中国科学院大学
University of Chinese Academy of Sciences

（2025）春季课程

人工智能安全与对抗

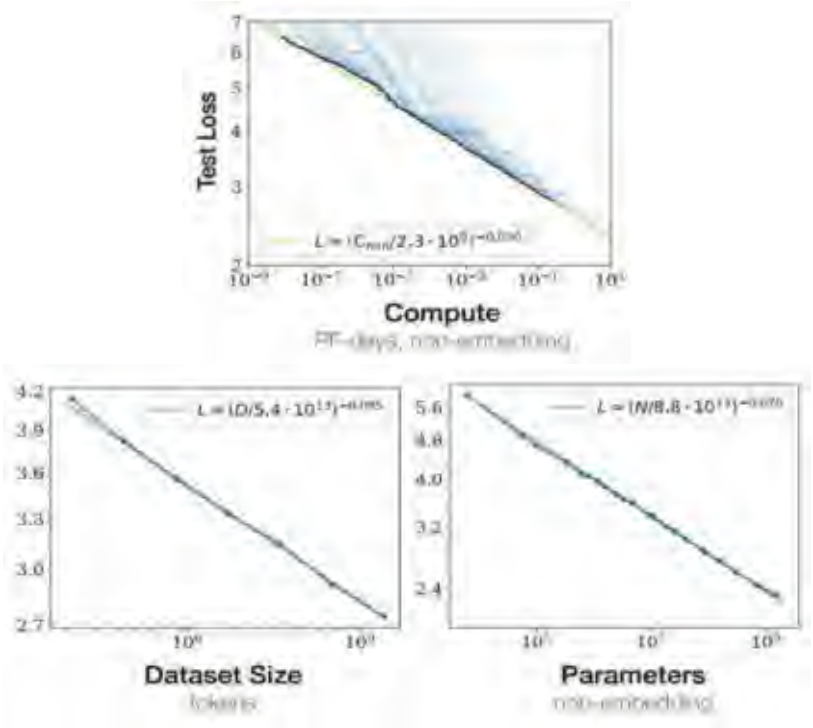
Artificial Intelligence Security, Attacks and Defenses

授课教师：董 晶　研究员
彭 勃　副研究员

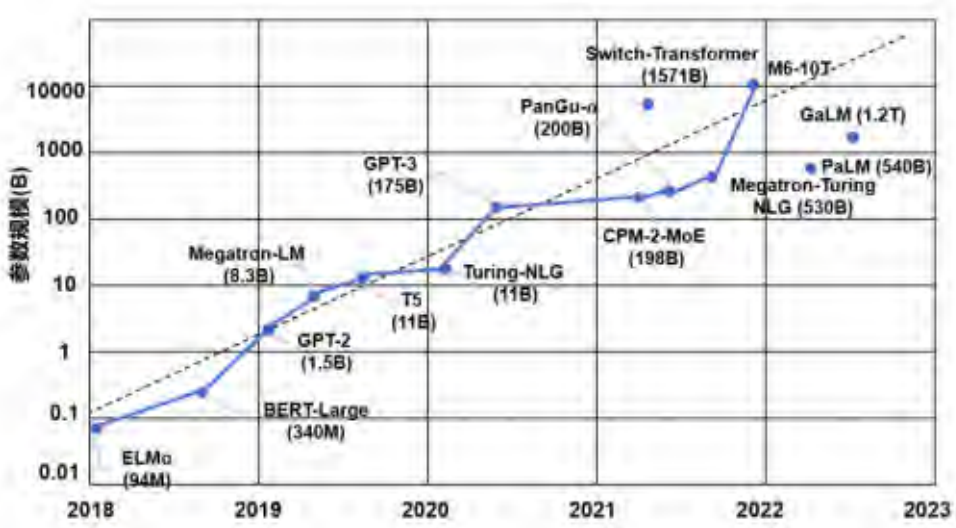
中国科学院自动化研究所

Recap：大模型技术发展的第一性原理：Scaling Laws

- 核心思想：测试Loss与模型参数量、数据集大小、训练计算量有幂律关系
- 应用于Pre-training阶段，提升模型基座能力；微调适配下游应用任务
- 本质：模型容量越大，记忆知识越多，模型能力越强



