

## 2026 人工智能发展白皮书

# The White Book on the Development of Artificial Intelligence in 2026

深圳市人工智能行业协会  
深圳市投控数字科技有限公司  
2026 年 6 月

# 2026

## The White Book on the Development of Artificial Intelligence

## 人工智能发展白皮书

深圳市人工智能行业协会、深圳市投控数字科技有限公司/组织编写

主 编 郎丽艳

编写人员 王潇桢 杨华健 宿亚浩

曾 鹏 陈丹霞

## 序 言

人工智能作为引领全球新一轮科技革命与产业变革的颠覆性技术，已成为国际竞争的新焦点、经济发展的新引擎、社会进步的新动能。2025 年，人工智能进入全面渗透、深度融合的新阶段，算力普惠、智能体自主执行、端侧智能普及、人形机器人商业化等趋势加速形成，全球人工智能竞争格局从技术竞赛演变为涵盖算力、算法、数据、能源、人才与治理的系统性工程。站在 2026 年这一关键节点，深刻把握人工智能发展态势，找准技术演进方向与产业变革路径，对于推动经济社会高质量发展具有重要意义。

本白皮书立足全球视野，系统勾勒人工智能行业发展全貌。第一章为年度复盘，以模型竞技、智能体爆发等六大领域的标志性事件，还原行业发展的年度坐标；第二章聚焦全球格局，拆解中美发展态势，剖析深圳产业实践，展现全球 AI 产业梯队与区域创新活力；第三章描绘产业生态，梳理 AI 产业链全景与代表企业，为理解产业提供微观视角；第四章深耕技术跃迁，围绕大模型、智能体、AI 芯片三大核心，剖析技术迭代的底层逻辑与突破路径；第五章围绕应用场景，以科学智能、工业智能、智能终端等领域为代表，呈现 AI 与经济社会深度融合的实践模式；第六章展望前沿未来，聚焦世界模型、太空算力、人形机器人等前沿方向，探索人工智能产业新可能；第七章回归风险思辨，研判人工智能发展挑战，探讨发展与安全动态平衡的治理路径。

技术演进永不止步，产业探索未有穷期。人工智能技术迭代的前方，既是规模化落地带来的广阔价值蓝海，也伴随着伦理、安全与可持续性的系统性命题。我们期待本白皮书能够成为政策制定者的决策参考、产业从业者的行动指南、学术研究者的观察窗口。我们愿与所有同行者一道，在创新中坚守安全底线，在发展中谋求全球共赢，共同推动人工智能向更普惠、更包容、更可持续的方向前行，为新一轮科技革命与产业变革贡献中国智慧与中国方案。

深圳市人工智能行业协会  
深圳市投控数字科技有限公司

## 2026 年人工智能七大趋势

### 趋势一：模型部署形态从云端向终端延伸

随着模型压缩、端侧芯片、终端操作系统和边缘计算能力的提升，AI 正与手机、电脑、汽车、可穿戴设备和各类物联网终端深度结合。端侧智能成为新趋势，原生 AI 终端日益普及，AI 助手从“语音助手”升级为“私人智能助理”，实现全场景、全时段的个性化智能服务。

### 趋势二：Token 趋于普惠，国产 Token 加速出海

智能体引爆 Token 消耗规模，伴随算法优化、硬件迭代及规模效应，Token 成本持续下降，近 2 年下跌 99.7%，预计未来将逐渐演变为“水电煤”化的普惠性基础要素。依托国内电力、算力成本优势，国产 Token 加速出海，有望打造数字服务贸易新形态。

### 趋势三：AI 能力从 AIGC 迈向 Agentic AI

AI 能力正在从 AIGC (生成式 AI) 阶段进一步走向 Agentic AI (代理式 AI) 阶段。过去几年，AI 能力主要体现在文本、图像、音频、视频和代码等内容生成能力，而 2025 年后，随着智能体技术的发展，AI 从辅助工具进一步升级为具备环境理解、自主决策与任务执行的“数字劳动力”。

### 趋势四：人形机器人迈入“商业化元年”

随着运动控制精度、续航能力、感知与决策能力的全面

升级，以及核心零部件国产化的重大突破，人形机器人整机成本显著下降，为大规模推广应用奠定基础，预计 2026 年我国人形机器人产能将达到 10 万至 20 万台，工业制造、仓储物流、汽车制造是当前核心落地场景。

### **趋势五：太空算力有望突破地面算力瓶颈**

人工智能算力需求的指数级增长构成了对能源、水、土地资源的巨大消耗。太空算力依托太空近乎无限的太阳能资源，以及真空低温环境构成的天然散热场，在绿色低碳方面优势显著。随着商业航天迈入成熟期，超大规模算力向太空迁移将成为必然趋势，算力产业有望开启“天地一体、协同计算”的新纪元。

### **趋势六：行业超额利润将逐渐转移到应用层**

Token 成本下降，推动 AI 产业的利润重心从模型层向应用层倾斜，溢价环节转移至能低成本解决垂直痛点、实现企业运营成本优化的 AI 企业。大模型与智能体在真实场景中的规模化应用成为产业界关注的焦点，预计未来 40% 的企业应用将嵌入任务型智能体，全面赋能千行百业。

### **趋势七：AI 治理成为全球合作新“交汇点”**

生成式 AI 的大规模应用，使得低质量 AI 内容、AI 幻觉、版权争议、数据合规、算法透明度、安全风险以及智能体自主行为失控等问题日益突出。各国正加速完善 AI 治理规则，力求实现发展与安全的动态平衡，我国主提的加强人工智能能力建设国际合作决议在联合国大会中获得广泛国际认可。

## 目 录

### 第一章 年度复盘：大事盘点与热词解析 ..... 1

#### 一、2025 年全球人工智能大事记 ..... 1

##### （一）模型竞技：开源崛起与性能突破 ..... 1

##### （二）智能体爆发：从“问答”进阶“执行” ..... 2

##### （三）具身智能落地：从“舞台”走向“工厂” ..... 4

##### （四）科学智能突破：AI 加速科学发现 ..... 4

##### （五）算力博弈：全球算力基建竞速 ..... 5

##### （六）全球治理：监管框架全面落地 ..... 6

#### 二、2025 年人工智能热词解析 ..... 7

##### （一）AI Agent 智能体 ..... 7

##### （二）Embodied Intelligence 具身智能 ..... 8

##### （三）Token 词元 ..... 8

##### （四）Multimodal AI 多模态大模型 ..... 8

##### （五）World Models 世界模型 ..... 9

##### （六）On-device AI 端侧智能 ..... 9

##### （七）Edge computing 边缘计算 ..... 10

##### （八）End to end 端到端 ..... 10

##### （九）Vibe coding 氛围编程 ..... 11

|                         |    |
|-------------------------|----|
| (十) AI Slop AI 泔水 ..... | 11 |
|-------------------------|----|

## 第二章 全球格局：双极引领与差异发展 ..... 12

### 一、美国人工智能发展态势 ..... 12

|                      |    |
|----------------------|----|
| (一) 市场规模位居全球首位 ..... | 12 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (二) 资本市场占据全球主导 ..... | 13 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (三) 学术领域主导地位减弱 ..... | 14 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (四) 算力产业生态持续领跑 ..... | 15 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (五) 软硬协同构筑生态壁垒 ..... | 16 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (六) 人工智能人才储备丰富 ..... | 16 |
|----------------------|----|

### 二、中国人工智能产业版图 ..... 17

|                      |    |
|----------------------|----|
| (一) 产业规模保持高速增长 ..... | 17 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (二) 创业浪潮呈现理性回调 ..... | 18 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (三) 技术创新能力全球领先 ..... | 19 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (四) 风险投资规模接近千亿 ..... | 20 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (五) 模型发布数量全球第一 ..... | 21 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (六) 算力规模位居全球第二 ..... | 22 |
|----------------------|----|

|                      |    |
|----------------------|----|
| (七) 数字经济生态持续活跃 ..... | 23 |
|----------------------|----|

### 三、深圳人工智能发展概况 ..... 24

|                      |    |
|----------------------|----|
| (一) 深圳人工智能战略部署 ..... | 24 |
|----------------------|----|

|                        |    |
|------------------------|----|
| (二) 核心产业规模快速增长 .....   | 26 |
| (三) 产业链条以应用为主导 .....   | 26 |
| (四) 人工智能创业热度回升 .....   | 27 |
| (五) 不同梯队企业协同并进 .....   | 28 |
| (六) 算力供给规模持续扩大 .....   | 31 |
| (七) 数据产业形成“三大名片” ..... | 35 |
| (八) 人才培养体系逐渐完备 .....   | 38 |
| (九) 人才需求侧爆发式增长 .....   | 43 |

### 第三章 产业全景：链条解析与生态构建 ..... 44

|               |    |
|---------------|----|
| 一、产业链全景 ..... | 44 |
| 二、基础层 .....   | 46 |
| 三、技术层 .....   | 47 |
| 四、应用层 .....   | 50 |

### 第四章 技术跃迁：关键领域的技术迭代 ..... 55

|                    |    |
|--------------------|----|
| 一、模型创新与演进 .....    | 55 |
| (一) 模型竞争格局 .....   | 55 |
| (二) 模型演进趋势 .....   | 59 |
| (三) 模型发展瓶颈 .....   | 64 |
| 二、智能体从概念走向实用 ..... | 68 |

|                                 |            |
|---------------------------------|------------|
| (一) 智能体产品演进 .....               | 68         |
| (二) 智能体技术架构 .....               | 73         |
| (三) 智能体落地场景 .....               | 74         |
| (四) 风险与趋势 .....                 | 77         |
| 三、AI 芯片国产替代加速 .....             | 80         |
| (一) AI 芯片竞争格局 .....             | 80         |
| (二) AI 芯片技术路线 .....             | 83         |
| (三) 发展瓶颈与趋势 .....               | 85         |
| <b>第五章 应用纵横：跨界融合与赋能实例 .....</b> | <b>88</b>  |
| <b>一、人工智能终端 .....</b>           | <b>88</b>  |
| (一) 发展概览 .....                  | 88         |
| (二) 细分领域进展 .....                | 91         |
| (三) 瓶颈与趋势 .....                 | 94         |
| <b>二、工业智能 .....</b>             | <b>95</b>  |
| (一) 发展概览 .....                  | 96         |
| (二) 技术路径 .....                  | 100        |
| (三) 重点场景赋能 .....                | 101        |
| (四) 趋势研判 .....                  | 103        |
| <b>三、智能增效 .....</b>             | <b>105</b> |

|                                 |            |
|---------------------------------|------------|
| (一) 发展概览 .....                  | 105        |
| (二) 重点场景赋能 .....                | 106        |
| (三) 瓶颈与趋势 .....                 | 108        |
| 四、智慧民生 .....                    | 111        |
| (一) 发展概览 .....                  | 111        |
| (二) 重点场景赋能 .....                | 112        |
| (三) 瓶颈与趋势 .....                 | 114        |
| 五、科学智能 .....                    | 115        |
| (一) 发展概览 .....                  | 115        |
| (二) 细分领域进展 .....                | 116        |
| (三) 瓶颈与趋势 .....                 | 120        |
| <b>第六章 未来展望：前沿突破与无限可能 .....</b> | <b>121</b> |
| 一、世界模型 .....                    | 122        |
| (一) 世界模型的概念 .....               | 122        |
| (二) 世界模型的技术路线 .....             | 123        |
| (三) 世界模型的发展趋势 .....             | 125        |
| 二、太空算力 .....                    | 128        |
| 三、人形机器人 .....                   | 131        |
| 四、量子计算 .....                    | 133        |

## 第七章 风险思辨：伦理、安全与可持续 ..... 136

### 一、资源消耗与可持续性 ..... 136

#### （一）资源消耗风险 ..... 136

#### （二）可持续发展路径 ..... 138

### 二、数字鸿沟与发展失衡 ..... 139

### 三、劳动转型与替代冲击 ..... 140

### 四、数据让渡与安全风险 ..... 142

### 关于协会 ..... 145

### 关于投控数科 ..... 147

### 法律说明 ..... 149

#### 一、版权声明 ..... 149

#### 二、免责声明 ..... 149

## 第一章 年度复盘：大事盘点与热词解析

### 一、2025 年全球人工智能大事记

#### （一）模型竞技：开源崛起与性能突破

1月20日，DeepSeek-R1发布，改写大模型“烧钱”规则。深度求索发布开源推理模型 DeepSeek-R1，单次训练成本仅为 29.4 万美元，其基础模型 DeepSeek-V3 的训练成本约 600 万美元，远低于行业水平。1月27日，受 DeepSeek 影响，英伟达市值单日蒸发约 5930 亿美元，创下美国上市公司单日市值损失最大的纪录。

4月5日，Meta 发布 Llama 4 系列模型，可一次性处理数百万文字。Llama 4 系列采用混合专家（MoE）架构，包含 Scout、Maverick 和 Behemoth 三款产品。Maverick 总参数 4000 亿，在创意写作、代码生成等基准测试中表现超过 GPT-4o 和 Gemini 2.0 Flash。Scout 支持 1000 万 token 上下文窗口，可一次性处理数百万字文本，能在单个英伟达 H100 GPU 上运行。

4月29日，阿里发布开源模型 Qwen3，被评为全球最强开源模型。旗舰模型 Qwen3-235B-A22B 采用混合专家（MoE）架构，总参数 2350 亿，激活参数量仅需 220 亿，预训练数据量达 36 万亿 tokens。该模型在 AIME25 奥数测评获 81.5 分，创开源纪录；LiveCodeBench 代码能力评测超 70 分，超越 Grok3。同年，Kimi、DeepSeek 等中国开源模型频繁霸榜 HuggingFace，中国开源模型全球使用份额由

2024 年 1.2% 升至 2025 年 30%。

8 月 5 日，OpenAI 发布开放权重推理模型 GPT-OSS。该模型包含 gpt-oss-120b 和 gpt-oss-20b 两个版本，采用 Apache 2.0 许可证，允许自由使用、修改和商业部署。CEO 阿尔特曼承认来自中国开源模型的竞争影响了这一决定，表示“如果我们不这样做，世界将主要建立在中国的开源模型之上”。

8 月 7 日，OpenAI 发布 GPT-5，整合推理与通用能力。GPT-5 整合了高效基础模型、深度推理模块和实时路由系统，能够根据不同情况判断何时该快速回应，何时应进行深度推理思考以提供专家级的答案。在编程基准 SWE-bench Verified 上得分为 74.9%，超过 o3 的 69.1% 和 GPT-4o 的 30.8%；在数学基准 AIME 2025 上达到 94.6%，展现极强的数学推理能力。

11 月 18 日，谷歌发布 Gemini 3，突破多模态极限。该模型发布后以 1501 高分迅速登顶 LMArena 排行榜，超越 GPT-5.1、Claude 4.5 与 Grok-4.1，刷新 LMArena 公开榜单历史最高分。在学术界用来测试 AI 系统是否具备博士级推理能力的“Humanity's Last Exam”（人类终极考试）中，Gemini 3 达到了 37.5% 的准确率。

## （二）智能体爆发：从“问答”进阶“执行”

3 月 6 日，全球首款通用 AI 智能体 Manus 发布，开启自主执行时代。Manus 由中国创业公司 Monica 研发，采用

多智能体架构，具备从任务规划到执行的全流程自动化能力，在 GAIA 基准测试中 Level 3 任务完成率达 47.6%，超越 OpenAI 同类产品。Manus 发布后 9 个月内实现 1 亿美元 ARR（年度经常性收入）。

**5 月，Anthropic 发布 Claude Code。**Claude Code 是 Anthropic 推出的终端原生 AI 编程智能体，能根据自然语言指令拆解任务、调用工具，帮助开发者高效完成代码编写、调试、重构和项目管理任务。在 SWE-bench 软件工程基准测试中 72.5% 的准确度，被业界评为“全球最强编程模型”，2025 年年化收入超过 25 亿美元。

**10 月，OpenAI 全面开放编程工具 Codex。**2025 年 5 月，OpenAI 发布 Codex 研究预览版，定位为基于云的软件工程智能体，优先向 ChatGPT Pro、Enterprise 和 Team 用户开放。10 月，OpenAI 在年度开发者大会（DevDay 2025）宣布 Codex 对一般用户开放使用，标志着 Codex 从前期的测试试点阶段走向了更广泛的应用场景。

**11 月，奥地利程序员 Peter Steinberger 发布 ClawdBot，开启 OpenClaw 雏形。**2026 年 1 月 30 日正式定名为 OpenClaw（国内惯称“龙虾”），2026 年 2 月 OpenClaw 凭借创新的自主决策架构与多模态环境适应能力火爆全球，2026 年 4 月，OpenClaw 的 Github 星标数突破 35 万，创下 Github 历史上最快星标增长速度，引发全球广泛关注，成为此后大量同类产品命名与形态上的源头。

### (三) 具身智能落地：从“舞台”走向“工厂”

3月，优必选 Walker S1 人形机器人批量进入汽车工厂。东风柳汽宣布将在其商用车智能制造工厂部署 20 台 Walker S1，并计划于 2025 年上半年完成部署，被业界视为全球首次人形机器人批量进入汽车工厂。2025 年全年优必选累计获得超过 500 台意向订单，主要来自汽车、3C、物流等行业。

12月8日，智元机器人第 5000 台通用具身机器人量产下线。这一里程碑标志着具身机器人已从技术验证阶段全面迈入规模商用时代。此前智元以初创公司身份反向控股二级市场上市公司上纬新材，引发市场广泛关注。

12月15日，中国首批 L3 级自动驾驶正式获批上路。工业和信息化部正式公布我国首批 L3 级有条件自动驾驶车型产品准入许可。长安汽车、北汽集团两款纯电动轿车获准在北京、重庆指定区域内开展上路通行试点。这一举措标志着我国自动驾驶技术从示范测试迈入真实道路运行和商业化探索的新阶段。

### (四) 科学智能突破：AI 加速科学发现

1月23日，“Humanity's Last Exam”基准测试发布。该测试是由人工智能安全中心（CAIS）与 Scale AI 联合发布的多模态基准测试，其题库包含 2500 至 3000 道高难度题目，覆盖数学、自然科学、人文社科等 100 多个学科，旨在评估前沿 AI 系统的综合能力。测试发布之初，顶级 AI 模型得分率不足 10%；截至 2025 年底，顶级 AI 模型得分已提升至

40%-50%。

2月，人工智能“从零开始”设计蛋白酶。诺贝尔奖得主、蛋白质设计先驱 David Baker 教授及团队吴可嘉、刘彩璇在《科学》期刊、《自然》期刊发表里程碑成果，利用 AI “从零开始”设计具有复杂活性位点的全新酶分子，突破了自然演化的结构限制，为药物研发和蛋白质工程开辟新路径。

9月，DeepSeek-R1 推理模型研究论文登上《自然》封面。论文披露 R1 单次训练成本仅 29.4 万美元，构建基础模型 DeepSeek-V3 的训练成本仅为 600 万美元，在国际学术界引起轰动。DeepSeek-R1 是全球首个经过同行评审的主流大语言模型。Nature 评论指出，目前几乎所有主流大模型尚未经历独立同行评审，而 DeepSeek 则率先打破了这一空白。

12月，StructBioReasoner 在药物设计领域取得突破。阿贡国家实验室与芝加哥大学团队推出针对“不可成药”靶点蛋白的生物制剂设计系统 StructBioReasoner。该系统依托锦标赛式推理框架，通过多智能体竞争生成并优化靶向方案。在针对 Der f 21 蛋白的 787 个候选设计中，系统超过 50% 的性能优于人类设计的参考方案，在难度更高的 NMNAT-2 靶点上，系统从 97066 个候选中筛出三类结合模式。该成果为依托百亿亿级超算开展 IDP 靶向药物筛选奠定了基础。

### **(五) 算力博弈：全球算力基建竞速**

1月21日，特朗普宣布“星际之门计划”启动。OpenAI、日本软银集团和美国甲骨文公司联合发起“星际之门计划”，

该项目的初始投资为 1000 亿美元，并计划在未来 4 年内扩展至 5000 亿美元。首个建设的数据中心将设在美国得克萨斯州，随后将扩展到其他州。特朗普称其“史上最大的 AI 基础设施投资项目”，预计创造超过 10 万个就业岗位。

4 月 10 日，华为发布昇腾 920 芯片。华为在云生态大会上宣布新一代 AI 芯片昇腾 920 (Ascend 920) 已研发成功，计划下半年正式量产。该芯片配备 HBM3 高带宽内存 (4TB/s)，BF16 算力达 900TFLOPS，训练效率较昇腾 910C 提升 30%-40%，专为 Transformer 和 MoE 大模型优化。其主要目标是替代受限的英伟达 H20 芯片，并在性能上达到或接近 H100 水平。

11 月 14 日，谷歌宣布投资 400 亿美元用于 AI 基础设施投资。谷歌宣布在德克萨斯州投资 400 亿美元，用于 AI 与云计算基础设施建设，计划 2027 年前建成三座新数据中心。该投资属于谷歌 “Investing in America” 计划，是其在美国单一州的最大规模投资。

## **(六) 全球治理：监管框架全面落地**

2 月 2 日，欧盟《人工智能法案》禁止性条款正式生效。被禁止的人工智能实践包括采用潜意识操纵技术、社会信用评分、公共场所实时远程生物识别等，违规罚款最高可达全球年营收 7%，这类禁止条款已于 2025 年 2 月起率先强制执行。

8 月 2 日，欧盟《人工智能法案》核心条款全面生效。

通用目的人工智能模型提供者义务、治理机制、处罚机制等核心条款正式适用，面向欧盟市场的模型提供者须履行透明度与合规义务。

8月，联合国建立两大 AI 治理机制。联合国大会通过决议正式设立“独立国际 AI 科学小组”和“全球 AI 治理对话”平台，首次实现 193 个成员国全覆盖参与 AI 规则制定。

9月1日，《人工智能生成合成内容标识办法》在中国正式施行。该办法由国家互联网信息办公室等四部门联合发布，覆盖所有利用 AI 生成的文本、图片、音频、视频、虚拟场景等内容，明确“显式标识”和“隐式标识”双轨机制，首次将传播平台纳入责任主体。

## 二、2025 年人工智能热词解析

### （一）AI Agent 智能体

一个能够理解目标、自主规划、调用工具、执行任务以及结果生成的 AI 系统。2025 年初，Manus 的走红使通用型 AI Agent 进入大众视野；随后，以扣子空间为代表的平台推动 Agent 从单一工具形态向工作空间与生态化应用演进；及至年末，OpenClaw 等开源项目的爆火也反映出开发者对可控、开放、本地化 Agent 的强烈诉求。当前，AI Agent 正在办公、编程、调研、客服、数据分析等场景加速落地，并推动软件从人机协作的方式转向 AI 全流程执行任务。其产业价值巨大，但稳定性、安全性和权限治理仍是关键挑战。

## （二）Embodied Intelligence 具身智能

具身智能是指拥有物理或虚拟身体，并能在环境中感知、决策和行动的智能系统。系统需要主动获取环境信息，基于不确定状态进行任务规划，并通过机器人、执行器或虚拟身体完成动作。随着多模态大模型、世界模型、VLA 模型（视觉—语言—动作模型）、模仿学习和强化学习的发展，具身智能成为 AI 从数字空间走向物理世界的重要方向。对产业而言，它将推动人形机器人、工业机器人、智能制造、物流、家庭服务和智能驾驶等场景升级。

## （三）Token 词元

指大模型处理语言和多模态信息的基本计量单位。2026 年 3 月，在中国发展高层论坛年会上 Token 被正式译为“词元”。大模型并不是直接按完整句子理解内容，而是将文本、代码、图片或语音等信息拆分为可计算的片段，再进行编码、推理和生成。一个 Token 可以是一个汉字、一个英文单词、一个符号、也可以是图像或者音频被离散化的表示单元。随着大模型应用普及，Token 成为衡量模型上下文长度、推理成本、响应速度和计费方式的重要指标。

## （四）Multimodal AI 多模态大模型

指能够同时理解和处理文本、图像、语音、视频等多种类型信息的 AI 模型。与主要依赖文字交互的传统大模型不同，多模态大模型更接近人类感知世界的方式，能够“看、听、读、说”并进行综合判断。随着 GPT-4o、Gemini、Qwen、豆包等多模态模型能力提升，多模态大模型成为基础模型竞

争的重要方向。它正在推动大模型从单一文本生成走向复杂场景理解，在教育、医疗影像、智能客服、工业质检、内容创作、智能驾驶和机器人等领域加速应用。其产业价值在于拓展大模型的感知边界和应用场景，但数据质量、算力成本、模型幻觉和跨模态一致性仍是主要挑战。

### **(五) World Models 世界模型**

指能够理解环境、预测未来、规划行动的生成式 AI 模型。它不仅识别当前环境中的对象、状态和关系，还能够推演“如果采取某种行动，环境会如何变化”。因此，世界模型被认为是 AI 从语言智能走向物理智能的重要基础。在自动驾驶和具身智能机器人快速发展的当下，世界模型成为行业关注的关键概念。它能够帮助 AI 在复杂环境中进行感知、预测、规划和决策，尤其适用于智能驾驶、机器人操作、仿真训练、游戏智能体和工业控制等场景。世界模型能够提升 AI 对真实世界的理解和泛化能力，具备极高的产业价值，但构建高质量世界模型面临数据获取、物理规律建模、实时推理和安全验证等挑战。

### **(六) On-device AI 端侧智能**

指 AI 模型在手机、电脑、汽车、可穿戴设备、机器人和物联网终端等本地设备上运行，而不完全依赖云端服务器完成推理。端侧智能是大模型从云端算力中心走向物理世界入口的重要口径，其合理性来自隐私安全、低延时、带宽成本和实时交互需求。随着 AI 手机、AI PC、智能座舱以及 OpenClaw 催生的本地硬件需求持续升温，终端设备正从云

端延伸转向具备感知、记忆与执行能力的智能载体。从产业视角来看，端侧智能将重构传统终端市场格局，带动 Soc、NPU、存储、操作系统和端云协同生态升级。

### (七) Edge computing 边缘计算

指将计算、存储和数据处理能力从中心云端下沉到靠近数据产生和使用现场的边缘节点，如基站、网关、工业设备、车辆、园区服务器和终端设备附近。它的核心价值在于降低网络时延、减少数据回传成本、提升实时响应能力，并增强数据安全与隐私保护。随着 AI 应用从云端走向物理世界，边缘计算成为端侧智能、工业互联网、智能驾驶、智慧城市和物联网场景的重要基础设施。

### (八) End to end 端到端

指 AI 系统从原始输入到最终输出由统一模型或整体架构直接学习和优化，减少中间人工干预和分模块处理。它强调用数据驱动方式完成从感知、理解到决策的完整流程。目前，端到端成为自动驾驶、具身智能机器人、语音识别、推荐系统和多模态模型中的高频概念。在智能驾驶领域，端到端方案被视为提升系统泛化能力和降低工程复杂度的重要方向；在具身智能机器人领域，端到端 VLA 模型也成为重要代表，即模型将视觉感知、语言理解与动作形成整合起来，使机器人能够根据环境信息和自然语言指令直接完成动作决策。就产业而言，端到端有助于减少传统模块化系统中的误差传递和规则维护成本，使 AI 更适应复杂动态场景。

## （九）Vibe coding 氛围编程

指开发者用自然语言描述产品想法、功能需求和交互效果，由 AI 辅助生成、修改和调试代码的一种编程方式。它弱化了传统编程中逐行写代码的过程，更强调意图表达、快速试错、持续调整的开发模式。2025 年，随着 Cursor、Claude Code、GitHub Copilot、Replit 等 AI 编程工具普及，Vibe coding 成为软件开发领域的高频热词。它降低了应用原型和小型产品开发门槛，使非专业程序员也能参与软件创造，并显著提升开发效率。但与此同时，代码质量、安全漏洞、架构可维护性和开发者对底层逻辑的理解仍是主要风险。因此，Vibe coding 更适合快速原型打造、个人工具和早期产品验证，而在生产级系统中仍需工程审查和专业把关。

## （十）AI Slop AI 泔水

指由人工智能批量生成的低质量、重复化、缺乏事实依据或审美价值的内容，常见于文章、图片、视频、评论、商品描述和社交媒体信息流中。它反映出 AIGC 普及后内容生产门槛下降后所带来的副作用。随着生成式 AI 工具的广泛应用，大量低成本 AI 内容涌入网络，可能稀释优质信息，制造误导，AI Slop 逐渐成为内容平台、媒体行业和搜索生态中的重要问题。

## 第二章 全球格局：双极引领与差异发展

2025 年，全球人工智能市场规模达 7575.8 亿美元，人工智能企业数量约 3.8 万家，产业进入技术、应用、生态协同共振的高速发展期。在这一进程中，全球 AI 竞争格局呈现出清晰的梯队格局特征。第一梯队由美国和中国组成，美国在市场规模、算力生态、资本积累、顶尖人才储备等维度占据全球领先地位，中国则凭借独特的场景、数据、成本、科研和体制优势，走出了一条差异化的发展道路。第二梯队包括英国、加拿大、以色列、法国、德国、日本、新加坡等国家，它们各具特色：英国侧重基础研究与伦理治理；加拿大与以色列强于学术创新与初创孵化；德日聚焦工业 AI 与机器人；新加坡则致力于国际化枢纽与合规落地。第三梯队为印度、韩国、澳大利亚等，正依托市场与资源发力垂直应用，但整体生态与核心技术仍有差距。

### 一、美国人工智能发展态势

#### （一）市场规模位居全球首位

美国凭借科技巨头主导的产业生态和深厚的资本积累，持续巩固其领先地位。2025 年美国人工智能企业数量预估在 13725-15500 家之间，占全球人工智能企业总数的 36%-40%。其中，约 7000 家为初创公司，占美国人工智能企业数量的 50% 左右。2023 至 2025 年，美国人工智能企业数量年复合增长率约为 18%-22%，2025 年美国新增获得融资的人工智能企业达 1953 家、新成立的人工智能初创企业 194 家。得

益于硅谷等创新生态的推动，以及风险资本对美国人工智能领域的持续投入，预测 2025 年美国人工智能市场规模将超过 3500 亿美元，占全球人工智能市场规模比例接近 50%。

## （二）资本市场占据全球主导

**美国人工智能领域风险投资占全球主导，资本向头部公司集中。**早期，美国人工智能领域风险投资规模小（2012 年单笔均值 608 万美元），2014 年金额同比激增 110.8%，源于核心技术突破和科技巨头布局。中期（2017-2021 年）进入高速扩张期，2021 年风投金额达 1138.4 亿美元，反映疫情催化的数字化转型红利及宽松货币政策驱动。但 2022 年金额骤降 47.2%，暴露资本泡沫：美联储激进加息导致流动性紧缩，经济不确定性加速资本退潮。2024 年市场信心恢复，投资再度活跃，2025 年金额骤增至 1803 亿美元，同比增长 82.9%，创历史新高，占全球 AI 风险投资金额超 70%，但风投数量仅增长 22.0%，平均单笔金额高达 5177 万美元，头部公司 OpenAI、Anthropic 和 xAI 融资合计约 600 亿美元，反映出三大现实：其一，美国占据全球 AI 风投主导地位；其二，资本向头部公司集中，市场倾向于在技术成熟期进行“重注押注”，但过度集中可能抑制其创新多样性。

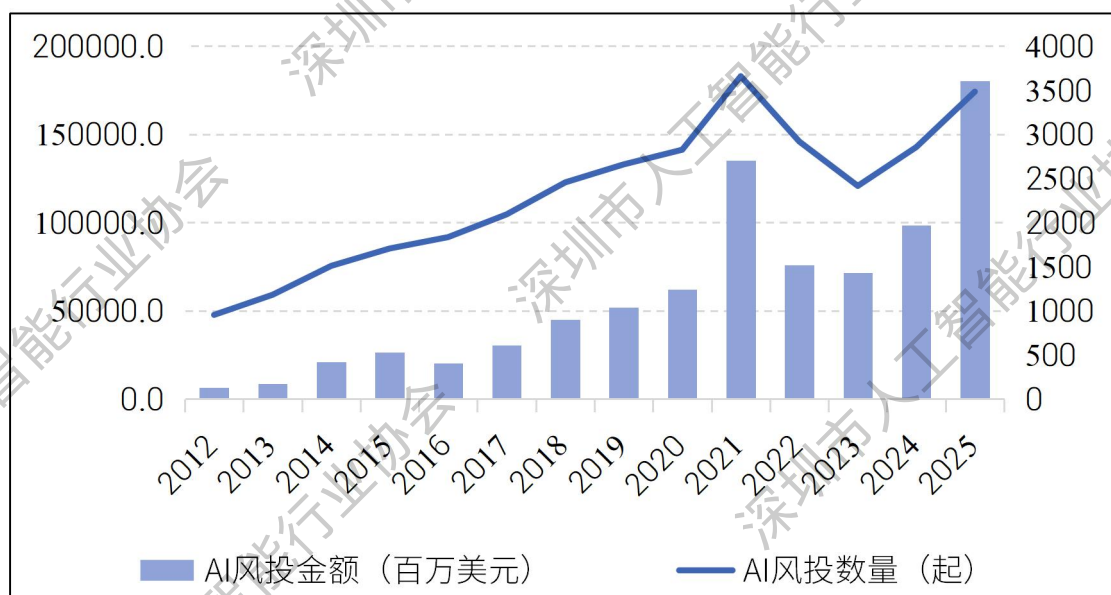


图 1 近年来美国人工智能领域风险投资变化（单位：亿美元，起）

数据来源：OECD.AI，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

### （三）学术领域主导地位减弱

美国人工智能出版物数量回升，但全球占比降低。前期（2012 年到 2021 年），美国人工智能出版物数量快速增长，由 3531.8 份攀升至 9847.3 份的峰值，其中 2018 年和 2019 年增速尤为显著，分别达到 35.7%和 25.6%，这得益于深度学习技术突破引发的学术热潮及美国国家科学基金会持续资助。2022 年出现负增长，跌幅达 22.7%，或与疫情后科研资源再分配、产业界人才向企业转移有关。2023 年短暂反弹 12.6%源于生成式 AI 突破带来的理论创新需求，2024 年暴跌 19.3%暴露学术研究体系的结构矛盾，2025 年显著回升 19.8%，达到 8284.9 份，反映出在全球生成式 AI 热潮下，美国政府、高校及企业加大基础研究与产业化布局，但全球占比下降至 9.5%，显示中国、印度加速追赶。

