittgenstein)

# 符号主义人工智能的逻辑推理：推理即计算



知识图谱：条理化结构化表示知识

推理就是计算 (reason is nothing but reckoning)：让一切描述同数学一样切实有形 (tangible)，这样我们就能一眼就找出推理的错误所在。在人们有争议之时，我们可以简单地说，让我们来计算 (calculamus) 一下，而无须进一步的忙乱 (ado)，就可知孰对孰错。

## 已知事实：

三峡大坝和葛洲坝同时位于长江流域、两者具有反调节关系；小浪底水库和三门峡水库同时位于黄河流域、两者具有反调节关系

## 归纳总结得到新知识：

任何两个水库如果位于同一个水域，则两者具有反调节关系

## 演绎推理得到新知识：

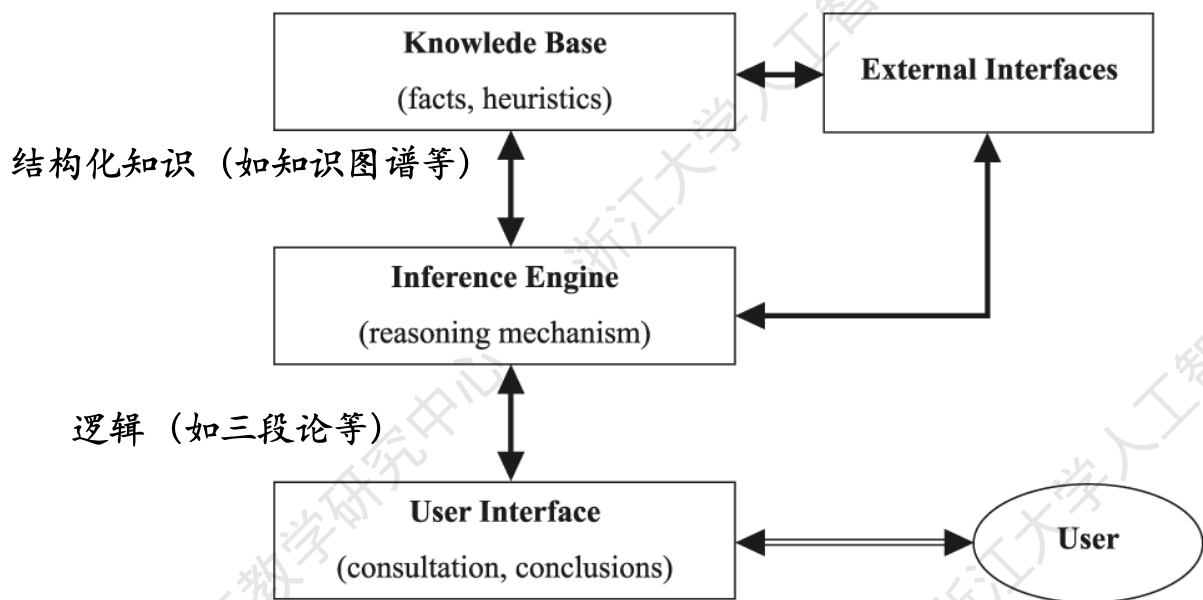
已知任何两个水库如果位于同一个水域，则两者具有反调节关系；已知大古力水坝和达拉斯水坝都位于哥伦比亚河流域

推理得到：大古力水坝和达拉斯水坝具有反调节关系



# 符号主义人工智能的逻辑推理：牛刀小试的知识工程

围绕某一特定领域（如牙病治疗、工具组装等）的应用，将人类专家知识转化为结构化知识，存储进入数据库，从而支持该领域应用，构建“知识水晶球”，这就是知识工程（knowledge engineering）和专家系统（expert system）的动机。

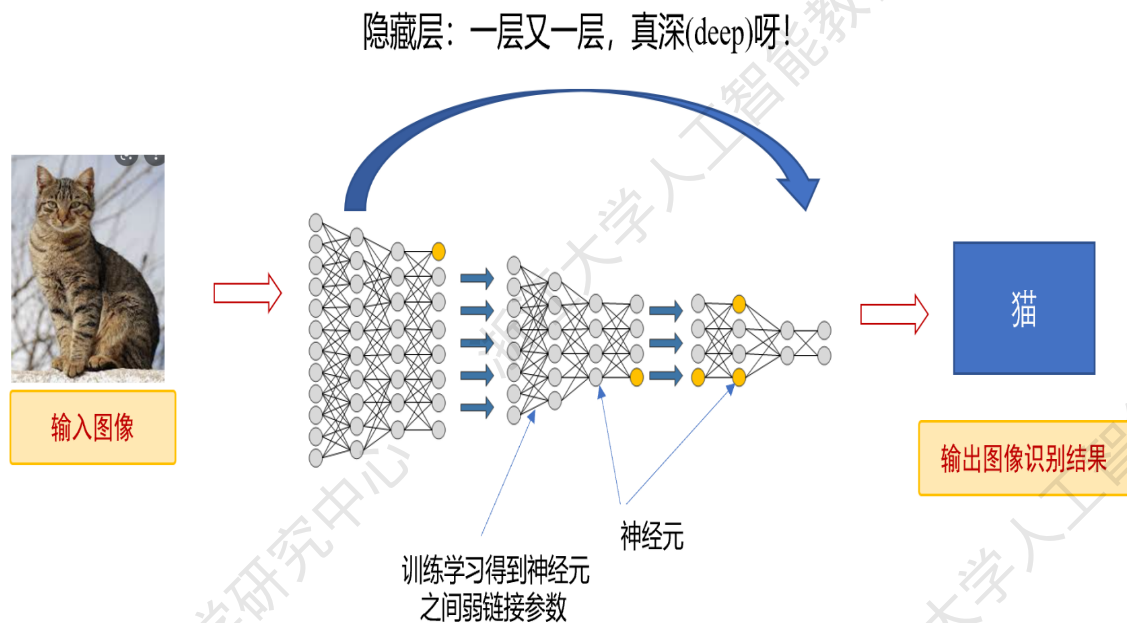


1965年，图灵奖获得者、斯坦福大学计算机科学家费根鲍姆（Edward Feigenbaum）和化学家勒德贝格（J. Lederberg）合作（1958年诺贝尔生理学或医学奖），结合化学领域的专门知识，研制了世界上第一个专家系统Dendral，进行分子结构分析。Dendral这一单词来源于古希腊语“树”，这也是继承和发展了雷蒙·卢尔（Ramon Llull）对知识进行规范化描述的“知识树或科学树”的努力。

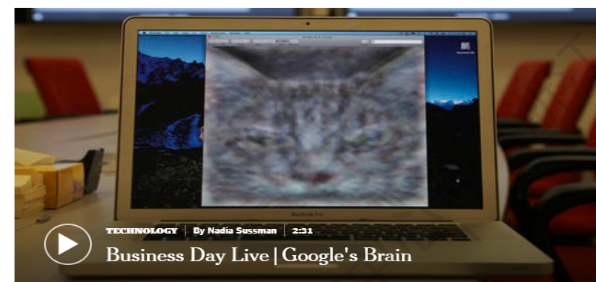


# 人工智能三剑客之二：连接主义人工智能的数据驱动

## 层层递进、逐层抽象



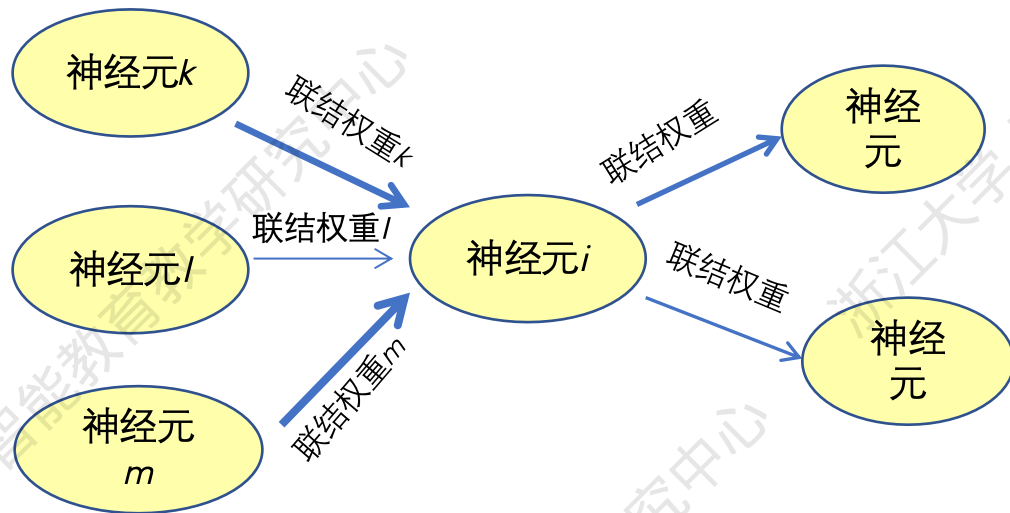
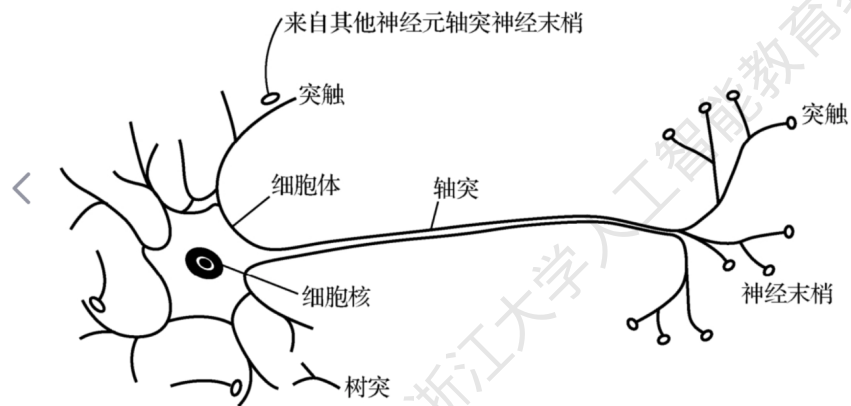
How Many Computers to Identify a Cat? 16,000



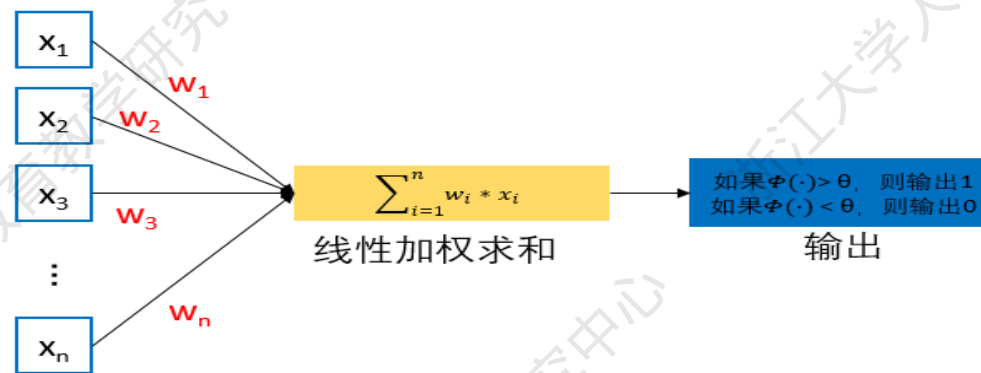
多少台机器可识别一只猫(2012.6)

深度学习的基本动机在于通过“端到端学习（end-to-end learning）”这一机制来构建多层神经网络，以学习隐含在数据内部的关系，从而使学习所得特征具有更强的表达能力

# 连接主义人工智能的数据驱动：解码神经元之奥秘

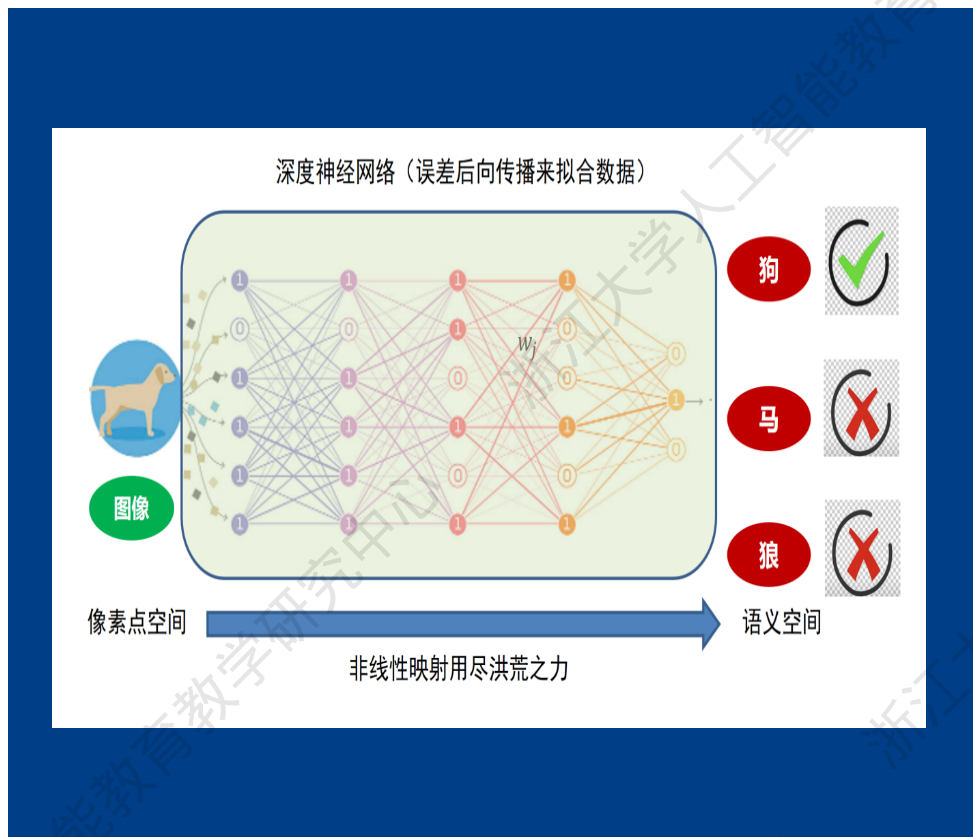


## 神经元工作机制：加权累加、阈值输出



1943年，神经科学家沃伦（Warren McCulloch）和逻辑学家沃尔特·皮兹（Walter Pitts）合作提出了以他们名字命名的“MCP神经元”模型：在科学史上第一次，我们知道了我们是怎么知道的。