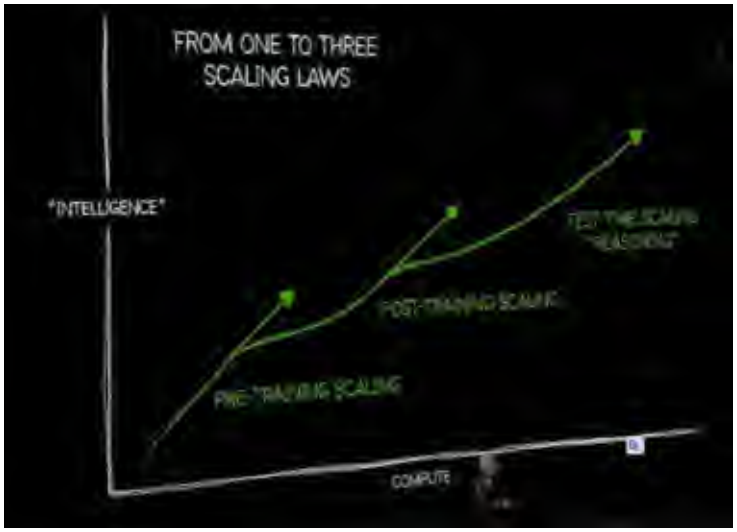
2018–2023 年模型参数规模变化图



[1] Kaplan J et al. Scaling laws for neural language models. OpenAI， 2020.
[2] 中国人工智能白皮书， 2023.

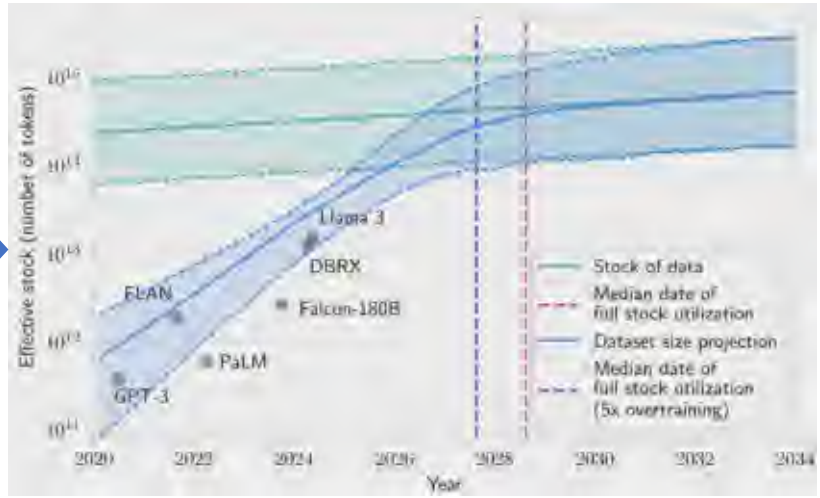
Recap：Scaling Laws并非免费的午餐！

- 背景：Scaling Laws增长速度放缓
- 原因：数据和算力瓶颈
- 解决方案：
  - 提示词工程、监督微调、知识增强等技术
  - Post-training Scaling Laws（后训练）

