CAICT 中国信通院

人工智能技术、产业和政策态势

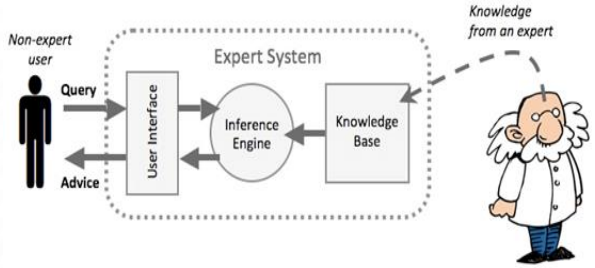
中国信通院 徐志发

2024.12

目录

AI大模型开启通往通用人工智能（AGI）之路

大模型代表人工智能技术演进进入新阶段，初步展现通用智能（Artificial General Intelligence）能力。

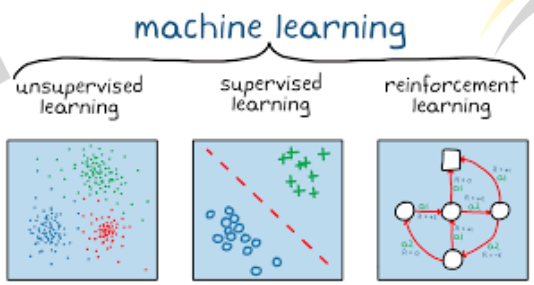


起步期
(1956-1980年代初)

- 1956年，AI概念首次提出
- 1970s-1980s，专家系统发展

基于规则 | 经验主义

依赖领域专家知识

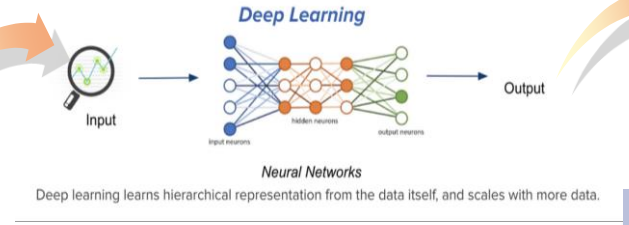


机器学习阶段
(1980年代初-21世纪初)

- 1980年代，机器学习成为独立分支

数据驱动 | 特征工程

基于数据手动构造特征



深度学习阶段
(2006年-2022年)

- 2006年，深度学习提出

神经网络 | 端到端学习

模型自动学习特征

大模型阶段
(2022年-)

- 2022年底ChatGPT发布

预训练 | 提示工程

知识迁移学习

DeepMind AGI 准则：

1. **关注能力而非过程**
2. **关注通用性和性能**
3. **关注认知和元认知任务**
4. **关注潜能而非部署**
5. **关注真实场景**
6. **关注通向AGI的路径，而非单一目标**

大模型加速人工智能技术迈向“大一统”

以Transformer架构为基础的大模型不断取得新突破，进一步确认了人工智能技术发展走向新范式

底层算法走向统一使得人工智能平台化成为可能，基础模型正在成为新的“操作系统”，创新不断提速

AI大模型的内涵与特征

- **AI大模型是人工智能预训练大模型的简称，包含了“预训练”和“大模型”两层含义**，二者结合产生了新的人工智能模式，即模型在大规模数据集上完成预训练后，仅需少量数据的微调甚至无需微调，就能直接支撑各类应用。

- 大模型通常具有多层神经网络结构，并使用高级的优化算法和计算资源进行训练，**具有强大的泛化性、通用性和实用性**，可以在自然语言处理、计算机视觉、智能语音等多个领域实现突破性性能提升。

泛化性

对新数据的适应能力
模型在从未见过的数据上能表现出良好的性能能力

通用性

解决多个任务的能力
模型能应用于不同的数据集或任务

实用性

应用时的可用性和效率
模型能以合理的时间和资源，快速处理数据并做出决策

“大模型”：在大规模无标注数据上进行训练，学习出一种特征和规则。基于大模型进行应用开发时，将大模型进行微调（在下游小规模有标注数据进行二次训练）或者不进行微调，就可以完成多个应用场景的任务，实现通用的智能能力。

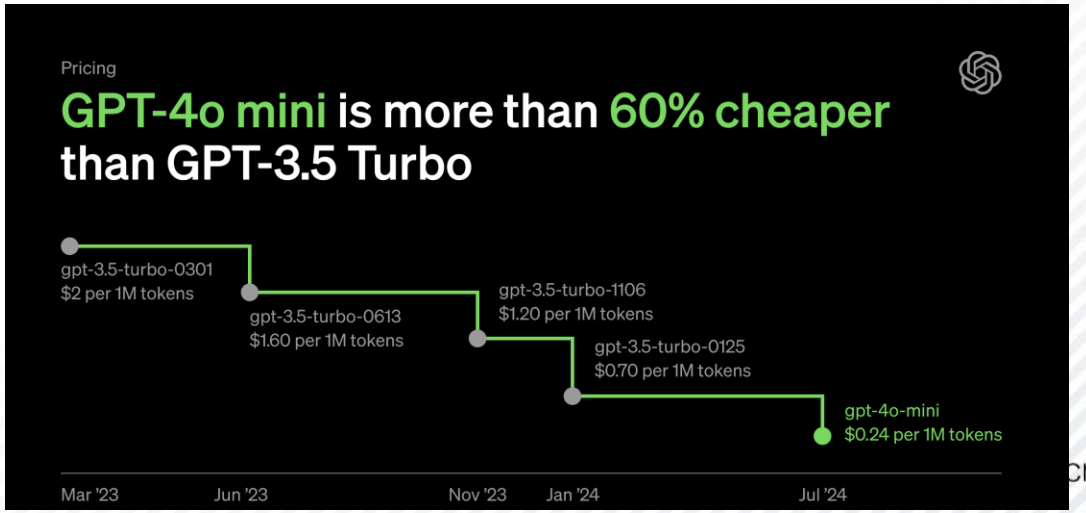
“小模型”：针对特定应用场景需求进行训练，能完成特定任务，但是换到另外一个应用场景中可能并不适用，需要重新训练（我们现在用的大多数模型都是这样）。这些模型训练基本是“手工作坊式”，并且模型训练需要大规模的标注数据。

主流大模型：语言类大模型GPT4

- GPT-4是OpenAI开发的最新语言模型，可以生成类似人类语言的文本，于2023年3月13日正式发布，用户可通过订阅ChatGPT-Plus 和Microsoft Copilot首先使用 GPT-4。除此之外，GPT-4 还以 API 的形式提供，开发者可以用来构建应用程序和服务。
- **GPT-4是在GPT-3.5基础上的一次重大升级**，在三个关键领域比之前的版本更加先进：创造力、视觉输入和更长的语境处理。

创造力方面	更长的语境处理	视觉输入
• GPT-4在生成创意项目和与用户合作方面表现更好，包括音乐创作、剧本写作、技术写作，甚至可以“学习用户的写作风格”。	• GPT-4可以处理多达128K个文本令牌，你甚至可以直接向 GPT-4 发送一个网页链接，并要求它与该网页上的文本进行交互。	• GPT-4 现在不仅能处理文字，还能接收图像作为输入。在 GPT-4 网站的一个示例中，用户上传了一些烘焙原料的图片，聊天机器人根据这些图片提供了可以制作的食谱。

- **GPT-4o mini 是OpenAI GPT-4 模型系列的最新版本。它是大型GPT-4o 模型的精简版**，适用于简单但需要大量处理的任务，这些任务更需要快速的响应速度，而不是整个模型的强大功能。
- GPT-4o mini 于2024 年7月发布，并取代 GPT-3.5 成为 ChatGPT 中的默认模型。
- 根据数据显示，GPT-4o mini 在 MMLU 推理基准中的表现明显优于类似的小型模型，比如谷歌的 Gemini 1.5 Flash 和 Anthropic 的 Claude 3 Haiku。



主流大模型：语言类大模型Llama3

- 作为开源领域的常青树，Llama系列一直在引领开源大模型的脚步，**Meta也被奉为与OpenAI齐名的巨头，分别代表着开源与闭源的技术走向**。
- 2024年7月，Mata发布Llama 3.1，包含8B、70B和405B三种参数规模，适用于多种场景，包括多语言代理、复杂推理和编码助手等。该系列模型上下文窗口增加到了128K，扩大16倍；支持多种语言，包括英语、德语、法语、意大利语、葡萄牙语、印地语、西班牙语和泰语等；提升了工具使用能力，支持搜索和Wolfram Alpha的数学推理；拥有更宽松的许可，允许使用模型输出改进其他LLMs。