# 连接主义人工智能的数据驱动：超越费曼极限的多者异也

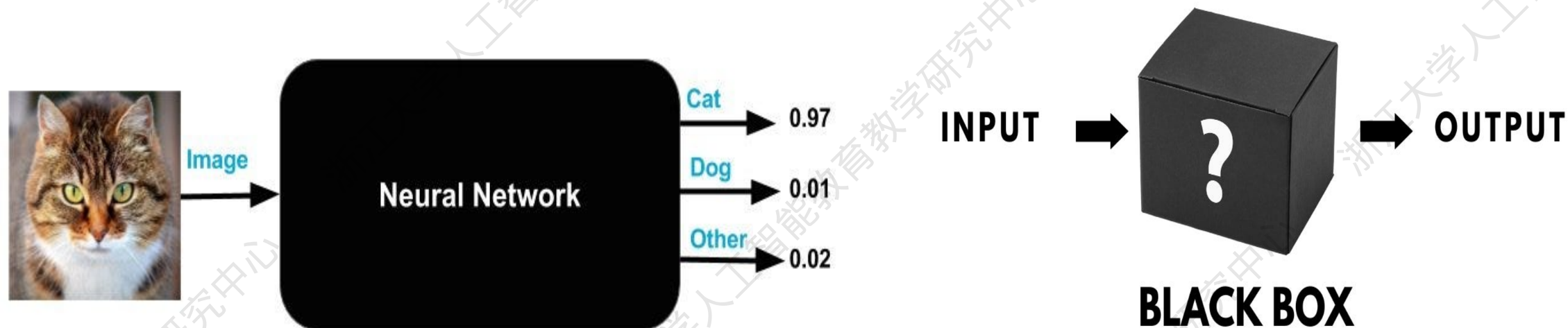


打通任督二脉

- **巴普洛夫条件反射定律：刺激说**  
大脑的一切心理活动都是对刺激的反应，包含有意识和无意识
- 理查德·费曼（Richard Feynman）在《物理学讲义》中曾提及到，在生物学、人类学或经济学等复杂系统中，很少有一种简洁的数学理论能与数学物理学理论中的数值精确度相媲美，其原因在于“其过于复杂，而我们的思维有限”，这被称为费曼极限。
- 美国物理学家安德森(Philip W. Anderson)（1977年诺贝尔物理学奖获得者）1972年在《科学》杂志上发表了一篇题为“More is different”（多者异也）的文章，深刻指出：还原论假说从来都不意味着建构论（constructionist）假说，将所有事物还原为简单的基本定律的能力并不意味着从那些基本定律出发并重建整个宇宙的能力。



# 连接主义人工智能：概率为王下黑箱效应之困惑



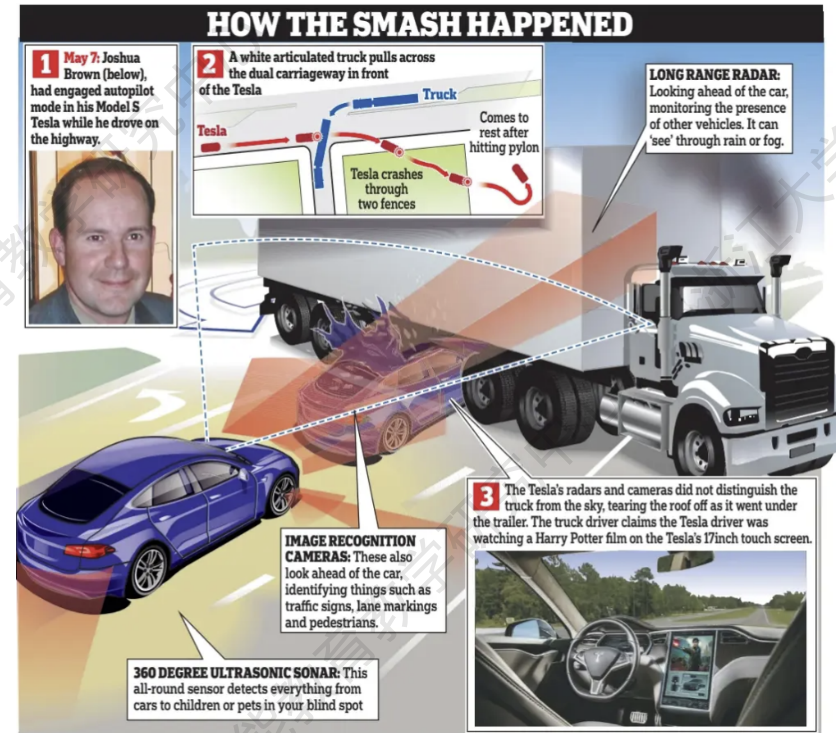
众多参数构成的复杂深度模型虽然在许多任务上取得了亮丽效果，但对模型知其然且知其所以然的理解却举步维艰，面对神经网络犹如炼金术一样的“黑箱效应”，不得不感叹复杂深度模型“**无他、但手熟尔**”，与知其然且知其所以然相去甚远。

# 连接主义人工智能：数据驱动下的滑铁卢

神经网络只是学会了区分阴天和晴天，而不是区分伪装而成的坦克和森林（数据拟合而非学习）



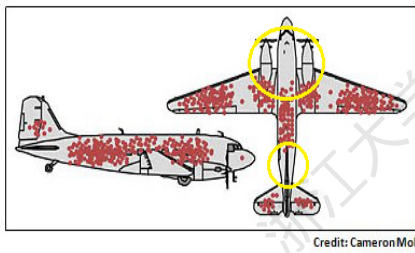
## 特斯拉的悲剧损失 (A Tragic Loss)



2016年7月，40岁的Joshua Brown驾驶特斯拉启动了Autopilot自动驾驶功能，在车内观看《哈利波特》时，没注意到前方迎面而来卡车，因而车毁人亡。

# 连接主义人工智能：拜托呀，数据

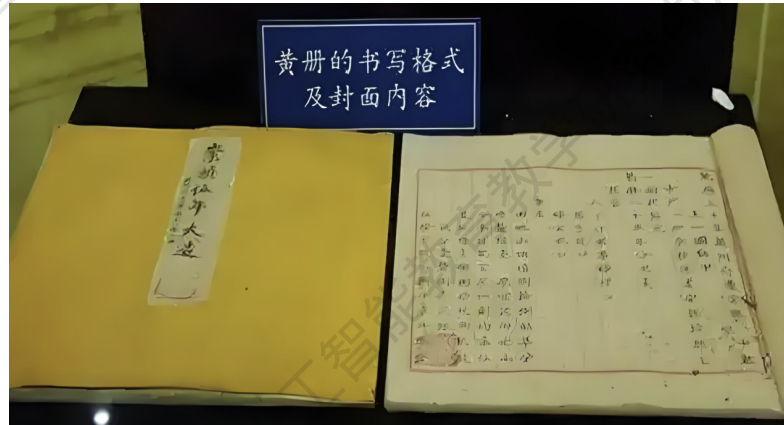
美国哥伦比亚大学统计学  
沃德教授(Abraham Wald)



Gentlemen, you need to put more armour-plate  
where the holes aren't because that's where the holes  
were on the airplanes that didn't return - Abraham  
Wald 1942.

幸存者偏差  
(survivorship bias)

后湖黄册库 (玄武湖)



明代黄册  
核实户口、征调赋役  
国家运行的根本

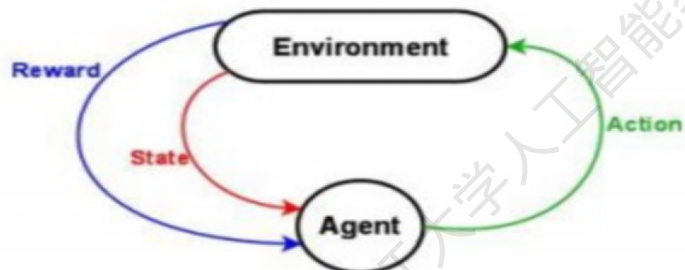


大数据杀熟/信息茧房

注：公元1645年，清军攻占南京后，对存放于玄武湖的黄册库很感兴趣，打开黄册库房看后，没想到黄册上所记录人口、田产等信息已经编排到了崇禎二十四年，明朝末代皇帝朱由检早于崇禎十七年留下“皆诸臣之误朕也”怨恨，在煤山自缢而死，黄册记录信息已“人为超前”了七年之久。

# 人工智能三剑客之三：行为主义人工智能的百折不挠

## 谋定而后动，知止而有得



从经验中的  
策略学习

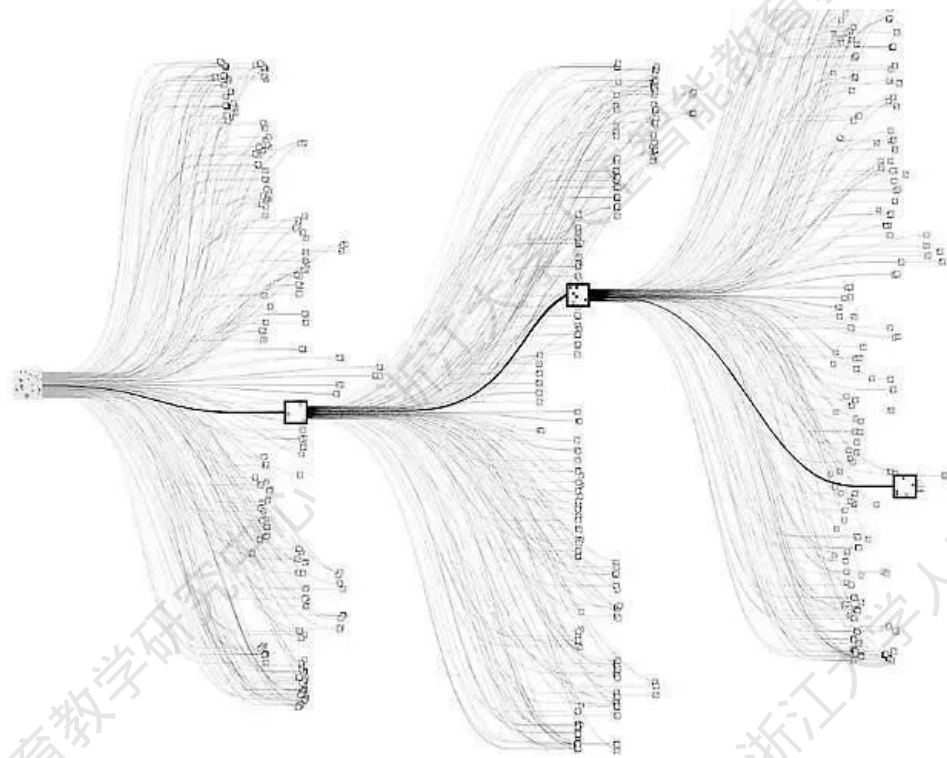
用问题引导  
(反馈牵引)

- 强化学习：人工智能算法在不断与其所处环境交互中进行学习，通过“尝试与试错”不断与环境交互，**形成序贯决策**，直至进入终止状态。
- 强化学习既不是从已有数据出发、也不是依赖于已有知识的学习方式，犹如“tabula rasa (拉丁语)”所蕴含“一张白纸绘蓝图”之义，从“授之以鱼”迈向“授之以渔”。