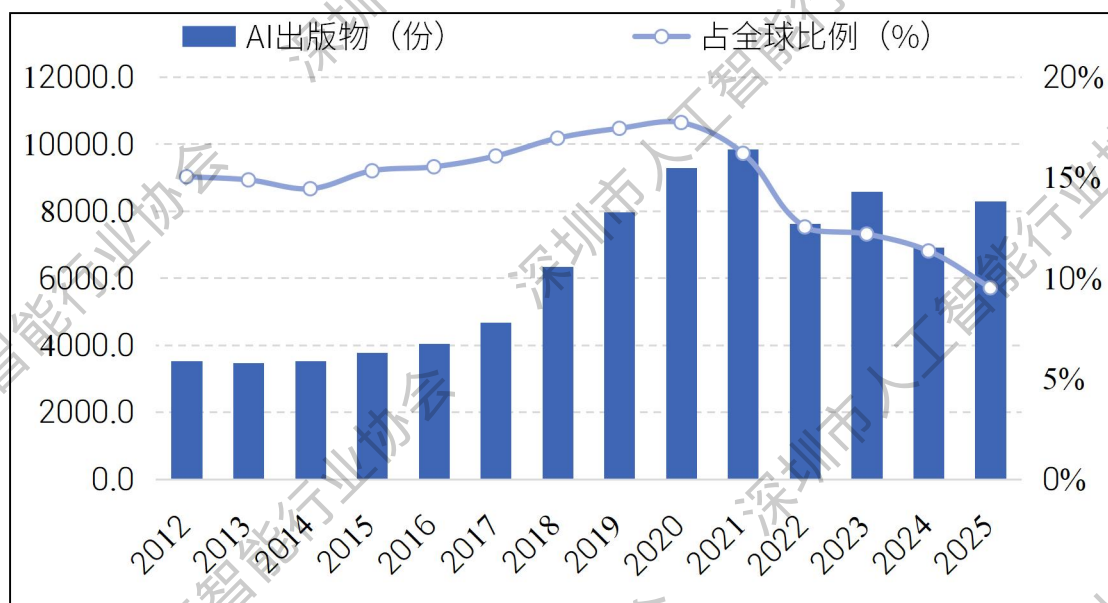


图 2 近年来美国人工智能出版物发表数量及份额变化(单位:份,%)

数据来源: OECD.AI, 深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

#### (四) 算力产业规模持续领跑

**美国主导全球 AI 算力生态,科技巨头竞逐百万卡智算集群。**美国拥有 5427 个数据中心, 占全球近 45%, 数量居世界首位, 算力容量占全球 60%; 在前沿 AI 算力领域, 美国控制了约 75% 的全球供给, 高端 GPU 集群容量同样占全球 75% 左右。在市场格局上, 美国算力生态呈现出由科技巨头主导的特征, 微软、亚马逊、谷歌与 Meta 四家公司掌控着美国超大规模数据中心超过 80% 的算力资源。同时, 美国科技巨头正推动智算集群从十万卡级向百万卡级迈进: xAI 公司 Colossus 集群已部署 23 万张英伟达芯片, 正在建设的 Colossus 2 超级集群首期规模为 55 万张, 建成后将成为全球首个吉瓦级智算集群; OpenAI 计划建设 10GW 数据中心,

等效 400-500 万张芯片，第一阶段预计 2026 年下半年部署。与此同时，美国面临的能源压力同样显著，数据中心激增的电力消耗已引发局部电网紧张。

### **（五）软硬协同构筑生态壁垒**

美国在 AI 芯片、软件生态等环节占据全球主导地位，形成了强大的生态垄断效应。英伟达占据全球 AI 芯片市场份额超过 80%，2026 年 2 至 4 月其数据中心业务（主要为 GPU 与加速器）收入达 752 亿美元，同比增长 92%。更具壁垒的是软件生态。英伟达的 CUDA（并行计算平台与编程模型）经过十余年积累，已覆盖全球 90% 以上的 AI 开发者，这种生态粘性使客户即便面临更具性价比的替代方案，也因巨大的迁移成本而难以切换：从 CUDA 生态切换到其他平台，意味着需要重写十年积累的经验、工具和代码。同时，PyTorch（Meta）、TensorFlow（Google）、Hugging Face 等主流 AI 开发框架与开发者社区也均由美国企业或机构主导，形成“硬件+软件+开发者社区”的闭环体系，进一步强化了美国在人工智能底层生态的整体控制力。

### **（六）人工智能人才储备丰富**

2025 年美国人工智能领域的人才规模同比增长 33%，达到 11.74 万人。从人才结构来看，产业界吸纳了约 9.5 万人，增长了 31%，学术界储备了约 2.24 万人，增长了 44.5%，为后续技术突破积蓄后劲。学术界人才中，有 2700 余人在 NeurIPS、ICML、ICLR 等顶级人工智能会议发表论文，占全球顶尖人工智能研究人员的 58.5%，其中有 1600 余人来自

于中国、印度及欧洲，反映了美国顶尖人才储备丰富，而这种优势建立在人才引进之上，对移民的限制将削弱人工智能研究能力。从技术栈分布来看，工程类人才占比达 73.8%，表明 AI 行业全面迈入大规模生产级部署周期。从企业分布来看，美国 AI 人才高度集中于科技巨头，Meta 以 9316 人位居榜首，紧随其后的是 Google（8704 人）、Amazon（5899 人）、Microsoft（3805 人）以及 Apple（3001 人），美国科技巨头核心人才 70% 的收入来自资本市场股票收益，全球资本的支持使其在人才争夺中形成“掐尖效应”。

## 二、中国人工智能产业版图

### （一）产业规模保持高速增长

中国人工智能产业规模保持高速增长。截至 2025 年底，我国人工智能企业数量超 6200 家，同比增长 27%，占全球总量的 16% 左右，产业集聚效应持续增强；人工智能核心产业规模预计超过 1.2 万亿元，同比增长 30%，净增长超 3000 亿元，年增量与 2020 年全年的产业规模相当，形成了覆盖基础底座、模型框架、行业应用的完整产业体系。在“人工智能+”行动全面实施的引领下，凭借全球最大制造业体系、最大数字消费市场、第二大商品消费市场的场景优势，我国人工智能产业正从技术探索向规模化商业应用加速跨越，深度融入制造、金融、政务、医疗、电商、出行等各领域，商业化效率全球领先。

## （二）创业浪潮呈现理性回调

创业活动呈现先陡升、后陡降、再企稳的特征，反映出行业从粗放增长向高质量发展过渡。2012 至 2018 年中国人工智能企业年成立数量从 155 家激增至 2124 家，年均复合增长率达 54.7%。这一阶段驱动因素包括：政策层面利好不断释放，技术层面深度学习算法突破，资本层面风险投资热潮推动创业活跃。2019 年成为行业拐点，当年新增企业同比减少 7.9%，此后进入长达六年的萎缩期，2022 年美联储加息引发全球流动性紧缩，叠加国内经济增速放缓，AI 创业进入“出清”阶段，企业成立数量腰斩式暴跌 50.1%。2023 年 ChatGPT 引爆全球大模型浪潮，AI 创业公司涌现，2025 年降幅收窄至 7.3% 以内，AI 创业活动进入理性回调阶段。该趋势反应出中国 AI 行业发展逻辑转变：从追求数量扩张转向质量提升，技术落地能力成为核心竞争要素。

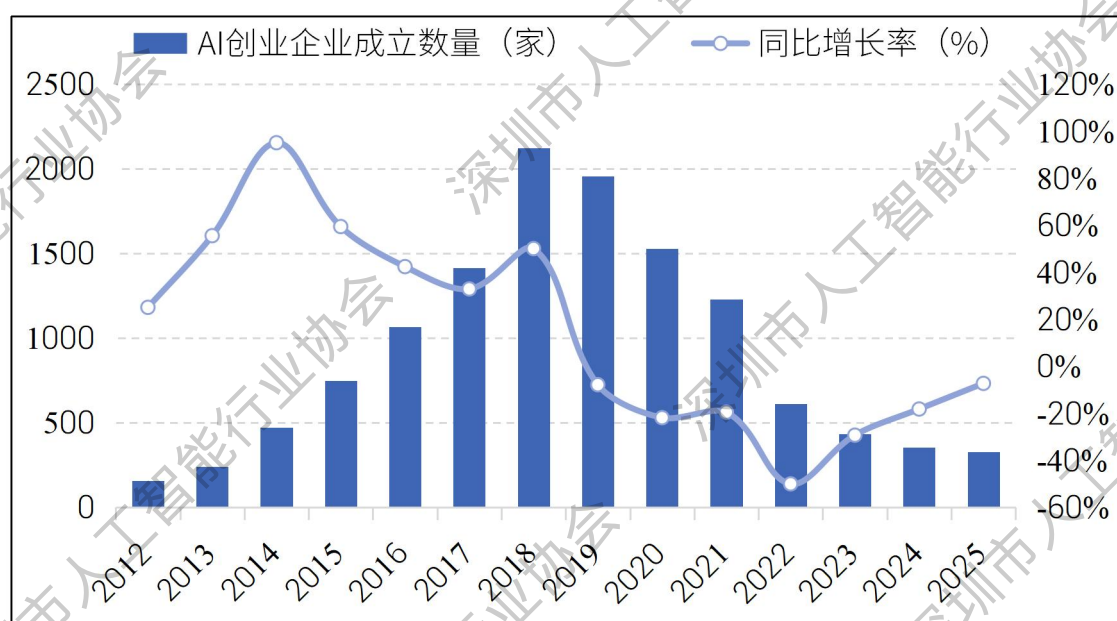


图 3 近年来中国人工智能企业成立数量变化（单位：家，%）

数据来源：IT 桔子，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

### （三）技术创新能力全球领先

**专利授权量、论文发表量和引用率均位居全球首位。**专利方面，世界知识产权组织（WIPO）及国家知识产权局数据显示，中国已成为全球人工智能专利最大拥有国，人工智能专利数量占全球总量的 60%，专利申请量和授权量均连续多年位居全球首位。AI 出版物方面，中国人工智能出版物产出呈现“波动增长—加速突破—高位回调”三阶段特征，2012 至 2016 年受国际竞争挤压及资金投入尚未到位影响，中国 AI 出版物发表数量小幅震荡，全球占比从 22.7% 降至 18.7%，2017 至 2024 年在政策牵引与算力基建支撑下，科研效能加速释放，2024 年 AI 出版物规模较 2012 年增长 361%，全球占比突破 30.7%。2025 年 AI 出版物数量增至 15696 篇，论文发表量、引用率均位居全球首位，但全球占比下降至 18%，反映出全球各国 AI 科研投入的竞争态势。

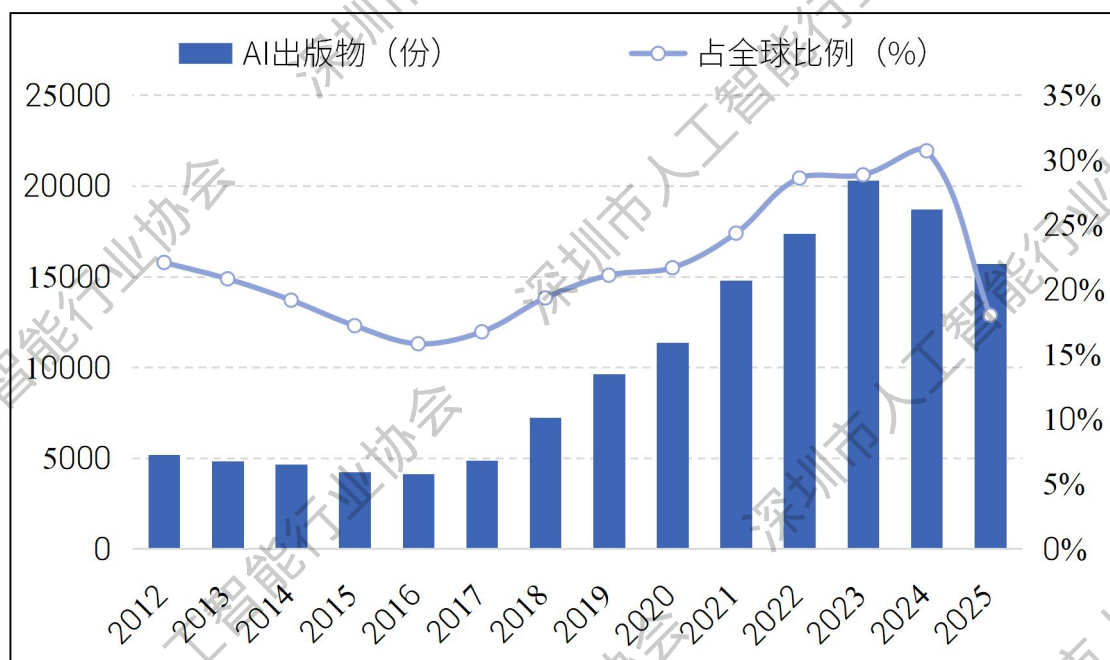


图 4 近年来中国 AI 出版物发表数量及份额变化（单位：份，%）

数据来源：OECD.AI，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

#### （四）风险投资规模接近千亿

中国人工智能领域风险投资金额下降但数量增长，资本青睐落地场景。2012 至 2021 年间，投资金额从 1.9 亿美元激增至 519.8 亿美元，复合增长率约 86.5%，投资数量增长 26 倍。这主要得益于深度学习技术突破、国家政策扶持及数字经济发展需求的叠加效应。2021 年投资额创历史新高，反映资本对 AI 在疫情催生的数字化转型中应用场景扩展的较高预期。但自 2022 年起，受全球经济承压、地缘政治（如美国颁布对华人工智能投资限制规则）、监管趋严及技术商业化瓶颈影响，投资金额连续四年收缩，2024 年较峰值已下降 68%，投资数量也呈现同步收缩态势。2025 年美国对华投资限制措施生效，导致美元风险投资萎缩，中国风险投资

金额下降至 156.9 亿美元，风投数量增长至 2098 起，单笔投资强度下降至 748 万美元，表明以中小额为主的人民币或国资基金成为主力，大额融资集中在 AI 芯片、大模型、具身智能、AI 软件应用等细分领域，反映出资本对人工智能基础设施以及在实际场景落地的支持。

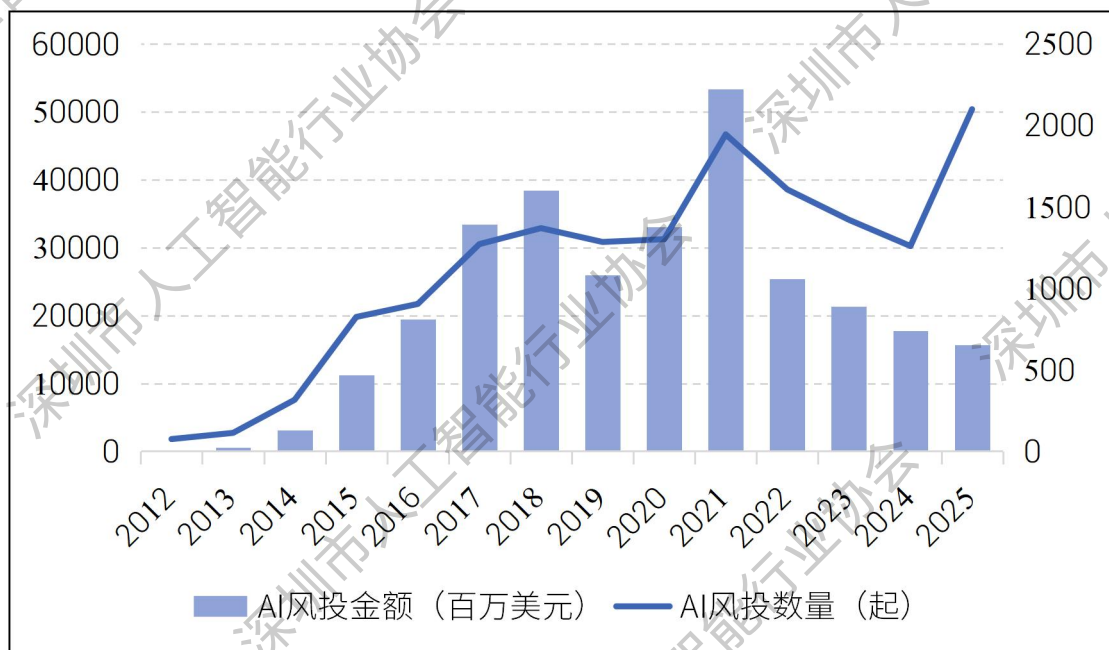


图 5 近年来中国人工智能领域风险投资变化（单位：亿美元，起）

数据来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

### （五）模型发布数量全球第一

截至 2025 年 7 月，全球共发布 3755 个人工智能模型，我国以 1509 个人工智能模型居首，占比 40%，覆盖 50 个重点行业领域、700 余个场景。在 Epoch AI 收录的重要人工智能模型数据集中，2025 年中国发布了 31 个重要人工智能模型，位列全球第二，仅次于美国（50 个）。在 MMLU、MMMU、MATH 和 HumanEval 等比较基准中，中美顶级大模型性能差距已从 2023 年的 17.5%、13.5%、24.3%和 31.6%，迅速

收窄至 2025 年的 0.3%、8.1%、1.6%和 3.7%。以 DeepSeek 为代表的国产大模型在训练效率上实现重大突破，DeepSeek V3 模型的单次训练成本约 600 万美元，仅为 GPT-4 的 1/10，成为“低算力高性价比”模型的典型代表。2025 年我国公有云大模型对客侧 Token 调用量有望达 2000 万亿，较 2024 年增长超 16 倍，头部云厂商调用量已处全球第一梯队。2025 年我国生成式人工智能用户规模达 6.02 亿人，较 2024 年底增长 141.7%；普及率达 42.8%，同比大幅提高 25.2 个百分点。

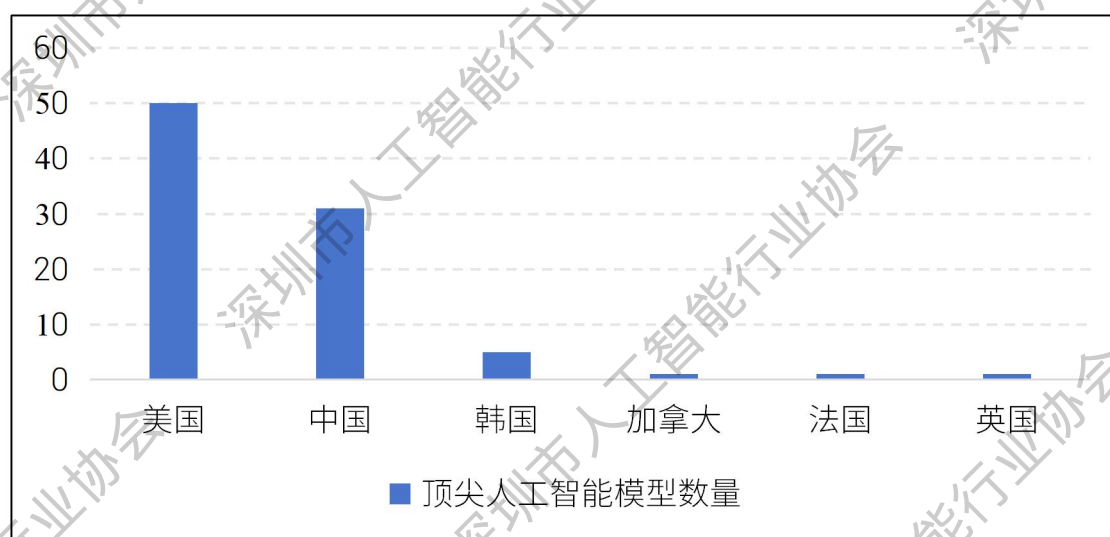


图 6 2025 年按国籍划分的重要人工智能模型分布

数据来源：Epoch AI，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

## （六）算力规模位居全球第二

在国家“东数西算”战略工程牵引下，全国一体化算力网络布局全面展开。截至 2025 年 12 月，我国已建成 42 个万卡智算集群，智能算力规模超过 1590EFLOPS（百亿亿次/

秒），全国在用算力中心标准机架超 1000 万架、已建和在建的智算中心超 1000 个，算力基础设施体系规模位居全球第二，智能算力在总算力中所占的比重从 2016 年的 3% 快速提升至 2025 年的 35%。与此同时，中国成本优势显著，工业平均电价约 0.08-0.1 美元/千瓦时，远低于美国的 0.12-0.15 美元/千瓦时，在西部等绿电资源丰富地区，电价可低至 0.05 美元/千瓦时，大幅降低数据中心能耗成本。

### （七）数字经济生态持续活跃

**中国拥有全球最大的网民规模和活跃的数字经济生态。**

从数据广度来看，截至 2025 年，我国总人口为 14.05 亿，移动电话用户总数达 18.18 亿户，网民规模超 11 亿人，每天都有大量的数据被生产和收集。根据国家数据局《全国数据资源调查报告（2025 年）》，2025 年我国数据生产总量达到 52.26ZB，相当于全国所有算力中心存储容量的近 30 倍，数据生产总量占全球比重约 27.44%，同比增长 27.28%。其中，工业数据、软件和信息技术服务业数据占比达到全行业的“半壁江山”；具身智能和低空经济领域数据生产量同比分别增长 477.78%、75%，表现突出。从数据多样性来看，我国是一个多民族、多语言、多文化国家，地域广阔，经济社会发展水平差异大，经济结构复杂，产业门类齐全，形成了多样化数据集合。从数据供给来看，截至 2026 年一季度，全国已建成高质量数据集超过 11.6 万个，总体量超过 960PB，相当于中国国家图书馆数字资源总量的 336 倍左右，数据供给规模及质量均显著提升，数据赋能人工智能成效日益显现。

### 三、深圳人工智能发展概况

#### （一）深圳人工智能战略部署

深圳在我国人工智能发展全局中扮演着先行示范的关键角色。2019 年，深圳获批建设国家新一代人工智能创新发展试验区，成为我国首批四个试验区之一。2021 年，深圳入选国家人工智能创新应用先导区，形成了“双区叠加”的战略优势。2023 年，深圳率先提出打造“全球人工智能先锋城市”目标。2025 年 3 月，深圳印发《深圳市加快打造人工智能先锋城市行动计划（2025-2026 年）》，明确提出到 2026 年“建成具有国际影响力的人工智能先锋城市”，形成“场景应用最开放、算力供给最普惠、产业生态最健全、创新创业最便捷”的产业发展环境，这一战略定位要求深圳在人工智能技术创新、产业集聚、应用示范、制度创新等方面发挥先行示范作用。

2026 年 2 月 9 日，深圳市七届人大七次会议开幕，深圳市市长覃伟中作政府工作报告，对深圳人工智能发展作出全面部署。报告明确提出，2026 年深圳人工智能产业集群增加值增长 10%以上，高标准打造光明实验室、福田实验室、领先边端智能开放研究院等创新平台，高水平建设深圳河套学院；加强全栈自主可控人工智能软硬件生态建设，集中力量攻坚算法理论、模型架构、智算芯片、基础软件、智能机器人等核心技术；建好用好鹏城云脑III等国家人工智能重大算力基础设施、国家超算深圳中心二期；深化拓展“人工智能+”，推动人工智能全域全时全行业高水平应用。

围绕上述战略目标，深圳发布多项专项行动计划，形成了多层次、分领域的政策矩阵。在人工智能终端产业方面，深圳印发《深圳市加快推进人工智能终端产业发展行动计划（2025-2026 年）》，明确提出到 2026 年全市人工智能终端产业规模达 8000 亿元以上、力争 1 万亿元，集聚不少于 10 家现象级人工智能终端企业，人工智能终端产品产量突破 1.5 亿台。在具身智能机器人方面，深圳发布《深圳市具身智能机器人技术创新与产业发展行动计划（2025-2027 年）》，明确到 2027 年在关键核心零部件、AI 芯片、多模态感知技术、灵巧操作技术等方面取得突破，新增培育估值过百亿企业 10 家以上，关联产业规模达到 1000 亿元以上。在“人工智能+”先进制造方面，2026 年 2 月印发《深圳市“人工智能+”先进制造业行动计划（2026-2027 年）》，到 2027 年开放百个应用场景，打造百个垂直行业模型及工业智能体，推广百个示范应用，形成“一基地、一中心、一联盟、百场景、多应用”支撑的发展格局。在 OPC 创业生态方面，2026 年 1 月印发《深圳市打造人工智能 OPC 创业生态引领地行动计划（2026-2027 年）》，到 2027 年底建成超 10 家面积均不少于 1 万平方米的 OPC 社区，培育超千家高成长性人工智能创业企业，集聚超万名人工智能创新创业人才。在 AI 服务器产业链方面，2026 年 3 月印发《深圳市加快推进人工智能服务器产业链高质量发展行动计划（2026-2028 年）》，推动 AI 服务器研发设计、核心零部件和整机制造、系统集成、配套服务等产业全链条创新集聚发展，打造全球领先的智算

集群研发制造枢纽。

## （二）核心产业规模快速增长

深圳人工智能产业综合实力稳居全国第一梯队。深圳全力建设人工智能先锋城市，2025 年深圳人工智能核心产业营收约 2200 亿元，规上人工智能企业超 2600 家，人工智能产业规模连续两年实现两位数增长，产业综合实力稳居全国第一梯队。深圳已形成以华为和腾讯等龙头企业引领、高成长性企业支撑、初创企业竞相发展的“雁阵式”梯队，智能原生企业加速聚集，2025 年深圳新注册的 AI 相关企业增长了 32%，其中 15% 是智能原生企业。智能硬件产品的规模化应用能力处于全国领先水平，催生出雷鸟 AI 眼镜、影石全景相机等全球爆款智能硬件产品。

## （三）产业链条以应用为主导

产业链条完整，60%的企业布局在应用层。从产业链来看，深圳形成了相对完整的人工智能产业链条，覆盖基础层、技术层、应用层全环节，其中超 60%的企业布局在应用层，展示了深圳 AI 发展的“应用驱动”特征。**基础层**方面，深圳数据服务、物联网、云计算、算力运营、智能传感器企业数量领先，为人工智能产业发展提供了坚实的基础支撑；智能芯片领域虽企业数量相对较少，但体现了深圳在硬件设计方面的潜力与布局。**技术层**中，计算机视觉和生物特征识别占据主导地位，这与应用层对图像、视频处理的强劲需求直接相关，技术成熟度和应用广度是关键动因；大模型企业加速集聚，深圳已有 49 款大模型通过国家网信办备案，显示深

圳正积极跟进生成式 AI 等前沿方向。**应用层**企业广泛布局于智能安全、智能制造、智能硬件、智能网联汽车、具身智能机器人、低空飞行器等领域，反映了深圳作为全球电子信息产业链最完整、配套最完善的地区，在 AI 终端规模化落地与生态构建方面具有不可替代的示范引领价值。

#### （四）人工智能创业热度回升

深圳人工智能创业生态活跃，企业成立数量呈现“波浪式”发展轨迹。2012-2017 年企业成立数量从 20 家激增至 129 家，年均增速超 45%，这一阶段主要受益于全球人工智能技术的突破、国家层面对新兴产业的鼓励、深圳强大的电子信息产业基础和浓厚的创新创业氛围。2018-2022 年企业成立数量连续五年下滑，反映出行业进入调整期：一方面，早期技术同质化导致市场竞争加剧；另一方面，美国对中国实施半导体出口管制加剧芯片供应压力；叠加疫情冲击，进一步压缩了中小企业生存空间。2022 至 2025 年企业成立数量回升，2025 年新增企业 55 家，同比增长 52.8%，标志着行业进入新阶段，这一阶段技术快速迭代、政策持续加码，同时深圳在算力基础设施、场景应用落地等方面形成独特优势。

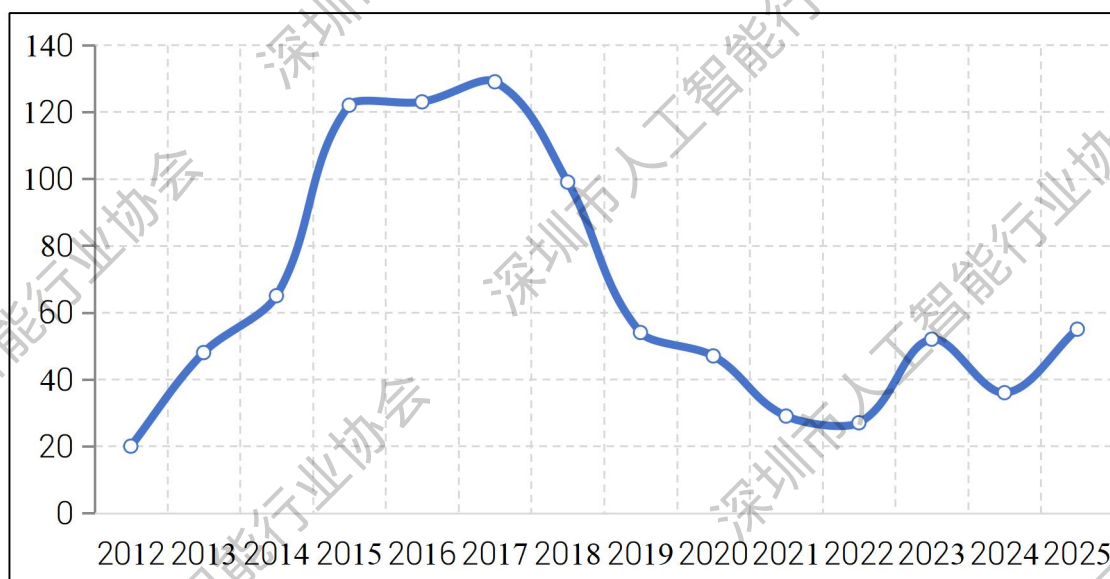


图 7 近年来深圳人工智能企业成立数量变化（单位：家）

数据来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

### （五）不同梯队企业协同并进

深圳企业在人工智能领域展现出强大的创新力和产业领导力，涵盖从底层硬件到行业应用的全产业链布局。这些企业既包括华为、腾讯、大疆创新、中兴通讯、优必选、比亚迪、平安科技、云天励飞、商汤科技等行业龙头，也涵盖了投控数科、鲲云科技、奥比中光、元戎启行、自变量机器人、乐聚机器人、帕西尼、众擎机器人等高成长企业。其中，华为通过“芯-端-云-智”一体化体系构建了完整的 AI 产业链，腾讯聚焦算力、算法与平台的深度融合，中兴通讯则构建了“端-管-云-智”全栈 AI 技术体系。优必选致力于人形机器人全产业链的发展，比亚迪推动智能化与电动化的深度融合，而大疆创新则在无人机领域引领 AI 应用。平安科技则依托全面的 AI 技术矩阵，推动金融、医疗与智慧城市的数字化转型。云天励飞通过边缘 AI 与大模型应用，助力行业智能化进程。

