

deepseek 大模型生态报告

赛迪研究院电子信息研究所

2025年2月

目 录

| | |
|--------------------|----|
| 一、DeepSeek 大模型背景 | 1 |
| (一) 深度求索公司情况 | 1 |
| (二) 模型迭代历程 | 2 |
| (三) 研发团队情况 | 3 |
| 1. 北京大学罗福莉 | 4 |
| 2. 北京大学高华佐 | 4 |
| 3. 北京邮电大学曾旺丁 | 5 |
| 4. 北京大学朱琪豪 | 5 |
| 5. 北京大学代达劭 | 6 |
| 6. 北京大学王炳宣 | 6 |
| 7. 浙江大学梁文锋 | 7 |
| 二、DeepSeek 大模型应用现状 | 8 |
| (一) 国内情况 | 8 |
| (二) 全球情况 | 10 |
| 三、DeepSeek 大模型技术特点 | 11 |
| (一) 技术创新 | 11 |
| 1. 压缩：从结构到量化 | 11 |
| (1) 多层注意力 MLA | 12 |
| (2) FP8 混合精度训练框架 | 13 |
| 2. 并行：对硬件的极限使用 | 16 |
| (1) DualPipe 跨节点通信 | 16 |
| (2) 无辅助损失的负载均衡策略 | 18 |
| (3) 底层通信优化 | 20 |
| 3. 总结 | 20 |
| (二) 训练成本 | 21 |
| 四、DeepSeek 大模型产业生态 | 22 |
| (一) 企业 | 22 |

| | |
|---|-----------|
| 1. DeepSeek 系列新模型正式上线昇腾社区，华为小艺已接入 DeepSeek..... | 22 |
| 2. 三大运营商全面接入 DeepSeek..... | 23 |
| 3. 多家公司宣布接入 DeepSeek..... | 23 |
| 4. 微软、英伟达、亚马逊全部接入 DeepSeek! | 25 |
| 5. 燧原科技实现全国各地智算中心 DeepSeek 的全量推理服务部署 | 26 |
| 6. 多家券商也积极“拥抱” DeepSeek..... | 26 |
| (二) 生态 | 27 |
| 1. Deepseek 应用 27 日登顶苹果中国地区和美国地区应用商店免费 APP 下载排行榜..... | 27 |
| 2. DeepSeek 在开源社区 GitHub 上的 Star 数首次超越 OpenAI | 28 |
| 3. 国家超算互联网正式上线 DeepSeek..... | 29 |
| 4. 传 DeepSeek 正准备适配中国国产 GPU... .. | 29 |
| 5. DeepSeek 的风，火速吹遍了终端企业..... | 30 |
| (三) 应用 | 32 |
| 1. 文学 | 32 |
| 2. 编程 | 32 |
| (四) 基础设施 | 35 |
| 五、DeepSeek 大模型产业影响 | 35 |
| (一) 对人工智能技术的影响 | 35 |
| 1. 吴恩达谈 DeepSeek: 中国 AI 迎头赶上，扩大算力并不是 AI 进步的唯一途径..... | 35 |
| 2. 蒸馏技术将成为行业突破口 | 36 |
| (二) 对产业生态的影响 | 37 |
| 1. DeepSeek 开源对于企业战略价值..... | 37 |
| 2. DeepSeek 是否真的绕开了 CUDA? | 39 |
| 3. DeepSeek 会导致对智算中心的需求大幅减少吗? | 39 |

| | |
|---|-----------|
| (三) 对国际竞争的影响 | 46 |
| 1. DeepSeek 的崛起, 冲击了全球科技供应链的 核心玩家 | 46 |
| 2. AI 低成本化, 最大输家可能不是 Nvidia, 而 是依赖大模型收费的 AI 公司 | 46 |
| 3. AI 进入创业黄金时代 | 47 |
| 4. DeepSeek 的崛起引发了全球资本市场的剧 烈震荡 | 47 |
| 六、国内相关表态 | 48 |
| (一) 国家层面 | 48 |
| 1. 中国常驻联合国代表傅聪回答记者提问 .. | 48 |
| 2. 外交部发言人郭嘉昆主持例行记者会 | 49 |
| (二) 地方层面 | 49 |
| 1. 广东 | 49 |
| 2. 郑州 | 49 |
| 3. 北京经济技术开发区 | 49 |
| 七、各国相关表态 | 50 |
| (一) 美国 | 50 |
| 1. 龙头企业单位 | 50 |
| (1) 英伟达 | 50 |
| (2) OpenAI | 51 |
| (3) 谷歌 | 52 |
| (4) 苹果 | 52 |
| (5) 微软 | 52 |
| 2. 美国联邦政府 | 52 |
| (1) 美议员推动立法促中美 AI 全面脱钩 | 53 |
| (2) 美国海军禁止官兵使用 DeepSeek 大 模型 | 53 |
| 3. 美国地方政府 | 54 |
| (1) 得克萨斯州宣布禁止政府设备使用 DeepSeek | 54 |

| | |
|---|----|
| (二) 英国及欧洲 | 54 |
| 1. 龙头企业单位 | 54 |
| (1) 英国阿兰·图灵研究所 | 54 |
| (2) 德国墨卡托中国研究中心 | 54 |
| (3) 英国利兹大学 | 55 |
| (4) 荷兰阿斯麦 | 55 |
| 2. 政府及公共机构 | 55 |
| (1) 意大利 | 55 |
| (2) 爱尔兰 | 55 |
| (3) 法国 | 55 |
| (三) 韩国 | 56 |
| 1. 龙头企业单位 | 56 |
| (1) 韩国 | 56 |
| (2) LG | 56 |
| (3) 三星 | 56 |
| 2. 政府及公共机构 | 56 |
| (1) 韩国代理总统崔相穆将 DeepSeek 称之为“新的冲击” | 57 |
| (2) 韩国祥明大学 | 57 |
| (四) 其他国家 | 57 |
| 1. 政府及公共机构 | 57 |
| (1) 澳大利亚 | 58 |
| (2) 日本 | 58 |
| (3) 印度 | 58 |
| 八、DeepSeek 大模型的经验启示 | 58 |
| (一) 技术 | 58 |
| 1. 中科院为何难现 DeepSeek 式突破 | 58 |
| (1) 制度基因的先天性分野 | 58 |
| (2) 创新效率的量子级差距 | 60 |
| (3) 体制突围的艰难探索 | 62 |
| (4) 破局之道：构建科技创新共同体 | 63 |
| (5) 超越二元对立的创新哲学 | 64 |

| | |
|-----------------------------|----|
| 2. 王凡、刘少山：中国如何引领全球具身智能？ | |
| DeepSeek 的战略启示 AI 观察..... | 64 |
| (1) DeepSeek 的突破对具身智能的启示 | |
| | 65 |
| (2) DeepSeek: 推动中国 AGI 领域自主 | |
| 创新 | 66 |
| (3) 具身智能：人工智能从虚拟世界走向 | |
| 物理世界 | 67 |
| (4) 上下文强化学习：具身智能的发展方 | |
| 向..... | 68 |
| (5) 具身智能的应用前景与行业变革. | 70 |
| (6) 政策建议：开源发展、标准化和产业 | |
| 化落地 | 71 |
| (7) 结论：把握技术机遇，领导未来的全 | |
| 球 AI 竞争 | 72 |
| (8) 本文作者 | 72 |
| 3. 专家观点 鲁传颖：人工智能重塑国家安全的 | |
| 范式和逻辑 | 72 |
| (1) 人工智能赋能国家安全 | 73 |
| (2) 人工智能技术安全对国家安全的深 | |
| 层次影响 | 77 |
| (3) 人工智能时代的国家安全治理体系 | |
| | 82 |
| 4. 巴黎 AI 峰会，李飞飞教授开幕演讲：当前 | |
| 才是“首个真正的 AI 时代” | 86 |
| (1) 人工智能的历史与起源 | 87 |
| (2) 现代 AI 三大关键要素 | 88 |
| (3) 2012 年之后的 AI 突破..... | 91 |
| (4) 大模型与近期 AI 进展..... | 92 |
| (5) 空间与具身智能的兴起 | 94 |
| (6) 以人为本的 AI 与三大核心价值... .. | 94 |
| (7) AI 治理：理性、务实与多方协作 | 96 |

| | |
|---|-----|
| 5. 人工智能驱动信息技术体系重构与再造——兼评《中国至 2050 年信息科技发展路线图》 | 98 |
| (1) 《路线图》和《三元融合》的重要预判与近 15 年发展的比较 | 99 |
| (2) 未来 10 年信息技术体系重构与再造的创新机遇与挑战 | 106 |
| (3) 发展对策 | 121 |
| (二) 企业 | 123 |
| 1. DeepSeek 的创新三重门 | 123 |
| (1) 小天才们的胜利 | 123 |
| (2) 华为式的军团平推 | 125 |
| (3) 原创（哲学性）思想 | 129 |
| (4) DeepSeek 跨越第三重门了吗？ | 130 |
| 2. DeepSeek 颠覆了什么？——大模型“国产之光”破局的启示 | 131 |
| (1) 击穿三大定式 | 132 |
| (2) 实现三大跃升 | 134 |
| (3) 超越技术的启示 | 136 |
| (三) 地方 | 138 |
| 1. 为什么 DeepSeek 没有出现在浦东张江 | 138 |

一、DeepSeek 大模型背景

（一）深度求索公司情况

DeepSeek，全称“杭州深度求索人工智能基础技术研究有限公司”，简称“深度求索”，于2023年7月17日由幻方量化创立，位于杭州市拱墅区，是一家创新型科技公司，专注于开发先进的大语言模型（LLM）和相关技术，法定代表人、执行董事兼总经理为裴湑。

2024年1月5日，该公司发布第一个大模型 DeepSeek LLM。之后，相继发布 DeepSeek-Coder、DeepSeek-VL、DeepSeek V2、DeepSeek V3 等模型。其中，DeepSeek V2 模型因在中文综合能力评测中的出色表现，且以极低的推理成本引发行业关注，被称为“AI 界的拼多多”。2025年1月20日，该公司正式发布 DeepSeek-R1 模型，并同步开源模型权重。DeepSeek-R1 性能逼近 OpenAI o1 正式版，推理成本却仅为 OpenAI o1 的几十分之一。1月27日，DeepSeek 应用登顶苹果美国地区应用商店免费 App 下载排行榜，在美区下载榜上超越了 ChatGPT。同日，苹果中国区应用商店免费榜显示，DeepSeek 成为中国区第一。1月28日，DeepSeek 发布开源多模态模型 Janus-Pro，其中 70 亿参数版本的 Janus-Pro-7B 模型在使用文本提示的图像生成排行榜中优于 OpenAI 的 DALL-E 3 和 Stability AI 的 Stable Diffusion。

2025 年 2 月 8 日，QuestMobile 数据显示，DeepSeek 在 1 月 28 日的日活跃用户数首次超越豆包，随后在 2 月 1 日突破 3000 万大关，成为截至 2025 年 2 月史上最快达成这一里程碑的应用。

（二）模型迭代历程

中国初创企业深度求索（DeepSeek）研发大语言模型时据报绕过了美国人工智能巨头英伟达的 CUDA 框架，为 DeepSeek 在未来适配中国国产芯片做好准备。