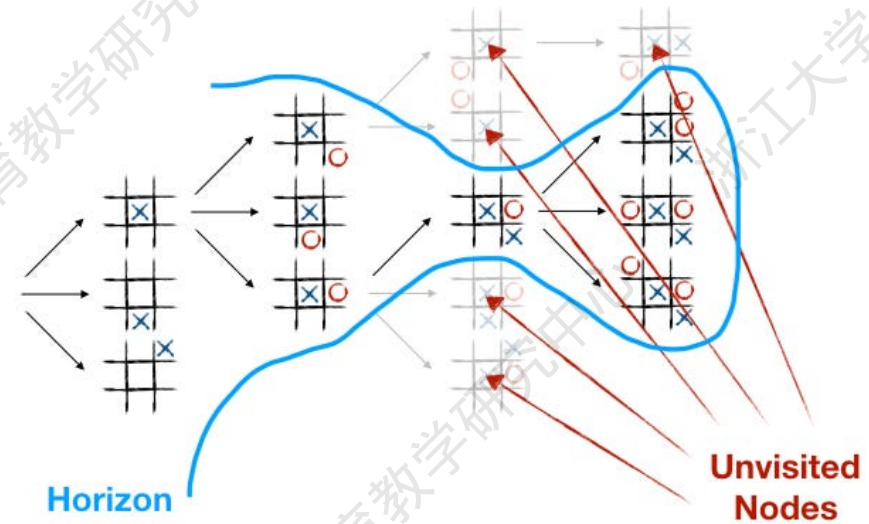
虽九死其犹未悔



# 地平线问题之困：强化学习中消失的雷达搜索信号



The Horizon Effect



李世石与AlphaGo第四局中第78步落子：上帝之落子（God 's Touch），这一步落子人类会不会选择，其发生的可能性只有万分之一。



# 人工智能：至小有内、至大无外

以“厚基础、强交叉、养品行、促应用”为理念

10大模块·63个知识点（含9个进阶知识点）



吴飞 潘云鹤，《人工智能引论》，高等教育出版社，教育部计算机领域本科教育教学改革试点工作计划（即101计划）核心课程教材

# 新一代人工智能迅速崛起

浙江大学潘云鹤院士提出驱动新一代人工智能发展的内因和外因论：回顾人工智能发展历程中的主要挫折，我们不难发现，当它与信息环境的变化趋势不符时，往往就会导致失败。促使人工智能变化的动力既有来自人工智能研究的内部驱动力，也有来自信息环境与社会目标的外部驱动力，两者都很重要，但相比之下，往往后者的动力更加强大。

## 信息 新环境 巨变

互联网、物联网和超级计算等

互联共融

## 社会 新需求 爆发

AI+X应用需求（医疗、教育和交通、基础科学等）

层出不穷

## AI 技术的新目标巨变

从“造人”到“赋能”

凡贵通者，贵其能用

Yunhe Pan, Heading toward Artificial Intelligence 2.0, 2(4): 409-413, 2015, Engineering

社会一旦有技术上的需要，这种需要就会比十所大学更能把科学推向前进！恩格斯



Research  
Artificial Intelligence—Perspective

人工智能走向 2.0

潘云鹤

Chinese Academy of Engineering, Beijing 100088, China

### ARTICLE INFO

Article history:  
Received 2 December 2016  
Revised 9 December 2016  
Accepted 13 December 2016  
Available online 16 December 2016

关键词  
人工智能 2.0  
大数据  
群体智能  
跨媒体  
人机混合智能  
无人智能系统

### 摘要

随着互联网的普及、传感网的渗透、大数据的涌现、信息社区的崛起，以及数据和信息在人类社会、物理空间和信息空间之间的交叉融合与相互作用，当今人工智能(AI)发展所处信息环境和数据基础已经发生了深刻变化，人工智能的目标和理念正面临重要调整，人工智能的科学基础和实现载体也面临新的突破，人工智能正进入一个新的阶段。这个源于传统而又与之不同的人工智能新阶段被称为人工智能2.0(AI 2.0)。本文从人工智能60年的发展历史出发，通过分析促成人工智能2.0形成的外部环境与目标的转变，分析技术萌芽，提出了人工智能2.0的核心理念，并结合中国发展的社会需求与信息环境特色，给出了发展人工智能2.0的建议。

© 2016 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

### 1. 引言

近年来，产业界、媒体、政界等都对人工智能产生了浓厚兴趣，人工智能的研究与应用在国内外迅速升温。

产业界首先对布局人工智能展开行动。根据美国风投数据公司CBInsight 2016年7月的报告[1]，从2011年以来的5年中，美国谷歌、微软、Twitter、Intel、Apple等IT巨头收购了约140家人工智能领域的创业公司。仅2016年上半年，资本市场对人工智能方向的投入就已经超过2015年全年。200家与人工智能相关的公司在股市上筹集到了15亿美元。

大量的并购与资本的涌入加速了人工智能技术更紧密地与应用结合，更快地向经济转化。以谷歌为例，2013年，当谷歌高价收购由多伦多大学的教授Geoffrey Hinton创办的只有3位员工的神经网络公司DNResearch

时，曾引起轩然大波，如今Geoffrey Hinton发起的深度学习，已经成为产业界最热门的技术，帮助谷歌切实提升了图像搜索的准确度，成为谷歌眼镜、谷歌无人车等项目的核心支撑。2014年，谷歌又以4亿英镑收购了DeepMind公司。2016年，DeepMind研制的AlphaGo战胜了围棋世界冠军而崭露头角。谷歌自称正在从一个移动优先(Mobile first)的公司进化到人工智能优先(AI first)的公司。

人工智能与产业需求的结合也催生出服务模式的重大的转变，如微软开发的小冰聊天机器人，正在引导传统使用方式从“图形界面”向“自然语言和情感理解交互界面”转变。微软在2016年6月以262亿美元收购社交网站LinkedIn[2]，准备利用人工智能技术来重建互联网社区。IBM开发的Watson系统，已经在医院运行，包括快速筛选癌症治疗史中的150万份患者记录、诊断疑难

E-mail address: zhs@cae.cn

2095-8099/© 2016 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).  
英文原文: Engineering 2016, 2(4): 409-413  
引用本文: Yunhe Pan. Heading toward Artificial Intelligence 2.0. Engineering, <http://dx.doi.org/10.1016/j.ENG.2016.04.018>

# 国务院《新一代人工智能发展规划》



## 六项重点任务

构建开放协同的  
人工智能科技创新体系

培育高端高效的  
智能经济

建设安全便捷的  
智能社会

加强人工智能领域  
军民融合

构建泛在安全高效的  
智能化基础设施体系

前瞻布局新一代  
人工智能重大科技项目

人工智能的迅速发展将深刻改变人类社会生活、改变世界

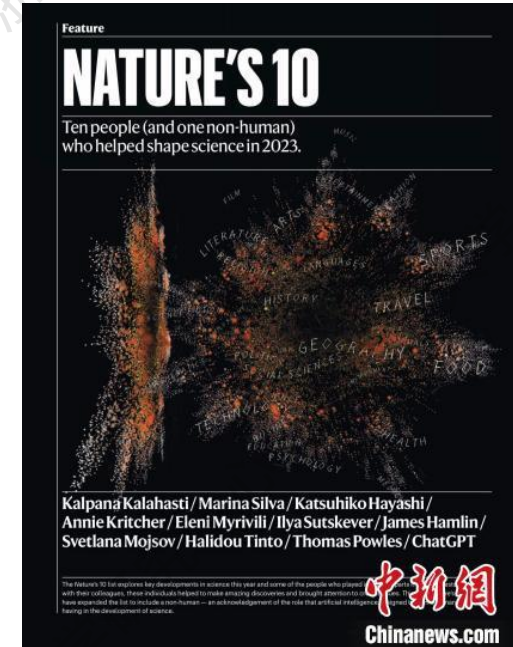
# 提 纲

- 1 从达特茅斯启航的人工智能三大主义
- 2 从 ChatGPT 到 DeepSeek
- 3 人工智能通识教育

# ChatGPT：人工智能的IPHONE时刻？



- 2007年1月9日，乔布斯发布第一代iPhone苹果手机，把iPod、电话、移动互联网设备等进行有机整合，推动了移动互联网进入了黄金发展年代。
- 今天大模型给人类社会诸多生产、生活模式带来一次大变革。2023年2月，英伟达创始人兼CEO黄仁勋提出随着ChatGPT为代表的大模型出现，我们已经进入“人工智能的iPhone时刻（iPhone moment of AI）”，这一观点受到美国《财富》杂志、华尔街时报等媒体的广泛认可并转载。

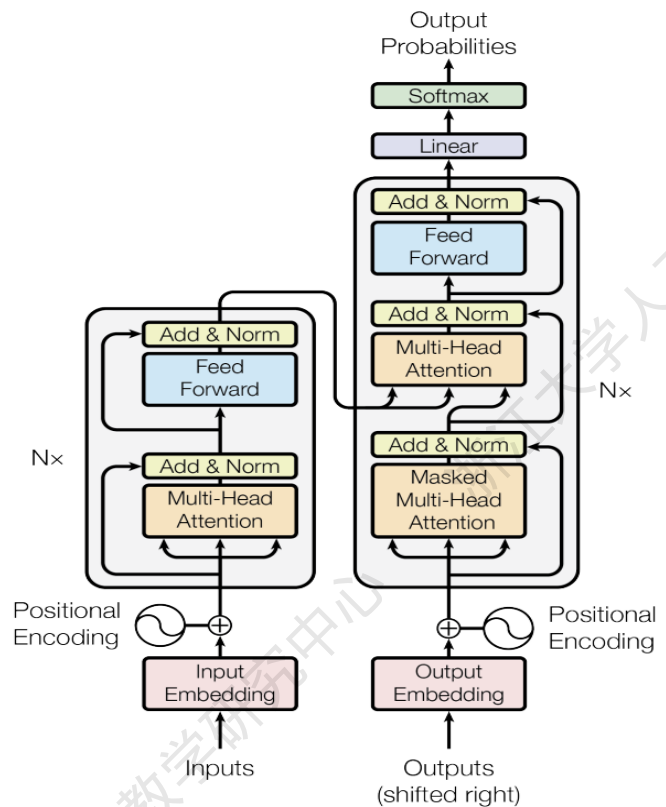


- 《自然》杂志列出2023年度十大人物(Nature's 10)，除了按惯例从全球的重大科学事件中评选出十位人物外，还有一位非人类——人工智能(AI)工具ChatGPT也“抢镜”上榜。



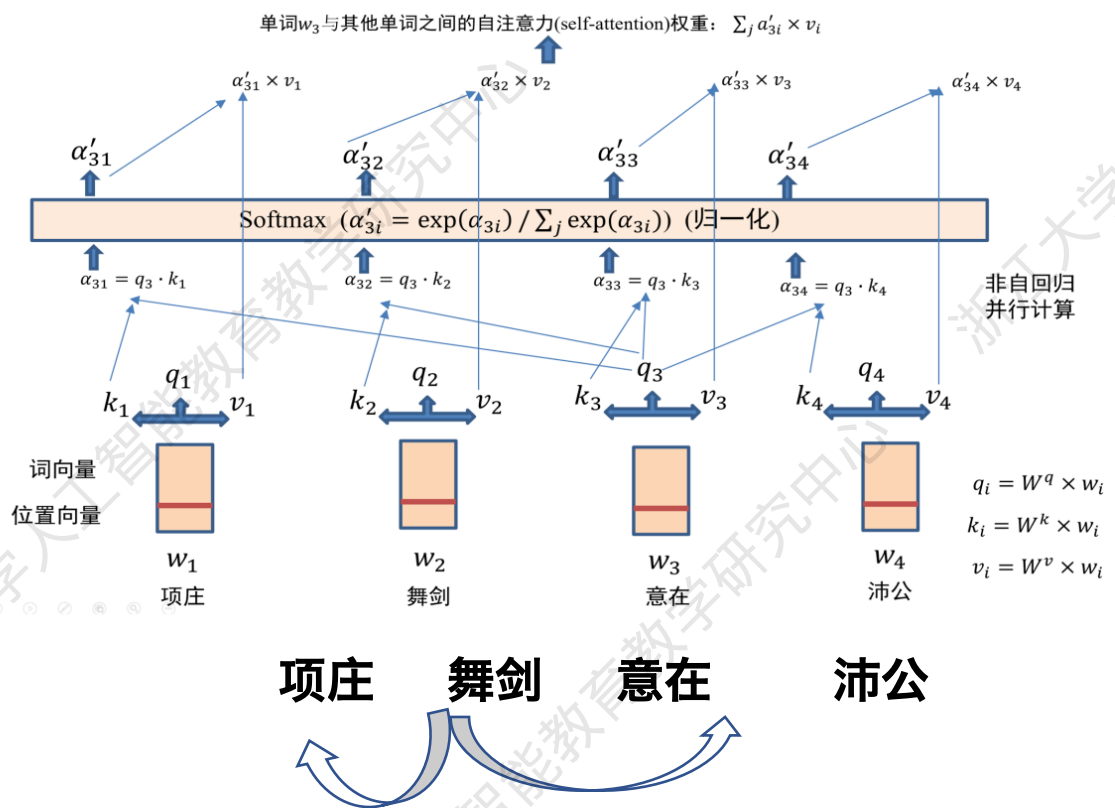
# GPT (generative pretraining transformer)