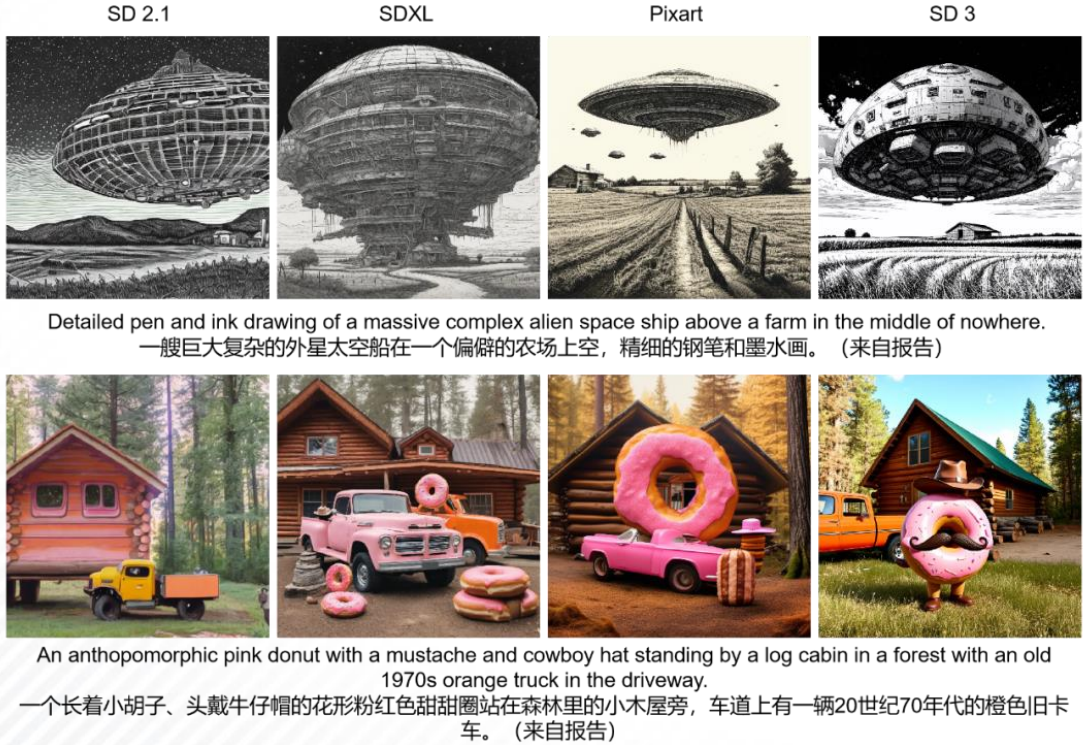
Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 (0125)	GPT-4o	Claude 3.5 Sonnet
General	MMLU (5-shot)	69.4	**72.3**	61.1	**83.6**	76.9	70.7	87.3	82.6	85.1	89.1	**89.9**
	MMLU (0-shot, CoT)	**73.0**	72.3△	60.5	**86.0**	79.9	69.8	88.6	78.7◁	85.4	**88.7**	88.3
	MMLU-Pro (5-shot, CoT)	**48.3**	–	36.9	**66.4**	56.3	49.2	73.3	62.7	64.8	74.0	**77.0**
	IFEval	**80.4**	73.6	57.6	**87.5**	72.7	69.9	**88.6**	85.1	84.3	85.6	88.0
Code	HumanEval (0-shot)	**72.6**	54.3	40.2	**80.5**	75.6	68.0	89.0	73.2	86.6	90.2	**92.0**
	MBPP EvalPlus (0-shot)	**72.8**	71.7	49.5	**86.0**	78.6	82.0	88.6	72.8	83.6	87.8	**90.5**
Math	GSM8K (8-shot, CoT)	**84.5**	76.7	53.2	**95.1**	88.2	81.6	**96.8**	92.3◇	94.2	96.1	96.4◇
	MATH (0-shot, CoT)	**51.9**	44.3	13.0	**68.0**	54.1	43.1	73.8	41.1	64.5	**76.6**	71.1
Reasoning	ARC Challenge (0-shot)	83.4	**87.6**	74.2	**94.8**	88.7	83.7	**96.9**	94.6	96.4	96.7	96.7
	GPQA (0-shot, CoT)	32.8	–	28.8	**46.7**	33.3	30.8	51.1	–	41.4	53.6	**59.4**
Tool use	BFCL	**76.1**	–	60.4	84.8	–	**85.9**	88.5	86.5	88.3	80.5	**90.2**
	Nexus	**38.5**	30.0	24.7	**56.7**	48.5	37.2	**58.7**	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	**95.2**	–	**95.2**	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	**83.4**	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	**100.0**	**100.0**	90.8
Multilingual	MGSM (0-shot, CoT)	**68.9**	53.2	29.9	**86.9**	71.1	51.4	**91.6**	–	85.9	90.5	**91.6**

- 评测结果显示，Llama 3.1 405B可与GPT—4o、Claude 3.5 Sonnet和Gemini Ultra等业界头部模型媲美，这也是Meta迄今为止最强大的模型。

- **Llama系列历经数次迭代，从Llama1到最新的Llama3，不仅在技术参数上实现了跨越式的提升，更通过开放源代码和数据集，深刻地改变了AI研究与应用的格局。**

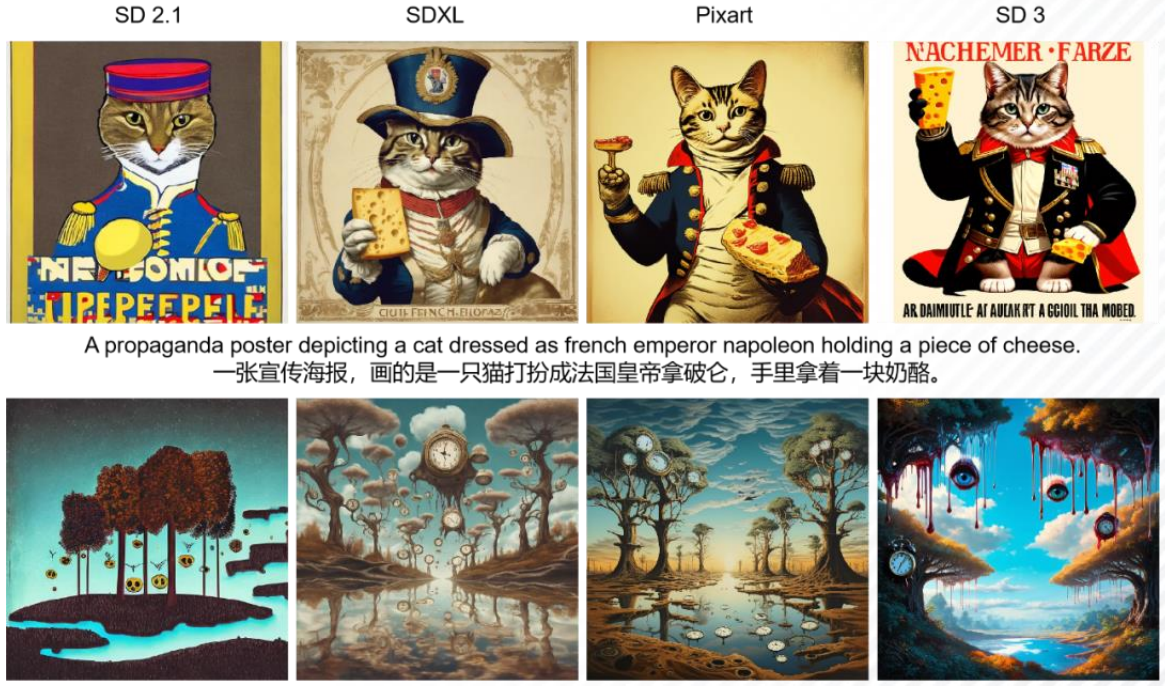
主流大模型：文生图大模型Stable Diffusion 3

- 2024年2月，Stability AI 发布了其第三代文生图大模型Stable Diffusion 3。该模型在排版和提示遵循等方面表现出色，超越了DALL·E 3、Midjourney v6 和 Ideogram v1 等最先进的文本到图像生成模型，为文本到图像生成技术带来了重大突破。



Detailed pen and ink drawing of a massive complex alien space ship above a farm in the middle of nowhere.
一艘巨大复杂的外星太空船在一个偏僻的农场上空，精细的钢笔和墨水画。（来自报告）

An anthopomorphic pink donut with a mustache and cowboy hat standing by a log cabin in a forest with an old 1970s orange truck in the driveway.
一个长着小胡子、头戴牛仔帽的花形粉红色甜甜圈站在森林里的小木屋旁，车道上有一辆20世纪70年代的橙色旧卡车。（来自报告）



A propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese.
一张宣传海报，画的是一只猫打扮成法国皇帝拿破仑，手里拿着一块奶酪。

A surreal landscape depicting melting clocks, floating eyes, and a sky filled with upside-down trees.
这是一幅超现实的风景画，描绘了融化的时钟、漂浮的眼睛和充满倒置树木的天空。