DeepSeek 通过优化算法架构，显著提升了算力利用效率，打破了算力至上的传统认知。DeepSeek-R1 的惊艳之处是通过重新设计训练流程、以少量 SFT 数据+ 多轮强化学习的办法，在提高了模型准确性的同时，也显著降低了内存占用和计算开销。DeepSeek-R1 的效果示意：更少的 GPU（或其他 AI 芯片）带来同样的效果高 DeepSeek-R1 提供了一种低成本训练的方法，而不是说只能通过低成本来进行训练。DeepSeek-R1 反而通过优化，DeepSeek-R1 可能实现了算力与性能的近似线性关系。每增加一张 GPU，模型推理能力可稳定提升，无需依赖复杂的外部监督机制。这一特性直接验证了“算力即性能”的 Scaling Law，为 AI 芯片市场及 AI Infra 市场提供了明确的增量需求逻辑。

DS 不仅在推理模型上实现了突破，在多模态方面也保持了进步，并于近日开源发布了 Janus-Pro 多模态模型。基于

算法工程方面的优化，DS 能够实现性价比更高的模型推理能力（价格低、效率高、性能强、可部署在端侧平台），并有望以强大的推理能力和多模态感知能力，重塑机器人交互与决策，同时通过低成本、高效率的解决方案，加速具身智能的普及与应用。

（三）研发团队情况

DeepSeek 团队最大的特点之一就是年轻。团队成员中，应届生和在读生占据了相当大的比例，他们活跃在公司的各个项目和研究领域中。这些年轻人思维敏捷、充满活力，对新技术和新方法有着敏锐的洞察力和强烈的探索欲望，为 DeepSeek 的发展注入了源源不断的动力和创新活力。在 DeepSeek 团队中，来自清华大学和北京大学的应届生尤为突出。他们不仅在学术上有着扎实的功底和丰富的研究成果，而且在实际工作中也展现出了卓越的能力和潜力。许多关键的技术创新和项目成果，都离不开这些清北应届生的贡献。他们将所学知识与实际问题相结合，不断突破技术瓶颈，推动 DeepSeek 在 AI 领域的持续进步。有些团队成员从 DeepSeekLLMv1 开始就加入了公司，并一直参与到了 DeepSeek-v3 的开发和研究中。他们见证了 DeepSeek 大模型从无到有、从弱到强的全过程，积累了丰富的经验和深刻的理解。这些全程参与的成员，对 DeepSeek 的技术发展和产品

迭代起到了至关重要的作用，他们的经验和见解为公司未来的发展提供了宝贵的参考和指导。

1. 北京大学罗福莉

罗福莉，女，95后，出生于四川，本科就读于北京师范大学计算机专业，硕士毕业于北京大学计算语言学研究所计算语言学专业。罗福莉2019年曾在国际顶级会议ACL上发表8篇论文，其中2篇为第一作者。罗福莉的职业生涯始于阿里巴巴达摩院，罗福莉主导开发了多语言预训练模型VECO，并推动了AliceMind的开源工作。2022年，罗福莉转战DeepSeek，参与了MoE大模型DeepSeek-V2的研发。2024年12月，网传罗福莉被小米创始人雷军以千万年薪招募，担任小米AI实验室的大模型团队负责人。12月31日，罗福莉的高中班主任告诉红星新闻记者，罗福莉还没决定去小米，还在考虑。

2. 北京大学高华佐

高华佐是MLA架构的关键创新者之一，他非常低调，目前只知道是北大物理系毕业。在DeepSeek团队中，高华佐凭借其扎实的物理基础和对AI技术的深刻理解，为MLA架构的提出和发展做出了重要贡献。他的创新思维和严谨态度，为DeepSeek的技术进步提供了有力的支持。另外，在“大模型创业六小强”之一阶跃星辰的专利信息中也可以看到高华

佐的名字，虽然暂不确定是否是同一人，但这也从一个侧面反映出他在 AI 领域的影响力和活跃度。

3. 北京邮电大学曾旺丁

曾旺丁来自北邮，研究生导师是北邮人工智能与网络搜索教研中心主任张洪刚。在 DeepSeek 团队中，曾旺丁主导了 GRPO 算法的研究和开发工作。他深入分析了强化学习中的关键问题和难点，提出了 GRPO 这一创新的算法，为 DeepSeek-Math 项目和整个强化学习领域的发展做出了突出贡献。曾旺丁的研究成果不仅在学术界得到了认可，也在实际应用中展现出了强大的潜力和价值。

4. 清华大学邵智宏

邵智宏是清华交互式人工智能 (CoAI) 课题组的博士生，师从黄民烈教授。他的研究领域包括自然语言处理、深度学习等，特别对如何构建一个稳健且可扩展的 AI 系统感兴趣。在 DeepSeek 团队中，邵智宏参与了 DeepSeekMath、DeepSeek-Prover、DeepSeek-Coder-v2、DeepSeek-R1 等多个项目，为公司在自然语言处理和 AI 系统构建方面的发展做出了重要贡献。他之前还曾在微软研究院工作过，积累了丰富的研究和实践经验。

4. 北京大学朱琪豪

在北京大学攻读博士期间，朱琪豪校友师从熊英飞副教授和张路教授，研究方向为深度代码学习技术，致力于探索

程序语言定义与深度学习技术的深度融合。他的科研成果在 ICSE、IJCAI、AAAI、ESEC/FSE、ASE 等国际顶级会议上频频亮相，共发表 CCF-A 类论文 16 篇。凭借基于语法制导的程序修复技术，他荣获 ESEC/FSE 2021 杰出论文提名奖，并多次斩获 ACM SIGSOFT 杰出论文奖。在博士阶段，朱琪豪校友不仅在学术上收获颇丰，还将研究成果成功应用于产业。他主导开发了国产开源代码大模型 DeepSeek-Coder-V1，达到国际领先水平，相关技术广泛应用于反编译、代码分析、代码修复等领域，受到全球学者和企业的高度关注。由此开发的修复工具 Recoder 在深度学习修复研究中首次超越传统方法，推动整个行业迈向深度学习的新时代。朱琪豪校友以其卓越表现，连续获得字节跳动奖学金、北京大学校长奖学金、国家奖学金及北京大学三好学生标兵等重量级荣誉。他的博士论文《语言定义感知的深度代码学习技术及应用》更是入选 2024 年 CCF 软件工程专业委员会博士学位论文激励计划，成为该领域具有里程碑意义的研究成果。

5. 北京大学代达劼

除了 DeepSeek-V2 MLA 和 DeepSeekMath GRPO 这两项关键成果，值得一提的是，还有一些成员从 v1 就加入其中，一直到 v3。代表人物之一是代达劼，2024 年博士毕业于北京大学计算机学院计算语言所，导师是穗志方教授。

6. 北京大学王炳宣

北大元培学院的王炳宣，来自山东烟台，2017年进入北大。硕士毕业后加入 DeepSeek，参与了从 DeepSeek LLM v1 开始的一系列重要工作。

7. 浙江大学梁文锋

梁文锋，1985年出生于广东湛江吴川，父母都是小学语文老师，这也使得他从小就成绩优异。“湛江发布”提及他一段成长往事：梁文锋性格很文静，但不是书呆子，在学习上很有自己的“一套方法”，初中时就学完了高中数学，甚至开始学大学数学。

2002年，梁文锋以吴川一中“高考状元”的成绩考上浙江大学，本科就读电子信息工程专业，之后又继续攻读浙江大学信息与通信工程专业研究生。大学期间，梁文锋就与同学一起开始积累市场行情数据和探索全自动量化交易。2015年，梁文锋硕士毕业后和几位浙大同学共同创立了量化对冲基金公司幻方量化，并于当年拿到第一张私募牌照，几年时间管理规模迅速超百亿。有意思的是，2016年幻方量化便首次上线 AI 策略，并在之后实现投资策略全面 AI 化，同时也是国内唯一公开宣称拥有万张英伟达 A100 显卡的企业。大学时期梁文锋便坚信“AI 一定会改变世界”，这些都为后来投身 AI 埋下伏笔。转折点是 2023 年，全球 AI 大模型风起云涌，国内大佬纷纷宣布 AI 创业，梁文锋也是其中一员。那年 4 月，梁文锋宣布正式进军通用人工智能领域，“追寻一直

以来的技术理想，超越投资去直面更大的课题”。很快，DeepSeek 在杭州应运而生同年 11 月，DeepSeek 发布了开源的代码大模型。

二、DeepSeek 大模型应用现状

（一）国内情况

随着 DeepSeek 的爆火，国产大模型云服务平台 SiliconCloud（硅基流动）开始被大家关注。

2 月 1 日，硅基流动和华为云宣布联合首发并上线基于华为云昇腾云服务的 DeepSeekR1/V3 推理服务。基于自研推理加速引擎加持，硅基流动和华为云昇腾云服务支持部署的 DeepSeek 模型可获得持平全球高端 GPU 部署模型的效果。

DeepSeek 对中国最大的启示，还是“开放”二字。无论是心态上还是方法上，都要毫不动摇地坚持开放，继续开放。DeepSeek 的胜利，就是开源打败闭源。OpenAI 以前是开源的，后来变成闭源的，这次 DeepSeek 爆火之后，OpenAI 的奥特曼也出来说话，承认闭源是一种战略错误。

开源不仅可以用我们举国的力量，还可以用全世界的力量。世界的科学共同体要联合起来，遏制资本、遏制某些政客的力量，促进人类的技术进步。中国过去很多技术成果和产业成功都是在开放条件下取得的，今后也必须如此。长三角、珠三角还有很多有潜力成为独角兽的企业，很多人没有资源，但是有技术，需要来自国家的资源。所以，科技要取

得长足进步，金融改革也要跟上。我们不要虚拟金融、泡沫金融，要遏制投机性的，但对新技术的投资，对真正创新的风投，我们一定要做好。

在吴恩达看来，本周围绕 DeepSeek 的热议，让许多人清晰地看到了几个一直存在的重要趋势：美国在 GenAI 领域的领先地位正在被中国迎头赶上，AI 供应链格局将被重塑；开放权重模型正在推动基础模型层商品化，为应用开发者带来新机遇；扩大规模并非通往 AI 进步的唯一途径。尽管算力备受追捧，但算法创新正在快速降低训练成本。

成立至今，DeepSeek 并未开放过外部融资，因此也并未引起创投圈的过多关注。直至 2024 年 5 月，DeepSeek 发布 DeepSeek-V2 开源 MoE 模型，因极度降低推理成本，意外掀起一场大模型“价格战”，开始出圈。隐秘低调，却聚集了一群 AI 天才。据悉，DeepSeek 包括创始人梁文锋在内，仅有 130 多名工程师和研究人员。不同于其他明星创业公司，DeepSeek 鲜有海归，团队成员几乎都来自清华大学、北京大学、中山大学、北京邮电大学等国内顶尖高校，不少还是在读博士。去年底，雷军开出千万年薪挖角 DeepSeek 研究员罗福莉。这一次，DeepSeek 正式浮出水面。上个月，梁文锋和团队开发的大模型 DeepSeek-V3 正式发布。起初在 AI 发烧友圈内流传，由于表现出色，一时轰动硅谷。2025 年 1 月 20 日，DeepSeek 正式发布 DeepSeek R1，开始席卷全球。

（二）全球情况

到今年 1 月 27 日，其最新发布的 R1 模型就已登顶苹果美国地区应用商店免费 APP 下载排行榜，超越了美国同类产品 ChatGPT。

DeepSeek 的成就至少让美国三种人感到尴尬。一是硅谷的技术精英们，他们耗费巨资拿出的产品，其性能和中国公司用“小米加步枪”的装备做出来的差不多，甚至创新性上还棋差一招。另一类人则是华尔街的金融大佬们，近些年他们通过炒作人工智能概念，把资产泡沫吹得五彩斑斓，但就在 DeepSeek 下载量超越 ChatGPT 之时，美国科技股遭受重创，英伟达股价一度重挫 17%。