通过K (key)、Q(query)和V (value)矩阵实现自注意力机制 (self-attention)



The Transformer - model architecture  
 Google (2017): Attention is all you need  
 (八位作者共同一作)

截至2月15日, 该文引用量次数152577



学习句子中单词和单词之间关联关系, 以“预测下一个Token”形式进行内容合成

# GPT的训练三板斧之一：完形填空形式下文字接龙

## 自 监 督 学 习

在训练时，人工智能模型会不断地在句子中“挖去”一个单词，根据剩下单词的上下文来填空，即预测最合适的“填空词”出现的概率，这一过程为“自监督学习”。

原话： 一辆 列车 缓慢 行驶 在 崎岖 的 山路上

移除单词： 一辆 列车 \_\_\_\_\_ 行驶 在 崎岖 的 山路上

预测填空： 一辆 列车 缓慢 行驶 在 崎岖 的 山路上

# GPT的训练三板斧之二:有监督微调 ( SFT )

## 人 教 机 学

让人工智能模型说人话、做人事 ( supervised fine-tuning )

手工设计提示和指令微调被誉为人工智能私语者 ( AI whisperer )

经师易得、人师难求

- “这部电影很精彩，因为其剧情[MASK1]”
- “这家餐馆很不错，因为其饭菜[MASK2]”
- 合成输出“引人入胜”和“可口”这样的内容
- 随后根据“引人入胜”和“可口”和分类标签的映射关系得到最后的预测结果“正向情感”

提示学习 ( 情感分类识别 )



- 第一轮: {“instruction”: 本院查明, 被告人酒后...  
请分析案情。  
“output”: 根据上述内容, 可以认定本案的核心要素包括醉酒驾驶、致人受伤、酒后逃逸...}
- 第二轮: {“instruction”: 根据上述分析, 请预测罪名。  
“output”: 本案预测的罪名是危险驾驶罪}
- 第三轮: {“instruction”: 请给出处罚意见。  
“output”: 结合嫌疑人逃逸的情节, 建议考虑拘役三个月, 并罚款6000元}

指令微调 ( 概念对比分析 )

# GPT的训练三板斧之三:人类反馈下强化学习

## 尝试与探索

### 谋定而后动，知止而有得

Step 1  
Collect demonstration data,  
and train a supervised policy.

A prompt is  
sampled from our  
prompt dataset.

Explain the moon  
landing to a 6 year old

A labeler  
demonstrates the  
desired output  
behavior.

Some people went  
to the moon...

This data is used  
to fine-tune GPT-3  
with supervised  
learning.

SFT

Step 2  
Collect comparison data,  
and train a reward model.

A prompt and  
several model  
outputs are  
sampled.

Explain the moon  
landing to a 6 year old

A B  
C D  
Explain gravity... Explain war...  
Moon is natural satellite of... People went to the moon...

A labeler ranks  
the outputs from  
best to worst.

D > C > A = B

This data is used  
to train our  
reward model.

RM  
D > C > A = B

Step 3  
Optimize a policy against  
the reward model using  
reinforcement learning.

A new prompt  
is sampled from  
the dataset.

Write a story  
about frogs

The policy  
generates  
an output.

PPO

Once upon a time...

The reward model  
calculates a  
reward for  
the output.

RM

The reward is  
used to update  
the policy  
using PPO.

$r_k$

根据中国婚姻法，离婚后子女  
抚养权如何确定？



A根据中国婚姻法，离婚后子女的抚养权主要考虑孩子的最佳利益。一般情况下，对于未满2周岁的婴儿，母亲有优先抚养权

C根据中国婚姻法，离婚后子女的抚养权确定主要遵循以下原则：子女年龄小于10岁的，一般由母亲抚养。因为此阶段孩子需要母亲的照料，但父母双方也可以协商由

D根据中国婚姻法，离婚后子女的抚养权主要是由父母双方协商确定的，如果协商不成，可以向人民法院提起诉讼，由法院根据具体情况作出裁决。

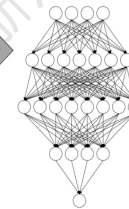
a) 训练数据生成



A > D > C > B

b) 人类偏好标注

A > D > C > B

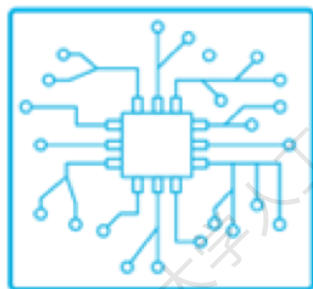


b) 奖励模型训练

与人类价值对齐  
训练反馈奖励模型

# AI有三宝：数据是燃料、模型是引擎、算力是加速器

算力



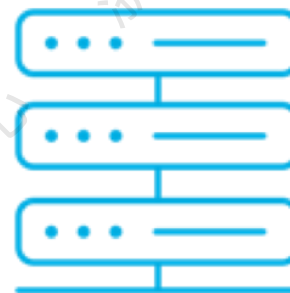
Computing Power

模型



Algorithm Power

数据



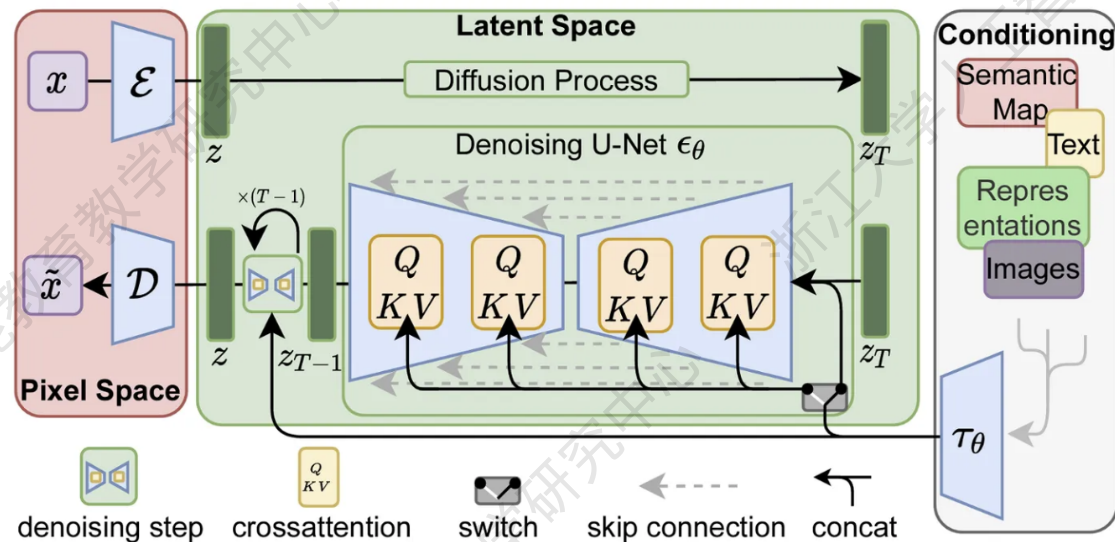
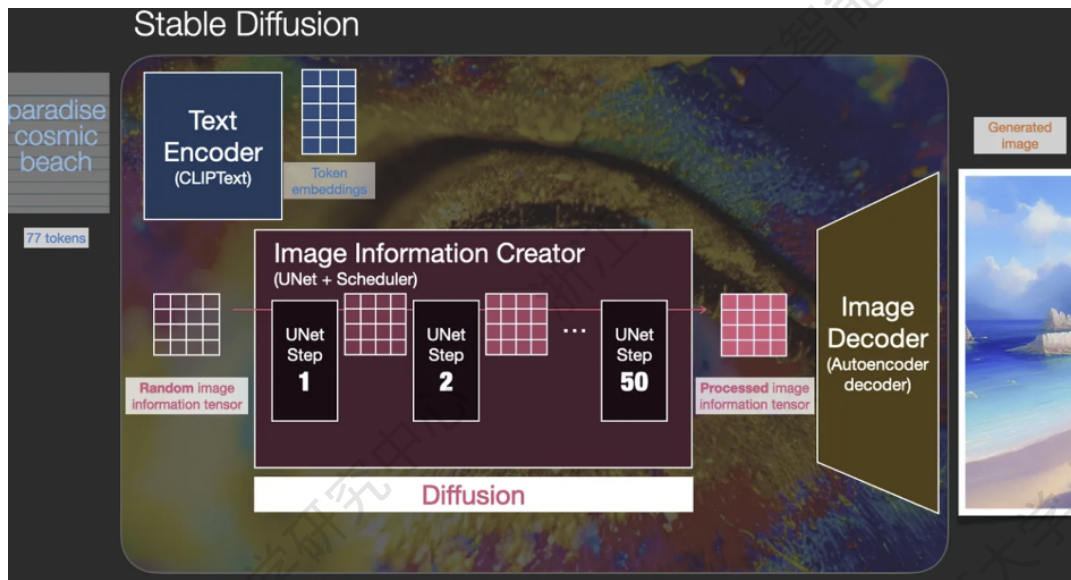
Data Availability

- **数据**：ChatGPT训练中使用了45TB数据、近 1 万亿个单词（约1351万本牛津词典所包含单词数量）以及数十亿行源代码。据估计全球高质量文本数据的总存量在5万亿token左右，人工智能算法可能在一个数量级内，耗尽世界上所有有用的语言训练数据供应。
- **模型**：包含了1750亿参数，将这些参数全部打印在A4纸张上，一张一张叠加后，叠加高度将超过上海中心大厦632米高度。
- **算力**：ChatGPT的训练门槛是1万张英伟达V100芯片、约10亿人民币，模型训练算力开销是 每秒运算一千万亿次，需运行3640天（ 3640 PetaFLOPs per day ）。
- 大数据、大模型、大算力下以“共生则关联”原则实现了统计关联关系的挖掘。



# SORA：从自然语言描述去合成视觉内容

从文本空间到视觉空间  
从图像合成到视频合成

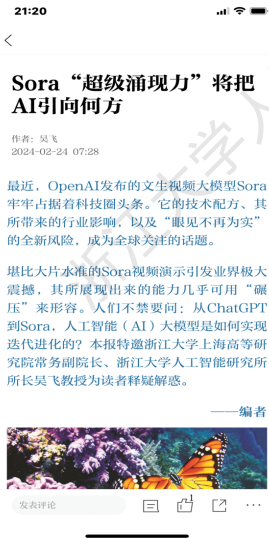


书同文、车同轨---》鲁班学艺---》行同伦

CLIPText 用于文本-视觉编码、U-Net + Scheduler 用于逐步处理/扩散被转化到潜空间中的信息、AutoEncoder Decoder (VAE: Variational AutoEncoder) 把隐性空间的运算结果解码成图像

# Sora “超级涌现力” 将把AI引向何方（文汇报：2024年2月24日）

## 从Chat到Sora: 对合成内容中最小单元进行有意义的关联组合，犹如昨日重现



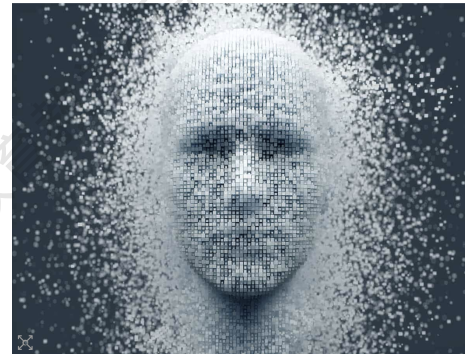
I am four years old.

There are five people in my family.

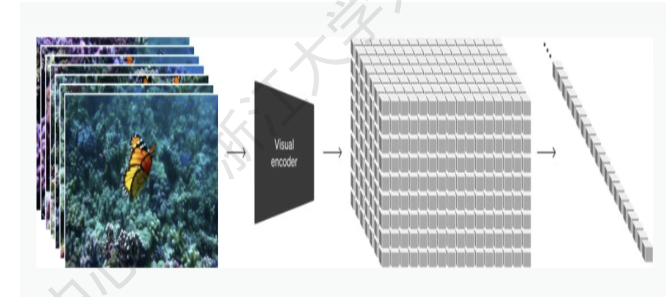
I am young, short and thin.

My dad is forty years old.

