机器学习系统 2026春

第一章 绪论

张燕咏 讲席教授 yyz@ustc.edu.cn
张午阳 特任教授 wuyangz@ustc.edu.cn

中国科学技术大学
University of Science and Technology of China

教师简介

张燕咏 讲席教授
中国科大人工智能与数据科学学院执行院长
中组部“千人计划”创新人才长期项目
yyz@ustc.edu.cn

- 2006， NSF Career Award
- 2015，Rutgers 教授
- 2017，IEEE Fellow
- 2018， 全职加入中国科大

张午阳 特任教授
中国科大人工智能与数据科学学院
国家青年人才
wuyangz@ustc.edu.cn

- 2020 美国罗格斯大学，博士
- 2020 日本东京大学访问学者
- 2020-2025 美国 Meta 主任级研究科学家

龙凯雪 课程助教
kaixue@mail.ustc.edu.cn

刘嘉政 课程助教
liujiazheng25@mail.ustc.edu.cn

课程前置要求

- Python基础编程
- 计算机体系结构
- 机器学习/算法导论

G2-B403
上课时间 周二 9:45-12:10

课程大纲

- 第一章 概述
- 第二章 机器学习与框架基础
- 第三章 GPU框架与CUDA编程
- 第四章 Transformer和大语言模型
- 第五章 混合专家模型
- 第六章 数据并行与模型并行
- 第七章 模型剪枝
- 第八章 模型量化
- 第九章 机器学习图优化
- 第十章 大语言模型部署
- 第十一章 大模型后训练
- 第十二章 客座讲座
- 第十三章 课程项目展示与总结
- 第十四章 课程项目展示与总结

3学分　40理论学时+20实验学时

课程考核

- Lab1 （个人） 10%：大模型qwen3 部署
- Lab2（个人） 20%：大模型的性能瓶颈分析
- Lab3（个人） 20%：基于vLLM的prefill/decode调度策略实现
- 课程大项目（2-3人团队） 40%： 一个高效大模型推理优化策略设计
- 课堂小测 x 2 10%

实验平台

- 本地环境

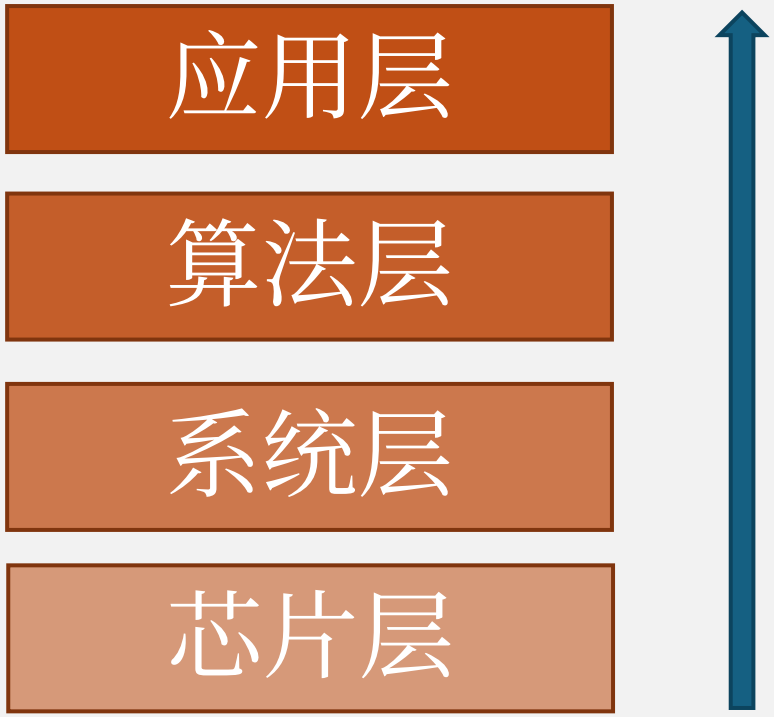
- 学院公共算力计算平台

http://user.saids.hpc.gleamoe.com/

本章节内容

- 为什么学院要开这门课
- 为什么同学来上这门课
- 人工智能发展背景
- 机器学习系统概述

人工智能技术分层



人工智能底层科技的缺失可能使得我国智能产业成为空中楼阁

人工智能方向应该培养什么样的人才?

两个参考问题

- 汽车专业应该培养什么样的人才?
  - 清华大学汽车专业培养目标：具有从事汽车、发动机和汽

人工智能方向应该培养人工智能（子）系统的设计者和研究者

- 计算机专业当培养计算机整机或子系统的设计者和研究者

人工智能方向应该培养什么样的人才?

- 只包含各类机器学习算法、视听觉应用这条软件线，只能算是“*人工智能应用专业*”或者“*人工智能算法专业*”
- 但当模型走向“大规模训练与部署”，决定成败的往往是 **系统能力**：**算力、内存、通信、软件栈、工程化与可复现**
- 谷歌有世界上最大的AI算法研究团队，然而
  - 董事长John Hennessy是计算机体系结构科学家，图灵奖得主
  - 谷歌AI的负责人Jeff Dean是计算机系统研究者
  - 谷歌AI 2025最重要的三个进展（Gemini3.0、TPU v7、Jax AI），不仅仅是某个特定算法，而是整个机器学习系统的联合优化
- OpenAI ChatGPT的成功很大程度上来源于系统发展
  - 大模型训练不是“写个算法就能跑”
  - **上万张 A100** 级别 GPU 的集群协同（计算 + 通信 + 容错 + 调度）
  - 单次训练成本可达 **千万美元级别**

课程开设背景

在高年级本科生、硕士研究生阶段，人工智能与数据科学学院设置一门**系统类课程**，能帮助同学实现对当前主流**智能软硬件体系的融会贯通**，具备自己动手完成一个**完整机器学习系统**的能力。

这门课程就是**机器学习系统**

什么是机器学习系统?

机器学习系统是智能的物质载体

它把算法、数据与算力组织起来，使模型能够在真实世界中可训练、可推理、可部署、可维护

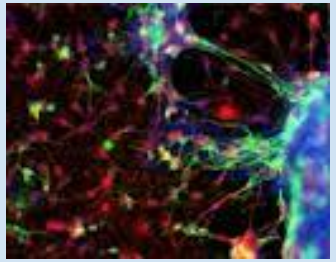
现阶段的机器学习系统通常是集成CPU和加速器的异构系统，软件上通常包括一套面向开发者的智能计算编程环境（包括编程框架和编程语言）

机器学习系统解决的不是 模型能不能算，而是 模型如何在约束下高效、稳定、可规模化地运行?

机器学习系统的形态

超级计算机

商业分析



药物研制

数据中心

Baidu 百度

YouTube

广告推荐

自动翻译

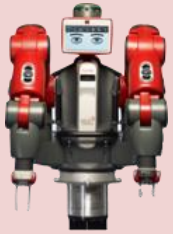
智能手机

语音识别

图像分析

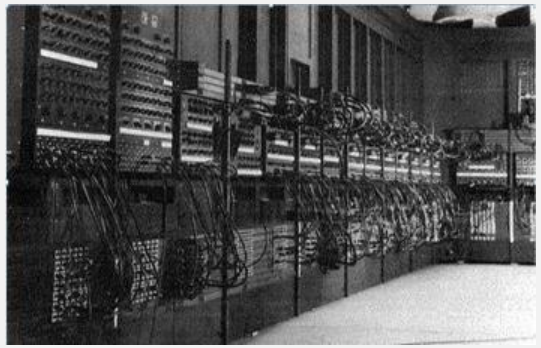
嵌入式设备

自动驾驶



消费类电子

机器学习系统具有重大价值



上世纪人类从工业时代过渡到信息时代

现在已经发展到向智能时代进化的拐点

中国需要一大批机器学习系统的开发者和设计者

本章节内容

- 为什么学院要开这门课
- 为什么同学来上这门课
- 人工智能发展背景
- 机器学习系统概述

为什么来上这门课?

好学生就是要上课?

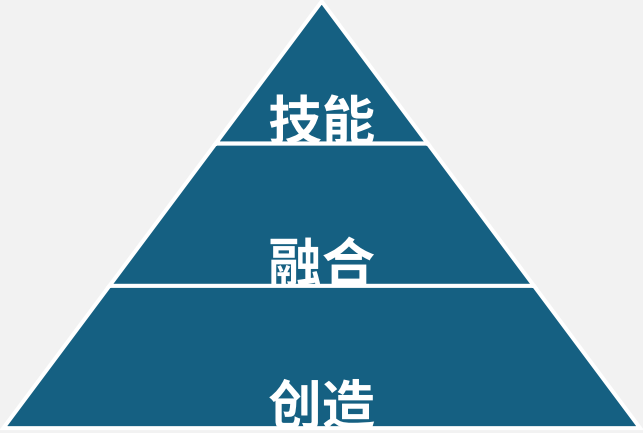
具备实际解决问题能力是上课的目标

应用驱动、全栈贯通的课程体系

一门帮助学生学以致用、形成全局系统观的工科课程

课程目标和目的

- 中国需要一大批智能基础设施的开发者和设计者
- 专业普及课程
  - 应用驱动，全栈贯通
- 机器学习系统
  - 建立机器学习系统设计及应用的知识体系
  - 掌握智能应用开发的基本技能
  - 培养开展机器学习系统基础研究的兴趣和能力



能用机器学习系统，能开发智能应用

实现对智能计算软硬件技术栈的贯通理解

创造未来机器学习系统的科研基础

机器学习系统课程对学生的价值

- **更全面的实践能力**
- 只会调参、缺乏系统知识的同学，往往对**耗时/耗电/显存/成本**缺乏直觉
- 结果是：算法在论文里很好看，但在真实平台上**跑不动、跑不起、跑不稳**
- 机器学习系统训练你把算法落地：从 **PyTorch → 运行时 → CUDA kernel → 分布式 → 部署推理**

- **更强的工程与产业竞争力**
- 会用 PyTorch，可能拿到**40 万人民币级**岗位
- 参与设计/优化 PyTorch（框架、编译、算子、并行、部署），往往进入**40 万美元级**赛道
- 核心差异：你解决的是“别人也能用”的问题，还是“别人用得更快、更省、更稳”的问题?

机器学习系统课程对学生的价值

- **更强的研究能力**
- 系统视角让你把研究目标从单一指标扩展为：
  - **准确率 × 延迟 × 吞吐 × 成本 × 能耗 × 稳定性 × 可复现性**
- 能提出更“真实”的研究问题：为什么某方法在小规模有效但大规模失效？瓶颈在哪一层？能否跨层级进行优化？

- **形成系统科研思维**
- 从优化模型结构升级为设计一个端到端系统：算法只是其中一环
- 系统思维带来更广阔的科研空间：框架、编译、硬件协同、并行、推理服务、能效优化等，都能做出硬成果

机器学习系统课程对老师的价值

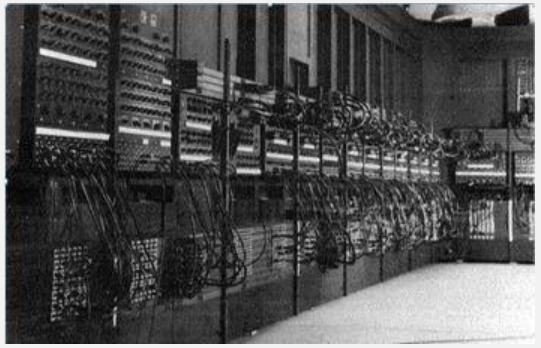
- 《礼记·学记》“教学相长”、开阔思路
- 美国计算机方向Top4高校Stanford、CMU、UC Berkley和MIT以及多个国际单位联合发布了白皮书——“SysML: The New Frontier of Machine Learning Systems”
  - 包括Yann LeCun、Michael I. Jordan、Bill Dally和Jeff Dean等
- 培养教授机器学习系统课程的教师，能抢占这一国际热点方向、也是未来重要学科增长点的先机

本章节内容

- 为什么学院要开这门课
- 为什么同学来上这门课
- 人工智能发展背景
- 机器学习系统概述

智能时代

蒸汽机



集成电路

机器学习系统

上世纪人类从工业时代过渡到信息时代
现在已经发展到向智能时代进化的拐点

中国需要一大批机器学习系统的开发者和设计者

全球国家AI战略

HOLONIQ. GLOBAL INTELLIGENCE

Holon IQ

Global AI Strategy Landscape

50 National Artificial Intelligence Policies as at February 2020.

Argentina
Drafting the "National Plan of Artificial Intelligence". Falls under the Innovative Argentina 2030 Plan and the 2030 Digital Agenda.

Australia
November 2019, AI Roadmap focused on specialization in health, infrastructure and natural resources. Planning for an additional 161,000 AI specialists by 2030.

Austria
June 2019, 'Artificial Intelligence Mission Austria 2030 (AIM AT 2030)'. Outlines seven fields for which AI will be critical

Belgium
March 2019, 'AI 4 Belgium' launched and includes seven major objectives.

Brazil
Consultation period ended January 2020. Building a network of eight research facilities focused on artificial intelligence.

Canada
2017 federal budget announced five-year, $125m plan. Led by CIFAR. Research and talent focus. First National AI Strategy.

Chile
Expected April 2020. Ministry of Science, Technology, Knowledge, and Innovation created a committee of 10 experts to develop.

China
July 2017, China launched the most comprehensive AI strategy globally with 2030 targets for a $1T RMB AI industry.

Colombia
November 2019, first draft issued for 'National Policy for Digital Transformation'. Medellin to become an AI & Robotics Centre of Excellence.

Czech Republic
May 2019, 'National Artificial Intelligence Strategy of the Czech Republic' was launched.

Denmark
March 2019, Denmark announced the 'National Strategy for Artificial Intelligence' with four key objectives.

Estonia – Kratts Strategy
May 2019, Estonian AI experts, led by government CIO produced a roadmap, later adopted as the Estonian National AI Strategy in July 2019.

Finland
June 2019, 'Leading the Way into the Age of Artificial Intelligence' identified 11 key actions following May 2017 Steering Group announcement.

France
€1.5 billion plan announced in 2018 influenced by the 'Villani Report' to transform France into a global leader in AI.

Germany
€3 billion plan announced Nov 2018 with a dedicated AI strategy to make Germany & Europe a global leader in AI.

Hungary
October 2019, Hungary announced an AI Action Plan, the first pillar of a national AI strategy, expected in 2020.

India
June 2018 working paper on using AI to ensure social growth, inclusion and positioning the country as a leader in AI.

Indonesia
Indonesia Artificial Intelligence Society (IAIS) inaugurated under Smart Indonesia in October 2019. National Strategy expected in 2020.

Ireland
Irish Economic Development Agency led process. AI Master program launched in 2018 and is 100% industry driven.

Israel
Innovation Authority, tasked with AI policies, has warned that a strategy is needed to prevent falling behind.

Italy
March 2018, AGID released a White Paper called "AI at the service of citizens," which was edited by the AI Task Force.

Japan
March 2017, Japan's AI policy, the 'Artificial Intelligence Technology Strategy', was announced second only to Canada with 'Society 5.0'.

Kenya
January 2018, government announced task force to create a five-year strategy on national use of emerging technologies.

Lithuania
April 2019, Artificial Intelligence Strategy announced "to modernize and expand the current AI ecosystem and ensure that the nation is ready"

Luxembourg
May 2019, launched 'Artificial Intelligence: a strategic vision for Luxembourg'.

Malaysia
2018, Malaysia revealed a National Artificial Intelligence Framework expanding the National Big Data Analytics Framework.

Malta
October 2019. 'A Strategy and Vision for Artificial Intelligence in Malta 2030' Malta.ai launched and aspiring to be the 'Ultimate AI Launchpad'.

Mexico
June 2018, 'Towards an AI Strategy in Mexico: Harnessing the AI Revolution', serves as a foundation for building full AI strategy.

Netherlands
November 2018. AINED published a roadmap for developing a full national strategy.

New Zealand
May 2018, AI Forum of New Zealand, released "Artificial Intelligence: Shaping a Future New Zealand."

Norway
January 2020, Norway issued its National Strategy for Artificial Intelligence.

Pakistan
Presidential Initiative for Artificial Intelligence launched December 2018, focused on training beginners in AI and advanced technology.

Philippines
Nov 2019. AIM, Aboitiz School of Innovation, Technology and Entrepreneurship (ASITE) appointed to craft an AI roadmap.

Poland
November 2019, 'Assumptions for the AI strategy in Poland' as an action plan towards developing an AI strategy.

Portugal
February 2019, 'AI Portugal 2030', seeks strengthen economic growth, scientific excellence, and human development using with AI.

Qatar
October 2019, National AI Strategy as a blueprint produced by Qatar Computing Research Institute (QCRI).

Russia
October 2019, Russia published its National Strategy for the Development of Artificial Intelligence by 2030.

Saudi Arabia
September 2019. Royal decree to establish an AI center, to align with the Kingdom's Vision 2030 program.

Singapore
May 2017. AI Singapore is a five-year, S$150 million national program launched in to enhance Singapore's capabilities in AI.

South Africa
Intsimbi Future Production Technologies Initiative" launched in 2018 with aim to advancing manufacturing sector.

South Korea
May 2018, five-year AI development plan launched with $1.95B budget.

Spain
March 2019, the Spanish Ministry of Science, Innovation and Universities launched the RDI Strategy in Artificial Inteligence.

Sweden
National Approach for Artificial Intelligence launched in May 2018.

Switzerland
An Artificial Intelligence (AI) expert group has published its recommendations for a Swiss AI strategy.

Thailand
Thailand's Digital Economy and Society (DES) Ministry has drafted the country's first artificial intelligence (AI) ethics guidelines.

Tunisia
AI Task Force and Steering Committee to develop a national AI strategy. The strategy was scheduled to be published in the first quarter of 2019.

United Arab Emirates
October 2017 announced strategy. First country to create a Ministry of AI and first in the Middle East to launch an AI strategy.

United Kingdom
April 2018, 'Sector Deal' announced. $1.24B funding as part of the UK's larger industrial strategy.

United States of America
February 2019 by Executive Order to promote and protect AI technology. AI.gov launched Mar 2019. Followed by the National Artificial Intelligence Research and Development Strategic Plan.

Vietnam
Ministry of Information and Communications developing a broad AI strategy.

Source: HolonIQ and source government strategy and policy papers.

www.holoniq.com

国家AI战略

China New Generation of Artificial Intelligence Development Plan

China announced its ambition to lead the world in AI in its July 2017 development plan, A Next Generation Artificial Intelligence. The plan is the most comprehensive of all national AI strategies, with initiatives and goals for R&D, industrialization, talent development, education and skills acquisition, standard setting and regulations, ethical norms, and security.

By 2030, the government aims to cultivate an AI industry worth 1 trillion RMB, with related industries worth 10 trillion RMB. In addition, the government has also partnered with national tech companies to develop research and industrial leadership in specific fields of AI and will build a $2.1 billion technology park for AI research in Beijing.

中华人民共和国中央人民政府
www.gov.cn

国务院关于印发
新一代人工智能发展规划的通知
国发〔2017〕35号

新一代人工智能发展规划

State Council Notice on the Issuance of the Next Generation Artificial Intelligence Development Plan
Completed: July 8, 2017
Released: July 20, 2017

A Next Generation Artificial Intelligence Development Plan

The rapid development of artificial intelligence (AI) will profoundly change human society and life and change the world. To seize the major strategic opportunity for the development of AI, to build China's first-mover advantage in the development of AI, to accelerate the construction of an innovative nation and global power in science and technology, in accordance with the requirements of the CCP Central Committee and the State Council, this plan has been formulated.

I. The Strategic Situation

The development of AI has entered a new stage. After sixty years of evolution, especially in mobile Internet, big data, supercomputing, sensor networks, brain science, and other new theories and new technologies, under the joint impetus of powerful demands of economic and social development, AI's development has accelerated, displaying deep learning, cross-domain integration, man-machine collaboration, the opening of swarm intelligence, autonomous control, and other new characteristics. Big data-driven cognitive learning, cross-media collaborative processing, and man-machine collaboration-strengthened intelligence, swarm integrated intelligence, and autonomous intelligent systems have become the focus of the development of AI. The results of brain science research inspired human-like intelligence that awaits action; the trends involving the chips, hardware, and platform have become apparent; the development of AI has entered into a new stage. At present, the development a new generation of AI and related disciplines, theoretical modeling, technological innovation, hardware and software upgrades, etc., all advance, provoking chain-style breakthroughs, promoting the acceleration of the elevation of economic and social domains from digitization and networkization to intelligentization.