这些企业不仅在国内市场占据重要地位，还在全球范围内树立了技术创新的标杆。

表 1 深圳市人工智能代表性企业列举

| 企业名称 | 核心业务与技术布局                                               | 代表产品与应用                                                                        |
|------|---------------------------------------------------------|--------------------------------------------------------------------------------|
| 华为   | 构建了“芯-端-云-智”一体化人工智能产业链，涵盖芯片、服务器、AI 框架、平台与大模型            | 昇腾 910/950 系列芯片、昇思 MindSpore 框架、盘古大模型，应用于智慧城市、智能制造、绿色能源、数字健康等领域                |
| 腾讯   | 聚焦算力、算法、平台与应用的全链条布局，推出混元大模型，致力于推动 AI 与各行业的深度融合          | 混元 Hy3 Preview 大模型、AI 原生助手“元宝”、WorkBuddy 智能体，应用于微信生态、企业知识引擎、AI 客服、金融风控、桌面办公等领域 |
| 中兴通讯 | 构建了“端-管-云-智”全栈 AI 技术体系，专注于 5G、云计算与 AI 的融合发展             | 昆仑服务器、大模型推理平台、智慧城市“数字星云”平台，应用于智能制造、智慧交通、智慧医疗等行业                                |
| 优必选  | 提供全栈式人形机器人解决方案，专注于人形机器人本体研发、核心零部件制造、AI 算法平台、操作系统及场景应用集成 | Walker 系列人形机器人、AI 教育机器人、物流搬运机器人，应用于工业制造、教育、康养等多个场景                             |
| 比亚迪  | 推动汽车智能化、电动化深度融合，通过璇玑 AI 大模型实现整车智能化                      | 璇玑 AI 大模型、智能驾驶平台，应用于智能电动汽车、智能制造、自动驾驶等领域                                        |
| 大疆创新 | 引领无人机领域的 AI 技术应用，专注于计算机视觉与机器学习，延伸至智能驾驶与 AI 教育等领域        | Phantom 4 无人机、D80 智能驾驶系统、AI 教育套件、ROMO 2 系列扫地机器人，应用于农业植保、智能驾驶、AI 教育等领域          |

| 企业名称 | 核心业务与技术布局                                               | 代表产品与应用                                                |
|------|---------------------------------------------------------|--------------------------------------------------------|
| 平安科技 | 通过预测 AI、认知 AI 与决策 AI 技术，提供金融、医疗、智慧城市等行业的全栈式 AI 解决方案     | 平安脑智能引擎、AI 风控决策模型、智慧城市平台，应用于金融科技、医疗健康、智慧城市等多个领域        |
| 云天励飞 | 聚焦“算法芯片化”与“端云协同”，提供边缘 AI 解决方案，推动 AI 大模型在各行业的应用          | DeepEdge10 芯片、云天天书多模态大模型、AI 智能眼镜，应用于公共安全、智慧城市、AI 教育等领域 |
| 投控数科 | 构建了数据技术、人工智能、企业数字化转型、信创技术四大服务体系，以 AI 赋能基层治理与行业数字化转型成效显著 | 基于 IoT+AI+CIM 的视频指挥调度系统，应用于防汛水涝监测、违停治理等城市治理场景          |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

## （六）算力供给规模持续扩大

深圳算力规模持续扩大，算力资源配置持续优化。当前，深圳已建和在建智能算力超 70EFLOPS。深圳通过“训力券”政策、智算中心集群建设等举措，在全国率先构建起覆盖算力供给、技术研发和场景应用的完整生态体系。随着鹏城云脑 II、深圳开放智算中心等智算中心的建成启用，标志着深圳算力建设取得重大突破性进展，全市智能算力集群进一步扩大，智能算力规模大幅提升。当前，深圳全市已建成/在建通用算力约 3.7EFLOPS、智能算力约 70EFLOPS、超算算力约 2.01EFLOPS。按规划，深圳将继续建设多个 10E 级智能算力集群，推动现有智能算力中心倍增扩容，并着力建设一批国产智能算力中心，打造南山、宝安、龙岗、龙华、前海、河套以及光明科学城等算力高地。预计 2026 年，全市实时可用智能算力将超 80E FLOPS，整体规模在国内主要城市中位居前列，展现出强大的发展势能，超智协同、异构融合、训推一体、普惠泛在的可持续算力供给体系基本形成。

表 2 深圳市智能算力基础设施一览表

| 序号 | 分类      | 算力基础设施                | 运营主体  |
|----|---------|-----------------------|-------|
| 1  | 高校及科研机构 | “鹏城云脑I”               | 鹏城实验室 |
| 2  |         | “鹏城云脑II”重大科技基础设施      |       |
| 3  |         | “鹏城云脑III”网络智能重大科技基础设施 |       |

| 序号 | 分类 | 算力基础设施                  | 运营主体                       |
|----|----|-------------------------|----------------------------|
| 4  |    | IDEA数字经济与金融超级计算集群       | 粤港澳大湾区数字经济研究院              |
| 5  |    | 光明实验室公共AI算力平台           | 人工智能与数字经济广东省实验室<br>(光明实验室) |
| 6  |    | 深圳大学智算中心                | 深圳大学                       |
| 7  |    | 国家工程实验室产业大数据服务中心GPU算力平台 | 深圳信息职业技术学院                 |
| 8  |    | 南方科技大学科学与工程计算中心         | 南方科技大学                     |
| 9  |    | 北京大学深圳研究生院AI4S高性能算力平台   | 北京大学深圳研究生院                 |
| 10 |    | 深圳湾实验室高性能计算与信息平台        | 深圳湾实验室                     |
| 11 | 企业 | 深圳百旺信云数据中心<br>(二期)      | 深圳易信科技股份有限公司               |
| 12 |    | 深圳百旺信云数据中心<br>(三期)      | 深圳易信科技股份有限公司               |
| 13 |    | 中国联通深圳同乐智算中心2号楼<br>(一期) | 中国联合网络通信有限公司深圳市分公司         |

| 序号 | 分类 | 算力基础设施                | 运营主体                         |
|----|----|-----------------------|------------------------------|
| 14 |    | 中国联通深圳同乐智算中心1号楼       |                              |
| 15 |    | 中国联通广东云数据中心<br>深圳坪山基地 |                              |
| 16 |    | 深圳开放智算中心              | 深圳市智运互联大数<br>据有限公司           |
| 17 |    | 中国移动-商汤科技人工智能计算中<br>心 | 深智城集团、中国移动<br>通信集团广东有限公<br>司 |
| 18 |    | 龙华区新型工业智算中心           | 中国移动通信集团广<br>东有限公司深圳分公<br>司  |
| 19 |    | 中国移动深圳宝观机楼一期智算中<br>心  | 中国移动通信集团广<br>东有限公司           |
| 20 |    | 宝城机楼                  | 中国移动通信集团广<br>东有限公司           |
| 21 |    | 光明数据中心                | 深圳耀德数据服务有<br>限公司             |
| 22 |    | 妈湾资源场站数据中心            | 深圳市智城能源云数<br>据中心有限公司         |
| 23 |    | 亚森工业园1号厂房             | 中国电信股份有限公                    |

| 序号 | 分类          | 算力基础设施                  | 运营主体               |
|----|-------------|-------------------------|--------------------|
| 24 |             | 沙河IDC中心                 | 司深圳分公司             |
| 25 |             | 缔息星河5G边缘算力中心            | 深圳缔息云联科技有<br>限公司   |
| 26 |             | 互盟5G数据交换中心              | 深圳互盟科技发展有<br>限公司   |
| 27 |             | 互盟(深圳)环球数据中心            |                    |
| 28 |             | 互盟(深圳)科兴数据中心            |                    |
| 29 |             | 南湾云产业基地                 |                    |
| 30 |             | 光明生命科学大数据中心             | 光明科学城发展建设<br>有限公司  |
| 31 |             | 云引擎算力中心                 | 深圳市云引擎网络科<br>技有限公司 |
| 32 |             | 前海深港人工智能算力中心            | 前海科技创新集团与<br>商汤科技  |
| 33 | 政府及事<br>业单位 | 龙华区“数字治理”AI支撑平台（二<br>期） | 龙华区政务服务和数<br>据管理局  |
| 34 |             | 深圳城市大数据中心（坂田）           | 深圳市大数据资源管<br>理中心   |
| 35 |             | 龙岗智慧中心机房                | 龙岗区大数据中心           |
| 36 |             | 龙华区“数字治理”AI支撑平台（一       | 龙华区政务服务和数          |

| 序号 | 分类 | 算力基础设施                   | 运营主体                   |
|----|----|--------------------------|------------------------|
|    |    | 期)                       | 据管理局                   |
| 37 |    | 深圳（东部）人工智能产业公共服<br>务平台一期 | 深圳市龙岗区城市建<br>设投资集团有限公司 |
| 38 |    | 中山大学附属第七医院智慧医疗联<br>合创新中心 | 中山大学附属第七医<br>院         |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

### （七）数据产业形成“三大名片”

**数据流通体系领先。**截至 2025 年 12 月 31 日，深圳市政府数据开放平台累计提供开放数据目录 5089 个，其中数据接口 4018 个，可下载文件数 21490 个，累计开放数据总量 28.6 亿多条。深圳数据交易所作为全国标杆，截至 2025 年 4 月底，累计上市交易标的 3218 个，吸引 4755 家参与主体，覆盖全国 34 个省级行政区。

**数据标注产业完善。**深圳依托腾讯、华为、平安等科技巨头，标贝科技、百川数安、点通数据等专业化数据标注企业，以及智影医疗、云天励飞、极视角等垂直领域数据标注企业，形成“综合型企业主导+专业化企业支撑+垂直领域企业深耕”的产业格局，涵盖语音标注、文本标注、图像标注、视频标注等多模态数据标注的类型。

表 3 深圳数据标注领域代表性企业发展情况

| 类型 | 企业简称 | 领域布局 | 数据标注产品/技术 |
|----|------|------|-----------|
|----|------|------|-----------|

| 类型         | 企业简称 | 领域布局                             | 数据标注产品/技术                 |
|------------|------|----------------------------------|---------------------------|
| 综合型数据标注企业  | 华为   | 多领域多业态布局，能实现多模态数据标注（语音、文本、图像、视频） | 华为云大数据标注平台                |
|            | 腾讯   |                                  | ModelArts                 |
|            | 平安   |                                  | 腾讯云数据标注平台<br>TI-DataTruth |
| 专业化数据标注企业  | 标贝科技 | 语音数据                             | 金融领域AI数据标注解决方案            |
|            | 百川数安 | 大模型数据                            | 语音标注平台、语音合成数据集            |
|            | 点通数据 | 多模态数据                            | 大模型数据标注服务                 |
| 垂直领域数据标注企业 | 云天励飞 | 智慧城市、智慧交通等                       | 多模态数据标注                   |
|            | 智影医疗 | 医疗影像                             | 智慧城市、智慧交通等图像、视频等标注        |
|            | 极视角  | 自动驾驶、医疗健康等                       | 医疗影像数据标注                  |
|            |      |                                  | 智能数据标注管理平台，自动驾驶、医疗健康等领域标注 |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

**数据治理体系健全。**立法引领层面，深圳率先出台国内数据领域首部基础性、综合性立法《深圳经济特区数据条例》，

在数据权益界定、公共数据开放、数据要素市场培育等方面作出多项创新性规定，为数据治理提供制度保障。**企业共治层面**，华为云、腾讯云分别推出了华为云数据治理中心 DataArts Studio 和腾讯云 WeData 等一站式数据治理平台，覆盖数据全生命周期的开发与运营，为企业提供全方位的数据治理解决方案；华傲数据、傲天科技、太极数智等专业数据治理企业则专注于政务、企业数据治理，提供从数据采集、标准制定到质量管理的全流程服务，形成了从综合型平台到垂直领域深耕的多层次企业生态。**平台赋能层面**，深圳数据交易所作为深圳核心数据交易平台，围绕数据交易前、中、后全周期，出台管理办法 19 项、规则指引 32 项、技术规范 9 项，保障数据要素市场规范发展。

表 4 深圳数据治理企业主要技术/产品情况

| 企业简称 | 主要数据治理技术/产品            | 技术/产品简介                                                                                            |
|------|------------------------|----------------------------------------------------------------------------------------------------|
| 华为云  | 数据治理中心 DataArts Studio | 数据治理中心(DataArts Studio)是华为云数据全生命周期一站式开发运营平台,提供数据集成、数据开发、数据治理、数据服务、数据可视化等功能,支持大数据存储、大数据计算分析引擎等数据底座。 |
| 腾讯云  | 腾讯云-WeData-一站式数据开发治理平台 | WeData 是腾讯云提供的一站式数据开发治理平台，融合数据集成、数据开发、数据资产管理等全链路能力，帮助企业实现数据价值最大化。                                  |
| 华傲数据 | 华傲数据治理平台               | 华傲政务数据融合治理方案包括一整套覆盖完整数据生命周期的数据智能产品族，提供数据汇聚、数据治理、数                                                  |

| 企业简称 | 主要数据治理技术/产品     | 技术/产品简介                                                                                                                 |
|------|-----------------|-------------------------------------------------------------------------------------------------------------------------|
|      |                 | 据融合、数据服务支撑等各项能力。                                                                                                        |
| 傲天科技 | 智制数据治理产品 (Z-DG) | 智制数据治理产品(Z-DG)以用户为中心，以元数据为基础，贯穿数据规划、数据采集、数据标准、数据质量、主数据管理、安全保障，数据运营管理、数据发布共享等环节，从数据源到大数据中心，经过数据治理之后，再到为数据应用提供数据支撑的全过程管理。 |
| 太极数智 | 数据质量治理平台        | 太极数智数据质量治理平台实现对数据从采集到共享的全过程质量治理。包括对元数据管理、数据质量管理、数据服务管理和系统管理等。                                                           |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

## (八) 人才培育体系逐渐完备

深圳地区拥有 17 所高等教育学校，形成较为完整的人工智能人才培养体系，呈现多点开花、协同发展的良好态势。目前，深圳大学、南方科技大学、深圳技术大学、深圳职业技术大学等高校均已设立人工智能专业，深圳大学、香港中文大学（深圳）、深圳职业技术大学、深圳信息职业技术学院、广东新安职业技术学院成立了人工智能学院，初步构建起从专科、本科、硕士到博士的完整人才培养体系，2025 年人工智能专业毕业生约为约 1100-1350 人。随着各高校新设专业的逐步扩招，未来几年深圳地区 AI 相关专业毕业生规模将持续增长；同时，多所高校积极探索创新教育模式，如深圳技术大学与百度共建“人工智能菁英班”，深圳理工大学与腾讯云设立腾讯云大模型特色班，深圳城市职业学院建

成商汤人工智能产业学院、优必选具身智能产教评基地等。

表 5 深圳市高等院校人工智能人才培养情况

| 学校          | 人工智能专业设置                                                 | 人工智能学院设置                                                                                | 在校生/招生规模                                                                              |
|-------------|----------------------------------------------------------|-----------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| 深圳大学        | 2024 年获批人工智能本科专业，招生专业为“人工智能（腾班）”与“人工智能（卓越班）”             | 2025 年成立人工智能学院，现有教教职工 80 人，专职教师 50 名（含两院院士 2 名、外籍院士 1 名、IEEE Fellow 2 名、国家杰青等国家级人才 3 名） | 现有学生 223 人（本科生 50 人、硕士生 166 人、博士生 7 人）；预计 2030 年达 2100 人（本科 1200 人、硕士 750 人、博士 150 人） |
| 南方科技大学      | 2025 年获批人工智能本科专业，首批选拔 25 人“人工智能卓越创新班”                    | 拥有图灵奖获得者 1 人、菲尔兹奖获得者 1 人、人工智能领域院士 18 人                                                  | 逐步扩大招生规模                                                                              |
| 深圳技术大学      | 2024 年获批人工智能本科专业，设人工智能基础模型与算法、人工智能技术与应用、大模型与提示工程技术应用三个方向 | 依托大数据与互联网学院，2024 年 3 月成立人工智能学科，现有专任教师 23 人（教授 7 人、副教授/高工 7 人、助理教授/讲师 9 人）               | 现有学生 125 人（本科生 89 人、硕士生 36 人）；预计 2030 年达 550 人                                        |
| 香港中文大学（深圳）  | 2023 年增设人工智能与机器人硕士研究生项目；2025 年成立人工智能学院                   | 现有教师 10 人（含两院院士 1 名、外籍院士 1 名、IEEE Fellow 1 名）                                           | 2025 年招生约 240 人（本科 200 人、硕士 10 人、博士 30 人）                                             |
| 哈尔滨工业大学（深圳） | 2024 年 9 月获批智能科学与技术一级学科博士学位授权点；2024 年 10 月成立智能科学与工程学院    | 现有专职教师 25 人（正高级职称 20 人、副高级职称 4 人）                                                       | 本科预计 2026 年招生，年招收本科 100-130 人                                                         |

| 学校              | 人工智能专业设置                              | 人工智能学院设置                                                                     | 在校生/招生规模                                                                                 |
|-----------------|---------------------------------------|------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| 中山大学深圳校区        | 2017 年设立智能科学与技术本科专业；2023 年获批智慧交通本科专业  | 智能工程学院现有教职工 147 人、博士后 16 人（国家级人才 4 人、省级人才 13 人）；网络空间安全学院现有教职工 50 余人、博士后 11 人 | 智能工程学院在校生 1377 人（本科 773 人、硕士 417 人、博士 184 人）；网络空间安全学院在校生 527 人（本科 413 人、硕士 88 人、博士 26 人） |
| 深圳理工大学          | 与腾讯云达成战略合作，设立腾讯云大模型特色班                | —                                                                            | —                                                                                        |
| 清华大学深圳国际研究生院    | 2019 年设立人工智能工程硕士项目                    | 现有 AI 学科方向相关教师 51 人（含两院院士 1 名、外籍院士 1 名、IEEE Fellow 2 名）                      | 2020 年以来共招收学生 546 人，已毕业 274 人                                                            |
| 北京大学深圳研究生院      | 2022 年开设智能科学（AI4S）导论课程；2025 年成立科学智能学院 | 现有教师 13 人（含中科院院士 2 名、IEEE Fellow 1 名、国家杰青等高层次人才 4 人）                         | 2025 年招收本科生约 40 人（从北大本部大三/大四学生中选拔）                                                       |
| 电子科技大学（深圳）高等研究院 | 2020 年设立软件工程专业，面向数字信息处理、智能计算等领域培养人才   | —                                                                            | —                                                                                        |
| 深圳职             | 职业本科：人工智能工程技术、软件工程技术、                 | 2019 年成立人工智能学院，2025 年成立人工智能                                                  | 专科在校 1427 人；本科在校 142 人；研究生 26 人；                                                         |

| 学校             | 人工智能专业设置                                                                                   | 人工智能学院设置                                                                                                   | 在校生/招生规模           |
|----------------|--------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|--------------------|
| 业技术<br>大学      | 大数据工程技术、计算机应用工程、信息安全与管理（5个）<br>专科：人工智能技术应用、软件技术、大数据技术、计算机应用技术、云计算技术应用、虚拟现实技术应用、区块链技术应用（7个） | 本科学院。现有专任教师114人（博士78人，占比68.4%），博士生导师3人，硕士生导师31人                                                            | 预计2030年本科在校生达3200人 |
| 深圳信息职业<br>技术大学 | 人工智能学院开设大数据技术、人工智能技术应用、云计算技术应用（专科）                                                         | 2024年8月组建人工智能学院，现有教职员工42名，专任教师博士(后)化率60%                                                                   | 专科在校834人，累计毕业900余人 |
| 广东新安职业<br>技术学院 | 2024年与普天和平科技合作创立“人工智能产业学院”                                                                 | —                                                                                                          | 专科在校100人           |
| 深圳城市职业<br>学院   | 信息与通信学院开设人工智能技术应用专业                                                                        | 建成华为现代产业学院、亚马逊云创技术学院、腾讯“校中企”人才培养基地、商汤人工智能产业学院、优必选具身智能产教评基地、大华物联网人才培养基地等。学院现有专任教师49名，副高及以上职称20名，博士10名，硕士30名 | —                  |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

### (九) 人才需求侧爆发式增长

深圳人工智能领域人才需求呈现爆发式增长态势。2024年以来深圳人工智能领域人才需求呈现爆发式增长态势，核心岗位招聘量同比增幅普遍超过 40%。一些前沿赛道的增速更为显著，其中大模型相关岗位增长 474.4%，智能芯片相关岗位增长 192.2%，虚拟/增强现实相关岗位增长 136.1%。2025 年 10 月深圳“AI 汇英才”校园招聘会中，超 330 家人单位释放超 1.8 万个 AI 岗位，其中半数以上岗位年薪超 30 万元，反映了深圳企业对 AI 人才的旺盛需求。

根据职友集基于近一年的 3.87 万条样本数据统计，**区域分布方面**，深圳人工智能领域人才需求方面主要集中在南山区，占比 31.6%，第二梯队为宝安区(16.6%)、龙华区(10.3%)、龙岗区(10.3%)、福田区(8.5%)。南山区高校、科研单位、互联网企业、AI 企业高度集中，是研发与创新人才聚集地，吸纳大量研究与高端工程岗；福田区金融科技企业集中，AI 在金融风控、智能客服、合规风控方向岗位较多；龙岗区、坪山区、宝安区依托工业制造场景应用需求，吸纳大量智能制造、智能硬件、嵌入式工程师、系统应用开发工程师、行业解决方案专家岗位。**薪资方面**，近一年深圳人工智能相关岗位平均月薪为 1.92 万元，其中月薪万元以下的岗位仅占 25%，1 万-2 万的岗位占比达到 30.4%，2 万-3 万、3 万-5 万的岗位占比均在 20%以上。**学历方面**，近一年深圳人工智能相关岗位学历要求中，本科学历占比 55.2%，硕士学历占

比 4.7%，博士学历占比 0.5%，大专学历占比约 25%，不限学历的岗位仅占 10.3%。

## 第三章 产业全景：链条解析与生态构建

### 一、产业链全景

人工智能产业链通常划分为基础层、技术层、应用层三级架构，基础层提供算力、算法和数据“三驾马车”，算力决定人工智能性能上限、算法优化资源调度、数据驱动模型进化，三者共同构成人工智能发展的基础设施；技术层基于基础层资源，通过算法模型开发感知、认知与决策能力，将原始数据转化为智能模型，链接基础资源与场景应用；应用层则实现技术与各行业的融合，形成可交付的软硬件产品，实现商业化落地。

全球人工智能产业链



图 8 全球人工智能产业链全景图

资料来源: 深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

## 二、基础层

基础层是人工智能发展的底层支撑部分，为下层技术和应用提供必要的硬件、数据和开发工具。它涵盖了硬件（智能芯片、智能传感器、网络设备、存储设备、信息收集设备等）、软件（操作系统、开发测试分析工具、AI 算法框架等）、算力、数据等细分领域。数据是人工智能的核心驱动力，数据服务为人工智能技术的研发与应用提供了基础资源，包括数据的收集、存储、清洗和标注等关键环节；智能芯片是人工智能计算的核心硬件组件，决定了人工智能模型的运行效率和能耗水平；算力为人工智能提供了强大的计算能力和存储资源，是人工智能技术得以规模化应用的保障；智能传感器是人工智能系统感知外部环境的关键组件，用于采集物理世界的的数据，如图像、声音、温度、压力传感器等。

表 6 人工智能产业链基础层代表企业

| 企业简称 | 成立日期       | 融资阶段 | 细分领域           | 所在城市 |
|------|------------|------|----------------|------|
| 华为   | 2004.10.18 | —    | 智能芯片、数据、<br>算力 | 深圳市  |
| 云天励飞 | 2014.08.27 | 上市公司 | 智能芯片           | 深圳市  |
| 地平线  | 2015.07.14 | 上市公司 | 智能芯片           | 北京市  |
| 寒武纪  | 2016.03.15 | 上市公司 | 智能芯片           | 北京市  |
| 四维图新 | 2002.12.03 | 上市公司 | 智能芯片           | 北京市  |
| 中芯国际 | 2000.12.21 | 上市公司 | 智能芯片           | 上海市  |

| 企业简称  | 成立日期       | 融资阶段 | 细分领域  | 所在城市 |
|-------|------------|------|-------|------|
| 镭神智能  | 2013.01.08 | 上市公司 | 智能传感器 | 深圳市  |
| 奥比中光  | 2013.01.18 | 上市公司 | 智能传感器 | 深圳市  |
| 万讯自控  | 1994.06.06 | 上市公司 | 智能传感器 | 深圳市  |
| 晶方半导体 | 2005.06.10 | 上市公司 | 智能传感器 | 苏州市  |
| 腾讯科技  | 2000.02.24 | 上市公司 | 数据    | 深圳市  |
| 平安科技  | 2008.05.30 | —    | 数据    | 深圳市  |
| 阿里巴巴  | 2007.03.26 | 上市公司 | 数据    | 杭州市  |
| 百度云   | 2012.01.05 | —    | 数据    | 北京市  |
| 无限迭代  | 2025.06.27 | —    | 数据    | 北京市  |
| 投控数科  | 2010.01.12 | —    | 数据    | 深圳市  |
| 阿里云   | 2008.04.08 | 战略融资 | 算力    | 杭州市  |
| 浪潮云   | 2015.03.09 | 上市辅导 | 算力    | 济南市  |
| 中科曙光  | 2006.03.07 | 上市公司 | 算力    | 北京市  |
| 润泽科技  | 2007.06.27 | 上市公司 | 算力    | 廊坊市  |
| 深智城   | 2018.12.18 | —    | 算力    | 深圳市  |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

### 三、技术层

技术层是人工智能产业链的核心环节，基于基础层的支撑，围绕算法模型研究展开。它包括机器学习、知识图谱、

大模型、智能体、自然语言处理、智能语音、计算机视觉和生物特征识别等细分领域。大模型是近年来人工智能技术的重大突破，通过大规模数据训练，具备更强的语言理解和生成能力；智能体是指能够感知环境、自主决策并采取行动以实现特定目标的计算系统，OpenClaw 的发布标志着智能体从概念走向实用，成为继大模型之后又一重要技术里程碑；智能语音技术包括语音识别、语音合成和语音交互，是人机交互的重要方式；计算机视觉使机器具备“看”的能力，包括图像识别、目标检测和视频分析等技术。

表 7 人工智能产业链技术层代表企业

| 企业简称 | 成立日期       | 融资阶段    | 细分领域         | 所在城市 |
|------|------------|---------|--------------|------|
| 腾讯   | 1998.11.11 | 上市公司    | 大模型/智能体/机器学习 | 深圳市  |
| 字节跳动 | 2012.03.09 | 战略投资    | 大模型/智能体      | 北京市  |
| 阿里巴巴 | 1999.06.28 | 上市公司    | 大模型/智能体      | 杭州市  |
| 百度   | 2000.01.01 | 上市公司    | 大模型/智能体      | 北京市  |
| 华为   | 1987.09.15 | —       | 大模型/智能体      | 深圳市  |
| 月之暗面 | 2023.04.17 | D轮      | 大模型/智能体      | 北京市  |
| 智谱AI | 2019.06.11 | 上市公司    | 大模型/智能体      | 北京市  |
| 稀宇科技 | 2021.11.03 | 上市公司    | 大模型/智能体      | 上海市  |
| 阶跃星辰 | 2023.04.06 | Pre-IPO | 大模型/智能体      | 上海市  |
| 深度求索 | 2023.07.17 | —       | 大模型          | 杭州市  |

| 企业简称 | 成立日期       | 融资阶段    | 细分领域     | 所在城市 |
|------|------------|---------|----------|------|
| 面壁智能 | 2022.08.12 | E轮      | 大模型      | 北京市  |
| 科大讯飞 | 1999.12.30 | 上市公司    | 大模型/智能语音 | 合肥市  |
| 网易智企 | 2015.10.13 | —       | 智能体      | 杭州市  |
| 声扬科技 | 2016.08.11 | B轮      | 智能语音     | 深圳市  |
| 云知声  | 2015.12.28 | 上市公司    | 智能语音     | 深圳市  |
| 百度   | 2000.01.18 | 上市公司    | 智能语音     | 北京市  |
| 出门问问 | 2017.03.13 | 上市公司    | 智能语音     | 北京市  |
| 捷通华声 | 2000.10.18 | 上市辅导    | 智能语音     | 北京市  |
| 云从科技 | 2015.03.27 | 上市公司    | 智能语音     | 广州市  |
| 奥比中光 | 2013.01.18 | 上市公司    | 计算机视觉    | 深圳市  |
| 商汤科技 | 2014.11.14 | 上市公司    | 计算机视觉    | 北京市  |
| 格灵深瞳 | 2013.08.16 | 上市公司    | 计算机视觉    | 北京市  |
| 旷视科技 | 2011.10.08 | E轮      | 计算机视觉    | 北京市  |
| 斗象科技 | 2014.05.14 | 战略融资    | 机器学习     | 上海市  |
| 思谋科技 | 2019.12.20 | 赴港IPO申报 | 机器学习     | 深圳市  |
| 英飞拓  | 2000.10.18 | 上市公司    | 生物特征识别   | 深圳市  |
| 汇顶科技 | 2002.05.31 | 上市公司    | 生物特征识别   | 深圳市  |
| 商汤科技 | 2014.11.14 | 上市公司    | 生物特征识别   | 深圳市  |

| 企业简称 | 成立日期       | 融资阶段 | 细分领域   | 所在城市 |
|------|------------|------|--------|------|
| 汉王科技 | 1998.09.11 | 上市公司 | 生物特征识别 | 北京市  |
| 依图科技 | 2012.09.29 | D轮   | 生物特征识别 | 上海市  |
| 云从科技 | 2015.03.27 | 上市公司 | 生物特征识别 | 广州市  |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

#### 四、应用层

应用层是人工智能技术的最终落地环节，涵盖了广泛的应用场景。基于基础层和技术层的支持，人工智能在多个行业实现了赋能，创造了巨大的商业价值。AI 终端包括 AI 手机、AI PC 等，成为人工智能技术进入日常生活的重要载体；工业 AI 通过人工智能赋能制造业，实现智能质检、预测维护和生产优化，提升生产效率；智能机器人集成了感知、决策和动作能力，应用于工业制造、家庭服务、医疗辅助等场景；智能医疗通过计算机视觉和深度学习等人工智能技术，实现疾病诊断、手术辅助和药物研发；智能金融应用于风险评估、智能投顾和反欺诈检测，助力金融服务智能化升级。

