



浙江大学
ZHEJIANG UNIVERSITY

DeepSeek

智能时代的全面到来和人机协作的新常态

孙凌云 教授

计算机科学与技术学院

2025年2月





一、智能演变

二、人机协作

三、产业现状

四、教育成长



GPT 背后的大数据

GPT训练数据规模史无前例

从 2018 年的 GPT-1
到 2020 年的 GPT-3,
模型预训练数据量从 4.6GB 增加到了 45TB
45TB 相当于三千万本《西游记》

主要模型数据集包括:

- 维基百科数据集 (庞大的客观知识)
- 书籍 (故事讲述能力与反应)
- 杂志期刊 (语言生成的严谨)
- Github代码等其他数据 (逻辑推理)

	维基 百科	书籍	杂志 期刊	Reddit 链接	Common Crawl	其他 数据	总计
GPT-1		4.6					4.6
GPT-2				40			40
GPT-3	11.4	21	101	50	570		753
The Pile v1	6	118	244	63	227	167	825
Megatron- 11B	11.4	4.6		38	107		161
MT-NLG	6.4	118	77	63	983	127	1374
Gopher	12.5	2100	164.4		3450	4823	10550

表: 主要数据集大小汇总, 以GB为单位。公开的数据集以粗体表示, 确定的数据以斜体表示。Common Crawl数据集过滤之前为45T

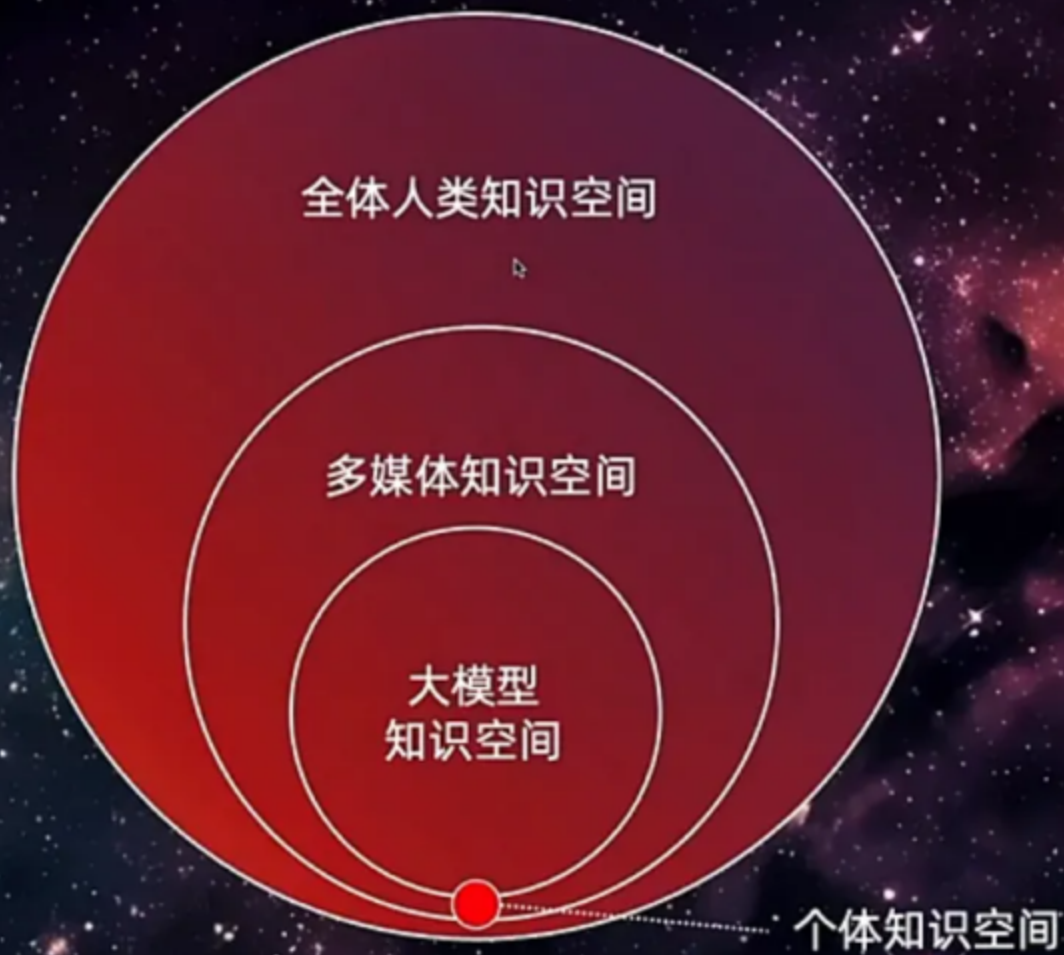
DeepSeek 背后的数据：营养搭配+错题训练+循序渐进



对比维度	GPT-3	DeepSeek-V3
发布时间	2020年6月	2024年12月
训练数据量	3000亿token	14.8万亿token
参数量	175B（密集架构）	671B（MoE架构）
训练成本	\$12M	\$5.57M
激活参数量	175B（全激活）	37B（5.5%激活率）
主要数据类型	通用互联网文本（含代码/数学）	强化代码/数学的高质量数据
官方来源	OpenAI技术论文	HuggingFace模型卡



大模型带来大知识





三体人如何学习地球知识？

电磁波

地球文明信息

原始数据、无目标性

探测器

探测器搜集数据集

精准观察、查漏补缺

拯救派

人类反馈实现对齐

动态反馈、价值博弈



三体人如何学习地球知识？

电磁波

地球文明信息
原始数据、无目标性

原始数据基建

探测器

探测器搜集数据集
精准观察、查漏补缺

定向能力强化

拯救派

人类反馈实现对齐
动态反馈、价值博弈

价值观校准





三体人如何学习地球知识？

电磁波

基础模型

DeepSeek-V3-Base

探测器 + 拯救派

强化学习

Reinforcement learning / GRPO

数学题自动判分 / 编程题实战演练 / 阶梯进化

DeepSeek-R1-Zero的训练过程
结合准确性奖励（数学、编程等任务的可验证结果）和格式奖励（强制输出结构化标签），通过GRPO算法优化模型
<https://huggingface.co/papers/2501.12948> 2025年1月22日



三体人如何学习地球知识？

电磁波

预训练

Pre-training

探测器

监督微调

Supervised Fine-Tuning
(SFT)

拯救派

人类反馈强化学习

Reinforcement learning
human feedback (RLHF)



三体人如何学习地球知识？

电磁波

基础模型

DeepSeek-V3-Base

探测器 + 拯救派

强化学习

Reinforcement learning / GRPO

数学题自动判分 / 编程题实战演练 / 阶梯进化

DeepSeek-R1-Zero的训练过程
结合准确性奖励（数学、编程等任务的可验证结果）和格式奖励（强制输出结构化标签），通过GRPO算法优化模型
<https://huggingface.co/papers/2501.12948> 2025年1月22日

强制输出过程

`<think> </think>` 标签内
是推理过程

`<answer> </answer>` 是
最终内容

为强制思考，模型第一个
输出单词必定为`<think>`

A conversation between User and Assistant.

The user asks a question, and the Assistant solves it.

The assistant first thinks about the reasoning process in the mind and then provides the user with the answer.

The reasoning process and answer are enclosed within `<think> </think>` and `<answer> </answer>` tags, respectively, i.e., `<think> reasoning process here </think> <answer> answer here </answer>`.

这是用户和助手之间的对话。

用户提出一个问题，助手解决它。

助手首先在脑海中思考推理过程，然后为用户提供答案。

推理过程和答案分别包含在 `<思考></思考>` 和 `<回答></回答>` 标签中。即：

`<思考>` 推理过程在这里 `</思考>`

`<回答>` 在这里回答 `</回答>`

Aha moment 顿悟时刻

人类长期思考后的灵感闪现
强化学习中的奖励信号引导

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...

$$(\sqrt{a - \sqrt{a + x}})^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

Table 3 | An interesting "aha moment" of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

DeepSeek-R1-Zero的训练过程

结合准确性奖励（数学、编程等任务的可验证结果）和格式奖励（强制输出结构化标签），通过GRPO强化学习策略后训练V3模型

<https://huggingface.co/papers/2501.12948> 2025年1月22日

实例：追及问题

求是小学组织去距离90公里的博物馆春游，全班同学8:00从学校坐大巴车出发。班主任老师因为有事，8:10自己自驾小车以大巴车快 $\frac{1}{3}$ 的速度追赶，结果比大巴车提前20分钟到。问：

- (1) 大巴和小车的速度各是多少？
- (2) 班主任老师追上大巴的地点距离博物馆还有多远？

你好，我是 DeepSeek-V3-241226

你可以试着问我

你能做什么？

写个故事

世界第一高山是什么？

求是小学组织去距离90公里的博物馆春游，全班同学8:00从学校坐大巴车出发。班主任老师因为有事，8:10自己自驾小车以大巴车快 $\frac{1}{3}$ 的速度追赶，结果比大巴车提前20分钟到。问：