AI has become a new focus of international competition. AI is a strategic technology that will lead in the future; the world's major developed countries are taking the development of AI as a major strategy to enhance national competitiveness and protect national security; intensifying the introduction of plans and strategies for this core technology, top talent, standards and regulations, etc.; and trying to seize the initiative in the new round of international science and technology competition. At present, China's situation in national security and international competition is more complex, and [China] must, looking at the world, take the development of AI to the national strategic level with systemic layout, take the initiative in planning, firmly seize the strategic initiative in the new stage of international competition in AI development, to create new competitive advantage, opening up the development of new space, and effectively protecting national security.

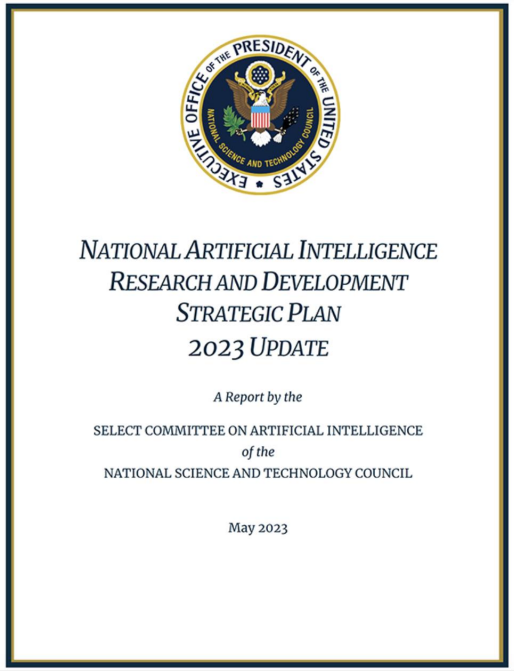
AI has become a new engine of economic development. AI has become the core driving force for a new round of industrial transformation, [which] will advance the release of the

全球国家AI战略

USA The National Artificial Intelligence R&D Strategic Plan

In February 2019, the United States launched the American AI Initiative, in the form of an executive order. This “whole-of-government strategy” aims at focusing federal government resources for investing in AI research, unleashing AI resources, setting AI governance standards, building the AI workforce and protecting the US AI advantage.

This plan (updated 2023) defines the major research challenges in AI to coordinate and focus federal R&D investments. It will ensure continued U.S. leadership in the development and use of trustworthy AI systems, prepare the current and future U.S. workforce for the integration of AI systems across all sectors, and coordinate ongoing AI activities across all federal agencies.

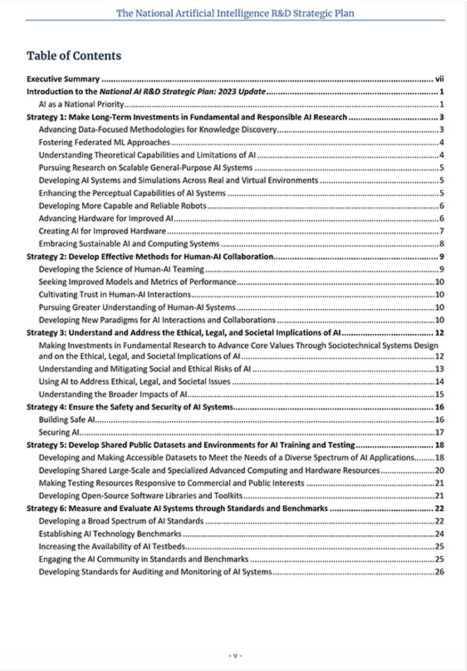


NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN 2023 UPDATE

A Report by the

SELECT COMMITTEE ON ARTIFICIAL INTELLIGENCE
of the
NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

May 2023



The National Artificial Intelligence R&D Strategic Plan

Table of Contents

全球国家AI战略

European Union Coordinated Plan on Artificial Intelligence

The 2021 Coordinated Plan on Artificial Intelligence, published in April of that year, builds on the collaboration established between the Commission and Member States during the 2018 Coordinated Plan.

It sets out the strategy to:

accelerate investments in AI technologies to drive resilient economic and social recovery aided by the uptake of new digital solutions;act on AI strategies and programmes by fully and timely implementing them to ensure that the EU fully benefits from first-mover adopter advantages;align AI policy to remove fragmentation and address global challenges.

In line with the above, the 2021 review of the Coordinated Plan provides ***an overview of actions taken*** since the adoption of the 2018 Coordinated Plan and sets out ***an outlook*** with concrete proposals and recommendations for further action, identifying areas where the partnership between the EU and the Member States is particularly effective in making Europe a hub for the development and use of cutting-edge, human-centric AI. The 2021 review aims to advance the objectives above and proposes 14 interrelated, joint action areas for collaboration between the European Commission and the Member States (seven horizontal and seven sectoral areas)[13]. As in the EU 2020 White Paper and the 2018 Coordinated Plan, the 2021 review of the Coordinated Plan does not address the development and use of AI for military purposes.

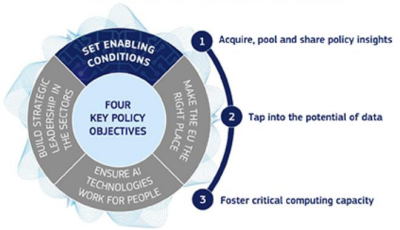
I. SET ENABLING CONDITIONS FOR AI DEVELOPMENT AND UPTAKE IN THE EU

In order to support the development and take-up of AI and to achieve the objectives of this Coordinated Plan, a number of enabling conditions are necessary. The first one is an appropriate **governance and coordination framework.** An efficient and functioning governance and coordination framework can help to build economies of scale, minimise information and transaction costs and facilitate synergies among Member States. The second enabling condition is **data.** The development of AI technologies often requires large, high-quality, secure and robust datasets. It is therefore important to ensure that data can 'flow' within the EU, with our trading partners and across sectors, in line with the EU acquis, including the General Data Protection Regulation for personal data and the Union's international commitments. Third is a **computation infrastructure.** This infrastructure is necessary for storing, analysing and processing the increasingly large volumes of data. In turn, this requires new developments and approaches to increase computing capabilities, e.g. through semiconductors that enable AI algorithms to store, run, and test data. Together, these three factors create broad enabling conditions for AI technologies to succeed in the EU.

Accordingly, to set enabling conditions for AI development and take-up and enhance cooperation among Member States and among Member States and the European Commission, the review proposes to focus on three key actions: to build a governance framework to effectively acquire, accumulate and share policy insights on AI; to tap into the potential of data to unleash its full potential; to foster critical computation infrastructure to support capacity building and enhance the development of AI.

[13] All actions must comply fully with the EU rules on competition law and notably State aid.

OUR KEY PROPOSALS TO SET ENABLING CONDITIONS



1. Acquire, pool and share policy insights

Knowledge is key. Sharing knowledge and policy insights, and coordinating policy actions and investments in a rapidly developing area such as AI can add an important competitive advantage. For this reason, in the 2018 Coordinated Plan the Member States and the Commission agreed on a governance mechanism for joint work and proposed two sets of actions to build policy insights and develop synergies. Member States were encouraged to put into place **national AI strategies** or programmes (or add an AI dimension to other relevant national strategies and programmes) and share these with each other and the Commission[14]; and the Commission pledged to **monitor developments and mobilise expertise.**

1.1. *Maximise the advantages from national strategies and accelerate the proposed actions*

Overview of actions taken

All Member States made substantial efforts to develop national strategies on AI or to include an AI dimension in their existing strategies and programmes[15]. The adoption of national strategies facilitated structured reflection on the priorities and objectives for the development and uptake of AI, and triggered wider public debate in many Member States. The exchange of information on the national strategies also fed into a structured dialogue between the Member States and the Commission.

As the analysis of national strategies indicates, the adoption of national strategies was an important first step to facilitate and streamline European effort on AI. This process helped to identify the priority sectors for joint actions, provided a solid mapping of the main investment priorities planned by the Member States and indicated possible steps forward for common multi-country projects and joint activities.

Outlook

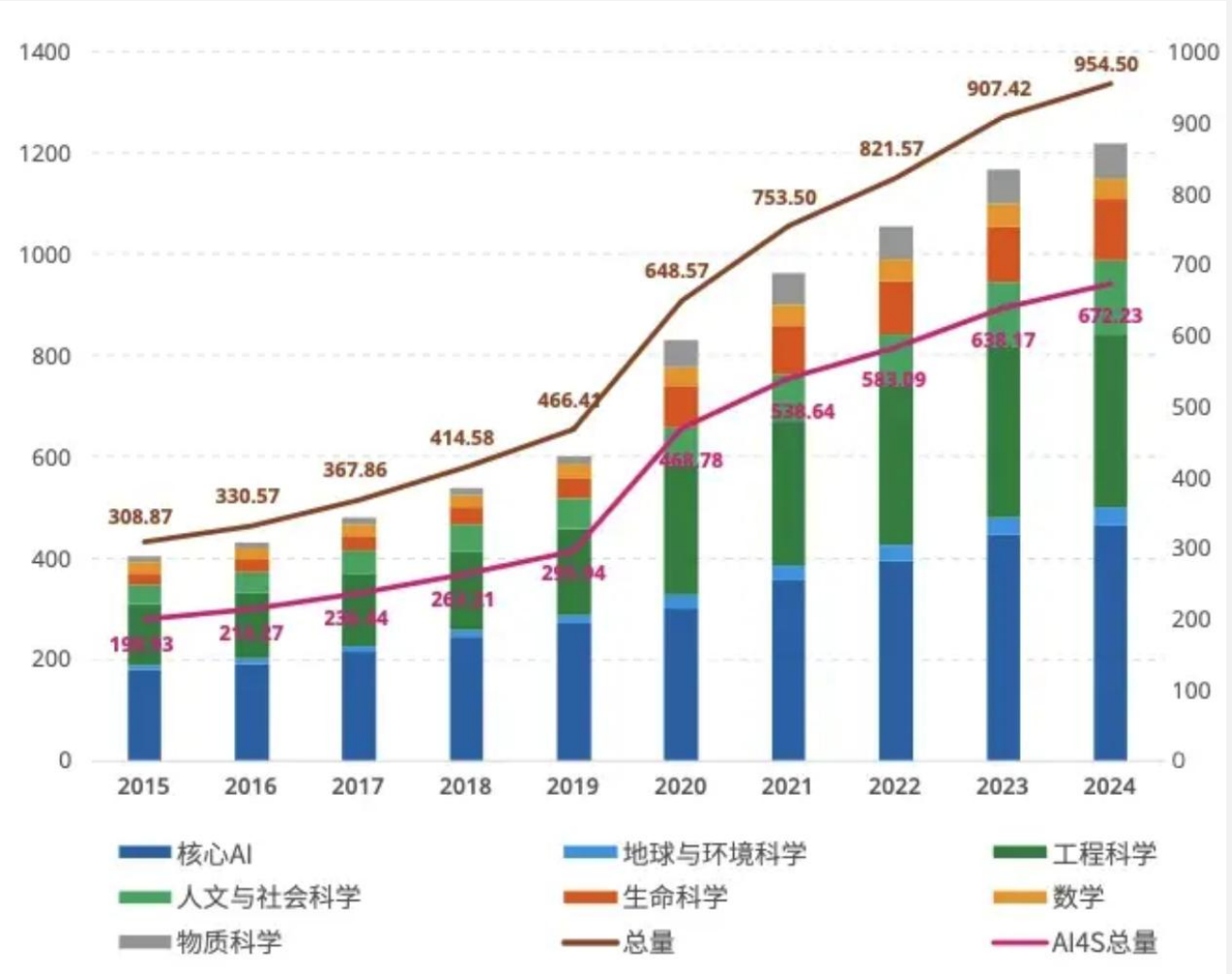
The next step is to ensure that the efforts invested by Member States in developing the

[14] This encouragement was also included in the February 2019 Council Conclusions; (Council of the European Union, *Conclusions on the coordinated plan on artificial intelligence*, 6177/19, 11 February 2019).

[15] See Appendix 1 to this document and JRC's forthcoming AI Watch report on AI national strategies (2021).

AI研究趋势

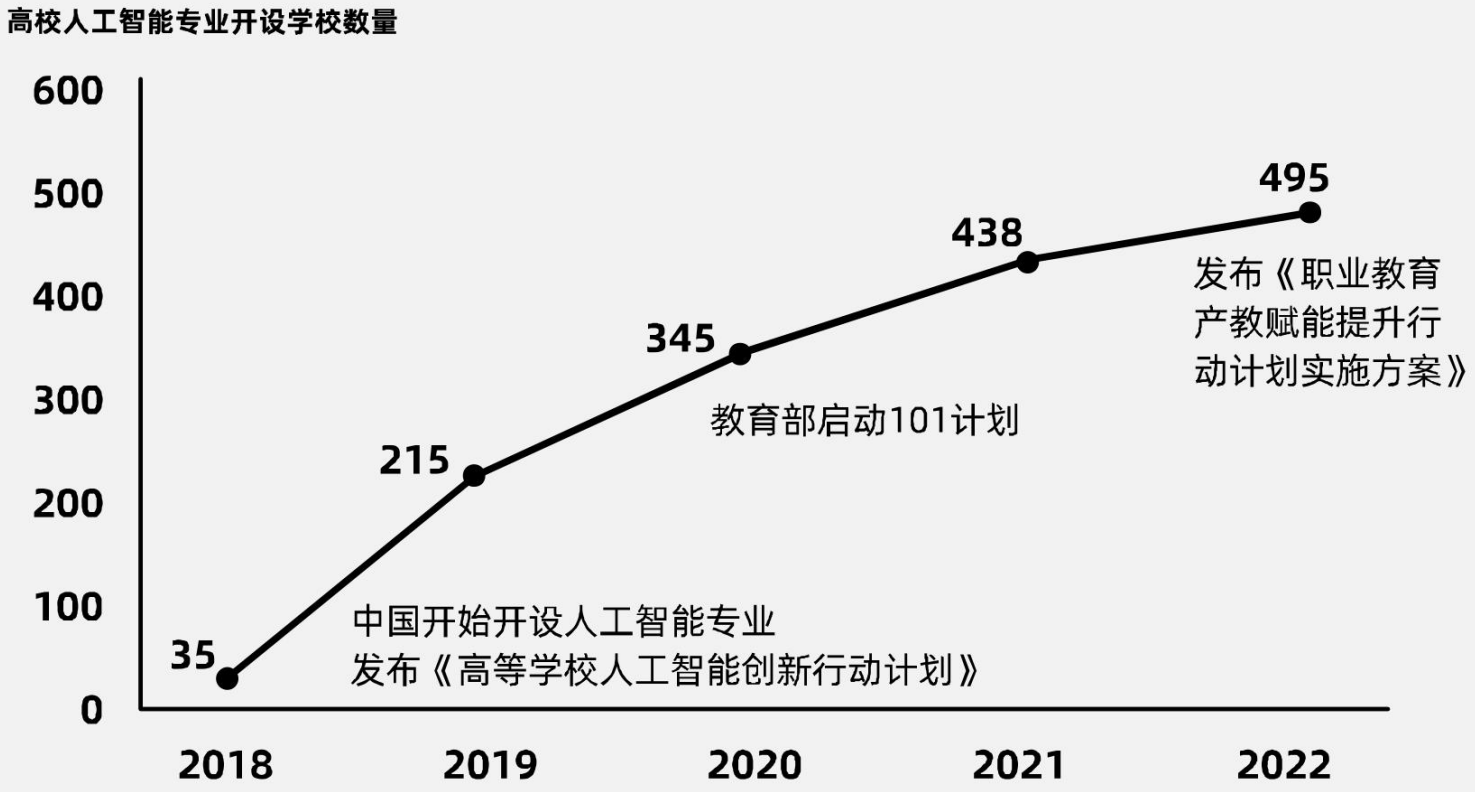
AI相关论文指数增长



AI出版物总量趋势与领域构成，来源：《科学智能白皮书》

AI 人才培养正在加速

中国人工智能专业建设情况



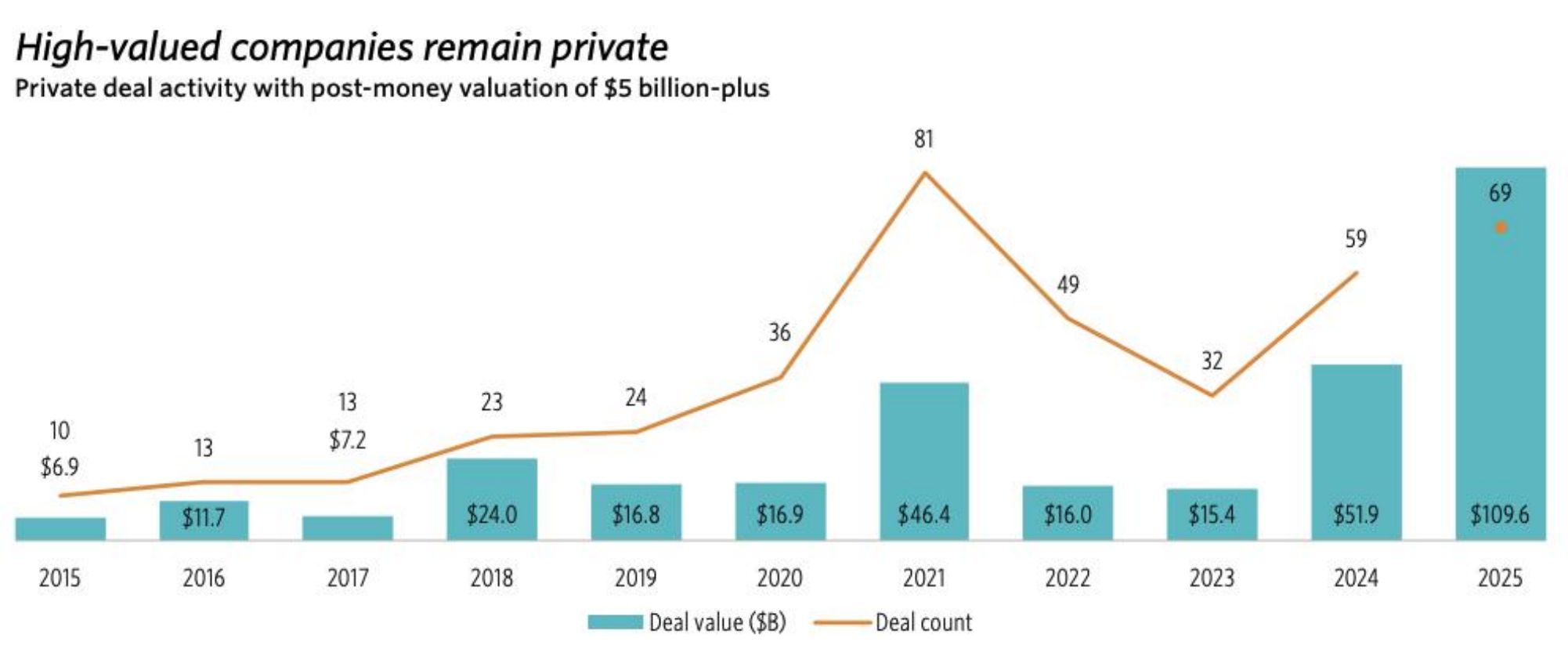
《加快数字人才培育支撑数字经济发展行动方案（2024－2026年）》

AI企业 （2025年战略）

Google	走“Gemini + 搜索/Workspace + TPU/云”一体化路线，强调模型能力与自研基础设施协同，把 AI 变成产品层的核心分发渠道
Microsoft	以“Copilot + Azure AI”做企业级入口，用云端算力与安全合规把生成式 AI 变成 Office/开发/业务流程的默认工作方式
amazon	用 AWS 提供“模型超市（Bedrock）+ 自研芯片（Trainium/Inferentia）+ 行业解决方案”，主打企业客户的成本、可控性与规模化部署
Meta	用“开源模型 + 大规模推理/推荐系统”驱动生态与应用分发，重点在把 AI 能力嵌入社交内容生产、推荐与广告效率
Alibaba	以“云 + 大模型”双轮驱动，把通义系模型与云基础设施/行业交付绑定，主攻企业智能化（客服、办公、研发、供应链）
Tencent	围绕“微信生态 + 企业服务”做 AI 产品化，把生成式 AI 融入内容生产、游戏与企业协作，同时强化云端推理与工具链
ByteDance 字节跳动	以“内容理解/生成 + 分发闭环”为核心，把大模型能力嵌入短视频/广告/创作工具，通过高频应用场景迭代模型与系统