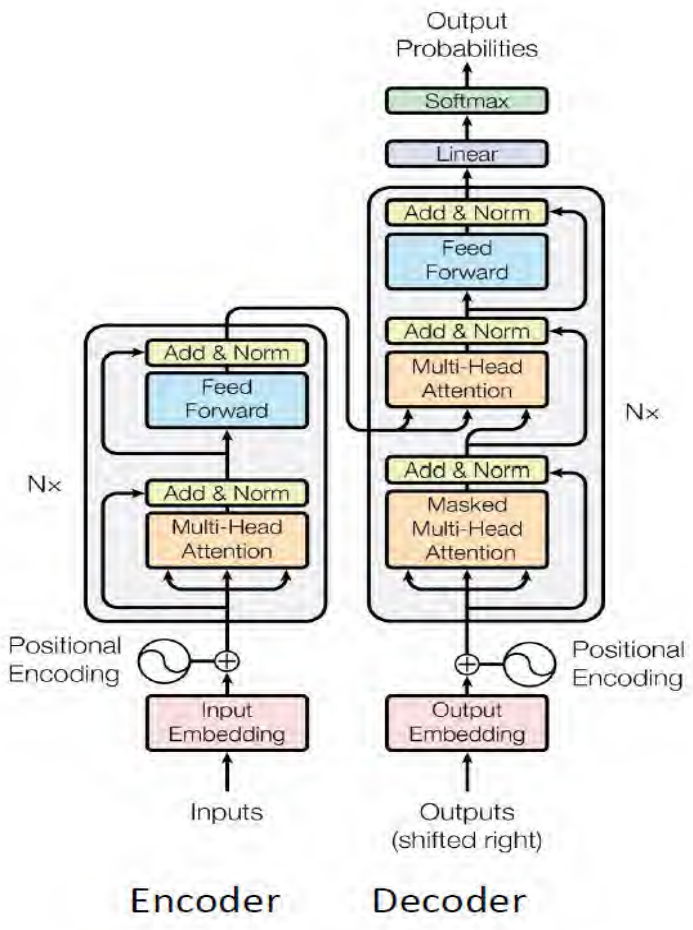


高质量数据受限

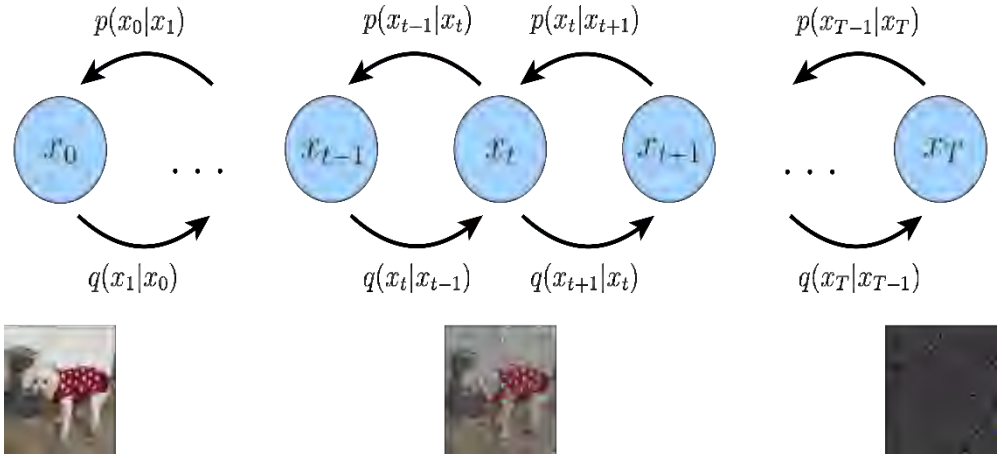
计算资源受限



Recap：大模型底层原理



Transformer



扩散模型 Diffusion Model

Recap： DeepSeek为什么引起关注？

模型（开源）	发布时间	模型类型	架构与参数	对标
DeepSeek-v1	2023.11	语言模型	稠密 67B	GPT 3.5
DeepSeek-v2	2024.05	语言模型	MoE 236B	GPT 4 Turbo
DeepSeek-v3	2024.12	语言模型	MoE 671B	GPT 4o
DeepSeek-R1	2025.01	推理模型	MoE 671B	OpenAI o1
Janus-Pro	2025.01	多模态生成模型	Unified 7B	LLaVa、SD等

- 核心贡献：模型训练与推理成本为对标模型或传统方案的1/10
- 逻辑能力（数学）和中文能力优势明显

Recap： DeepSeek为什么引起关注？

模型（开源）	发布时间	模型类型	架构与参数	对标
DeepSeek-v1	2023.11	语言模型	稠密 67B	GPT 3.5
DeepSeek-v2	2024.05	语言模型	MoE 236B	GPT 4 Turbo
DeepSeek-v3	2024.12	语言模型	MoE 671B	GPT 4o
DeepSeek-R1	2025.01	推理模型	MoE 671B	OpenAI o1
Janus-Pro	2025.01	多模态生成模型	Unified 7B	LLaVa、SD等

- 核心贡献：模型训练与推理成本为对标模型或传统方案的1/10
- 逻辑能力（数学）和中文能力优势明显

Recap：大模型风险类型

- 大模型涉及的安全风险类型主要包括以下几个方面
  - 技术安全风险
    - 因模型设计、训练或部署中的技术漏洞引发的安全问题。
    - **E. g. 大模型幻觉、对抗攻击、越狱攻击、后门攻击、数据重构（窃取）攻击、RAG攻击**
  - 数据隐私风险
    - 涉及训练数据及用户交互中的隐私泄露问题。
  - 伦理社会风险
    - 由模型输出引发的社会公平、偏见或伦理争议。
  - 恶意使用风险
  - 法律合规风险
  - 系统漏洞风险

Recap：大模型防御

- 外挂式安全
  - 核心思想：对输入和输出进行检测或预（后）处理，去除不安全内容。效果可靠，成为现有商业大模型的主要安全措施
  - 局限性：增加系统时延
- 内生式安全
  - **大模型的安全对齐**
    - 基于人类反馈的对齐、基于AI反馈的对齐、基于社会交互的对齐
  - 纠正模型的偏见歧视
  - 纠正模型的不良内容
  - 消除模型的有害知识
  - 标记内容的生成来源
  - 大模型安全评测技术