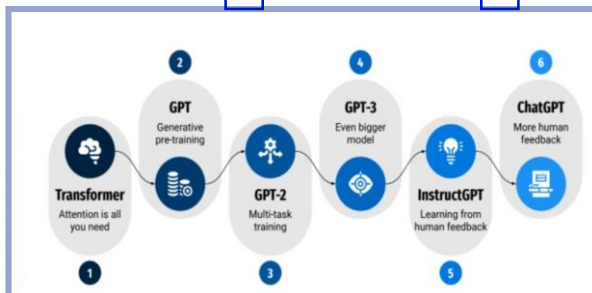
国内外大模型发展呈现“一横一纵”两条路径

横向路径：大模型以通用人工智能为目标，从通用大模型出发向更强能力、更通用的方向发展；

纵向路径：大模型构筑了智能基座，结合模型微调步骤，赋能更多行业与场景的应用。

路径一：横向发展

Agent　GPTs　API　插件



泛化能力

可扩展性

学习能力

多模态能力

横：通用大模型走向AGI

- 大模型朝着参数量更庞大、模型能力更强、效果更通用方向发展，需要持续高强度投入资源。

OpenAI　Microsoft　Meta

路径二：纵向发展

纵向：行业大模型

数智融合	金融新闻理解	虚拟口语教练	生成巡检报告
智能客服	金融市场评论	AI作文指导	意图识别
电信大模型	金融大模型	教育大模型	电力大模型

基础大模型

- 大模型构筑基础的支撑能力，赋能若干行业大模型和场景大模型，使模型应用门槛更低、见效更快。

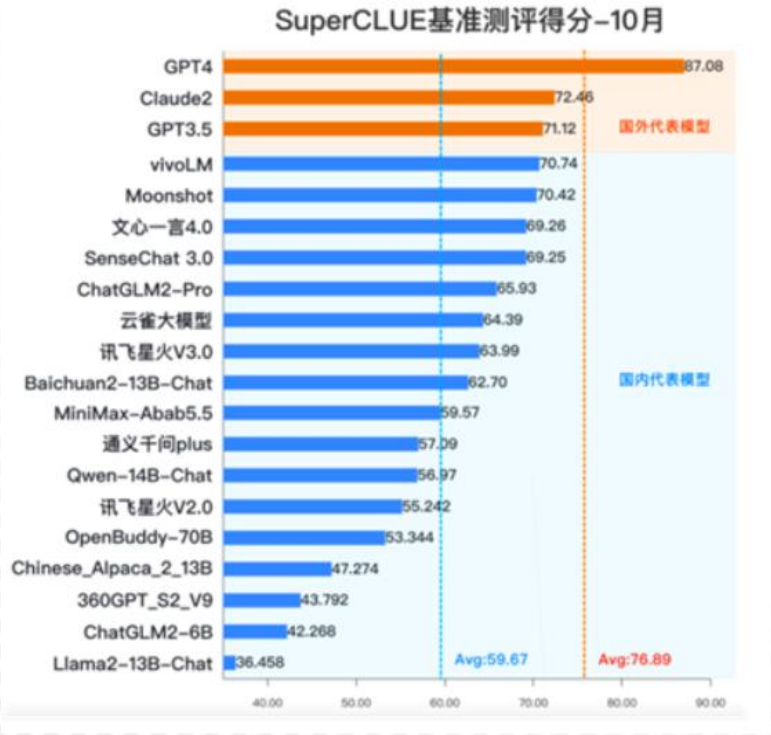
HUAWEI　Baidu 百度　中国电信 CHINA TELECOM

我国通用大模型处于技术追赶阶段，部分能力达GPT3.5水平

□ 头部国产大模型总体上与GPT-3.5能力极为接近，但在复杂逻辑推理、小样本自主学习、超长文本处理、跨模态统一理解等维度能力与GPT4依然存在差距。

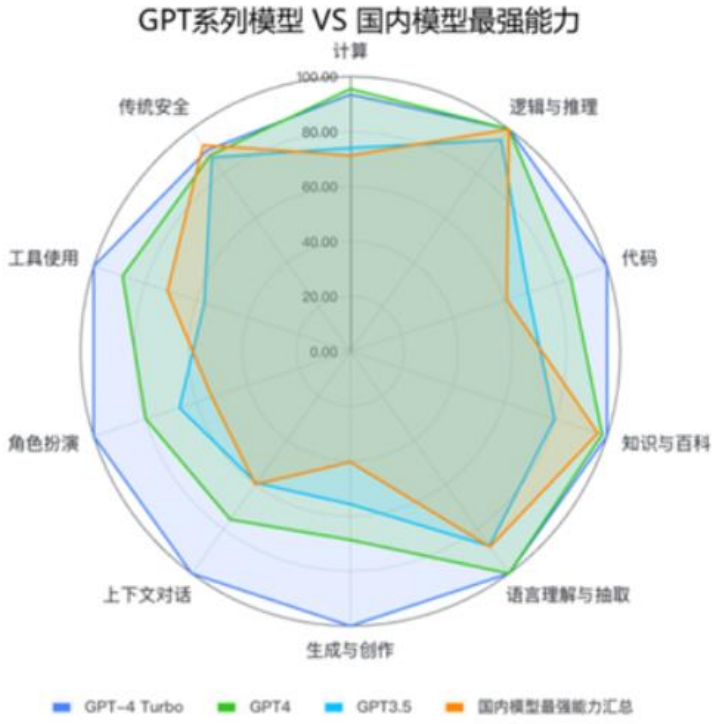
头部国产大模型与GPT-3.5能力极为接近

- 头部国产大模型总体上已经与GPT3.5能力极为接近，在48项任务测评上表现相当。



国产大模型较GPT-4V仍有差距

- 国产大模型在复杂逻辑推理、小样本自主学习、超长文本处理、跨模态统一理解等维度能力上与GPT4依然存在差距。



国产大模型网络结构有套壳之嫌

- 主流开源协议允许代码再发布，目前国产大模型网络结构是否原创仍有待考证。

Apache Licence

当前开源代码主流协议，鼓励代码共享，允许代码修改、再发布为开源或商业软件。

国产开源大模型与Llama参数量高度一致

	模型名称	发布时间	参数规模
美	LLaMA	2023年2月	70、130、340、650亿参数
中	Aquila-7B	2023年6月	70亿参数
	Aquila2-7B, 34B	2023年10月	70亿、340亿
	InternLM-7B	2023年6月	70亿参数
	baichuan-7B	2023年6月	70亿参数
	baichuan2-7B,13B	2023年9月	70亿、130亿参数
	Qwen-7B	2023年8月	70亿参数

大模型演进对智算能力提出更高要求

- 大模型的背后意味着巨大的计算资源，模型大小和训练数据大小是决定模型能力的关键因素。**因此，AI大模型时代，智能算力就是生产力。**

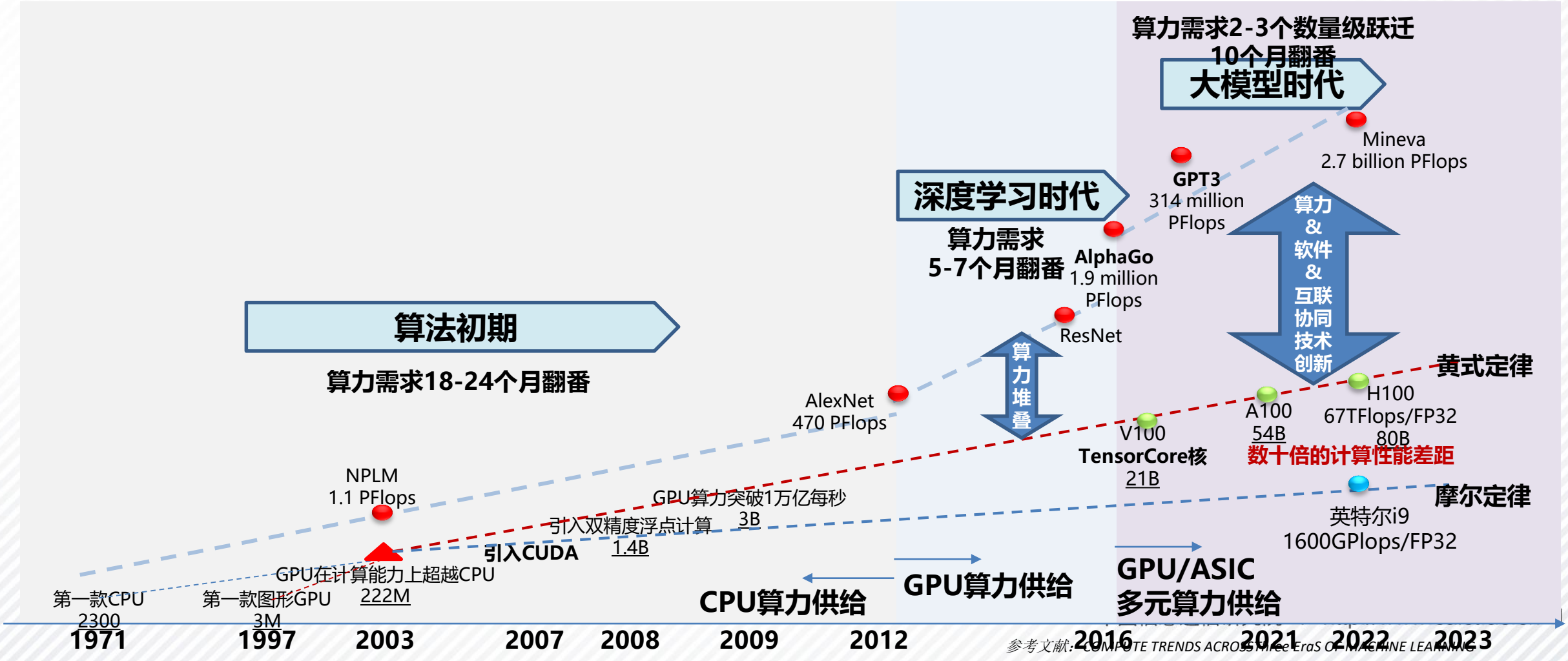
模型	模型参数（个）	训练集（Token数）	训练阶段 *训练所需的浮点运算次数=6×模型参数规模×训练集大小*					推理阶段	
			训练完成所需算力估算（PFLOPS）	服务器需求数量估算（台/每天）	电费（万元/天）	400P智算中心耗时（天）	100P智算中心耗时（天）	日常并行访问每秒所需算力估算（PFLOPS）	秒级推理响应所需算力规模（PFLOPS）
文心一言	2.60E+11	1.00E+12	5.17E+09	11957	121.24	149	598	239	0.26
GPT3	1.75E+11	3.00E+11	1.04E+09	2414	24.48	30	121	161	0.18
ChatGPT	1.75E+11	4.10E+11	1.43E+09	3300	33.46	41	165	161	0.18
GLM-130B	1.30E+11	4.00E+11	1.03E+09	2391	24.24	30	120	120	0.13
盘古	1.00E+11	3.00E+12	5.96E+09	13797	139.90	172	690	92	0.10
LLaMA-65B	6.50E+10	1.40E+12	1.81E+09	4185	42.44	52	209	60	0.07
BloombergGPT	5.00E+10	7.08E+11	7.03E+08	1628	16.51	20	81	46	0.05
LLaMA-7B	7.00E+09	1.00E+12	1.39E+08	322	3.27	4	16	6	0.01

注:1.统一换算成英伟达DGX A100服务器，8卡，640G，服务器总功耗6.5Kw
2.英伟达A100单卡支持FP16（半精度浮点运算）算力可达312TFlops，1PFlops算力约3.3张A100卡
3.电费按照0.65元/千瓦时

推理所需的浮点运算次数=2×模型参数规模×对话token大小

智能算力供需缺口日益扩大

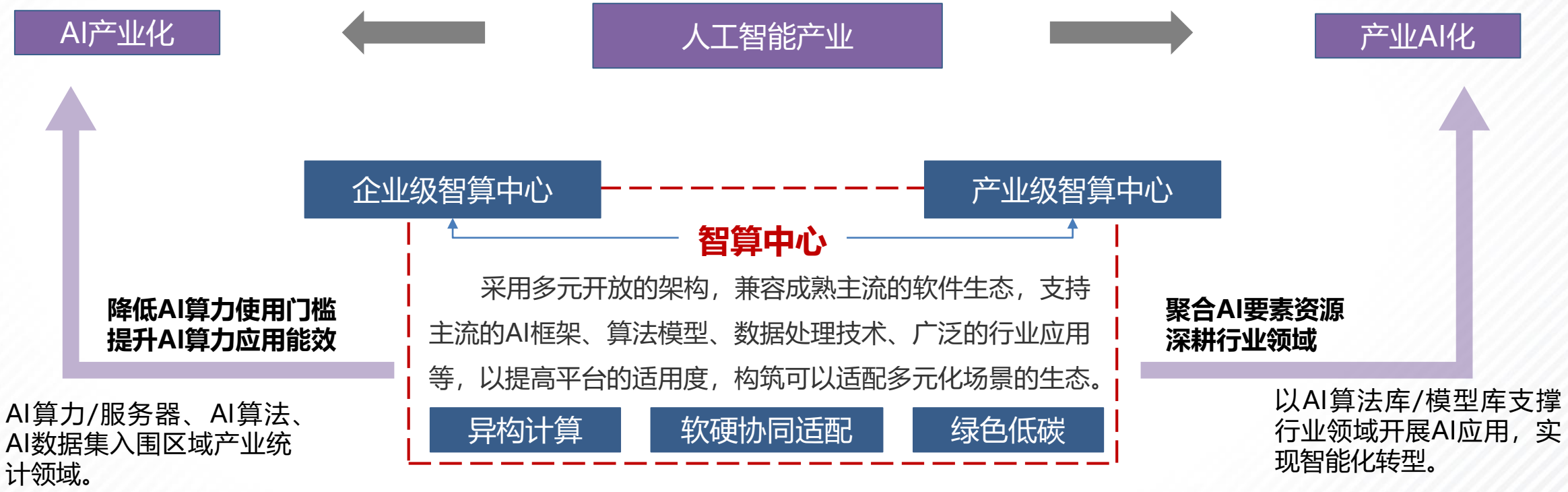
- **大模型技术及生态迭代，对智能算力提出更强更大规模的需求，智算成为支撑AI大模型的重要基石。**
- 巨量参数已成为人工智能大模型开发过程的必经之路，大模型算力需求超过半导体增长曲线，算力需求增长与芯片性能增长之间尚不匹配。