当然，最“愤怒”的还是美国的政客们，他们搞技术围堵已经好几年了，但中国的高科技产品一项接一项推出。围堵不但没能压制中国智造，反而加速了中国科技的突破，这昭示所谓的围堵其实已经破产。当 DeepSeek 登顶下载排行榜第二天，就遭遇了来自美国 IP 地址的持续网络攻击，并且在随后几天，攻击量暴增百倍。而一些西方政客以各种理由将 DeepSeek 列入“实体清单”，或要求封锁该软件。同时他们还利用舆论，声称 DeepSeek “蒸馏”了美国技术。

DeepSeek 以更低的算力要求和更低的成本，达到了与美国顶级 AI 大模型相当的效果，展现出了强大的技术竞争力。

它的出现，不仅打破了原有的市场格局，更是被认为对美国在 AI 领域的领导地位构成了实实在在的威胁。

“DeepSeek 可能会开辟一个全新的由中国引领的技术生态系统，这会让全球的开发者因 DeepSeek 开发的低成本、更高效的人工智能产品而受益。”一位 TMT 领域投资人如是说，“这种竞争从某种程度来说也将打破美国科技巨头的市场主导地位。”

三、DeepSeek 大模型技术特点

（一）技术创新

大模型训练降低成本主要靠两招：压缩、并行和提升硬件使用效率。DeepSeek-V3 这次所用的方法基本上就是猛挥这三板斧。

1. 压缩：从结构到量化

压缩很容易理解，就是把大的东西压缩成小的。对于模型训练来讲，压缩之后，运算单元（GPU 和 CPU）需要进行的运算数据量就会减少，运算速率必然会提升。另一个重要影响是，内存占用和缓存会减少，这样训练同样大小的模型所需要的硬件规模也可以大幅减少。而在训练模型的过程中，内存占比最高的就是向量数据。DeepSeek-V3 这次一次用了两种方法去压缩向量数据，一是 MLA 多层注意力架构，另一个就是 FP8 混合精度训练。

(1) 多层注意力 MLA

多层注意力 (Multi-Layer Attention, MLA) 架构设计的核心在于在 Transformer 架构中引入动态层聚合机制。传统 Transformer 中每一层都需要完整的计算和存储, 其中的 Key 和 Value 矩阵往往占用大量内存空间。而 MLA 通过动态合并相邻层的特征来减少计算量。

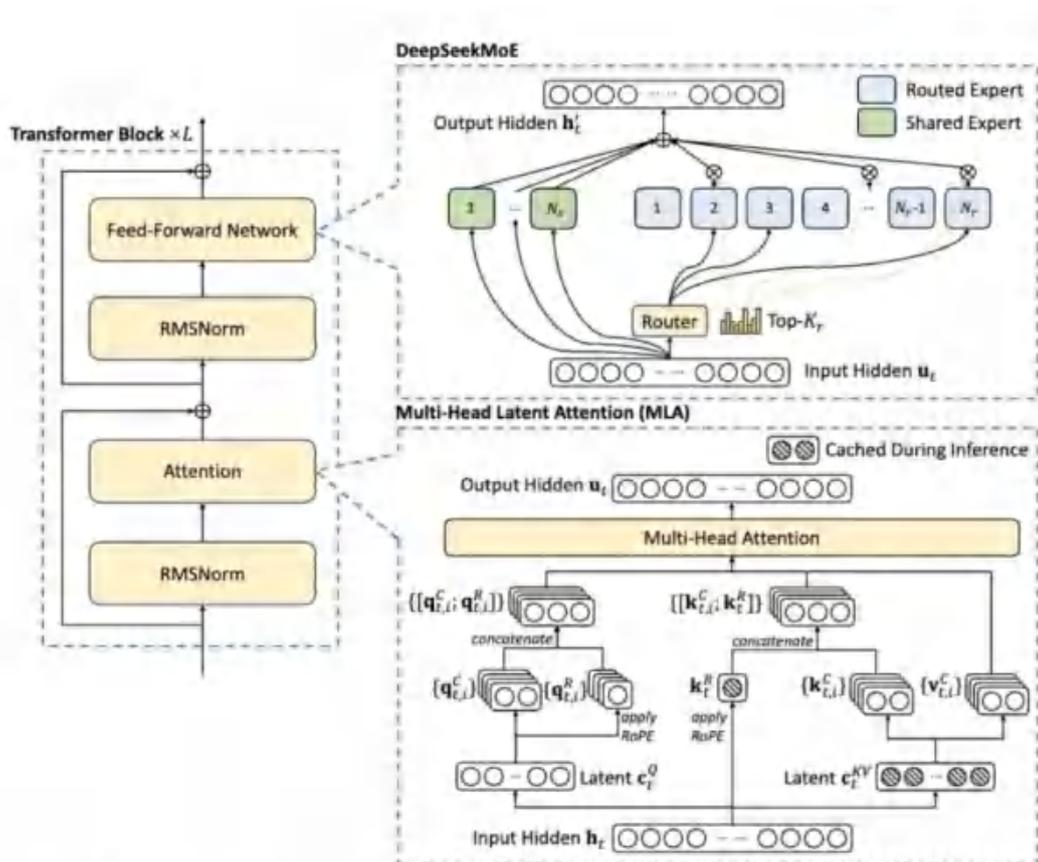


Figure 2 | Illustration of the basic architecture of DeepSeek-V3. Following DeepSeek-V2, we adopt MLA and DeepSeekMoE for efficient inference and economical training.

MLA 通过压缩和复用前序层的 K、V 来减少内存占用和计算量。具体来说, 将连续几层的 K、V 进行合并压缩成一组共享表示。打个比方, 如果把注意力机制比作图书检索系统, 传统方法相当于为每本书都建立完整的索引卡片 (Key)

和内容摘要 (Value), 而 DeepSeek 的方法则像是建立了一个智能的分类系统, 不记具体信息, 而是记一个简单的"标签"(压缩的 Key/Value), 需要时再从标签还原出详细信息。就像把"计算机技术, 三楼右侧第二排"简化成"C2-3"这样的编码。在这个过程中, DeepSeek 使用了低秩压缩技术(可以理解为将高维矩阵压缩为若干个低维矩阵的乘积), 将 KV 压缩到 512 维度, 远小于原始维度。通过 Key/Value 的低秩压缩使得的训练内存占用减少了 20-30%。在 Query 端的优化对训练效率也非常有意义。Query 可以理解为用户的检索请求, 传统方法会为每个请求都分配大量计算资源。DeepSeek 通过对 Query 的低秩压缩, 减少了计算过程中的激活内存占用。虽然这种优化对推理阶段的影响相对较小, 但在训练过程中发挥了重要作用, 显著提升了训练效率。这就像是优化了图书检索系统的查询处理机制, 使得系统能够更快速地处理大量并发的检索请求。而 DeepSeek-V3 巧妙的找到了一些平衡, 让这些压缩技术几乎没有影响模型的性能。

(2) FP8 混合精度训练框架

MLA 方法是从 DeepSeekV2 开始就采用的方法, 本次只是进行了优化调整。而在 DeepSeek-V3 里引入了一种 FP8 混合精度训练框架, 并首次在超大规模模型上验证了其有效性。FP8 就是用 8 个二进制位来表示数字的格式, 相比传统的 32 位 (FP32) 和 16 位 (FP16) 格式, 精度低了很多, 但是占用

空间小，计算快。就像用"约 350 人"代替"准确的 358 人"，牺牲一些精度来换取效率。虽然不够精确，但在很多场景下已经够用了，而且能大大提升运算速度和节省内存。DeepSeek 在采用 FP8 格式时，采用了"混合精度"的方案。在训练时，它的大部分核心计算内核均采用 FP8 精度实现。包括前向传播、激活反向传播和权重反向传播都用了 FP8 作为输入，并输出 BF16 或 FP32 格式的结果。这一设计理论上使计算速度相较于原始的 BF16 方法提升了一倍。此外，DeepSeek 中的向量激活值以 FP8 格式存储，供反向传播使用，从而显著降低了内存消耗。

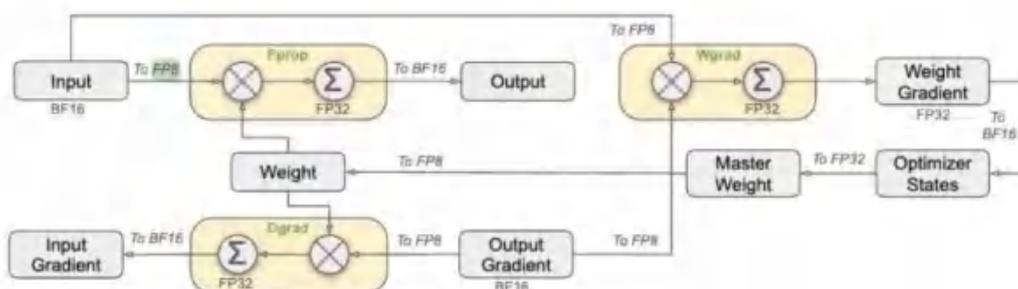


Figure 6 | The overall mixed precision framework with FP8 data format. For clarification, only the Linear operator is illustrated.

针对某些对低精度计算敏感算子和一些低成本算子，比如嵌入模块、输出头、MoE 门控模块、归一化算子以及注意力算子保留了 FP16 乃至 FP32 的精度。这样能保证数据的精确性。同时为了保证数值稳定性，DeepSeek 还将主权重、权重梯度和优化器状态以更高精度存储。就像一个精打细算的

主厨：日常备菜用普通的厨具就够了，但到了关键的烹饪步骤，就会换上最好的刀具。

在模型训练中，大部分的前向运算都使用 FP8 来处理，这样可以大大节省显存和计算资源，让整个训练过程跑得更快。但他们也很清楚哪些地方不能省：比如最后的调味、摆盘（对应嵌入模块、输出头等），就一定要用精密的工具（FP16 或 FP32 精度）。过去使用 FP8 模式的时候，最大的困难出现误差累计。就像普通计算器（Tensor Cores 的 FP8）只能显示到小数点后两位，而科学计算器（CUDA 核心的 FP32）能显示到小数点后六位。当你需要加很多个小数时，用普通计算器会逐渐累积误差，最后结果可能差异很大。

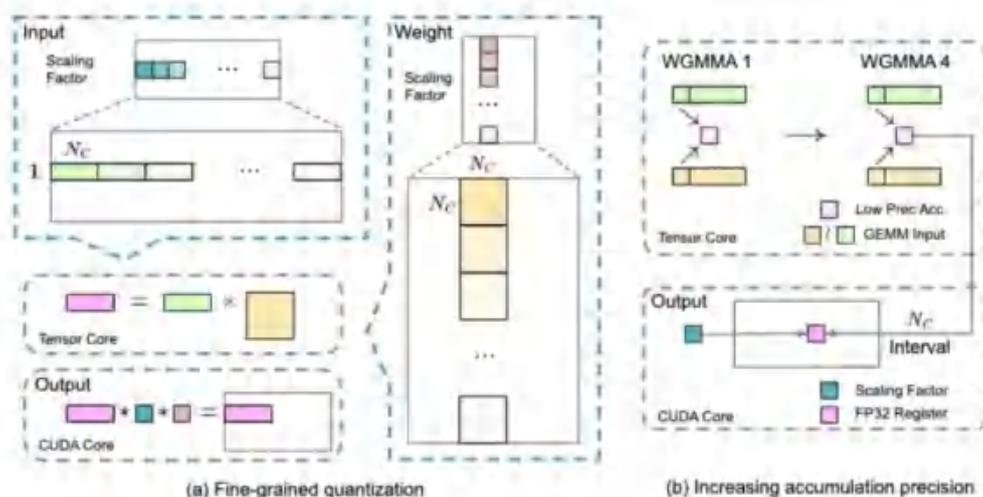


Figure 7 | (a) We propose a fine-grained quantization method to mitigate quantization errors caused by feature outliers; for illustration simplicity, only Fprop is illustrated. (b) In conjunction with our quantization strategy, we improve the FP8 GEMM precision by promoting to CUDA Cores at an interval of $N_c = 128$ elements MMA for the high-precision accumulation.