AI新创公司

2025 年 AI 创业呈现“高集中度 + 大额融资”特征

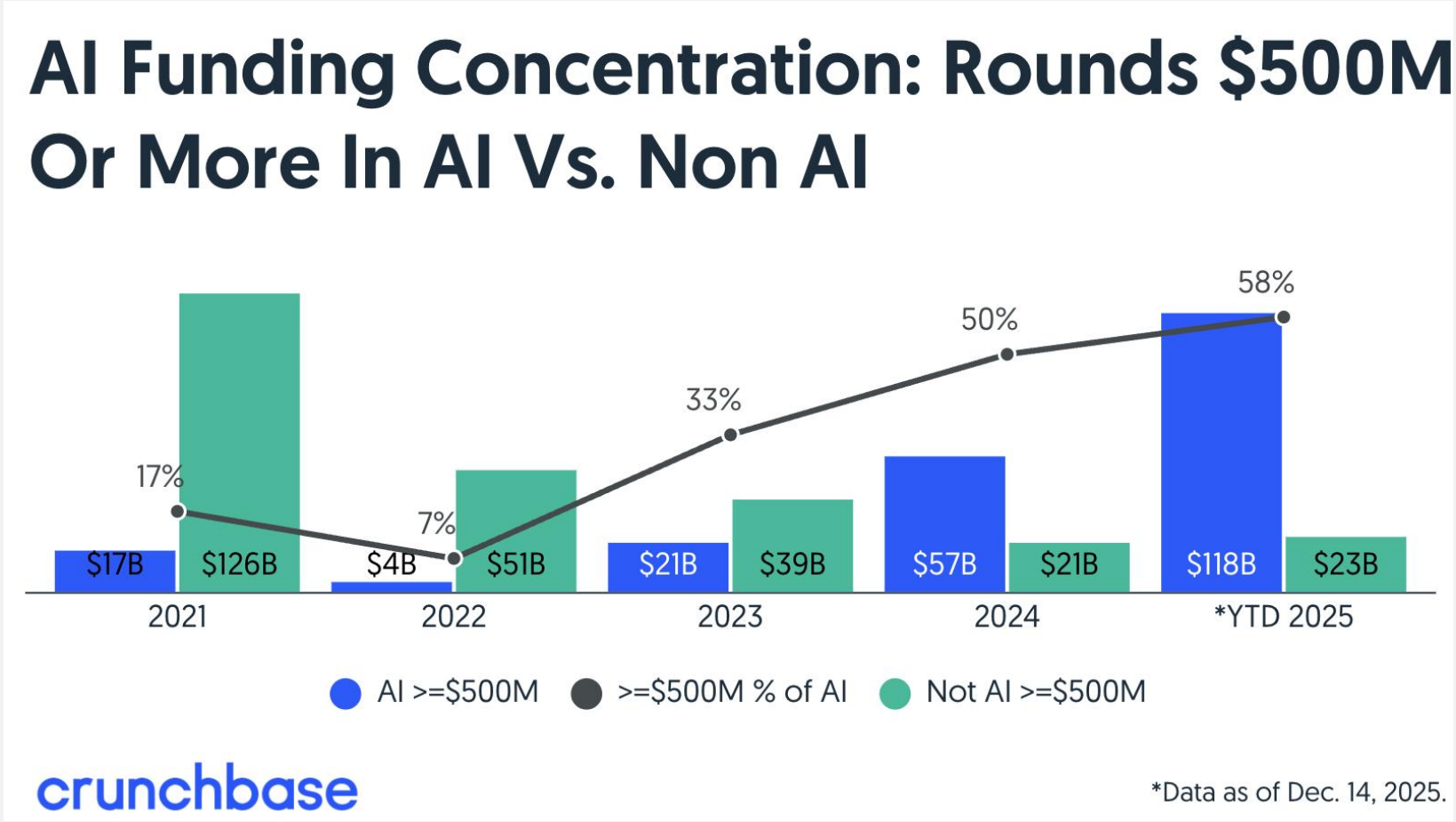


AI&ML VC deal value / deal count（2015–2025）

https://nvca.org/wp-content/uploads/2025/10/Q3-2025-PitchBook-NVCA-Venture-Monitor.pdf

AI投资规模

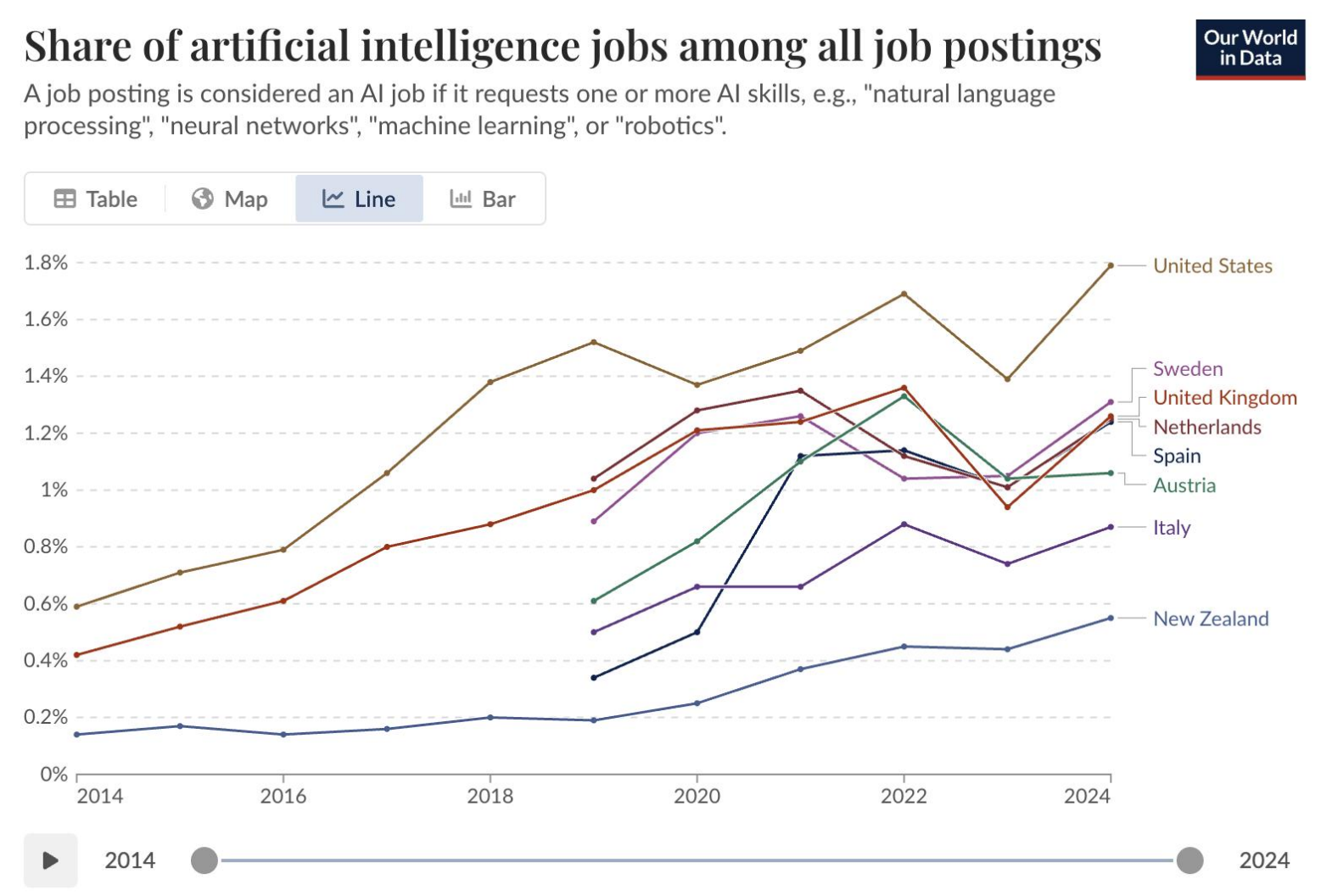
AI 已成为 2025 年全球风险投资的“主赛道”，投资从应用层进一步向基础设施与基础模型集中



https://news.crunchbase.com/ai/big-funding-trends-charts-eoy-2025/?utm_source=chatgpt.com

AI就业机会

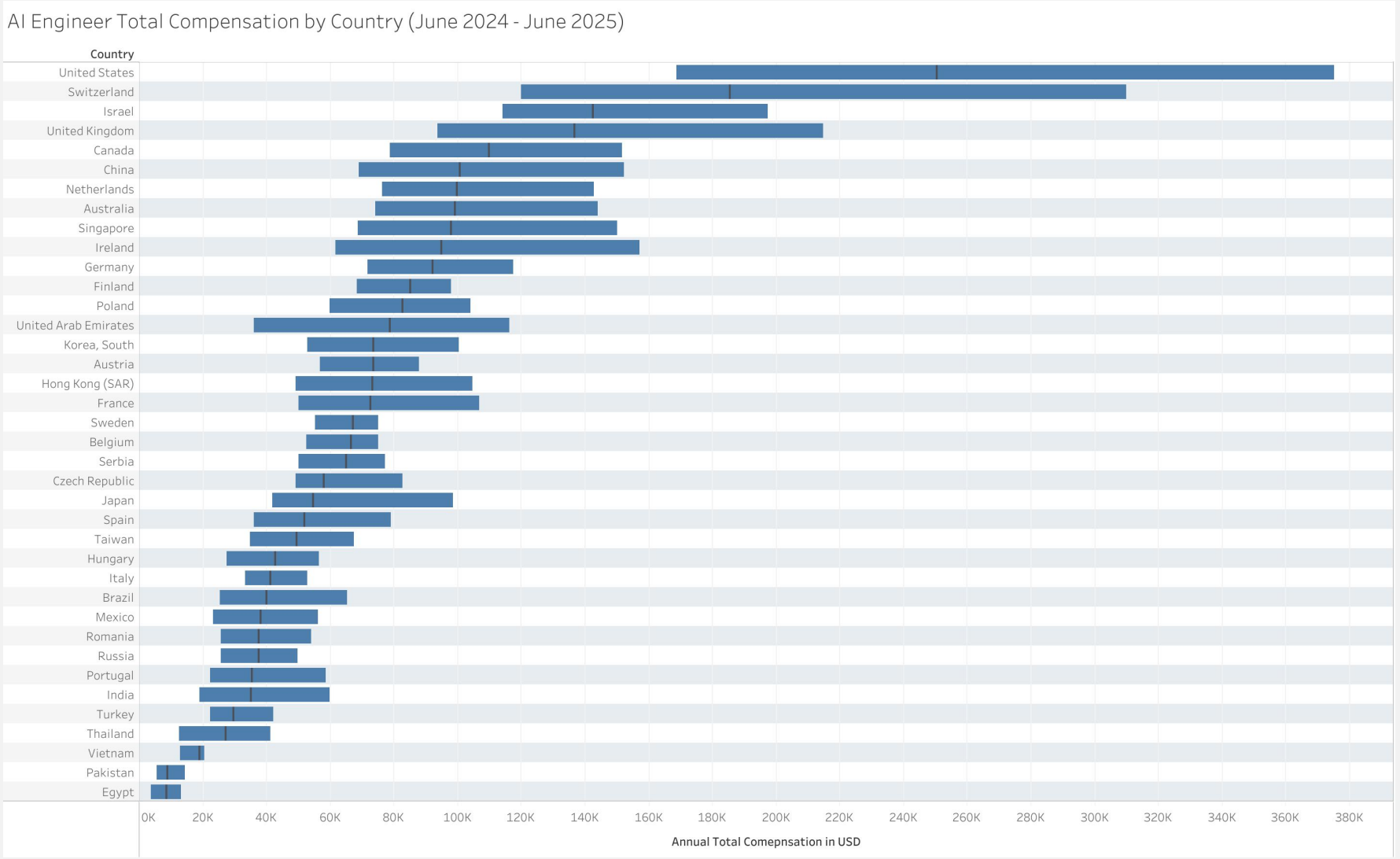
AI 正在重塑就业结构：企业对 AI 技能的需求持续上升，AI 相关岗位在整体招聘中的占比呈长期增长趋势



https://ourworldindata.org/grapher/share-artificial-intelligence-job-postings?tab=map

AI就业薪资

AI领域人才紧缺度增加，研发岗位平均薪资涨幅明显



https://www.levels.fyi/blog/ai-engineer-compensation-trends-q3-2025.html

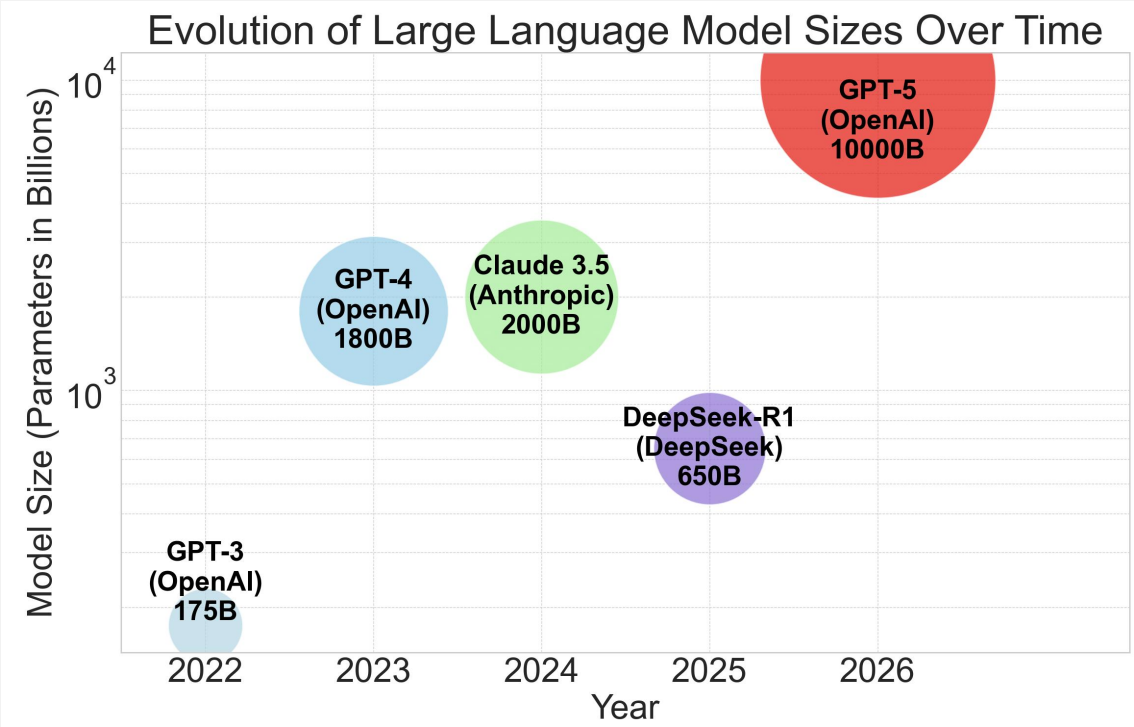
大模型参数规模

大模型规模高速增长

模型规模持续增长，或迎来十万亿模型

集群规模需千倍增长

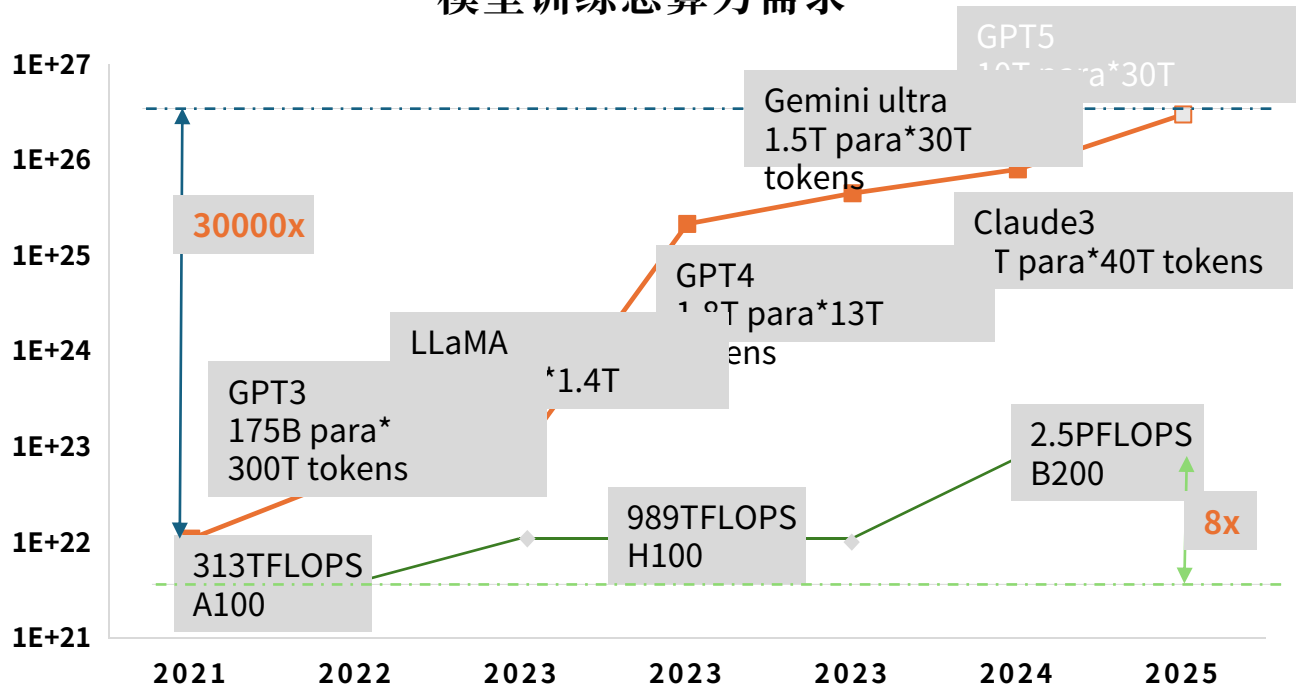
§ LLM的scaling law持续，模型规模走向十万亿



§ 模型训练总算力需求4年增长速度（4年3万倍）远快于单卡算力增长速度（4年8倍），需要更大规模的集群

总算力需求增长（万倍）= 单卡算力（10倍）× 集群规模（千倍）

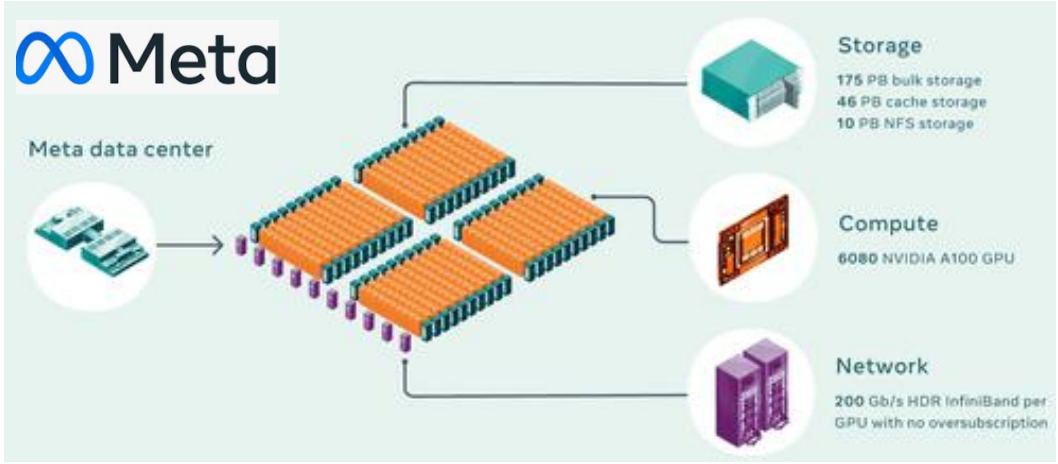
模型训练总算力需求



大模型算力要求

计算、网络资源高效使用在十万卡集群中成为核心竞争力

超10万卡集群不仅仅是愿景



2022.1
- 16000 A100
- 训练Llama，Llama2

2024.3
- 两个24576 H100集群
- 训练Llama3（16K）

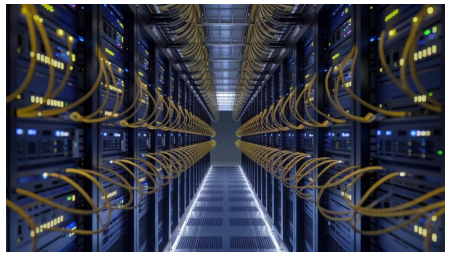
未来
- 继续扩大至350k H100
- 产品组合的计算能力相当于600k H100

- 马斯克7月23日宣布xAI（Twitter）的10万张H100集群上线
- 微软工程师透露，和openAI合作，计划建10万H100集群训练GPT6

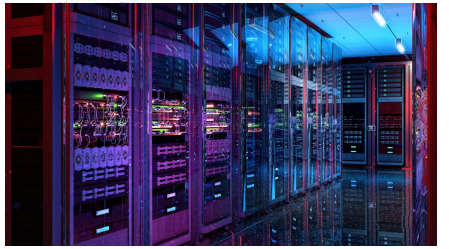
国内2030年10万卡集群是标配

- 当前国内多家厂商建设集群都超万卡
  - 字节万卡集群[MegaScale, NSDI24]
  - 阿里万卡集群[HPN, SIGCOMM24]

ByteDance



幻方量化 deepseek

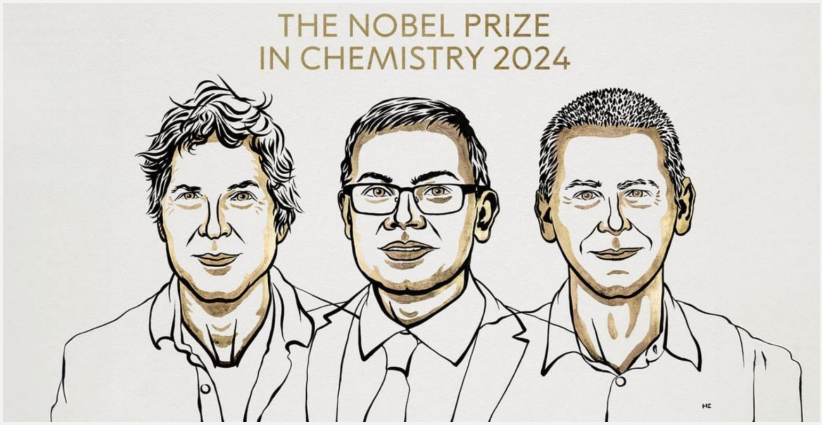


Alibaba Cloud

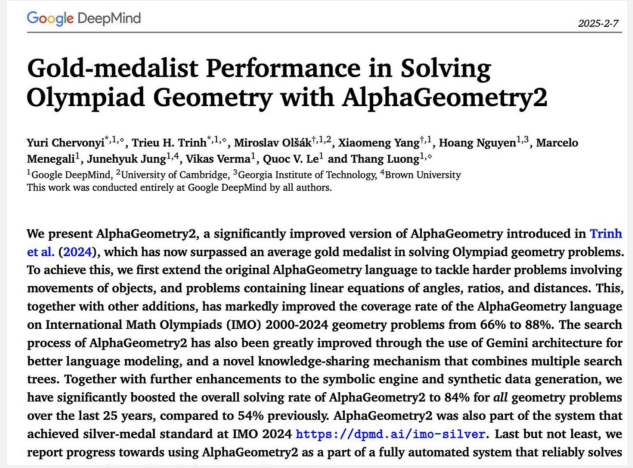
腾讯云

AI算法应用

Vibe coding



2024 年诺贝尔化学奖
蛋白质结构预测
（AlphaFold）



Google DeepMind 2025-2-7

Gold-medalist Performance in Solving Olympiad Geometry with AlphaGeometry2

Yuri Chervonyi[*,1,◦], Trieu H. Trinh[*,1,◦], Miroslav Olšák[†,1,2], Xiaomeng Yang[†,1], Hoang Nguyen[1,3], Marcelo Menegali[1], Junehyuk Jung[1,4], Vikas Verma[1], Quoc V. Le[1] and Thang Luong[1,◦]

[1]Google DeepMind, [2]University of Cambridge, [3]Georgia Institute of Technology, [4]Brown University
This work was conducted entirely at Google DeepMind by all authors.