第五周作业

1.大模型Scaling Law的内涵是什么？推理过程的Scaling Law又是什么？

	Scaling Law	推理过程的 Scaling Law
内涵及核心	大模型技术发展的第一性原理，测试 Loss 与模型参数量、数据集大小、训练计算量有幂律关系。	模型规模对推理效率的影响，主要涉及延迟、吞吐量、显存占用等指标。
规律	通过增加模型、数据、计算规模提升性能，但需平衡三者的比例。 1.计算规模：计算量越大，模型性能越好。 2.数据规模：训练数据量需与模型参数量匹配，避免过拟合或欠拟合。 3.模型规模：参数越多，模型容量越大，但需与计算和数据同步增长。	模型参数量与显存占用呈线性关系，大模型推理需要并行以分摊显存压力。 单次推理的计算量与模型参数量和输入长度成正比。 延迟随模型规模增长，但可通过优化缓解。 吞吐量受硬件计算能力和批处理大小影响。 大模型因显存限制，通常无法使用大数据量，导致吞吐量低于小模型。
优化方式	三者按固定比例缩放时（例如计算量翻倍，数据和参数量也需相应增加），性能提升最优。	通过量化、蒸馏降低推理成本，但可能损失性能。 利用模型激活稀疏性减少计算量。 硬件适配：针对特定硬件优化计算量。

第五周作业

1.大模型Scaling Law的内涵是什么？推理过程的Scaling Law又是什么？

大模型Scaling Law

损失函数与模型规模的关系	模型性能(如Test Loss)与模型参数量，数据集大小，训练计算量之间的幂律关系如下： $L(N,D)=L_0+a\cdot N^{\alpha}+b\cdot D^{\beta}$
体现阶段	①预训练阶段：主要应用阶段，通过增加模型规模与计算资源，构建强大的基础模型，学习更广泛、通用的知识。 ②微调阶段：通过微调模型特定部分，使模型适配下游应用任务。 ③后训练：预训练后通过强化学习、测试时间扩展等方法进一步提升模型性能。
本质	模型容量越大，能够记忆的知识越多，在合适的推理策略帮助下，其能力也越强。

相应的，推理过程的Scaling Law研究模型在推理阶段的性能随规模变化的规律，总结如下。

推理过程的Scaling Law

内涵	在模型实际应用阶段，通过增加计算资源或优化推理策略，进一步提升模型性能的规律。
核心思想	随着推理阶段投入的算力与响应时间的增加，模型性能也随之增长。
本质	通过引导模型进行更深入、更详细的思考，实现更准确、更可靠的输出结果，即在推理阶段"放大"了模型的能力。
具体方法	提示词注入，Process Reward Model+搜索方法等

通过结合上述不同阶段的Scaling策略，有望获得更强大的AI模型。

第五周作业

2. Transformer的自注意结构相比于CNN的卷积结构，其不同点主要体现在哪里？

核心区别：

1. 依赖关系建模方式

   CNN：通过局部感受野（卷积核）逐层提取特征，依赖堆叠层数捕捉长距离依赖，存在"信息传递路径长"的问题。

   Transformer：通过自注意力机制直接计算全局依赖关系，任意两个位置的信息可一步交互，更适合长序列建模（如文本、高分辨率图像）。

2. 计算复杂度

   CNN：计算复杂度为 $O(n \cdot k^2)$（n 为输入长度，k 为卷积核尺寸），对长序列友好。

   Transformer：自注意力复杂度为 $O(n^2)$，序列较长时计算开销显著增加，需通过稀疏注意力或分块优化。

3. 参数共享与平移不变性

   CNN：卷积核参数在空间上共享，天然具备平移不变性（适合图像任务）。

   Transformer：无参数共享约束，需通过位置编码（如正弦函数、可学习向量）显式引入位置信息。

4. 并行化能力

   CNN：局部卷积操作可并行计算，但深层网络需顺序执行。

   Transformer：自注意力的全局计算可完全并行化，训练效率更高。

小结：Transformer 通过全局注意力突破了 CNN 的局部性限制，但牺牲了计算效率；CNN 则凭借局部性和平移不变性，在图像领域仍具优势。

第五周作业

2. Transformer的自注意结构相比于CNN的卷积结构，其不同点主要体现在哪里？

二者核心差异在于：CNN通过归纳偏置实现高效特征提取，适合图像等高维规则数据；Transformer以动态计算全局关系为核心，通过减少先验假设换来更强的灵活性，在训练数据充足时表现通常更好。对二者不同点的总结如下表所示：

特性	卷积结构	自注意结构
信息交互范围	局部，通过堆叠多层实现全局	单层即可实现全局
参数共享	卷积核共享	权重各自独立
计算复杂度	线性O(n)	二次O(n^2)
序列/空间敏感性	局部平移不变，隐式位置建模	依赖位置编码，显式位置建模
适用场景	图像视觉任务为主	NLP，视觉任务皆均可
训练效率	参数相对少，训练稳定高效	计算资源需求大，训练难度大
可解释性	卷积核可视化，可解释性较好	全局交互复杂，解释性弱