DeepSeek 发现了一个巧妙的解决方案：不等到最后再算总和，而是每加 128 个数就把当前结果转移到科学计算器上继续计算。为了让这个过程不影响速度，他们利用了

H800GPU 的特点：就像有两个收银员，当一个在结算购物篮的时候，另一个可以继续扫描新商品。这样在提高精度的同时，基本不影响处理速度。这一策略使得模型训练速度大幅提升，毕竟核心计算能提升 100% 的速度，而显存使用减少也非常明显。并且模型最终的效果精度损失能做到小于 0.25%，几乎无损。

2. 并行：对硬件的极限使用

要实现更快的训练速度，最有效的方法就是增加并行计算的规模，让更多的计算单元同时处理不同的数据或任务。而在并行中，需要解决的问题就是尽可能的有效利用计算资源，让它们都高负载的工作。

在系统架构层面，DeepSeek 就使用了专家并行训练技术，通过将不同的专家模块分配到不同的计算设备上同时进行训练，提升了训练过程中的计算效率。

但这种简单的并行还远不够。DeepSeek 这次对算力做的是极限压榨：如果把训练过程当成一个工厂的话，他们主要做的就是不让流水线上没有闲人，再加上尽可能优化工序，让工件（数据）进入流水线时直接就可以被操作（计算）。

（1）DualPipe 跨节点通信

优化流水线流程的主要模式是 DeepSeek 创新的 DualPipe 方法。在计算和通信重叠方面，DualPipe 采用了类似于“多任务并行处理”的思路。

就像现代计算机能够在下载文件的同时处理文档一样，DualPipe 让模型在进行计算的同时，后台已经开始准备下一步需要的数据传输。这种设计确保了通信开销被很大程度地隐藏在计算过程中，极大提升了整体效率。

传统的训练信息流水线并行就像一条产品装配线，每个工位按顺序处理任务。当数据包从一个阶段传递到下一个阶段时，往往会产生等待时间，这就是所谓的"流水线气泡"。这些气泡会导致计算资源的浪费，就像流水线上的工人不得不等待上游工序完成才能开始工作。此外，不同节点之间的数据传输时间也会成为性能瓶颈，就像工位之间传递零件的时间过长会影响整体生产效率。

而 DualPipe 引入了双重流水线的概念，就像在同一条生产线上同时处理两批产品。当一个计算阶段在等待数据传输时，可以立即切换到处理另一批数据，这样就能充分利用原本的空闲时间。

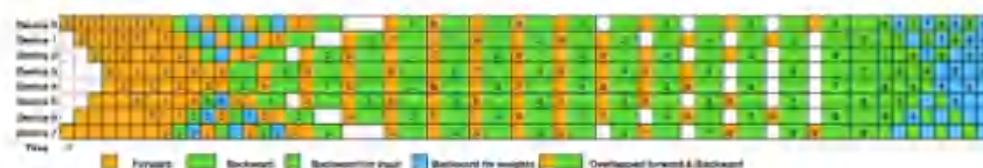


Figure 5 | Example DualPipe scheduling for 8 PP ranks and 20 micro-batches in two directions. The micro-batches in the reverse direction are symmetric to those in the forward direction, so we omit their batch ID for illustration simplicity. Two cells enclosed by a shared black border have mutually overlapped computation and communication.

这是让流水线上没有“闲人”。除此之外，还要尽量压缩取件到操作的过程。因为 DeepSeek 对流水线的特殊设计，使得通信和计算的过程可以重叠。当一个节点在进行当前批次

数据的计算时，系统已经开始准备下一批次需要的专家参数传输。当前向计算完成时，下一步需要的数据已经就位，几乎不会产生等待时间。大部分数据传输时间被"隐藏"在了计算过程中，就像在无缝衔接的装配线上，零件的运送时间对整体生产效率几乎没有影响。

DualPipe 正是通过精确控制这种重叠过程，实现了在大规模分布式训练中接近零通信开销的理想状态。

根据 DeepSeek 的技术报告，DualPipe 算法减少了 50% 的计算气泡，有效隐藏了通信开销。跨节点通信优化则提升了带宽利用率，减少了 20% 的通信开销。这就基本相对传统方式提高了一倍的算力使用效能。

| Method | Bubble | Parameter | Activation |
|-----------------|--------------------------------------|-----------|------------|
| 1F1B | $(PP - 1)(F + B)$ | 1× | PP |
| ZB1P | $(PP - 1)(F + B - 2W)$ | 1× | PP |
| DualPipe (Ours) | $(\frac{PP}{2} - 1)(F + B + B - 3W)$ | 2× | PP + 1 |

Table 2 | Comparison of pipeline bubbles and memory usage across different pipeline parallel methods. F denotes the execution time of a forward chunk, B denotes the execution time of a full backward chunk, W denotes the execution time of a "backward for weights" chunk, and $F\&B$ denotes the execution time of two mutually overlapped forward and backward chunks.

(2) 无辅助损失的负载均衡策略

无辅助损失的负载均衡策略是 DeepSeek-V3 一个让训练过程中工人各展所能的调整。负载均衡策略在 V2 时代已经被引入，但在这一代更进一步。

在专家混合系统(MoE)中，负载均衡一直是个关键挑战。因为 MoE 有很多专家模型，怎么能让该上的专家不闲着，没人无事可做对训练和模型效率都很关键。

传统方法通常需要引入额外的辅助损失项来平衡专家的使用，就像在工厂中人为设置配额来确保各条生产线的负载均衡。这种方法不仅增加了训练的复杂性，还可能影响模型的本地优化目标。

DeepSeek 的创新在于实现了无辅助损失的自然均衡。系统会根据专家的历史利用率动态调整其"接收容量"。当某个专家持续过载时，系统会自动降低其接收新任务的概率；反之，对于利用率低的专家，系统会提高其接收任务的机会。既考虑专业匹配度，也考虑当前的工作负荷。这种自适应机制确保了长期来看的负载平衡。这更像是市场经济，而非计划经济。这个改进让训练过程更稳定，大家都有机会训练，也提高了训练效率。

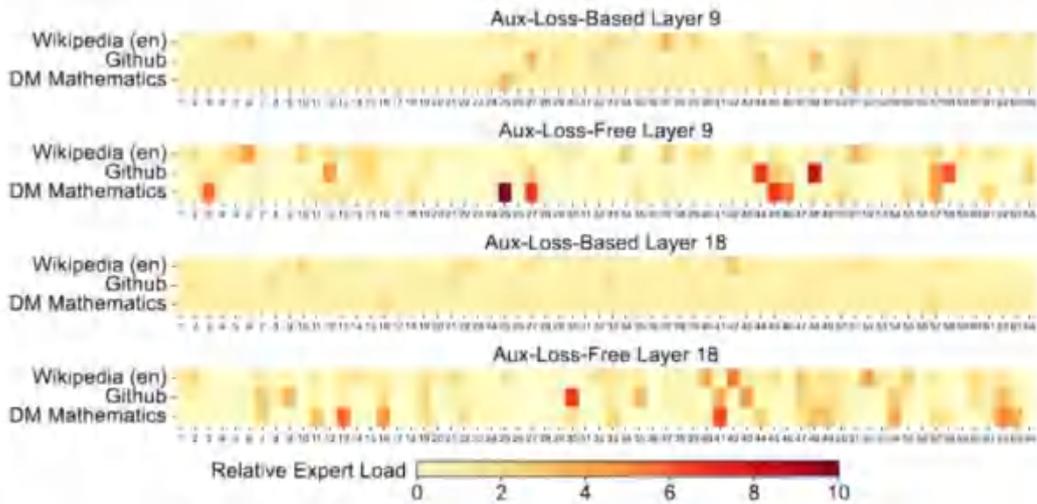


Figure 9 | Expert load of auxiliary-loss-free and auxiliary-loss-based models on three domains in the Pile test set. The auxiliary-loss-free model shows greater expert specialization patterns than the auxiliary-loss-based one. The relative expert load denotes the ratio between the actual expert load and the theoretically balanced expert load. Due to space constraints, we only present the results of two layers as an example, with the results of all layers provided in Appendix [C](#)

(3) 底层通信优化

对于模型训练来讲，底层通讯也是个大问题，很多时候硬件间通讯不畅就会使得训练产线出现局部停工，无活儿可干的事儿。

DeepSeek 在这方面也做了相当的优化，专门开发了高效的跨节点全对全通信内核。这就像是在高速公路系统中建立了更智能的红绿灯调度系统，能够充分利用 InfiniBand 和 NVLink 这些高速通道的带宽。这些优化确保了数据在不同计算节点之间的传输始终保持在最高效率。

3. 总结

以上这些还不是 DeepSeek 在训练上采用的所有提效手段，只是相对大胆创新的部分。目前其他训练在架构中常用的移除 LayerNorm 中的 bias 项、在 FFN 后引入 scale 因子、采用 RoPE 相对位置编码等方式，DeepSeek-V3 也都有所采用。而在训练策略上，DeepSeek 还采用了 ALiBi 位置编码预训练、Flash Attention 2 实现、序列长度动态扩展等已有先进技术。

DeepSeek-V3 这回真的可以说是在训练工程上无所不用其极。总结下来，最重要的包括以下这么几个方面。

| 技术创新 | 训练效率影响 | 推理效率影响 | 模型能力影响 | 部署友好性影响 |
|-----------------|-------------------|----------|-------------|----------|
| MLA架构 | | | | |
| - Key/Value低秩压缩 | 减少训练内存和计算开销 | 降低KV缓存占用 | 保持模型性能 | 降低部署内存需求 |
| - Query低秩压缩 | 降低激活内存占用 | - | 保持模型性能 | - |
| 无辅助损失负载均衡 | 提升训练稳定性；避免序列内平衡开销 | - | 增强专家分化效果 | 简化部署流程 |
| 多token预测 (MTP) | 提升数据效率 | 推理加速1.8倍 | 增强模型预测能力 | - |
| FP8混合精度训练 | 简化训练流程；提升训练稳定性 | - | 精度损失 <0.25% | - |
| DualPipe算法 | 减少50%流水线气泡；优化通信开销 | - | - | - |
| 跨节点通信优化 | 提升带宽利用率 | - | - | 支持大规模部署 |
| 分阶段部署策略 | - | 提升推理吞吐量 | - | 支持灵活部署配置 |

不管是 MLA、FP8 和 Daulpipe 算法，都是非常大胆的使用当下降低训练成本的前沿技术。这些基础技术方向都已经是主流认知的可能性，但 DeepSeek 做到了经过精心调教优化设计，使其可用，且发挥最大的能力。既然 GPU 少，那就卷工程，DeepSeek 这回确实是用东亚魔法打破西方垄断。

（二）训练成本

DeepSeek-V3 的横空出世，用一组的数据完美诠释了没有 GPU Poor，只有卷得不够多。当 o1、Claude、Gemini 和 Llama 3 等模型还在为数亿美元的训练成本苦恼时，DeepSeek-V3 用 557.6 万美元的预算，在 2048 个 H800 GPU

集群上仅花费 3.7 天/万亿 tokens 的训练时间，就达到了足以与它们比肩的性能。这意味着每万亿 tokens 仅需 180K 个 H800 GPU 小时，总计 278 万 GPU 小时的训练成本。而 Llama 3.1 的训练使用了 16,384 块 Nvidia H100 GPU，总计 2100 多万 GPU 小时，翻了十倍。通过 671B 的总参数量，在每个 token 激活 37B 参数的精准控制下，DeepSeek-V3 用 14.8 万亿高质量多样化 token，构建出了一个能够超越所有开源模型，直逼 GPT-4 和 Claude-3.5 的 AI 巨人。