- (1) 大巴和小车的速度各是多少？
- (2) 班主任老师追上大巴的地点距离博物馆还有多远？

开启联网 90 / 65536 tokens



思维链 Chain of Thoughts

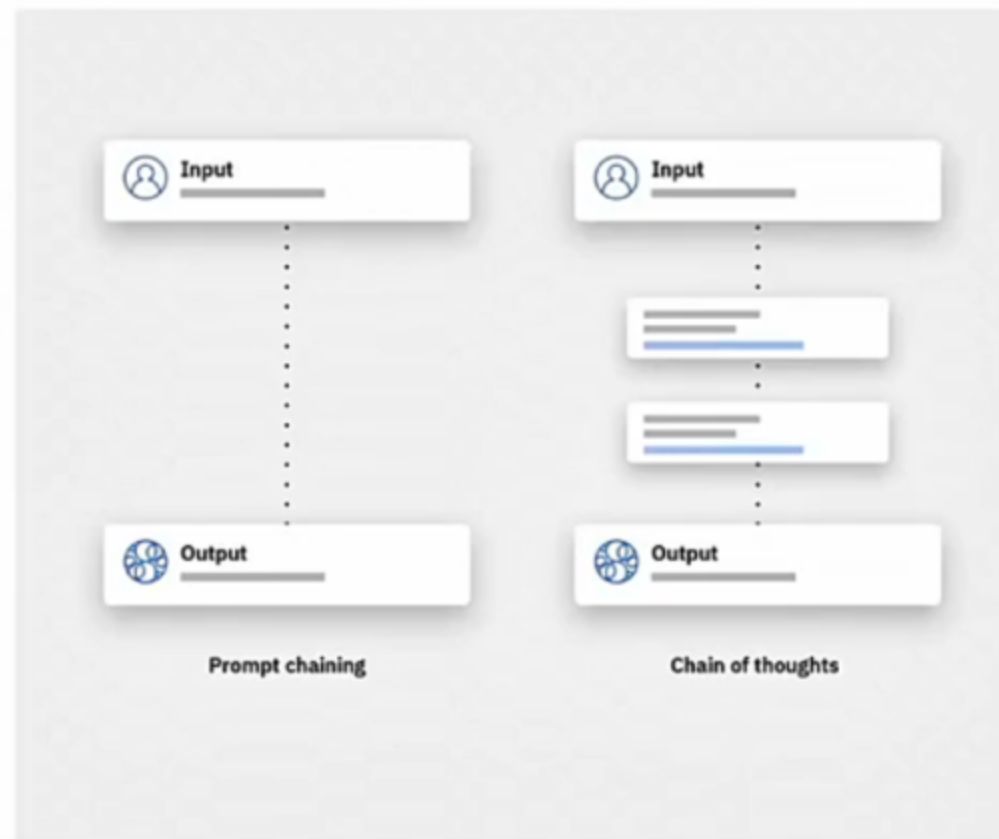
解题步骤更重要

思维链 (Chain of Thought, CoT)

- 通过模拟人类逐步推理过程来提升人工智能模型复杂任务处理能力的技术
- 核心是将问题拆解为多个中间步骤，引导模型生成逻辑链条，从而增强推理的准确性和可解释性

基本原理

- 分布推理：思维链通过将复杂问题分解为更小的子问题，逐步构建答案
- 显示中间过程：与传统直接输出答案的方式不同，CoT要求模型引入推导过程



思维链 Chain of Thoughts

新对话

介绍下deepseek的意义吧



已深度思考 (用时 19 秒) ^

DeepSeek R1

首个将思维链显式展示的开源模型

好的，我现在要介绍一下DeepSeek的意义。首先，我需要确认DeepSeek具体指的是什么。可能有不同的实体或项目使用这个名字，比如公司、产品或技术等。根据用户的问题，可能指的是深度求索（DeepSeek）这家中国的公司，专注于AGI（通用人工智能）领域的研发。

接下来，我需要收集关于深度求索（DeepSeek）的基本信息。根据之前的知识，这家公司成立于2023年，目标是实现AGI，并且已经推出了多个模型，比如DeepSeek-R1和DeepSeek-Chat。他们的模型在数学和代码能力上表现不错，而且开源了部分模型，如DeepSeek-MoE-16b。此外，他们可能还参与了多个应用场景，比如智能助手、教育、医疗等。

现在要分析其意义。首先，从技术层面来看，DeepSeek在推动大模型技术的发展，尤其是在多模态和MoE架构上的探索。其次，开源模型促进了技术共享和社区发展。再者，作为中国在AGI领域的参与者，DeepSeek可能对全球AI竞争格局有影响，推动本土技术创新，减少对外依赖。另外，他们的应用场景解决实际问题，比如教育、医疗，有社会价值。还有，推动AGI发展本身具有长远意义，可能带来技术革命。

需要确认信息的准确性，比如成立时间、模型名称等是否正确。可能存在的误区是，是否真的有开源模型以及模型的具体表现。需要查证。另外，AGI的定义和当前进展是否匹配，DeepSeek是否真的在AGI方向有实质性突破，还是更专注于当前AI技术。

可能的遗漏点：DeepSeek与其他公司的区别，比如与OpenAI、Google的差异，或者在中国市场的定位。另外，商业模式的创新，或者合作伙伴关系，是否在教育、医疗领域有实际合作案例。

总结来说，DeepSeek的意义可能包括技术创新、开源贡献、推动中国AGI发展、应用场景落地以及促进AI伦理和安全发展。需要将这些点组织成结构化的回答，分点阐述，并确保信息准确。

思维链 Chain of Thoughts



对比维度	白盒派	黑箱派
核心特征	强制展示推理过程链 支持步骤回溯	仅输出最终结果 决策过程不可见
典型代表	DeepSeek V3 IBM Explainable AI	OpenAI GPT o1 Meta Llama 3
优势	<ul style="list-style-type: none">- 错误可追溯- 合规性强- 调试效率高	<ul style="list-style-type: none">- 计算效率高- 商业保密性强- 处理开放性任务更灵活
缺陷	<ul style="list-style-type: none">- 推理延迟增加- 部分复杂任务表现受限	<ul style="list-style-type: none">- 存在"聪明汉斯"效应（正确结论错误归因）- 合规成本高（欧盟AI法案罚款风险）
关键技术	链式思维提示（Chain-of-Thought） 动态置信度标注	黑盒优化 隐式知识蒸馏
适用场景	教育、医疗诊断、金融风控	创意生成、游戏NPC、舆情分析
伦理风险	过度透明可能导致模型被逆向工程	决策偏见难追溯（如招聘AI的性别倾向性）