单词有意义的线性组合



像素点有意义的空间组合

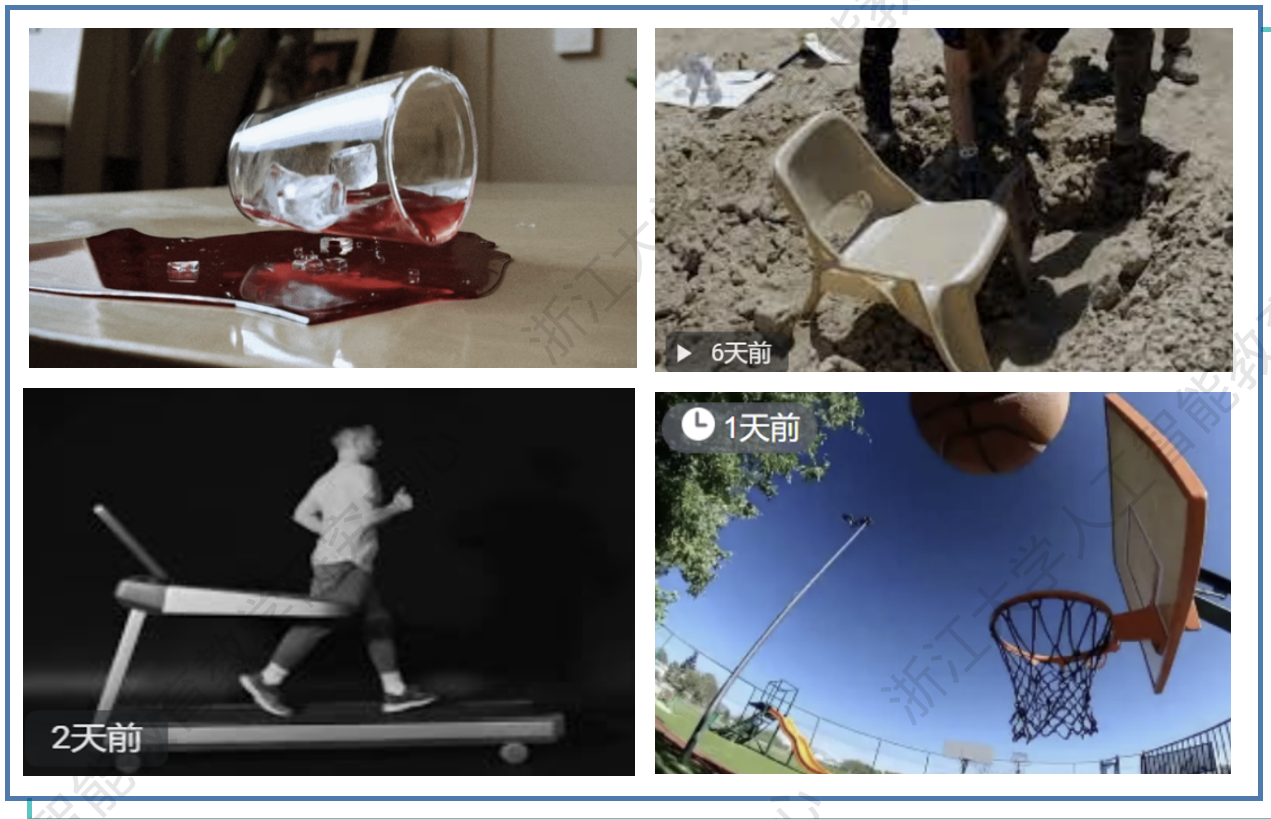


时空子块有意义的时序组合

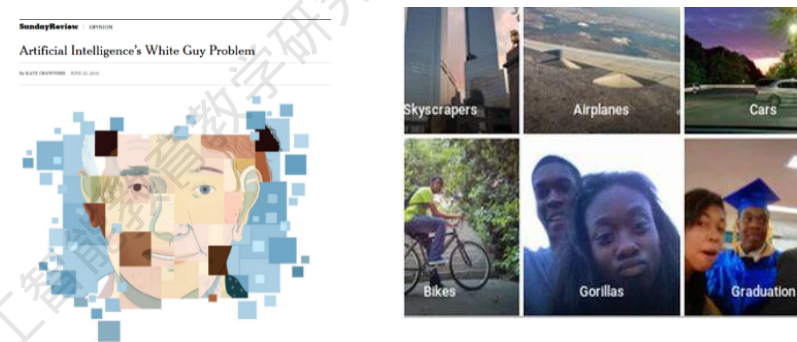
在保持连贯的上下文语境中，对若干个单词进行有意义线性组合，从而连缀成一个会意句子；在保持合理的空间布局下，对众多图像小块进行有意义结构组合，拼合为一幅精彩图像；在保持一致的连续时空内，对一系列时空子块进行有意义时空组合，从而拼接成一段动感视频。



# Sora:合成失误的视频



谷歌暂停有关人物图像合成的服务（2024.2）



谷歌图像标注系统暂停服务人脸标注（2016.6）

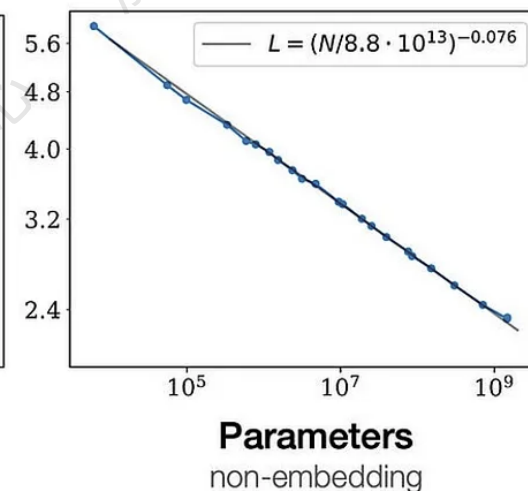
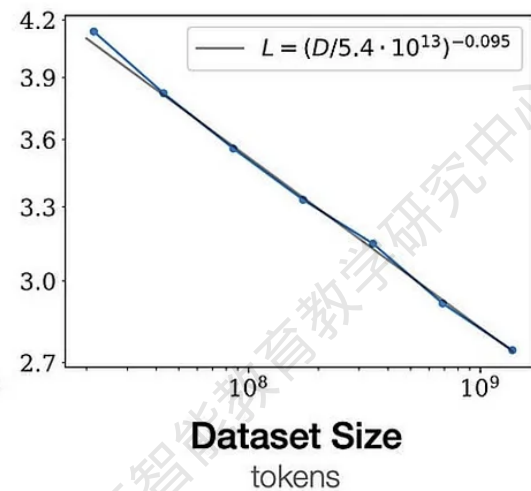
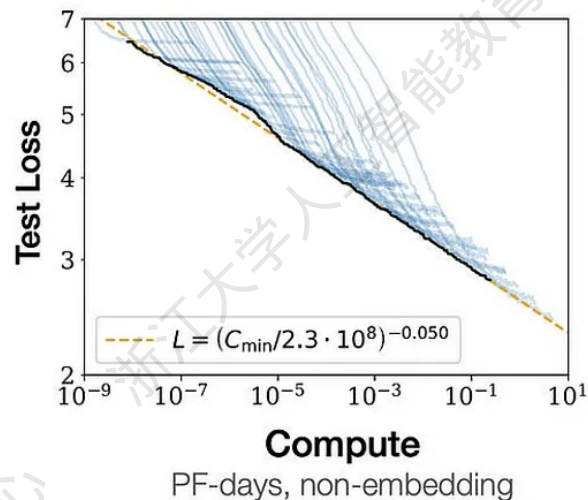
# 大模型的扩展定律

- 每个单词要记住越来越多不同语境下的“左邻右舍”，以组合意义下“昨日重现”方式合成众所周知的语言内容，导致模型参数不断增多而导致模型规模不断增长，随之出现了大模型的“扩展定律”（scaling law），即随着**模型规模**、**训练数据**和**计算资源**的增加，**模型性能**会得到显著提升，并且这些关系遵循可预测的模式。
- 英伟达创始人兼首席执行官黄仁勋提出“黄氏定律”(Huang's law)：在“计算架构改进”的推动下，人工智能芯片的性能每年可提升1倍，远远超过了摩尔定律（算力霸权）。

数据规模

$$L(N, D) = \left[ \left( \frac{N_C}{N} \right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_C}{D} \right]^{\alpha_D}$$

模型参数      损失函数



# 人工智能的苦涩经验

## THE BITTER LESSON IN AI



- 2019年，人工智能领域强化学习鼻祖、DeepMind研究科学家、加拿大阿尔伯塔大学计算机学教授理查德·萨顿（Rich Sutton）2019年发表了一篇被称为《苦涩的教训》（The Bitter Lesson）的文章，认为“纵观过去70年的AI发展历史，想办法利用更大规模的算力总是最高效的手段”。也正是在模型规模不断增长理念下，OpenAI 极度注重算法的工程化和工程的算法思维，搭建了工程算法紧密配合的团队架构和计算基础设施，实现了ChatGPT和Sora等核心产品。
- 这一通过不断扩充模型规模而形成“无他、但手熟尔”合成能力的思路一定存在天花板，因为“化繁为简、大巧不工”是推动“机器学习”迈向“学习机器”的初心。

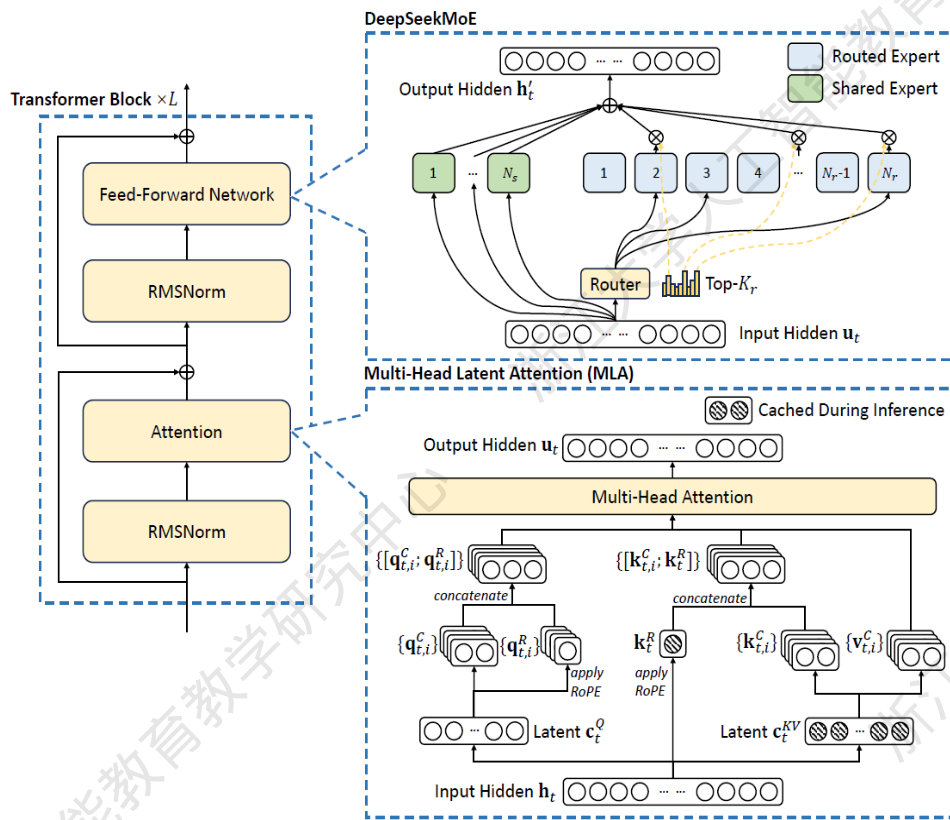


# DeepSeek崛起之因：模型算法和工程优化的系统级协同创新

- DeepSeek模型仍是基于美国谷歌公司于2017年提出的Transformer架构，虽没有实现改变游戏规则的颠覆性基础理论创新，但它在模型算法和工程优化方面进行了系统级创新，在2048块英伟达H800 GPU（针对中国市场的低配版GPU）集群上完成训练，**打破了大语言模型以大算力为核心的预期天花板，为在受限资源下探索通用人工智能开辟了新的道路。**
- 能用众力，则无敌于天下矣；能用众智，则无畏于圣人矣。  
**DeepSeek的精彩表现在于其对算法、模型和系统等进行的系统级协同创新，是众智和众力相互叠加的成果。**