此外我注意到，近年来对二者的融合也是重要的研究方向。

第五周作业

3. 简述本节课介绍到的几种大模型安全风险。

- **技术安全风险：** 包括对抗攻击、越狱攻击、后门攻击、模型窃取攻击等。对抗攻击通过输入prompt的微小改变来操纵模型的输出响应，如白盒攻击、灰盒攻击和黑盒攻击；越狱攻击通过设计输入内容、格式或模板，引导模型持续输出有害内容，绕过安全机制；后门攻击在训练过程中对模型植入后门，使模型在特定触发条件下输出恶意标签；模型窃取攻击通过蒸馏等技术获取模型的结构和参数，可能导致模型被免费使用或用于更危险的攻击。
- **数据隐私风险：** 涉及训练数据及用户交互中的隐私泄露问题，如数据重构攻击，利用模型的记忆特性提取训练数据，产生隐私威胁。
- **伦理社会风险：** 由模型输出引发的社会公平、偏见或伦理争议，如生成辱骂仇恨、违法犯罪、偏见歧视、敏感内容等有害信息。
- **恶意使用风险：** 不法分子可能利用大模型进行违法犯罪活动，如WormGPT、FraudGPT、Evil-GPT等模型被用于网络攻击、诈骗等。
- **系统漏洞风险：** 大模型可能存在系统性的漏洞，被攻击者利用来破坏模型的正常运行或窃取信息。

第五周作业

4. 大模型安全防御的主要思路是什么？并进一步阐释。

(1) 外挂式安全：外挂式安全的核心思想是对输入和输出进行检测或预（后）处理，去除不安全内容。这种方法效果可靠，成为现有商业大模型的主要安全措施。例如，LatentGuard 在隐空间训练一个不良语义分类器，实现不安全输入滤除；Ethical-Lens 对输入输出不合适内容进行检测和修改。然而，这种方法的局限性在于会增加系统时延。

(2) 内生式安全：内生式安全通过大模型的安全对齐，确保大模型行为与人类价值观一致，避免有害内容或有害行为。安全对齐分为三类方法：

① 基于人类反馈的对齐：将人类偏好整合到模型的训练过程中，主要依赖强化学习和监督学习手段。例如，基于人类反馈的强化学习（RLHF）将人类偏好固化到模型，以强化学习范式融入模型微调过程中；监督微调（SFT）使用带有人类偏好的数据对模型进行微调；直接偏好优化（DPO）评估完整的生成内容，并优化模型使得最大化偏好内容的生成概率。

② 基于 AI 反馈的对齐：引入基于 AI 反馈的对齐方法，替换人类反馈。例如，Constitutional AI 在监督学习和强化学习阶段利用 AI 反馈进行对齐。