参考文献：COMPUTE TRENDS ACROSS THREE ERAS OF MACHINE LEARNING

智能计算中心已成为AI落地赋能的核心设施之一

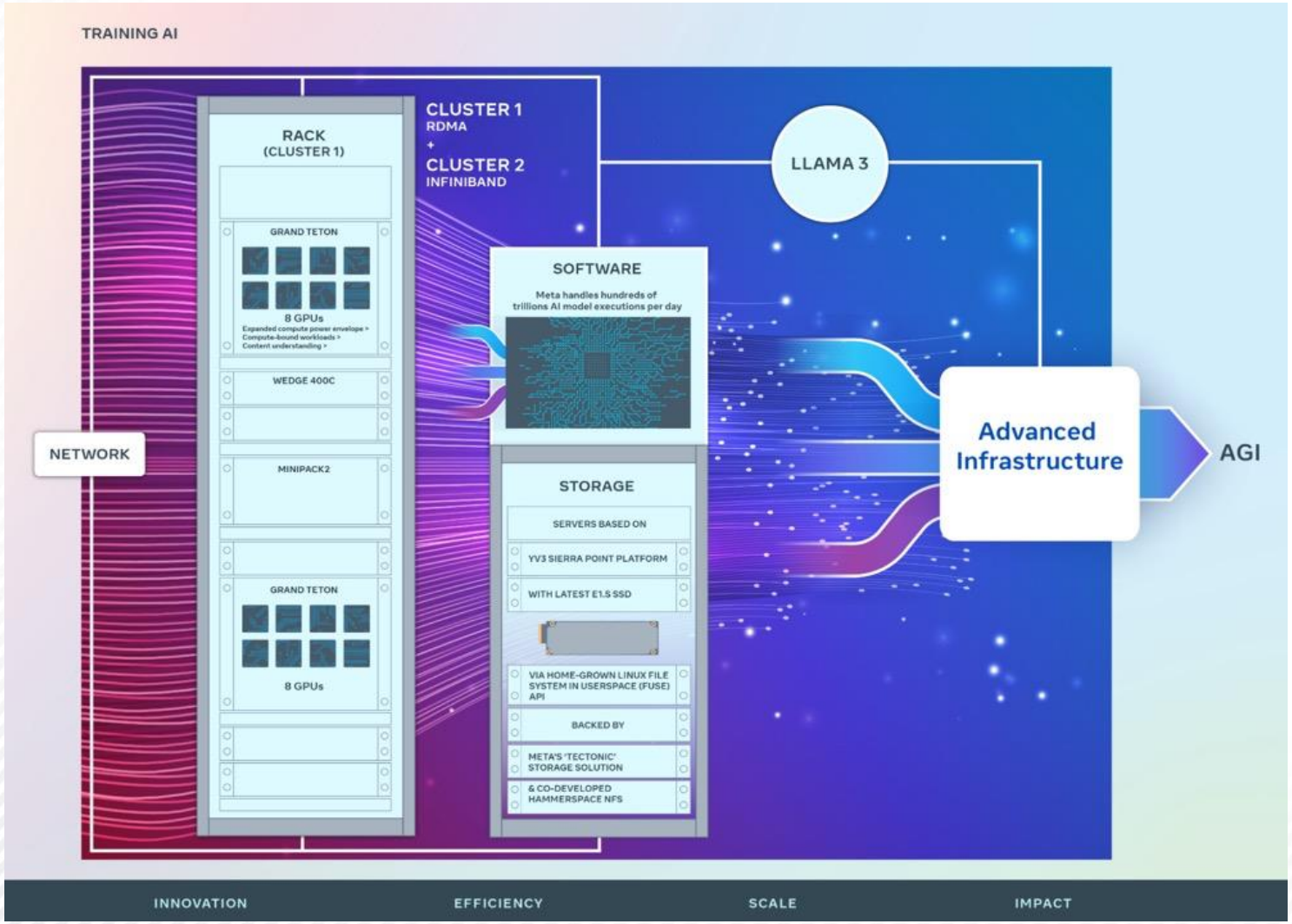
- **从总体定位看，智算中心成为地方发展人工智能产业、发展产业数字化的重要创新载体，具有重要价值。**
- **从落地运营看，智算中心是加快AI产业化和产业AI化的重要战略支撑。**



- **智算中心是涵盖了软硬件、解决方案为一体的技术创新综合体。未来的智能型企业、公共服务都将建构在智算中心的基础上，供给形态可以有自建、云服务、公共基础设施等多种形式。企业和组织将根据自己的需求和资源禀赋，选择匹配自身需求的服务形态。**

更大规模的智算集群成为大厂的追逐目标

- **大模型演进，对智算能力提出更高要求，对智算集群提出更高要求。市场的主力玩家们利用数万个GPU构建大型人工智能集群，以训练LLM。**



- Meta用2.2万个V100 GPU构建了第一代人工智能集群。 — 2017年
- Meta用9920个A100 GPU加速器构建了研究超级集群 (RSC) — 2023年5月
- **Meta推出了两个拥有2.4万个GPU的智算集群（总计49152个H100）** — 2024年3月
- Meta 计划，**到2024年底将部署 35 万个 NVIDIA H100 GPU** — 2024年12月

头部主体积极推动万卡智算集群建设

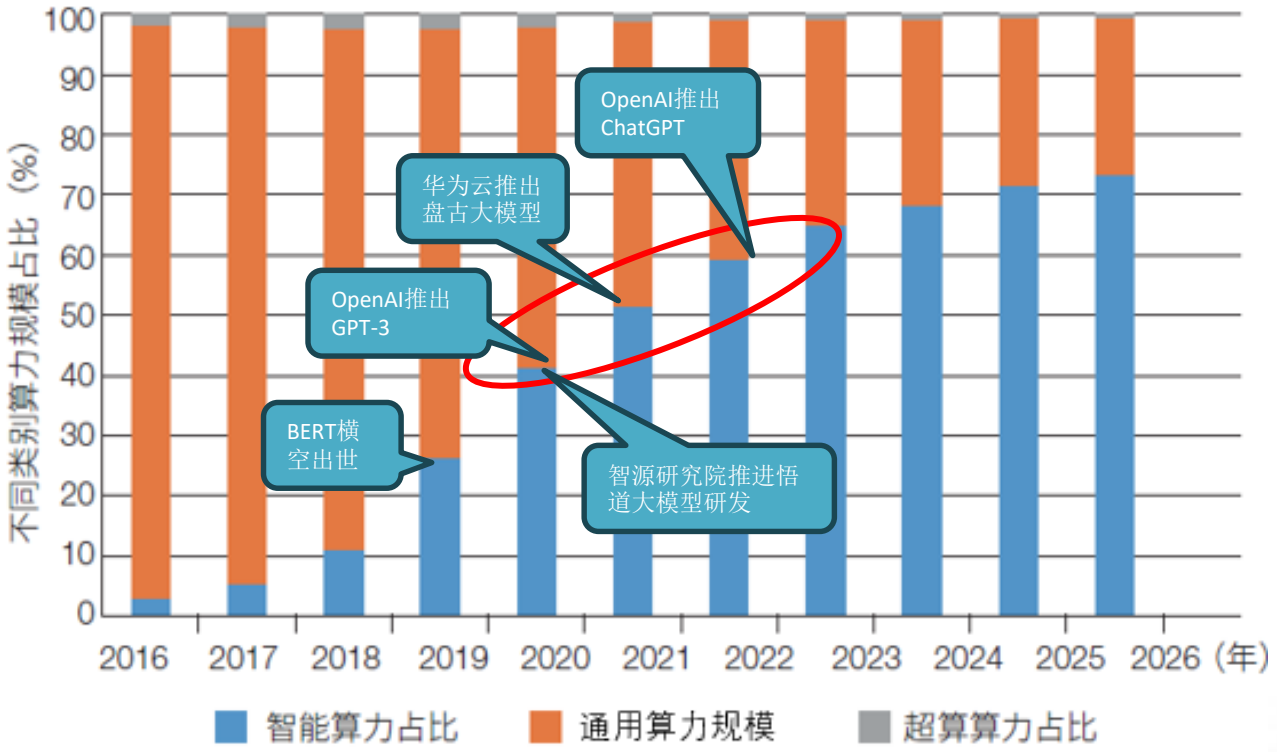
- 通信运营商、头部互联网、大型AI研发企业等均在发力超万卡集群的建设，比如中国移动、华为、字节跳动、阿里巴巴、百度、科大讯飞等。

序号	智算中心名称	城市	投资主体	算力卡规模
1	中国移动呼和浩特智算中心	呼和浩特	中国移动	20000
2	中国移动哈尔滨智算中心	哈尔滨	中国移动	20000
3	中国移动贵阳智算中心	芜湖	中国移动	20000
4	中国电信天翼云上海临港国产万卡算力池	上海临港区	中国电信	15000
5	中国联通上海临港国际云数据中心	上海临港区	中国联通	10000
6	商汤新一代人工智能计算与赋能平台	上海临港区	商汤科技	20000
7	深圳鹏程云脑	深圳	鹏程实验室	16000
8	科大讯飞飞星一号	合肥	科大讯飞	10000
9	蚂蚁集团万卡异构算力集群	上海	蚂蚁集团	10000
10	字节跳动	北京	字节跳动	12000
11	百度智算集群	北京	百度	10000

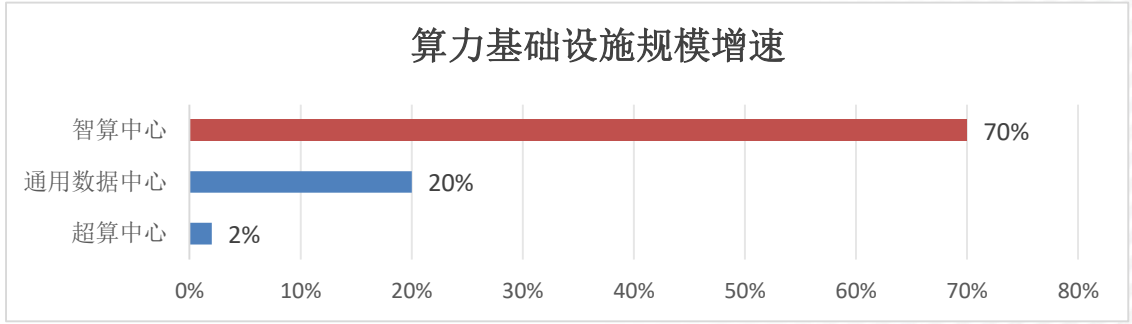
一个 10 万 GPU 的集群需要超过 150MW 的数据中心容量，一年的消耗就是 1.59TWh（15.9 亿度电）。

大模型爆发，助推智算中心正成为我国算力基础设施最主要增长点

- 随着人工智能技术在经济社会的加速渗透，加之以大模型为代表的人工智技术快速迭代，人工智能计算需求快速提升，智能算力规模持续提升，正成为算力基础设施发展的重要方向。
- 据工信部数据显示，截至2023年底，我国智能算力规模达到70EFlops，智能算力占比算力总规模30.4%，年增速超过70%。**在大模型技术创新与深度赋能态势下，智能算力占比将逐年上升，预计2026年占比将达到73%，达到365EFlops（FP32）。**