# DeepSeek V3: 混合专家模型灵致、模型参数低秩压缩以及工程化努力

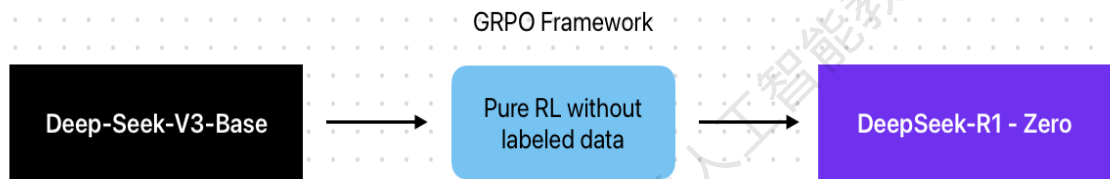


The basic architecture of DeepSeek-V3

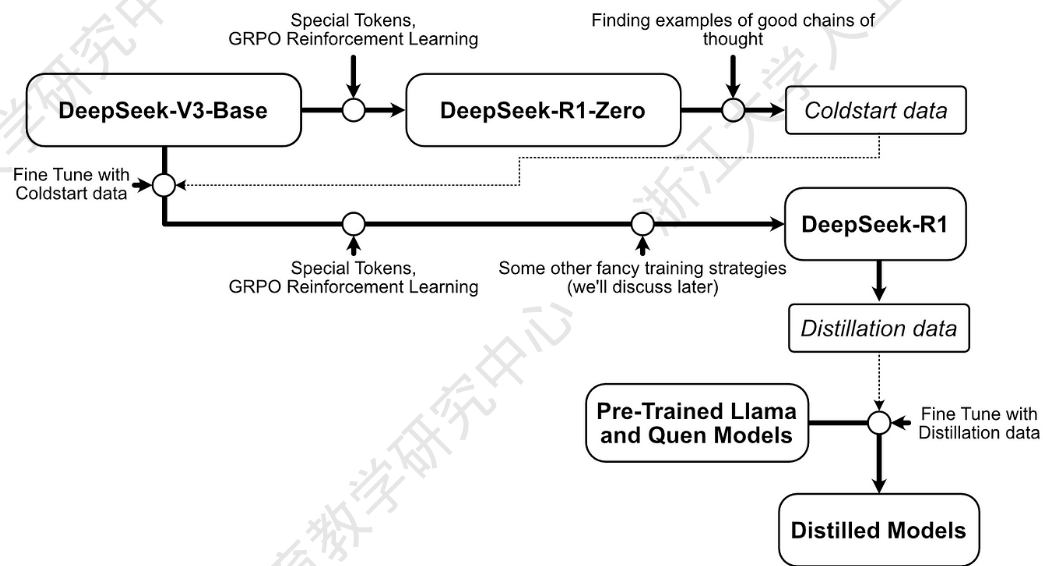
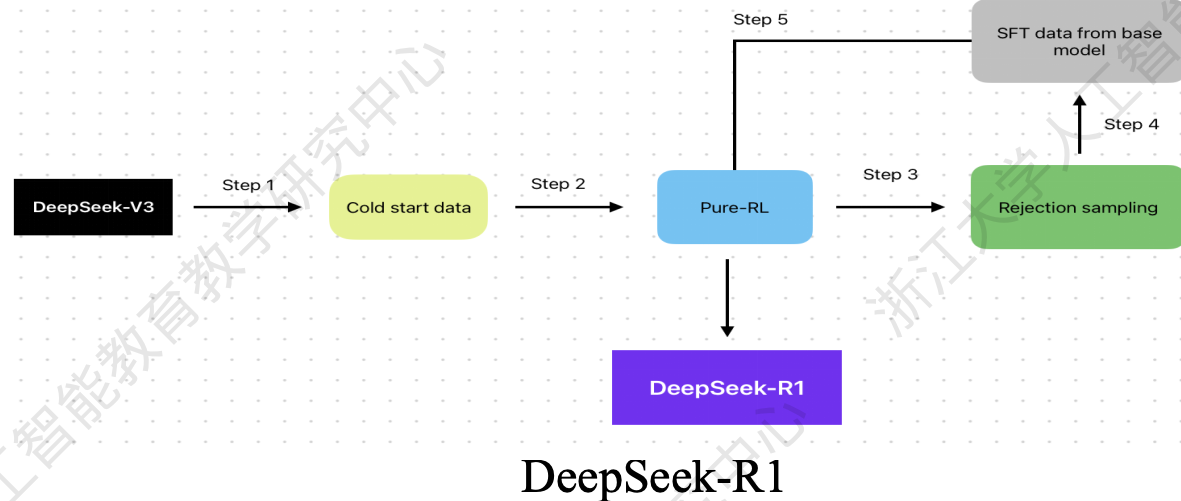
- **混合专家模型:** 不同于传统稠密模型架构（如FFN），DeepSeek的61个Transformers层中58个Transformer层各自包含256个专家和1个共享专家，V3基座模型总共有6710亿参数，但是每次token仅激活8个专家、370亿参数（**~5.5%**）。
- **多头潜在注意力机制:** 引入了低秩这一概念，对巨大的注意力机制矩阵进行压缩，减少参与运算的参数数量，显存占用**仅为其他大模型的5%-13%**。
- **工程化努力:** 使用FP8混合精度加速训练并减少GPU内存使用，使用DualPipe算法（即将前向和后向计算与通信阶段重叠以最大限度地减少计算资源闲置）提升训练效率，进行极致内存优化。

# DeepSeek R1: 强化学习推理和小模型蒸馏

## DeepSeek-R1-Zero (纯粹强化学习)



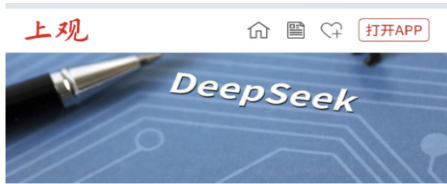
组相对策略优化算法(Group Relative Policy Optimization, GRPO)  
Aha Moment (顿悟时刻)



被DeepSeek R1 蒸馏的小模型

大模型：学而不思则罔；小模型：思而不学则殆

# DeepSeek: 迈向全社会分享的普遍智能 (文汇报: 2025年2月7日)



DeepSeek: 迈向全社会分享的普遍智能

科学新知  
2025-02-03 20:23

来源: 上观新闻 作者: 浙江大学 吴飞

将使AI像水和电一样触手可及。

文匯·上觀

近期, 杭州深度求索人工智能(AI)基础技术研究有限公司(DeepSeek)先后上线大语言基座模型DeepSeek V3, 以及基于V3训练、专为复杂推理任务设计的DeepSeek R1模型, 并同步开源。它们以卓越的性能超越或媲美了全球顶级的开源及闭源模型。



网络版文章刊于立春之日(2025年2月3日, 正月初六)

- DeepSeek的开源之举将使得AI像水和电(以及网络)一样触手可及, 为实现“时时、处处、人人可用的普遍智能”带来曙光。1月28日, 美国“外交学者”(The Diplomat)网站发表题为《中国的DeepSeek是美国人工智能的“斯普特尼克时刻”》的文章指出, DeepSeek此次的开源之举延续了OpenAI的初心使命——**为了人类利益推动人工智能发展。**
- 任何人均可从DeepSeek网站自行下载与部署相关模型。可以预见在不久将来, DeepSeek不同大小模型将被部署为不同场景中的人工智能基座, 大家都可通过行业自有数据、知识和经验进行专业训练和微调, 从而创造无限可能。如果说, 传统大模型遵循的是一条“由通到专”的人工智能发展思路, 那么DeepSeek的做法将**推动形成一条“由专到通”的人工智能发展路径**, 进一步牵引人工智能软硬件技术生态健康发展, **迈向全社会分享的普遍智能之路。**

注: 前苏联于1957年发射人类首颗人造卫星“斯普特尼克1号”, 触发了美苏两霸科技与军事竞赛

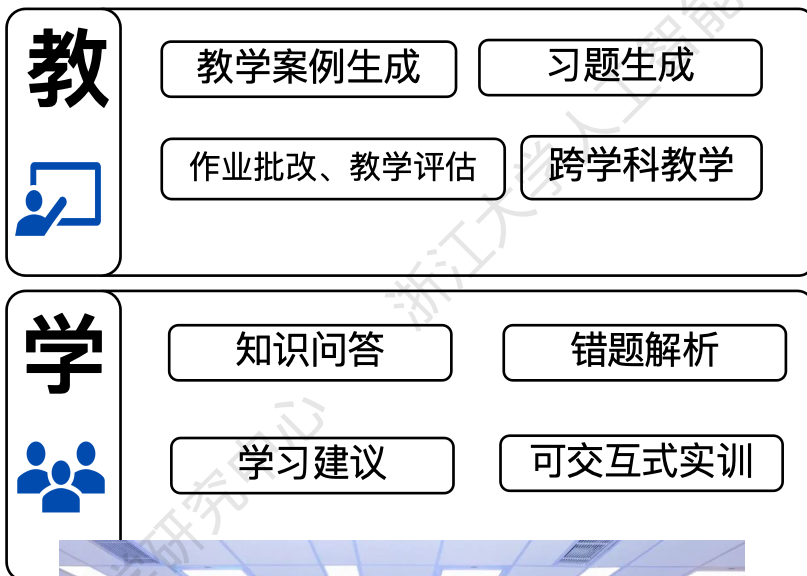
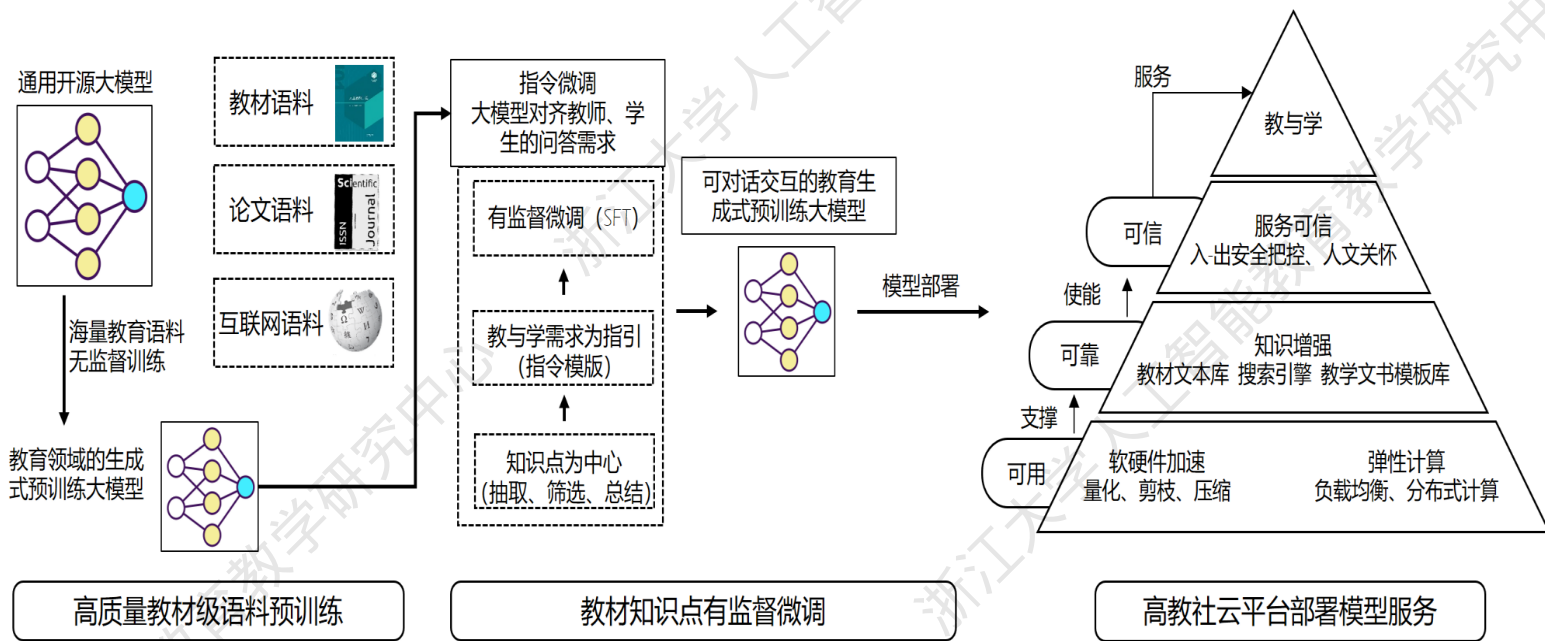
## DeepSeek：极致工程成就大事、开源开放促进信任

- **工程化的创新是大事**：中国工程院院士、中国工程物理研究院研究员李幼平曾经讲过一个故事：他曾请教我国“两弹一星”元勋、两院院士朱光亚先生，为什么九院称“工程物理研究院”。朱光亚先生回答：“物理是深度，工程是规模——没有规模，做不成国家大事”。由此可见，在算力成本呈指数级增长的人工智能领域，通过算法优化、架构突破和工程创新降低大语言模型成本，这本身就是技术实力的体现，是难能可贵的大事。
- **开源开放促进了科技创新**：通过开源开放，技术创新的门槛才能被有效降低，才能汇聚更多的开发者和人才，推动产业的整体进步。由埃里克·雷蒙提出，以Linux创始人林纳斯·托瓦兹（Linus Torvalds）命名的林纳斯定律告诉我们“**只要有足够多的眼睛关注，任何漏洞都无处隐藏**”。因此，人工智能开源将给AI模型注入透明度，并向算法的使用者提供了安全保障，从而构建更安全、更可信和更合乎伦理道德的人工智能技术。
- **DeepSeek并没有解决深度学习面临困惑**：当前生成式人工智能底层逻辑是通过概率合成内容，不可避免会存在不可解释性、AI幻觉、黑箱效应等不足，仍须推动“机器学习”迈向“学习机器”。

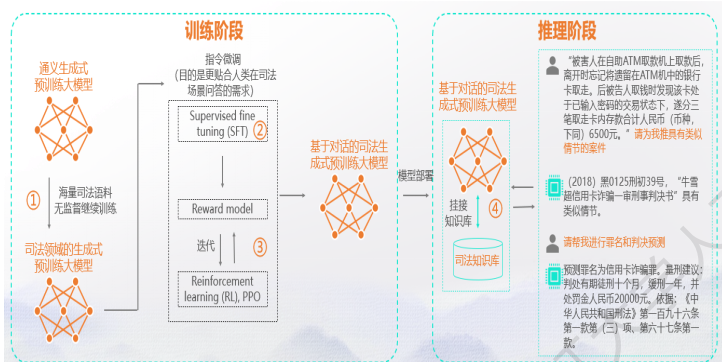


# 垂直领域学科大模型：智海三乐（教育领域）

浙江大学与高等教育出版社、阿里云计算有限公司和华院计算基于通义千问联合研发人工智能领域大模型智海-三乐 (<https://sanle.hep.com.cn>)，提供人工智能知识的智能问答、试题生成、学习导航、教学评估等服务