We present AlphaGeometry2, a significantly improved version of AlphaGeometry introduced in Trinh et al. (2024), which has now surpassed an average gold medalist in solving Olympiad geometry problems. To achieve this, we first extend the original AlphaGeometry language to tackle harder problems involving movements of objects, and problems containing linear equations of angles, ratios, and distances. This, together with other additions, has markedly improved the coverage rate of the AlphaGeometry language on International Math Olympiads (IMO) 2000-2024 geometry problems from 66% to 88%. The search process of AlphaGeometry2 has also been greatly improved through the use of Gemini architecture for better language modeling, and a novel knowledge-sharing mechanism that combines multiple search trees. Together with further enhancements to the symbolic engine and synthetic data generation, we have significantly boosted the overall solving rate of AlphaGeometry2 to 84% for *all* geometry problems over the last 25 years, compared to 54% previously. AlphaGeometry2 was also part of the system that achieved silver-medal standard at IMO 2024 https://dpmd.ai/imo-silver. Last but not least, we report progress towards using AlphaGeometry2 as a part of a fully automated system that reliably solves

奥数级数学推理：AI 达到
IMO “金牌标准”

人工智能算法在多种应用上接近或超过了人类水平

百花齐放的AI应用

人工智能已实现多领域的深度渗透与场景革新

文本生成

写作

阿里妈妈 | 爱创作 | AI帮个忙 aibang.run | 彩云小梦 | 笔灵AI | 笔墨公文 | 光速写作 | 百度作家平台 | Pitaya | 火山写作 | lingo 灵构笔记 | 写作猫 | 深言达意 | 序列猴子 奇妙文 | 搜外内容管家 | 万彩AI | 万兴智演 | 悟智写作 | FlowUs AI | 小鱼AI写作 | 新华妙笔 | 万能小in 我的个人AI助手 | Effidit | WriteWise | FRIDAY | GiiSO 写作机器人 | Verse AI写作 | 易撰

阅读知识库

报告说 | 通义智文 | 氢刻 | 星火文档问答 SPARK DOCQA | 有道速读 | CUBOX | 知我AI

智能搜索

360AI搜索 | 秘塔AI搜索 | 天工

语音生成

配音/变音

讯飞智作 | 大饼AI变声 | 魔音工坊 | reecho.AI | 悦音配音 | 制片帮 ZHIPIANBANG.COM

音乐

BGM猫 | 天谱 | 腾讯音乐娱乐集团 TENCENT MUSIC ENTERTAINMENT | X Studio

音频后期

音剪

对话式助理

ChatBot

360智脑 | 百川大模型 BAICHUAN | Chat JD | Chato | 豆包 | 腾讯混元 | 灵沐AI | 天工 | 文心一言 | 通义千问 | 讯飞星火 | X 元象大模型 | 紫东太初 | BetterYeah | Kimi Chat | MOSS | 智谱清言 | 商汤 sensetime | SenseChat

角色聊天

通义星尘 | 网红秀 | 乌托邦 | gemsouls

生活助理

CHAT LAW | 海瑞智法 | 万木健康 | Aicy | 合同嗖嗖 | 灵医智惠 | 华为小艺 | 医真·AI+ | 张三 | 蓝心小V | 智能罐子 AI JAR | 左手医生 | Timtalk | SpeakG | 小爱同学 | ChatPaper | 松鼠AI 智适应教育 | 通义法睿 | 新小布 1.0

视频生成

视频生成

度加创作工具 | 开拍 | 奇妙元 | 深氧AI | 腾讯智影 | 万彩微影 | 万兴播爆 | 万兴喵影 | 一起剪 | 一帧秒创

视频后期

AlibabaWOOD beta | WinkStudio | 绘影字幕 | 快剪辑 · SaaS | 剪映 | 闪剪 | 新简

图片生成

文生图

360智绘 | 触手AI | 秒画 | 腾讯混元助手 | 文心一格 | 通义万相 | 言之画 | 无限画 | Dreamina | MiracleVision 奇想智能 | 智谱·AI

商业设计

爱设计 iSHEJI.COM | 标智客 | 创客贴 | Canva可画 | 堆友相机 | 稿定AI | 即时AI | redoon.ai 灵动AI | 鹿班 | 羚珑 | 美间 | 模袋云AI | 美图设计室 | 墨刀 | 木疑 | 腾讯云智绘 | 莫高设计 MasterGo | Motiff | AIDesign | PIC COPILOT | Pixso

图片后期

改图 | 悟空图像 Photosir | 美图云修 | 佐糖 | 360智图 | 像素蛋糕 | 咻图AI

生产力工具

思维导图

Chatmind | GitMind | ProcessOn | TreeMind 树图 新一代思维导图 | 小画桌 | edraw亿图 | 印象图记 | 知犀AI

会议助理

飞书妙记 | 麦耳会记 | 腾讯会议 | 通义听悟 | 讯飞听见 | 会议

编程助手

AIXcoder | 通义灵码 | 代码小浣熊 | CHATDEV | CodeArts Snap | CODEFUSE | CodeGeeX | DEC3 | iFlyCode | SKYCODE

PPT

AiPPT.cn | CHATPPT | iSlide | MINDSHOW 上海所思所见科技有限公司 | 比格PPT | 歌者PPT | 美图设计室 AI PPT | 轻竹办公 | PLAY | 讯飞智文

办公助手

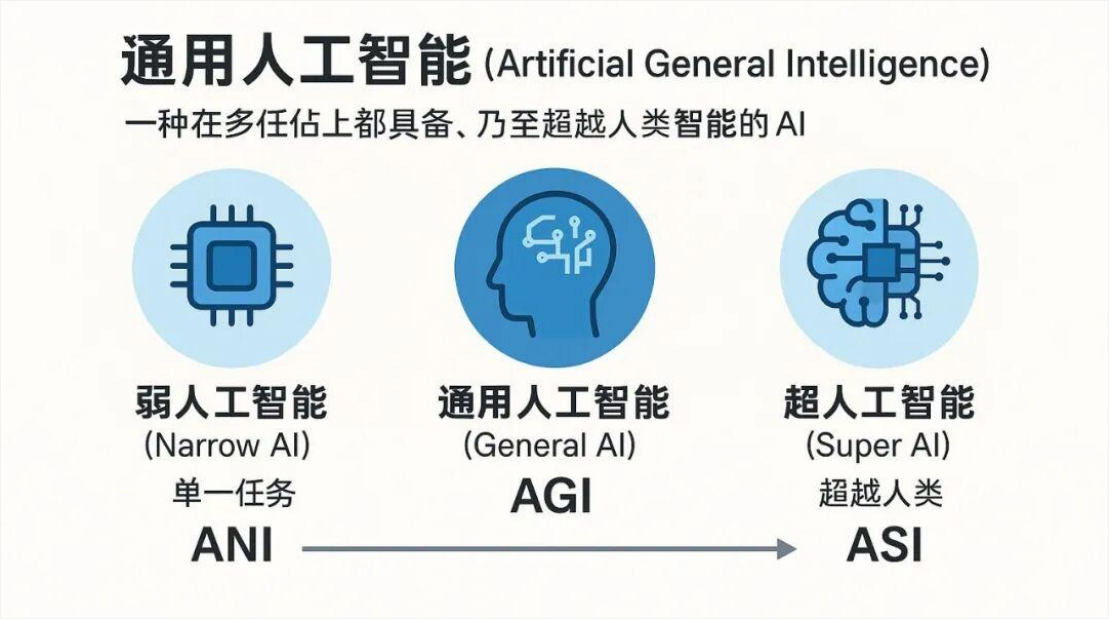
My AI | WPS AI | 办公小浣熊 | 钉钉智能助手

什么是人工智能?

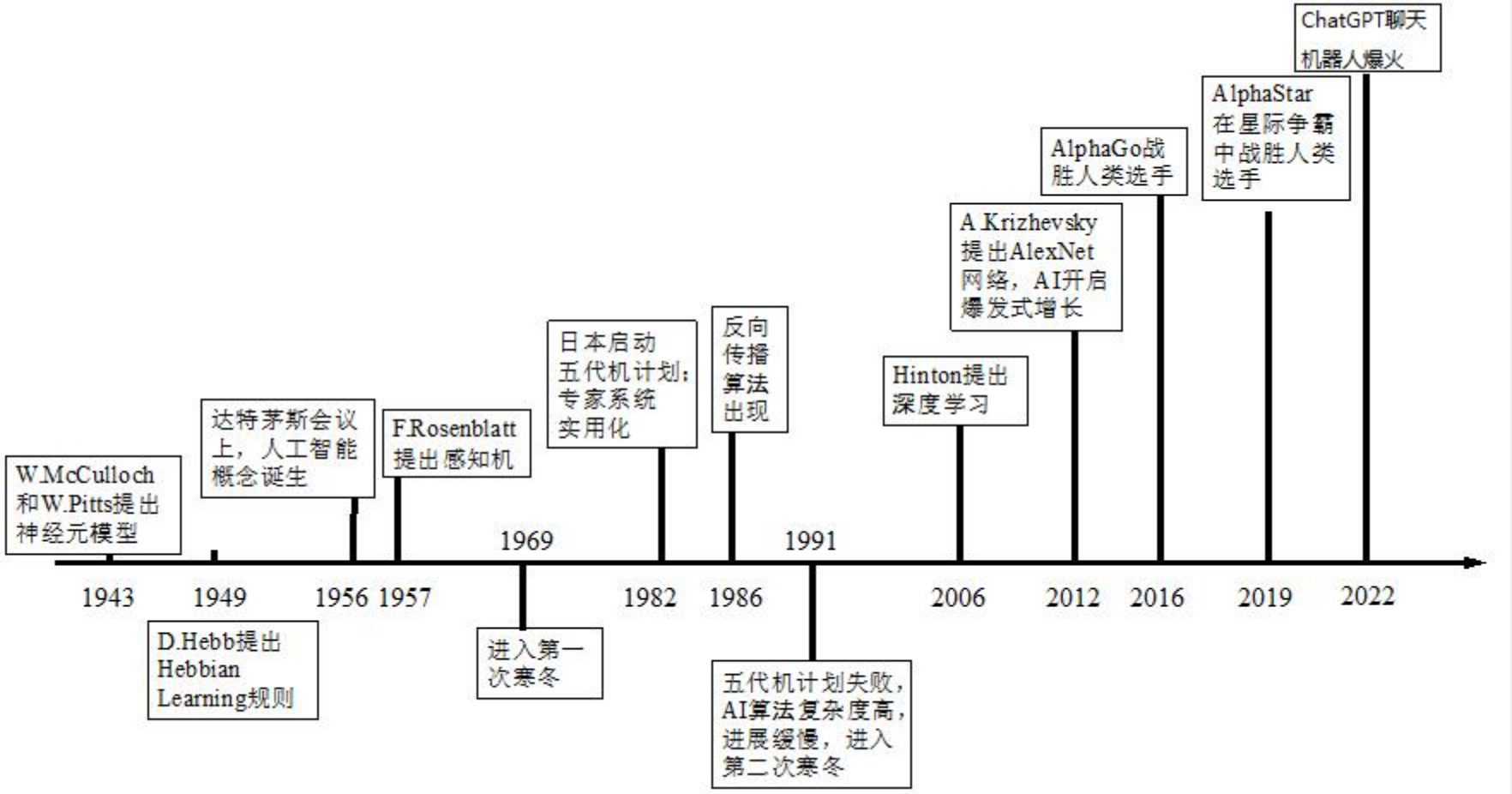
- **人工智能**：人制造出来的机器所表现出来的智能，能够感知信息、理解与推理，并做出决策与行动。

- **强人工智能**或**通用人工智能**：具备与人类相当或超越人类的通用认知能力，能在不同领域中完成广泛任务，并表现出人类通常具备的多种智能行为。

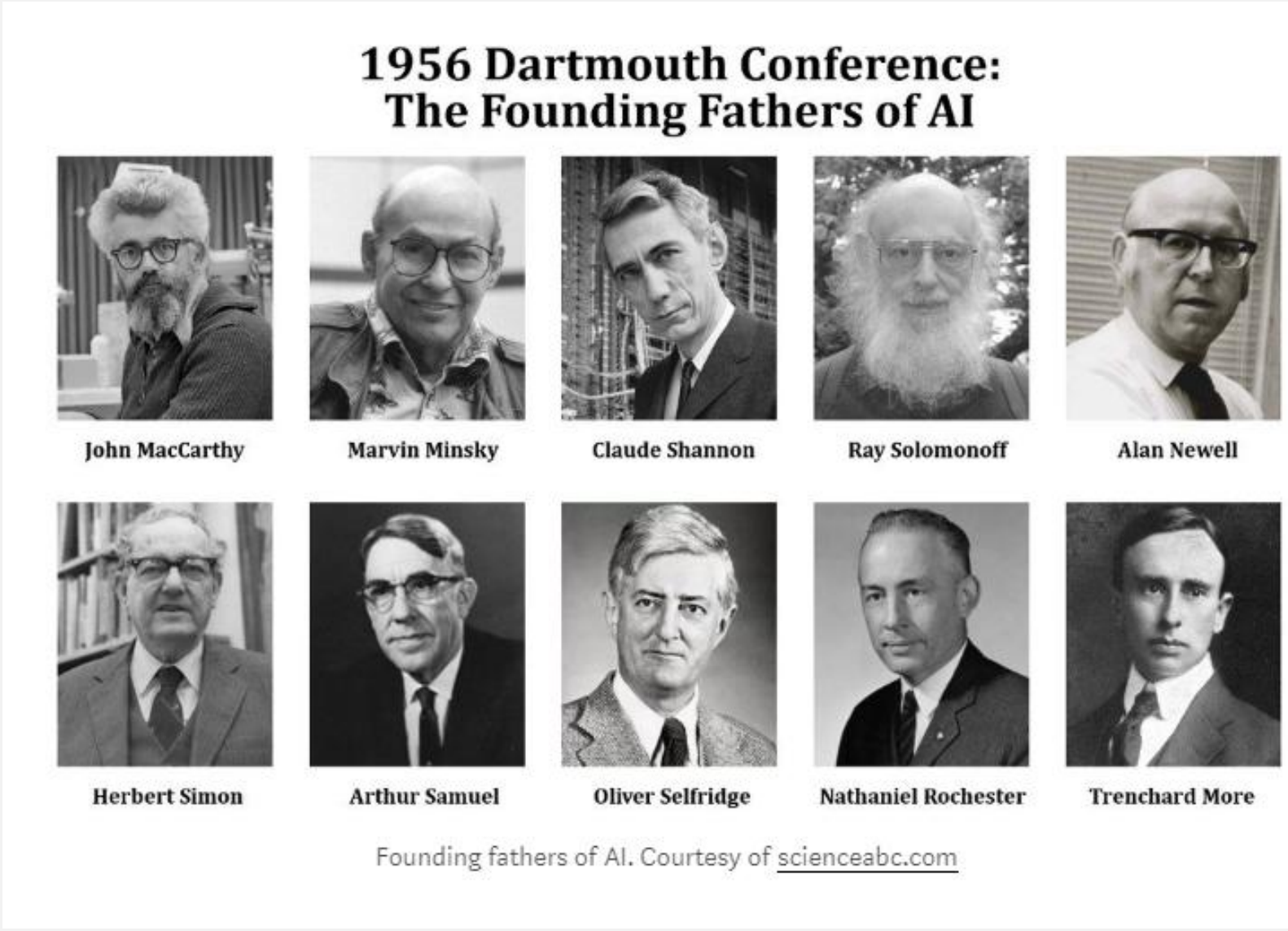
- **弱人工智能**：面向特定任务的人工智能系统，在某一类问题上达到很高性能。



人工智能的三次热潮

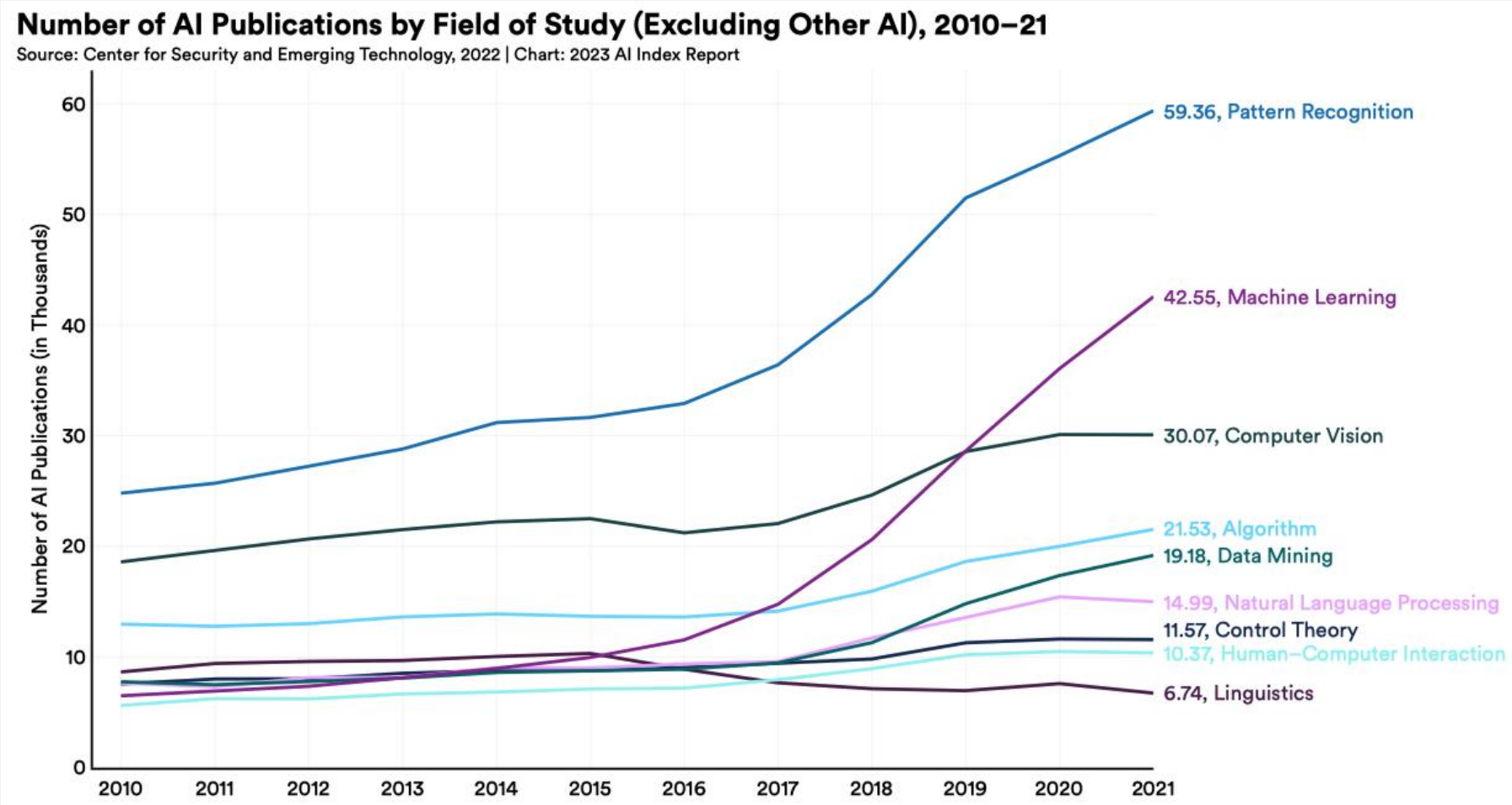

ChatGPT聊天机器人爆火
AlphaStar在星际争霸中战胜人类选手
AlphaGo战胜人类选手
A.Krizhevsky提出AlexNet网络，AI开启爆发式增长
Hinton提出深度学习
反向传播算法出现
日本启动五代机计划；专家系统实用化
F.Rosenblatt提出感知机
达特茅斯会议上，人工智能概念诞生
W.McCulloch和W.Pitts提出神经元模型
D.Hebb提出Hebbian Learning规则
进入第一次寒冬
五代机计划失败，AI算法复杂度高，进展缓慢，进入第二次寒冬
1943
1949
1956
1957
1969
1982
1986
1991
2006
2012
2016
2019
2022

1956年达特茅斯人工智能研讨会



Founding fathers of AI. Courtesy of scienceabc.com

人工智能都在研究什么?