四、DeepSeek 大模型产业生态

（一）企业

1. DeepSeek 系列新模型正式上线昇腾社区，华为小艺已接入 DeepSeek

2025 年 2 月 4 日，DeepSeek-R1、DeepSeek-V3、DeepSeek-V2、Janus-Pro 正式上线昇腾社区，支持一键获取 DeepSeek 系列模型，支持昇腾硬件平台上开箱即用，推理快速部署，带来更快、更高效、更便捷的 AI 开发和应用体验，欢迎广大开发者体验使用。

华为小艺已接入 DeepSeek。2 月 5 日，华为宣布，基于原生鸿蒙操作系统的小艺助手 App 已经接入 DeepSeek。目前，纯血鸿蒙小艺智能体已经支持 DeepSeek-R1，已升级至原生鸿蒙的华为用户可以通过“小艺助手 App-发现-智能体广场”与 DeepSeek 进行对话，实现更加无缝的 AI 体验。为

了体验这一全新功能，用户需要将小艺助手升级到 11.2.10.310 版本及以上。升级完成后，用户只需在底部导航栏中点击“发现”选项，即可轻松进入智能体广场，探索 DeepSeek-R1 Beta 版带来的丰富功能与全新体验。

2. 三大运营商全面接入 DeepSeek

2月8日，工业和信息化部运行监测协调局发布2025年春节通信业务相关情况。2025年春节期间，基础电信企业积极利用AI开展创新服务。其中，中国移动、中国电信、中国联通三家基础电信企业均全面接入 DeepSeek 开源大模型，实现在多场景、多产品中应用，针对热门的 DeepSeek-R1 模型提供专属算力方案和配套环境，助力国产大模型性能释放。

3. 多家公司宣布接入 DeepSeek

随着 DeepSeek 热度飙升，其“朋友圈”也在快速扩容。2月5日，京东云宣布正式上线 DeepSeek-R1 和 DeepSeek-V3 模型，支持公有云在线部署、专混私有化实例部署两种模式。前段时间，阿里云、百度智能云、华为云、腾讯云、火山引擎、天翼云已接入了 DeepSeek 模型。海外的亚马逊 AWS、微软 Azure 等云巨头同样官宣支持。有 IDC 从业者认为，DeepSeek 的火爆有望改善目前智算中心的经营模式，提高国产算力的利用率。随着三大运营商接入 DeepSeek，有助于加速大模型技术与产业场景的融合，推动 AI 技术普惠应用。

4. 多家车企宣布与 DeepSeek 模型深度融合

近日，多家车企宣布完成与 DeepSeek 模型的深度融合，DeepSeek 正加速 AI 大模型上车。吉利是最早官宣的车企。2 月 6 日，吉利汽车宣布，其自研的星睿大模型与 DeepSeek-R1 深度融合。吉利汽车将利用 DeepSeek-R1 模型对星睿车控 FunctionCall 大模型、汽车主动交互端侧大模型等进行蒸馏训练。2 月 7 日，极氪也官宣旗下自研 Kr AI 大模型与 DeepSeek R1 大模型完成了深度融合，并且智能座舱助手 AI Eva 也已集成并即将上线。同日，岚图汽车也称其智能座舱已经与 DeepSeek 完成深度融合，并计划在 2 月 14 日开启 DeepSeek 全民知识蒸馏训练，岚图知音和梦想家将率先搭载。岚图知音将成为汽车行业首个融合 DeepSeek 的量产车型。

2 月 8 日，又有 5 家车企宣布接入 DeepSeek。宝骏汽车宣布，其灵语智舱与 DeepSeek 大模型已完成深度融合，宝骏享境完成实车装载。未来，宝骏云海、宝骏悦也 Plus 等车型也将陆续通过 OTA 实现应用。同属上汽集团旗下的智己汽车宣布，其智能座舱也已深度引入 DeepSeek 大模型，并与豆包、通义等大模型合作，通过深度联合训练，构建多场景插拔式 AI 矩阵平台。东风汽车宣布，公司旗下自主品牌已完成 DeepSeek 全系列大语言模型接入工作，并将于近期陆续搭载应用在东风岚图、东风猛士、东风奕派、东风风神、东风纳米等公司旗下自主品牌车型。零跑汽车宣布，部署 DeepSeek-R1 的零跑全新座舱即将上线。目前，小零 GPT 大

模型已接入 DeepSeek-R1。同时，DeepSeek-R1 大模型已在零跑内部 IT 团队运营部署，辅助工作提效。长城汽车 CTO 吴会肖 2 月 8 日在微博上发文宣布，DeepSeek 的 demo 在长城汽车上已经跑通，Coffee Agent 已完成融合适配。

业内普遍认为，2025 年将是高阶智能驾驶技术全面 AI 化的元年，智能化是新能源汽车产业发展的下半场，而 AI 智驾将成为行业的发展方向。小鹏汽车董事长何小鹏日前在《行稳致远，2025 向蓝海进发》的开工信中表示，下一个十年，AI 会驱动汽车产生巨大变革，也会驱动全球范围内的硬件与软件产生巨变。他也提到了 DeepSeek，称 DeepSeek 大模型给全球科技圈带来震撼——它既实现了媲美 OpenAI 的体验，又将成本压缩到极低。蔚来董事长李斌也在 2 月 8 日晚间的抖音直播间提到，1 月体验 DeepSeek 实在是太魔幻现实主义了。

4. 微软、英伟达、亚马逊全部接入 DeepSeek!

一开始便对 DeepSeek 赞赏有加的英伟达，刚刚宣布：「DeepSeek-R1 正式登陆 NVIDIA NIM」。据介绍，在单个 NVIDIA HGX H200 系统上，完整版 DeepSeek-R1 671B 的处理速度可达 3,872 Token/秒。同在今天，亚马逊也在 Amazon Bedrock 和 SageMaker AI 中，上线了 DeepSeek-R1 模型。曾经冲出来和 OpenAI 一起高调质疑 DeepSeek「偷窃」数据的

微软，甚至在前一天就把 DeepSeek-R1 抢先部署在了自家的云服务 Azure 上。

除了科技大厂，AI 初创们也没有放过如此上好的机会。Windsurf 编辑器同时集成了 DeepSeek-R1 和 V3 模型，并且第一次在编程智能体中，实现了 R1 的工具调用。Cerebras 不仅实现了比 GPU 快 57 倍的推理速度，而且还报告称，自己部署的 70B 模型在准确率上要比 GPT-4o 和 o1-mini 更高。

5. 燧原科技实现全国各地智算中心 DeepSeek 的全量推理服务部署

作为国产算力领军企业，燧原科技完成了对 DeepSeek 全量模型的高效适配，包括 DeepSeek-R1/V3 671B 原生模型、DeepSeek-R1-Distill-Qwen-1.5B/7B/14B/32B、DeepSeek R1-Distill-Llama-8B/70B 等蒸馏模型。整个适配进程中，燧原 AI 加速卡的计算能力得到充分利用，能够快速处理海量数据，同时其稳定性为模型的持续优化和大规模部署提供了坚实的基础。

6. 多家券商也积极“拥抱” DeepSeek

截至发稿，至少有 16 家券商宣布已完成 DeepSeek-R1 模型的本地化部署，包括国泰君安、国金证券、中泰证券、兴业证券、国元证券、华福证券、光大证券、华安证券、广发证券、国盛证券、中金财富证券、华西证券、国信证券、中信建投、东兴证券、西南证券等。

多家券商称，已将 DeepSeek 融入公司多个核心业务领域，将赋能信息检索、文档处理、行业研究、市场研判、辅助软件研发、辅助制定营销方案、合规问答、业务办理指引等多个业务场景。

例如，2月8日，中金财富宣布完成与 DeepSeek-R1 大模型深度融合。据介绍，此次技术应用不仅实现智能投顾助手 IC-Copilot 的升级迭代，更开创性地构建公司投顾服务领域“热点发现—资讯处理—策略生成”三位一体的服务生态。

光大证券介绍，近日光大证券 AI（人工智能）中台新增 DeepSeek 大模型本地化部署和多场景应用测试，并基于华为 NPU 算力平台实现国产化适配。据悉，光大证券此次部署以“自主可控+创新应用”为核心，依托国产化算力基础设施，自主研发知识库构建引擎和多模态数据处理框架，大幅降低大模型应用成本。

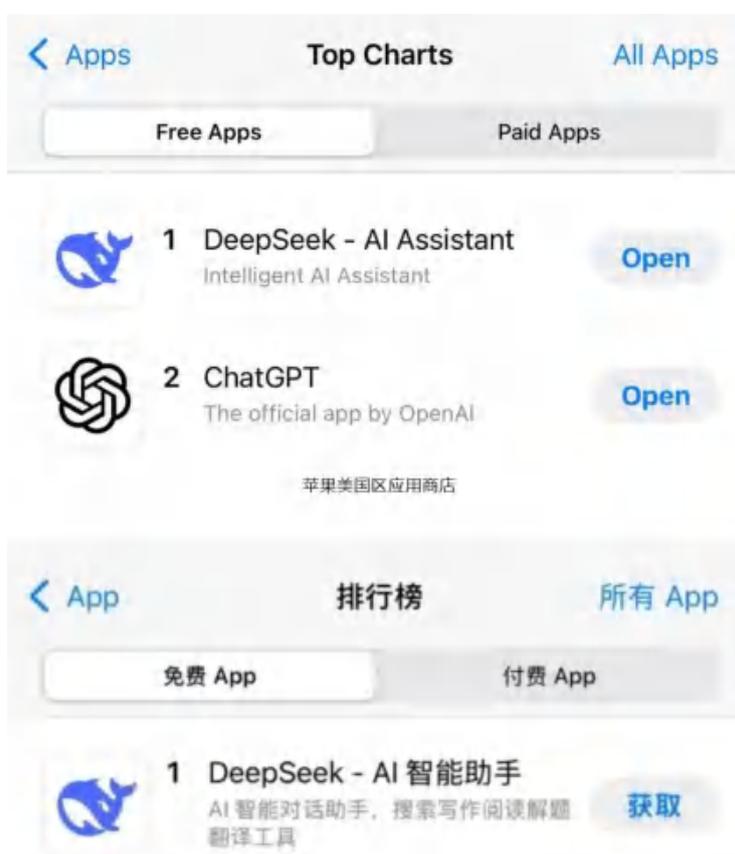
华安证券也在近日完成 DeepSeek 的本地化部署及场景适配。华安证券认为，DeepSeek-R1 大模型“性能倍增、成本递减”的双重优势将加速人工智能在垂直领域的规模化应用，目前公司已初步打造 AI 中台，以大模型作为中枢，整合、调配各类 AI 能力，实现 AI 资源的高效利用与协同运作，为后续业务的智能化升级和创新提供有力支撑。