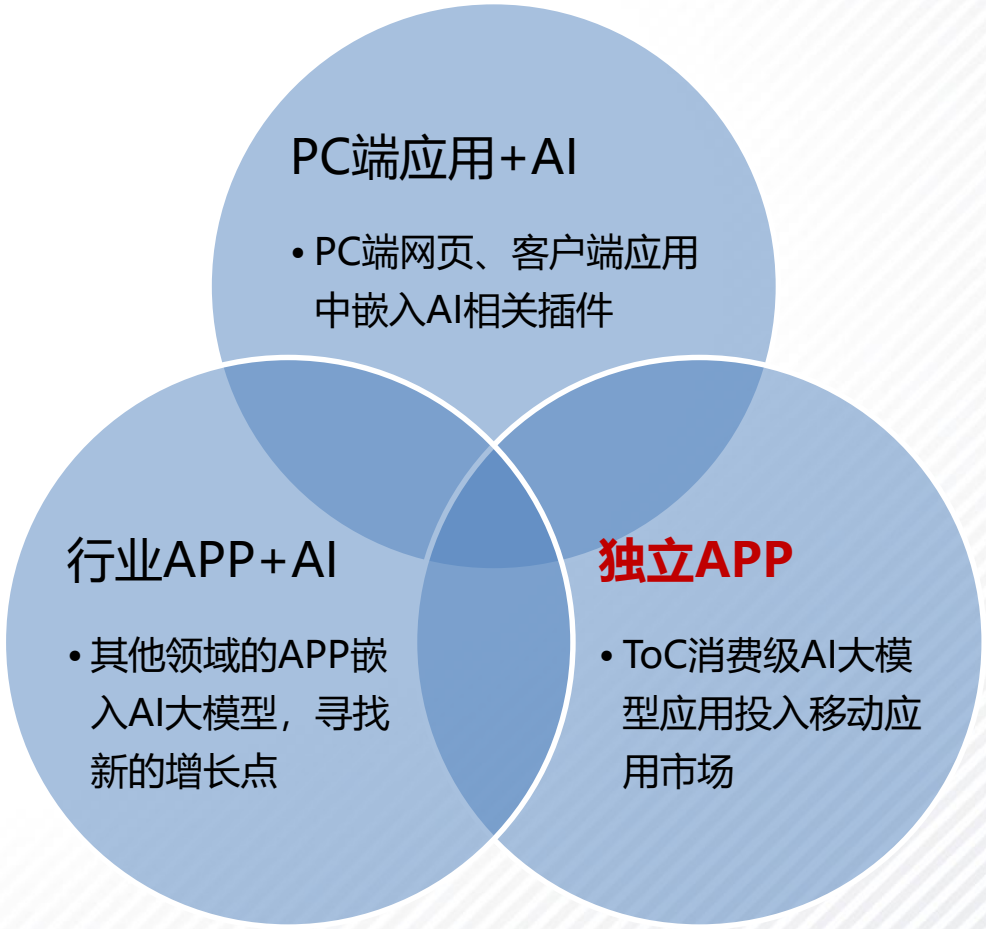
数据来源：中国信息通信研究院



数据来源：中国信息通信研究院

人工智能产品形态持续演进

- AI产品形态向内容生成、知识洞察、智能助手以及数字代理等方向演进，后续将着重在AIGC产品领域进行进一步的延展。



人工智能产品形态：PC端应用+AI

- **信息生成：**围绕工作场景，用户可借助AI生成所需的文本、图片等内容信息，也可生成办公文档、辅助计算，缩短工作时长，提高效率

AI生成文字、图片等信息

使用AI大模型网页端可生成所需的文本、图片内容

AI生成办公文档

通过AI功能生成所需的PPT、EXCEL等办公文档

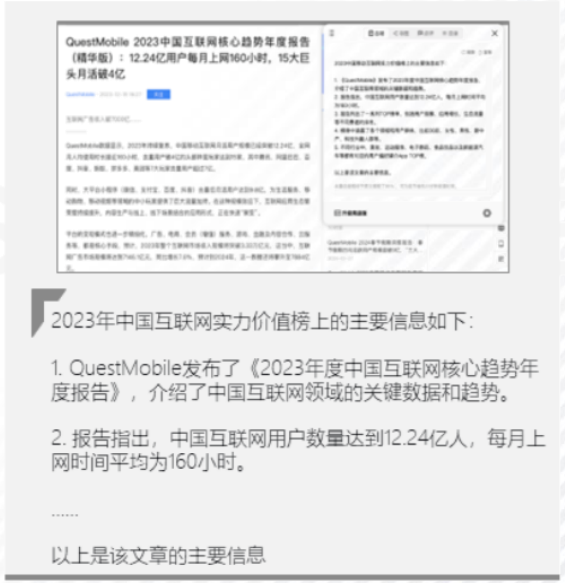
- **内容总结：**借助浏览器AIGC插件总结文字、视频等内容，提升阅读效率，聚焦核心信息，并以标签形式提炼，助力工作提效

AI总结网页视频内容

浏览器中的AI插件可对视频内容进行文字总结

AI总结报告内容

浏览器中的AI插件可对报告内容进行文字总结



人工智能产品形态：独立APP

- 生成式AI行业移动端渗透率上升至12%，月活已超4000万，日均新增规模保持稳定，用户新增和粘性情况稳中有升。现阶段头部应用普遍聚焦在文本和图像信息模态生成。

- 据极光大数据旗下月狐数据发布的《2024年5月中国生成式AI行业市场热点月度分析》报告显示，移动端主流生成式AI应用当中，**文心一言实现了今年月活用户连续5个月的持续增长，月活用户达到1105.6万，稳居国内厂商第一**，并且断档领先天工、智谱清言、Kimi等一众生成式Ai应用（kimi的月活甚至不及文心一言的1/4）。

移动端主流生成式AI应用活跃情况（2024年5月）

序号	应用名	厂商	活跃用户数（万）
1	文心一言	百度	1105.6
2	豆包	字节跳动	1049.8
3	天工	昆仑万维	573.5
4	智谱清言	智谱清言	340.2
5	Kimi	月之暗面	313.7
6	光速写作	作业帮	221.3
7	讯飞星火	科大讯飞	197.4
8	星野	MiniMax	132.6
9	通义千问	阿里	107.9
10	海螺AI	MiniMax	57.7

人工智能产品形态：行业APP+AI

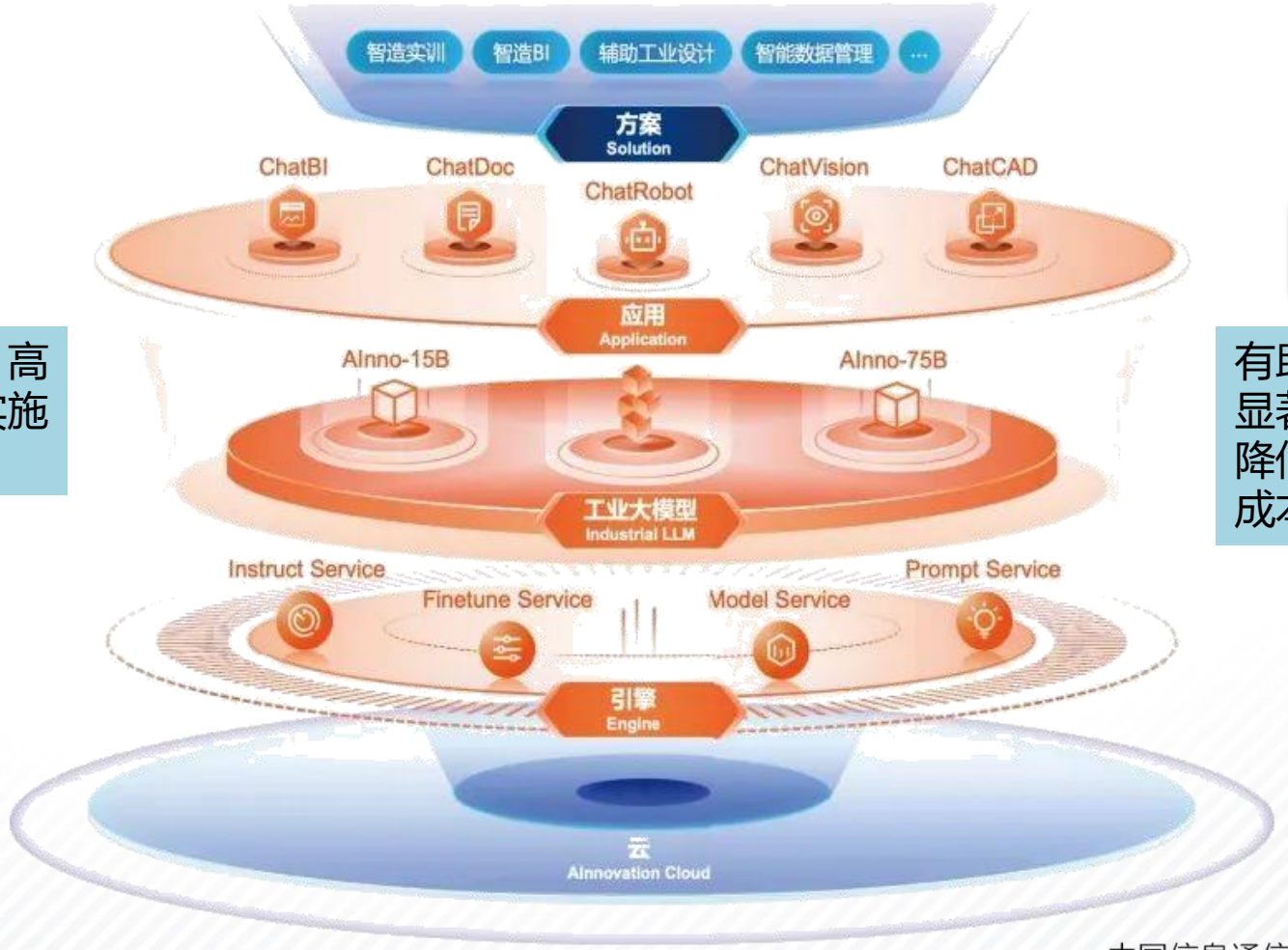
- 各行各业“垂直行业模型+APP”模式发展已初见格局，以AI能力提升自身原有应用体验、重构交互模式成为目前较为明确的发展思路，在原有应用强大的流量水平加持下，AI能力有望更快达成提升交易效率、生产和服务效率等目标。

大模型应用：奇智孔明AInnoGC工业大模型

- AInnoGC平台由工业大模型、大模型服务引擎和ChatX系列生成式应用构成。大模型参数量级在750亿以上，具备工业知识问答、数据分析、代码生成、任务编排、海量知识管理、复杂逻辑推理、长流程任务编排、Agent智能体以及更多工业模态的生成能力。

应对需求

工业小模型面临数据质量低、高标注成本、泛化能力有限和实施难度大等问题。



产生效益

有助于提高生产力和效率；可以显著减少人工更新和维护的需要；降低总体运营成本，提供更高的成本效益。

大模型应用：商汤大医

- “大医”基于80亿医学文本训练，能够提示工程自定义、长程记忆存取、医学知识库查询总结、多智能体调度等功能，内嵌了13个预设医疗场最，以满足新华医院的特定需求。