Number of AI Publications by Field of Study (Excluding Other AI), 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report
Number of AI Publications (in Thousands)
0
10
20
30
40
50
60
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
59.36, Pattern Recognition
42.55, Machine Learning
30.07, Computer Vision
21.53, Algorithm
19.18, Data Mining
14.99, Natural Language Processing
11.57, Control Theory
10.37, Human–Computer Interaction
6.74, Linguistics

人工智能的三个流派

- **行为主义**：基于控制论，构建感知-动作型控制系
  - 闭环控制与实时决策
  - **Norbert Wiener**：控制论（Cybernetics）奠基
  - 反馈控制、PID/状态反馈

- **符号主义**：基于符号逻辑的方法，用逻辑表示知识和求解问题。
  - 可解释的结构化推理
  - John McCarthy：Lisp、形式化 AI 思想
  - 搜索：A*、启发式搜索、博弈树搜索

- **连接主义**：基于大脑中神经元细胞连接的计算模型，用人工神经网络来拟合智能行为。
  - 从数据中学习表示
  - Yann LeCun / Geoffrey Hinton / Yoshua Bengio：深度学习三巨头（卷积、表示学习、深度网络等）
  - MLP/反向传播、CNN、RNN/LSTM

符号主义的困难：逻辑，常识，求解器

- 逻辑：未找到能表述世间所有知识的简洁逻辑体系
- 常识：无穷无尽的常识
- 求解器：命题逻辑判定NP完全，一阶谓词逻辑不可判定

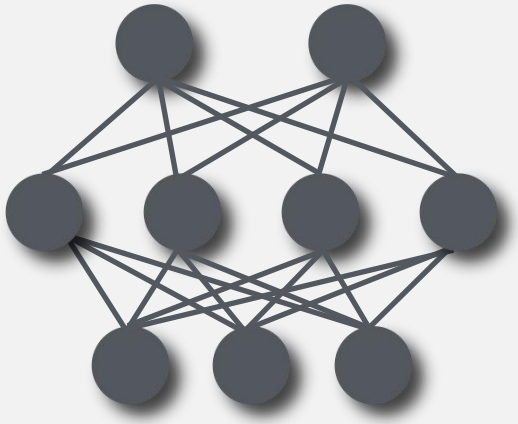
更本质的问题

- 人的智能主要是符号智能吗?

小丽、小玲、小娟三个人一起去商场里买东西。她们都买了各自需要的东西，有帽子，发夹，裙子，手套等，而且每个人买的东西还不同。有一个人问她们三个都买了什么，小丽说："小玲买的不是手套，小娟买的不是发夹。"小玲说："小丽买的不是发夹，小娟买的不是裙子。"小娟说："小丽买的不是帽子，小娟买的是裙子。"
她们三个人，每个人说的话都是有一半是真的，一半是假的。那么，她们分别买了什么东西？

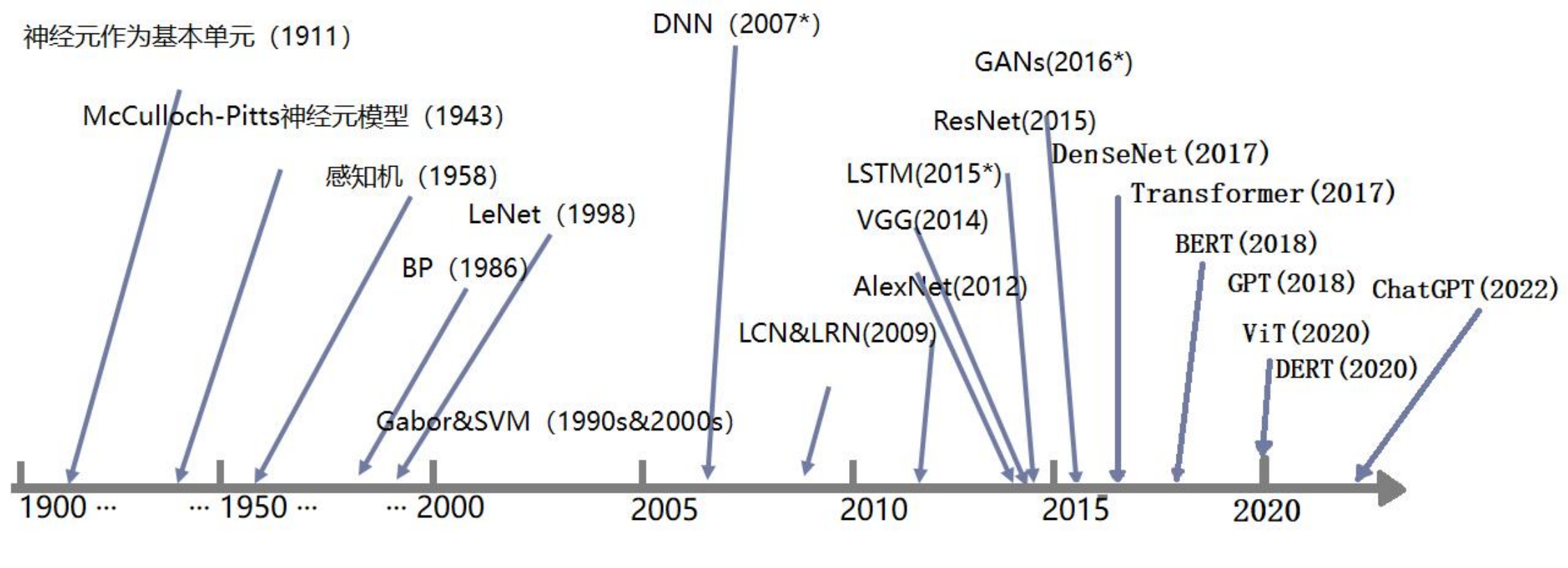
- 符号主义最本质的问题是只考虑了理性认识的智能。人类的智能包括感性认识（感知）和理性认识（认知）两个方面
  - 人类语言的例子：词汇，时态，格，数字

连接主义：人工神经网络



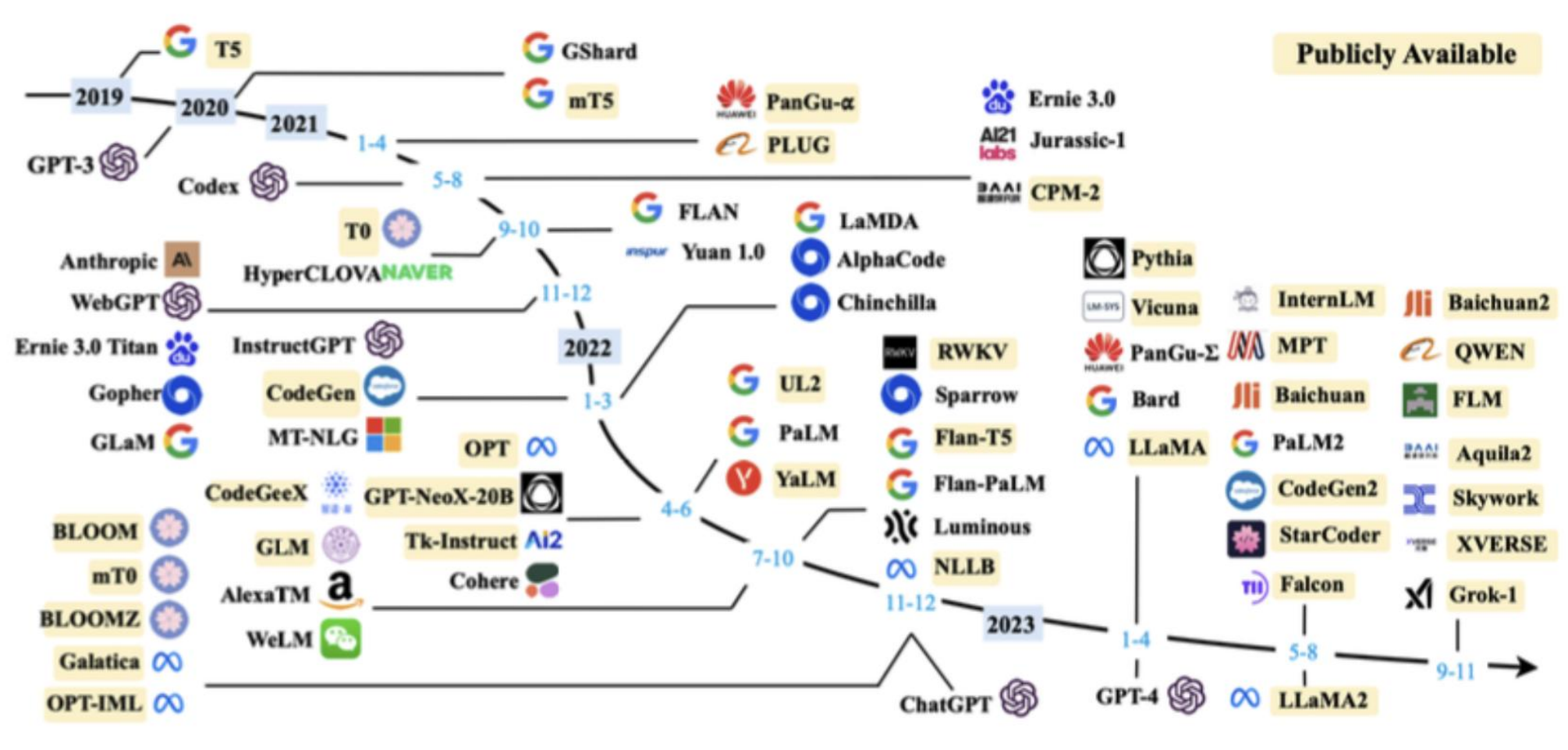
- 1943， 神经元模型， McCulloch 和 Pitts，第一波神经网络
- 1949，《The Organization of Behaviour》，Hebb学习
- 1958，感知机模型（preceptron），Rosenblatt
- 1986，BP反向传播训练方法， Rumelhart、Hinton 和 Williams，第二波神经网络
- 1998，卷积神经网络， Lecun
- 2000，自然语言模型， Bengio
- 2006，深度置信网络（ DBN ），Hinton, 第三波神经网络
- 2012，AlexNet （Dropout），Hinton团队赢得ImageNet比赛ILSVRC的冠军
- 2015，Deep Residual Network，AlphaGo
- 2016，生成对抗网络，GAN
- 2017, Transformer自然语言处理
- 2018，BERT, GPT自然语言处理
- 2020，ViT, GPT图像处理
- 2022，ChatGPT聊天机器人
- 2023，Llama 2 开源发布（开放权重大模型生态成型）
- 2024，GPT-4o（实时多模态：文本+图像+音频）
- 2025，DeepSeek-R1

人工神经网络发展线路



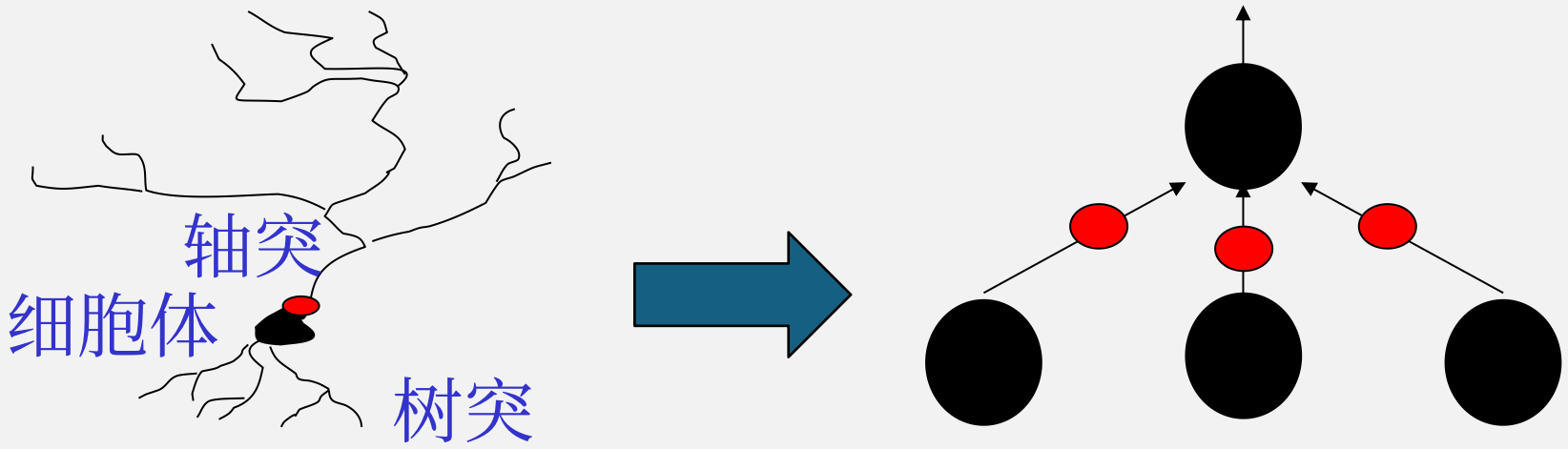
*开始流行时间

人工神经网络发展线路

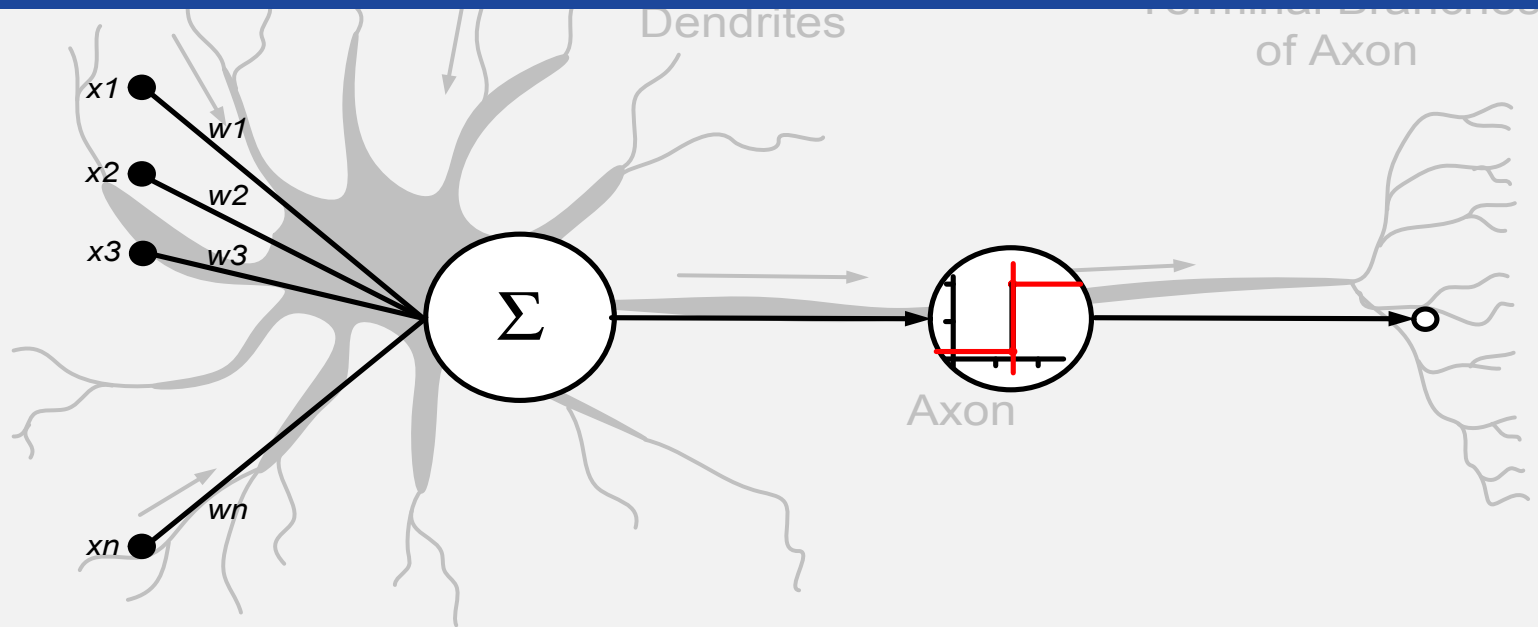


*开始流行时间，2022年后大模型时代

人工神经元

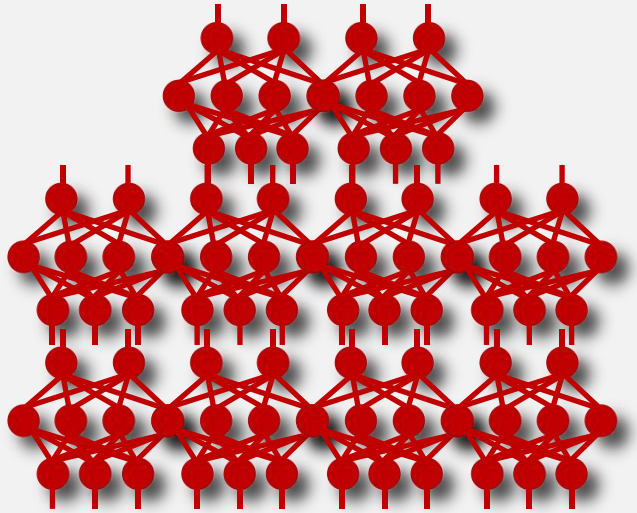


生物神经元：人工神经元=老鼠：米老鼠

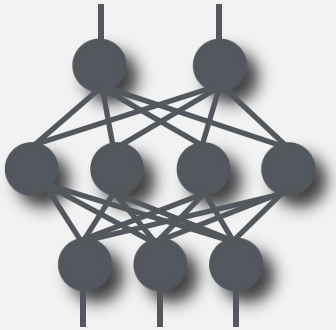


多层+多个神经元

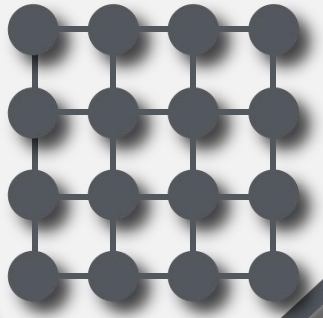
第三代，深度神经网络



第二代，MLP

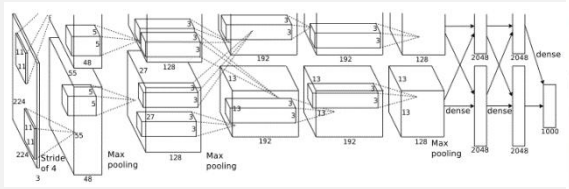


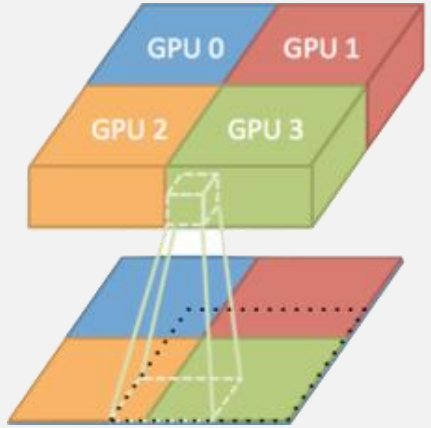
SVMs



如今

深而大的深度神经网络





6千万参数　　十亿参数　　110亿参数　　1750亿

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1–9).

Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., ... Ng, A. Y. (2012). Building High-level Features Using Large Scale Unsupervised Learning. In International Conference on Machine Learning.

Coates, A., Huval, B., Wang, T., Wu, D. J., & Ng, A. Y. (2013). Deep learning with cots hpc systems. In International Conference on Machine Learning.

Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

本章节内容

- 为什么学院要开这门课
- 为什么同学来上这门课
- 人工智能发展背景
- 机器学习系统概述

什么是机器学习系统?

机器学习系统是智能的物质载体

它把算法、数据与算力组织起来，使模型能够在真实世界中可训练、可推理、可部署、可维护

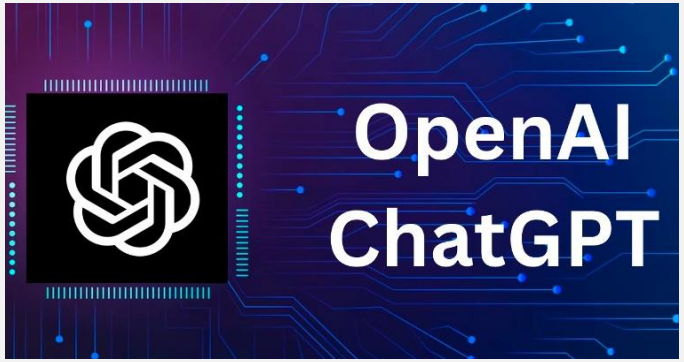
现阶段的机器学习系统通常是集成CPU和加速器的异构系统，软件上通常包括一套面向开发者的智能计算编程环境（包括编程框架和编程语言）

机器学习系统解决的不是 模型能不能算，而是 模型如何在约束下高效、稳定、可规模化地运行?

人工智能的物质载体

1.6万个CPU核学一周识别猫脸的谷歌大脑

和李世石下一盘围棋电费数千美元的AlphaGo



1万个A100训练1个月的ChatGPT

人工智能必须有其核心物质载体

异构机器学习计算系统

- 为什么采用异构机器学习计算系统?

  近十年来通用 CPU 的计算能力增长近乎停滞，而机器学习计算能力的需求在不断以指数增长，二者形成了剪刀差

  - 例如，寒武纪深度学习处理器能够以比通用 CPU 低一个数量级的能耗，达到 100 倍以上的处理的速度

- 异构系统在提高性能的同时，也带来了编程上的困难
  - 一般会集成一套编程环境，方便程序员快速便捷地开发高能效的智能应用程序
  - 深度学习编程框架包括 PyTorch、TensorFlow 等
  - 深度学习编程语言包括 CUDA 语言和 BCL 语言等

三代智能计算系统

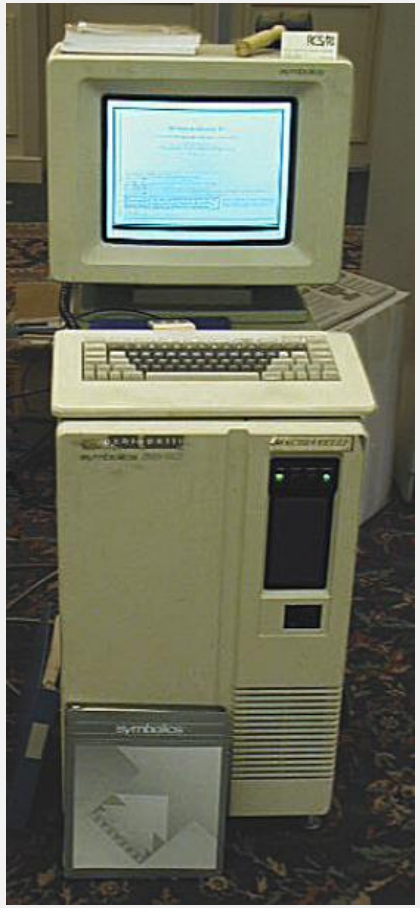
- 第一代智能计算系统：1980年代，面向符号主义智能处理的专用计算机（Prolog机，LISP机）
- 第二代智能计算系统：2010年代，面向连接主义智能处理的专用计算机（深度学习计算机）
- 第三代智能计算系统：未来强人工智能/通用人工智能的载体

第一代智能计算系统

- 1975，MIT AI Lab的Greenblatt研制成功LISP机CONS
- 1978，MIT AI Lab发布CONS的后继，CADR
- 1980s，发展高峰
  - Symbolics (3600, 3640, XL1200, MacIvory）
  - Lisp Machines Incorporated (LMI Lambda)
  - Texas Instruments (Explorer and MicroExplorer）
  - Xerox (Interlisp-D workstations)
  - 日本，五代机
  - Prolog机，1983，David H. D. Warren Warren Abstract Machine
- 1980s末到1990s初，AI winter，第一代智能机市场坍塌

第一代智能计算系统

LISP机（MIT博物馆）



Symbolics 3640

第一代智能计算系统

- High-level language computer architecture
  - OS的编程语言和硬件“统一”化，如LISP
  - 只针对特定语言的优化
- 局限性
  - 没有太多的实际应用需求
  - 由于摩尔定律发展，性能比不上CPU
    - 贵，几十万美元一台

第二代智能计算系统

- 面向连接主义（深度学习）处理的计算机或处理器

- 第二代智能计算系统的优势
  - 深度学习有大量实际的工业应用，已经形成了产业体系，因此相关研究能得到政府和企业的长期资助
  - 摩尔定律在 21 世纪发展放缓，通用 CPU 性能增长停滞，专用智能计算系统的性能优势越来越大

第二代智能计算系统


物端设备


移动设备


客户端


汽车


服务器


超级计算机

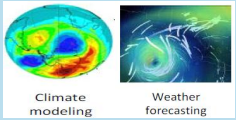

图像识别


语音识别


游戏竞技


自动驾驶

广告推荐

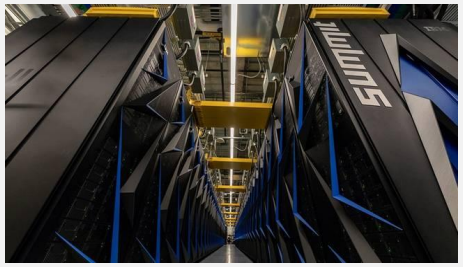


气象预报

TensorFlow caffe dmlc mxnet DRAGON ONNX PyTorch

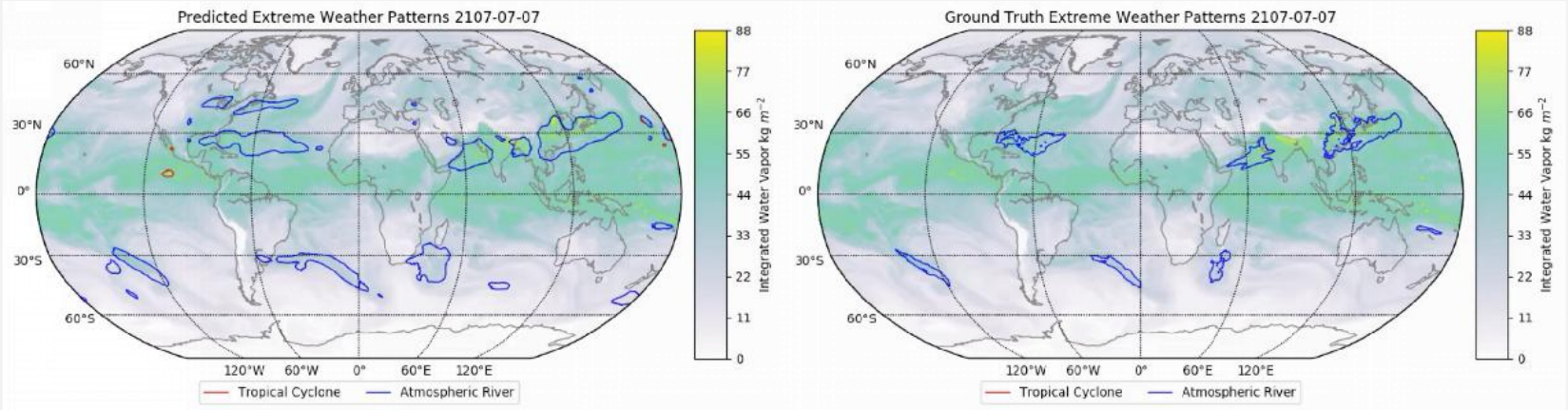
智能计算系统

第二代智能计算系统



美国智能计算系统代表“顶点”（Summit）
浮点运算速度峰值达每秒20亿亿次
（200PFlops）

中国智能计算系统代表“曙光7000”
浮点运算速度峰值达每秒30亿亿次
（300PFlops）



2018年**戈登•贝尔奖**由**劳伦斯伯克利国家实验室**和**NVIDIA公司**的联合研究团队使用**Summit智能计算平台**完成，获奖原因为“Employing **Deep Learning Methods** to Understand Weather Patterns”，该模型使用了混合精度进行训练，峰值算力达到了**1.13Eflops**

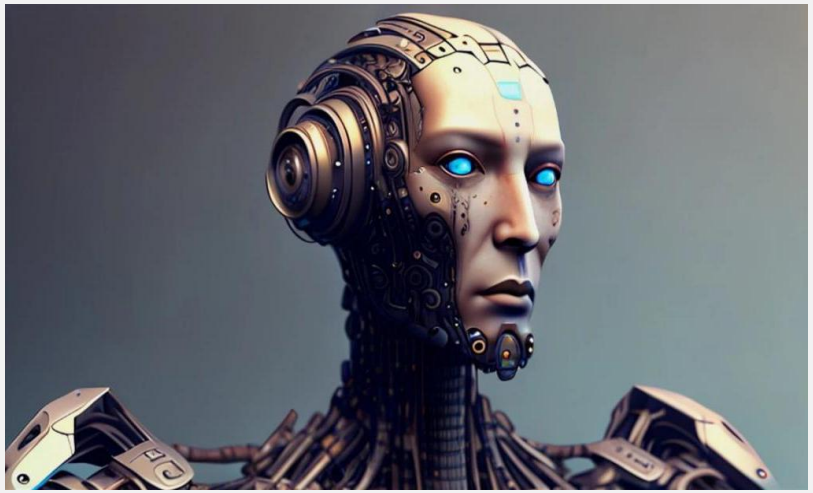
第二代智能计算系统

代表性深度学习处理器/计算机

时间	深度学习处理器/计算机	研制单位	特点
2013 年	DianNao[19]	中科院计算所	国际上首个深度学习处理器架构
2014 年	DaDianNao[20]	中科院计算所	国际上首个多核深度学习处理器架构
	cuDNN（深度学习库）	NVIDIA	升级 GPU 用于深度学习
2015 年	PuDianNao[21]	中科院计算所	国际上首个通用机器学习处理器
	ShiDianNao[22]	中科院计算所	端侧视频图像处理
2016 年	Cambricon[23]	中科院计算所	国际上首个深度学习指令集
	Cambricon-X[24]	中科院计算所	国际上首个稀疏神经网络处理器
2017 年	TPU[25]	Google	基于脉动阵列架构
	FlexFlow[26]	中科院计算所	动态数据流结构
2018 年	TPUv3 cloud	Google	基于 TPUv3 芯片的云计算
	DGX-2 服务器	NVIDIA	16 块 NVIDIA v100 显卡
	Summit 超级计算机	IBM	27684 块 NVIDIA v100 显卡
	MLU100	Cambricon	基于寒武纪云端智能芯片
2019 年	E-RNN[27]	Syracuse 大学	循环神经网络加速器
	Cambricon-F[28]	中科院计算所	分形冯诺依曼架构
	Float-PIM[29]	UCSD	支持训练的存内计算架构
2020 年	Azure	Microsoft	10000 块 NVIDIA 显卡，用于 GPT 系列研发
	DGX A100 Superpod	NVIDIA	140 个节点，1120 块 NVIDIA A100 显卡
2021 年	Frontier	Oak Ridge Leadership Computing Facility	8472 个节点，37888 块 AMD MI250X 加速器
2022 年	DGX H100 服务器	NVIDIA	8 块 NVIDIA H100 显卡
2023 年	DGX GH200	NVIDIA	256 块 NVIDIA Grace Hopper 超级芯片，900 GB/s 卡间互联

第三代智能计算系统

- 大模型的发展为第三代智能计算系统的发展提供了一种可能
- 随着智能计算系统计算能力的逐步增强，深度学习大模型可以变得越来越大，甚至在规模上超过人脑，这将不仅仅是把个别弱人工智能问题做得更好，而是能逐步逼近强人工智能，从而像人一样在各种简单问题上表现出好的效果
- 若我们能使大模型进一步拥有推理和涌现等高级认知智能，或许强人工智能有可能成为现实
- 第三代智能计算系统应当具有超强计算能力，从而能涌现出强人工智能的系统



第三代智能计算系统

总体思路：具备全面感知能力和超大规模硬件的原始智人是怎样一步步获得智能的?

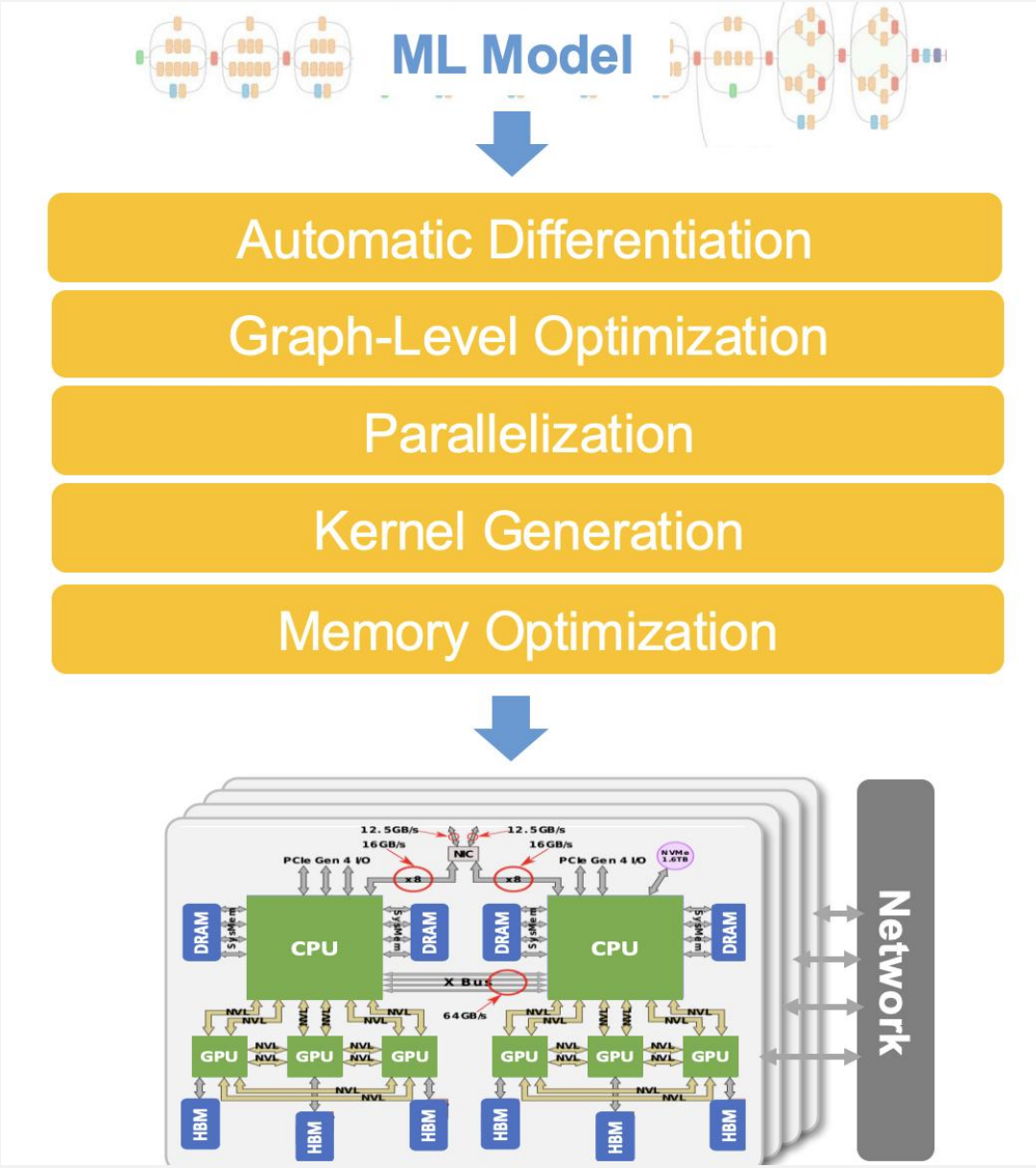
- **体系结构**：面向海量并发认知智能计算线程和超大规模虚拟环境的计算机和芯片
- **算法**：有限延迟的认知智能算法，能自主产生语言和文字，从本能之上建立起自己的知识图谱，打通感知到逻辑的鸿沟
- 编程框架，操作系统，网络等等都将为之巨变

第三代智能计算系统

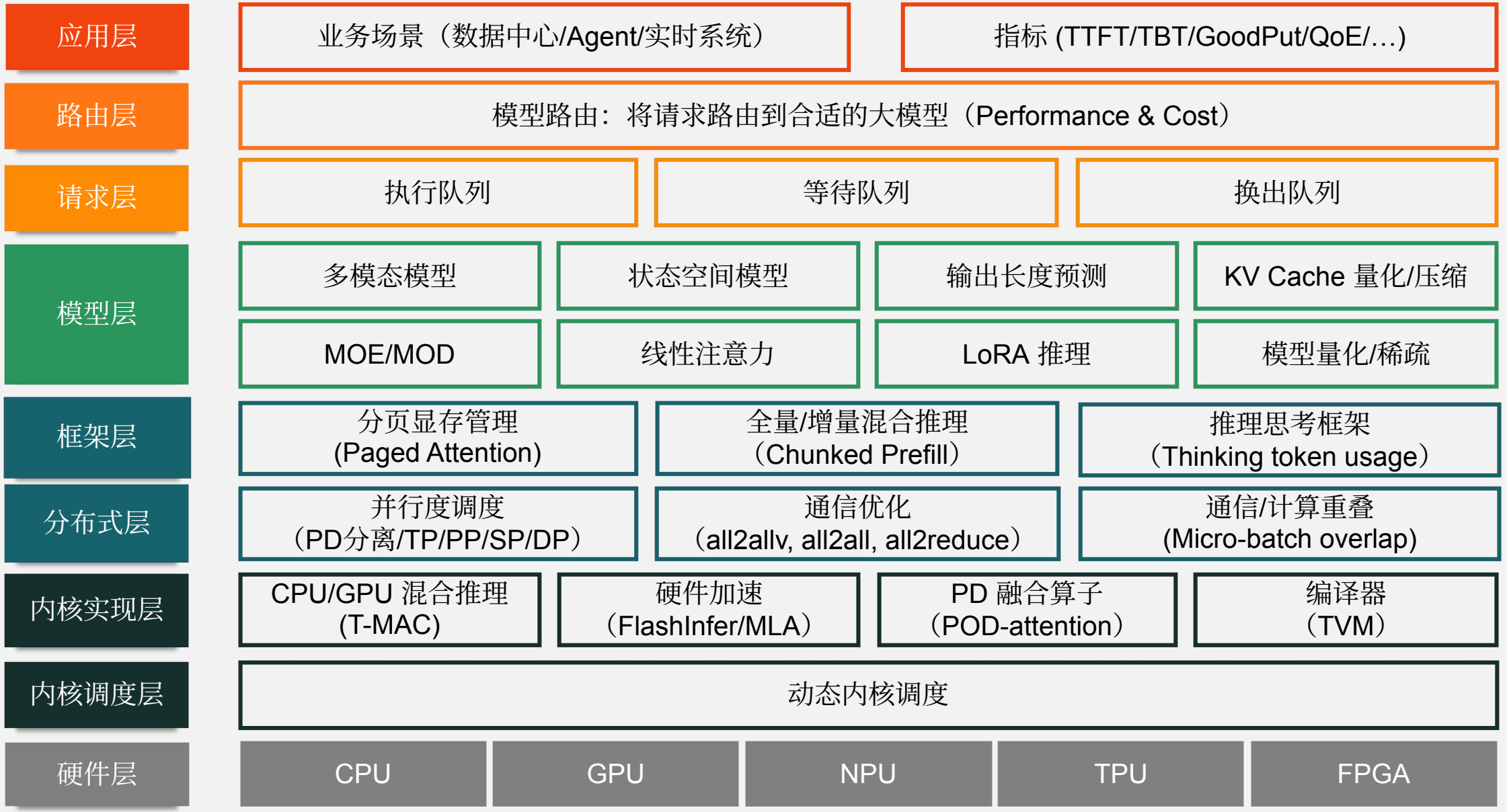
- 编程框架，操作系统，网络等等都将为之巨变

https://o.doubao.com/

机器学习系统：优化层介绍



大模型训练推理加速平台框架



大模型系统常见的问题

- **序列变长后无法处理**：计算与显存平方级增长
- **算力很强但跑不快**：常见卡在**内存带宽/访存模式**
  - **算子很多很碎**：kernel launch + 读写中间结果导致吞吐下降
- **训练吞吐看起来很好，部署延迟却很差**：目标函数不同

为什么发布常强调上下文窗口大小?