表 8 人工智能产业链应用层代表企业

| 企业简称  | 成立日期       | 融资阶段     | 细分领域  | 所在城市 |
|-------|------------|----------|-------|------|
| 科沃斯   | 1998.03.11 | 上市公司     | 智能机器人 | 苏州市  |
| 乐聚机器人 | 2016.03.24 | 创业板IPO申报 | 智能机器人 | 深圳市  |
| 优必选   | 2012.03.31 | 上市公司     | 智能机器人 | 深圳市  |

| 企业简称 | 成立日期       | 融资阶段     | 细分领域   | 所在城市 |
|------|------------|----------|--------|------|
| 普渡科技 | 2016.01.13 | D轮       | 智能机器人  | 深圳市  |
| 石头科技 | 2014.07.04 | 上市公司     | 智能机器人  | 北京市  |
| 宇树科技 | 2016.08.26 | 科创板IPO申报 | 智能机器人  | 杭州市  |
| 大疆创新 | 2006.11.06 | C轮       | 智能运载工具 | 深圳市  |
| 小马易行 | 2021.04.08 | —        | 智能运载工具 | 深圳市  |
| 元戎启行 | 2019.02.18 | 战略投资     | 智能运载工具 | 深圳市  |
| 华为   | 1987.09.15 | —        | 智能终端   | 深圳市  |
| 中兴通讯 | 1997.11.11 | 上市公司     | 智能终端   | 深圳市  |
| 易瞳科技 | 2015.02.12 | Pre-A轮   | 智能终端   | 深圳市  |
| 华大基因 | 2010.07.09 | 上市公司     | 智能医疗   | 深圳市  |
| 智影医疗 | 2014.11.26 | A轮       | 智能医疗   | 深圳市  |
| 可孚医疗 | 2009.11.19 | 上市公司     | 智能医疗   | 长沙市  |
| 好未来  | 2008.05.08 | 上市公司     | 智能教育   | 北京市  |
| 点猫科技 | 2015.03.09 | 战略投资     | 智能教育   | 深圳市  |
| 平安科技 | 2008.05.30 | —        | 智能金融   | 深圳市  |
| 微众银行 | 2014.12.16 | A轮       | 智能金融   | 深圳市  |
| 中电金信 | 2011.01.26 | —        | 智能金融   | 深圳市  |
| 蚂蚁金服 | 2014.10.16 | 战略投资     | 智能金融   | 杭州市  |

| 企业简称  | 成立日期       | 融资阶段    | 细分领域 | 所在城市 |
|-------|------------|---------|------|------|
| 来画    | 2017.06.14 | 战略投资    | 数字人  | 深圳市  |
| 幻影未来  | 2020.06.08 | —       | 数字人  | 深圳市  |
| 相芯科技  | 2016.03.10 | A+轮     | 数字人  | 杭州市  |
| 硅基智能  | 2017.08.08 | D轮      | 数字人  | 南京市  |
| 元始智能  | 2023.06.15 | 天使轮     | 智能服务 | 深圳市  |
| 言域科技  | 2023.12.13 | 天使轮     | 智能服务 | 深圳市  |
| 快手    | 2015.03.20 | 已上市     | 智能服务 | 北京   |
| 趣丸网络  | 2014.12.13 | Pre-IPO | 智能服务 | 广州市  |
| 得理    | 2018.01.25 | —       | 智能法律 | 深圳市  |
| 法狗狗   | 2016.05.26 | Pre—A轮  | 智能法律 | 深圳市  |
| 北大法宝  | 1999.12.29 | 战略投资    | 智能法律 | 北京市  |
| 幂律智能  | 2017.09.29 | Pre-B轮  | 智能法律 | 北京市  |
| 云天励飞  | 2014.08.27 | 上市公司    | 智能政务 | 深圳市  |
| 雄帝科技  | 1999.04.03 | 上市公司    | 智能政务 | 深圳市  |
| 蜜度    | 2009.11.27 | B轮      | 智能政务 | 上海市  |
| 平安智慧城 | 2018.09.04 | 战略投资    | 智慧城市 | 深圳市  |
| 图元科技  | 2006.02.09 | 战略投资    | 智慧城市 | 深圳市  |
| 深智城集团 | 2018.12.18 | —       | 智慧城市 | 深圳市  |

| 企业简称 | 成立日期       | 融资阶段 | 细分领域 | 所在城市 |
|------|------------|------|------|------|
| 延华智能 | 2001.12.04 | 上市公司 | 智慧城市 | 上海市  |
| 农博创新 | 2015.11.30 | A轮   | 智慧农业 | 深圳市  |
| 托普云农 | 2008.04.07 | 上市公司 | 智慧农业 | 杭州市  |
| 极飞科技 | 2012.04.13 | C+轮  | 智慧农业 | 广州市  |
| 佳格天地 | 2016.02.17 | A轮   | 智慧农业 | 北京市  |
| 工业富联 | 2015.03.06 | 上市公司 | 智能制造 | 深圳市  |
| 科瑞技术 | 2001.05.23 | 上市公司 | 智能制造 | 深圳市  |
| 中控技术 | 1999.12.07 | 上市公司 | 智能制造 | 杭州市  |
| 创新奇智 | 2008.02.06 | 上市公司 | 智能制造 | 青岛市  |
| 深城交  | 2008.01.14 | 上市公司 | 智能交通 | 深圳市  |
| 金溢科技 | 2004.05.20 | 上市公司 | 智能交通 | 深圳市  |
| 佳都科技 | 2001.09.30 | 上市公司 | 智能交通 | 广州市  |
| 海信网科 | 2000.12.14 | —    | 智能交通 | 青岛市  |
| 量云能源 | 2015.08.13 | —    | 智能能源 | 深圳市  |
| 夸克云智 | 2017.07.19 | —    | 智能能源 | 深圳市  |
| 国能日新 | 2008.02.02 | 上市公司 | 智能能源 | 北京市  |
| 明阳智能 | 2006.06.02 | 战略投资 | 智能能源 | 中山市  |
| 顺丰科技 | 2009.04.07 | —    | 智能物流 | 深圳市  |

| 企业简称 | 成立日期       | 融资阶段 | 细分领域 | 所在城市 |
|------|------------|------|------|------|
| 今天国际 | 2000.10.19 | 上市公司 | 智能物流 | 深圳市  |
| 德马科技 | 2001.04.29 | 上市公司 | 智能物流 | 湖州市  |
| 兰剑智能 | 2001.02.23 | 上市公司 | 智能物流 | 济南市  |
| 深信服  | 2000.12.25 | 上市公司 | 智能安全 | 深圳市  |
| 信义科技 | 2006.12.20 | 战略投资 | 智能安全 | 深圳市  |
| 海康威视 | 2001.11.30 | 上市公司 | 智能安全 | 杭州市  |
| 大华股份 | 2001.03.12 | 上市公司 | 智能安全 | 杭州市  |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

## 第四章 技术跃迁：关键领域的技术迭代

### 一、模型创新与演进

#### （一）模型竞争格局

2025 年以来，AI 模型竞争从单点能力突破进入系统化落地阶段。参数规模和榜单名次仍然重要，但已经不足以单独解释产业格局。决定模型价值的因素，正在扩展为推理能力、多模态理解、长上下文处理、工具调用、代码生成、端侧部署、私有化交付、成本控制和安全治理的综合能力。换言之，AI 模型不再只是一个对话入口，而是逐步成为智能体应用、企业智能化、科学智能、工业智能和端侧 AI 的共同底座。

从供给规模看，中国是全球最活跃的模式发布市场。截至 2025 年 7 月，全球已发布大模型数量达到 3755 个，其中中国发布大模型 1509 个，占比约 40.2%。若以“重要模型”作为统计口径，全球模型供给格局呈现更高集中度：Epoch AI 根据是否为最新技术、是否具有历史意义或引用率等标准，定义了一个囊括 900 余个重要 AI 模型的数据库，2025 年产业界贡献超过 90% 的重要 AI 模型；按国家/地区看，美国生产 50 个重要模型，中国生产 31 个。这说明前沿模型竞争高度集中在美国、中国及少数具备算力、数据、资本和产品入口的机构。

从备案数据看，国内以北京、上海、广东强势领跑。国家网信办 2026 年 5 月公告显示，截至 2026 年 4 月 30 日，

全国累计 868 款生成式人工智能服务完成备案，530 款应用或功能完成登记；相比 2025 年 8 月 538 款服务备案、263 款应用或功能登记，服务化和应用化供给仍在快速扩张。北京、上海仍是大模型备案高地，合计占比接近 50%，依托雄厚的科研资源和完善的产业生态，成为我国大模型技术发展的主引擎；广东更偏场景导向，行业大模型占比超 50%，广泛渗透到政务、教育、交通、工业、农业、物流、气象等多个垂直领域；浙江依托杭州强大的数字经济，更偏云服务、开源生态和模型分发。

表 9 全国重点区域生成式 AI 服务备案数据对比

| 区域 | 数据                                          | 解读                                           |
|----|---------------------------------------------|----------------------------------------------|
| 全国 | 截至2026年4月30日：868款生成式AI服务备案，530款应用或功能登记      | 供给从模型发布转向备案服务、应用功能和行业产品；备案数不等于真实活跃量和收入       |
| 北京 | 截至2026年6月2日：241款生成式AI服务备案                   | 通用基座、科研机构、平台型公司和政策监管资源集中，仍是模型供给高地            |
| 上海 | 截至2026年6月1日：169款生成式AI服务备案                   | 多模态、企业服务、金融科技和长三角产业协同突出，适合系统工程与企业级交付         |
| 广东 | 截至2026年4月21日：广东142款生成式AI服务备案；深圳54款生成式AI服务备案 | 行业大模型占比超50%，广泛渗透到政务、教育、交通、工业、农业、物流、气象等多个垂直领域 |
| 浙江 | 截至2026年4月13日：79款生成式AI服务备案                   | 互联网平台、云服务、开发者社区和商业化分发能力强，高性价比与开放权重生态扩散是主要看点  |

资料来源：国家网信办，深圳市人工智能行业协会、深圳市投控数字科技有

限公司整理

从 AI 支出来看，资金并非主要沉淀在模型层本身。Gartner 预测，2026 年全球 AI 总支出约 2.59 万亿美元，同比增长约 47%；其中 AI 模型支出约 326 亿美元，只占总支出的约 1.26%。产业资金更大比例流向 AI 基础设施（55.1%）、服务（22.6%）、软件（17.5%），三者合计超过 95%。这意味着 AI 落地的主要成本并不只在“买一个模型”，而在于建设算力底座、接入企业系统、改造软件流程、保障数据安全，并把模型能力嵌入真实业务。模型仍是 AI 产业的关键，但真正决定产业价值释放的，是模型能否通过云平台、软件产品和行业应用形成持续可付费的解决方案。

表 10 2026 年全球 AI 支出构成（预测）

| 类别   | 金额（亿美元） | 占比     | 说明             |
|------|---------|--------|----------------|
| 基础设施 | 14300   | 55.10% | 数据中心/芯片/服务器/网络 |
| 服务   | 5855    | 22.60% | 云服务/咨询/集成/运维   |
| 软件   | 4532    | 17.5%  | 企业软件/AI功能订阅    |
| 安全   | 513     | 2.0%   | 安全/治理/风控/合规    |
| 模型   | 326     | 1.26%  | AI Models模型层收入 |

资料来源：Gartner，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

从模型能力来看，中美模型多次在评测榜单顶部互换位置。不同评测榜单的排序并不完全重合，LMArena 更适合观

察通用对话体验和用户偏好，Artificial Analysis 更适合观察综合能力、价格和速度，SWE-bench、LiveCodeBench 等更偏向代码与可验证任务，说明大模型竞争已经分化为体验、推理、代码、成本、开放权重和产业生态等多个维度。以 LMArena 为例，闭源旗舰榜仍以美国模型为主，开放权重与低成本部署榜中 GLM、Kimi、MiMo 和 DeepSeek 等中国模型进入较高区间，显示出中国模型在高性价比和长上下文方向的竞争力。

表 11 主要大模型评测榜单结果（截至 2026 年 6 月）

| 榜单                     | 当前结果                                                                                                                                            | 说明                                     |
|------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------|
| LMArena Text<br>Arena  | Text Overall 前列仍以 Anthropic、Google、OpenAI、Meta 等美国机构模型为主；开放/中文场景中中国模型影响力更突出                                                                     | 适合观察通用对话体验、人类偏好和模型交互风格                 |
| Artificial<br>Analysis | Claude Opus 4.6、Claude Opus 4.7、Claude Opus 4.8、Muse Spark、GPT-5.5、Gemini 3.1 Pro Preview 位居前列；开放权重前列包括 Kimi K2.6、MiMo-V2.5-Pro、DeepSeek V4 Pro | 适合同同时观察综合能力、价格、速度、上下文窗口和开放权重格局         |
| SWE-bench<br>Verified  | 榜单前列包括 live-SWE-agent/Sonar+Claude 4.5 Opus (79.2%)、TRAE+Doubao-Seed-Code (78.8%)、Gemini 3 Pro Preview (77.4%) 等                                | 适合观察真实代码仓库修复能力，以及模型、工具链和 Agent 框架的组合效果 |
| LiveCodeBench          | 榜单前列包括 o4-mini (High, 87.3%)、o3 (High, 84.7%)、DeepSeek-R1-0528 (84.4%)、                                                                         | 适合观察代码生成、算法题和可验证任务能力                   |

| 榜单 | 当前结果                                                     | 说明 |
|----|----------------------------------------------------------|----|
|    | Gemini-2.5-Pro-06-05 (84.3%)、<br>Qwen3-235B-A22B (80.4%) |    |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

综合来看，AI 模型产业已形成“数量扩张、前沿集中、榜单分化、备案提速”的竞争格局。全球 AI 模型供给数量持续增加，但真正能够定义前沿能力和商业模式的仍是少数机构；中国在数量、备案、开源和应用扩散上进展显著，但在闭源旗舰、全球开发者生态、统一评测影响力和高端算力供给上仍面临约束。

## （二）模型演进趋势

模型正在从“规模扩张”走向“系统优化”。能力提升不再只依赖参数增长，而是由架构效率、训练范式、推理机制、上下文长度、多模态能力、工具调用、部署方式和商业交付共同决定。MoE、稀疏计算、低精度训练和缓存复用降低单位推理成本；后训练和强化学习强化复杂任务求解；长上下文和原生多模态扩大任务入口；工具调用与 Agent 机制则把模型能力接入软件、数据和业务流程。

### 1.架构与成本：从密集扩张转向稀疏激活

在架构层面，MoE、分组查询注意力、稀疏注意力、低精度训练、量化推理和缓存复用成为兼顾能力与成本的主线。DeepSeek-V3、Llama 4 Maverick、Kimi K2 等模型路线都体现出同一方向：总参数继续扩张，但推理时只激活部分专家，并通过 FP8、MLA、量化和缓存压缩降低训练与部署压力。

开源和开放权重模型的竞争，已经从单点追赶转向能力、成本、上下文和执行效率的综合竞争。

这种变化的产业意义在于，模型能力正在从“训练一次、回答一次”变成“按任务动态分配计算”。简单对话可以交给轻量模型，高难推理、代码、科研和复杂业务流程则交给旗舰模型或推理模型。企业部署时也不再只比较单一模型的能力上限，而是同时评估任务复杂度、单次调用成本、延迟、稳定性、私有化能力、合规要求和迁移成本。

## 2.后训练与推理：从人类偏好对齐转向复杂任务能力塑造

后训练正在从早期的“人类偏好对齐”转向“复杂任务能力塑造”。ChatGPT 早期阶段，SFT、RLHF、DPO 等方法主要用于提升指令遵循、回答质量和人类偏好一致性；进入推理、代码、工具调用和 Agent 任务后，单纯让模型“回答得像人”已不足以支撑复杂任务交付。DeepSeek-R1 的意义在于，它展示了可验证奖励强化学习在数学、代码、竞赛题等任务上的扩展潜力：模型可以通过规则奖励、结果验证和多轮采样强化推理策略。

到 2025-2026 年，后训练已成为由偏好优化、可验证奖励、工具使用训练、长程任务评估和安全约束共同组成的工程体系。OpenAI、Anthropic、Google 等闭源模型把推理、代码、Computer Use、知识工作和多模态任务纳入核心能力；DeepSeek、Qwen、Kimi、GLM 等国产模型则在开源推

理、Agent 任务、工具交互、代码工程和低成本部署上持续推进。后训练的目标已经不只是“回答更好”，而是让模型能够拆解目标、规划步骤、调用工具、验证结果、控制成本并遵守安全边界。

### 3.长上下文与多模态：任务入口从文本扩展到连续内容

长上下文正在从卖点变成基础能力。百万级上下文、稀疏注意力、提示缓存和检索增强正在被放入同一个工程问题中处理，目标不是无限堆长，而是在成本可控的前提下处理整本书、完整代码仓库、长视频转录、多份企业文档和跨部门知识库。长上下文的价值不在于一次塞入更多文本，而在于让模型在跨文档、跨系统和长周期任务中维持稳定状态。

多模态方向也从“看图说话”走向原生多模态和连续内容处理。以 Gemini、GPT 系列、Qwen-VL、InternVL、SenseNova 等为例，文本、图像、音频、视频被纳入统一任务入口，使模型能够从理解图片、语音和文档，进一步扩展到生成内容、编辑内容、操作界面和辅助决策。多模态能力的产业价值不只在内容生产，也在教育培训、工业质检、医疗影像、自动驾驶、机器人和数字孪生等场景中打开新入口。

### 4.工具化与 Agent：从生成答案转向完成任务

工具调用和 Agent 机制把模型能力从语言空间接入软件、数据和业务流程。MCP 解决模型与工具、数据库和企业系统的连接问题，A2A 推动不同 Agent 之间发现、委托和协作；编码 Agent、办公 Agent 和数据分析 Agent 则把这类能力推

向真实 workflows。模型不再只生成一段回答，而是在多步骤任务中调用工具、执行动作、记录日志并校验结果。

Agent 的核心竞争力不在语言流畅度，而在任务能否被可靠完成。代码修复、数据分析、网页浏览、表格处理、客服决策、办公自动化和研发辅助等任务，要求模型在计划、执行、观察、修正和交付之间循环推进。未来两三年，企业选择模型时会更关注任务成功率、工具兼容性、错误恢复能力、权限边界、日志审计和单位任务成本。

### 5. 产业交付与端侧部署：从云端 API 走向多层分发

产业交付正在从单一云端 API 调用走向三层分发结构。第一层是闭源或托管旗舰模型，承担高复杂度推理、专业知识工作、复杂代码和企业级 Agent 任务；第二层是 DeepSeek、Qwen、Kimi、Llama、Mistral 等开放或开放权重模型，支撑私有化部署、二次开发、本地推理和行业模型；第三层是经量化、蒸馏等压缩技术和框架优化后的轻量模型，体积更小、算力门槛更低，进入 AI PC、手机、车载、机器人和边缘服务器。

端侧部署的意义不是完整替代云端旗舰，而是承担高频、低延迟、隐私敏感和弱网环境下仍可用的任务。AI 手机、AI PC、企业内网助手、代码补全、文档问答、产线边缘质检、机器人和车载终端，都需要更小、更快、更可控的模型。端云协同将成为主流形态：云端模型负责复杂推理和全局知识，端侧模型负责实时响应、隐私保护和本地执行。

表 12 模型分发形态、适用场景与主要约束

| 分发形态    | 代表路径                                   | 适用场景与价值                                 |
|---------|----------------------------------------|-----------------------------------------|
| 闭源/托管旗舰 | OpenAI 、 Anthropic 、 Google，以及国内云端平台模型 | 复杂推理、长文档、代码工程、在线研究、企业级Agent；能力上限高，工具链完整 |
| 开放/开放权重 | DeepSeek、Qwen、Llama、Kimi、Mistral等路线    | 私有化部署、行业微调、本地知识库、内网Agent；可控性强，适合敏感数据场景  |
| 端侧轻量模型  | 量化、蒸馏、小模型、SLM与NPU适配模型                  | AI PC、手机、车载、机器人、边缘服务器；低延迟、低成本、隐私友好      |
| 混合部署架构  | 云端旗舰+开放权重+端侧模型组合                       | 金融、医疗、工业、政务和企业内网；按任务分层兼顾能力、成本和合规        |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

## 6.VLA 与具身智能：从理解世界走向作用于世界

VLA 模型（视觉—语言—动作模型）代表大模型向物理世界延伸：模型不只识别图像和理解指令，还要把感知结果转化为动作决策。海外案例包括 Google RT-2/RT-X 把网页和机器人动作数据结合用于机器人控制，Figure Helix 用于人形机器人双手协作；国内小鹏、理想等车企将 VLA 思路用于智能驾驶感知、规划和控制，机器人企业也在搬运、巡检、抓取等场景尝试“看懂环境再行动”。

VLA 模型仍处在早期产业化阶段，难点在于真实世界数据、物理规律建模、长程安全验证和低延迟控制。它不会在短期内替代通用大模型竞争主线，但会成为机器人、自动驾

驶、工业操作、家庭服务和空间智能的重要底层技术。对产业而言，VLA 意味着模型从“数字世界的助理”进一步走向“物理世界的执行系统”。

表 13 模型演进主线、典型案例与产业含义

| 主线        | 代表方向                  | 典型案例                                               | 产业含义                        |
|-----------|-----------------------|----------------------------------------------------|-----------------------------|
| 架构效率      | MoE、稀疏激活、低精度、缓存复用     | DeepSeek-V3、Llama 4 Maverick、Kimi K2、Qwen、Mistral等 | 降低单位推理成本，提升私有化和规模化部署能力      |
| 复杂推理      | RLVR、推理时计算、长程任务评估     | DeepSeek-R1、OpenAI 推理模型、Claude扩展思考                 | 从回答质量转向任务成功率和可验证交付          |
| 原生多模态     | 文本、图像、音频、视频、屏幕状态统一处理  | Gemini、GPT 系列、Qwen-VL、InternVL、SenseNova等          | 扩大模型进入教育、工业、医疗、内容生产和办公流程的入口 |
| 工具与 Agent | 工具调用、浏览器操作、工作流编排、权限控制 | MCP、A2A、OpenAI Agents SDK、企业编排 Agent、编码Agent       | 模型从内容生成转向跨系统执行              |
| 端侧部署      | 量化、蒸馏、NPU 适配、端云协同     | MiniCPM 车载端侧模型、AI手机、AI PC、车载、机器人、边缘服务器             | 降低延迟和隐私风险，提升高频场景可用性         |
| VLA       | 视觉—语言—动作统一建模          | Google RT-2/RT-X、Figure Helix、小鹏 VLA、理想MindVLA-o1  | 大模型从理解数字内容走向驱动物理行动          |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

### （三）模型发展瓶颈

大模型发展瓶颈已经从回答能力本身，转向能否稳定、低成本、可治理地完成真实任务。随着模型进入企业系统、

行业流程和物理世界，问题不再只是参数、榜单和演示效果，而是数据、成本、可靠性、安全、透明度、迁移成本和生态控制权的综合约束。

## 1. 高质量数据与验证环境不足

高质量数据正在从“可抓取资源”转为“需授权、可追溯、可审计的生产要素”。2025-2026 年，数据瓶颈的核心不只是公开文本存量下降，而是网页抓取、版权授权、训练数据披露和高质量专业数据获取同步收紧。Cloudflare 在 2025 年推出 AI 爬虫默认权限控制和“Pay Per Crawl”机制，反映网站、出版方和平台正从默认开放抓取转向显式授权、计费 and 审计；欧盟《人工智能法案》也已把通用模型训练数据摘要、版权政策和系统性风险治理纳入义务。《The 2025 Foundation Model Transparency Index》与《AI Index Report》的相关观察同样显示，前沿模型在训练数据来源、算力使用和部署影响披露上仍不充分，透明度与合规压力持续上升。

因此，数据瓶颈更准确地说，是专业数据、授权数据、多模态数据、长程任务轨迹和可验证环境同时稀缺。市场端，Reddit、新闻出版机构、代码问答社区和专业数据库与模型厂商的数据授权合作增多，说明高质量语料正在资产化。对企业而言，合成数据可以使用，但必须嵌入专家反馈、授权数据、数据血缘和可验证任务集的闭环；否则容易出现长尾分布丢失、错误累积和性能退化。

## 2.训练、推理与能源成本约束

推理模型和 Agent 模型通常需要更长上下文、更多工具调用和更多结果校验，单位任务成本显著高于普通对话。IEA 2026 年专题报告将 AI 能耗问题从模型层拉到数据中心、电力系统和能源安全层面；LBNL/DOE 报告则显示，美国数据中心用电已从 2023 年的 176TWh、占全国用电量 4.4%，走向 2028 年 325-580TWh、占 6.7%-12.0% 的预测区间。这说明模型竞争的边际成本不只在 GPU 和 API 价格，也在电力接入、冷却、网络、土地和配套基础设施的可获得性。Meta 路易斯安那 AI 数据中心配套天然气发电项目获批的案例，也表明大型算力投资正在外溢为区域电网扩容和能源项目压力。

中国侧的约束同样清晰，但更多体现为算电协同、绿色算力和集约化智算基础设施。国家能源局公开信息显示，2025 年我国已建成 42 个万卡级智算集群，全国算力中心总用电量约 1700 亿千瓦时，占全社会用电量约 1.6%；预计“十五五”时期全国算力用电量年均新增 1000 亿千瓦时以上，到 2030 年占全社会用电量约 6%。2026 年发布的《促进人工智能与能源双向赋能行动方案》进一步把保障算力设施能源供给、推动绿色低碳转型、促进算力电力高效经济协同列为重点任务。企业端，阿里巴巴未来三年 3800 亿元 AI 和云基础设施投入、中国移动 2025 年算力领域 373 亿元资本开支，也说明推理侧成本已经从模型调用价格，扩展为算力、能源、网络、冷却和资本开支的综合约束。

### 3.长程可靠性与 Agent 失误成本

公开报道、当事人复盘和机构研判指向同一结论：长程 Agent 的风险不在“不会生成代码”，而在接入生产系统后的权限、隔离、确认、回滚和审计机制不足。2025-2026 年，Replit 相关数据库误删、PocketOS 经由 Cursor/Claude 删除生产库及备份、DataTalks.Club 在 Terraform 迁移中误删生产基础设施等事件，共同说明生产环境中的 Agent 事故往往是模型行为、工具权限、云平台接口、备份机制和人工确认缺失叠加的结果，不能简单归因于单一模型能力。Gartner 也提示，成本上升、业务价值不清和风险控制不足，会使相当一部分 Agent 项目被取消或降级。企业级 Agent 落地应把权限分级、沙箱环境、人工确认、破坏性操作拦截、结果验证和日志审计作为默认设计。

### 4.透明度、安全和知识产权压力

模型能力越强，透明度和安全治理压力越大。Stanford CRFM《The 2025 Foundation Model Transparency Index》显示，主要模型开发者平均透明度分数从 2024 年的 58/100 降至 2025 年的 40/100；接受联系并提交透明度报告的公司比例也从 2024 年的 74% 降至 2025 年的 30%。欧盟《人工智能法案》下通用 AI 模型义务已于 2025 年 8 月 2 日开始适用，要求围绕技术文档、训练内容摘要、版权政策和系统性风险模型建立更明确的披露与治理机制。训练数据来源、授权方式、算力使用、风险评估和部署后影响披露不足，正在成为金融、医疗、教育、政务和内容平台落地时的共同约束。

## 5.生态锁定与迁移成本

模型竞争正在从“单模型能力”延伸到生态控制权。企业一旦围绕某一模型建立提示词、知识库、插件、Agent 工作流、权限体系和评测体系，后续迁移成本会快速上升。2025-2026 年的企业部署趋势已经从“押注单一旗舰模型”转向混合策略：高价值任务使用闭源旗舰，量大、敏感或内部任务使用开放权重和私有化模型，并通过统一路由、标准接口和评测体系降低供应商锁定、数据出站费用和重构成本。

## 二、智能体从概念走向实用

过去一年，智能体的形态正从对话工具演进为可自主执行任务的系统。智能体的特点在于：用户给出目标后，系统能自行调用工具、操作软件并完成任务，过程留有记录、可供复核。判断智能体是否成熟，关键在于能否安全接入企业系统、稳定完成任务，并带来可计量的回报。这也是 2025 年至 2026 年行业关注的焦点。

### （一）智能体产品演进

智能体走向规模化部署，市场规模高速增长。麦肯锡 2025 年全球调查显示，约 88% 的组织已在至少一个业务职能中常规使用人工智能；在智能体方面，约 39% 的组织已开始试验、约 23% 的组织已在企业内规模化部署智能体系统。Grand View Research 估算 2025 年全球智能体市场规模约 76.3 亿美元，其中北美约占 39.6%，亚太增速最快；预测 2026 年市场规模将达到 109.1 亿美元，2026-2033 年的年均复合

增速约 49.6%。国内方面，据科智咨询测算，中国企业级智能体市场规模由 2024 年约 86 亿元升至 2025 年约 212 亿元，预计 2026 年约 449 亿元、2029 年约 3320 亿元，2024-2029 年的年均复合增速约 107%。

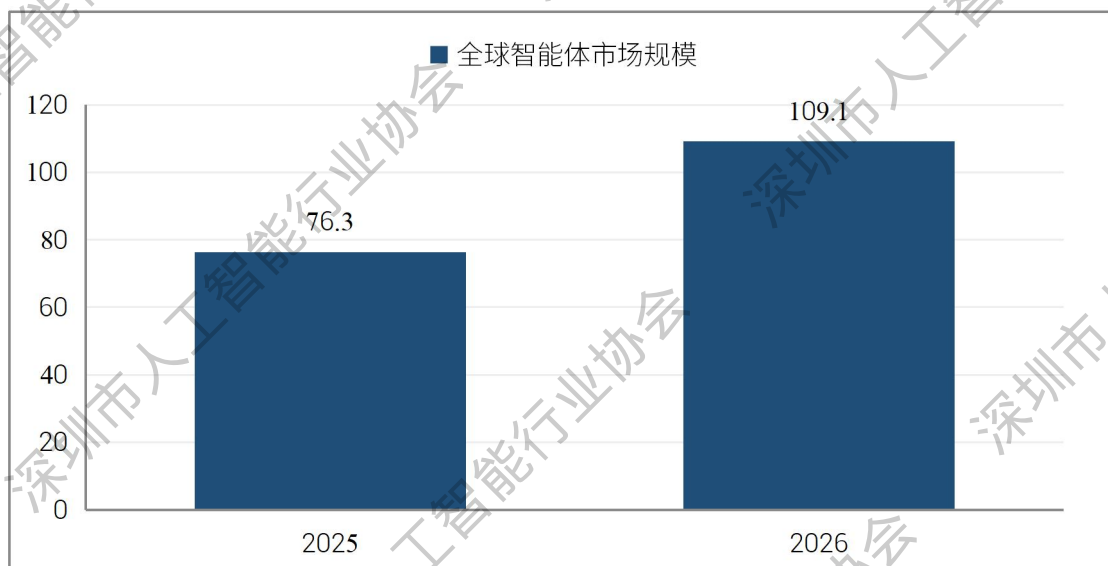


图 9 近年来全球智能体市场规模（单位：亿美元）

数据来源：Grand View Research，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

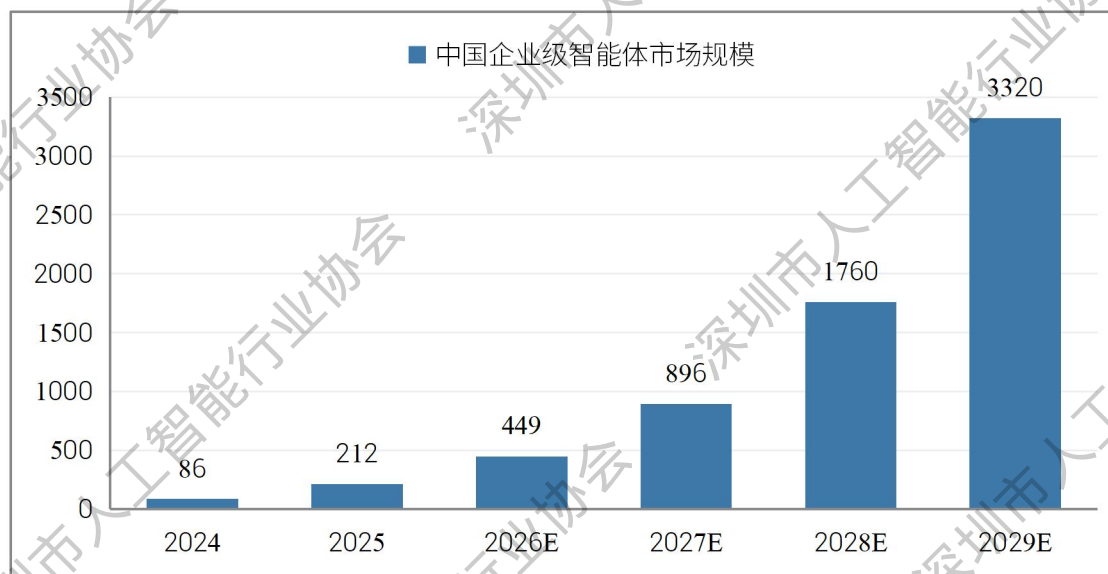


图 10 近年来中国企业级智能体市场规模（单位：亿元）

数据来源：科智咨询，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

**OpenClaw 成为智能体从对话走向执行的重要标志。**  
OpenClaw（国内惯称“龙虾”）由 Peter Steinberger 发起，采用 MIT 协议，支持本地自托管，用户可通过 WhatsApp、Telegram、微信、飞书等应用作为交互入口，接入 Claude、OpenAI、DeepSeek、Kimi、GLM 等多家模型，凭借其自主调用工具、执行复杂任务的能力，为用户提供了可自主部署、跨平台使用的个人智能体。OpenClaw 的核心价值在于填补了智能体从“对话”到“执行”的最后一公里，截至 2026 年 4 月，OpenClaw 的 Github 星标数突破 35 万，创下 Github 历史上最快星标增长速度，引发全球广泛关注，成为此后大量同类产品命名与形态上的源头。

**基于 OpenClaw 开源生态，国内厂商快速跟进。**2026 年上半年，多家中国企业相继推出本土化的“类自主智能体”，命名上普遍沿用 OpenClaw 的“Claw”，如月之暗面的 KimiClaw、腾讯的 QClaw、阿里巴巴的 JVS Claw、百度的红手指 Operator、字节跳动的 ArkClaw、智谱的 AutoClaw 等，主要面向桌面操作与个人助理场景。本土化集中体现在三个方面：以微信、钉钉、飞书等为主要入口；默认接入 DeepSeek、Kimi、GLM 等国产模型；提供一键安装包以降低非技术用户门槛。

**智能体迭代速度进入每日小版本更新的高频节奏。**  
OpenAI Codex 从 2025 年 4 月以命令行工具首次亮相，到

2026 年 2 月推出桌面应用（支持多任务并行）Codex 与新一代长时运行模型 GPT-5.3（可长时间自主运行、并支持随时介入），前后不到一年；Anthropic Claude Code 自 2025 年 2 月研究预览起步，其后以月为单位持续融入 MCP（模型上下文协议）、子智能体、联网搜索和可观测性等新能力，至 2026 年 5 月随 Claude Opus 4.8 发布动态 workflow 预览，距离初始版本仅 15 个月。进入 2026 年，主流智能体产品普遍进入接近每日小版本更新的高频迭代节奏，能力边界持续快速扩展，单次评测结论往往很快被新版本刷新。