（二）生态

1. Deepseek 应用 27 日登顶苹果中国地区和美国地区应

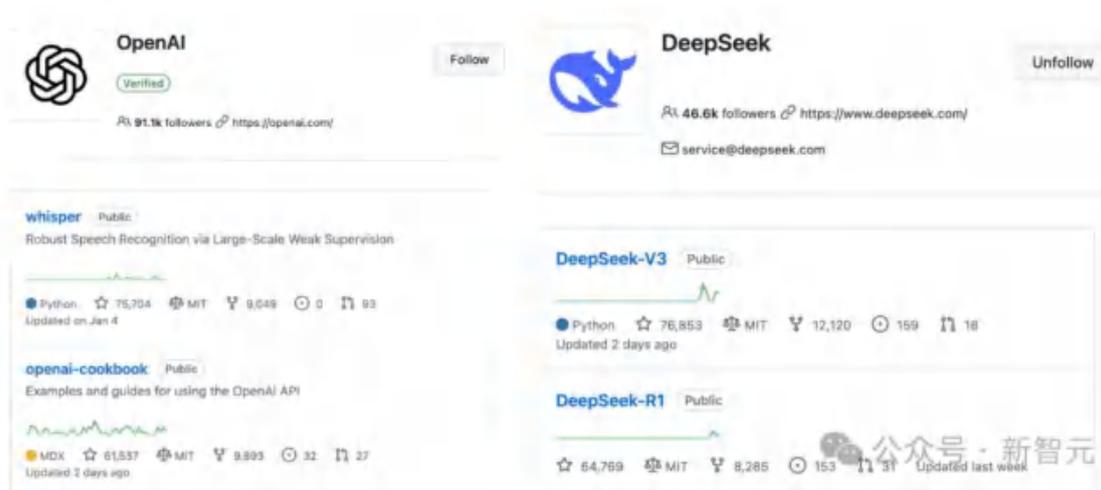
用商店免费 APP 下载排行榜

由国产大模型公司杭州深度求索开发的 Deepseek 应用 27 日登顶苹果中国地区和美国地区应用商店免费 APP 下载排行榜，在美区下载榜上超越了 ChatGPT，在中国、美国的科技圈受到广泛关注。环球时报 27 日称，Deepseek 被认为是大模型行业的最大“黑马”，在外网被不少人称为“神秘的东方力量”。



2. DeepSeek 在开源社区 GitHub 上的 Star 数首次超越 OpenAI

DeepSeek 项目在 GitHub 平台上的 Star 数，已经超越了 OpenAI。热度最高的 DeepSeek-V3，Star 数如今已达 7.7 万。



而 DeepSeek-R1,更是仅用了3周时间,就超越了「openai-cookbook」。前有 App Store 登顶,今有 GitHub 超越,网友们高呼:永远不要低估开源社区的力量。

3. 国家超算互联网正式上线 DeepSeek

国家超算互联网平台已正式上线 DeepSeek-R1 模型的 1.5B、7B、8B、14B 版本,并将于近期陆续更新 32B、70B 等版本。DeepSeek-R1 小版本模型提供一键推理服务,无需下载本地;还可根据私有化需求,引入专有数据,对模型进行定制化训练和开发。

除 R1 模型外,国家超算互联网平台还上线了 DeepSeek-V3、DeepSeek-v2.5 系列、DeepSeek-coder 系列、DeepSeek-math 系列 (7b) 和 DeepSeek-v2 系列 (Lite) 等模型。

4. 传 DeepSeek 正准备适配中国国产 GPU

然而,据美国科技网站 Tom's Hardware 报道,韩国未来资产证券在对 DeepSeek 技术论文进行分析时发现,该模型的硬件效率之所以能比 Meta 等高出 10 倍,因为“他们从头

开始重建了一切”。分析指出，DeepSeek 在使用英伟达的 H800 晶片进行训练时，使用了英伟达的底层硬件指令 PTX（Parallel Thread Execution）语言，而不是高级编程语言 CUDA。由于 CUDA 是通用型编程框架，会导致训练模型时损失一些灵活性。中国网媒“快科技”分析，DeepSeek 的做法相当于绕过了硬件对训练速度的限制，意味着其他模型需要训练 10 天，而 DeepSeek 只需要五天。然而，这种编程非常复杂且难以维护，行业通用的做法是使用 CUDA 这类高级编程语言。快科技和腾讯网引述的消息人士指出，DeepSeek 内部拥有一些擅长写 PTX 语言的内部开发者，假如 DeepSeek 未来有意改适配中国国产的 GPU，在硬体适配方面将会更得心应手。

5. DeepSeek 的风，火速吹遍了终端企业

近期，消费电子类、显示类终端企业“紧跟热点”，纷纷高调宣布接入 DeepSeek，DeepSeek 的风，终于吹到了终端领域（其他领域接入 DeepSeek 情况，请参考本报早前发表文章《关于 DeepSeek，多家算力芯片公司宣布……》《DeepSeek 为何“多云”？》）。如此快速的“结盟”究竟是基于业务发展的技术升级，还是蹭流量的营销动作？

华为：2 月 5 日，华为宣布，基于原生鸿蒙操作系统的小艺助手 App 已经接入 DeepSeek。目前，纯血鸿蒙小艺智能体已经支持 DeepSeek-R1，已升级至原生鸿蒙的华为用户可

以通过“小艺助手 App-发现-智能体广场”与 DeepSeek 进行对话，实现更加无缝的 AI 体验。

荣耀：2 月 8 日，荣耀 YOYO 智能体商店上线 DeepSeek-R1 尝鲜版，成为第二家与 DeepSeek 达成合作的手机厂商。荣耀 PC 产品线总经理朱臣才发文称，2 月的新品 MagicBook Pro 14 将率先全面接入 DeepSeek。

OPPO：2 月 8 日，OPPO 宣布 OPPO 系手机将正式接入 DeepSeek-R1 人工智能推理模型。

视源股份：2 月 8 日，主营显示、交互控制的解决方案提供商视源股份宣布，公司学习机全系列产品目前已接入 DeepSeek，希沃电脑助手、希沃白板、希沃课堂智能反馈系统等产品也将逐步全面接入 DeepSeek。

魅族：2 月 8 日，魅族 FlymeAIOS 团队宣布：确认已完成 DeepSeek-R1 大模型接入，魅族 21 系列、Lucky08 率先上线尝鲜，星纪魅族 StarVAR 眼镜等 AI 生态产品将陆续接入。

洲明科技：2 月 10 日，领先的 LED 应用厂商洲明科技宣布 Agent 平台已接入 DeepSeek 的 V3 及 R1 模型，通过结合 DeepSeek 在自然语言处理和多模态推理上的优势，优化教育场景的多 Agent 交互体验。

利亚德：2 月 10 日，另一家 LED 应用领头羊企业利亚德在互动平台表示，旗下虚拟动点动作大模型 LYDIA 已正式接入 DeepSeek。通过这一整合，LYDIA 在空间动作数据的

认知、感知、理解和生成能力上得到了进一步升级，尤其在动作精准度及动作效果层面得到了进一步提升。

海信视像:2月11日,海信电视宣布正式接入 DeepSeek。海信电视正在进行用户终端的升级发布,预计3—5天内,所有搭载智能体的海信电视将全面支持 DeepSeek。接下来,即将上市的海信电视 2025 年新品也将全面支持 DeepSeek 服务

(三) 应用

1. 文学

根据扬子晚报消息,2月5日,接连两家企业宣布接入 Deepseek 以辅助网文创作。国内网文头部企业阅文集团宣布,旗下作家辅助创作产品“作家助手”应用已集成独立部署的 DeepSeek-R1 大模型,这是 deepseek 首次在网文领域应用。

2. 编程

DeepSeek 的编程能力优秀。DeepSeek 团队在编程方面进行了较长时间的积累和研发,两年前就曾推出过 DeepSeek-Coder 编程模型,而近期广受关注的 DeepSeek V3 和 R1 模型在编程方面的性能得到了大幅的提升,这里可以参考权威的 Aider LLM Leaderboards 榜单(部分结果):

| 模型 | 成功率 | 格式正确率 | 费用 |
|---|-------|--------|---------|
| DeepSeek R1 + claude-3-5-sonnet-20241022 | 64.0% | 100.0% | \$13.29 |
| o1-2024-12-17 (high) | 61.7% | 91.5% | \$186.5 |
| o3-mini (high) | 60.4% | 93.3% | \$18.16 |
| DeepSeek R1 | 56.9% | 96.9% | \$5.42 |
| claude-3-5-sonnet-20241022 | 51.6% | 99.6% | \$14.41 |
| DeepSeek Chat V3 | 48.4% | 98.7% | \$0.34 |
| gpt-4o-2024-08-06 | 23.1% | 94.2% | \$7.03 |
| qwen-max-2025-01-25 | 21.8% | 90.2% | \$0.0 |
| DeepSeek Chat V2.5 | 17.8% | 92.9% | \$0.51 |

在这个专门针对编程场景设计的评测榜单中，DeepSeek R1 取得了 56.9% 的成功率，格式正确率达到了 96.9%，而完整测试集的费用仅为 \$5.42，远低于其他竞争对手，如 Claude 3.5 (\$14.41) 和 OpenAI o1 (\$186.5)。这意味着 DeepSeek 在提供高质量代码生成的同时，也能保持极高的性价比。DeepSeek Chat V3 虽然在成功率上低于 R1，仅为 48.4%，但格式正确率仍达到了 98.7%，并且其调用成本仅为 \$0.34，是榜单中最便宜的模型之一，适合对价格敏感的开发人员使用。从成功率来看，DeepSeek R1 + Claude 3.5 组合达到了 64.0%，这里是使用 DeepSeek R1 推理进行高层次设计，用 Sonnet 完成基层的代码编写，在高准确率的同时，进一步提高了代码生成的可靠性。这表明 DeepSeek R1 具备较强的抽象推理能力，并且在与其他强大模型配合使用时可以取得更好的效果。值得注意的是，GPT-4o 和 Qwen Max 2025 在成功率上明显落后于 DeepSeek，尤其是 GPT-4o 的成功率

仅有 23.1%，而 Qwen Max 甚至只有 21.8%，这说明 DeepSeek 的模型在代码相关任务上的优化更为到位。此外，DeepSeek Chat V2.5 的成功率仅为 17.8%，表明在更新到 V3 版本后，DeepSeek 取得了显著的提升。总体来看，DeepSeek R1 和 V3 在代码生成领域的表现相当优秀，特别是在性价比和格式正确率方面占据明显优势。这使得 DeepSeek 成为开发者在 AI 编程助手领域中的一个有力选择，特别是在需要大规模调用 API 或者希望控制成本的场景下，其价值更加突出。

除了正确率高、价格低廉之外，DeepSeek 模型用于编程还有其它一些优势：多语言支持丰富。DeepSeek 在多种编程语言和开发场景中表现出色，支持 C/C++、Java、Python、JavaScript/TypeScript、MATLAB、R 等 70 余种语言，适用于前端、后端、数据分析及系统脚本编写。相比于其他 AI 编程助手，DeepSeek 经过深度优化，能够理解复杂语法和编程库，并提供精准的代码补全、优化和重构建议。够用的长上下文。得益于其 64K tokens 的长上下文窗口，DeepSeek 能够有效分析大规模代码库，支持跨文件代码分析与优化，使其在实际开发中的应用更加广泛。此外，DeepSeek 在中文编程环境下表现尤为突出，能够理解中文注释和代码需求，降低语言障碍，提高团队协作效率。支持项目级的编程辅助。从应用趋势来看，AI 编程助手正在深度融入开发流程，开发

者需要的不仅仅是代码补全，更希望 AI 能够理解项目、协助调试、优化性能，并提供代码审查建议。DeepSeek 具备这些能力，配合上 Cursor、Aider 这些新一代的编程助手，能够在整个软件开发生命周期中充当智能助手，极大提升开发效率。

（四）基础设施

加拿大网络安全公司 Feroot Security 对 Web 版 DeepSeek 代码的分析显示，该公司使用了中移动的基础设施。中移动是最大的移动通讯服务商，同时也是最大的网络服务商之一，DeepSeek 使用中移动的基础设施并不出人意料。问题在于中移动受到了美国的有限制裁，美国人被限制投资中移动。这项分析没有涉及移动版的 APP，DeepSeek 的 AI 助手是苹果和 Google 应用商店下载量最高的 APP 之一。