③ 基于社会交互的对齐：模拟社会运作体系，从社会视角来评估和优化模型。这种方法考虑了整个社会价值观的视角，弥补了个体维度偏好对齐的不足。

课程大纲

1 AI技术与AI安全概述（3学时）

2 对抗机器学习-I（3学时）

3 对抗机器学习-II（3学时）

4 对抗机器学习-III（3学时）

5 大模型及其安全问题（3学时）

6 具身智能及其安全问题（3学时）

具身智能及其安全

0 具身智能概述

1 具身感知

2 具身推理

3 具身执行

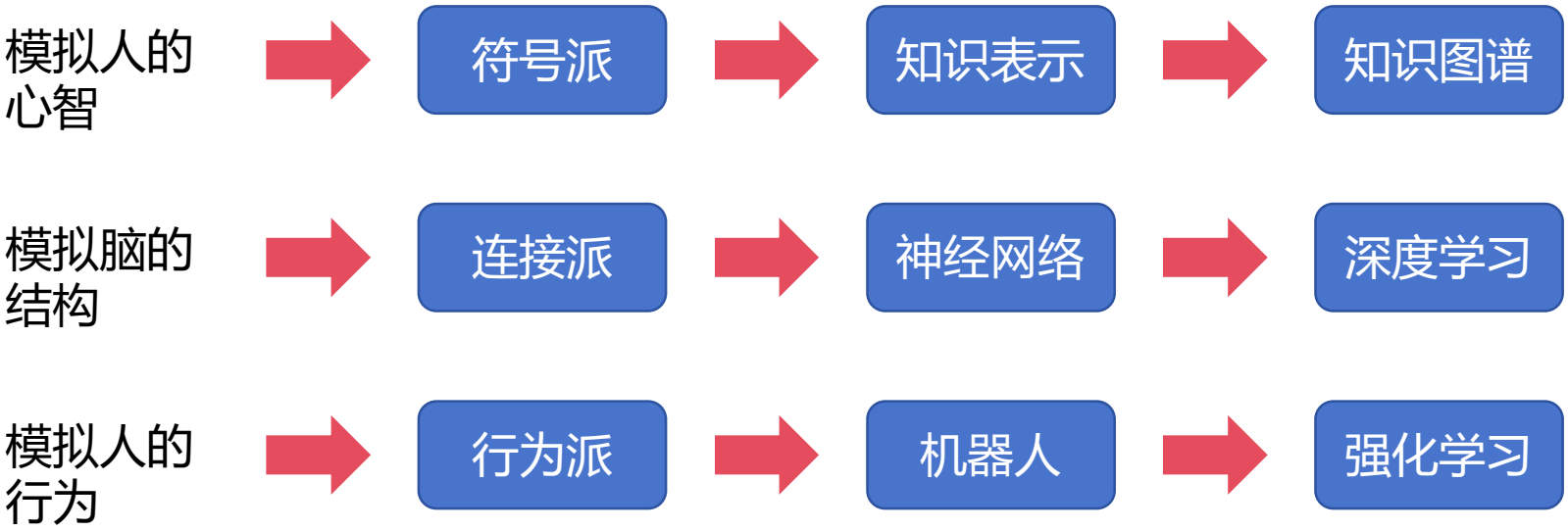
5 具身智能安全

6 具身智能安全评测

什么是智能机器人?