**智能体能力的一项观测指标是任务完成时长阈值。** METR 的研究提供了一个观察智能体能力跃迁的量化指标，即智能体能够独立完成的任务的时间跨度。按照这一指标，前沿模型的能力边界已从早期只能完成秒级、分钟级任务，提升到可完成数小时级任务。METR 还指出，2019-2025 年间前沿模型能够独立完成的任务的时间跨度大约每 7 个月翻番；而在 2023 年后，这一速度加快为约每 4 个月翻番。时间跨度的延长意味着模型不只是回答单个问题，而是开始具备跨步骤规划、调用工具、持续调试和完成真实交付的能力。

表 14 全球代表性智能体产品

| 产品          | 推出方       | 起始      | 应用场景                           |
|-------------|-----------|---------|--------------------------------|
| Codex       | OpenAI    | 2025.04 | 面向软件开发场景，适用于代码理解、修改、测试、调试和工程协作 |
| Claude Code | Anthropic | 2025.02 | 面向开发者终端和代码库，适用于大型项目理解、代码解释、重构与 |

| 产品            | 推出方               | 起始      | 应用场景                                |
|---------------|-------------------|---------|-------------------------------------|
|               |                   |         | 调试                                  |
| OpenClaw      | Peter Steinberger | 2025.11 | 面向通用智能体开发与部署，适用于工具调用、任务执行和本地系统自动化   |
| Hermes Agent  | Nous Research     | 2026.02 | 面向复杂任务处理，适用于多步骤规划、跨工具协同和流程自动化       |
| KimiClaw      | 月之暗面              | 2026.02 | 面向知识处理与长文本任务，适用于资料分析、文档整理和信息检索      |
| WorkBuddy     | 腾讯                | 2026.03 | 面向办公协同场景，适用于文案生成、数据处理、信息整理和工作流自动化   |
| JVS Claw      | 阿里云               | 2026.03 | 面向企业级智能体应用，适用于业务流程编排、系统集成和任务自动化     |
| 红手指 Operator  | 百度智能云             | 2026.03 | 面向云手机与移动应用操作场景，适用于移动端任务执行和自动化操作     |
| ArkClaw       | 字节跳动-火山引擎         | 2026.03 | 面向多模型智能体应用，适用于工具调用、任务规划和跨平台执行       |
| AutoClaw      | 智谱                | 2026.03 | 面向自动化任务执行，适用于网页操作、数据采集、流程处理和重复性工作   |
| Minimax Agent | 稀宇科技              | 2025.06 | 面向 AI 原生工作台，适用于网页生成、图表制作、内容创作和多模态办公 |

| 产品     | 推出方  | 起始      | 应用场景                                      |
|--------|------|---------|-------------------------------------------|
| Manus  | 蝴蝶效应 | 2025.03 | 面向通用任务代理，适用于市场调研、数据分析、简历筛选、旅行规划和文档处理。     |
| miclaw | 小米   | 2026.03 | 面向手机、汽车与智能家居生态，适用于语音交互、设备控制、生活助理和多模态场景理解。 |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

## （二）智能体技术架构

智能体本质上是“大模型+运行框架”构成的复合系统。大模型负责理解、推理与生成，运行框架负责将模型能力接入工具、流程、记忆、权限与验证环节。当前的技术架构大致形成七层能力栈：上下文接入层读取企业知识库、业务系统、文档、邮件、日程与代码仓库；任务规划层将自然语言目标转为结构化任务单，明确目标、约束、输入输出、工具列表与人工确认点；工具协议层连接 API、浏览器、代码执行与企业软件；执行环境层提供云沙箱、虚拟桌面、测试环境与版本控制；长期记忆与技能层留存任务状态、用户偏好与可复用技能；验证与可观测层记录调用过程、校验输出并支持失败回滚；权限与治理层限定访问数据与可执行动作并设置人工确认节点。上述任一环节缺失，系统都将由可控的执行体退化为不可控的自动化脚本。

工具协议层逐步成为连接智能体与外部世界的基础设施。不同协议分别承担不同职能：MCP（模型上下文协议）

实现模型连接工具与数据源的标准化，已被主流编程及办公产品采用，活跃公共服务器超过 1 万个；A2A（智能体间通信协议）规范跨智能体的发现、通信与协作，支持组织已逾 150 家；AP2（智能体支付协议）面向智能体在用户授权下的支付执行，参与组织逾 60 家；UCP（通用商务协议）由 Shopify 与 Google 共同开发，面向商品发现与交易协同，解决“怎么在商业场景中落地”的问题。但协议解决的是连接问题，系统能否稳定交付，仍取决于其上的记忆、验证与权限机制。

**智能体在部署形态上存在两条路径。**其一为运行于个人设备的个人智能体，侧重处理邮箱、日程、浏览器与本地文件，侧重长期记忆与可扩展技能；其二为企业平台化，侧重将知识库、权限体系、审计链路与流程节点嵌入既有系统。受数据安全、个人信息保护、行业监管与政企采购要求的约束，中国市场更倾向于先把知识库、权限体系、流程节点与审计链路嵌入办公、客服、制造、政务等既有系统，再在可控场景中逐步扩大执行范围；代表性企业级智能体包括飞书 aily、扣子、阿里云百炼等，多沿企业知识库与流程嵌入的路径展开。

### （三）智能体落地场景

判断一项任务能否交给智能体自主执行，主要看三个条件：任务边界是否清晰、所需工具是否可达、执行结果是否可验证。三者都满足，才有可能形成可量化的回报并安全进入生产环境；缺了任何一项，即便演示时效果亮眼，实际运

行中也难以稳定。从当前落地情况来看，智能体主要落地于以下方向：代码开发、界面操作（即通过模拟人机交互操作现有软件界面）、客户服务、企业办公、平台生态嵌入。

**编程与代码开发最先具备智能体自主执行的条件**，因其具备测试、版本控制、代码审查与回滚等完整的验证链路：智能体完成代码修改后，可以自动运行测试、提交审查请求、交由人工复核，并在必要时回滚变更。目前，Codex 与 Claude Code 均已覆盖从仓库操作、终端执行到测试运行和审查提交的完整流程，远不止于代码补全或解释说明；其效果以测试通过率、PR 采纳率、缺陷修复率与回滚率衡量，并依赖仓库质量、测试覆盖与权限配置。

**客服与办公场景是代码开发之外落地最快的领域**。从国内外实践来看，美国财税服务公司 1-800Accountant 在报税高峰期将约七成对话咨询交由智能体自主处理，法律案件管理公司 Engine 应用智能体后的案件处理时长下降约 15%；海底捞“大明白”日均服务逾 6000 人次，解答率约 87%；美宜佳智能客服覆盖约 3.6 万家门店，响应时长下降约 58%。与此同时，智能体的应用形态也正在从软件向软硬一体延伸——网易智企旗下云信业务将实时对话智能体嵌入硬件终端，实现传统硬件的智能化改造，在企业会议中，实时对话智能体可以实时记录会议要点、提供决策支持和后续行动项跟踪。在编程场景，使用者借助 AI 硬件与客户端、APP 连接，便可以通过对终端口述需求，生成开发代码、确认开发进程、获取开发结果，从而让 AI 的思考状态具象化。

界面操作技术使智能体得以通过模拟人机交互操作现有软件界面，但目前仍处早期阶段。根据 OpenAI 公布的测试数据，其操作型智能体在 WebVoyager 基准（网页导航任务）上成功率约 87%，在 WebArena 基准（更复杂的网页交互）上约 58.1%，而在 OSWorld 基准（桌面环境操作）上仅约 38.1%。可以看到，任务越接近真实桌面环境，成功率越低。因此，现阶段该技术更适合用于后台填表、订单查询、网页检索、费用报销、供应商门户操作等重复性高、容错空间大的长尾任务。同时，须配合沙箱隔离、域名白名单限制以及对敏感操作的人工确认等控制手段，以应对误点击、页面结构变化和提示注入等风险。

在金融授信、医疗诊疗、工业控制、城市运行等高责任场景，智能体目前适合承担的职能限于资料整理、风险提示、建议生成与审计材料准备，最终决策仍需人工确认，并明确责任归属与合规留痕。从国内实践看，智能体平台化落地通常遵循一条路径：先上线合规的模型与应用，再将办公、客服、制造、政务等场景逐步嵌入，最后推动企业流程的可执行化改造。需要注意的是，平台能力不等同于落地效果，实际成效仍需结合具体业务流程和关键指标来验证。

表 15 智能体落地场景、代表能力、关键指标与落地条件

| 落地场景 | 代表能力                   | 关键指标                     | 落地条件                           |
|------|------------------------|--------------------------|--------------------------------|
| 软件工程 | 修改代码、测试验证、缺陷修复、提代码合并请求 | 测试通过率、代码合并请求采纳率、缺陷修复率、回滚 | 软件工程天然具备代码评审、测试与版本回退等验证链路；实际效果 |

| 落地场景                   | 代表能力                      | 关键指标                                         | 落地条件                                                           |
|------------------------|---------------------------|----------------------------------------------|----------------------------------------------------------------|
|                        | (Pull Request)            | 率                                            | 取决于代码仓库质量、测试覆盖率及人工审查机制                                         |
| 客户服务与办公                | 工单分流与处理、知识问答、文书生成、流程办理    | 一次解决率、转人工率、平均处理时长                            | 场景流程清晰、指标可量化、结果易归因；涉及复杂判断、投诉纠纷或合规事项时仍需人工介入                     |
| 界面操作<br>(Computer Use) | 后台填表、订单查询、网页检索、费用报销、跨系统操作 | 任务成功率、误操作 (WebVoyager 约 87%、OSWorld 约 38.1%) | 适用于缺乏 API 接口的存量系统，可通过界面自动化完成操作；需设置沙箱环境、权限白名单，并对敏感动作进行人工确认      |
| 高责任决策                  | 资料整理、风险识别、建议生成、审计材料准备     | 人工确认率、留痕完整度、建议采纳率、复核通过率                      | 在金融、医疗、工业、城市治理等高责任场景中，智能体应定位为辅助工具；关键决策需保留人工审核，并满足可解释、可追溯、可回滚要求 |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

#### (四) 风险与趋势

智能体进入真实业务环境，面临多重风险。任务链条越长，越容易出现目标偏离、步骤遗漏和上下文遗忘；工具调用越频繁，成本和延迟随之上升；权限越开放，越容易发生越权操作和数据错误回写；流程越是跨部门，业务、IT、法务、合规等多方协同设计的难度也越大。因此，企业级智能

体部署的成败，更多取决于企业能否将既有流程重新设计为适合智能体执行的形态，模型本身的能力高低反而退居次要。

随着自主程度上升，安全风险随之放大。开源个人智能体生态已出现恶意技能、越权访问与安全漏洞；以 OpenClaw 为代表的开源框架，因系统权限较高、默认安全配置薄弱，已被国家互联网应急中心（2026 年 3 月）和中国互联网金融协会相继发布风险提示，指出其面临提示词注入、插件投毒和漏洞利用等攻击风险，可能被用于窃取敏感数据或非法操控交易。国内已限制国企、政府机构和银行在办公电脑上运行此类工具。企业侧同样面临“影子 IT”问题——员工擅自部署智能体，可能因网页、邮件或文档中的恶意指令而引发越权操作；多智能体协作链路越长，责任归属越复杂。

表 16 智能体部署风险点与默认治理配置

| 风险点    | 典型表现                    | 默认治理配置                    | 预期效果          |
|--------|-------------------------|---------------------------|---------------|
| 可靠性/幻觉 | 长链任务目标漂移、步骤遗漏、结果不可复现    | 任务拆分、多模型校验、反馈循环、阶段性检查点    | 降低执行偏差，提高可复核性 |
| 成本累积   | 长上下文、多工具调用与重复校验推高单位任务成本 | Token 预算、模型路由、缓存复用、轻重模型分层 | 控制单位任务经济性     |
| 权限与安全  | 越权访问、误删数据、错误回写或泄露敏感信息   | 最小权限、沙箱环境、操作白名单、审计日志      | 防止高风险操作失控     |
| 提示注入   | 网页、邮件、文档中的恶意指令诱导越权      | 内容隔离、指令分级、外部内容不可信、敏感动作二次确 | 降低外部内容劫持风险    |

| 风险点    | 典型表现                | 默认治理配置                    | 预期效果          |
|--------|---------------------|---------------------------|---------------|
| UI 误操作 | 误点按钮、误填字段、误提交表单     | 预览确认、只读优先、关键动作人工批准、回滚机制   | 控制浏览器/桌面智能体风险 |
| 多智能体失控 | 子代理重复调用、互相误导、责任不清   | 身份标识、任务边界、消息日志、主控代理与熔断器   | 保持协作链路可追踪     |
| 组织适配   | 责任不清、人工接管缺位、ROI 难量化 | 明确责任人、失败回滚、业务指标闭环、上线前流程重构 | 让智能体从试点平稳进入流程 |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

### 智能体规模化落地过程中，渗透速度与失败率同步高企。

一方面，渗透速度极快——Gartner 预计到 2026 年底，约 40% 的企业应用将内置任务型智能体（2025 年这一比例不足 5%），到 2028 年，约三分之一的交互将由智能体前端替代原生应用界面。另一方面，失败率同样高企——预计到 2027 年底，超过 40% 的智能体项目可能因成本上升、价值不清或风险失控而被叫停或取消。成本、价值与治理，是决定能否从试点走向规模化的核心变量。

未来两至三年，智能体更可能以“人机协同”的形态进入企业，而非直接完全接管业务。其中，低风险业务率先实现自主运行，高责任业务保留人工确认；通用办公场景先提升效率，金融、医疗、工业与城市运行等场景更强调可解释、可审计与可回滚。评价标准亦转向任务成功率、人工接管率、错误恢复率、单位任务成本与审计完整性。智能体的成熟度，

最终取决于能否在清晰边界内稳定交付可验证的任务结果，自主程度本身不是衡量标准。

### 三、AI 芯片国产替代加速

#### （一）AI 芯片竞争格局

全球 AI 芯片市场处于高速扩张阶段。据 Precedence Research，2025 年全球 AI 芯片市场规模约为 944 亿美元，预计 2026 年将超过 1200 亿美元。国产替代亦在加速推进。据 IDC 数据，2025 年中国 AI 芯片总出货量约 400 万张，其中国产 AI 芯片出货量约 165 万张，市场份额从 2024 年的约 30% 提升至 41%，出货量较 2024 年（约 82 万张）增长约一倍，进口芯片占比约 59%。进口芯片中，英伟达仍居首位，出货约 220 万张，市占率降至 55%；AMD 出货约 16 万张，占比约 4%。国产阵营中，华为昇腾以约 81.2 万张的出货量位居国产第一，占据国产总出货量的近一半，市占率达 20%。在出口管制与自主可控需求的推动下，国产替代既获得了市场空间，供给安全也愈发凸显。

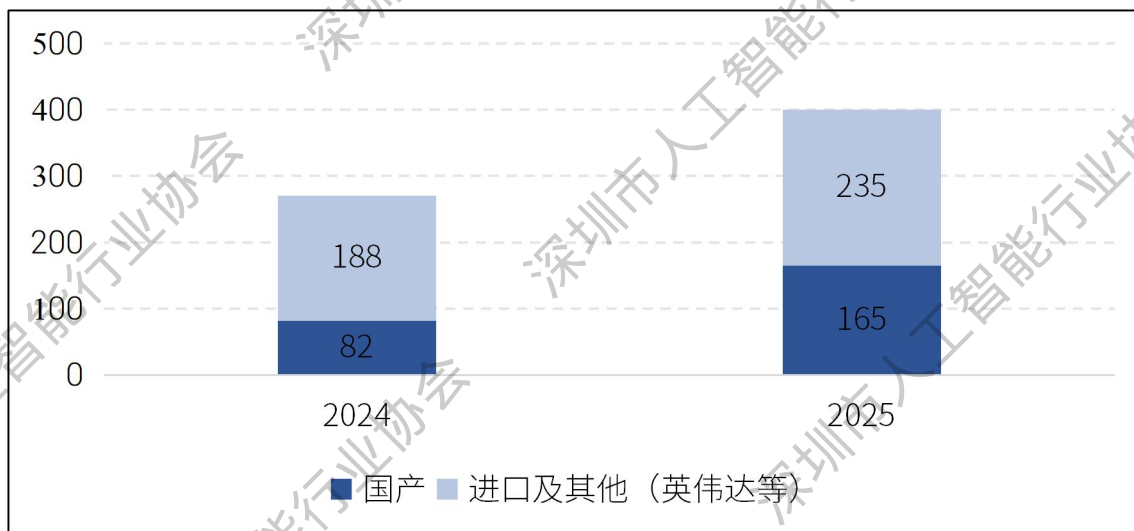


图 11 中国 AI 芯片出货量：国产与进口（单位：万张）

数据来源：IDC，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

海外 AI 芯片的竞争格局已相对清晰。旗舰 GPU 用于最前沿的训练，云厂商自研的 ASIC 芯片承担自家的固定负载，大显存 GPU 主要用于推理；竞争的重心也由单颗芯片转向整机柜与整集群的系统能力。英伟达仍处于领先地位，2026 年 2-4 月数据中心业务（主要为 GPU 与加速器）收入达 752 亿美元、同比增长 92%；AMD 主打大显存推理芯片，Google、AWS 则以自研芯片组建大规模集群。总体而言，海外 AI 芯片按训练、推理与专用负载形成了更细的分工，芯片间的互联带宽已与单卡性能同等重要。

国产 AI 芯片加速进入行业落地阶段。受出口管制、政企采购、运营商云部署与地方算力基础设施规划政策影响，国产 AI 芯片优先布局政企云、行业推理与私有化部署等场景，而最前沿的训练卡仍以海外产品为主。2025 年中国市场 AI

芯片出货约 400 万张，其中国产厂商合计约占 41%，国产厂商出货量同比增长约一倍，表明国产替代已加速进入行业落地阶段，但尚未实现对海外高端 GPU 的整体替代。从产品看，当前国产出货以华为昇腾为主。以昇腾 910C 为例，采用 7nm 制程，单卡约 128GB HBM、带宽约 3.2TB/s、BF16 算力约 780TFLOPS，据 SemiAnalysis 等机构研判，其综合性能接近英伟达 H100 的 80%。

受企业端突破与外部管制收紧双重推动，国产替代空间明确。2025 年，寒武纪营业收入约 65 亿元并首次实现年度盈利，海光信息持续盈利；摩尔线程、沐曦、壁仞、燧原等国产 GPU 企业亦相继上市或推进 IPO。与此同时，美国对华芯片出口管制进一步强化了国产替代动力。英伟达在最新业绩展望中已不再计入来自中国的数据中心收入；对国内客户而言，采购高端 AI 芯片需综合考量许可证、交付周期与供给连续性，而不再仅取决于价格与性能。在此背景下，国产芯片在政务、金融、运营商、能源等场景获得了更为明确的替代空间。



图 12 对华 AI 芯片出口管制时间线

资料来源：美国商务部产业安全局（BIS）等，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

## (二) AI 芯片技术路线

AI 芯片各技术路线的核心差异，主要在于其适配的应用场景。训练芯片追求大规模集群的协同算力，对显存、芯片间互联与长时间稳定运行要求较高；推理芯片更看重响应速度、并发处理能力与单位成本；面向终端的芯片，则需在功耗、成本、隐私与实时响应之间取得平衡。就路线分工而言，GPU 仍是通用训练与通用推理的主力，软件生态最为成熟、完整；国产 NPU 与加速卡更强调自主可控，适合政企云与行业私有化部署；云厂商自研的 ASIC 芯片适合支撑自家固定的大规模业务；面向手机、汽车、机器人与工业等终端的芯片则以低功耗与本地运行见长。因此，衡量一款芯片是否合适，关键在于其在真实业务中的实际速度、稳定性、能耗、迁移工期与长期供应，仅看峰值算力并不充分。

表 17 AI 芯片各技术路线的代表案例与适用场景

| 技术路线    | 代表案例                                  | 最适合场景                       | 主要优势                              | 主要边界                      |
|---------|---------------------------------------|-----------------------------|-----------------------------------|---------------------------|
| 通用GPU   | 英伟达 H/B系列、<br>AMD MI300/MI350         | 前沿训练、通用<br>推理、多模型混<br>部     | 生态成熟，工具<br>链完整，跨云和<br>本地交付能力<br>强 | 成本、功耗、<br>HBM和出口限<br>制压力高 |
| 云厂商ASIC | Google TPU、AWS<br>Trainium/Inferentia | 云内训练/推<br>理、固定负载、<br>大客户长期算 | 可围绕自有云<br>和模型服务深<br>度优化成本/能       | 生态锁定较强，<br>跨平台迁移成<br>本高   |

|            |                        | 力                   | 效                    |                       |
|------------|------------------------|---------------------|----------------------|-----------------------|
| 国产NPU/加速卡  | 昇腾、昆仑芯、寒武纪、海光等         | 行业推理、私有化部署、政企云、运营商云 | 供应可控，适合与行业系统和国产软件栈绑定 | 训练生态、长尾算子和大规模调优经验仍需积累 |
| 边缘AI SoC   | 车载、机器人、工业视觉和AI PC/手机芯片 | 端侧感知、本地推理、低时延任务     | 低功耗、低延迟、隐私保护和量产导入快   | 不适合承接前沿大模型训练          |
| 存算一体/专用低功耗 | 语音唤醒、传感器侧AI、穿戴设备       | 常驻小模型、低功耗连续感知       | 能效高、体积小、适合电池设备       | 模型规模、通用性和开发工具链受限      |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

海外技术路线已从单颗芯片转向系统级整合。英伟达将 Blackwell 架构、NVLink 与液冷机柜整合为“AI 工厂”，并以 GB300 及下一代 Rubin 延续迭代；Google 的 Ironwood TPU 将单个 Pod 扩展至 9216 颗芯片；AWS 则以逾 50 万颗 Trainium2 组成大规模训练集群。其共同特征，是将竞争由单卡峰值推向“整机柜—整集群—整生态”的系统能力。

中国方案则以系统规模弥补单卡性能的差距，强调“芯片—整机—云平台—行业负载”的纵向贯通。以华为 CloudMatrix 384 超节点为例，其将 384 颗 910C 以全光方式互联，按厂商披露的峰值口径，BF16 总算力约 300 PFLOPS、

约为英伟达 GB200 NVL72 的 1.7 倍，显存容量与带宽约为英伟达的 3.6 倍和 2.1 倍；代价是整机功耗与每 FLOP 能耗约为英伟达 4.1 倍和 2.5 倍，反映了当前以规模与电力换取性能的技术路径。在实际负载中，该超节点已实测可稳定支撑 DeepSeek-R1 等大模型的推理。昆仑芯、寒武纪、海光等也分别在云内推理、推训一体与行业部署等方向布局。国产芯片当前规模化落地场景，多来自“确定负载、可控环境与行业合规”，而非直接挑战最前沿的训练集群。

### （三）发展瓶颈与趋势

芯片制造与制程工艺是国产替代的核心短板。理解国产替代的瓶颈，需先区分芯片设计与制造这两个环节。芯片设计是绘制电路与架构，制造则是在晶圆厂将芯片设计实际生产出来。国内 AI 芯片企业（如华为海思、摩尔线程、沐曦等）多采用“fabless”（无晶圆厂）模式，即自身专注于芯片设计，而将制造委托给台积电、中芯国际等代工厂。芯片设计环节已形成多条技术路线，并非当前国产芯片的主要瓶颈；真正的制约在于芯片制造与制程工艺。国内代工方面，中芯国际已具备 7nm 级工艺的量产能力、5nm 级工艺仍在突破之中；但在缺乏 EUV 光刻设备的条件下，良率与成本均明显受限，先进制程晶圆与高带宽显存（HBM）尚未形成真正的国产闭环。

此外还需正视高带宽显存等“隐形依赖”。国产高端芯片所用的高带宽显存（HBM）长期依赖三星、SK 海力士的进口存货，国产长鑫存储的 HBM3 量产已推迟至 2026 年之

后，与三星、SK 海力士仍有约 2-3 年的代际差距，华为已公布昇腾 950 系列搭载自研 HBM 的计划，但在此之前，高带宽显存仍是制约国产高端卡有效算力的主要约束之一；先进制程所需的 EDA 设计工具约八成市场由国际三巨头（Synopsys、Cadence、Siemens）占据，先进光刻设备亦在禁运之列。据摩根士丹利等机构测算，2024 年中国半导体整体自给率约 22%-23%、其中制造环节仅约 18%；若将进口裸片与 HBM 封装而成的成品也计入“国产”，这一比例还会被明显高估。

**软件生态的完善度，直接影响国产芯片的规模化落地。**目前推理层的迁移门槛正在降低，vLLM、SGLang 等推理框架已支持多芯片部署，华为 CANN 生态于 2025 年 8 月开源后，DeepSeek 新模型已实现原生适配，企业在推理场景下切换到国产芯片的技术障碍正在减少。但训练侧仍需解决长尾算子覆盖不全、分布式并行方案不成熟与故障恢复机制不完善等问题，由 CUDA 生态迁移的成本依然较高。

表 18 国产 AI 芯片主要瓶颈的可缓解性研判

| 瓶颈    | 短期 | 中期 | 长期 | 结论                        |
|-------|----|----|----|---------------------------|
| 先进制程  | 低  | 中  | 中高 | 最核心、最硬约束，短期难靠单点工程完全<br>补齐 |
| HBM供给 | 低  | 中低 | 中  | 对训练和长上下文推理影响最大            |
| 高速互联  | 中  | 中高 | 高  | 封闭栈内可先优化，跨厂商标准仍需时间        |

|          |   |    |   |                   |
|----------|---|----|---|-------------------|
| 软件生态     | 中 | 中高 | 高 | 推理先缓解，训练后缓解       |
| 客户迁移与可靠性 | 中 | 高  | 高 | 依赖真实项目、SLA和运维体系沉淀 |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

从国产替代节奏看，未来一至两年内最现实的突破口在于推理、政企私有化与行业场景；三至五年内，国产芯片有望在运营商、金融、政务、能源等行业训练集群中形成规模化的应用；更大规模的通用基础模型训练，则仍取决于先进制程、HBM、互联与软件生态能否协同突破。判断产业替代不应只看单卡参数，更要考察“芯片—服务器—云平台—模型—应用”这一链路能否稳定交付、可运维、可复购。

## 第五章 应用纵横：跨界融合与赋能实例

### 一、人工智能终端

人工智能终端，是指搭载 AI 芯片与本地模型、具备端侧感知与推理能力、并可在端云协同下提供智能服务的终端设备，涵盖 AI 手机、AI PC、AI 眼镜、智能网联汽车与具身智能等品类，是人工智能触达用户的下一代入口。《深圳市加快推进人工智能终端产业发展行动计划（2025-2026 年）》进一步将人工智能终端细分为手机、计算机、平板、大模型一体机、可穿戴、影像、全屋智能、工业级与其他新型九大品类，并提出 2026 年全市产业规模力争万亿元。

但终端设备具备 AI 能力，并不等于用户已经用起来。硬件出货快速放量，真实使用却未同步：S&P Global 调查显示，仅约 47% 用户知道自己的设备内置 AI；在知道的用户中，约 87% 的三星、73% 的苹果用户仍认为现有手机 AI 价值有限。可见衡量 AI 终端是否成熟，发布会参数与出货量并不够，关键在于能否嵌入日常任务、形成持续使用与真实的任务执行。这一进程在手机、PC、智能眼镜与车载/机器人等品类间并不同步。

#### （一）发展概览

**市场规模快速增长。**IDC 预测全球 AI 手机出货由 2024 年约 2.34 亿台（占整体约 19%）升至 2025 年超 3.70 亿台

(约 30%)、2026 年有望达约 6.88 亿台；Gartner 预测全球 AI PC 出货由 2024 年约 3814.5 万台升至 2025 年约 7779.2 万台、2026 年约 1.431 亿台 (约占 PC 市场 54.7%)；Omdia 估算 2025 年全球 AI 眼镜出货约 870 万台、同比增长约 322%，2026 年预计逾 1500 万台。AI 手机基数最大，AI PC 进入换机周期，AI 眼镜增速最快但仍属小品类。

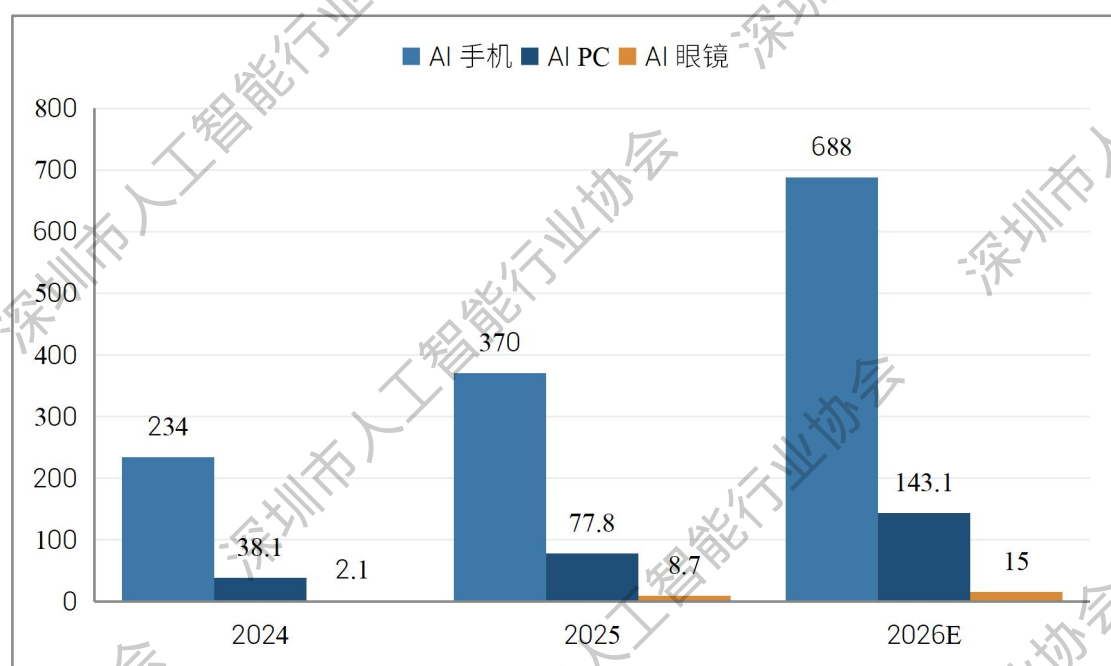


图 13 近年来全球 AI 终端年度出货量 (单位：百万台)

数据来源：IDC、Gartner、Omdia，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

AI-ready 与 AI-active 之间存在明显落差。AI 终端出货量仅反映设备是否具备 AI 能力，真实成熟度取决于用户是否持续使用。S&P Global 调查显示，受访用户中约 47% 知道设备内置 AI，其中约 48% 经常使用，折合口径内约 22%；Counterpoint/SellCell 调查中约 87% 三星、73% 苹果用户认

为现有手机 AI 功能价值有限。落差的另一面亦有积极信号：QuestMobile 数据显示，2025 年末六家手机厂商 AI 助手月活合计约 5.59 亿，AI 原生 App 使用时长同比增长约 22.3%。总体而言，人工智能终端设备能力与出货量持续提升，而用户认知、使用频率相对滞后。

表 19 AI 终端用户 AI 认知度与使用率调查

| 指标             | 数值        | 口径/来源                      |
|----------------|-----------|----------------------------|
| 用户知道设备内置 AI    | 约47%      | S&P Global 451 Research    |
| 知道者中经常使用 AI    | 约48%      | S&P Global 451 Research    |
| 三星用户觉得 AI 价值有限 | 约87%      | Counterpoint/SellCell 2025 |
| 苹果用户觉得 AI 价值有限 | 约73%      | Counterpoint/SellCell 2025 |
| 六厂商 AI 助手月活合计  | 约5.59亿    | QuestMobile 2025.12        |
| AI 原生 App 时长   | 同比增长22.3% | QuestMobile 2025.12        |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

**中国市场硬件普及快、高频应用仍在形成。**据 IDC 统计，2025 年全球智能眼镜（含 AI/AR/VR/MR 眼镜）出货约 1477.3 万台、同比增长约 44.2%，其中中国智能眼镜出货约 246 万台、同比增长约 87.1%，增速明显高于全球；同期中国智能眼镜中支持大模型语音助手的产品比例已达约 50.5%，但功能多停留于问答、翻译等通用场景。《人工智能终端智能服务能力分级指南》（T/TAF 289—2025）将人工智能终端智能服务能力划分为 L1 响应级至 L5 自主级，判断 2025 年 AI 手

机整体达到 L3 “智能助理级”。

## (二) 细分领域进展

**AI 手机已形成体系化布局。**苹果 Apple Intelligence、三星 Galaxy AI、华为小艺、荣耀 YOYO、OPPO 小布、VIVO 蓝心小 V 等系统级助手均已内置，成熟场景集中在影像编辑、实时翻译、摘要、相册检索、日程整理与跨应用搜索。中国信息通信研究院测试显示，AI 手机的辅助通话与影像处理在 2025 年取得突破，4 款主流终端达 L3 “智能助理级”，但跨应用权限、隐私与价值感仍是难点。现阶段 AI 手机的可靠价值，仍停留在高频功能更顺滑，而非代替用户办事。