序列长度 n 是“平方杀手”

- 自注意力核心：QK^T 产生 n×n 的相关矩阵 → **计算量 $O(n^2)$**
- 同时带来**显存/带宽压力**：注意力中间张量、KV cache 占用随 n 上升
- 结果：当 n 从 2k→8k，系统不是“慢一点”，而是可能**直接 OOM/抖动/吞吐断崖**

你会看到的症状（现象→指标）

- 吞吐下降、显存飙升、OOM
- GPU 利用率不稳定：算力空转但访存/调度变重
- 推理端：prefill（一次性处理长 prompt）特别慢

系统对策

- **减少 n 或减少参与注意力的 token**：稀疏/分块/选择性计算
- **减少中间结果存储**：重计算、KV 管理
- **并行与分片**：张量并行/序列并行分摊 n^2

Claude's context window size is 200K, meaning it can ingest 200K+ tokens (about 500 pages of text or more) when using a paid Claude plan.

Claude

have access to a 500K context window when . See What is the Enterprise plan? for more

Ultra-Long Context Support

KIMI Platform

- `kimi-k2.5`, `kimi-k2-0905-Preview`, `kimi-k2-turbo-preview`, `kimi-k2-thinking`, and `kimi-k2-thinking-turbo` models all provide a 256K context window.

为什么“内存带宽”经常是瓶颈

不是 FLOPs 不够，是数据喂不进去

- 很多算子是**低算术强度（Arithmetic Intensity）**：每读写 1 字节只做很少计算
- GPU 峰值 FLOPs 高，但**HBM 带宽**有限 → 大量算子变成“带宽受限”

典型症状

- GPU SM 利用率不高，但 HBM/DRAM 读写很忙
- 算子时间几乎与 FLOPs 无关，反而与 tensor 大小/访存模式高度相关

常用系统优化方向

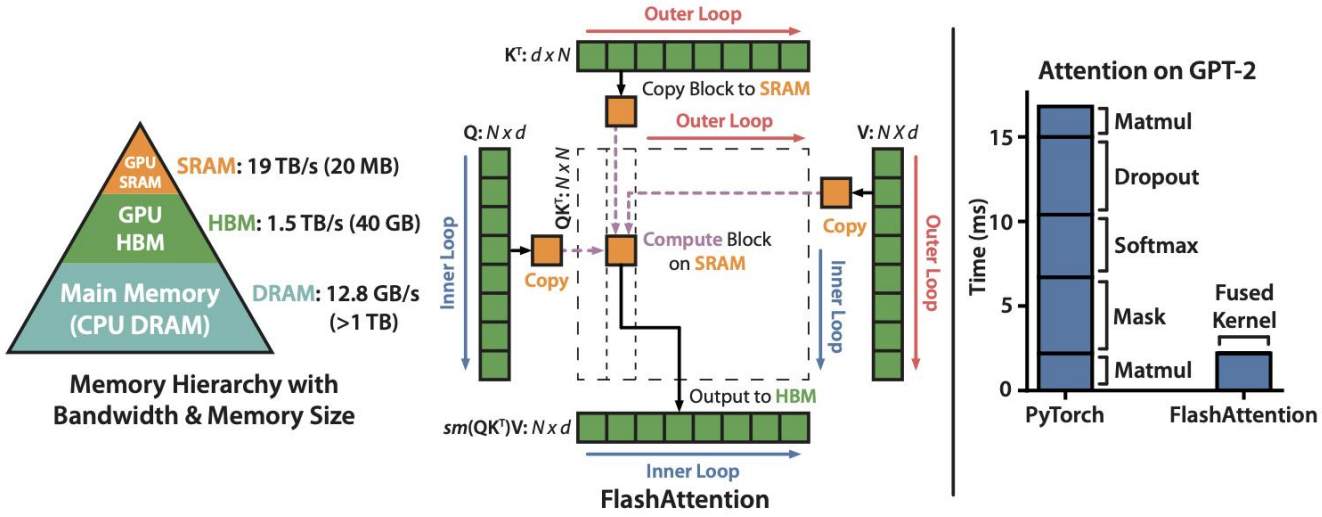
- **访存合并（coalescing）**、减少随机访问
- **减少读写次数**：融合算子、in-place、减少中间张量落地
- **提升数据复用**：tiling、共享内存/缓存利用、重排 layout
- **混合精度/量化**：减少字节数 → 直接释放带宽压力

FLASHATTENTION: Fast and Memory-Efficient Exact Attention with IO-Awareness

Tri Dao†, Daniel Y. Fu†, Stefano Ermon†, Atri Rudra‡, and Christopher Ré†

†Department of Computer Science, Stanford University
‡Department of Computer Science and Engineering, University at Buffalo, SUNY

{trid,danfu}@cs.stanford.edu, ermon@stanford.edu, atri@buffalo.edu, chrismre@cs.stanford.edu



本课程学习产出目标

大模型部署性能瓶颈分析

- 能从 算子 / 计算图 / 内存 / 通信 / 并行 / 精度 等维度拆解分析
- 能读懂并解释关键指标：吞吐、延迟、显存占用、带宽利用率、算力利用率

跨模型系统的多层级优化

- 掌握系统化优化流程：**定位 → 提出假设 → 测试 → 提出方案 → 优化实现 → 复测验证**
- 能在不同层面实施优化：改图（融合/重排）/ 改 kernel（访存/并行度）/ 改并行（DP/MP/PP）/ 改量化（精度-性能权衡）

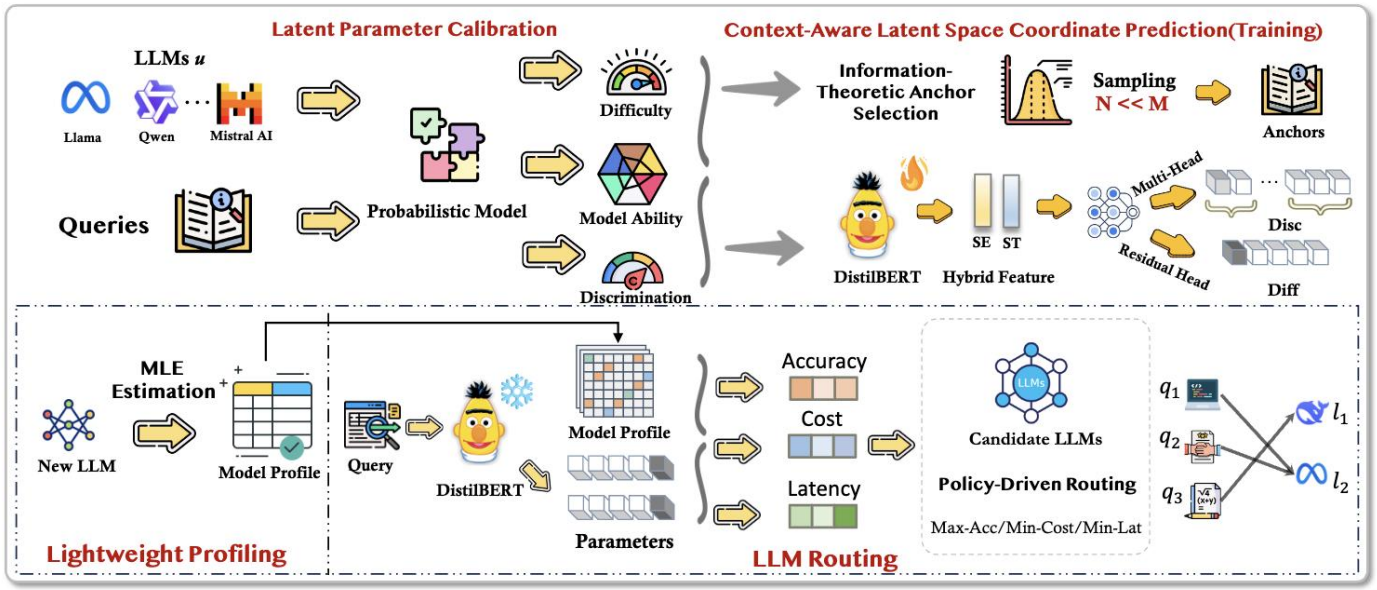
落地到真实系统栈

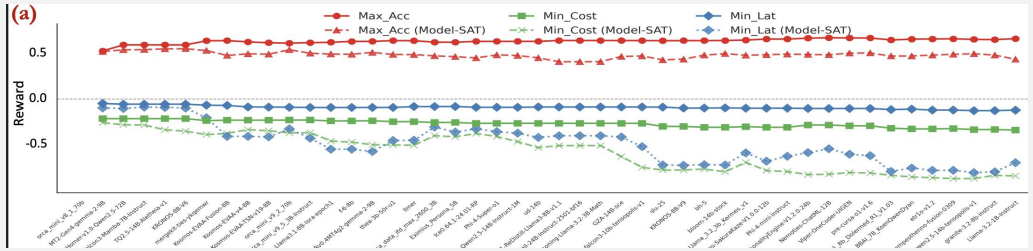
- 理解并贯通完整软硬件栈：**算子图 → 编译/图优化 → CUDA kernel → 运行时 → 分布式训练 → 部署推理**
- 能把模型从“能跑”做到 可复现、可扩展、可部署、可观测、可调优

机器学习系统工作范例

云端大模型推理加速 （路由层）

挑战：精准预测输入prompt复杂度以及模型能力，动态匹配服务的推理大模型，在保证任务精度的情况下大幅降低大模型推理成本25%。





测试近三年发布的60个大模型

- **通用难度层级：**
  - 利用项目反应理论估计**模型无关**的数据难度值；
  - 采样少至100个代表样本即可**高效评估模型**，获取新集成模型的难度-性能及难度-成本关系。

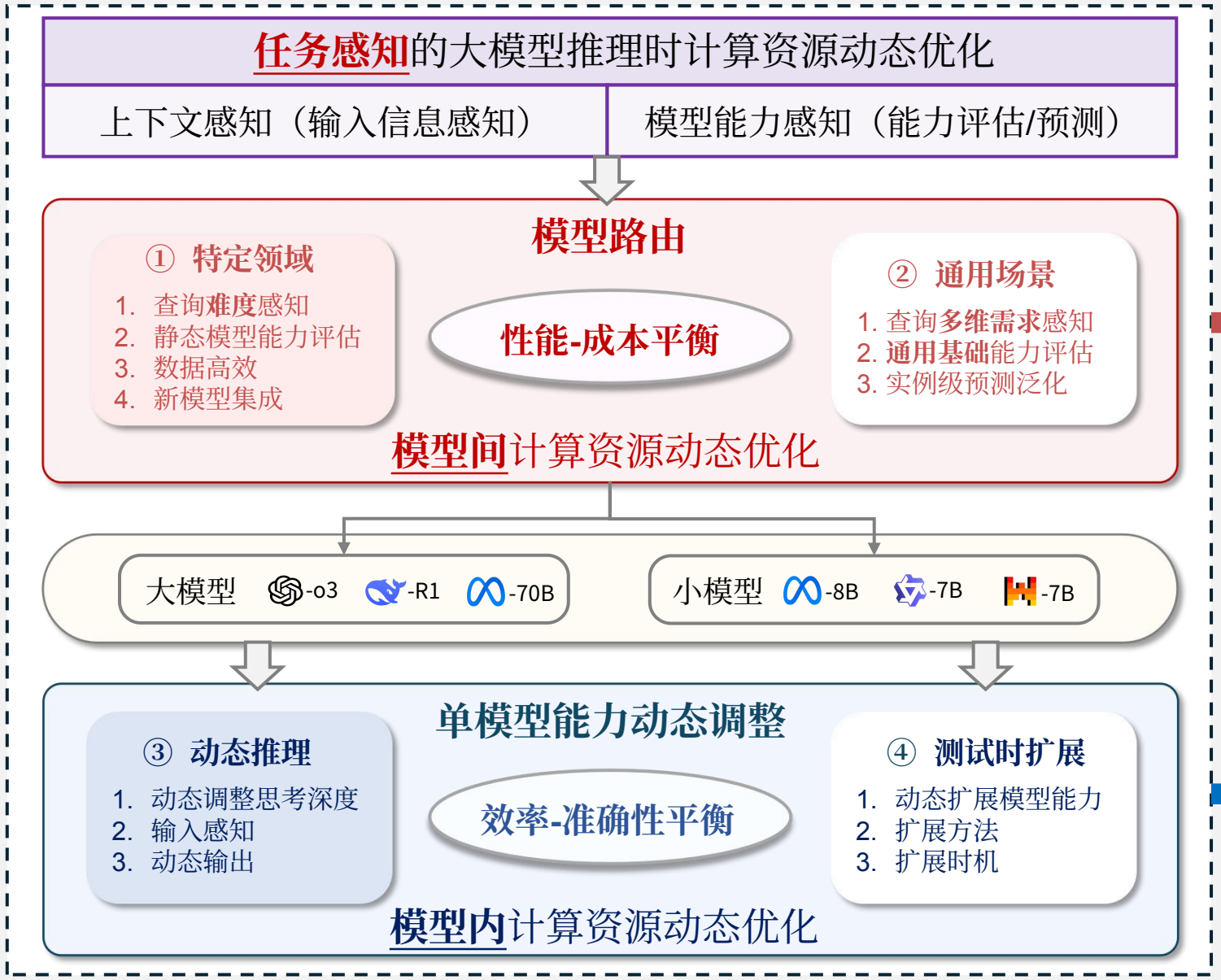
- **上下文感知的难度预测器：**
  - 提取查询文本的语义深度和结构复杂度两种互补表征，通过轻量化的DistilBERT模型**实时为未见过的查询**预测难度分数。

- **成本感知的路由策略：**
  - 利用线性整数规划算法，实现不同成本/性能约束下的模型分配策略，以做到**成本与准确性的平衡**。

高效性与扩展性　　实时性与泛化性　　动态性

Breaking Model Lock-in: Zero-Shot Difficulty-Aware Routing for Cost-Efficient LLM Services

任务感知的大模型推理时计算资源动态优化 （数据层）



一、从**难度感知**角度解决选择哪个模型的问题

- **查询难度感知**
- **模型高效评估**
- **新模型集成**

二、解决模型选择中的通用场景泛化问题

- **建立基础通用评估体系**
- **实例级预测泛化**

三、解决模型思考深度的问题

- **输入属性感知**
- **平衡效率与准确性**

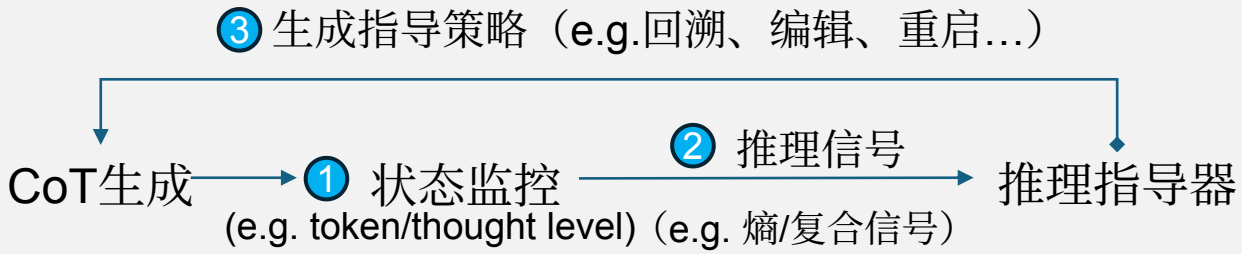
四、解决模型**如何扩展及何时扩展**的问题

- **动态能力扩展**
- **扩展时机决策**

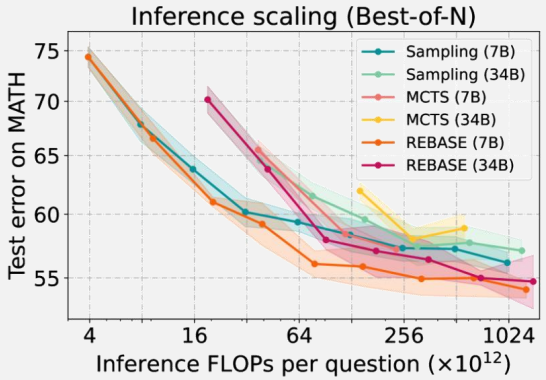
高效推理

优化中的关键问题：

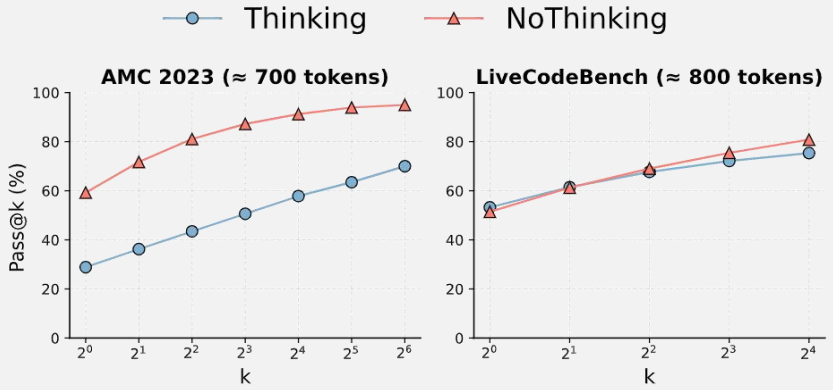
- **量化推理效用：**如何设计**评估指标（e.g. 熵、一致性）**，量化思维链中每一步推理状态（思考的合理性、过度思考或欠思考程度），作为指导LRM后续推理的**信号**。



- **计算最优的测试时扩展：**量化分析测试时**计算量扩展与LRM性能增益的关系**，系统地权衡成本-性能的关系。
- **跨任务泛化性：**考虑不同任务特性对思考深度需求的**差异性**。



MATH inference scaling across model sizes [1]



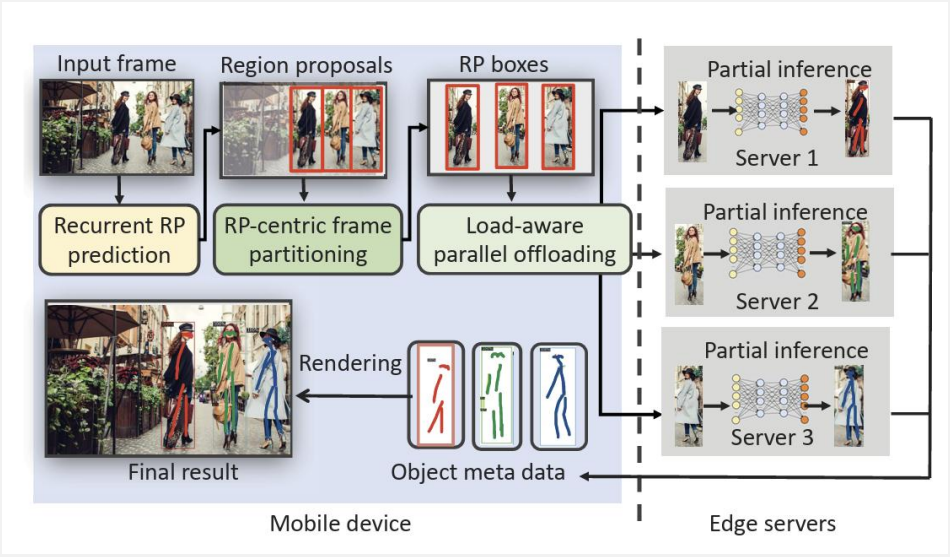
Thinking vs. NoThinking as k increases [2]

[1] Wu Y, Sun Z, Li S, et al. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving[C], ICLR. 2025.
[2] Ma W, He J, Snell C, et al. Reasoning models can be effective without thinking[J]. arXiv preprint arXiv:2504.09858, 2025.