Grok 3 发布

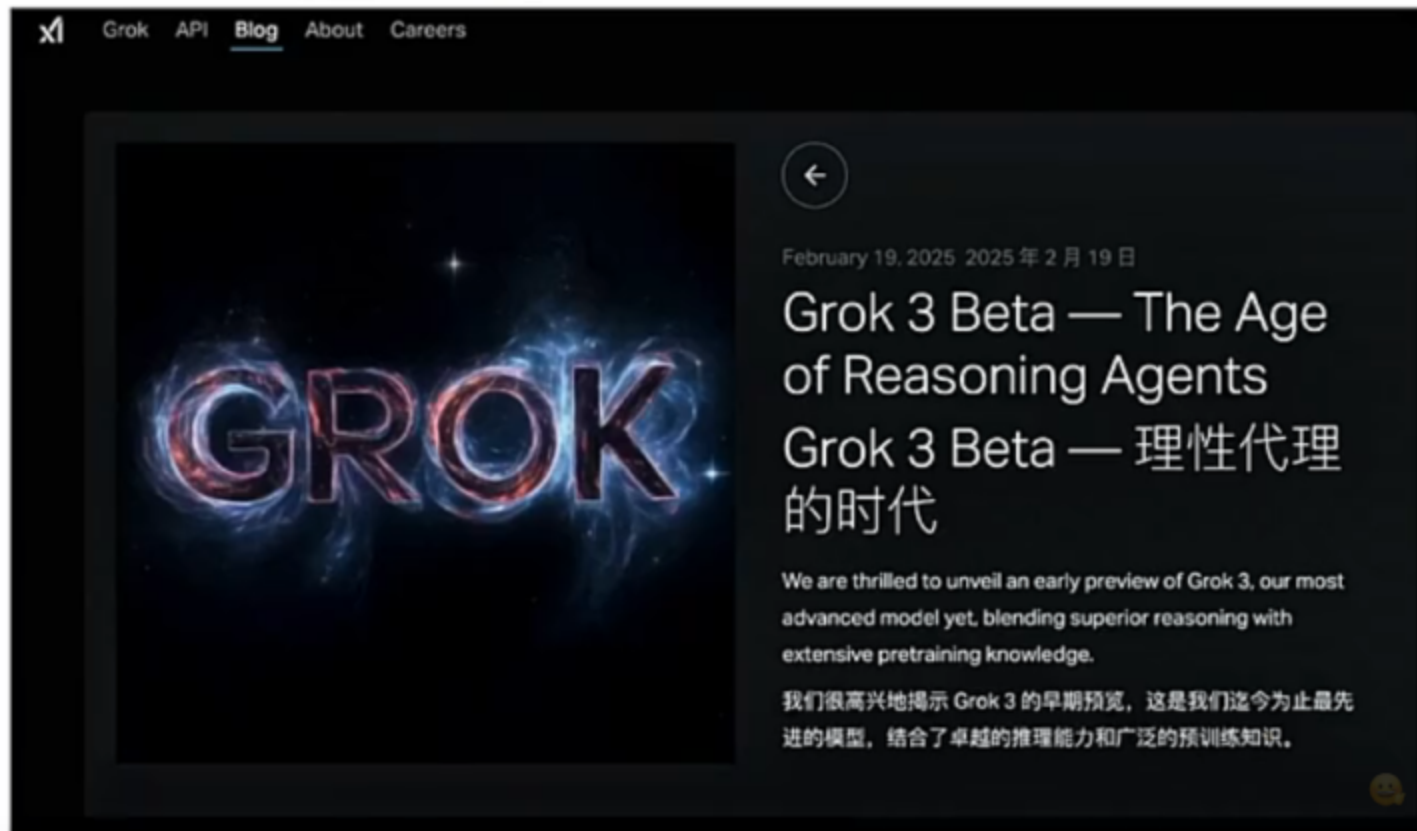
马斯克发布的大模型

2月18日正式发布

- 通过X平台直播
- 展示火星轨道计算、游戏开发等实际应用场景
- 计算能力为Grok 2的10倍
- 展示深度搜索

2月20日：开放免费试用（至服务器超负荷为止）

<https://x.ai/blog/grok-3>



Grok 3 发布

xAI 的定制超级计算机

世界最大的AI训练集群之一

第一阶段：仅 122 天内完成，
部署了 100,000 个 H100 GPU

在第二阶段：在另外 92 天内将
计算能力翻倍。

允许持续训练，这意味着 Grok
3 在更多用户与其互动时仍在
实时改进



Grok 3 发布

对比维度	Grok 3 (xAI)	DeepSeek V3 (DeepSeek)	参考资料
算力路线	20万H100 GPU集群	2000颗H800 GPU	xAI发布会 , DeepSeek白皮书
训练成本	估算2.1亿 / 5亿美元, 含基础设施	558万美元 (单次训练成本)	Axios成本报告
技术特性	2.7万亿参数 动态架构通过Think Mode分解问题 (如火星轨道计算调用9个子模型协同)	671B参数 采用MoE架构 仅激活少量参数完成同类任务	xAI技术文档 , DeepSeek-V3技术表
数学推理	AIME 52%	AIME 39.2%	xAI基准测试 , DeepSeek-V3数据表
代码生成	LiveCodeBench 57%	LiveCodeBench 40.5%	Radargit性能对比
训练能耗	相当于30万户家庭年用电	相当于于3000户家庭年用电	xAI发布会实录
推理成本	\$0.015/次	\$0.0018/次	Radargit成本对比
生态策略	封闭生态 X平台, 但提供实时数据接口	开源策略 允许开发者定制思维链展示规则	Radargit行业分析

推理模型与非推理模型

对比维度	非推理模型 Non-reasoning Models	推理模型 Reasoning Models
典型产品	DeepSeek-V3 GPT1-4全系列 / Meta LLaMA3 / Gemini Pro / Grok 2	DeepSeek-R1 GPT-o1 / 通义千问QwQ-32B / 天工Skywork o1 / 讯飞星火X1 / Gemini Ultra / Grok 3
侧重点	<ul style="list-style-type: none">- 语言模式匹配与统计规律学习- 上下文连贯性与生成流畅度- 大规模数据泛化	<ul style="list-style-type: none">- 复杂逻辑分析与多步骤推理- 符号化规则与结构化思维链构建- 逻辑验证准确性
核心特点	<ul style="list-style-type: none">- 基于Transformer的并行化生成- 端到端模式识别- 依赖提示工程提升表现	<ul style="list-style-type: none">- 思维链 (CoT)、思维树 (ToT) 等技术增强- 原子推理模块组合 (如批判性思维、逐步推导)- 需强化学习优化推理路径
典型应用	<ul style="list-style-type: none">- 开放域对话 (客服机器人)- 文本创作 (小说/诗歌生成)- 语义分类 (情感分析)	<ul style="list-style-type: none">- 数学定理证明 (如代数方程求解)- 动态规划算法设计- 实时决策系统 (自动驾驶)

DeepSeek模型本身的贡献（V3为例）



类别	贡献方向	具体技术/成果
训练成本	硬件成本	<ul style="list-style-type: none">- 采用MoE架构（混合专家模型），仅激活部分参数（如V3激活37B/671B参数）- FP8混合精度训练减少内存占用与计算量- 训练成本显著降低（如V3成本557.6万美元，仅为Llama3的7%）
	算力利用率	<ul style="list-style-type: none">- DualPipe双向流水线优化计算与通信重叠- 跨节点MoE通信优化实现全计算通信折叠- 自适应计算分配机制提升GPU利用率
	流程周期	<ul style="list-style-type: none">- 强化学习直接用于基础模型（无需监督微调）- 知识蒸馏技术缩短后训练周期- 动态知识库与多模态引擎加速迭代
算法架构	训练范式	<ul style="list-style-type: none">- 多令牌预测（MTP）提升数据效率与推理速度- FP8混合精度框架首次验证超大规模模型可行性- 负载均衡策略减少性能损失
	突现能力	<ul style="list-style-type: none">- 数学推理（Math 7b接近GPT-4水平）- 自我验证与长思维链生成（R1-Zero模型）- 代码生成超越GPT4-Turbo（Coder-V2）
	模型压缩	<ul style="list-style-type: none">- 知识蒸馏将大模型能力迁移至小模型（如15B参数模型）- MLA机制压缩KV缓存93.3%- 动态稀疏注意力降低计算复杂度

注：使用秘塔搜索（R1推理模型）总结，内容未经人工确认，格式经人工调整。详情参考：<https://metaso.cn/s/ZYdtday>



帮我总结下deepseek对于大模型领域的贡献。初步想法是：训练成本：1硬件成本，2算力利用率，3流程周期；算法架构：1训练范式，2突现能力，3模型压缩。请确认上述内容是否合理。最终给我总结表。注意来源主要应该是：近2年的学术文献、权威的科技博客。

思考中...

步骤拆解 ^

好的，我现在需要帮用户分解关于总结DeepSeek在大模型领域贡献的问题。用户已经给出了初步的想法，分为训练成本和算法架构两个部分，每个部分有三个子点。我需要确认这些分类是否合理，并最终生成总结表。

首先，用户提到的训练成本包括硬件成本、算力利用率和流程周期。这部分可能需要搜索DeepSeek在硬件优化、算力效率提升以及训练流程优化方面的具体措施或成果。比如是否有论文提到他们如何降低硬件成本，或者提高算力利用率的技术。

然后是算法架构部分，涉及训练范式、突现能力和模型压缩。这里需要查找DeepSeek在训练方法上的创新，比如是否采用了新的训练策略，或者在模型压缩方面有什么突破。突现能力可能指的是模型在训练过程中出现的意外能力，比如零样本学习等，需要确认是否有相关研究。

接下来，用户要求来源主要是近两年的学术文献和权威科技博客。因此，每个子问题都需要通过搜索来获取最新的信息。例如，搜索DeepSeek在硬件成本方面的研究，或者他们在模型压缩方面的论文。

需要注意的是，用户希望最终得到一个总结表，所以分解的问题需要覆盖所有提到的方面，并确保每个部分都有足够的信息支撑。可能还需要分析这些搜索结果，综合起来形成

现在，分解步骤应该是：

视频案例网址：

<https://metaso.cn/s/TM0D0r>



政治局会议首提AGI

2023年4月28日，中共中央政治局召开会议分析研究当前经济形势和经济工作：

要巩固和扩大新能源汽车发展优势，加快推进充电桩、储能等设施建设和配套电网改造。

要重视通用人工智能发展，营造创新生态，重视防范风险。

CCTV 13
新闻

中共中央政治局召开会议 分析研究当前经济形势和经济工作 中共中央总书记习近平主持会议

汽车发展优势，加快推进充电桩、储能等设施建设和配套电网改造。**要重视通用人工智能发展，营造创新生态，重视防范风险。**

会议强调，恢复和扩大需求是当前经济持续回升向好的关键所在。积极的财政政策要加力提效，稳健的

新闻联播
XINWENLIANBO

通用人工智能

人类能力标尺

1. 能力而非过程
2. 广度和深度
3. 认知和元认知任务
4. 潜力而非部署
5. 生态效度
6. 关注通向AGI的路径

来源: Google DeepMind团队, Levels of AGI:
Operationalizing Progress on the Path to AGI
<https://arxiv.org/pdf/2311.02462.pdf>
2023年11月4日

Performance (rows) x Generality (columns)	Narrow	General
	clearly scoped task or set of tasks	wide range of non-physical tasks, including metacognitive abilities like learning new skills
Level 0: No AI	Narrow Non-AI calculator software; compiler	General Non-AI human-in-the-loop computing, e.g. Amazon Mechanical Turk
Level 1: Emerging <i>equal to or somewhat better than an unskilled human</i>	Emerging Narrow AI GOFAI ⁴ ; simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	Emerging AGI ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023)
Level 2: Competent <i>at least 50th percentile of skilled adults</i>	Competent Narrow AI toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLI (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	Competent AGI not yet achieved
Level 3: Expert <i>at least 90th percentile of skilled adults</i>	Expert Narrow AI spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	Expert AGI not yet achieved
Level 4: Virtuoso <i>at least 99th percentile of skilled adults</i>	Virtuoso Narrow AI Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016, 2017)	Virtuoso AGI not yet achieved
Level 5: Superhuman <i>outperforms 100% of humans</i>	Superhuman Narrow AI AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	Artificial Superintelligence (ASI) not yet achieved