**AI PC 是端侧 AI 成熟度最高的品类。**按端侧能力评级，AI PC 是唯一达到 A+级的：NPU 算力门槛明确( $\geq 40$  TOPS)，可在本地运行 7B-13B 模型，代表产品包括微软 Copilot+PC 及联想、惠普、戴尔等机型，会议纪要、文档摘要、本地知识库与代码辅助最易沉淀使用。按广义口径统计，2025 年 AI PC 在 PC 市场的渗透率已达 41.5%，预计 2026 年将达到 54.7%；按狭义口径统计（具备本地模型运行能力，NPU 算力 $\geq 40$  TOPS），2025 年 AI PC 在 PC 市场的渗透率为 19.6%，2026 年预计将达 30%，相较 2024 年的 0.5%已大幅度跃升。

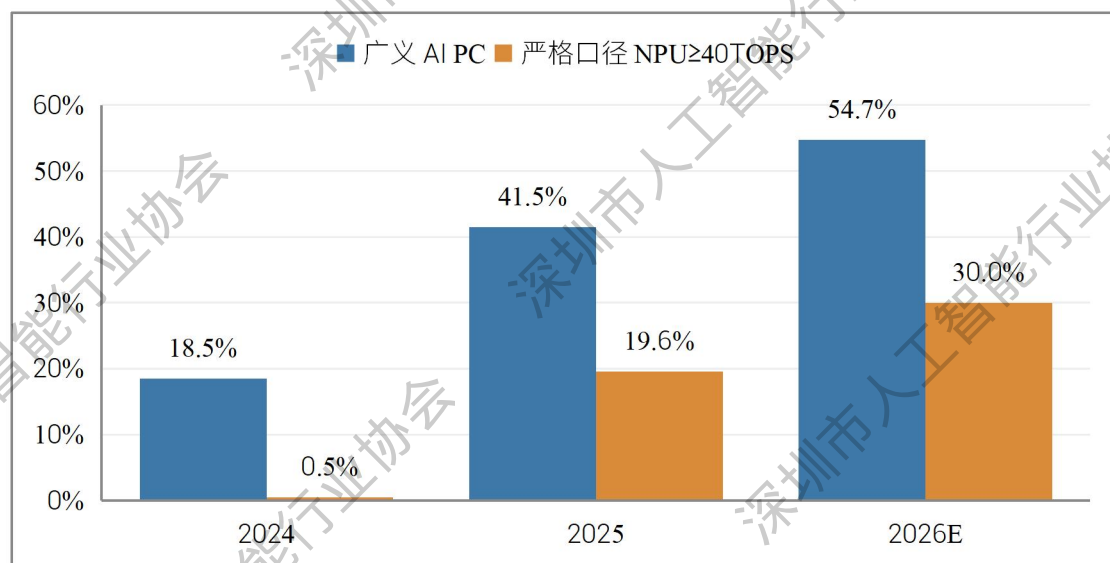


图 14 近年来全球 AI PC 市场渗透率（单位：%）

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

**AI 眼镜增长最快，但留存率尚未跟上。**AI 眼镜的代表产品有 Ray-Ban Meta，以及国内的雷鸟、Rokid、XREAL、影目、小米等，场景集中在第一视角拍摄、语音记录、实时翻译与导览。调研显示，行业平均退货率约 30-50%、Ray-Ban Meta 的 AI 功能使用率不足 2%、逾 67% 早期用户尝鲜后停用。产品结构正由听觉助手转向第一视角，2025 年带摄像头的拍摄眼镜占比由一季度约 7% 升至四季度约 39%。会议记录、维修辅助、文旅导览较易落地，人脸识别、无提示拍摄则受隐私与权限约束。

**智能座舱、具身智能把 AI 带入物理世界，安全与责任是首要约束。**智能座舱渗透最快，2025 年中国智能座舱市场约 1545 亿元、前装搭载率突破 40%、域控出货约 1330 万台，高通骁龙 8397 等舱驾一体芯片已支持百亿参数级模型上车，离线语音与多模态座舱助手成为标配方向。人形机器人算力

要求更高，以端侧 VLA（视觉—语言—动作）自治为特征、单机算力达 100-2000TOPS，2025 年全球出货超 1.3 万台、中国占比约 80%（宇树科技与智元机器人合计约占国内 56%）。与消费电子不同，这类终端必须服从实时性、安全性、低功耗与责任边界，不能把“上线后再学习”当作主要安全策略。

**端云协同决定了 AI 终端的运行方式。**本地模型适合处理高频、低延迟、隐私敏感与弱网任务（语音唤醒、实时翻译、相册检索、会议摘要、本地知识问答），云端模型则承担复杂推理、大规模检索与多模态生成，再由策略层决定哪些数据可以上传、哪些任务必须留在本地。能效是这一分工的现实约束：中国信息通信研究院以 1.5B 开源模型在手机、平板连续运行 6 小时的测试显示，端侧推理在整机电量中占比可观，而 2025 年主流机型电池虽已普遍超过 4800mAh、受体积所限仍难无限扩容——因此高耗能任务必须谨慎决定是否留在本地。

**系统级 AI 入口正在重构软件分发。**过去，用户需通过屏幕菜单完成 App 操作；如今，用户可以直接用自然语言、语音、视觉或跨应用指令来让 AI 完成任务——荣耀曾以 YOYO 演示“一句话下单 2000 杯咖啡”，体现智能体对传统 App 交互模式的调度能力。谁掌握系统级 AI 入口，谁就更可能掌握数据、交互入口与服务分发权。与之配套，MCP、A2A、AP2 等智能体协议，正推动跨应用执行走向标准化。

在使用层面，AI 终端的改变可从三个层面观察。对消费

者，它重塑拍照修图、实时翻译、搜索入口、日程整理、购物决策与陪伴交互；对企业，它重塑会议纪要、文档摘要、代码助手、销售陪访、门店培训与现场巡检；对产业，它推动软件分发由“App 菜单”转向“系统级 AI 入口”。其中，低风险、高频、可即时反馈的任务最先形成粘性，跨应用执行与物理动作则更依赖权限、审计与责任机制。

### （三）瓶颈与趋势

当前瓶颈集中在五处。一是使用率，AI-active 指标尚未统一，高频使用仍集中于摘要、翻译、影像、会议与搜索。二是成本与算力，端侧主流设备的算力与内存仍是约束，模型常驻进一步带来发热、续航与存储压力，并通过 DRAM、SSD 成本传导至终端价格。三是标准，能力分级、意图框架与评测标准仍在建立。四是隐私，终端越了解用户，越须明确数据在何处处理、由谁调用、能否删除与审计。五是真实任务闭环，跨应用执行涉及支付、位置、通讯录与第三方权限，须以审计与人工确认兜底。

未来三年，AI 终端不会只比拼出货量，而会比拼三件事：AI 能否嵌入日常任务、形成持续使用与真实的任务执行，AI 功能是否高频使用，端云协同是否能在隐私、成本与任务完成率之间取得平衡。AI 终端真正进入主流的标志，是入口、功能、数据、权限与业务流程共同过关——竞争的重心，已由硬件出货增长，转向入口重构、软件分发重构、端云协同与真实使用率之争。

## 二、工业智能

工业智能是人工智能从数字世界走进实体经济的关键场景，以智能模型、数字孪生和工业智能体等技术为底座，将人工智能系统融入研发设计、生产制造、质量管理、设备运维、能耗优化和供应链协同等场景。其核心在于融合现场感知数据、工业机理模型与领域专家知识，将海量数据转化为可验证的经营改进，使工业系统具备诊断、预测、优化和协同决策的闭环智能能力。它叠加在自动化、工业互联网和数字孪生之上，补上诊断、预测、优化和协同决策这一层。

依托制造业规模、工业数据、设备密度、场景复杂度和产业链完整度优势，中国发展工业智能具备独特基础。但工业智能越深入生产现场，越需要工业软件、数据接口、边缘设备和工程系统协同支撑。其中，高端工业软件和工业知识沉淀不足，仍是制约能力提升和规模化复制的短板。因此，中国推进工业智能应坚持场景牵引和工程落地并重，先从痛点突出、价值明确的场景切入，打通数据采集、模型应用、现场反馈和业务改进的可验证的闭环，同时补齐工业软件、数据标准和工业知识库等基础。衡量工业智能成效，应看缺陷率、停机时间、单位能耗、交付周期和设备综合效率等硬指标是否改善，而非模型演示是否惊艳。



图 15 工业智能闭环示意图

### （一）发展概览

中国工业智能的发展，建立在庞大的制造业规模与场景之上。据国家统计局数据，2025 年中国工业增加值约 41.7 万亿元、同比增长 5.8%，制造业增加值约 34.7 万亿元、同比增长 6.1%，制造业规模有望连续 16 年位居全球第一。如此庞大而门类齐全的制造体系，不断沉淀出海量工业数据与高密度应用场景，而这正是工业智能区别于通用人工智能、得以扎根生长的关键资源。

中国的工业自动化基础同样位居世界前列。国际机器人联合会（IFR）《World Robotics 2025》显示，2024 年全球工业机器人新装约 54.2 万台，其中中国约 29.5 万台、占全球约 54%；中国在役工业机器人已超过 200 万台，本土供应商在国内市场的份额升至 57%，为工业智能在生产现场落地提供了重要基础。

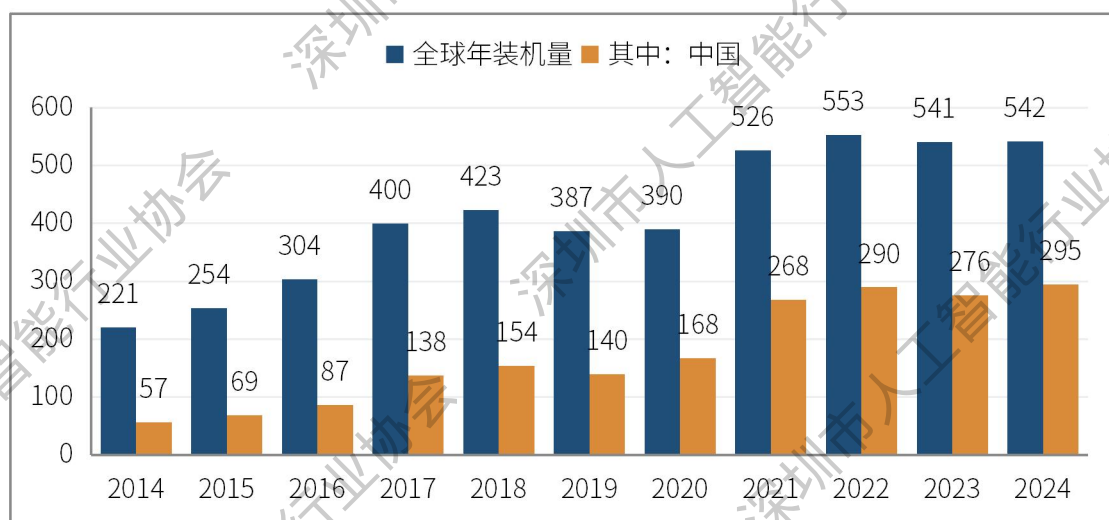


图 16 近年来全球及中国工业机器人年装机量（单位：千台）

数据来源：国际机器人联合会（IFR）《World Robotics 2025》，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

智能工厂加速推进更能说明中国制造正在从设备自动化走向系统智能化。据工业和信息化部披露，截至 2025 年初，全国已建成基础级智能工厂 3 万余家、先进级智能工厂 1200 余家、卓越级智能工厂 230 余家；其中卓越级智能工厂覆盖超过 80% 的制造业行业大类，累计形成智能仓储、在线检测、数字化研发设计、智能排产调度、质量追溯等近 2000 个优秀场景。更重要的是，这批卓越级智能工厂已展现出明确的经营成效：产品研发周期平均缩短 28.4%，生产效率平均提升 22.3%，不良品率平均下降 50.2%，碳排放平均减少 20.4%。这些成效表明，智能化演进正在为中国制造带来更高水平的生产效率和综合绩效。

工业和信息化部数据显示，2024 年规模以上工业企业的数字化研发设计工具普及率达 84.1%，关键工序数控化率达

66.4%；按相关政策部署，到 2027 年两项指标将分别提升至 90%和 75%以上，同步推动约 200 万套工业软件与 80 万台套工业操作系统的更新换代。这意味着，工业智能的首要任务并非急于堆叠模型，而是先补齐数据采集、软件系统、边缘控制与工业操作系统这一整套工程底座。

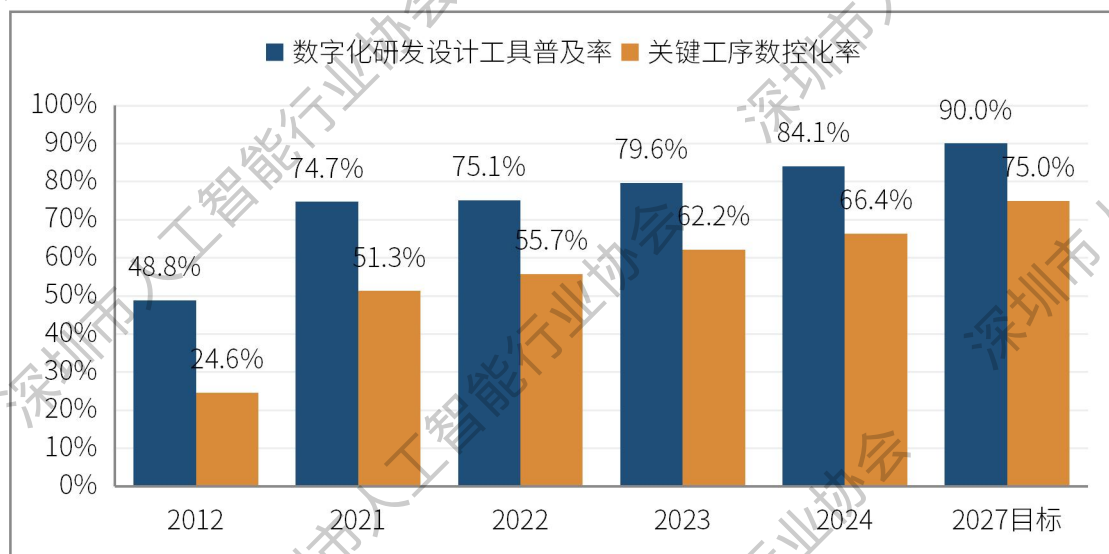


图 17 近年来规模以上工业企业智能制造普及率（单位：%）

数据来源：工业和信息化部，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

从全球范围看，制造业对人工智能的投入正在进入长期增长周期。据 MarketsandMarkets 估算，全球制造业人工智能市场规模将由 2025 年的约为 341.8 亿美元增至 2030 年的约 1550.4 亿美元，2025-2030 年的年均复合增速约 35.3%。相比市场预测，灯塔工厂数量和典型案例更能反映人工智能在生产现场的实际落地。灯塔工厂是世界经济论坛（WEF）与麦肯锡共同遴选的全球灯塔网络成员工厂，是全球范围内

在数字化、智能化和绿色化转型方面具有示范意义的标杆工厂。2026 年 1 月，全球灯塔网络新增 23 个工业站点，总数达到 224 个，其中新增站点有 16 个位于中国。工业和信息化部也于 2025 年发布 151 项“人工智能赋能新型工业化”典型应用案例，推动单点探索逐步沉淀为可复制、可推广的场景经验。

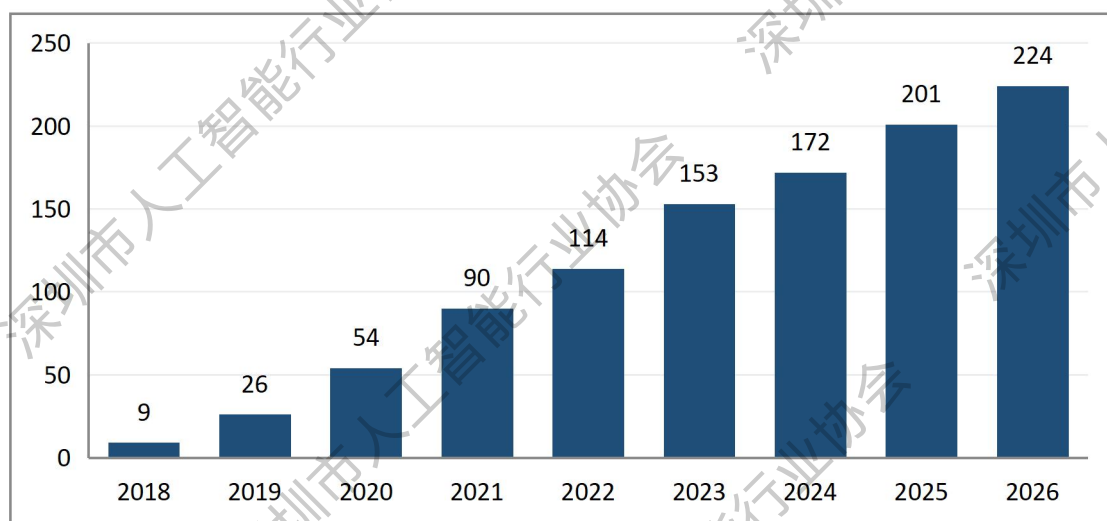


图 18 近年来全球灯塔工厂累计家数（单位：家）

数据来源：世界经济论坛，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

综合来看，中国发展工业智能的基础，来自庞大完整的制造体系、持续提升的自动化和数字化水平，以及不断积累的智能工厂和典型应用案例。这些条件使中国具备工业智能规模化落地的独特优势，也让场景、数据和工程验证成为中国工业智能发展的重要资源。但与此同时，优势更多集中在应用端，基础层面的短板仍需补齐。高端工业软件、数据标准、系统接口和工业知识复用能力，仍制约着工业智能从单

点应用走向系统能力。下一阶段，中国推进工业智能，应在场景牵引中跑通可验证闭环，并同步补强工业软件和工程系统底座，使模型能力真正进入研发设计、生产控制和运营管理流程，转化为可持续、可复制的产业能力。

## （二）技术路径

从技术路径看，工业智能是一个由现场数据、工业模型、应用执行和治理安全共同构成的架构。底层是现场数据层，需要通过设备、控制系统和业务系统，将设备运行、工艺过程、能耗计量等信息沉淀为可计算、可追溯的数据基础；其上是工业模型层，涵盖视觉检测、预测维护、工艺与排产优化，以及工业知识库与工业大模型；再向上是应用执行层，负责把模型结果嵌入 MES、ERP、SCADA、PLC、APS 等软件与边缘设备；最上层是治理安全层，处理实时性、可解释性、网络安全、权限与责任边界等问题。

这一技术分层也决定了大模型在工业场景中的位置。现阶段，大模型更适合承担知识检索、人机交互和任务编排的工作，例如查阅手册、解释报警信息、生成工单、辅助编写代码、调用工业软件。场景中，真正承担核心计算与现场决策的，仍是机器视觉、时序预测、异常检测、约束优化、数字孪生以及既有的工业软件。毫秒级的控制与安全联锁，依旧由 PLC、DCS 和保护系统负责，大模型不宜在未经验证的情况下改写关键工艺参数。McKinsey 对全球灯塔网络用例的梳理也印证了这一判断。2025 年灯塔网络的代表性高频用例中，约 77% 由分析型人工智能支撑，仅约 9% 使用生成式

人工智能。

### (三) 重点场景赋能

质量检测是工业智能最先成熟的典型场景之一。在电子、光伏、电池、钢铁等制造环节，视觉模型识别表面缺陷、尺寸偏差与装配异常，并与边缘设备和质检系统联动，将质量管控从事后抽检前移至在线监测和实时拦截。它之所以率先成熟，是因为缺陷样本相对明确、损失可直接折算为成本、判断结果能够当场验证。在世界经济论坛的灯塔工厂中，通威太阳能借此将电池转换效率提升约 12%、缺陷率降低约 41%；亿纬锂能荆门工厂的缺陷率则降低约 52%。

预测性维护是另一类高价值场景。它通过串通连接设备的振动、温度、电流、报警记录与维修工单等数据，在故障发生前识别异常趋势并给出预警，使设备维护从被动抢修转向主动安排。这类应用尤其适合能源、化工、轨道交通、重型装备等行业。因为这些行业的非计划停机成本高昂，且关键设备通常已部署较多传感器，能够持续采集温度、压力、振动、电流等运行数据，为模型判断设备健康状态提供基础。例如，卡塔尔的 Shell GTL 工厂通过部署 45 项第四次工业革命技术用例，提升了装置现场吞吐量 9%，并将运行可靠性提高至 99%，延长了设备寿命 50%。

工艺、能耗与排产优化是更复杂的场景。难点在于系统不能只判断状态或提前预警，还要在安全、质量、设备能力和工艺规则的约束下，给出下一步可执行的调整方案。因此，

单靠预测模型远远不够，需要将预测模型、优化算法、仿真或数字孪生、安全规则与人工审批结合起来，形成可验证、可执行的决策流程。其收益主要体现在良率提升、能耗下降和交付周期缩短等方面，但对数据质量、模型稳定性和工程集成能力的要求也更高。在工艺与能耗优化方向，已有部分钢铁、化工与能源企业，将空压站、锅炉、窑炉、压缩机等高耗能环节作为试点入口。

供应链与仓储的复杂性在于跨系统、跨环节的协同执行。订单预测、库存补货、路径规划与异常处置，往往需要跨 ERP、WMS、MES 等多个系统协同决策，其价值不在单点算法的精度，而在能否把订单、排产、库存、配送和人工审批串联成一条可执行的链路。美的厨卫电器芜湖工厂的灯塔案例即体现了这一特点。面对五级分销网络与小批量订单带来的复杂性，工厂通过实时客户订单管理、AI 先进计划排程系统和供应链“控制塔”等数字化用例，使端到端交货周期缩短 39%、库存周转天数减少 30%、市场缺陷率下降 86%。

行业大模型和工业智能体是中国正在探索的系统化路径。以中国有色金属行业的“坤安”大模型为例，它由行业协会与龙头企业联合主导、依托国产云与国产大模型搭建，已在矿山、冶炼、加工、安全环保与材料研发等环节推进百余场景、并建成 8 个行业高质量数据集。其价值在于把行业知识、数据集、算力和行业经验沉淀为可复用的能力底座。政策导向同样体现了这一思路。“人工智能+制造”专项行动提出，到 2027 年推动 3 至 5 个通用大模型在制造业深度

应用、推出 1000 个工业智能体、打造 100 个高质量工业数据集、推广 500 个典型应用场景。

从这些场景可以看出，中国工业智能的优势在于制造场景密度高、机器人和自动化基础较好、地方产业集群完整，能够为模型训练、场景验证和规模化推广提供土壤。但多数企业仍面临数据基础薄弱、工业软件系统分散以及复合型人才不足等约束。因此，对多数企业而言，务实的推进路径是先选定损失清晰、价值明确的场景，完成数据接入、业务基线与投资回报核算，再逐步向车间级、工厂级扩展，而不是一开始就自主建立工业大模型。

#### （四）趋势研判

工业智能的普及之所以慢于智能增效，根源在于它必须先把物理世界转化为可观测的数据，再将数据与业务动作连接起来，这远比处理文本和软件流程困难。当前真正制约其落地的瓶颈，大致可归为四类。其一是数据与系统集成：工业现场新旧设备并存、各家 MES 与 ERP 标准不一，缺乏统一的资产台账、数据口径与权限体系，导致模型容易止步于演示阶段。其二是实时性与安全边界：控制回路、安全联锁与高危工艺不允许带着风险上线试错，人工智能只能先从建议、计划和辅助决策环节切入。其三是投资回报与组织协同：工业智能项目往往横跨设备、工艺、质量、维修、采购与财务多个部门，一旦缺少共同的指标与明确的业务负责人，模型结果便难以进入主流程。其四是数据安全和知识产权：设备参数、配方与工艺诀窍均属高敏感资产，工业助手必须做

到权限隔离、来源可追溯，并确保客户数据不被用于训练底层模型。

展望未来三年，工业智能大致会沿三条路径推进。第一，从单点的视觉质检与预测维护，走向产线级的工艺与排产优化；第二，在交互方式上，工业智能从以问答和辅助分析为主的大模型应用，走向能够读取系统、调用工具、生成工单并发起审批的工业智能体；第三，工业智能将从少数示范工厂走向更多企业。随着行业平台、工业软件服务与边缘智能方案向中小企业扩散，有助于这些企业降低系统建设、算法开发和运维成本，成为工业智能规模化落地的重要路径。

工业大模型的价值不应止步于对话更流畅、报告更美观，而是在三个更具难度的方向上发力。其一，接入设备、产线与工业软件的现场数据，使模型基于真实工况而非通用语料运作；其二，把行业机理、工艺知识与质量经验沉淀进模型，使其能够理解约束、给出符合工艺的建议；其三，与 MES、PLC、数字孪生等系统打通，在真实 KPI 上形成可验证、可审计的闭环。沿着这一方向推进，工业大模型才能从知识层和交互层逐步进入诊断与决策支持环节。“坤安”大模型这类由行业牵头、自建数据集与场景库的实践，正是朝这一门槛迈出的起点。据 IDC 估算，2024 年中国工业大模型应用市场规模约为 12 亿元，相对庞大的制造业体量，这一规模仍处于起步阶段，后续增长取决于数据接入、系统集成和闭环验证能否跑通。

总体而言，工业智能并非一蹴而就的“工厂大脑”工程，

而是制造企业把现场数据、工业软件、模型、机器人与治理机制逐步连成闭环的长期过程。未来企业之间的差距，将更多取决于谁能更快地把数据、流程、责任与投资回报，沉淀为可复制的工业能力。

### 三、智能增效

智能增效指人工智能进入企业生产环节、提升岗位与流程效率的过程。在编程、客服、办公、营销等环节，AI 已带来可计量的效益；但这些效益能否转化为组织层面的持续回报，取决于企业是否相应重构了工作流程与考核体系，而非仅取决于所采购工具的数量。

#### （一）发展概览

企业 AI 支出持续高速增长，向贴近业务的应用层倾斜。据 Menlo Ventures，2025 年企业 AI 支出从 2023 年的 17 亿美元飙升至 370 亿美元，目前占据全球 SaaS（软件即服务）市场的 6%，增长速度超过历史上任何软件类别。其中，约 190 亿美元流向了面向用户的产品和软件，也即应用层，占比超过 50%，表明企业倾向于即时的效益提升，而非长期基础设施投资；面向特定岗位的部门 AI 支出达 73 亿美元，同比增长 4.1 倍，其中编程岗位的 AI 支出占比最高，达到 40 亿美元，占部门 AI 支出的 55%，其次为 IT 运维、营销、客服。采购方式也由自建转向外购，外购占比由 2024 年的约 53% 升至 2025 年的约 76%，反映商业化 AI 应用趋于成熟、可直接嵌入既有流程。

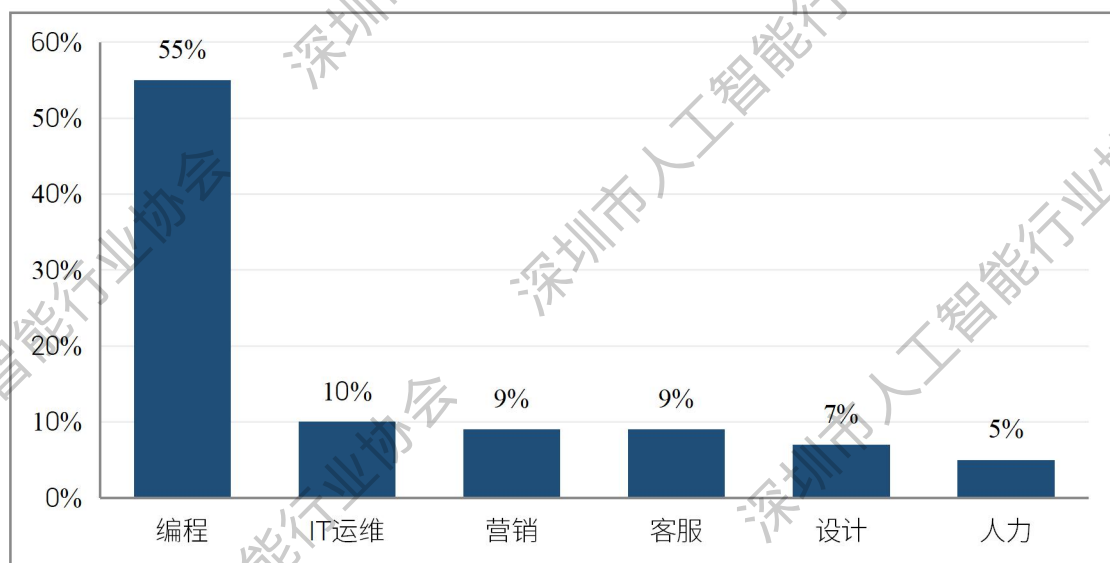


图 19 按职能分布的部门 AI 支出（单位：%）

数据来源：Menlo Ventures，深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

采购面虽广，但 AI 价值兑现高度集中于对工作流程进行重塑的企业。麦肯锡 2025 年覆盖 105 国、近 2000 份问卷的调查显示，仅约 6% 的企业属于由 AI 带动 EBIT（息税前利润）提升超过 5% 的高绩效企业，多数企业尚未实现规模化回报。差异主要体现在使用方式上：这 6% 的企业对工作流程进行重塑的比例约为其他企业的近三倍。采购与使用之间也存在落差：截至 2026 年 4 月，微软 AI 办公助手 Microsoft 365 Copilot 的付费企业已超过 2000 万，但第三方对企业部署的复盘普遍显示，实际持续使用比例明显低于已购数量。

## （二）重点场景赋能

编程是智能增效最为突出的应用场景，但提速幅度与代码库的成熟度高度相关。从规模看，谷歌于 2026 年 Cloud

Next 大会披露其新增代码约 75% 由 AI 生成，兴业银行于 2025 年年报披露其 AI 编程助手已覆盖约 90% 的研发人员，国内通义灵码、字节 TRAE、腾讯 CodeBuddy 等 AI 编程助手的使用规模也在快速扩大。从提速幅度看，在新功能开发等边界清晰的任务上，AI 辅助编程可明显缩短工时，受控实验显示提速约一半；而在规模大、迭代时间长、依赖复杂的成熟代码库中，METR 2025 年中的随机对照实验显示，熟悉项目的资深开发者使用 AI 后平均反而慢约 19%，且其事前预期与事后自评均为提速，未能察觉效率下降。原因在于成熟代码库的隐含约定较多，AI 容易生成需要返工的改动。因此编程提效较为稳定的环节，是检索、补全、测试与文档生成等可即时验证的部分，核心逻辑的改动仍需人工审查。

**客服场景是智能增效最先规模化应用的领域之一。**客户场景的优势在于成效可直接量化，解决率、转人工率、平均处理时长等均为现成指标。从行业实践看，成效已初步显现。兴业银行 AI 数字客服已累计服务约 5500 万人次；客服产品 Intercom Fin 平均解决率已提升至约 67%，并从按会员收费转为按“成功解决一次”收费（约 0.99 美元/次）；广东公牛集团基于飞书自建的“公牛智服”智能体可 7×24 实时响应上千家经销商咨询，使客服接待能力提升约 30 倍；美宜佳将智能客服“小佳”用于 3.6 万家门店，活动规则和商品信息等问题响应时长下降 58%。但客服场景并非可完全无人化，AI 适合先处理可标准化、知识库完备、规则明确的高频问题，复杂判断仍需人工把关。

**办公场景中，智能增效正向具体业务场景渗透。**华恒生物用 AI 实验记录员把语音转写为实验参数并回写系统，每年节省人力成本约 145 万元；网易智企推出企业级 AI Agent 管理平台帝王蟹（ClawHive），支持云端与本地部署，集成技能资产中心、AI 算力网关、知识库等模块，其“销售宝”解决方案提供客户跟进、策略规划与对练陪练等能力，帮助销售脱离日常重复性工作，有效项目跟进与成果转化率从 20%-40% 提升至 70% 以上，AI 对练陪练功能替代传统讲师培训，培训成本降低 50% 以上，新销售独立产出周期从 6 个月缩短至 2-3 个月；Meta 披露已有超过 100 万家企业在 WhatsApp 和 Messenger 上使用 Meta Business Agent。