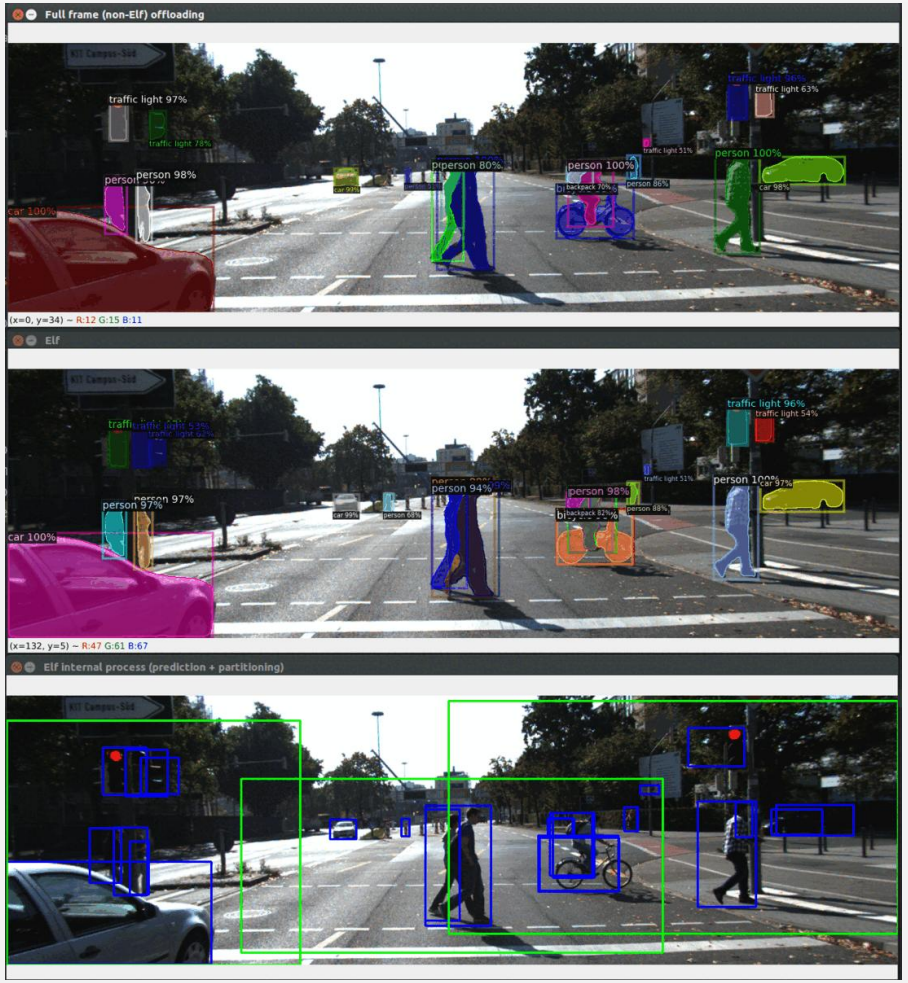
云端大模型推理加速 （数据层）

挑战：针对高分辨率视频流推理图像识别、分割任务进行基于数据并行的分布式计算。

创新方法：通过负载均衡感知的动态视频帧切割，将单帧视频图并行卸载到多计算节点并行计算。



成效：高分辨率视频推理速度提升4.85X

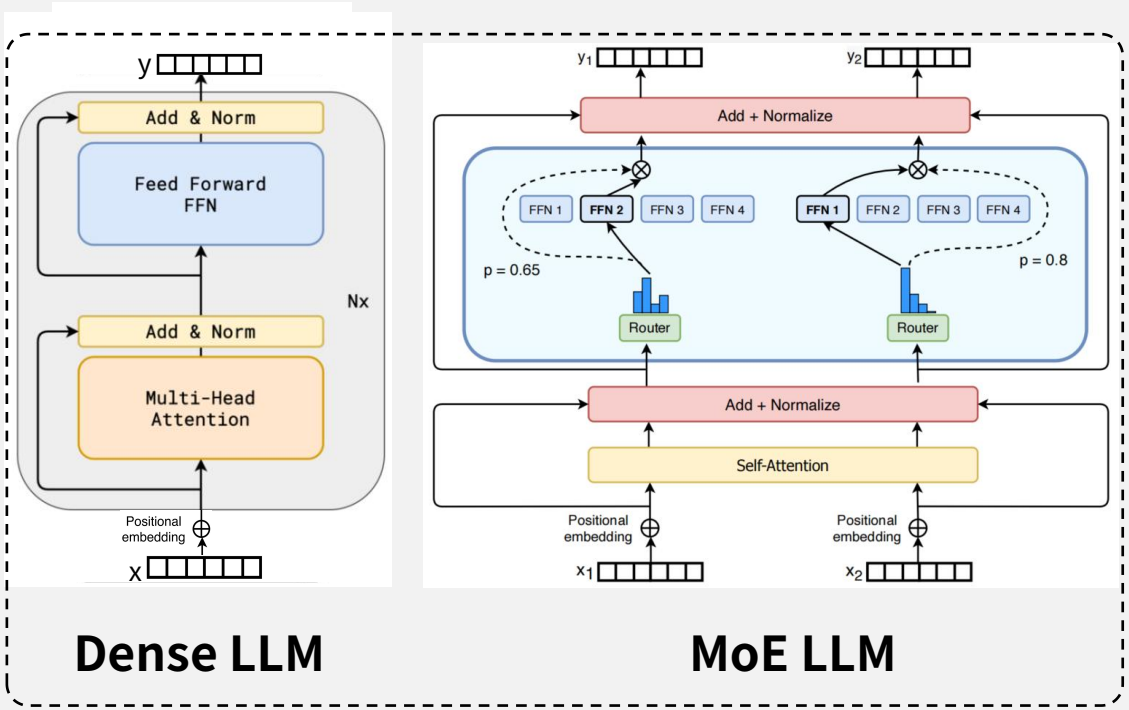


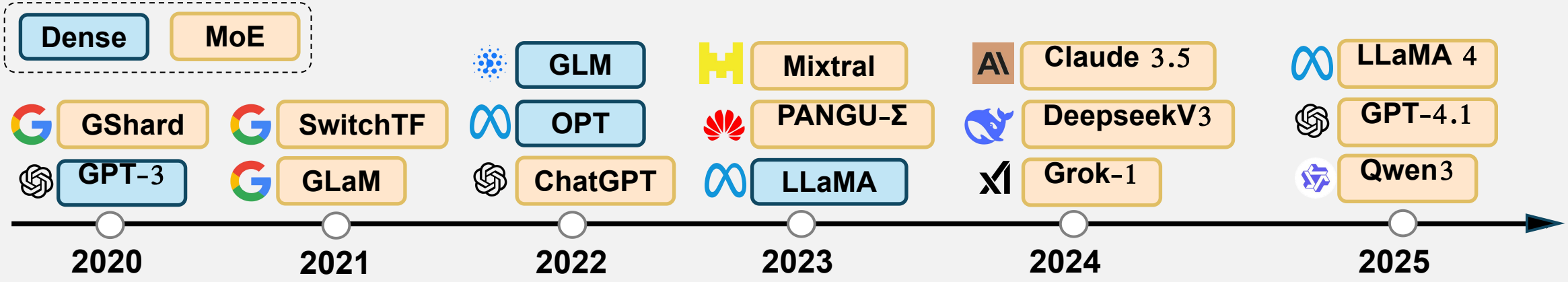
Elf: accelerate high-resolution mobile deep vision with content-aware parallel offloading MobiCom21

LLM架构演进：从密集模型到混合专家模型

迈向动态、模块化和专业化的计算新范式

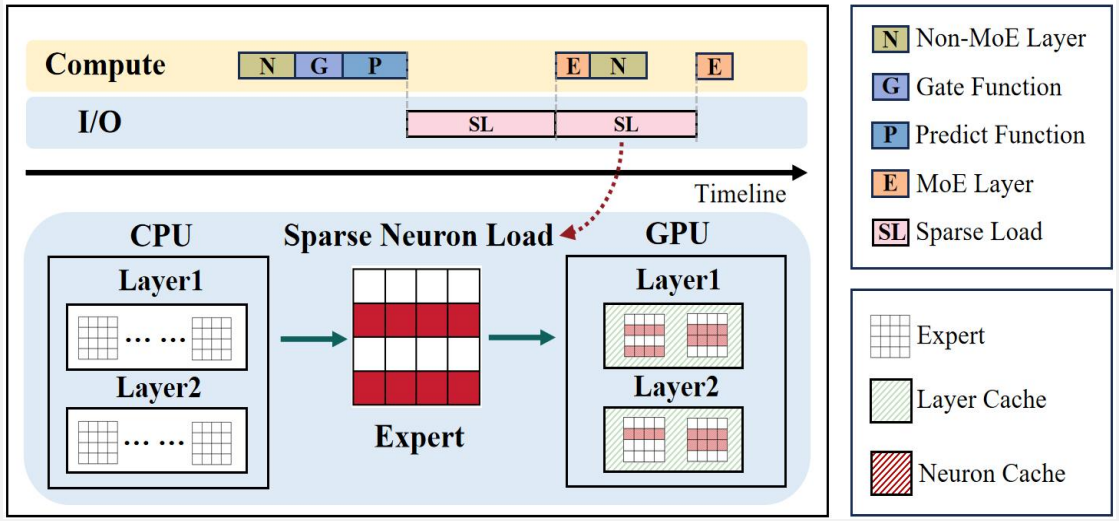
- 缓解指数级增长的计算开销
- 促进模型内部的功能特化
- 向条件计算范式演进

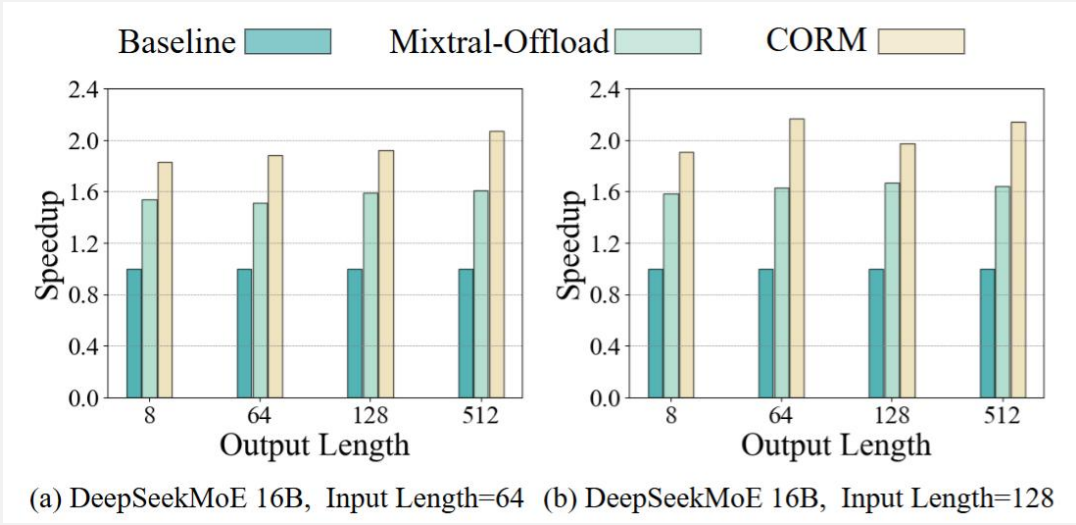




云端大模型推理加速（模型层）

挑战：MoE LLM的规模往往超出边缘设备内存的限制。在已有的解决方案中，将**模型卸载**到低级存储，推理时动态加载的方法导致**高延迟推理**；**量化方法**导致模型**精度大幅下降**。





创新方法：

- 利用SMoE LLMs的**双层稀疏性**来加速推理，包括MoE模型的稀疏路由策略和LLM的神经元稀疏特性。
- 通过稀疏神经元加载技术实现**粗粒度到细粒度**的权重加载。利用残差网络性质进行**稀疏性预测**，实现权重加载和计算并行。

成效：

- 在开源的SOTA MoE LLM上验证，相比 SOTA 方法端到端推理加速比达**2.01×**。

CORM: COARSE-TO-FINE-GRAINED OFFLOADING FOR SMOE LLM INFERENCE ON CONSUMER-GRADE GPU' **ICASSP 26**

云端大模型推理加速 （模型层）

ICML 25

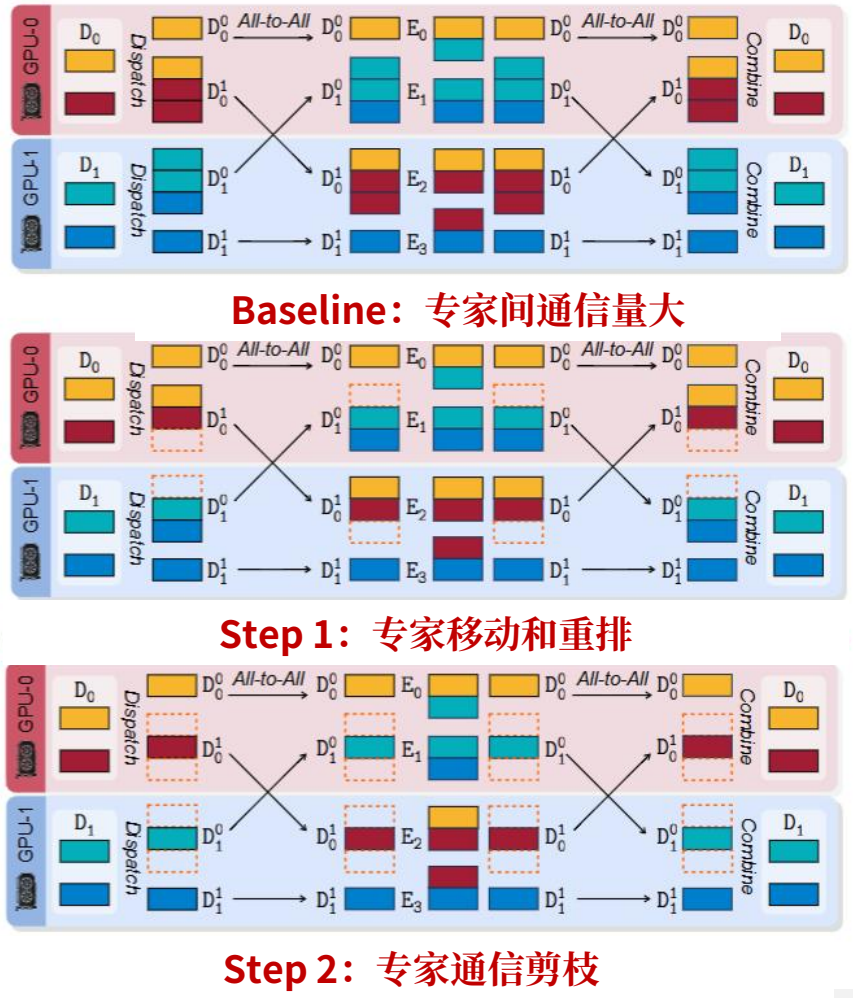
针对 MoE 专家并行中 All-to-All 通信开销巨大 的瓶颈，提出了一种“**专家协作**”机制。

- **通信冗余发现:** 深入分析 All-to-All 模式，发现大量 Token 在设备间传输是不必要的。我们重新定义通信范式，利用**设备内**来减少数据搬运。
- **专家布局重构**：基于“专家协作度”进行**静态重排**，将经常共同响应的Token的专家部署在同一 GPU 上，变“**远程通信**”为“**本地内存访问**”。
- **协作通信剪枝:** 对于必须跨设备的低收益专家请求，使用**本地相似专家**进行替代，在极微损精度的前提下极致压缩通信量。

成效：

- 在开源的SOTA MoE LLM上验证，相比 SOTA 方法端到端推理加速比达**1.51×~8.66×**

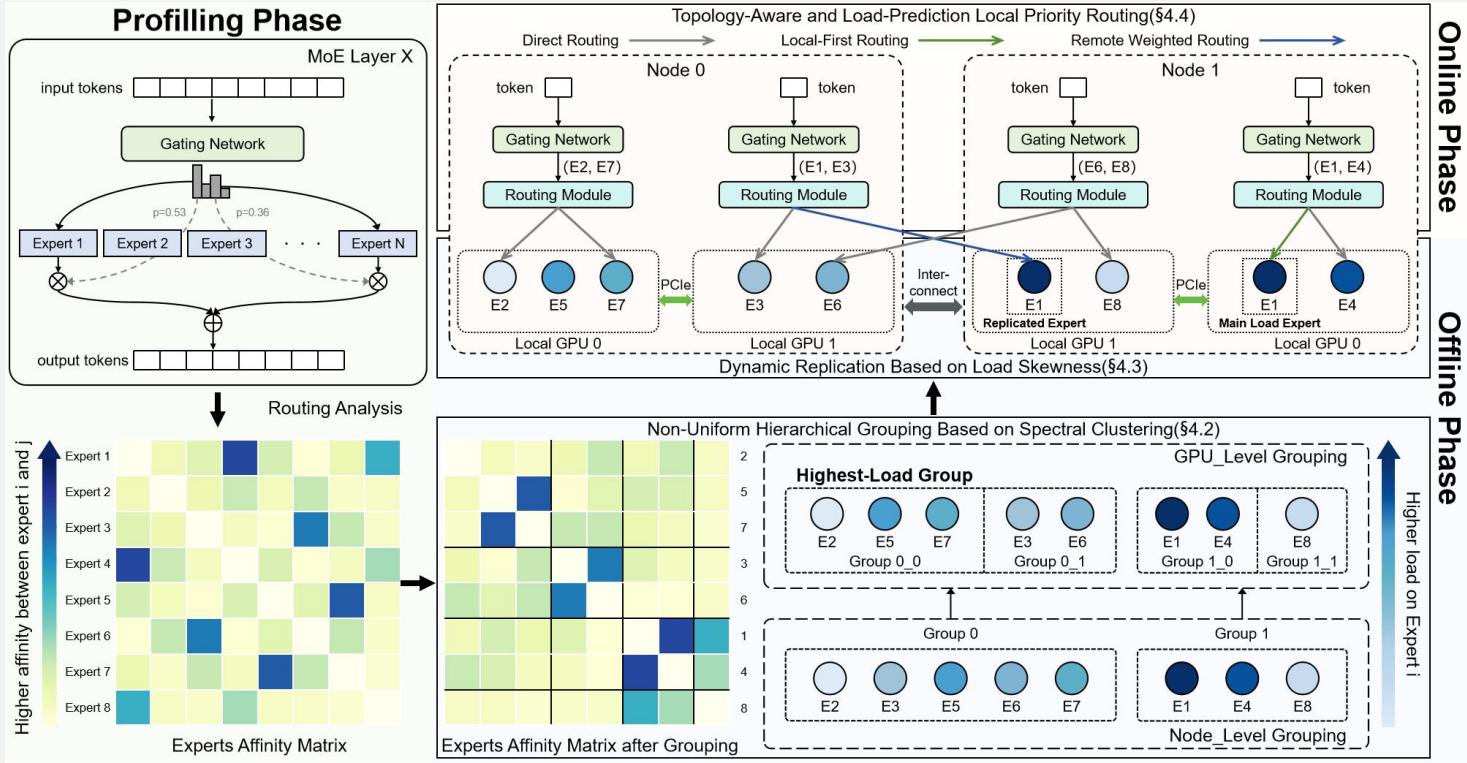
■ **系统框架**



云端大模型推理加速 （模型层）

- **痛点**: 现有的分组方法（如项目二）为了负载均衡，强制每个 GPU 放一样多的专家，但这违反了专家的自然属性。
- **非均衡谱聚类: 打破规模一致性约束**，完全基于专家间的**亲和力**进行物理聚类。允许“热门专家组”包含更多专家，从源头消除跨设备通信。
- **偏度感知动态复制**: 针对非均衡放置可能引发的计算热点，实时监测负载偏度，仅对高频访问的“核心专家”进行动态复制，以空间换时间，填平负载差异拓扑路由
- **本地优先路由**: 引入拓扑感知机制，在主专家与副本之间动态决策。优先将 Token 路由至通信开销最小（即本地或近端）的专家副本，实现带宽与算力的双重最优。

■ 系统框架



成效：

- 在开源的SOTA MoE LLM上验证，相比 SOTA 方法端到端推理加速比达**1.75×~2.31×**