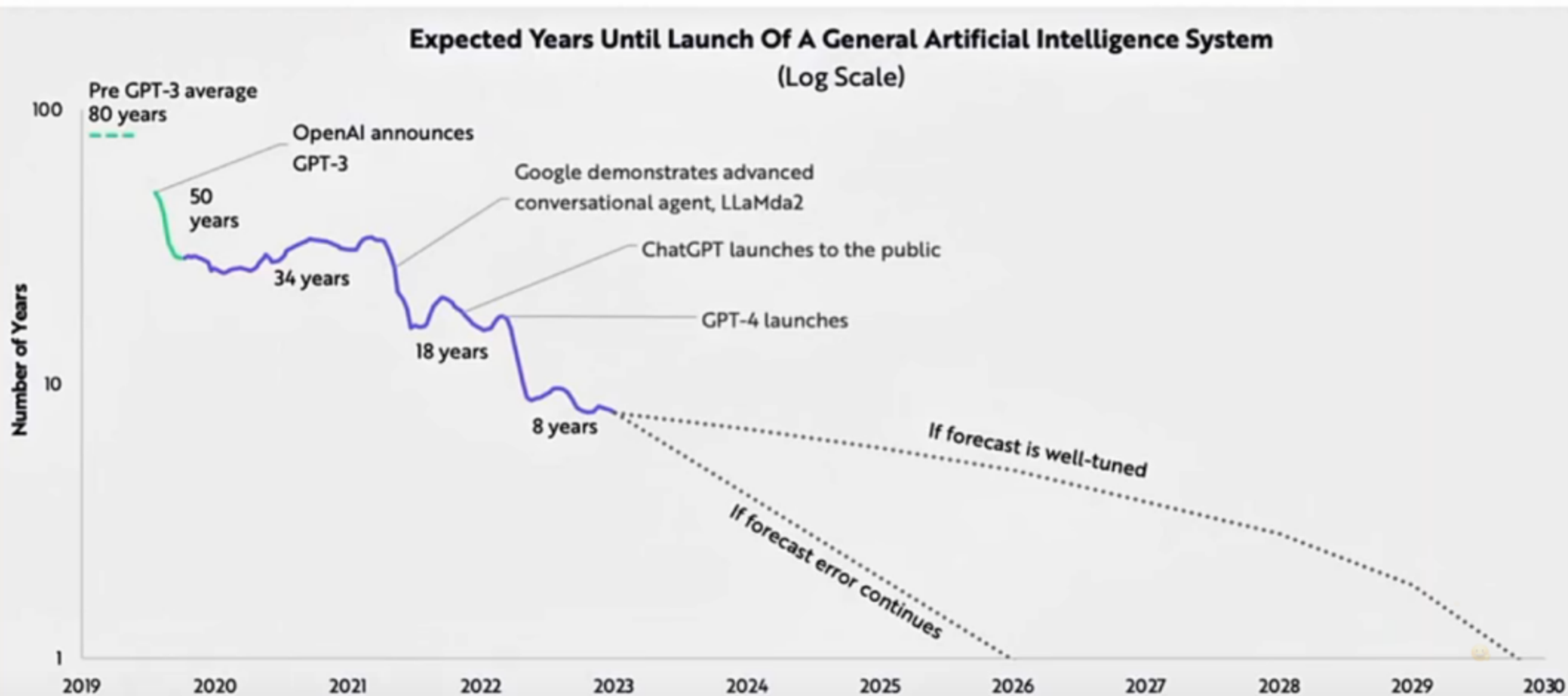
任务复杂度标尺

OpenAI 2024年7月公布
用于追踪人工智能向人工通用智能（AGI）发展的进程
将AI能力分为从基础对话到全面组织管理的不同阶段
没有纠结是否真正理解世界
关注在场景中的能力

等级	能力特征	典型应用场景	当前进展
L1	聊天达人（自然语言交互）	ChatGPT等聊天机器人	已实现
L2	解题高手（博士级问题解决）	复杂数据分析	部分达到
L3	持久管家（多天自主决策）	自动化商业流程	开发中
L4	创新引擎（原创性发明）	药物研发/科学突破	理论阶段
L5	组织大脑（企业级系统替代）	战略规划与资源分配	远期目标

来源： OpenAI's 5 Levels Of 'Super AI' (AGI To Outperform Human Capability) <https://www.forbes.com/sites/jodiecook/2024/07/16/openais-5-levels-of-super-ai-agi-to-outperform-human-capability/> 2024年7月16日

通用人工智能：能力涌现不可逆+社会影响倒计时





能力涌现不可逆 社会影响倒计时 人机协作新常态





浙江大学
ZHEJIANG UNIVERSITY

一、智能演变

二、人机协作

三、产业现状

四、教育成长



人机协作加速社会发展

新工作机会和体面工作

诺贝尔经济学奖得主 克里斯托弗·皮萨里德斯
较长的转型过渡期

- AI融入生产过程不会那么迅速
- 技能提升是关键
- 大部分人要做的是技能提升，即学会使用AI
- 新工作机会出现
- 工作毁灭 (Job Destruction)
- 工作创造 (Job Creation)

ChatGPT opens door to four-day week, says Nobel Prize winner

Tom Rees

Apr 6, 2023 - 9:18am



London | The [ChatGPT revolution](#) opens the door to a four-day week by providing a big productivity boost for swathes of jobs, according to a Nobel Prize-winning labour economist.

Christopher Pissarides – a professor at the London School of Economics who specialises in the impact of automation on work – said the labour market could adapt quickly enough to artificial intelligence-backed chatbots. His remark tamped down concerns that rapid advances in technology could bring mass job losses.



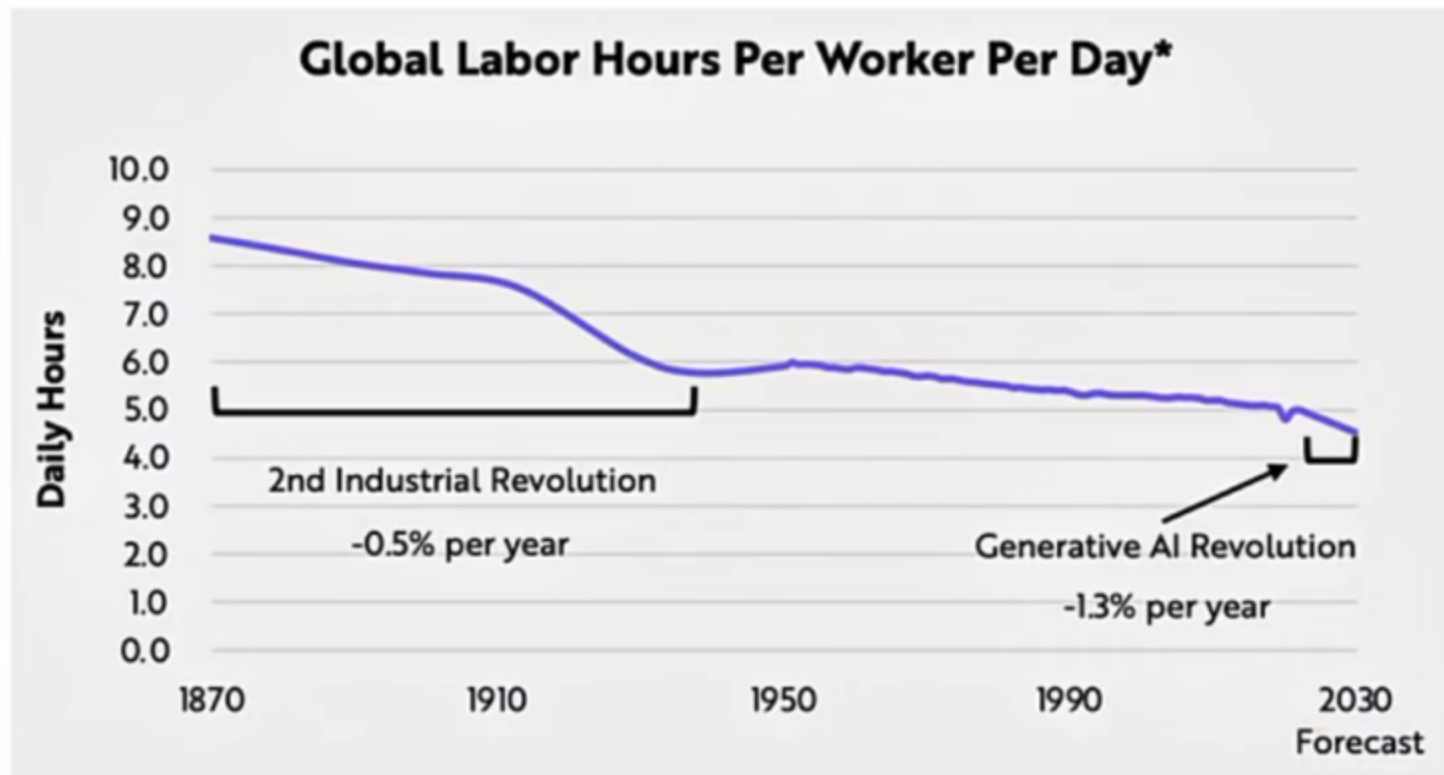
人机协作加速社会发展

日均工作时间下降

从第二次工业革命到二战结束的80年间，全球每个工人的劳动时间每年减少0.5%

生成式人工智能可以将每个工人的平均劳动时间降低1.3%

从2022年每天5小时到2030年的4.5小时



实例：人机协作 - 机器人设计

来源: Stella, F., Della Santina, C. & Hughes, J. How can LLMs transform the robotic design process. Nature Machine Intelligence. 5, 561-564 (2023). <https://doi.org/10.1038/s42256-023-00669-7> 2023年6月7日

a

"One way to build a gripper for handling crops would be to use **silicone or rubber**, which are both soft and flexible, helping to reduce the risk of crushing the crops"

"Overall, a **Dynamixel motor** is suitable for actuating a gripper for picking tomatoes"



"One possible shape for a gripper that is designed to grasp tomatoes could be a **round or oval shape** that is slightly larger than the tomato"