五、DeepSeek 大模型产业影响

（一）对人工智能技术的影响

1. 吴恩达谈 DeepSeek：中国 AI 迎头赶上，扩大算力并不是 AI 进步的唯一途径

近年来，围绕“扩大模型规模能够推动 AI 进步”的观点备受追捧。事实上，我曾是这一理念的早期支持者。一些公司利用这一叙事成功融资数十亿美元，声称只要有足够资本，就能通过 (i) 扩大规模和 (ii) 预测性改进来持续提升 AI 性

能。因此，行业对扩大算力的关注远超对其他创新路径的探讨。然而，美国对中国的 AI 芯片出口限制促使 DeepSeek 团队进行大量优化，使其能够在 H800 GPU(算力远不如 H100) 上高效运行，从而在不计研发成本的情况下，仅用不到 600 万美元的计算成本完成模型训练。DeepSeek 的成功经验是否意味着降低对算力的需求仍有待观察。有时，当某项技术变得更便宜时，反而会促使整体投入增加。我认为，从长期来看，人类对智能和算力的需求几乎没有上限，因此我仍然看好 AI 计算需求的持续增长。我在社交媒体上看到人们对 DeepSeek 进展的解读各不相同，仿佛是一场 Rorschach test，每个人都在其上投射自己的理解。我认为 DeepSeek-R1 在地缘政治上的影响尚有待厘清，同时它也为 AI 应用开发者带来了巨大机遇。我的团队已经开始头脑风暴，探索因开源高阶推理模型的可及性而催生的新应用方向。现在依然是一个绝佳的 AI 创新时代！

2. 蒸馏技术将成为行业突破口

DeepSeek 的蒸馏技术更是这一领域的佼佼者，它不仅攻克了传统蒸馏的瓶颈，还在多模态数据处理等前沿领域取得了突破性进展。DeepSeek 的蒸馏技术将数据蒸馏与模型蒸馏相结合，实现了从大型复杂模型到小型高效模型的知识迁移。这种结合方式不仅提升了模型的性能，还显著降低了计算成本。DeepSeek 在知识迁移策略上进行了多项创新，以实现高

效的知识传递和模型优化。DeepSeek 的蒸馏模型架构设计充分考虑了效率与性能的平衡，通过精心设计的模型结构，实现了从大型复杂模型到小型高效模型的知识迁移。DeepSeek 的蒸馏模型训练过程包括多个关键步骤，通过精心设计的训练策略和优化方法，确保了模型的高效训练和性能提升。DeepSeek 的蒸馏模型在推理效率方面表现出显著的提升，这主要得益于模型结构的优化和蒸馏技术的应用。通过将知识从大型复杂模型（教师模型）迁移到小型高效模型（学生模型），DeepSeek 的蒸馏模型在计算资源、内存使用和推理速度方面都实现了显著的优化。尽管 DeepSeek 的蒸馏技术在提升模型性能和降低计算成本方面取得了显著成效，但蒸馏技术仍面临“隐性天花板”的挑战。这一挑战主要体现在学生模型的性能难以超越教师模型的固有能力，限制了模型在新领域或复杂任务中的扩展性。

（二）对产业生态的影响

1. DeepSeek 开源对于企业战略价值

以 DeepSeek 为代表的开源人工智能模型的成功，充分展示了开源模式在企业战略中的巨大潜力。企业通过积极参与开源生态，不仅可以降低研发成本，加速技术迭代，还能借助全球开发者的智慧，提升产品质量和市场竞争力。

DeepSeek 基于开源战略打造人工智能开放协作生态圈。当人工智能模型成为数字时代的核心基础设施，开放源代码

实质是在定义行业标准的话语权争夺。DeepSeek 通过开放模型权重和代码，推动去中心化的技术创新，使开发者、研究机构和企业能够在共享框架上快速迭代，进一步使企业从技术竞赛的参与者转变为游戏规则的制定者。同时，DeepSeek 的开源对闭源大模型形成强烈冲击，打破了封闭式平台的技术壁垒，使企业无需依赖昂贵的 API 付费模式即可获取高质量模型。相比闭源模式依赖的封闭商业生态，开源带来的全球协作效应使技术优化更快，挑战了闭源大模型企业在市场上的主导地位。

DeepSeek 开源模式为企业带来创新速度非线性跃升。传统封闭式研发受限于企业内部资源，而开源社区则形成了一种“分布式创新网络”，使技术迭代呈指数级增长。深度求索的开源实践显示，其模型在开源后 6 个月内获得的优化建议，相当于原团队 3 年的技术积累。通过开源，企业不仅能加速技术演进，还能在市场竞争中占据更主动的战略位置。

DeepSeek 开源助力企业社会责任的结构化转型。通过开源，DeepSeek 在打造更透明的人工智能生态的同时，也推动了伦理治理模式的创新——借助社区共治机制，降低算法偏见、提升人工智能透明度。企业不再只是社会问题的解决者，而是通过技术民主化，让社会各界共同参与治理，塑造更具包容性的人工智能伦理体系。例如，DeepSeek 在开源后，允

许不同文化、不同价值观的社群基于自身需求部署专属人工智能模型，从而摆脱单一价值体系的垄断。

2. DeepSeek 是否真的绕开了 CUDA?

DeepSeek 没有直接使用 CUDA 的 API,而是直接使用 PTX, PTX(Parallel Thread Execution)语言是英伟达(NVIDIA)开发的，专门为其 GPU 设计的中间指令集架构，位于高级 GPU 编程语言（如 CUDA C/C++或其他语言前端）和低级机器代码（流处理汇编或 SASS）之间。PTX 是一种接近底层的指令集架构，将 GPU 呈现为数据并行计算设备，因此能够实现寄存器分配、线程/线程束级别调整等细粒度优化，这些是 CUDA C/C++等语言无法实现的。

直接使用 PTX 的优势是可以对计算过程进行更细粒度的控制。因为 PTX 作为中间语言，能够更接近底层硬件的执行逻辑。DeepSeek 可以根据模型的特点和需求，对 PTX 代码进行针对性的优化，从而更好地发挥英伟达 GPU 的性能。比如，针对 DeepSeek 模型中某些特定的矩阵运算或者算法逻辑，通过直接编写和优化 PTX 代码，可以减少不必要的计算开销，提高计算效率。但是使用 PTX 也有缺点，就是这种编程非常复杂且难以维护，要求技术人员对 GPU 工作机制非常熟悉。从上述内容可以看出，DeepSeek 还是基于 CUDA 的能力，并没有完全抛开 CUDA。

3. DeepSeek 会导致对智算中心的需求大幅减少吗?

近日，关于 DeepSeek 技术是否会引发智算中心停建的话题在业界引起了广泛关注。DeepSeek 作为一种前沿的人工智能算法，其在数据处理和分析方面展现出强大实力。然而，有人担忧其发展将导致相关智算中心建设停滞。对此，CDCC 邀请多位业内专家发表了自己的看法，从不同角度分析了 DeepSeek 技术对智算中心的影响，以期为我国人工智能产业发展提供有益参考。

(1) 韩玉，CDCC 算力中心规划设计研究员，“2018 数据中心科技成果奖青年科技人才奖”获得者。

当阿里去 IOE 代为 PC 服务云计算做数据架构+算力+存储的时代，阿里的基础设施规模从一年几个单体数据中心交付改为两位数的交付，并且快速成为云计算服务行业和服务器规模体量国内第一。

杰文斯悖论 (Jevons Paradox) 揭示了技术进步反而最终造成能源消耗总量增加，因为技术进步带来的效率提升和成本降低可以让公众服务需求成为更快普及。

DeepSeek 的火爆是一种开源模式和闭源模式的竞争分水岭，也是一种效率和成本解决某种方案更优解，但是 DeepSeek 也是站在巨人肩上 (比如使用了 GPT 的部分标注数据和一些其他网上数据-应该非侵权的)，这只是 AI 发展阶段的一个阶段里程碑，即从算力资源的占有和抢时间领先改变为算力性能效率和成本优化整体架构成本优化竞争领

先的阶段，最终谁走到最后领先行业风骚尚早，但是反而成为百花齐鸣的开始。

算力的竞争门槛被大幅降低，智算中心的春天反而会快速扩展，超大资源规模智算中心是否必须建设其实是业务发展的阶段，即使 10 万卡的数据中心也仅仅是 100-200MW 的规模，只是头部客户需求的几分之一，但是数十 MW 的智算中心会更多发展，更多智算企业可以入局；至于边缘计算，从来不是智算中心的范围，仅仅就是边缘配套也就是运营商接入和布局的范畴，中大型的智算中心的春天开始并不远，相信 AI 的前景就相信中大型智算中心的明天。

(2) 唐虎，CDCC 专家技术组委员、“2023 数据中心科技成果奖青年科技人才奖”获得者。

《DeepSeek 让多地智算中心停建》文章大方向没问题，但多少有点儿“震惊体”的感觉。还是要回归现实，智算的缓建是没有看明白，之前的“疯狂建”也是没有看明白，是“一窝蜂”特质多一些，看明白的都是按部就班的。

DeepSeek 的意义就犹如其创始人所讲的“自信”。宏观上，其复现了证明了 OpenAI 的工作。微观上，它就是看清楚，按部就班提质增效的典范。更重要的影响来自于，它的完全开源的胸襟。

关于 DeepSeek 的成熟制程，本人也多次给国产 GPU 伙伴提及过，但这个系统变革的自信是缺乏的。DeepSeek 给大

家带来了自信，并且让需求更加释放了，AI 才刚开始。对于数据中心行业，也要学习 DeepSeek，按部就班的自信。

在半导体产业的后摩尔时代，数据中心向算力中心演进的时代，需要数据中心、服务器及半导体的国人从业者，拿出自己的按部就班的自信的路径。

(3) 官伟达，CDCC 专家技术组委员、世纪互联 DC 运维产品负责人。

“冰水为之而寒于水”，在西方国家算力限制的情况下，DeepSeek 通过“开源+算法优化+端侧革命”，对半导体、AI 等行业影响深远。打破 CUDA 生态限制，为中国芯片设计带来机遇；动摇“制程决定论”，推动算力需求结构迁移；触发端侧硬件生态链式反应，重塑全球半导体价值链，对于国内数据中心建设将带来巨大的变化：

建设规划方面，规模规划调整：大规模智算中心建设需求需要观望，需要重新思考和规划建设规模与方式。布局倾向分布式：推理算力需求会快速井喷式上升，分布式机房内部署 1-10 台 GPU 服务器的小规模集群需求爆发，数据中心布局会更倾向于分布式，以满足就近访问需求。

技术架构方面，ARM 架构渗透：“低成本 AI 算力新范式”，会推动以卓越能效比和性价比著称的 ARM 架构加速渗透至数据中心，数据中心为降本增效，可能更多采用 ARM 架构服务器。算力优化升级：为支持 DeepSeek 等高效 AI 模型，

数据中心会加快硬件升级，采用更先进的芯片、存储和网络设备，提高整体算力和性能。

能源与散热方面，降低电力需求压力：**DeepSeek** 以较低能耗就能实现较高性能，有助于降低数据中心过度供应的风险，减少数据中心的能源消耗，降低电力需求压力，数据中心电力申请需求，将需要重新考量。散热技术仍受挑战：尽管整体能耗可能降低，但在满足算力需求时，仍可能因高密度计算设备的使用，使得数据中心的散热需求上升，需采用液冷等绿色化散热技术。