# 垂直领域学科大模型：智海录问（司法领域）



浙江大学、阿里巴巴达摩院与华院计算合作  
研制司法领域大模型

2023年8月发布智海-录问

阿里巴巴魔搭（ModelScope）社区开源

国家重点研发计划项目：智慧司法智能化感知交互技术研究  
(负责人：吴飞，2021-2023)

Github开源地址：

<https://github.com/zhihaiLLM/wisdomInterrogatory>

Modelscope开放地址：<https://modelscope.cn/models/wisdomOcean/wisdomInterrogatory>

<https://modelscope.cn/studios/wisdomOcean/wisdomInterrogatory>

- 研制“智海-录问”和 LegalMind法律大模型，在浙江省高级人民法院和浙江省司法厅等部署，月均使用量超过150万余次
- 在浙江省高级人民法院等60家法院落地使用，覆盖民事、刑事和行政等45种案由，辅助庭审案件超过1.5万件，当庭宣判率达90%以上，裁判文书完整度达95%以上，提高了审判效率近40%。
- 智慧司法智能化支撑平台与示范应用获得中国人工智能学会2024年度吴文俊人工智能科技进步一等奖

# 垂直领域学科大模型

## 智海系列垂直领域大模型 | 已发布

教育领域	司法领域	金融领域	财会领域
 智海-三乐	 智海-录问	 智海-金磐	 智海-大禹
<a href="https://sanle.hep.com.cn/">https://sanle.hep.com.cn/</a>	<a href="https://modelscope.cn/studios/wisdomOcean/wisdomInterrogatory/summary">https://modelscope.cn/studios/wisdomOcean/wisdomInterrogatory/summary</a>	<a href="https://jinpan.mjyos.com">https://jinpan.mjyos.com</a>	<a href="https://modelscope.cn/studios/DayuLLM/DayuLLM-v1.0">https://modelscope.cn/studios/DayuLLM/DayuLLM-v1.0</a>

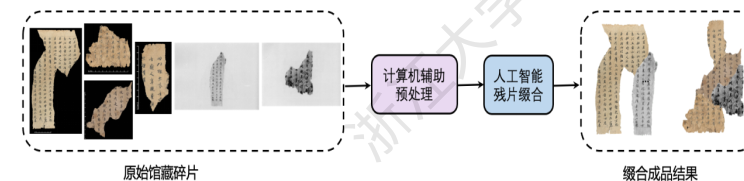
## 智海系列垂直领域大模型

凡贵通者，贵其能用之也  
人工智能作为通用目的使能技术与垂直领域结合将推动该领域范式变革

<b>智海-三乐</b> 教育大模型	<b>智海-敦煌</b> 敦煌文化大模型
<b>智海-录问</b> 司法大模型	<b>智海-知行</b> 大模型协作系统
<b>智海-金磐</b> 金融大模型	<b>伏羲-万象</b> 地理科学大模型
<b>智海-大禹</b> 财会大模型	<b>智海-悟化</b> 药物大模型
	<b>智海-烛照</b> 影视大模型
	<b>智海-开物</b> 设计大模型



影视作品生成



敦煌残片缀合



药物逆合成

人工智能本身颇具学科交叉内在禀性，垂直领域智能基座模型（foundation model）将深度学习架构在人类不同学科专业知识之上，开辟了数据驱动和知识引导的人工智能研究新模式，促进科学发现、社会治理、司法创新和探索等迈向赋能新时代

# 提 纲

- 1 从达特茅斯启航的人工智能三大主义
- 2 从 ChatGPT 到 DeepSeek
- 3 人工智能通识教育

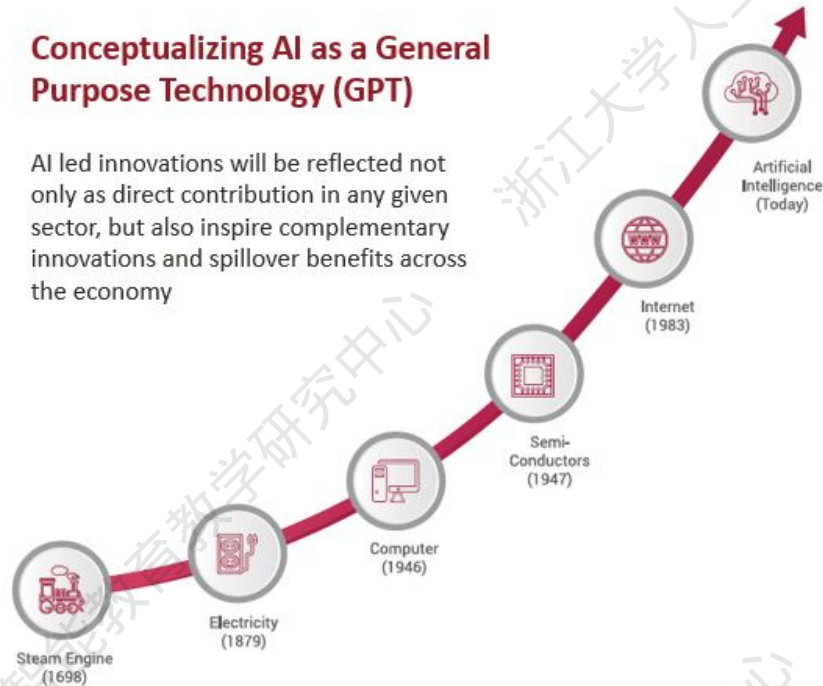


# 人工智能：通用目的技术（GPT）

1987年诺贝尔经济学奖获得者美国经济学家索洛研究表明，国民经济最终会达到这样一种稳态发展阶段：在那个阶段以后，经济增长将只取决于技术的进步。虽然技术进步是经济增长的源泉，但是长期的经济增长是由少数几种被称之为“通用目的技术”（general purpose technologies, 简称GPTs）而驱动。

## Conceptualizing AI as a General Purpose Technology (GPT)

AI led innovations will be reflected not only as direct contribution in any given sector, but also inspire complementary innovations and spillover benefits across the economy



人类不同时代的通用目的技术

- 在人类发展历史上，蒸汽机、电力、计算机、半导体和互联网等与人工智能一样，都是通用目的技术，具有普遍适用性、动态演进性和创新互补性特点。
- 美国历史学家斯塔夫里阿诺斯（Leften Stavros Stavrianos）在《全球通史》一书赞誉为“蒸汽机的历史意义无论怎样夸大都不过”。蒸汽机与纺织、交通和冶金等工业结合，推动人类迈入工业革命时代。
- 通用目的技术是最核心创新要素，但不是完整的最终解决方案。比如瓦特在1795年改良了蒸汽机，但是直至这之后近百年，当蒸汽机与纺织、交通和冶金等工业紧密结合，使得机械动力迅速取代了人力、风力、水力和畜力，蒸汽机对劳动生产率的贡献才达到顶峰，推动人类迈入工业革命时代，突破了“马尔萨斯陷阱”。



# 浙江大学成立人工智能教育教学研究中心

2024年3月，浙江大学成立人工智能教育教学研究中心，面向全校本科生开设人工智能通识必修课程，打造人工智能通识课程体系和实训范式，直面“智能时代、教育何为”挑战，让更多人成为人工智能这一通用智能技术的受益者。。



课程体系

构建人工智能类本研公共课程体系

实践平台

出版人工智能系列高水平教材

培养项目

打造人工智能系列人才培养项目

师资队伍

组建跨学科高水平师资队伍

实践平台

强化核心实践创新能力培养

赋能教学

推进人工智能赋能教育教学

素养要求

发布人工智能系列研究报告

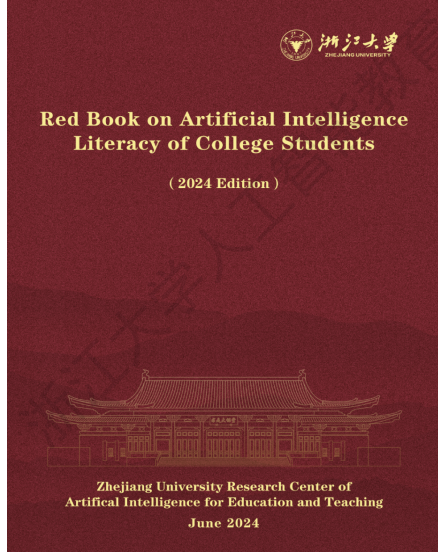
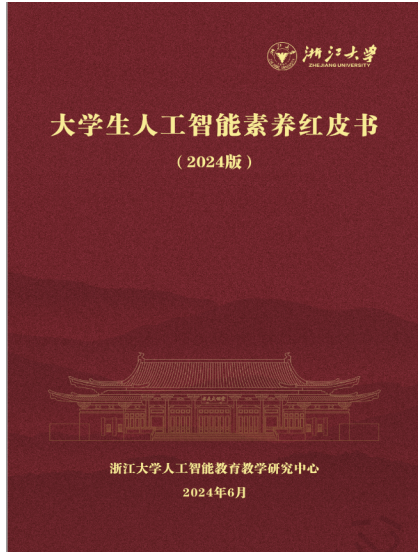
# 浙江大学人工智能通识必修课程

在教育部101计划核心课程《人工智能引论》的建设基础上，浙江大学推出了“人工智能基础”系列通识课程，分别面向理工农医类、社会科学类和人文艺术类专业学生，出版人工智能通识课程教材，培训人工智能通识课程师资。

课程名称	面向专业	学分
人工智能引论	计算机类 (信息安全/人工智能/计算机图灵班、工业设计)	3
人工智能基础 (A)	理工农医类 (不含计算机类及开设人工智能必修课的专业)	2
人工智能基础 (B)	社会科学类 (需先修程序语言基础)	2
人工智能基础 (C)	人文艺术类 (无需修读程序语言基础)	2

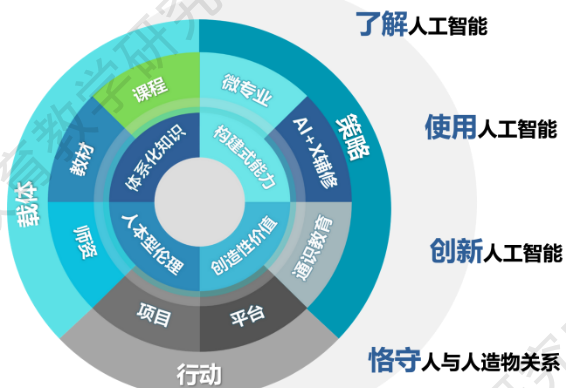


# 浙江大学发布《大学生人工智能素养红皮书》



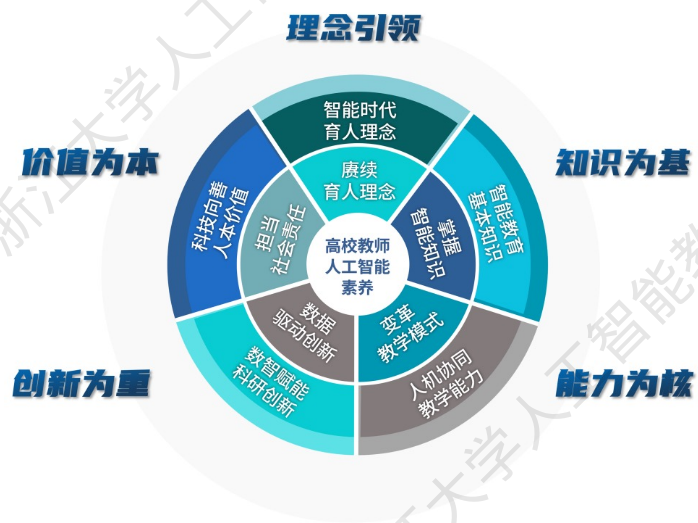
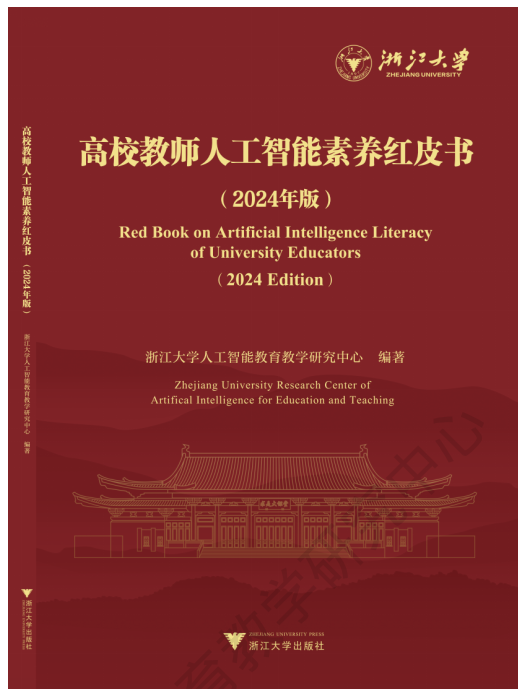
先前的技术发明从机械化增强角度提升了人类与环境的互动能力，然而，人工智能的出现挑战了人类的根本，它深刻改变了人类与环境互动的能力和角色。近来生成式人工智能的出现使得智能机器成为知识生产的辅助者，对个体学习者的自主思考、判断、学习能力乃至伦理道德观提出了挑战。