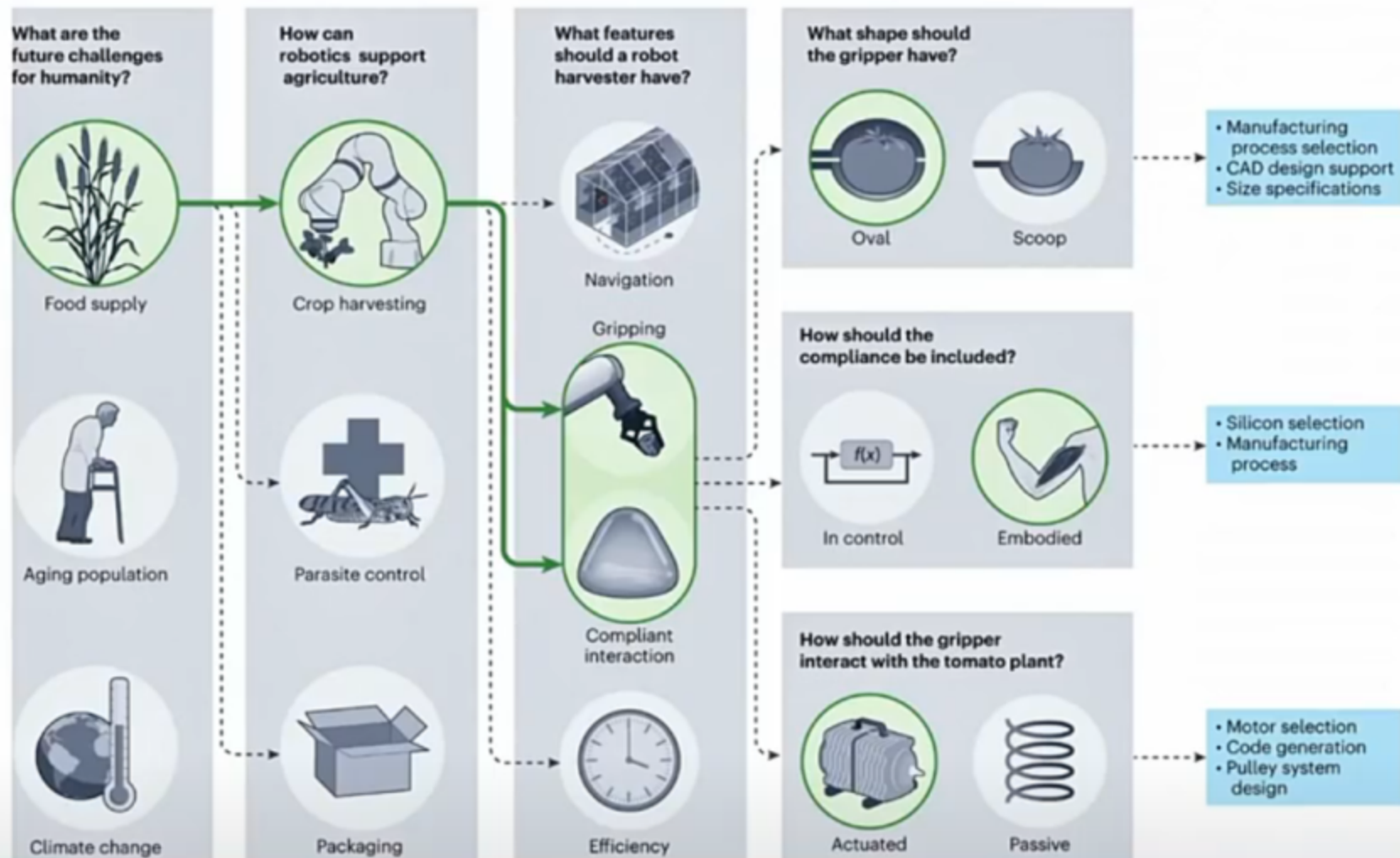
"One way to connect the oval shape of a gripper to a motor would be to use a gear or **pulley system**. This could involve attaching a pulley to the output shaft of the motor, and then connecting the gripper to the pulley **using a linkage** such as a belt, chain, or **cable**"

b



实例：人机协作 - 机器人设计

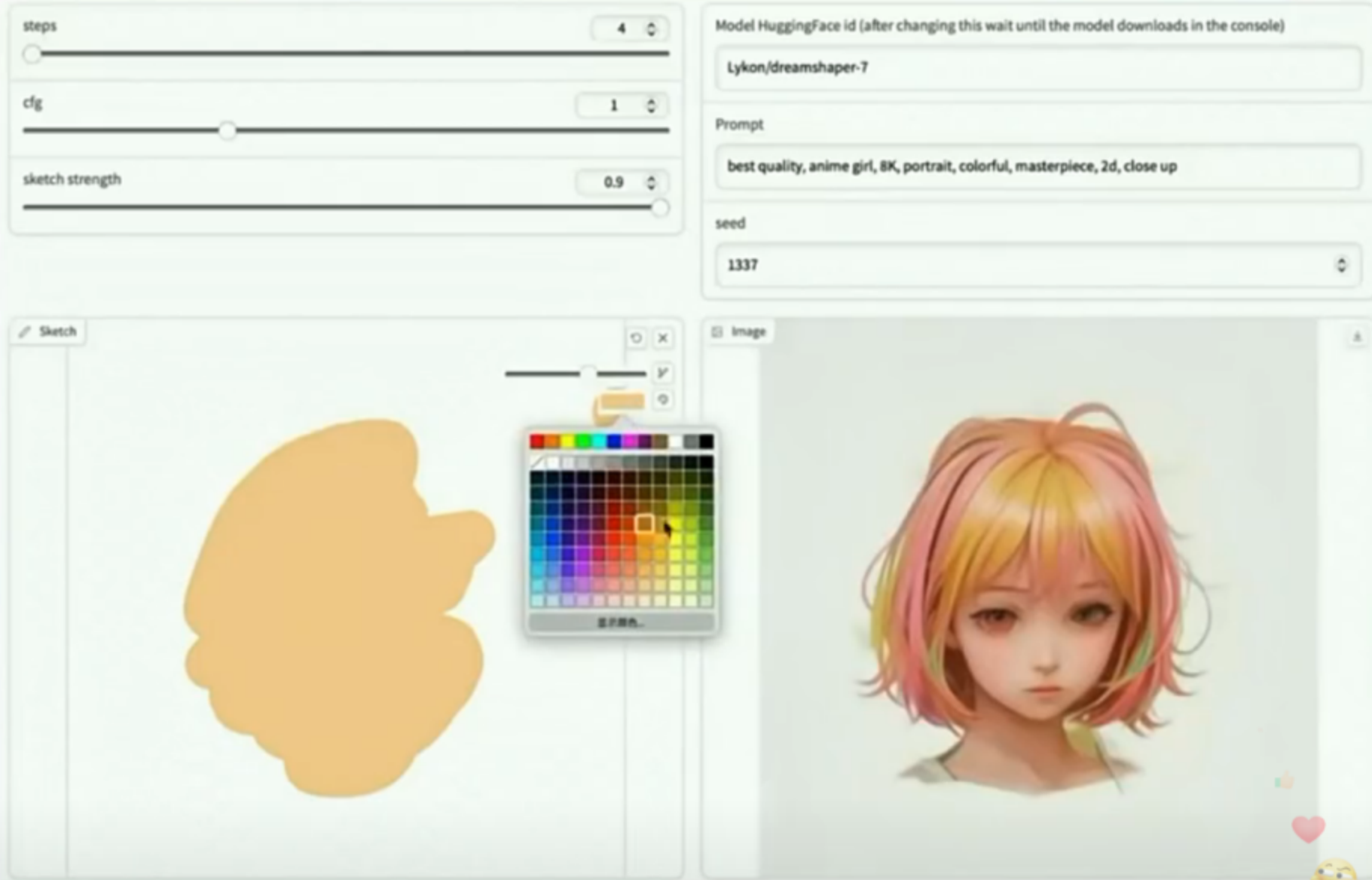
来源：Stella, F., Della Santina, C. & Hughes, J. How can LLMs transform the robotic design process. Nature Machine Intelligence. 5, 561-564 (2023). <https://doi.org/10.1038/s42256-023-00669-7> 2023年6月7日



实例： LCM

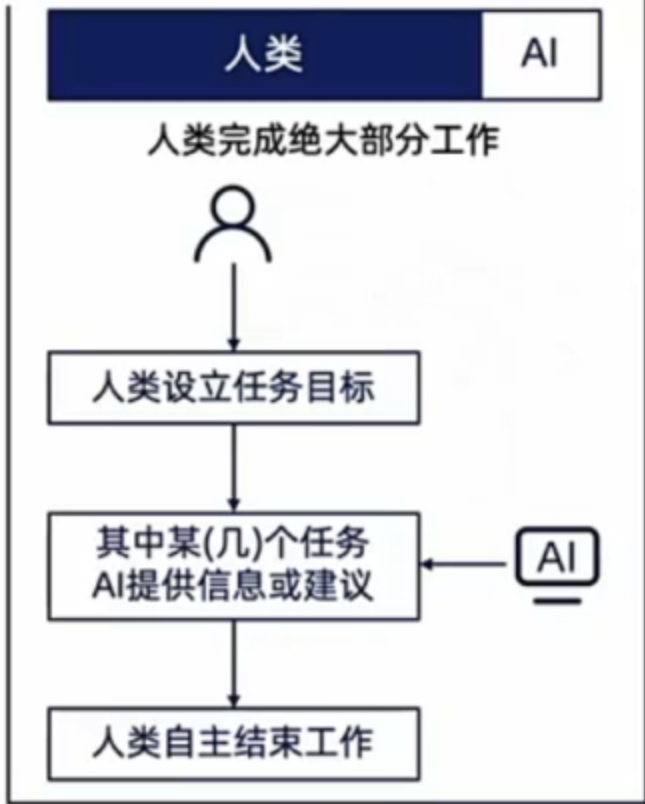
Latent
Consistency
Models
潜在一致性模型

来源：
https://twitter.com/gong_cn/status/1726501516285264303
https://huggingface.co/blog/lcm_lora
2023年11月9日