应对需求

新华医院希望在全院推动“医疗大模型”技术融合，重构诊前、诊中、诊后全流程，提高医疗服务的整体效率，改善患者的就医体验，优化医院资源分配，同时面向未来构建数据驱动的“健康生活管理”创新医疗模式。

产生效益

对提升医院运营指标具有长期正面影响和巨大发展空间；对于数据的高效利用将推动医疗服务智能化创新；数字孪生、元宇宙等模式对老龄化社会的居家照护具有现实意义。

大模型应用：丰登种业大语言模型

- 基于书生浦语2.0强大的基座模型能力，丰登通过注入我国种业相关的科研文献、科技书籍、种企报告等数据，**使大模型获得了理解和分析育种相关专业问题的能力**，拓展了大模型助力生物育种的探索路径。

应对需求

育种信息分散；缺乏统一的数据；知识平台数据的孤岛性、分散性等限制了生物育种的工作效率；育种技术的学习有着显著的行业和知识壁垒。

产生效益

提高获取种业信息的工作效率；降低了育种知识的学习门槛；填补了我国在种业专业领域的大模型空白；是对我国育种领域采纳新技术的一次大胆尝试。

大模型应用：蜜巢大模型

- 通过运用蜜巢政务大模型的智能知识管理、智能舆情分析能力，并引入先进算法，将业务工作层、数据分析层、治理决策层进行一体化整合，逐一击破业务痛点。**蜜巢政务大模型帮助市民服务热线实现智能化转型**，同时有效降低了人力成本，助力业务效率整体提升80%以上。

应对需求

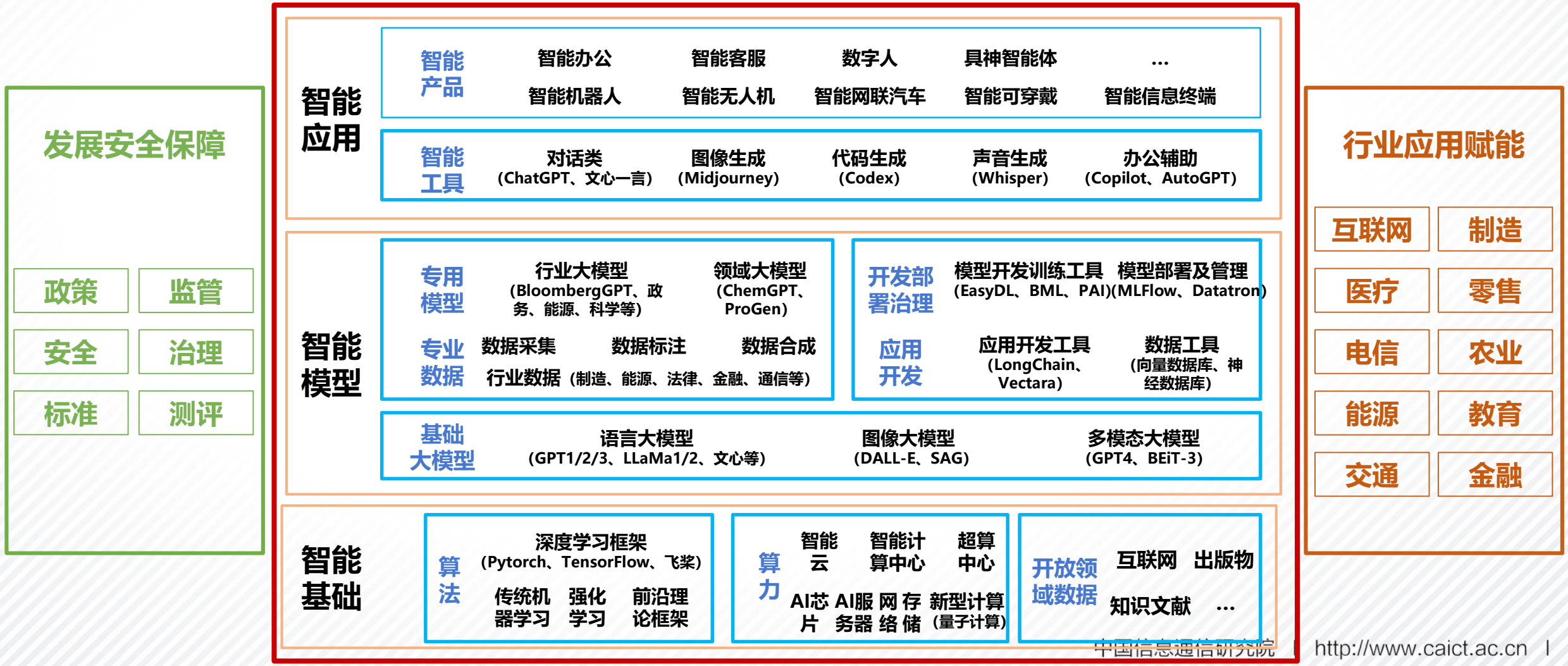
在市民服务热线场景中，工作人员面临着热线量大、工单快速分类难、诉求分析少、辅助决策难等痛点。

产生效益

帮助提升政企办公场景日常工作效能；推动各行业向智能化、高效化、数字化方向迈进。

人工智能新的产业链和产业生态日益成熟

- 产业链：“基础软硬件构成根系、大模型作为主干、工具和产品枝繁叶茂”
- 产业生态：“以安全保发展，以赋能为使命”



目录

大模型：网信办牵头推进生成式人工智能服务管理

□促进生成式人工智能健康发展和规范应用，是《生成式人工智能服务管理暂行办法》的核心主旨，为此《办法》通过 “划定合规底线”、 “采取有效措施鼓励发展”、 “采取更精细化监管举措”这三个层面推进落实。

2023年七部委《生成式人工智能服务管理暂行办法》

应用场景优化
应用生态构建
基础技术创新
基建平台建设

鼓励措施

监管举措

安全评估
算法备案
信息披露
数据分级

合规底线

基本法律、行政法规、社会公德、伦理道德

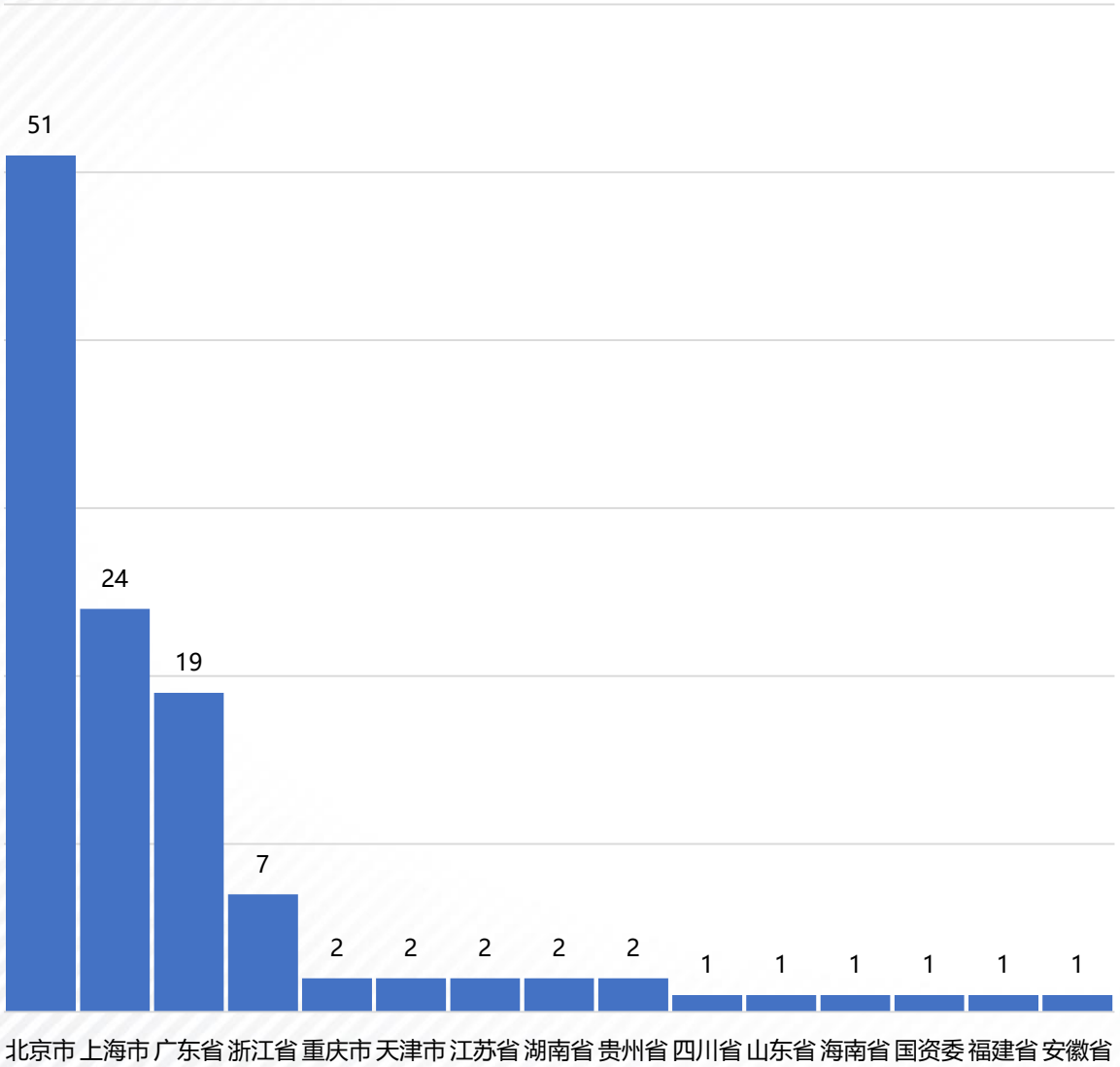
生成式人工智能服务已备案信息（2024年3月）

序号	属地	模型名称	备案单位	备案号	备案时间
115	广东省	CVTE大模型（自研）	广州视源电子科技股份有限公司	Guangdong-shiyuan-20240314	2024/3/28
116	广东省	YOYO助理（PC版）	深圳荣耀软件技术有限公司	Guangdong-YOYOPC-20240314	2024/3/28
117	广东省	YOYO助理（移动版）	深圳荣耀软件技术有限公司	Guangdong-YOYOM-20240314	2024/3/28

序号	属地	模型名称	备案单位	备案号	备案时间
87	国资委	九天自然语言交互大模型	中国移动通信有限公司	ZhongYangQiYe-JiuTian-20240123	2024/2/7