**在设计与营销环节，AI 正在引发岗位重组。**低复杂度素材的生成成本急剧下降，品牌策略、转化评估和质量把关等高阶环节的价值占比则相应上升。从实践来看，生成式 AI 工具 Adobe Firefly 累计生成内容（含图像与视频）已超过 220 亿项；Klarna 借助 AI 将广告图片的制作周期从约六周压缩至约一周，外部代理开支下降约 25%，每年节省数千万美元。与此同时，岗位结构开始出现变化——平面设计类岗位数量有所收缩，用户体验、用户界面和品牌系统类岗位的需求则明显增长。需要指出的是，现阶段 AI 在该环节仍以放大人力产能为主，尚不能完全替代人工。

### **（三）瓶颈与趋势**

**智能增效潜力可观，但伴随着深刻的就业结构调整。**低技能、重复性岗位加速被替代，世界经济论坛《未来就业报

告 2025》预计，到 2030 年全球新增约 1.7 亿、减少约 9200 万、净增约 7800 万个岗位，受冲击较大的是数据录入、行政文书等岗位，而复核、判断与担责的岗位重要性上升。在企业层面，这一过程往往伴随管理层级精简，亚马逊 2025 年裁减约 1.4 万名企业员工，并明确将其与广泛应用 AI 带来的效率提升相关联。

### **增效收益仍集中于少数完成流程与治理重构的企业。**

MIT NANDA 研究指出，约 95% 的组织尚未从 AI 投入中获得可衡量回报。管理层最容易误判的是把 AI 当成单个软件采购，实际上智能增效往往需要同时重建流程、权限和责任：谁能调用客户数据，哪些任务允许自动执行，哪些任务必须人工确认，模型输出如何标注来源，错误如何上报和纠正，员工如何学习新的工作方式。没有这些配套，AI 工具会制造更多初稿、更多待审核内容和更多信息噪声；有了这些配套，AI 才可能真正减少重复劳动。

**“影子 AI”与数据合规风险突出。**Software AG、毕马威全球调查显示，约 50% 的员工使用过未经批准的 AI 工具，约 24% 的员工曾将公司专有数据输入公共 AI 系统，导致客户信息、商业数据、代码与合同条款等敏感内容面临泄露风险。对此，企业需要统一入口、权限隔离、日志审计、数据脱敏和高风险任务复核等机制，否则“增效”会变成新的合规和安全成本。

**未来三年，智能增效将经历三个阶段。**第一阶段以个人 AI 办公助手普及为主；第二阶段，客服、研发、销售、财务

等部门各自形成 AI workflow；第三阶段，企业知识中台、模型调度、Agent 编排与岗位分工、业务系统、关键经营指标协同运行，成为组织运转的基础设施。唯有进入第三阶段的企业，才能真正把 AI 从软件功能转化为组织生产力。衡量智能增效的标准，将从“部署了多少 AI 功能”，转向 AI 嵌入工作流程的深度、关键业务指标的改善幅度，以及组织内部的实际使用率。

## 四、智慧民生

### （一）发展概览

智慧民生，是将人工智能嵌入医疗、教育、政务与城市治理等公共服务，从而改善服务体验、提升运行效率。2024 至 2025 年，全球智慧城市市场规模由约 396 亿增至约 506 亿美元，智慧医疗市场规模由约 267 亿增至约 370 亿美元，智慧教育市场规模由约 52 亿增至约 70.5 亿美元。国内方面，医疗、教育、政务、城市治理等领域已被国务院列为深度实施“人工智能+”行动的重点方向；截至 2025 年底，国家中小学智慧教育平台注册用户超 1.63 亿，医生使用 AI 的比例由 2023 年的 38% 升至 2026 年约 81%，覆盖面已具备相当规模。

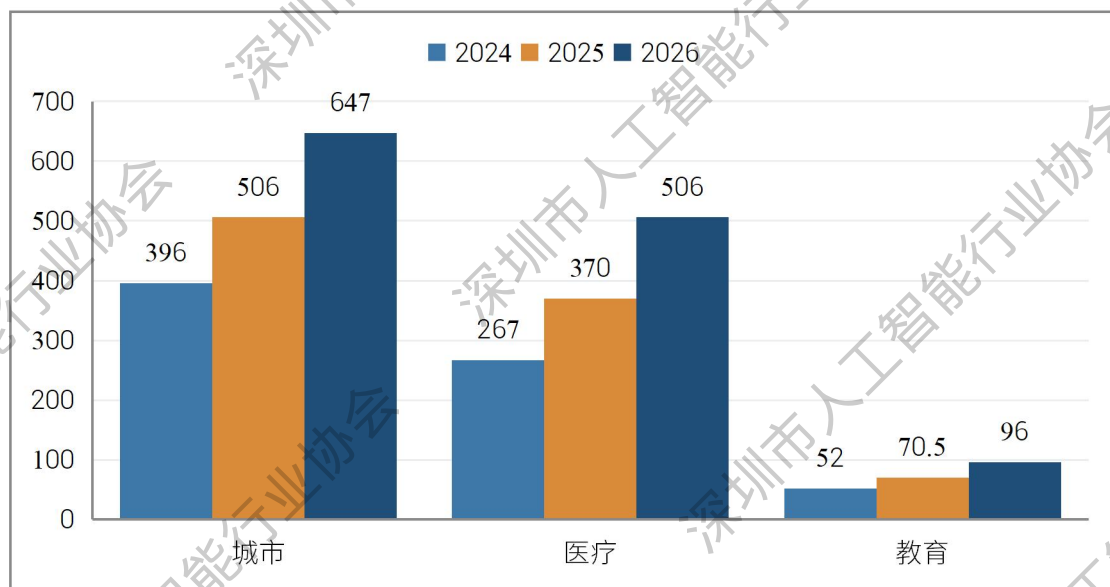


图 20 全球智慧民生细分领域市场规模（单位：亿美元）

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

## （二）重点场景赋能

医疗领域，人工智能主要应用于分诊导诊、病历文书、影像辅助与患者沟通等环节，在缩短就诊等待、减轻文书负担、辅助影像诊断方面成效明显。微软临床 AI 助手 Dragon Copilot 一个月辅助超 300 万次医患对话、每次就诊节省约 5 分钟文书时间，改善医生疲劳和患者体验；美国食品药品监督管理局累计授权的 AI 医疗器械由 2015 年的 6 个增至 2025 年的约 1450 个、影像与放射类占七成以上，国内联影智能、数坤科技、推想医疗、鹰瞳科技等企业也围绕胸部 CT、心脑血管、眼底和多病种影像辅助形成产品线。

教育领域，人工智能主要应用于备课、批改、个性化辅导与区域教学平台，在减轻教师负担、促进个性化学习方面成效较为显著。英国国家教育研究基金会（NFER）随机对照实验显示使用 AI 辅助教学可使教师每周备课时间减少约 31%；安徽蚌埠在 875 所学校落地人工智能智慧教育模式，教师备课时间减少 33%、英语作文批改时间减少 67.26%、学生无效练习降低 49.6%，使学生有更多时间用于全面发展；深圳龙岗区 AI 个性化学习项目规模化覆盖 50 余所学校，2025 年全年智能采集作业学情 197.2 万份，AI 作文批改达 19.2 万人次，已深度融入日常教学，生成超 7 万册个性化手册；横岗高级中学高三“一生一机”AI 平板教学试点，大幅提升课堂效率与学生学习效果，近半年来高三年级联考平均分提高 10 分，深一模本科率提高 16 个百分点。

政务领域，人工智能主要应用于政策问答、材料预审、

工单分派与办事协同，有效简化办事流程、减少材料提交、提升响应速度。深圳福田区约 70 名“AI 数智员工”覆盖 240 个政务场景、公文修正准确率超 95%；深圳龙岗区打造商事登记“智能审批助手”，已累计帮助工作人员完成商事登记审查业务近 3300 件，共审核审查要点近 20000 条，准确率稳定在 96% 以上，审批效率提升超 30%，大幅降低人工审核压力；张家口市 12345 政务服务便民热线系统接入 DeepSeek 大模型，工单准确率提升至 99.5%，分拨准确率提升至 95% 以上。

城市治理领域，人工智能覆盖交通、能源、建筑、水务、应急、城管等环节，价值集中体现在运营指标改善。交通方面，成都二环高架依托智能交通管控系统与匝道管控，使高峰时段通行效率提升了 40% 左右；能源方面，南方电网以 AI 负荷预测服务南方五省六十余地市，深圳虚拟电厂聚合充电与储能资源超 50 兆瓦，支撑新型电力系统调度；建筑方面，DeepMind 使谷歌数据中心制冷能耗降低约 40%；水务方面，鹰潭市智慧水务建设将管网漏损率由 35% 降至 8.59%、年节约成本约 600 万元；城市运行方面，武汉城运智能体联通 14 个市直部门和 16 个区调度系统，实时监测 1500 余项城市体征指标和超千万条城市运行事件数据，推动城市事件处置效率提升 60%；应急管理方面，投控数科在深圳宝安区搭建了全国首个 IoT+AI+CIM 的防汛视频指挥调度体系，通过 AI 视频分析强化对积水事件的主动识别能力，结合三维地理信息数据构建全要素回溯模型，解决防洪排涝难点，2024

至 2025 年汛期 AI 新发现积水易涝点 20 个，积水点处置时长从 2-3 个小时缩短到 1 小时以内。

### （三）瓶颈与趋势

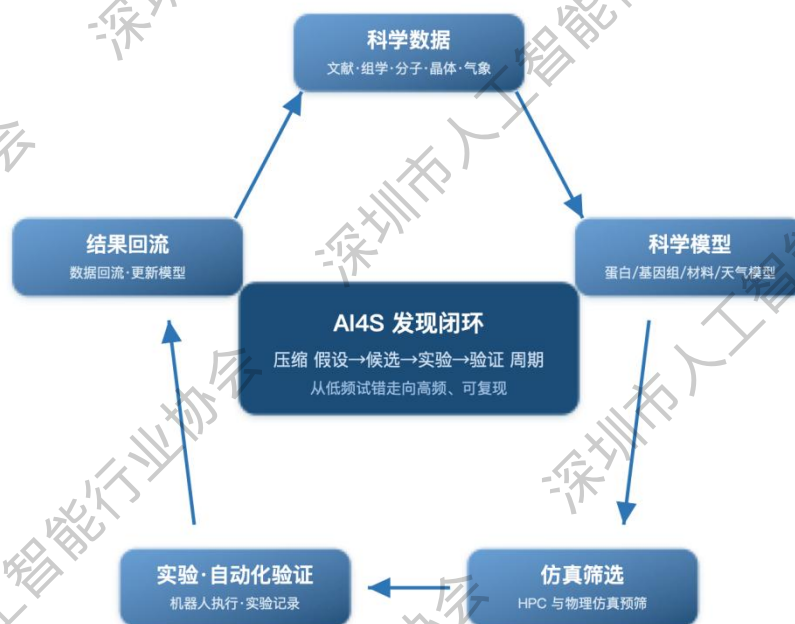
当前瓶颈主要集中在三处：一是服务对象的“数字鸿沟”，民生服务覆盖老年人、残障人士、偏远地区居民等“数字弱势群体”，当前标准交互环境以普通话、主流交互习惯、稳定网络与设备为假设，易形成公共服务供给的事实性不平等。二是公共服务决策不可完全自动化。民生领域部分决策具有法律效力或重大利益影响，例如低保资格认定、医疗诊断、保障房分配等，要求责任可追溯、结果可申诉，AI“黑箱”特性无法满足上述要求，仍需人工复核。三是财政与运营的可持续投入。民生服务的社会效益往往是长期的、难以量化的。财政决策倾向于支持看得见的硬件建设，导致许多智慧民生项目建成后陷入“建而不用、用而不活”的困境。

从发展态势来看，AI 将驱动民生服务从“被动响应”转向“主动预见”与“个性化服务”。在医疗领域，AI 正推动慢病管理从“被动治疗”转向“主动预防”，华为预测到 2035 年 AI 将助力预防超过 80% 的慢性病；IDC 则预计到 2030 年多模态 AI 将在症状出现前预测 45% 的慢性病和罕见病。在教育领域，AI 将支撑大规模因材施教，实现精准教学、减负增效。在城市治理领域，AI 将推动城市管理从人力密集、经验驱动向知识驱动、AI 研判跃升，让城市运行实现自我感知、自我决策、自进化治理。

## 五、科学智能

### （一）发展概览

科学智能（AI for Science, AI4S）正在由科研辅助工具演进为新型科研基础设施。其核心机制在于，将科学数据、模型推理、仿真计算、自动化实验和专家反馈连接为连续发现流程，从而压缩“提出假设—筛选候选—设计实验—验证结果”的周期。2024 年诺贝尔物理学奖授予利用人工神经网络实现机器学习的奠基性发现，化学奖则授予以 AI 为核心工具的蛋白质设计与蛋白质结构预测，标志着 AI4S 取得阶段性突破，正在推动科学研究的范式转变。在应用广度上，AlphaFold 数据库已覆盖约 2 亿个蛋白质结构，被全球 190 多个国家的数百万研究者使用，成为生命科学领域的基础设施级工具；在开放程度上，以 Boltz-1、Evo 2 为代表的开源模型全面公开权重，使得前沿 AI4S 能力加速扩散，进一步夯实了其作为科研基础设施的属性。



闭环越完整，越接近科研基础设施；闭环断裂，AI 易停留在“看似合理的建议”。

图 21 科学智能发现闭环

## (二) 细分领域进展

### 1. 生命科学从结构理解走向功能设计

生命科学是 AI4S 进展最成熟、同时也最容易被高估的方向。相关能力已经完成“结构理解—相互作用预测—新功能生成”的关键跃迁。AlphaFold 将蛋白结构预测转化为基础工具，AlphaFold 3 进一步支持蛋白与 DNA、药物等复合物预测；ESM-3 生成了与天然 GFP 序列同一性仅约 36%、且能在实验中发光的新荧光蛋白，显示 AI 正在从复现已知结构走向设计未知功能。基因组方向的 Evo 2 也开始把问题从单个蛋白推向序列与功能关系。需要明确的是，这些能力仍主要停留在计算预测层。模型提供的是可检验假设，能否进入诊断与治疗流程，仍取决于湿实验、临床证据和监管审评。

其产业价值主要体现在压低研发前端搜索与筛选成本，帮助研究者更快排除低价值方向，而不是直接给出生命科学问题的最终答案。

## 2. 科研智能体从辅助推理进入实验闭环

科研智能体的价值不在于替代科学家独立完成研究，而在于承担科研流程编排功能。它可以将文献检索、假设生成、验证设计、工具调用、实验执行和方案修正等环节连接起来，减少人工在重复性组织工作上的消耗。斯坦福 Virtual Lab 提供了较具代表性的案例，多角色智能体协作设计并实验验证了 92 个靶向新冠变异株的纳米抗体，说明其已具备贯通“假设—设计—可验证结果”链条的能力。但科研智能体距离真正自主仍有明显差距。同类 AI Scientist 虽能完成论文级闭环，第三方评测显示约 42% 的实验因代码错误失败。现阶段，科研智能体更接近高阶科研协作者，可作为科研工作流的中间层，上接文献与模型，下接实验与验证；物理实验和关键判断仍需要人类参与，可靠性是其进入生产流程的核心门槛。

## 3. AI 制药临床验证强化，后期试验仍是关键

AI 制药是市场预期分化最明显的方向，应以临床证据和平台复用能力作为主要判断依据。Insilico 的 Rentosertib (生成式 AI 参与设计的 TNIK 抑制剂) 在特发性肺纤维化 IIa 期出现初步疗效，用药组 12 周用力肺活量较基线改善约 98 毫升，安慰剂组下降约 20 毫升 (Nature Medicine 2025)，提供了 AI 设计候选进入人体验证的实证。与此同时，

BenevolentAI 的 BEN-2293 在特应性皮炎 IIa 期未胜过安慰剂，公司随后退市，说明靶点发现与分子生成仍难以跨越临床生物学的不确定性。

综合来看，AI 最有可能压缩的是靶点判断、分子生成和候选筛选等研发前端环节；人体临床的样本规模、周期、终点设置和审评要求仍难以显著压缩。因此，评判 AI 药企更应关注临床推进、失败复盘和平台复用，而非融资额、管线数量或客户清单。

#### **4.材料科学领域重构研发前端，量产验证仍需推进**

AI 对于材料科学领域的核心贡献在于重构研发前端，将经验试错转化为计算筛选与实验验证相结合的流程。微软与 PNNL 的固态电解质项目从约 3260 万个候选中筛选出 18 个，并在约 9 个月内做出电池原型，新材料用锂量约减少 70%，较完整地走通了“AI 筛选—实验合成—原型验证”流程；GNoME 将候选晶体搜索空间扩展到百万量级，说明 AI 能够显著放大材料发现的候选空间。需要注意的是，相关成果大多仍处于实验室原型阶段，能否进入量产取决于合成工艺、成本控制、稳定性和工程放大。对中国而言，应用案例集中在高分子材料、特钢、碳捕捉、锂电电解液等方向，价值主要体现在缩短候选到验证的循环、减少实验轮次，与中国制造业和工程体系优势具有较高适配度。

#### **5.气象科学领域具备最清晰的市场化路径**

气象科学领域是 AI4S 市场化最清晰的方向。ECMWF 在

2025 年将 AIFS 投入业务运行，并与传统物理预报系统并行，表明 AI 预报已开始进入公共服务流程。AI 并不取代物理模型，而是以更高速度、更低算力成本和概率化输出补强传统系统。该方向率先成熟，主要得益于三项条件，包括长期积累的高质量再分析数据、清晰公开的评估基准，以及可与现有业务系统并行试运行的场景。这些条件正是医疗、材料等领域仍需补齐的基础。中国在该方向上也具备一定竞争力，盘古、伏羲、风乌等模型已在中期和极端天气预报中形成影响，深海、遥感、灾害预警也成为国内 AI4S 中较少数可评测、可上线的场景。其边界在于，极端事件、物理一致性和长期气候模拟仍需传统模式校验。

## 6. 自动化实验室与工程仿真形成验证闭环

自动化实验室与工程仿真决定 AI4S 能否形成从计算预测到物理验证的闭环。自动化实验室的关键作用，是将模型建议放回真实实验台，并将结果回流至模型系统。只有形成这一闭环，科学智能才具备持续迭代能力。MIT 的 CRES<sub>t</sub> 提供了代表性案例，约 3 个月完成 900 多种催化剂配方和 3500 多次电化学测试，筛出一种单位成本功率密度约为纯钯 9 倍的八元合金。其价值不在单次结果，而在于证明主动学习与机器人实验能够持续产出可验证数据。工程仿真则是另一条已被验证的工程化路径，AlphaChip 用强化学习开展芯片布局并进入多代 TPU 设计，AIIA 案例中的动车组气动、客机翼型等也表明其已是国内较成熟的应用场景。此类应用必须先界定工况、精度和安全责任，仿真加速不能替代关键工况下

的实测。

### （三）瓶颈与趋势

科学智能下一阶段竞争的核心，不在于单一模型规模，而在于能否将“数据—模型—仿真—实验—验证”连接为可复用、可审计、可持续迭代的闭环。技术上看，它可以拆分为科学数据、科学模型、模拟与实验、验证反馈、平台与标准五层。决定成败的关键不在于某一层能力单独领先，而在于闭环是否完整。模型提出候选，仿真筛选方案，实验验证少数高价值候选，结果再回流，才构成真正的科研生产系统。



图 22 科学智能五层基础设施

当前制约闭环形成的因素主要有三类。第一是数据，高质量科学数据并非天然开放，失败实验数据沉淀不足；第二是验证，湿实验、临床和工程试制周期难以显著压缩，越接近现实决策越需要独立验证和责任边界；第三是人才，AI4S是算法、领域科学、实验平台与工程体系的交叉工程，任何一端不足都会影响落地。对中国而言，与其在最前沿的通用

科学模型上全面正面追赶，不如优先布局能够形成闭环的方向，包括气象、材料、自动化实验室和重大科研基础设施，通过工程能力和场景优势逐步积累原创能力与国际影响力。

表 20 AI4S 对科研流程的重构

| 科研环节    | 传统模式       | AI4S 模式            | 价值           | 风险控制         |
|---------|------------|--------------------|--------------|--------------|
| 文献与知识整理 | 人工检索、阅读和综述 | 跨学科文献挖掘、证据链整理、问题推荐 | 缩短进入新领域的准备时间 | 防止引用幻觉和过期文献  |
| 假设生成    | 依赖经验和灵感    | 多模型生成候选假设并排序       | 扩大搜索空间       | 保留专家筛选       |
| 候选筛选    | 逐个实验或仿真试错  | 模型预筛、代理模型和物理仿真组合   | 减少无效实验和计算开销  | 明确适用工况和误差边界  |
| 实验设计    | 人工设置条件和参数  | 在约束条件下生成实验组合并调用设备  | 提高实验效率       | 安全、伦理和可复现性审查 |
| 结果验证    | 人工统计与复核    | 实验结果自动记录、异常检测、数据回流 | 形成可追踪闭环      | 审计日志和独立复现    |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

## 第六章 未来展望：前沿突破与无限可能

### 一、世界模型

#### (一) 世界模型的概念

世界模型（World Model）是指能够理解环境、预测未来、规划行动的生成式 AI 模型，其终极目标是让智能体具备像人类一样“想象”行动后果并提前规划的能力，是实现通用人工智能（AGI）的关键路径之一。其思想最早可追溯至 1943 年英国心理学家 Kenneth Craik 的理论：人在对现实作出反应之前，会先在大脑中构建一个“小规模的世界模型”，用它来模拟可能发生的过程，再据此选择行动。2018 年，David Ha、Jürgen Schmidhuber 共同发表了论文《World Models》，正式提出“世界模型”这一名称，并给出了“世界模型=观察世界（V）+预测世界（M）+在内部世界中学习行动（C）”的简易理解框架，自此世界模型成为 AI 研究领域的重要方向。

表 21 世界模型与大语言模型的区别

| 维度   | 大语言模型              | 世界模型                        |
|------|--------------------|-----------------------------|
| 预测目标 | 下一个词或 token        | 下一秒世界状态（下一帧画面、下一步动作）        |
| 训练数据 | 以文本为主的静态内容         | 视频、传感器、动作结果等动态时序数据          |
| 输出结果 | 语言或图像内容            | 未来状态预测、行为模拟、可执行行动方案         |
| 学习方式 | 通过语言间接理解世界，是“知识容器” | 通过交互和推演直接理解世界，能“看见”“预测”“干预” |

| 维度   | 大语言模型       | 世界模型               |
|------|-------------|--------------------|
| 应用场景 | 对话、写作、翻译、问答 | 机器人、自动驾驶、物理模拟、决策系统 |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

综合学术界的定义与产业界的实践，研究者们普遍认为世界模型应具备三大核心特质：其一，表示世界。模型能够理解所处的环境里有什么、物体在哪里，以及物与物之间是什么关系。其二，预测未来。它能够对事件进行模拟和生成——如果我推一下杯子、打开一扇门、往前走两步，世界会发生什么样的改变。其三，在世界里规划和行动。当能预测接下来会发生什么之后，模型应能据此采取行动，实现从认知到决策的闭环。正如 Meta 产品设计负责人 Yiqi Zhao 所概括的：“世界模型是把世界抽象到一个潜在的、被压缩过的空间里，在这个潜在空间里，你能够通过学到的物理规律，去做对未来的预测，形成一个对真实世界的模拟器。相当于它是一个模拟系统，有点像是个缩小的平行宇宙。”

## （二）世界模型的技术路线

当前世界模型主流技术路线可划分为表征派、生成派。

表征派主张世界模型是一个“认知引擎”，代表架构是 Yann LeCun 提出的 JEPA（联合嵌入预测架构），核心理念可概括为“预测抽象表征，而非具体像素”。它主要由编码器与预测器构成，其根本优势是大幅提升计算效率——省去纹理、光照等像素级细节的算力，聚焦于结构规律。Meta 的 V-JEPA 2 在视频预测中以少量标注样本达到高水平并展现强泛化能

力。在机器人领域，Dreamer 系列将 JEPA 与强化学习结合，在抽象空间中“想象训练”：智能体用少量真实交互学习世界模型表征，再在想象空间里进行数百万次试错，最后将策略迁移回真实环境，极大提高样本效率，开辟了高效学习路径。

生成派主张世界模型是“模拟引擎”，能够生成可供人类或智能体直接交互的显性虚拟世界。内部分为三条技术路线：**其一，视频生成路线。**以 Sora 和 Genie 为代表，目标是让模型直接生成连续、可控的视频世界。其技术本质是学习视频帧之间的时序转换规律，使模型能够根据初始帧或文本提示，逐帧预测世界的后续演化。Sora 生成的视频展现出令人惊讶的物理一致性——物体运动符合惯性、光影变化遵循几何规律、遮挡关系保持稳定。这证明大规模视频模型已经内化了部分物理直觉。Genie 系列进一步实现用户交互实时生成视频帧，且长时保持场景一致性，意味着模型内部已构建可持续追踪的状态表征。**其二，3D 生成路线。**以李飞飞创办的 World Labs 为代表，主张世界模型应生成显式的 3D 几何结构。World Labs 发布的 Marble 模型能够从单张图片或文本描述中重建 3D 场景，具有“可编辑性”和“可计算性”优势：用户可任意修改物体位置、光照等并实时更新，且天然支持物理模拟。这使得 3D 生成路线成为游戏开发、影视制作、工业仿真等领域的理想选择。近期研究开始探索将视频生成与 3D 生成深度融合，这种混合路线被认为是通往“可交互世界模型”的重要方向。**其三，智能体训练路线。**

该路线将世界模型视为智能体的“训练场”，让 AI 在模拟环境中大规模试错，最终学会可在真实世界迁移的能力。Google SIMA 大量 3D 游戏环境中训练 AI 模型，新版 SIMA 2 用 Genie 3 生成无限场景，已展现零样本迁移到新游戏的“通用行动能力”。在机器人领域，世界模型用于解决仿真到现实迁移难题，通过持续学习校正偏差，Dreamer 系列模型在抽象空间中训练机器人，使其在仅需少量真实交互的情况下，学会复杂的操作技能，如叠放积木、开门、倒水等。

当前各技术路线的产业落地呈现出差异化格局。视频生成路线因结果直观、数据易得、落地场景明确，成为商业化进展最快的方向——Runway、Pika、快手可灵、腾讯混元等产品已进入影视制作、广告营销、社交娱乐等市场。3D 生成路线技术难度更高、成本更大，但因其对几何结构的精确控制能力，在游戏开发、工业设计、数字孪生等专业领域具备不可替代的价值，World Labs 的最新融资达 10 亿美元，估值 50 亿美元，反映了资本市场对这一路线的长期看好。智能体训练路线目前仍以实验室研究和企业战略布局为主，SIMA 等项目的目标是探索通用智能体的可能性，距离成熟商业产品尚有距离。表征派路线则因其“不可见性”在产业落地上面临挑战，但其高效率、强泛化的特性使其成为机器人、自动驾驶等嵌入式智能领域的理想选择。

### （三）世界模型的发展趋势

**产业竞争格局正加速形成。**美国以 Google DeepMind、NVIDIA、Meta 为代表，在视频生成、物理仿真、3D 交互环

境、机器人训练平台等领域全面发力；英国 Wayve 则聚焦自动驾驶世界模型，推出 GAIA-2 生成式 AI 世界模型，用于多摄像头视角、高时空一致性的复杂驾驶环境生产；中国厂商加速追赶，商汤绝影开悟世界模型、华为盘古世界模型、蔚来 NWM 世界模型等已在智能驾驶、仿真测试和具身智能平台方向形成代表性案例，腾讯混元、快手可灵、昆仑万维等在视频生成方向取得突破，字节跳动、阿里巴巴等在大模型与世界模型融合领域持续投入。开源生态同步成长，Meta 的 V-JEPA 2、DeepSeek 系列模型等大幅降低了研究门槛。未来三至五年将是产业格局定型的关键窗口期。

**多模态融合正重塑 AI 认知架构。**大语言模型与世界模型的融合已成为产业共识，二者形成互补的能力拼图：大语言模型提供符号化的语言理解和常识推理，世界模型则赋予系统对物理时空的直觉把握。融合架构使 AI 能够同时处理“是什么”和“会怎样”两类问题，例如，自动驾驶车辆不仅能识别交通标志，还能预判行人突然横穿时的运动轨迹；服务机器人既能理解“倒杯水”的指令，又能在头脑中预演抓取杯子、避开障碍物、控制倾倒角度的完整物理过程。

**从专用到通用的跃迁正在发生。**当前世界模型大多面向特定领域设计——自动驾驶有专门的世界模型，机器人有专门的仿真引擎。但通用世界模型正从理想走向现实，这类模型能够在同一框架内处理城市街道、家庭客厅、工厂车间等截然不同的环境，并具备跨场景迁移能力。2025 年以来，Google DeepMind 的 Genie 3 和李飞飞 World Labs 的

Marble 已展现出这一趋势的雏形。展望未来，通用世界模型有望成为智能系统的“操作系统级”基础设施，为各类应用提供统一的“世界理解”能力。

**行业应用场景正在全面拓展。在自动驾驶领域**，世界模型率先进入规模化验证与商业化应用阶段。目前，商汤绝影已与智己汽车共建端到端数据工厂，依托世界模型绝影开悟针对高风险长尾场景进行规模化生成，补齐智驾系统训练与验证所需的关键数据，为智能驾驶的快速量产落地提供核心支撑。**在具身智能领域**，过去要花工程师几十小时调参的任务，机器人在模拟世界里自主练习就能掌握。以 Meta 的 V-JEPA 2 为例，通过自监督视频学习，验证了世界模型在机器人抓取、导航和规划任务中的潜力。无论是家庭服务机器人、仓储机器人、工厂协作机器人，还是专业级的巡检、医疗辅助手臂，世界模型都可能成为它们跨过智能门槛的钥匙。**在内容生成领域**，世界模型让内容创作只需要给一个世界观、一个任务、一个初始状态，模型就能自动“长出一个世界”。华为云盘古世界模型已覆盖图像理解、图像生成、视频生成和互动娱乐场景，并与万兴科技合作推进音视频垂类模型能力。**在游戏行业**，世界模型意味着游戏世界不需要“制作”，而可以自动生成和进化，设计师只需要设定规则、生态、冲突，AI 就能生成出森林、河流、生物群落、文化、经济系统，甚至 NPC 的性格和行为以及社会关系的演化。DeepMind 的 Genie 3 已展示出从文本或图像提示生成实时交互 3D 世界的能力，可用于游戏原型、智能体训练和虚拟环境探索。**在穿**

**戴式设备领域**，世界模型能让设备真正读懂用户的 3D 世界，实时推断空间结构、物体关系和潜在风险，从“信息终端”变成随身的“世界理解引擎”。

## 二、太空算力

人工智能的爆发式增长推动全球算力需求呈指数级攀升，然而传统地面数据中心的扩张已难以匹配这一增速，并对能源、水资源与土地资源构成巨大压力。相比之下，太空环境具有近乎无限的太阳能资源，能源边际成本趋近于零，且太空的真空低温环境是天然的绝佳散热场，无需消耗水资源进行冷却，在绿色低碳方面优势显著。这些特点让太空算力作为全新计算范式应运而生。

太空算力是指在近地轨道卫星等太空空间平台，部署分布式计算网络、太空数据中心或太空边缘计算，目标是实现数据在采集源头（太空）的实时处理、智能分析与自主决策，推动“天感地算”（卫星负责采集和传回原始数据，由地面数据中心处理）向“天数天算”（卫星在轨道上直接对原始数据进行实时处理和智能分析，只下传高价值结果）跃迁。太空算力的在轨实时处理、广域覆盖等优势，可提升数据处理效率、全域覆盖和抗干扰能力。

技术架构层面，太空算力系统包括四大关键模块：**一是算力模块**，需采用抗辐射加固的高性能 CPU、GPU 或专用 AI 加速芯片，保障其在太空平台严苛的体积、重量与功耗约束下稳定工作；**二是能源系统**，主要依赖太阳能电池阵、电

源控制器和储能电池，太阳能电池阵在晨昏轨道等特定轨道可实现近乎 24 小时不间断供电，电源控制器用于调配平台电力，储能电池保障卫星进入背阴面时的持续电力供应；三是**散热系统**，在太空真空环境中，热量主要依靠热辐射散热，通常采用热管/泵驱两相流体回路与大型辐射散热板方案，将芯片废热排向深空，摆脱对水冷系统的依赖；四是**通信链路**，包括用于多星协同、构建分布式网络的星间链路，以及连接天基算力与地面用户的星地链路。其中，星间激光通信的单链路传输速率已达到 100-400Gbps，凭借高带宽、低延迟及强抗干扰特性，有望为在轨协同计算提供高速骨干网支撑。

市场布局层面，中美两国作为太空算力的核心引领者，正沿着截然不同的路径推进。美国受日益严峻的能源消耗和散热瓶颈驱动，由 SpaceX、亚马逊、谷歌、英伟达等科技巨头依托成熟的商业航天生态布局，通过星链 V3、Starcloud 试验卫星等项目加速在轨计算验证与星座部署，追求短期内的技术突破与商业化落地。中国则采取政府牵头、产学研协同的模式，由国家级实验室与顶尖高校（如中科院计算所、武汉大学、之江实验室等）引领，通过“三体计算星座”、“东方慧眼”等重大项目，强调技术自主可控与天地一体化架构，目前正处于体系化快速推进与在轨试验验证的关键阶段。