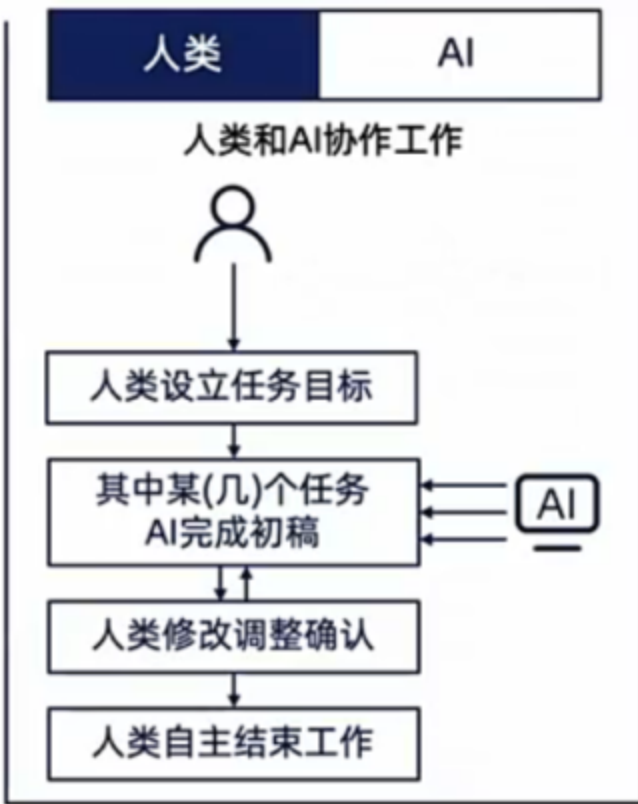
三种协作模式

Embedding模式



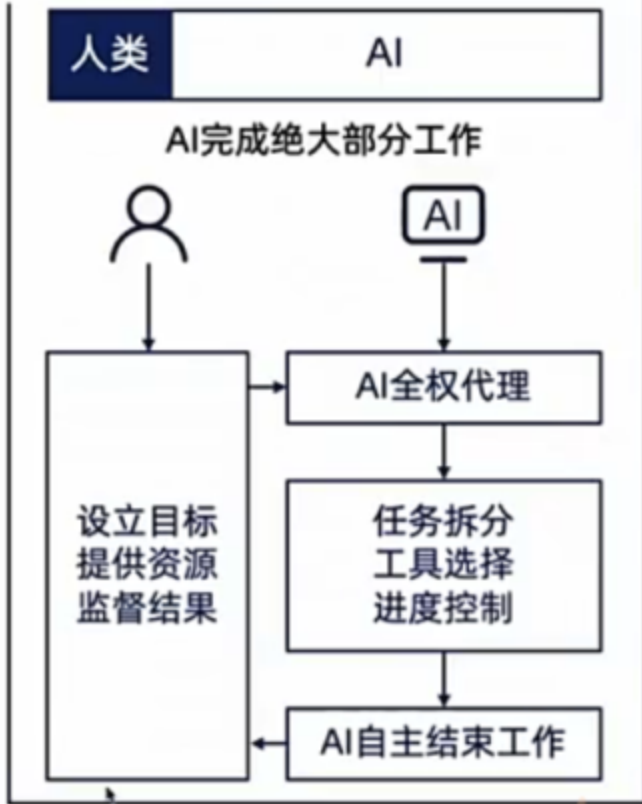
小助理

Copilot模式



副驾驶

Agents模式



代理人

AI的供给已经成为社会基础设施

第一次工业革命

蒸汽时代 (1760-1840)

水力、纺织、铁

第二次工业革命

电气时代 (1840-1950)

蒸汽、铁路、钢

电、化学品、内燃机

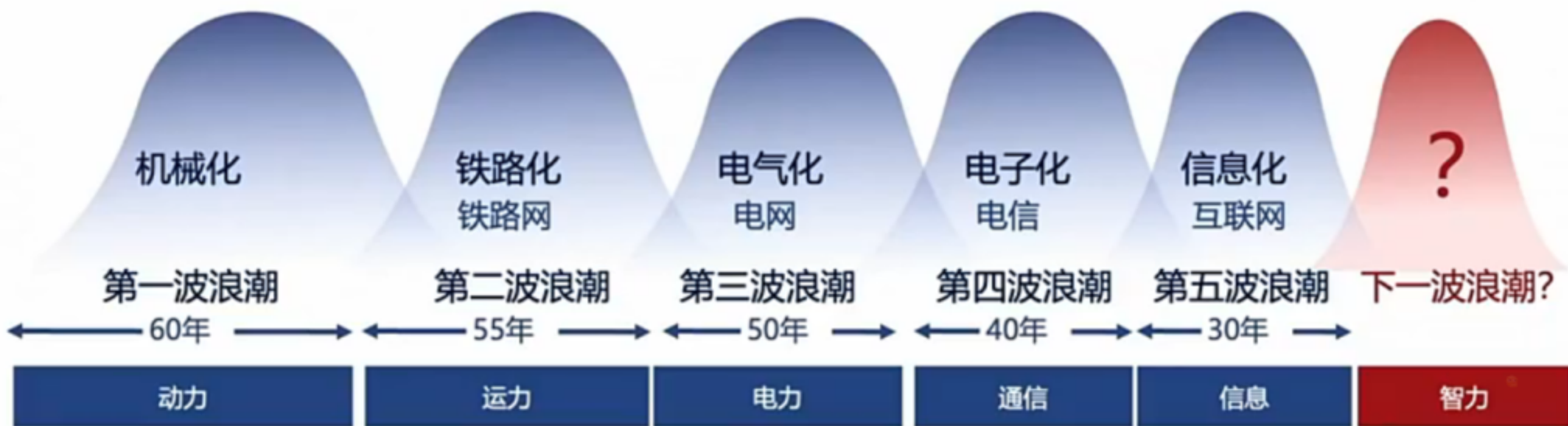
第三次工业革命

信息时代 (1950-?)

石油、天然气、硅电子

互联网

以人工智能为代表的
新型基础设施





浙江大学
ZHEJIANG UNIVERSITY

- 一、智能演变
- 二、人机协作
- 三、产业现状**
- 四、教育成长



1 基础模型 高速迭代

AI大模型发布情况

斯坦福大学 2024 AI Index Report

- 数据截至2024年1月31日
- 全球发布149个大模型（其中中国20个）

根据LifeArchitect.ai（持续更新）

- 2025年2月20日
- 观测534个大模型（其中中国117个）

来源:

Artificial Intelligence Index Report 2024

<https://aiindex.stanford.edu/report/>

Models Table Rankings <https://lifearchitect.ai/models-table/>

2025年2月

#	Model	Lab	Date	MLU
1	o1	OpenAI	Dec/2024	92.3
2	DeepSeek-R1	DeepSeek-AI	Jan/2025	90.8
3	Claude 3.5 Sonnet (new)	Anthropic	Oct/2024	90.5
4	R1 1776	Perplexity	Feb/2025	90.5
5	Sonus-1 Reasoning	Rubik's AI	Jan/2025	90.15
6	Hunyuan-Large	Tencent	Nov/2024	89.9
7	Doubao-1.5-pro	ByteDance	Jan/2025	88.6
8	Llama 3.1 405B	Meta AI	Jul/2024	88.6
9	MiniMax-Text-01	MiniMax	Jan/2025	88.5
10	Qwen2.5-Max	Alibaba	Jan/2025	87.9
11	Grok-2	xAI	Aug/2024	87.5
12	Kimi k1.5	Moonshot AI	Jan/2025	87.4

#	Model	Lab	Date	GPQA
1	o3	OpenAI	Dec/2024	87.7
2	Grok-3	xAI	Feb/2025	84.6
3	o1	OpenAI	Dec/2024	79
4	o3-mini	OpenAI	Jan/2025	77
5	DeepSeek-R1	DeepSeek-AI	Jan/2025	71.5
6	Sonus-1 Reasoning	Rubik's AI	Jan/2025	67.3
7	QwQ-32B	Alibaba	Nov/2024	65.2
8	Claude 3.5 Sonnet (new)	Anthropic	Oct/2024	65
9	Doubao-1.5-pro	ByteDance	Jan/2025	65
10	Gemini 2.0 Pro	Google DeepMind	Feb/2025	64.7
11	Gemini 2.0 Flash exp	Google DeepMind	Dec/2024	62.1
12	Qwen2.5-Max	Alibaba	Jan/2025	60.1

2 智能应用 蓬勃发展

AI工具迅猛增长

根据Toolify.ai统计（2025年2月20日访问）

- 23915个AI工具
- 233个分类
- 162万AI API

其中：

- 2023年8813个
- 2024年13,795个（57%）



DeepSeek开源推动传统应用智能化

AI工具迅猛增长

DeepSeek的开源策略正在加速AI技术普惠化

- 头部应用正在/即将完成深度集成（社交/搜索/推荐），聚焦：交互体验
- 企业级场景在电力、政务、医疗等垂直领域取得突破，聚焦：系统级优化

瓶颈：算力成本、数据合规性和人才储备

类型	应用/平台	集成时间	核心功能与技术特性	信息来源
消费级	百度搜索	2025年2月18日	集成DeepSeek-R1提供AI搜索服务，支持复杂问题推理和深度搜索功能	百度官方公告
消费级	微信搜索	2025年2月16日	灰度测试DeepSeek-R1，提供AI搜索入口和深度思考服务	腾讯集团官方声明
消费级	支付宝百宝箱	2025年2月11日	支持DeepSeek全系列模型，用户可构建智能体并一键发布到小程序	支付宝官方新闻
消费级	抖音内容审核	2024年11月	引入DeepSeek-V3增强多模态内容审核能力	字节跳动技术白皮书
消费级	高德地图导航	2024年10月	集成DeepSeek-LLM优化实时路径规划和拥堵预测	高德地图2024年度技术报告
企业级	南方电网	2025年2月12日	部署千亿参数电力大模型，实现故障预测和能耗优化	南方电网AI创新平台公告
企业级	华为云	2025年2月12日	推出混合云DeepSeek本地部署方案，支持全系列模型并优化推理性能	华为云官方新闻
企业级	深圳福田政务	2025年2月18日	上线70个AI数智员工，覆盖240个政务场景	深圳福田政府公告
企业级	吉利汽车	2025年2月13日	星睿大模型与DeepSeek-R1深度融合，优化车控交互	吉利汽车技术发布会
企业级	中国移动	2025年2月5日	移动云全量支持DeepSeek模型，覆盖13个智算中心	中国移动官方声明