□ 截至2024年3月，已有117个大模型成功完成备案，其中北京、上海和广东三个地区共计拥有94个大模型，上海以24个大模型的数量位居第二。

大模型：当前备案情况



序号	属地	模型名称	备案单位	备案号	备案时间
1	北京市	文心一言	北京百度网讯科技有限公司	Beijing-WenXinYiYan-20230821	2023/8/31
2	北京市	智谱清言（ChatGLM）	北京智谱华章科技有限公司	Beijing-ChatGLM-20230821	2023/8/31
3	北京市	云雀大模型	北京抖音信息服务有限公司	Beijing-YunQue-20230821	2023/8/31
4	北京市	百应	北京百川智能科技有限公司	Beijing-BaiYing-20230821	2023/8/31
5	北京市	紫东太初大模型开放平台	中国科学院自动化研究所	Beijing-ZiDongTaiChu-20230821	2023/8/31
6	上海市	abab	上海稀宇科技有限公司	Shanghai-Abab-20230821	2023/8/31
7	上海市	日日新	上海商汤智能科技有限公司	Shanghai-RiRiXin-20230821	2023/8/31
8	上海市	书生·浦语	上海人工智能创新中心（上海人工智能实验室）	Shanghai-ShuShengPuYu-20230821	2023/8/31
9	安徽省	星火认知大模型	科大讯飞股份有限公司	Anhui-XingHuoRenZhiDaMoXing-20230823	2023/9/4
10	天津市	360智脑大模型	三六零科技集团有限公司	Tianjin-360ZhiNaoDaMoXing-20230831	2023/9/11
11	浙江省	通义千问大模型	阿里巴巴达摩院（杭州）科技有限公司	ZheJiang-TongYiQianWen-20230901	2023/9/12
12	广东省	腾讯混元助手大模型	深圳市腾讯计算机系统有限公司	Guangdong-TencentHunyuan-20230901	2023/9/14
13	贵州省	华为云盘古NLP大模型	华为云计算技术有限公司	Guizhou-HuaWeiYunNLPDaMoXing-20230908	2023/9/19
14	江苏省	智慧助手（小艺）大模型	华为软件技术有限公司	Jiangsu-ZhiHuiZhuShou-20230911	2023/9/27

大模型：当前备案发展态势

◆ **大模型领域的监管逐渐走向常态化和有序化。**

2023年8月31日，百度文心一言、商汤大模型“商量SenseChat”、百川智能的百川大模型、智谱华章的“智谱清言”等第一批通过备案的公司官宣向广大用户开放。除此以外，抖音云雀大模型、智谱AI“GLM”大模型、中科院紫东太初大模型，MiniMax“ABAB”大模型、上海人工智能实验室书生通用大模型、360公司的“360智脑”等大模型也陆陆续续对公众开放。

2023年11月初，第二批通过备案的公司开始密集宣布，产品将面向广大用户开放。其中包括网易有道“子曰”教育大模型、昆仑万维“天工”大模型、知乎“知海图AI”模型、金山办公“WPS AI”、好未来“MathGPT”大模型、面壁智能“面壁露卡Luca”、出门问问“序列猴子”、月之暗面“Moonshot”等公司的大模型产品。

2023年12月底，第三批国产大模型通过备案，其中包括京东“言犀大模型”、抖音“福禄瓜视觉大模型”、快手“快意大模型”、红棉小冰科技“小冰大模型”、澜舟科技“孟子GPT大模型”、中科闻歌“雅意大模型”、深言科技“语鲸大模型”、云知声“山海大模型”和聆心智能“CharacterGLM”等大模型产品。

◆ **垂直行业大模型逐步成为大模型监管备案的重点主体。**

大模型：业界对大模型备案的认知定位

四川经济日报　第二版：要闻　2024年04月18日

我省首个人工智能大模型在国家网信办成功备案

四川经济日报讯（记者 廖振杰）4月16日，记者从四川省科技厅获悉，国家互联网信息办公室日前发布生成式人工智能服务已备案信息公告，我省首个人工智能大模型“长虹云帆”成功备案。

据悉，截至2024年3月，全国共有117个人工智能大模型备案。四川长虹电器股份有限公司（下称“长虹公司”）将生成式人工智能应用在电视终端，通过自主研发的AI技术，融合训
加速智慧家庭领
务管理、多模交
下一步，长
智能制造、智能
业高端化、智能

通信产业报网　+关注

24-4-7 10:11　来自 微博网页版

【央企首个！中国移动九天大模型通过双备案】4月2日，国家网信办公布已备案大模型清单，中国移动“九天自然语言交互大模型”名列其中，这标志着中国移动九天AI大模型可正式对外提供生成式人工智能服务，也成为了同时通过国家“生成式人工智能服务备案”和“境内深度合成服务算法备案”双备案的首个央企研发的大模型。网页链接

杭州市经济和信息化局

《关于支持人工智能全产业链高质量发展的补充意见》

支持模型合规备案。鼓励企业自研模型申请模型备案，对获得中央网信办生成式人工智能模型备案的企业，给予模型评测等相关费用一次性奖励50万元。（责任单位：市委网信办、市经信局）

成都市经济和信息化局

《成都市进一步促进人工智能产业高质量发展的若干政策措施》

支持企业、高校院所开展行业大模型研发应用，对性能先进且成功通过国家大模型备案登记的前十名，给予100万元一次性奖励。

智能算力：工信部&发改委从不同角度推进智能算力科学布局

- 国家部委从不同角度推进智能算力科学布局。

- **工信部关注智能算力规模占比**
  - 《算力基础设施高质量发展行动计划》，提出到2025年，算力规模超过300EFLOPS，东西部算力平衡协调发展。并且提出，“**重点在西部算力枢纽及人工智能发展基础较好地区集约化开展智算中心建设，逐步合理提升智能算力占比**。”

- **国家发展改革委重视算力一体化建设**
  - 《关于深入实施“东数西算”工程 加快构建全国一体化算力网的实施意见》，**提出加强通用计算、智能计算、超级计算等多元算力资源的科学布局**，提升国家枢纽节点各类算力资源的综合供给水平；**要提升智能算力在人工智能等领域适配水平**，增强计算密集型、数据密集型等业务的算力支撑能力。

国家部委严密推进算力网络建设，促进东西部算力高效互补和协同联动

提升算力高效运载能力，探索构建布局合理、泛在连接、灵活高效的**算力互联网**，探索算力协同调度机制，开展算网融合发展行动

提出要统筹东中西部算力的一体化协同，提升算力网络传输效能，**探索算网协同运营机制**，构建跨区域算力调度体系

智能算力：央、地高度重视智能算力绿色低碳、技术创新发展

绿色低碳

- **中央层面** 逐步对以智算中心为代表的算力基础设施的**绿色化、高效化、集约化发展**提出更细致的目标和要求。
  - 《信息通信行业绿色低碳发展行动计划（2022-2025年）》提到全国新建大型、超大型数据中心电能利用**效率（PUE）降到1.3以下**。
  - 深入实施“东数西算”工程加快构建全国一体化算力网的实施意见，提到推进**数据中心用能设备节能降碳改造，推广液冷**等先进散热技术。
- **地方层面** 北京、上海、浙江、贵州、四川、广东等省市纷纷发布相关政策，对算力基础设施减碳措施集中于通过开展**节能降碳技术**的应用**、利用清洁能源、余热回收**等手段减少碳排放。

技术创新

- **中央层面**
  - 《新型数据中心发展三年行动计划(2021-2023年)》强调部署数据中心发展时，推动**CPU、GPU等异构算力**提升，强调逐步提高**自主研发算力的部署比例**。
  - 《算力基础设施高质量发展行动计划》提出创新驱动原，则鼓励在关键信息基础设施中使用**自主的存储设备，带动关键存储部件的国产化应用**。
- **地方政府举措**

技术迭代——紧抓智算产业发展机遇，**增强大模型、高性能计算**等技术创新能力。

技术应用——关注智算技术在**重点应用场景**中的融合升级，着力推动科技成果向现实生产力转化。

技术国产化——通过奖励、退税等措施**支持智算中心国产化建设**，推动提升智能计算自主可控水平。

融合赋能：多部委密集关注AI融合赋能

2019年中央深改委
《关于促进人工智能和实体经济融合的指导意见》

促进人工智能和实体经济深度融合，要把握新一代人工智能发展的特点，坚持以市场需求为导向，以产业应用为目标，深化改革创新，**优化制度环境，激发企业创新活力和内生动力，结合不同行业、不同区域特点，探索创新成果应用转化的路径和方法**，构建数据驱动、人机协同、跨界融合、共创分享的智能经济形态。

2022年六部委
统筹推进人工智能场景创新

科技部、教育部、工业和信息化部、交通运输部、农业农村部、国家卫生健康委等六部门印发《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》，以促进人工智能与实体经济深度融合为主线，**以推动场景资源开放、提升场景创新能力为方向**，强化主体培育、加大应用示范、创新体制机制、完善场景生态，加速人工智能技术攻关、产品开发和产业培育，**探索人工智能发展新模式新路径**，以人工智能高水平应用促进经济高质量发展。

2024年工信部
推进人工智能赋能新型工业化

加快制造业关键环节、重点行业和重点产品智能化升级。深化人工智能技术在制造业全流程融合应用，大幅提升研发、中试、生产、服务、管理等环节智能化水平。面向对国民经济影响大、带动力强、数字化基础好的重点行业，**开展人工智能赋能新型工业化专项行动，加强供需对接、标准宣贯、应用推广，加快重点行业智能化升级**，提升高端装备、关键软件、智能终端等重点产品和装备智能化水平。

区域示范：工信部国家人工智能创新应用先导区

国家人工智能创新应用先导区

指导思想

是贯彻党的十九届历次全会和中央经济工作会议精神，落实中央全面深化改革委员会第七次会议部署，促进人工智能和实体经济深度融合的重要举措。

立足定位

谋改革 **促应用** **导经验**

主要任务

强化应用导向，挖掘应用场景，带动新技术、新产品落地应用，培育新经济增长点

深化改革创新，在政策机制层面勇于突破、先行先试，探索更多新模式

加强部省联动，调动产业、科技、教育、金融等力量协同，整合资源、凝聚工作合力

发挥产业基础，找准优势领域和发力重点，因地制宜实现差异化、特色化发展

工信部推进先导区建设整体安排

第一批（2019年）

	上海（浦东新区）	济南—青岛	深圳
优势利用	产业布局、基础建设、标准体系构建、知识产权交易	制造业基础 大数据资源与应用场景	电子信息与通信基础 创新生态、企业发展
聚焦领域	制造、医疗 交通、**金融**	制造业、医疗 **家居**、轨道交通	医疗健康、**金融** **供应链**、交通、制造

第二批（2021年）

	北京	天津（滨海新区）	杭州	广州	成都
优势利用	技术原创、产业生态、人才基础 发展环境	中国（天津）自由贸易试验区政策	城市数字治理 先进制造	产业链条齐全 创新要素汇集 应用场景丰富	“一带一路”重要枢纽、“成渝”双城经济圈
聚焦领域	制造、**网联汽车** 智慧城市 **“科技冬奥”**	制造 **智慧港口** 智慧社区	城市管理 智能制造 **智慧金融**	**智能关键器件** **智能软件** 智能设备	**医疗**、**金融** 带动区域融通