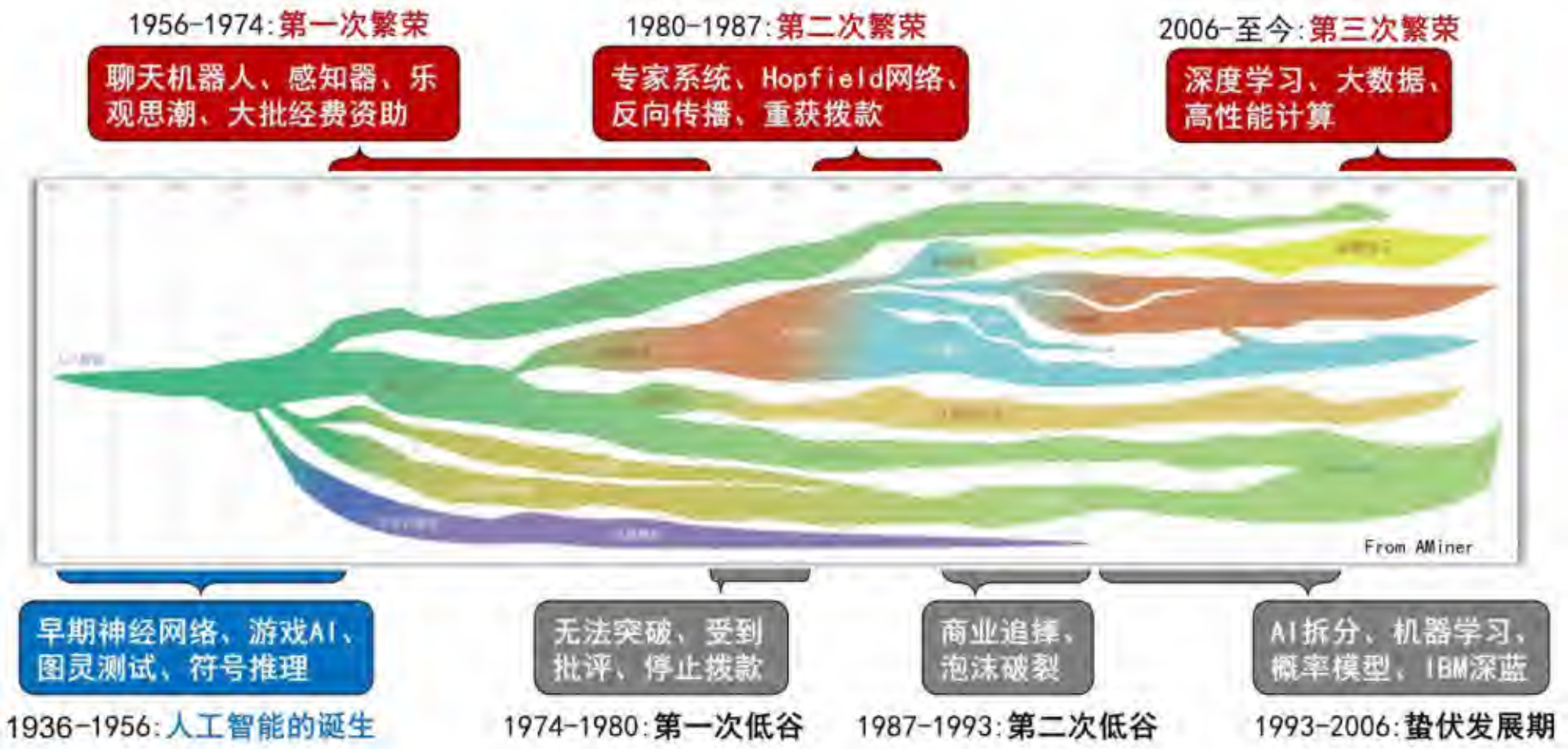
人工智能的三种学派

- 60多年历史，三种学派，三次浪潮



人工智能的三次浪潮

- 60多年历史，三种学派，三次浪潮



21世纪，工业机器人已经相对成熟，人们开始探索更多场景、更智能的机器人

医疗微创机器人

物流运输机器人

智能家居机器人

更好的自主性：应对的场景和任务更复杂，涉及多机器人协调

机器人 → **智能**机器人

① **自主能力**：尽可能少的人类干预

② **泛化能力(通用能力)**：具备强大的

机器人 → **智能机器人 ≈ 人类**

↓

一种算法 人工智能？

人工智能真的让机器人智能了吗?

人形机器人===智能机器人?

什么是具身智能?

人形机器人===具身智能?

我们设想中的智能机器人是什么?

像人类一样工作的机器人?

各方面强于人类的机器人?

有意识和情感的机器人?

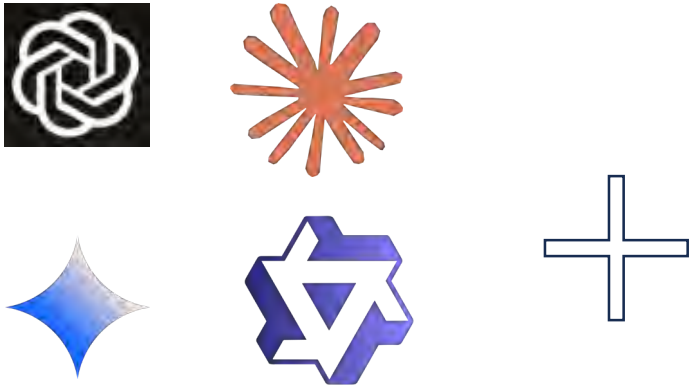
纵观人工智能发展

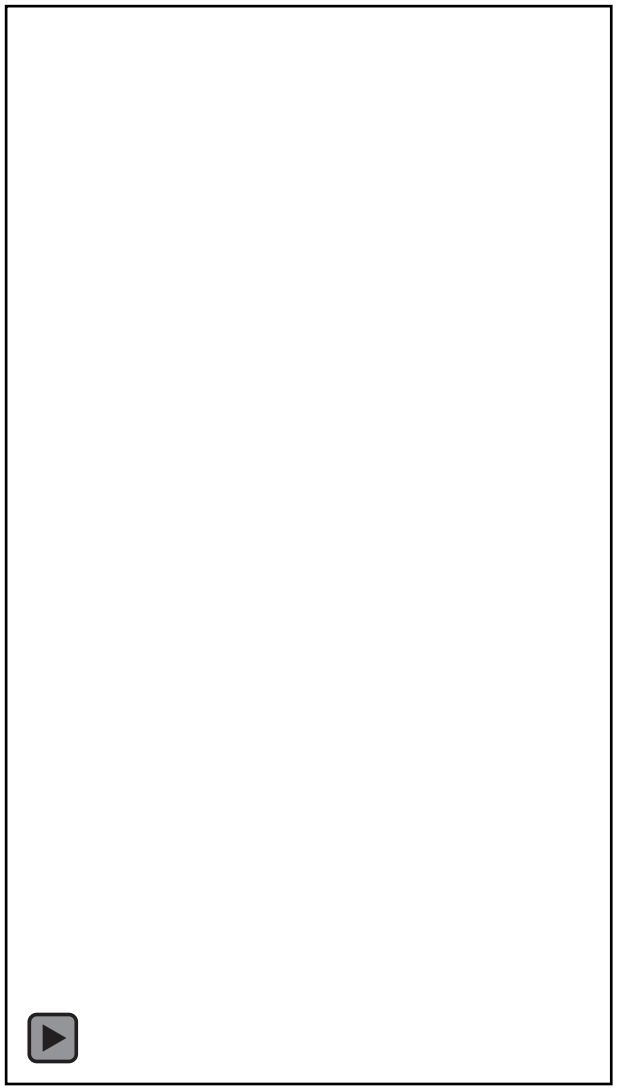
- 1956年—20世纪60年代初，使用人工智能做符号推理，进行**数学证明**
- 20世纪60年代—70年代初，启发式的搜索算法能力有限
- 20世纪70年代初—80年代中，构建专家系统处理医疗、化学、地质等特定领域应用
- 20世纪80年代中—90年代中，专家系统需要海量的专业知识，实用价值有限
- 20世纪90年代中—2010年，机器学习算法处理实际问题
- 2011年之后，深度学习算法用于图像、文本、语音等信息处理
- 2022年之后，可以处理通用任务的大模型
  - ✓ 一定的**自主能力**
  - ✓ 一定的**泛化能力**

但离我们设想的智能还有多远?