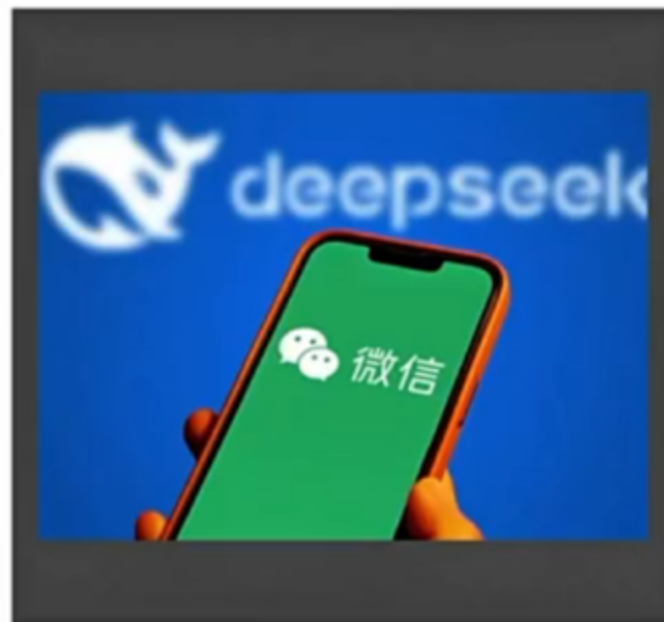
头部应用智能化



百度（网页端）



支付宝-百宝箱（网页端）



微信（搜索-AI搜索）



技巧：优化AI大模型三个路径



微调

Fine-tuning

微调

- 针对特定任务或领域定制，显著提高任务性能
- 适应特定领域的细微差别、语气或术语
- 依赖足够高质量标注数据
- 适合目标任务明确、数据相对静态的场景

训练实习生

检索增强生成

Retrieval-Augmented Generation, RAG

- 利用外部知识库增强答案
- 提供支持性的证据
- 能够动态检索相关知识，适合数据频繁更新的场景
- 成本效益高
- 更可信、更可解释、时效性和准确性高

提供项目资料

提示词工程

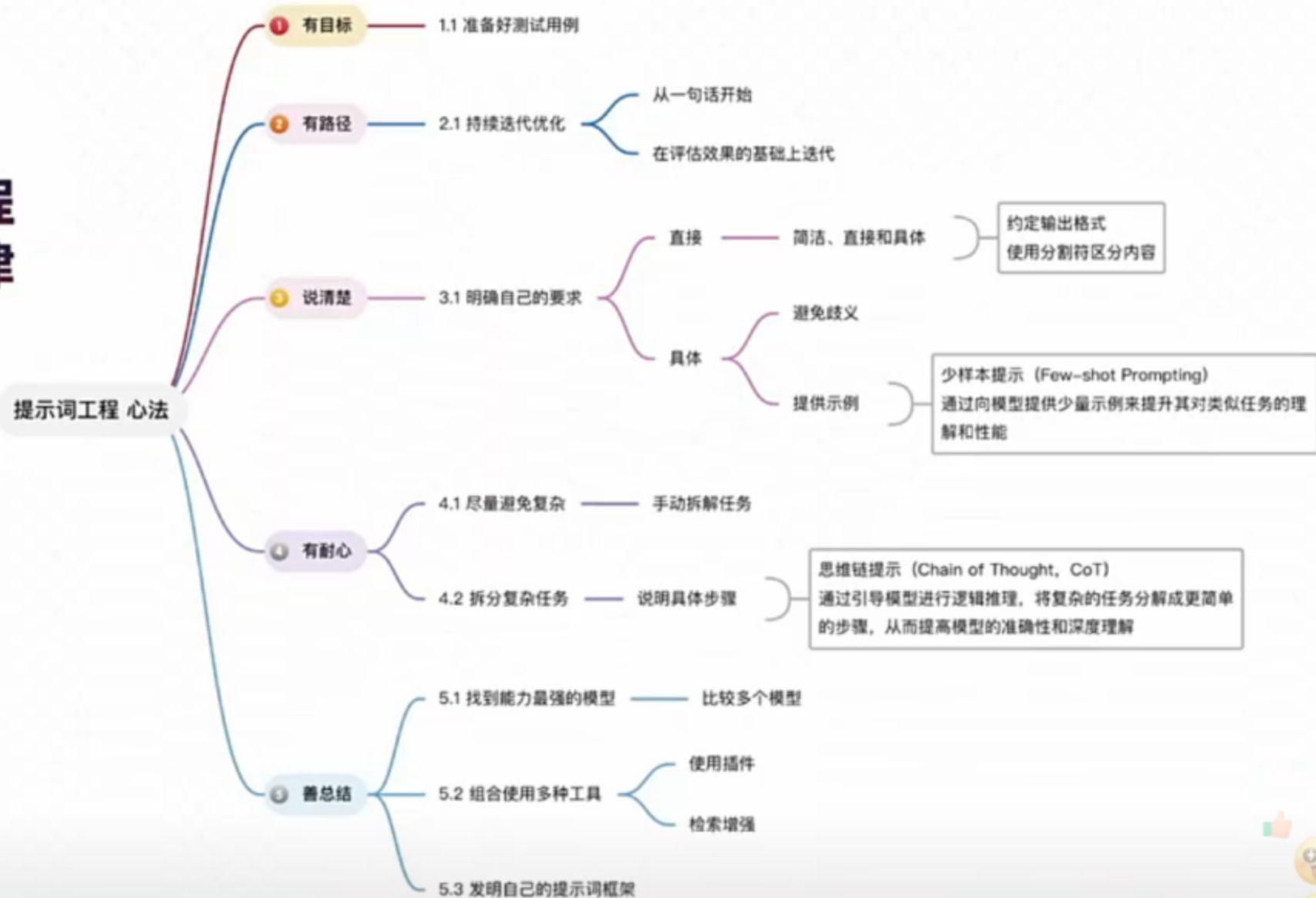
Prompt Engineering

提示词工程

- 通过精心设计输入提示，引导LLM生成期望的高质量输出
- 无需重新训练模型，不需要标注数据
- 一些提示工程框架可以显著增强LLM的推理能力
- 灵活性、使用简单

任务布置清晰

提示词工程 的经验规律



提示词依然重要

DeepSeek降低了基础提示词的使用门槛。但特定领域中，提示词可以带来显著效果提升：

- 专业领域任务
- 复杂逻辑推演
- 特定格式输出

随着AI大模型的发展，提示词工程将更侧重“精准表达”而非“技巧堆砌”，成为人机协作的语义校准工具。

需求类型	传统提示词技巧	DeepSeek适配方案
专业内容生成	角色扮演（如“扮演经济学家”）	细化场景（如“为90后设计小红书3C产品种草话术”）
逻辑推理任务	思维链（Chain-of-Thought）	内置推理框架，自动生成完整逻辑链
格式控制	严格模板约束	自然语言描述格式需求（如“用对比表格呈现”）

实例：偶然事件推理

从图像中推理事件

即使最普通的环境，也存在各种可能性。

实例：扫地机器人是否足够智能了？

使用AI大模型：

- 预测会发生什么事情
- 分析发生了什么事情
- 推理应该如何应对



你



发生什么事了？



ChatGPT

看起来扫地机器人在清洁时遇到了一堆散落的东西，它在尝试清理时反而将其推得到处都是。如果这些散落的物体是宠物排泄物，这样的情况是可能发生的，因为普通的扫地机器人不具备识别此类物质的功能。如果确实是这种情况，机器人和地板都需要被彻底清洁，以避免污染和气味的扩散。