未来，随着商业航天迈入成熟期，太空算力将极具性价比。研究表明，一旦商业航天的发射成本降至约 200 美元/公斤的阈值，太空算力的大规模商业化将成为可能。随着以

SpaceX 可重复使用火箭为代表的商业航天技术进步，业界普遍预测这一临界点有望在 2030 至 2035 年间实现。根据 StarCloud 的推演，当商业航天迈入成熟期，太空算力的成本将极具竞争力：以单个 40MW 集群为例，其在太空持续运营十年的预估总开支约 820 万美元，而地面运营成本高达 1.67 亿美元（美国），开源证券测算中国市场对应运营成本约 0.97 亿美元，均显著高于太空算力的运营成本。

表 22 中美太空算力布局对比

| 对比维度 | 美国                                          | 中国                                          |
|------|---------------------------------------------|---------------------------------------------|
| 发展模式 | 科技巨头主导的商业化路径                                | 政府、科研院所与地方国资协同的工程化推进模式                      |
| 参与主体 | SpaceX、亚马逊、谷歌、英伟达、Starcloud等                | 中国科学院计算所、武汉大学、之江实验室、国星宇航、北邮、北京星辰未来空间技术研究院等  |
| 代表项目 | 星链V3+星舰、亚马逊Leo星座、Starcloud在轨AI服务器、谷歌“捕日者计划” | 三体计算星座、星算计划（含梁溪星座）、天算星座、东方慧眼星座、晨昏轨道GW太空数据中心 |
| 战略目标 | 突破地面算力能源瓶颈，将AI算力大规模迁移至太空，构建太空商业算力标准         | 建设天地一体化算力网络，实现太空计算自主可控，保障国家战略支撑能力           |
| 核心优势 | 商业航天运力强、科技生态成熟、商业化节奏快                       | 全产业链整合能力强、多计划并行推进、注重技术自主与长期战略布局             |
| 主要挑战 | 商用芯片太空环境可靠性仍需验证，星间组网与散热等技术有待突破              | 商业航天运力与美国存在差距，大规模星座组网部署进度受限于发射能力            |

资料来源：深圳市人工智能行业协会、深圳市投控数字科技有限公司整理

### 三、人形机器人

人形机器人被认为是具身智能的终极形态。人形机器人的最大优势在于与人类环境的天然适配性，凭借高度仿生的体型与关节构型，它们能够直接使用现有的工具无缝接入工位及空间设施，从而免去了对传统产线进行高昂改造的成本。当前主流的技术路线是“大脑-小脑”协同架构：“大脑”相当于机器人的智慧中枢，利用多模态大模型、世界模型、VLA模型（视觉—语言—动作模型）等算法处理智能传感器收集的信息，不仅能听懂人话，还能自主规划并拆解复杂任务；“小脑”则是运动控制中心，负责精准调度全身关节，保证动作的流畅与安全。这两个系统通过高速的数据通道实时联动，让机器人真正实现了“知行合一”。

**2025 年全球人形机器人市场实现爆发式增长。**2025 年全球人形机器人出货量超 1.3 万台，较 2024 年增长近 5 倍，市场规模达 6 亿美元。国内人形机器人整机企业已超 140 家，发布产品达 330 多款，2025 年装机量超 1.2 万台，占全球人形机器人装机量的 80%，智元、宇树、优必选等头部企业均实现了千台级甚至五千台级的交付量；订单金额亦创新高，优必选 2025 年末订单金额接近 14 亿元，沃克系列累计订单金额已突破 8 亿元，标志着行业规模化落地进程加速。

**2026 年人形机器人正迎来从“量产元年”到“商业化元年”的关键转折。**根据工信部在国新办新闻发布会上的明确

表态，2026 年将成为政策与市场协同发力、实现突破的关键一年。人形机器人已从技术验证阶段加速向真实场景落地，迈入规模化应用的关键阶段。工信部强调，将重点推动人形机器人在工业制造、仓储物流等场景的规模化应用，力争实现从“能用”到“好用”的跨越。预计 2026 年中国人形机器人产能将突破 10 万台，全球人形机器人市场规模有望从 2025 年的 6 亿美元跃升至 2030 年的 60 亿美元。

### **技术突破、成本下降是人形机器人商业化的核心驱动力。**

当前，人形机器人的运动控制精度已达到工业级标准，能满足超高精度操作的需求。同时，续航能力得到了全面升级，以特斯拉 Optimus Gen3 为代表的新一代产品能实现 8-10 小时的连续工作能力。在智能化层面，感知与决策能力正经历革命性飞跃。端到端 VLA 模型的普及，使机器人具备了更强的环境理解、任务规划和自主决策能力，从简单的程序执行者向能适应复杂场景的“具身智能体”演进。更为关键的是，核心零部件国产化取得重大突破，驱控一体关节、减速器等关键部件国产化率已达 70%，带动整机成本显著下降。目前，工业级人形机器人、消费级人形机器人售价全面下探，宇树科技的 R1 AIR 以 2.99 万元的超低售价打破了行业对人形机器人的价格认知，特斯拉也明确将 Optimus 的目标售价定在 2 万至 3 万美元，为大规模推广应用奠定了基础。

### **市场渗透呈现出工业场景先行、消费场景培育的格局。**

工业制造、仓储物流、汽车制造是当前核心落地场景，在一汽红旗、蔚来等汽车产线，乐聚人形机器人承担料箱搬运、

SMT 料盘分拣、质检等实操任务；在比亚迪新能源汽车车间，Walker S2 机器人单日可完成 200 台车辆装配，效率相当于 3 名熟练工人；在十五运会、春节联欢晚会等国家级舞台，人形机器人也发挥了文娱表演、火炬传递、颁奖仪式服务等作用。人形机器人正从简单的执行者升级为具备自主决策能力的智能体，不仅能完成高强度、高精度的工业作业，还能通过灵巧手实现拧瓶盖、叠衣服等精细动作。

**行业的成熟与规范发展，离不开标准与认证体系的快速建立。**中国在标准体系建设上全面提速，制定全球首个《人形机器人伦理基本要求》国家标准，构建覆盖全产业链的五级标准体系。与此同时，全球形成了中国 CR、欧盟 CE、美国 FCC 三大认证体系并存的格局，确保产品的质量与安全，中国已有产品率先集齐了这三项国际主流认证。

#### **四、量子计算**

量子计算凭借理论上指数级的计算加速潜力，成为全球科技竞赛的焦点之一。量子计算的优势源于其与经典计算截然不同的信息处理范式。经典计算机使用非 0 即 1 的比特进行运算，而量子计算机的基本单元量子比特，可以处于 0 和 1 的叠加态。当多个量子比特相互纠缠时，其形成的叠加态数量将随比特数呈指数增长，使得量子计算机能够并行处理海量可能性。这一特性使其具备经典计算机无法比拟的运算速度，可显著缩短模型训练时间、降低计算成本，对大语言模型、深度神经网络等计算密集型任务意义重大。当前量子计算产业规模增长迅速，已形成以中美为主导的竞争格局。根

据麦肯锡测算，2025 年全球量子计算市场规模约为 10 亿美元，预计到 2028 年将快速增长至 32 亿至 44 亿美元，年复合增长率高达 47%。

**量子计算整体仍处于从“原理验证”迈向“实用探索”的发展阶段。**在技术路线上，超导路线是目前工程化最成熟的路径，其优势在于扩展性较好且与现有集成电路工艺有一定兼容性，国际商业机器公司、谷歌以及中国的本源量子等机构在此方向领先，但其系统需运行在接近绝对零度的极低温环境下，运维成本高昂且纠错效率较低。光量子路线以中国科学技术大学的“九章”系列为代表，能在室温下运行，适宜构建分布式量子计算网络，且可与现有光纤通信基础设施兼容，无需进行信号转换便能实现扩展。中性原子与离子阱路线则在量子比特的精度、相干时间或室温操作灵活性上各有优势，是全球研发和投资的重点方向，正致力于解决规模扩展与高精密操控等工程难题。

**量子计算与人工智能的“双向赋能”，是当前产业演进中最具战略价值的方向之一。**人工智能可作为量子计算的工程辅助工具，用于量子芯片校准、噪声建模、控制参数优化和量子纠错解码，帮助提升量子系统的稳定性和可操作性。例如，谷歌 Quantum AI 在超导量子芯片研发中持续推进自动化校准、误差抑制和量子纠错研究。随着量子硬件规模和容错能力提升，量子计算有望在复杂采样、组合优化、分子模拟、材料数据生成等子任务中为 AI 提供新的算力工具，尤其适用于传统计算成本较高的科学计算和优化场景。总体来

看，量子计算更可能作为专用加速能力，与 GPU、超算和云计算共同构成面向未来 AI 发展的混合算力基础设施。

展望未来，预计量子计算将在 2035 年左右进入通用容错阶段（指能够执行任意量子算法，且在硬件存在噪声和错误的情况下仍能可靠运行的量子计算系统）。届时，量子计算产业将实现飞跃式增长，不仅将重构计算技术本身，更将为金融服务、材料科学、药物研发等特定领域提供加速引擎，并对人工智能的未来发展产生深远影响。

## 第七章 风险思辨：伦理、安全与可持续

随着人工智能技术在全球范围内的深入应用，其在赋能经济社会发展的同时，也引发了一系列复杂的伦理、安全与可持续性挑战。本章将从资源消耗与可持续性、数字鸿沟与发展失衡、劳动转型与替代冲击、数据让渡与安全风险四个维度，探讨人工智能发展中的关键风险及治理路径，为构建安全、可信、包容、可持续的人工智能治理体系提供参考。

### 一、资源消耗与可持续性

人工智能既是优化资源配置、提升能源效率的赋能者，其自身指数级增长的算力需求也构成了对能源、水资源乃至关键矿物的巨大消耗，引发了能源供需与环境承载的双重压力，同时也催生了绿色转型的创新路径。

#### （一）资源消耗风险

人工智能是能源密集型产业，算力发展对能源供应存在高度依赖。算力基础设施的核心部件是芯片，其研发制造、计算运行需要庞大且稳定的电力资源支撑。以英伟达 H100 GPU 为例，单个 GPU 峰值功耗约 700 瓦，若按全年 61% 的使用时间计算，全年耗电量约 3740 度，相当于一个美国家庭的平均年用电量。一座大型数据中心的 GPU 机柜数量可达数千个，年耗电量可能在数亿度，相当于数万户家庭一年的用电量。国际能源署预测，至 2030 年，全球数据中心的年耗电量预计将达到 945 太瓦时，约占全球用电量的 3%。

然而，人工智能用电需求与现有电力供应间存在多方面的矛盾。首先，电力基础设施的规划建设周期长，难以匹配算力需求的爆发式增长。在弗吉尼亚、阿姆斯特丹、伦敦等地区，已因电网负荷放缓新项目审批。其次，数据中心需 7×24 小时高可靠供电，而风能、太阳能等具有间歇性、波动性的特征，难以独立保障稳定电力输出。此外，电力资源与算力需求存在地理错配：东部地区算力需求旺盛但电力紧张，西部可再生能源丰富但需求较少。尽管“东数西算”工程旨在优化布局，但受网络时延等因素制约，低时延业务仍需布局于东部，导致电力供需在区域上存在矛盾。

被长期忽视的水资源压力正成为新的瓶颈。数据中心服务器运行时产生大量热量，传统冷却方式（如冷却塔、蒸发冷却）依赖水的蒸发吸热来散热，耗水量巨大。超大规模数据中心年耗水量可达数亿升，相当于数万人的日常用水量。然而，全球水资源分布不均衡，许多算力枢纽地区（如美国西部、中东）本身水资源有限，这引发了数据中心与当地居民、农业之间日益紧张的水资源争夺。水资源可用性正演变为决定数据中心选址、运营许可的环境考量。

除此以外，人工智能行业对稀土、锂、钴、铜等矿物资源需求，正在引发多重影响：一方面，这些矿物的开采与提炼本身即是高耗能、高水耗、高污染的过程，随着易开采的高品位矿石逐渐枯竭，开采转向低品位或偏远地区，导致成本上升、环境压力增大。另一方面，关键矿物的开采和加工高度集中于少数国家，这种地理集中性导致供应链易受地缘

政治冲突、政策变动或社会动荡的影响，加剧了资源供应的不确定性。

## （二）可持续发展路径

政策层面，我国发布《数据中心绿色低碳发展专项行动计划》，要求到 2025 年底，新建大型数据中心 PUE 需降至 1.25 以内，国家枢纽节点需低于 1.2。在绿色电力方面，我国明确要求国家枢纽节点新建数据中心的绿电占比超过 80%，鼓励数据中心通过参与绿电绿证交易等方式提高可再生能源利用率。

在技术革新方面，行业已突破传统风冷模式，形成以液冷为核心、多技术并存的格局。其中，冷板式液冷因技术成熟度高，正成为新建智算中心处理高功耗 GPU 的主流方案；浸没式液冷、余热回收技术也已从技术验证走向实际运营。与此同时，业界正通过改进模型架构与算法从根本上提升计算能效，例如通过混合专家模型架构、模型压缩技术减少计算负载等。

在能源结构方面，寻求稳定的低碳基荷能源已成为全球性战略议题。在这一背景下，核能因其接近零排放且能提供不间断电力的特性，获得了新的战略关注。包括微软、谷歌在内的全球科技巨头，已将投资下一代核能技术纳入其长期能源战略，为其日益扩大的云计算和人工智能业务提供稳定、不间断的清洁电力。

人工智能正反向赋能能源治理。一方面，人工智能技术

通过优化工业流程、交通系统与建筑能效，赋能传统产业节能减排；另一方面，人工智能技术应用于能源领域，正在提升风电与光伏发电的预测精度、加速新型电池或碳捕集材料的研发、以及优化电网的实时调度与稳定性。这些应用展现出人工智能创造环境效益的潜力，其带来的碳减排量，有望在未来超过人工智能基础设施自身的运营碳足迹。

## 二、数字鸿沟与发展失衡

人工智能的全球扩张，加剧了长期存在的数字鸿沟，并催生出更具结构性的发展失衡。研究显示，2025 年全球北方地区的 AI 应用率增速接近全球南方地区的两倍，增速最快的十个国家无一例外是高收入经济体。与此同时，全球仍有近 26 亿人未接入互联网，其中绝大多数集中在最不发达国家和内陆发展中国家。这种数字基础设施的悬殊直接限制了后者参与人工智能生态的可能性。另外，算力资源高度集中于美国、中国与欧洲，部分发展中国家缺乏自主研发 AI 能力，不仅难以训练适配本土需求的大模型，甚至在使用开源模型时也因电力不稳、网络延迟和硬件匮乏而举步维艰。

同时，全球主流开源大模型训练数据严重依赖英语语料，如 OpenAI 约 90% 的训练数据由英文文本构成，导致非英语母语群体在使用生成式 AI 时面临理解偏差、文化错位。这种语言与文化维度的边缘化，使数字鸿沟从技术问题转化为认知与话语权的剥夺。此外，人才储备的断层亦不容忽视：全球顶尖 AI 研究人员中，来自低收入国家的比例不足 3%，本土化技术团队的缺失使得技术引进难以转化为可持续的内

生能力。

在中国，数字鸿沟虽因“宽带中国”“东数西算”等国家战略显著缓解，但区域、城乡与代际间的不平衡依然突出。东部沿海地区已进入 AI 融合应用阶段，而中西部部分县域仍面临数字技能薄弱等问题。尤其是老年群体的“智能边缘化”——尽管国务院 2020 年即出台《关于切实解决老年人运用智能技术困难的实施方案》，但 2025 年调研显示，65 岁以上人群中仍有近四成难以独立完成医疗挂号、移动支付等操作。

弥合数字鸿沟是一个系统性工程，需从国际合作、技术共享与资源开放等多个维度协同推进。中国通过“一带一路”数字经济合作倡议，已帮助非洲建成 15 万公里通信骨干网，服务近 7 亿用户终端，并推动轻量化大模型在资源受限地区落地。世界互联网大会倡导的“开放资源”理念亦鼓励，企业适度开放算法、数据与模型资源，为中小企业和研究机构提供可及的技术工具。2025 年以来，Hugging Face 等全球性开源平台上的开源中文模型数量增长 3 倍，为非英语国家提供了更多选择。

### 三、劳动转型与替代冲击

人工智能技术正以前所未有的深度和广度重塑全球劳动力市场，引发关于人类劳动价值与未来工作形态的思考，呈现出替代与创造并存、挑战与机遇交织的复杂态势。一方面，人工智能孵化出新场景与新业态，催生对 AI 芯片与硬件研发、算力运维、数据服务等专业人才的需求；另一方面，

人工智能也带来了岗位替代风险，高盛研究显示，在人工智能冲击下，全球约 3 亿全职岗位面临替代风险，美国约三分之二的职业可通过 AI 实现部分自动化。随着 AI 技术与各行业的深度融合，替代效应不断强化。

岗位替代风险呈现结构性特征。生成式人工智能擅长编程、数据分析、内容生成等技能，使软件开发、文案撰写、平面设计等岗位面临替代压力。研究表明，软件开发岗位因低具身性需求，71% 的专业技能暴露于替代风险；而护理岗位依赖肢体操作与情感支持，仅 32.9% 的技能可能被自动化。安永分析进一步指出，数学与编程技能在生成式 AI 威胁暴露指数中高居首位，而人际互动与情感理解技能则处于安全区。另外，随着智能体、机器人技术的持续突破，制造、物流、服务业中的中低技能劳动力也面临被替代的压力。基于人工智能形成的新岗位多为高技能要求，被替代的中低技能劳动者难以胜任，导致其在承担失业风险的同时，在收入分配中仍处弱势，资本所有者则成为技术红利的主要受益者，加剧贫富分化。

人机协作成为新的工作模式。劳动者正与人工智能工具深度融合，例如，程序员借助人工智能生成代码框架，专注于逻辑设计和问题解决；设计师通过人工智能生成创意灵感，再结合自身审美进行细化；客服人员借助人工智能快速响应常见问题，专注于处理复杂投诉。此外，技术快速迭代使得个人知识技能更新周期迅速缩短，倒逼全社会进入技术转型加速期。从个体层面，劳动者需持续学习人工智能技能，强

化社会情感能力、创造性思维等机器难以替代的能力，主动拥抱人机协同工作模式，灵活调整职业规划；从社会层面，应加强全民智能素养教育，完善社会保障与就业服务，为转岗劳动者提供技能培训、失业保障，如欧盟的数字欧洲计划和新加坡的技能创前程计划，通过税收激励和再培训基金降低转型成本。

#### 四、数据让渡与安全风险

人工智能的持续演进高度依赖高质量、大规模数据的持续输入，数据让渡已成为模型训练、迭代与部署不可或缺的前提。然而，随着生成式 AI 深度融入社会运行，数据让渡过程中的安全风险已从隐私泄露扩展至系统性社会威胁。据国家网信办 2025 年 12 月发布的《2025 年数据安全年度报告》，2025 年全国共发生涉及人工智能系统的数据安全事件 1,842 起，较 2024 年上升 37%，其中因第三方数据合作或 API 接口漏洞导致的供应链攻击占比达 58%，成为最主要的风险来源。

用户在使用 AI 服务时，往往在缺乏充分知情的情况下让渡行为轨迹、生物特征等敏感信息，而这些数据一旦被滥用，可能会引发诈骗、身份盗窃等严重问题，对个人隐私和安全构成威胁。2025 年 12 月公安部通报的典型案例（公通字〔2025〕127 号）显示，犯罪分子利用合成语音冒充企业高管实施“AI 换脸诈骗”，单次诈骗金额超千万元，该案例已纳入《2025 年全国电信网络诈骗案件典型案例汇编》。这直接印证了数据滥用在诈骗场景中的现实危害。

值得警惕的是，大模型训练数据中混入的恶意内容可引发“模型污染”，导致 AI 系统在金融风控、医疗诊断等关键领域输出错误决策。2025 年 11 月，国家网信办联合工信部发布的《网络数据安全风险评估办法（征求意见稿）》首次将“模型训练数据完整性”纳入强制评估范畴，反映出监管层对源头风险的高度警觉。

为应对挑战，我国正构建多维度治理框架。技术层面，2025 年 10 月实施的国家标准《数据安全技术 电子产品信息清除技术要求》明确了设备退役时的数据彻底清除责任；隐私计算技术应用加速落地，截至 2025 年底，全国已有 43 个地市政务平台采用联邦学习架构处理跨部门数据协作。制度层面，2025 年多地出台专项政策，北京、上海推进开源软件安全审查机制，防范供应链投毒。执法层面，2025 年网信部门开展“清源”专项行动，累计下架违规收集个人信息的 AI 类 APP 217 款，处罚未履行数据安全保护义务的企业 89 家，最高罚款达上年度营业额 5%。

尽管治理成效初显，深层次矛盾依然存在。2025 年工信部《工业领域数据安全能力提升实施方案》显示，近六成中小制造企业因无力承担数据分类分级与风险评估成本，被迫放弃 AI 转型。在数据要素市场化改革背景下，2026 年被国家明确为“数据要素价值释放年”，强调以“安全为发展赋能”。未来路径需聚焦三点：一是推动数据“可用不可见”技术规模化应用，降低流通风险；二是建立分级分类的 AI 数据使用授权机制，保障用户知情与控制权；三是通过公共

数据开放与可信数据空间建设，缓解企业对私有数据的过度依赖。唯有在释放数据价值与筑牢安全底线之间实现动态平衡，才能确保人工智能真正成为高质量发展的可靠引擎。

## 关于协会

深圳市人工智能行业协会（简称 SAIIA）经深圳市民政局批准于 2017 年 5 月成立，是全国首家专注人工智能领域的行业协会，被评为深圳市 5A 级社会组织。协会本着“立足深圳，深化 AI，服务中国，辐射全球”的宗旨，开展资源对接、成果推广、产业研究、标准制定、项目评审等服务。

协会是 AI Alliance Network 创始单位、联合国大学（UNU）全球人工智能网络成员，近年来被认定为首批广东省智能制造生态合作伙伴、深圳市高技能人才培训基地、深圳市人大常委会立法联系点、知识产权保护工作站、公平贸易工作站、海智工作站、产业集群公共服务平台共建单位、深圳行业协会商会高质量 100 等。

协会拥有 1000 余家会员单位，囊括人工智能基础层、技术层和应用层各个领域的优秀企业；协会设立专家委员会、专业委员会，吸引国内外 300 余位人工智能专家和 16000 位各领域专家加入。

协会设立研究部，日常开展市场调查、数据统计、项目评审、标准制定、政策献言等工作，输出行业研究报告、可行性研究报告、产业园区规划等成果。目前，协会已完成 100 多份研究成果。

协会主办的深圳国际人工智能展（简称 GAIE），是全国

首个覆盖人工智能全生态链的专业博览会，自 2020 年创办以来已成功举办七届，累计集聚 1600+ 参展企业，展示科技成果近万件，举办高层次论坛 190 多场，吸引专业观众超 38 万人次，被海内外 500 多家媒体报道，已成为粤港澳大湾区人工智能第一展，是人工智能领域辐射全国、接轨国际的重要展示交易平台。

## 关于投控数科

深圳市投控数字科技有限公司成立于 2010 年 1 月，注册地位于宝安区新桥街道创新智慧港，是全市较早从事信息化、数字化、智能化技术服务的国有企业，是深圳市投资控股有限公司服务数字经济发展的重要抓手平台。近年来，公司抢抓数字经济发展机遇，携手政府及企业客户，探索数据要素价值化和人工智能技术创新应用，共创美好数字生活。

公司按照深投控赋予的定位、使命和国家、省、市的政策要求，扎实在数据技术服务、人工智能技术服务、企业数字化转型技术服务、信创技术服务等领域创新、积累，形成核心竞争力。

数据技术服务方面，数据生产、整合、加工、产品化、产品应用及运营能力基本建成，知识库、标签系统等产品初步完成研发，在多个客户、场景进行测试性推广。

人工智能技术服务方面，以“大模型+小模型+田间地头创新”模式初步建成的人工智能技术服务平台和运营能力，在多个客户和场景中应用，多次荣获奖项。

企业数字化转型技术服务方面，OA、财务核算、考核、合同管理、数据资产建设、云网安全服务、新能源消费平台等服务和产品均已在深投控和其它国企广泛推广。

信创技术服务方面，基于鸿蒙的软件开发、数据服务能

力初步形成，公司新产品、新应用全栈技术基于鲲鹏创新改造，基于鲲鹏云提供技术服务已成为公司的标配能力。

## 法律说明

### 一、版权声明

本白皮书由深圳市人工智能行业协会制作，其版权归属深圳市人工智能行业协会。未经深圳市人工智能行业协会许可，任何组织和个人不得以任何形式擅自发布、出版、改编本白皮书内容。转载、引用或利用其它方式使用本白皮书内容的，应注明“来源：深圳市人工智能行业协会”。任何未经授权使用本白皮书的相关商业行为都将违反《中华人民共和国著作权法》和其它法律法规以及有关国际公约的规定。违反上述版权声明者，本协会保留追究其相关法律责任的权利。

### 二、免责声明

本白皮书为深圳市人工智能行业协会研究员采用信息研究法、大数据捕捉法、专家会议法、调查研究法等方法研究所得。本白皮书的部分内容参考自公开资料，部分数据基于协会行业数据库通过统计预测模型估算获得。

受研究方法和信息资源的限制，本白皮书只提供给用户作为参考资料。深圳市人工智能行业协会对本白皮书内容的准确性、及时性、完整性和可靠性尽最大努力的追求，但不作任何保证。本白皮书所载的任何意见、推测仅反映我会于本白皮书发布日期当天的判断，本协会可以在不发出通知的

情况下做出更改。在任何情况下，本白皮书中的信息或所表述的观点均不构成任何建议，深圳市人工智能行业协会及相关人员对于阅读或使用本白皮书而作出的决定或采取的行动不承担任何责任。



# 第八届 THE 8th SHENZHEN GLOBAL ARTIFICIAL INTELLIGENCE EXPO 深圳国际人工智能展览会

2027年5月13-15日 深圳会展中心（福田）

2027年5月13-15日，第八届深圳国际人工智能展览会（GAIE2027）将在5万平方米的展馆内迎来800多家参展企业。作为亚洲AI圈不容错过的年度盛会，这里不仅有最新技术的集中亮相，更是一次产业应用的集中演练。

GAIE2027将重点展示“AI+千行百业”的融合成果，让深圳成为全球AI资源对接的核心枢纽。与其说这是一个展会，不如说它是一个巨大的实战现场——每一场发布都在探索商业落地的可能，每一次交流都在解决技术变现的难题，GAIE正推动AI行业从单纯的技术秀，转向真正的价值变现。

## 人工智能全生态链专业展览会

50000m<sup>2</sup>

展览面积

2026年携手 500+ 家参展企业共襄盛举  
分别来自 15 个国家/地区

2000+

合作项目

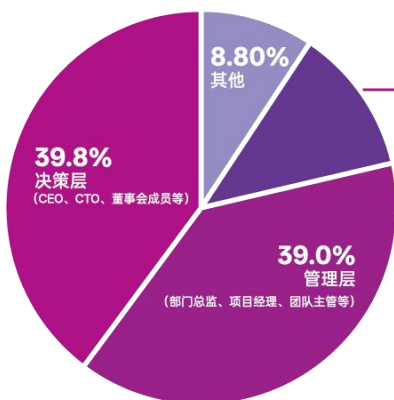
累计发布 6 份行业白皮书  
推动超 800+ 个技术成果转化落地

80+

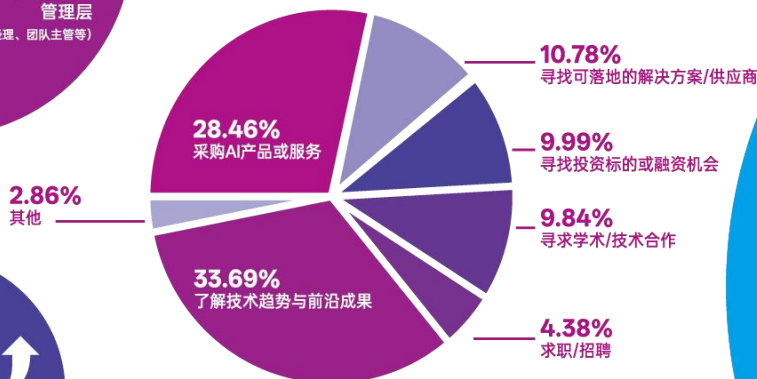
配套活动

预计吸引超过 20 个国家及地区  
300+ 位全球顶尖专家深度对话

### 观众职位层级



### 观众参观目的



47.6%↑

观众人数显著攀升  
同比增长47.6%

2026年展会观众人数 93827  
100+ 专业观众团体

每10位观众中，就有近6位正在积极  
寻求采购、投资或技术合作伙伴。



## 算力基座展区

展示AI芯片（GPU、NPU、ASIC、存算一体芯片）、智算服务器、液冷整机柜、边缘计算设备、云计算与智算中心解决方案、算力调度平台、传感器、激光雷达、减速器、伺服电机、灵巧手、机器视觉模组、绿色数据中心技术等。

## 大模型与AIGC应用展区

展示通用大模型、行业垂类大模型、AI AGENT智能体平台、AIGC工具（文生文、文生图、文生视频、AI编程、AI办公）、数字人、虚拟主播、AI客服、智能体行业应用等。

## AI+行业深度融合展区

展示智能制造（工业质检、预测性维护、数字孪生工厂）、智慧医疗（AI影像、辅助诊断、手术机器人）、低空经济（无人机、EVTOL、空域管理系统）、自动驾驶（ROBOTAXI、无人配送车、车路协同）、智慧金融（智能风控、智能投顾、反欺诈）、智慧政务、智慧城市、智慧教育解决方案等。

## AI终端与智能硬件展区

展示AI手机、AI PC、AI平板、AI眼镜、智能手表、可穿戴设备、AI耳机、AI玩具、AI陪伴设备、智能家居、智能家电、全屋智能方案、智能座舱、车载终端、端侧大模型模组、AIOT开发板等。

## 具身智能与机器人展区

展示人形机器人、轮式/足式机器人、工业协作机器人、AGV/AMR物流机器人、服务机器人（清洁、配送、导览、康养陪护）、特种机器人（巡检、消防、水下、排爆）、机器视觉与运动控制系统等。

## 未来创新与国际生态展区

展示高校及科研院所前沿成果（脑机接口、类脑计算、量子AI）、初创企业创新产品、海外展团国际成果、AI安全与数据合规方案、AI伦理治理、投融资机构路演对接等。

120000+

专业观众

2026年来自 81 个国家及地区，  
超9万名观众到场

1亿+

媒体曝光量

深度合作 350+ 媒体与 500+ 记者  
2000+ 篇稿件覆盖海内外媒体矩阵

看底座：全球顶尖算力生态汇聚

看大脑：百模千用，智能体全面爆发

看身体：具身智能，机器人走进生活

看落地：垂直行业数字化转型全案呈现

抢占黄金展位，  
定义AI新纪元！





# 深圳市人工智能行业协会

Shenzhen Artificial Intelligence Industry Association

[ 5A级社会组织 ]

人工智能行业一站式服务平台

立足深圳、深化AI、服务中国、辐射全球

## 协会简介 ASSOCIATION INTRODUCTION

深圳市人工智能行业协会成立于2017年5月，会员单位涵盖人工智能全生态链各领域优质企业等1000余家，目前已有中科院、清华大学、北京大学、新加坡南洋理工大学、上海交通大学、哈尔滨工业大学、深圳大学、中山大学等400多位专家学者及17000位各领域专家。协会被评为5A级社会组织、联合国大学（UNU）全球人工智能网络成员、“AI ALLIANCE NETWORK”创始单位、首批广东省智能制造生态合作伙伴、深圳市高技能人才培训基地“数字经济”专项基地、深圳市人大常委会立法联系点、知识产权保护工作站、公平贸易工作站、海智工作站、产业集群公共服务平台共建单位、深圳行业协会商会高质量100、人工智能营商环境监测站等多个荣誉资质。



## 研究成果 RESEARCH ACHIEVEMENT



## 协会活动 ACTIVITY



深圳国际人工智能展览会



人工智能生态大会



人工智能创业者大会



2026年APEC中小企业工商论坛  
人工智能赋能业太中小企业产业发展论坛

## 联系我们

电话: 0755-82711965/23991694  
邮箱: cindy\_lang@saiia.org.cn

地址: 深圳市福田区梅康路新一代产业园3栋15楼1503  
网址: <https://www.saiia.org.cn>



协会公众号

电话：0755-82547884

电话: 0755-83218979