大模型与人形机器人结合形成智能机器人

- 上个世纪对未来人工智能的幻想，主要表现为智能人形机器人，但目前人工智能技术仍然停留在电脑屏幕，没有以实体的方式进入物理世界
- 目前智能程度最强的**大模型**，与目前最先进的**人形机器人**，能否结合形成**智能机器人**？





构建智能机器人(以人形机器人为例)

硬件方面

运控技术相对来说已经较为成熟

软件及算法方面

大脑

- 收集所有传感器采集的环境信息和自身状态。并综合分析当前所有状态**(具身感知)**
- 根据当前状态，对自身下一步的运动做出决策和规划**(具身推理)**

小脑

- 向下位机下发送运动指令**(具身执行)**
  (形式包括代码、技能库API、关节旋转角度等)
- 下位机通过运控技术执行指令

还存在诸多问题

当前人工智能这几个方面存在哪些问题?

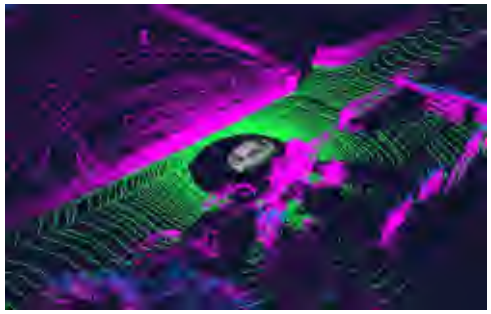
- 收集所有传感器采集的环境信息和自身状态。并综合分析当前所有状态**(具身感知)**

多模态大模型已能做到:

但实际场景远比此复杂

这是什么?如何打开它?

这些手势是什么意思?



3D点云图如何理解

当前人工智能这几个方面存在哪些问题?

- 根据当前状态，对自身下一步的运动做出决策和规划**(具身推理)**

来看目前大模型在EgoPlan比赛任务样例上的表现:

问题:根据视频中显示的当前状态，请问接下来我应当进行那个动作来制作拿铁?

Question: Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to *make coffee with milk (Task Goal)*?

Candidate Actions:
A. pick milk bottle B. open fridge
C. put down mug D. close milk bottle **Answer:** D

候选: A.拿起牛奶瓶 B.打开冰箱 C.放下马克杯 D.盖上牛奶瓶 **答案:D**

主流大模型在该数据集上的表现:

Table 1: Performance of MLLMs on EgoPlan-Val.

Model	LLM	Acc%
BLIP-2	Flan-T5-XL	26.71
InstructBLIP	Flan-T5-XL	28.09
InstructBLIP Vicuna	Vicuna-7B	26.53
LLaVA	LLaMA-7B	27.00
MiniGPT-4	Vicuna-7B	28.11
VPGTrans	LLaMA-7B	27.38
MultiModal-GPT	Vicuna-7B	27.81
Otter	LLaMA-7B	28.08
OpenFlamingo	LLaMA-7B	27.67
LLaMA-Adapter V2	LLaMA-7B	27.81
GVT	Vicuna-7B	27.87
mPLUG-Owl	LLaMA-7B	27.63
Kosmos-2	Decoder only 1.3B	26.97
Qwen-VL-Chat	Qwen-7B	27.69
LLaVA-1.5	Vicuna-7B	27.81
VideoChat	Vicuna-7B	27.51
Video-ChatGPT	LLaMA-7B	27.33
Valley	LLaMA-13B	27.27
Video-LLaMA	LLaMA2-Chat-7B	28.58
SEED-LLaMA	LLaMA2-Chat-13B	29.93
CogVLM	Vicuna-7B	27.48
DeepSeek-VL-Chat	DeepSeek-LLM-7B	27.57
mPLUG-Owl-2	LLaMA2-7B	27.84
Yi-VL	Yi-6B	28.67
Gemini-Pro-Vision	-	30.46
SEED-X	LLaMA2-Chat-13B	31.07
XComposer	InternLM-7B	37.17
GPT-4V	-	**37.98**

当前人工智能这几个方面存在哪些问题?

- 向下位机下发送运动指令**(具身执行)** (形式包括代码、技能库API、关节旋转角度等)

对于生成关节旋转角度形式的运动指令:

多模态大模型

倒水

关上抽屉

扩散小模型

转移红色方块

	执行的成功率	执行的流畅度	泛化能力
多模态大模型	较低(60%~70%)	不够流畅	**物品泛化**
多模态大模型	较高(90%以上)	流畅	**位置泛化**或无泛化

技能泛化
场景泛化
物品泛化
位置泛化
无泛化

泛化能力

对于生成技能库API或代码API形式的运动指令：**现实世界场景过于复杂，构建完整的技能库几乎不可能**