实例：糖画机器人

糖画打印

使用AI大模型：

- 输入文字或照片
- 智能生成糖画模型
- 糖画机器人打印输出



实例：智能三维设计

智能3D打印机

DeepSeek 等AI大模型支持的
零门槛三维建模

- 手绘/文本/图片
- 二维概念图
- 三维建模
- 3D打印输出





- 一、智能演变
- 二、人机协作
- 三、产业现状
- 四、教育成长**

变化之一：效率提升



浙江大学
ZHEJIANG UNIVERSITY

全面提升教和学的效率

教师

生成教案
润色课件
出题组卷
自动阅卷

便捷搜索
快速答疑
创作辅助
实时评价

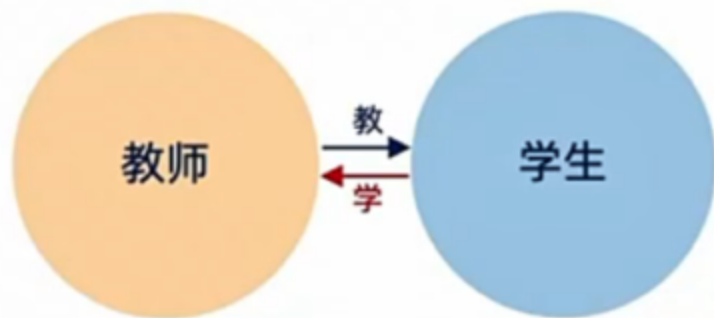
学生

变化之二：模式变革



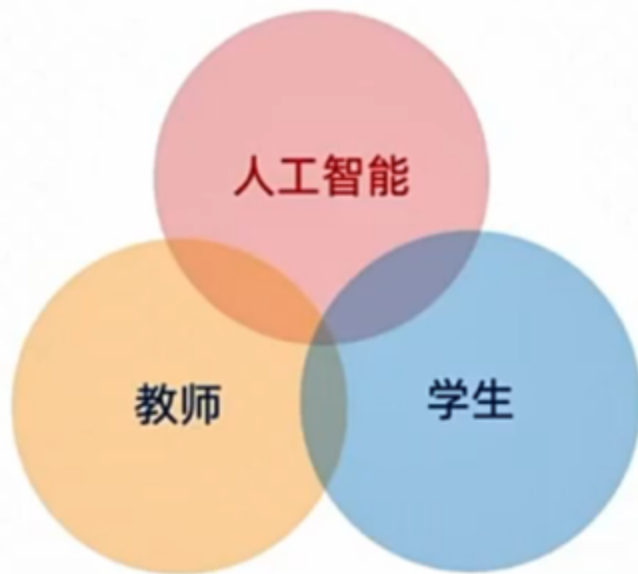
浙江大学
ZHEJIANG UNIVERSITY

实例：师生关系



变化之二：模式变革

实例：师生关系



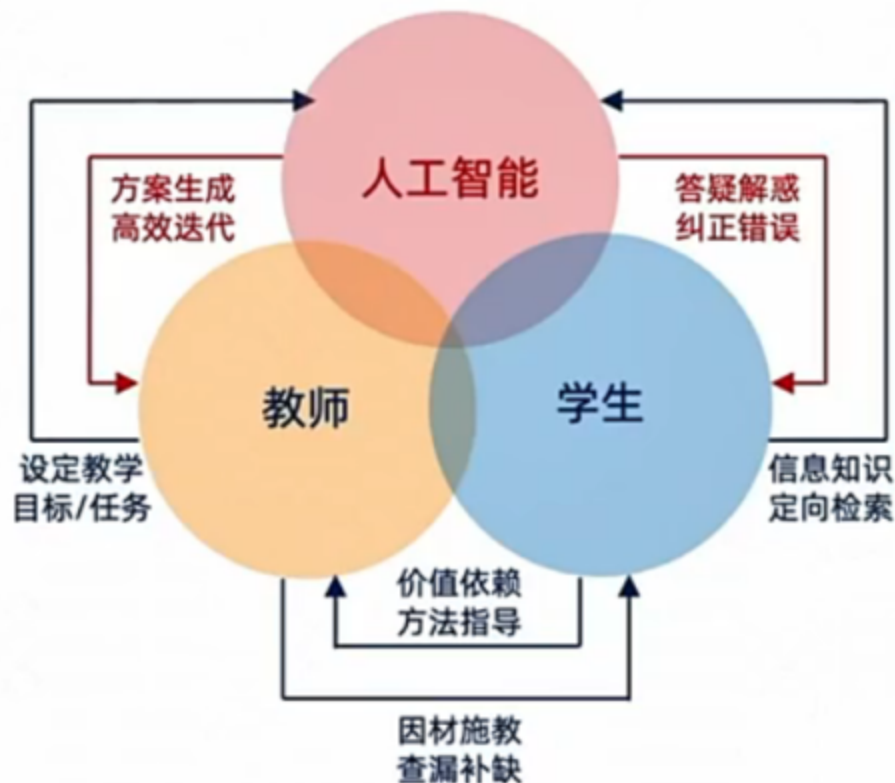
由：教师-学生 二元结构 转变为：人工智能-教师-学生 三元结构

变化之二：模式变革



浙江大学
ZHEJIANG UNIVERSITY

实例：师生关系



由：教师-学生 二元结构 转变为：人工智能-教师-学生 三元结构

变化之二：模式变革



浙大大先生 (chat.zju.edu.cn)

AI大模型支持的个性化学习、课程设计、教学管理



智海MO平台 (mo.zju.edu.cn)

AI大模型支持的人工智能课程学习

变化之三：范式重构

全面引入AI能力

替代你的不是AI，而是会用AI的人。

知识依然是教育的关键载体

知识的生产、传播、传授方式急剧变迁

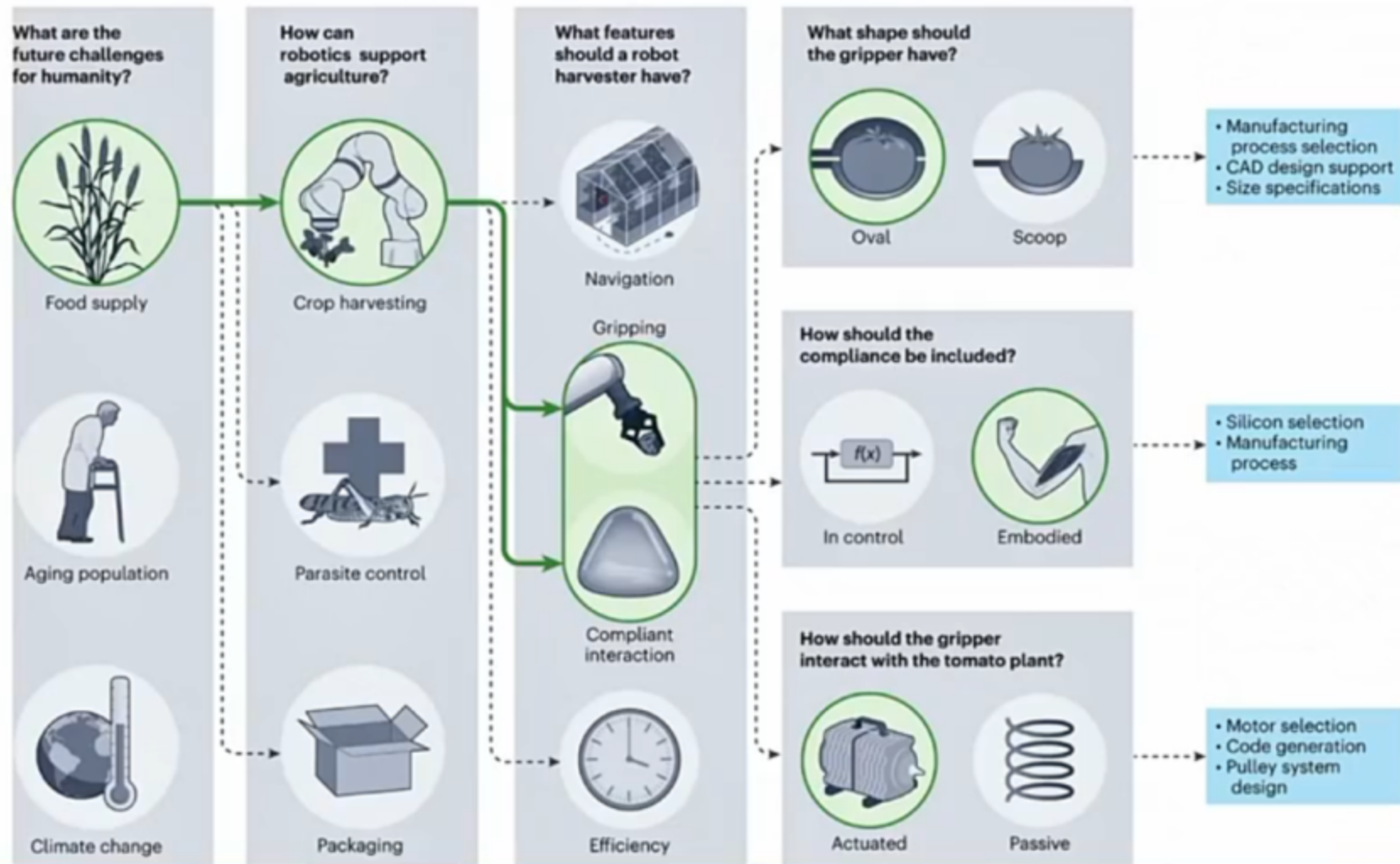
- 围绕知识的方法论、系统论愈发重要
- 知识的系统化、因果化、立体化、网络化
- 依稀记得、通识

围绕知识的高级能力和素养愈发重要

- 逻辑思维、批判性思维、创造性思维
- 终身学习、灵活学习
- 沟通能力、动手能力、创新能力、领导能力
- 价值观和人格

实例：人机协作 - 机器人设计

来源：Stella, F., Della Santina, C. & Hughes, J. How can LLMs transform the robotic design process. Nature Machine Intelligence. 5, 561-564 (2023). <https://doi.org/10.1038/s42256-023-00669-7> 2023年6月7日





内容由AI生成，请仔细甄别

我们是否提出好问题？我们能否做出好判断？



图灵测试

1950年提出

人工智能领域最具争议性与启发性的评估范式

持续75年的思想实验

见证了AI从规则系统到神经网络的范式迁移

见证了人类对自身智能本质的认知突破

见证了智能时代的到来



DeepSeek

智能时代的全面到来和人机协作的新常态

一、智能演变

大数据带来大知识、通用人工智能

二、人机协作

小助理、副驾驶和代理人

三、产业现状

基础模型、智能应用和智能硬件

四、教育成长

效率提升、模式变革、范式重构