第三批（2022年）

	南京	武汉	长沙
优势利用	科教资源丰富 软件产业基础扎实	中部崛起&长江经济带交汇点 老工业基地和新兴产业基地	产业基础厚 基础设施强
聚焦领域	制造、**能源** **文旅**、**消费**、教育	智能制造、**智能建造** 智慧教育、**智慧医疗**	制造、**家居**、**智能工程机械** **智能文创**、**智能安全**

国资委：推动中央企业在人工智能领域实现更好发展、发挥更大作用

国资委召开中央企业人工智能专题推进会

加快推动人工智能发展，是国资央企发挥功能使命，抢抓战略机遇，培育新质生产力的必然要求。

- 中央企业要主动拥抱人工智能带来的深刻变革，**把加快发展新一代人工智能摆在更加突出的位置**，不断强化创新策略、应用示范和人才聚集。
- **着力打造人工智能产业集群**，发挥需求规模大、产业配套全、应用场景多的优势，带头抢抓人工智能赋能传统产业，**加快构建数据驱动、人机协同、跨界融合、共创分享的智能经济形态**。

中央企业要把发展人工智能放在全局工作中统筹谋划，深入推进产业焕新，加快布局和发展人工智能产业。

- 要夯实发展基础底座，把主要资源集中投入到最需要、最有优势的领域，**加快建设一批智能算力中心**，进一步深化开放合作，更好发挥跨央企协同创新平台作用。
- **开展AI+专项行动，强化需求牵引，加快重点行业赋能**，构建一批产业多模态优质数据集，打造从基础设施、算法工具、智能平台到解决方案的大模型赋能产业生态。

中央企业人工智能专题推进会：
10家中央企业签订倡议书，将主动向社会开放人工智能应用场景。

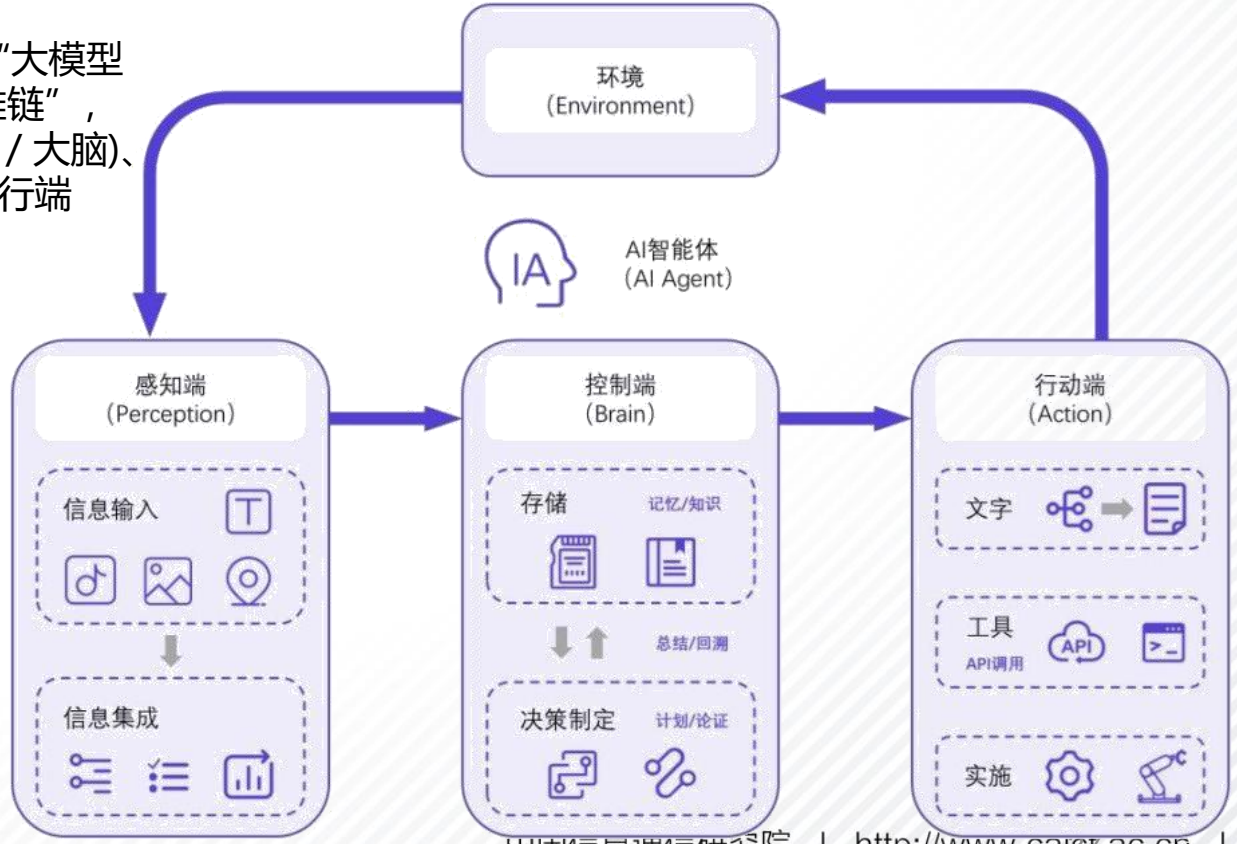
中央企业大规模设备更新工作推进会：
加快人工智能等新技术与制造全过程、全要素深度融合

目录

一、AI 智能体成为大模型应用新热点

- 从大模型应用案例征集情况看，AI Agent（智能体）相关案例占比超1/5（23%），成为大模型应用热点关键词。
- AI Agent是指人工智能代理（Artificial Intelligence Agent），它的官方定义是一种能够感知环境、进行决策和执行动作的智能实体。**以大型语言模型(LLM)作为其核心引擎**，它们能够感知其环境，做出决策，并执行任务以实现特定的目标，AI Agent的设计理念是赋予机器自主性、适应性和交互性，使其能够在复杂多变的环境中独立运作。

Agent 其实基本就等于 “大模型 + 插件 + 执行流程 / 思维链”，分别会对应控制端 (Brain / 大脑)、感知端 (Preception)、执行端 (Action) 环节。



二、知识库成为大模型落地主要辅助手段

- 知识库问答是知识及检索之上的一个应用场景，知识库和检索作为中间的能力层，可以应用的点会很多。因此，**知识库也是企业用户在落地的过程中，优先考虑以及选择比较多的能力场景。基于RAG技术搭建行业知识库，成为大模型落地主要辅助手段。**
- **RAG(检索增强生成)是一种结合信息检索和文本生成的技术**，旨在提高自然语言处理任务的性能。它可以从一个大型知识库中检索与输入相关的信息，并将这些信息作为上下文和问题一起输入给模型进行处理。通过这种方式，RAG可以帮助模型生成更加准确、相关和丰富的回答。简单说，**RAG就是先检索后生成，让模型“有据可依”**。

- **在传统企业知识库构建的整体流程中，** 企业需要人工整理常见问题解答（FAQ）并训练QA机器人。但由于它主要依赖预先设置的答案进行训练，导致机器人在实际应用中频繁出现语义解读错误，无法给出精确答案。此外，这类机器人的功能模型泛化能力较差，例如无法进行文本摘要、内容扩展或文本润色。用户需要在检索文档后，手动浏览获取相应信息。

- **使用大模型知识助手时，** 用户可以在查询时采用自然语言进行多轮自然交互，企业无需提前配置任务型的问答流程。大模型能够依据知识库的内容，做出更准确的回答。基于策略不同，大模型不仅支持独立推理并作出回答，同时支持完全基于常见问题解答（FAQ）给出所需答复。同时，加入大模型后，知识库类应用还可支持具有扩展性的问答，用户可以使用知识助手解决相对复杂的问题。

三、人工智能大模型助力企业向智能型企业演化

■ **依托AI大模型，实现企业生产效率和体验效果的“双效提升”。**

四、基础设施：智算基建的构建与运营需重点关注四个核心性能

- 智算集群的构建与运营，需要从**规模化算力部署**的角度，统筹考虑大模型分布式训练对于计算、网络和存储的需求特点，并集成平台软件、结合应用实践，充分关注数据传输、任务调度、并行优化、资源利用率等，设计和构建高性能、高速互联、存算平衡可扩展集群系统，以满足AI大模型的训练需求。

1 有效性

- 降低AI加速卡协同过程中的算力损耗，尽可能提升集群整体的有效算力。

2 稳定性

- 降低集群运行故障率，提升集群故障恢复效率，支撑大模型高效可靠训练。

3 低碳性

- 降低集群能耗成本，推进集群绿色低碳运行。

4 可用性

- 构筑全栈技术能力，提升集群服务性能，方便用户使用智能算力。

五、人工智能赋能应用场景持续拓展方向

人工智能应用场景不断丰富，在企业产品研发设计、生产制造、仓储物流、销售服务、经营管理各个环节逐步渗透，以降低生产成本、提高经营效率，实现智能化生产运营。

研发设计	生产制造	仓储物流	销售服务	经营管理
• **复杂产品研发设计**中，医药研发是典型领域：靶点药物研发、利用AI大模型进行软件开发等。 • **创成式设计**，复杂机械装备产品设计已具备创成式设计特点，可应用在汽车制造、机器人制造等领域。 • **产品可视化仿真**在产品高度定制化、市场竞争激烈、消费者决策过程复杂的领域应用广泛。	• **制造工艺优化**：通过数据挖掘和分析、优化工艺参数实现工艺最优化。 • **大规模个性化定制**：运用大数据与智能决策技术收集和分析客户的数据，实现个性化定制生产。 • **复杂产品质量检验**借助深度神经网络集成技术，用以识别产品表面缺陷。 • **智能工业协作机器人**具备学习能力与交互能力，与人类工人共同作业。	• **智能仓储**即实现仓库物品的自动识别、分类、数量统计等功能。 • **智能配送系统**是通过大数据分析、路线规划等技术，实现配送路线的优化、配送时间的预测等功能。 • **智能调度系统**则是实现车辆的自动调度、任务分配和路径规划等。	• **精准营销**通过提取分析海量的用户数据，把握消费者需求特点并预测客户需求。 • **预测性维护**利用人工智能技术将故障、异常情况以及维护方案提供给用户，帮助制定维护方案。 • 基于人工智能的**聊天机器人/智能客服**通过语音识别和语音合成技术，在预测分析和机器学习技术支持下，预测未来的需求，并提供相应的解决方案。	• **智能安全管理场景**中，通过智能化系统监测设备的工作状态和生产参数，及时发现异常情况，预测可能发生的故障和事故。 • **企业财务会计与财务管理**领域的人工智能应用催生了业财一体化软件、财务机器人等工具，以帮助财务自动化处理重复的工作。

CAICT 中国信通院

谢　谢

THANKS