围绕“智能时代、教育何为”这一命题，本红皮书提出大学生人工智能素养的构成内涵、培养的目标与愿景以及培养的载体、行动与策略，认为大学生人工智能素养是由体系化知识、构建式能力、创造性价值和以人为本型伦理构成的有机整体，其中知识为基、能力为重、价值为先、伦理为本。





# 浙江大学发布《高校教师人工智能素养红皮书》



本红皮书旨在提出高校教师人工智能素养的概念与内涵以及提升的目标、路径与保障，认为高校教师人工智能素养是指在高校从事教学科研工作的教师为了在智能时代胜任教书育人、科研创新、社会服务及文化传承等工作而应具有的一种与人工智能应用相关的专门素养，它包含**赓续育人理念（何为师）、学习智能知识（以何为师）、变革教研模式（何以成师）和担当社会责任（师者为何）**等能力，具体包括智能时代育人理念、智能教育基本知识、人机协同教学能力、数智赋能科研创新以及科技向善人本价值等五个纬度的内容，其中，**理念引领、知识为基、能力为核，创新为重、价值为本。**



# 从新一代人工智能系列教材迈向新一代人工智能通识系列教材

## 新一代人工智能系列教材（理论系列，26本）

教材名	作者	作者单位
人工智能导论：模型与算法	吴飞	浙江大学
可视化导论	陈为、赵焱、张嵩、鲁爱东	浙江大学、密西西北州立大学、北卡罗来纳大学夏洛特分校、肯特州立大学
智能产品设计	孙凌云	浙江大学
自然语言处理	刘挺、秦兵、赵军、黄童青、车万翔	哈尔滨工业大学、中科院大学、复旦大学
人脸图像合成与识别	高新波、王楠楠	西安电子科技大学
物联网安全	徐文源、冀晓宇、周歆妍	浙江大学、宁波大学
人工智能伦理导论	古天龙	暨南大学
金融智能：理论与实践	郑小林、朱梦莹、陈超超	浙江大学
模式识别	周杰、郭振华、张林	清华大学、同济大学
赋能：人工智能与数字经济	王延峰、于晓宇、史占中、吴明辉、李泉、周耀、俞凯、惠慧、熊友军	上海交通大学
机器学习基础	李宏亮、孟凡满、吴庆波	电子科技大学
自主智能运动系统	薛建伟	西安交通大学
深度学习基础	刘运超	哈尔滨工业大学
人工智能芯片与系统	王刚可、李莹、李英明	浙江大学
计算机视觉	程明明	南开大学
神经认知学	唐华锋、潘炳	浙江大学
人工智能伦理与安全	秦湛、潘恩荣、任章	浙江大学
媒体计算	韩亚洪、李泽超	天津大学、南京理工大学
人工智能逻辑	廖备水、刘奋荣	浙江大学、清华大学
数字生态：人工智能与区块链	吴超	浙江大学
人工智能内生安全	姜育刚	复旦大学
数据科学前沿技术导论	高云君、陈曦、苗晓晖、张天明	浙江大学、浙江工业大学
遥感图像智能分析与处理	尹继家、罗晓燕、飞絮教材编写组	北京航空航天大学
具身智能导论	刘华平、鄧迪、孙富春	清华大学
因果发现与推断	李康	合肥工业大学
*《人工智能引论》	吴飞、潘云鹤	浙江大学

\*教育部计算机101计划核心课程（教材）

## 新一代人工智能系列教材（实践系列，11本）

教材名	作者	作者单位
智能之门：神经网络与深度学习入门（基于Python的实现）	胡晓武、秦婷婷、李超、邹欣	微软亚洲研究院
人工智能基础	徐增林、康昭	哈尔滨工业大学（深圳）
跨媒体移动应用理论与实践	张克俊、叶雨晴、吴宋若理、俞佳兴	浙江大学
计算机视觉理论与实践	刘家琛、杨帅、杨文瀚、段凌宇	北京大学
语音信息处理理论与实践	王龙标、党建忠、于强	天津大学
机器学习	胡清华、杨柳	天津大学
深度学习技术基础与实践	吕建成、段磊、张卫华、聂永胜、耿天玉	四川大学
自然语言处理理论与实践	黄河燕、史树敏、李洪政	北京理工大学
人工智能导论：案例与实践	朱强、飞絮教材编写组	浙江大学、百度
智能驾驶技术与实践	黄宏成	上海交通大学
人工智能芯片编译技术与实践	蒋力	上海交通大学



- 在成立于2018年的新一代人工智能系列教材编委会指导下，出版了理论教材26本、实践教材11本，形成了各具优势、衔接前沿、涵盖完整、交叉融合的人工智能教材体系。
- 2024年12月28日，启动新一代人工智能通识系列教材编写工作。

# 国家教材建设重点研究基地（高等学校人工智能教材研究）

## 入选“国家教材建设重点研究基地（高等学校人工智能教材研究）”

1

开展人工智能及其交叉学科的知识体系及教材研究

2

构建人工智能高水平教材体系，建设相应数字化资源

3

建设人工智能教材研究队伍

4

培养人工智能课程教材建设专业人才

5

交流传播人工智能教材研究成果

### 中华人民共和国教育部

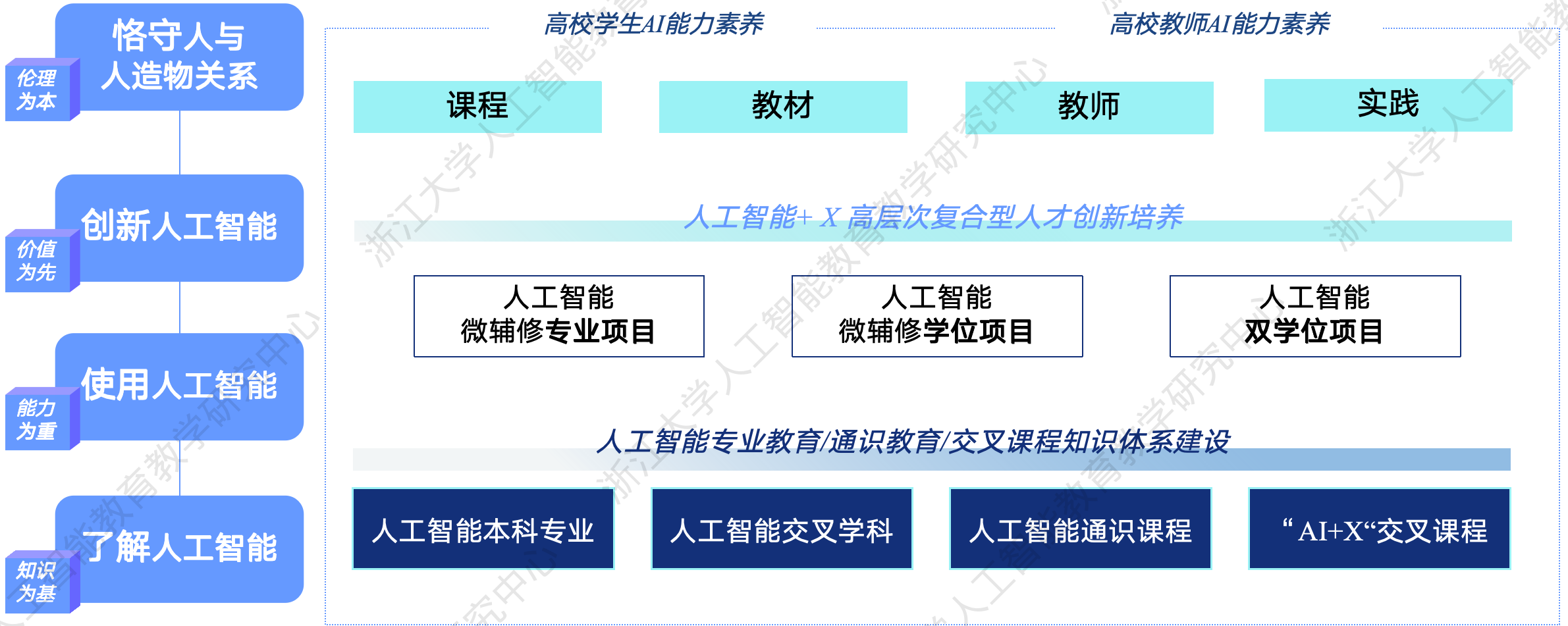
教材函〔2024〕1号

#### 教育部关于2024年度国家教材建设 重点研究基地认定结果的通知

各省、自治区、直辖市教育厅（教委），新疆生产建设兵团教育局，有关部门（单位）教育司（局），部属各高等学校、部省合建各高等学校，有关直属单位：

根据《教育部办公厅关于组织申报国家教材建设重点研究基地的通知》（教材厅函〔2024〕5号），经有关单位自主申报，第三方专业机构组织评审，教育部组织公示，现认定北京师范大学申报的国家教材建设重点研究基地（大中小学德育一体化教材研究）等31个基地为2024年度国家教材建设重点研究基地（名单见附件）。

# 人工智能体系化人才培养载体





# 基础教育中人工智能通识教育

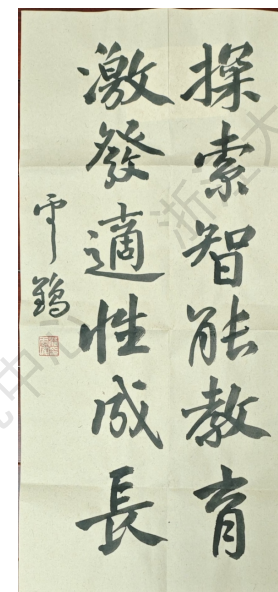
与杭州外国语学校联合成立“人工智能赋能基础教育创新研究中心”；与求是教育集团联合组建求是人工智能创新实践中心；与建兰中学建设“建兰人工智能创新探究中心”



高中：形成一批适用于基础教育的生动活泼人工智能教学案例与课程

开篇智能教育  
传承求是精神  
甲辰宁鹤

小学：通过人工智能案例体验，激发小学生对人工智能的兴趣



初中：通过人工智能与学科课程内容的结合，激发适性成长



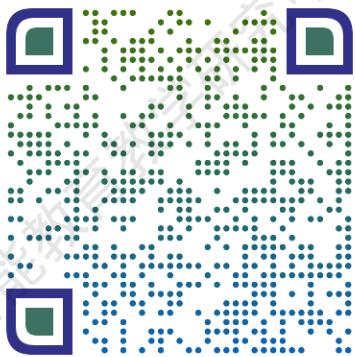
# 人工智能科普通识读物

在高等教育出版社支持下，打造了原创人工智能前沿科普有声通识数字栏目《走进人工智能》和科普通识读物《走进人工智能》，**潘云鹤院士作序：未来将是人和人工智能共同进化的时代...科学普及将人类进化中累积知识转化为人和人造物的力量。**

有专业高度、显学理深度、含人文温度

## 《走进人工智能》章节目录

- 前言：走进人工智能
- 第一篇 从机器人偶到图灵机模型：迈向自动计算时代
- 第二篇 原点：从达特茅斯启航
- 第三篇 从专家系统到深蓝：在逻辑推理与优化搜索中成长
- 第四篇 从信息载体到智能燃料：数据的蝶变
- 第五篇 从机器学习到学习机器：智能算法的心之所向
- 第六篇 从华丽转身到炼金术之困：深度学习
- 第七篇 从最优解到均衡解：博弈论拥抱人工智能
- 第八篇 从个体智能到群体智能：整体大于部分总和
- 第九篇 从知其然到知其所以然：因果推理
- 第十篇 从单通道独奏到多通道协同：跨媒体计算
- 第十一篇 从摩尔定律到黄氏定律：人工智能算力之源
- 第十二篇 从硅基之力到碳基之巧：智能的混合增强
- 第十三篇 科学第三极：美美与共的科学计算
- 第十四篇 从预测决策到内容合成：ChatGPT的涌现之力
- 第十五篇 人工智能的双重属性：技术价值与社会价值的统一
- 第十六篇 人工智能、教育先行：成天下之才



截止目前，共32万+人次收听

入选2024年度第八届  
“中国科普作家协会优秀科普作品奖”银奖

致天下之治者在人才，成天下之才者在教化，教化之所本者在学校

人工智能  
教育先行  
產學協作  
引領創新

雲鶴

使能技术、赋能社会：人工智能是引领这一轮科技革命、产业变革和社会发展的战略性技术，具有溢出带动性很强的头雁效应，其作始也简，其将毕也必巨