市场竞争方面，加剧芯片市场竞争：**DeepSeek** 的低成本对 **AMD** 等芯片制造商的数据中心业务产生影响，将推动数据中心硬件市场的竞争更加激烈，促使相关企业不断创新和优化产品，国产芯片（产能优势）即将登上世界历史舞台。数据中心竞争格局生变：能快速适应 **DeepSeek** 等新技术，提供更高效、低成本算力服务的数据中心企业，将在市场竞争中占据更有利地位。

（4）车凯，联通数字科技有限公司安全管理、高级工程师。“2021 数据中心科技成果奖青年科技人才奖”获得者。

DeepSeek 等大模型在资源占用和能效方面的优化，确实可能对智算中心等建设产生一定影响，存在波动，但从整体和长期看，依然会推动基础设施的持续发展。

1) 通用领域成熟，垂直领域仍在高速发展。大模型已经在文字、图片、视频等形式方面，以及聊天、检索、生成等通用领域取得了较高的应用和较大的成功，相对比较成熟，因此这些领域已经从高速发展向着高质量发展转型，能效显得非常重要。而医疗、制造业等垂直行业相关领域正处于发展阶段，训练和推理需求正在呈指数级增长，资源需求必将增加。

2) 技术效率提升，将向着更高质量发展。效率提升引发成本的降低，通常情况下，输出质量越高，消耗的资源就越多。节省的资源可能被重新投入，以提升输出的质量。我们现在觉得 DeepSeek 效率已经很高，但是离真正高质量还相差甚远。

3) 技术效率提升，将刺激需求的增加。大模型能效提高可能降低单任务成本，但也进一步刺激了应用的普及，用户群体和应用场景的增加。以 DeepSeek 自身为例，随着用户数量的激增，我们经常会遇到“服务器繁忙，请稍后再试”的情况，效率再高，也挡不住这种 DDoS 攻击，因此持续建设是必然。

资源节约型 AI 应该是算力基础设施可持续发展的推动条件，而非抑制其增长的因素。数据中心行业在关注 DeepSeek 等效率提升的同时，还是应该更多关注自身的高效能转换和绿色化发展，避免单纯的规模性扩张。

(5) 俞佳炆，华信咨询设计研究院有限公司（中通服数字基建产业研究院）副主任工程师，中通服金牌内训师，天翼云高级解决方案架构师。

DeepSeek-R1 的出现，对于算法的优化是否会产生算力过剩？个人认为可能恰恰相反，DeepSeek-R1 的突破降低了大模型的边际成本，这促使过去因成本过高而被抑制的需求得以释放，催生更多应用场景，应用的繁荣一定是牵引出更多算力需求，就如同内燃机、发电机效率的优化，不会导致原料降价，只会催生出更多需求。

不管怎么样 DeepSeek 并没有打破 scaling law，通往 AGI 的路程还是很漫长，关于基础大模型的角逐会因为鲶鱼效应继续保持火热。DeepSeek 的开源做出了很大贡献，其中之一是把大家注意力拉回到国产大模型+国产卡+算法优化的正途上。这种先行者的成功是很振奋人心的，从数据看，截止 2 月 6 日，已经有华为、沐曦、天数、摩尔、海光等 10 大国产芯片支持了 DeepSeek，各大云商也纷纷跟进，一定会掀起“模型/算法定义芯片”的热潮，对于国产算力是利好。

对于效率至上的云商和模型服务提供商而言，DeepSeek 在 MoE 架构、低精度训练、通信效率等方面做出了优化，这些对于 Infra 的影响都决定着智算中心从芯片选取到整体架构设计的各个环节，例如：如何平衡算力和显存的取舍、低

精度算力芯片的发展趋势、专家并行方式对于网络架构的影响，这些问题都值得重新审视一番，给出自己的回答。

(三) 对国际竞争的影响

1. DeepSeek 的崛起，冲击了全球科技供应链的核心玩家

Nvidia: AI 训练不再需要超大规模 GPU，市场开始质疑 Nvidia 高端 AI 芯片的需求是否会放缓。AI 训练成本下降 \neq AI 计算需求减少，只会让计算变得更分散、更普及。未来 AI 可能更倾向于“终端推理+云端微调”模式，而不是完全依赖数据中心训练。台积电、ASML: 如果 AI 计算需求下降，先进制程芯片的增长预期可能被重新评估。微软: 投资数百亿美元押注 OpenAI，但 DeepSeek 的低成本路线可能会影响 OpenAI 的商业化策略。

2. AI 低成本化，最大输家可能不是 Nvidia，而是依赖大模型收费的 AI 公司

OpenAI、Anthropic、Google DeepMind 等公司的盈利模式依赖于高昂的训练成本，只有少数公司能做，AI 被作为稀缺资源 高价出售。DeepSeek R1 把 AI 价格压到 OpenAI o1 的 3%，这可能会彻底 颠覆大模型 API 的商业模式。由于影响实在太太大，我们已经可以看到美国（及其他国家）的各种反制，未来各种制裁可能进一步升级：Anthropic CEO 发万字檄文，希望白宫进一步加码对华芯片出口管制 OpenAI

称 DeepSeek 通过“蒸馏”(distillation)技术侵权使用 openAI 的模型意大利下架 Deepseek (以数据隐私为由)。

3. AI 进入创业黄金时代

大模型创业已成为显学，无论是 SaaS、教育、医疗，还是内容创作、游戏，都在快速 AI 化。过去 AI 创业受限于高昂的训练成本，现在成本下降，创业门槛极大降低。AI API 价格下降 95%，独立开发者也能负担得起 AI 训练。2025 年：AI 爆款应用将提前 1-2 年到来 AI 完全自动化的内容创作(文章、视频、音乐、游戏)将进入 C 端市场。AI 智能助手将更贴近人类思维，从工具变成真正的“数字劳工”。AI 将进入工业、医疗、教育等核心行业，提高效率，重塑职业分工。如果说 2023 年是“大模型元年”，那么 2024-2025 年就是“AI 商业化元年”。这场革命，不是遥远的未来，而是正在发生。

4. DeepSeek 的崛起引发了全球资本市场的剧烈震荡

1 月 27 日，美国纳斯达克综合指数显著下跌，市值排名前十的科技公司开盘总计蒸发近 1 万亿美元。其中，美国芯片巨头英伟达股价单日最大跌幅达 17%，超威半导体、微软等巨头股价同步下跌 2%~6%。这场冲击波迅速蔓延至全球资本市场，台积电、三星电子等亚洲半导体企业股价同步走低，与美股科技股形成强烈共振。

受 DeepSeek 影响，英伟达在 1 月 24 日至 2 月 3 日期间，股价累计跌幅超 20%。对此，英伟达曾紧急对外发表声

明，试图以“DeepSeek 证明未来需要更多芯片”的论调稳定投资者信心。但市场仍普遍担忧，DeepSeek 以更低的成本开发出与 OpenAI 等行业巨头相媲美的产品，将颠覆传统 AI 依赖“高资本、高算力”“大力出奇迹”的烧钱模式。而算法进步将大幅降低算力需求，动摇英伟达等硬件厂商的业绩增长。

与此同时，中国 A 股市场却呈现出另一番景象：A 股 DeepSeek 指数在春节后三个交易日累计涨幅达 27.48%。2 月 5 日，有 11 只 DeepSeek 相关概念股开盘即涨停。其中，并行科技创下“30cm”涨停纪录。安凯微、优刻得、安恒信息、青云科技等“DeepSeek 概念股”一度连续“20cm”涨停。在港股市场，“DeepSeek 概念股”也表现活跃，部分股票涨幅惊人。

六、国内相关表态

（一）国家层面

1. 中国常驻联合国代表傅聪回答记者提问

当地时间 2 月 3 日，在美国纽约联合国总部，中国常驻联合国代表傅聪举行记者会。当被问到中国人工智能企业深度求索（DeepSeek）发布的模型及中美人工智能合作问题时，傅聪表示：“从华为到 TikTok，再到 DeepSeek，美国还想禁多少？”“永远不要低估中国科研人员的聪明才智。DeepSeek 引发全球轰动和一些人的焦虑恐慌，说明技术遏制和技术限制无法奏效，这是全世界、特别是美国需要学习的一课。”

2. 外交部发言人郭嘉昆主持例行记者会

在中国外交部 2 月 6 日举行的记者会上，有记者提到，韩国一些政府部门已屏蔽对 DeepSeek 的访问，之前意大利、澳大利亚、印度、美国、日本等国也传出禁止或限制使用 DeepSeek 的消息，一些企业也出现屏蔽 DeepSeek 访问的情况。对此，中国外交部发言人郭嘉昆表示，我想强调的是，中国政府高度重视并依法保护数据隐私和安全，从来没有也不会要求企业或个人以违法形式采集或存储数据。中方一贯反对泛化国家安全概念、将经贸科技问题政治化的做法。同时，中方也将坚定维护中国企业的合法权益。

（二）地方层面

1. 广东

根据广东省人民政府官网，2 月 5 日，广东省委书记黄坤明同志在广东省高质量发展大会上强调，“深度求索（DeepSeek）以不怕虎的勇气锐气撼动美国人工智能巨头、产生连锁效应”。

2. 郑州

2025 年 2 月 5 日，郑州市委书记安伟在优化营商环境大会上指出，要加快引入 DeepSeek 等大模型。

3. 北京经济技术开发区

北京经济技术开发区表示：“支持区域科技企业与 DeepSeek 多元合作，推动我国大模型产业从单点突破转向体系化创新。”

七、各国相关表态

（一）美国

1. 龙头企业单位

（1）英伟达

2025 年 1 月 27 日，Deepseek 引发了美国纳斯达克震动，尤其受冲击最大的英伟达股价下跌 16.97%，上一个交易日下跌 3.12%。据路透社报道，在中国人工智能（AI）初创公司 DeepSeek 崛起冲击美股之际，美国芯片巨头英伟达当地时间 27 日股价暴跌约 17%，随后该公司发布声明，称 DeepSeek 的进步表明了其芯片在中国市场的有效性，未来将需要更多的英伟达芯片来满足对 DeepSeek 的服务需求。

路透社称，英伟达发表上述声明前，该公司股价暴跌约 17%，至 118.58 美元每股，原因是投资者担心 DeepSeek 使用的英伟达芯片远少于美国公司，但同时也能与美国开放人工智能研究中心（OpenAI）等竞争对手不相上下。此外，英伟达的竞争对手超威半导体公司（AMD）的股价同日也下跌超 6%，至 115.01 美元每股。

报道称，英伟达在 27 日的声明中表示，“DeepSeek 的表现说明了如何利用技术创建新模型，（也就是如何）利用广泛可用的模型和完全符合出口管制的计算（创建新模型）。”

报道称，英伟达认为，DeepSeek 取得的进步表明市场对其芯片仍有需求，“推理需要大量的英伟达 GPU 和高性能网络。”

（2）OpenAI

2 月 4 日，据参考消息援引德国之声电台网站 2 月 3 日报道，美国开放人工智能研究中心（OpenAI）首席执行官奥特曼 2 月 3 日表示，该公司没有计划起诉中国 AI 初创公司杭州深度求索人工智能基础技术研究有限公司（DeepSeek），“我们打算继续打造出色的产品，并在模型能力方面引领世界，我认为这样会很好。”

奥特曼称：“DeepSeek 是一个令人印象深刻的模型，但我们相信我们会继续开拓前沿并推出出色的产品，所以很高兴能有另一个竞争对手。他表示：“我们之前有过很多竞争对手，我认为继续推进并保持引领地位符合每个人的利益。”此前，OpenAI 在 1 月 29 日表示，有证据表明 DeepSeek 使用 OpenAI 专有模型来训练自己的模型，并暗示这可能违反了 OpenAI 服务条款，但没有进一步提出证据。

(3) 谷歌

谷歌首席执行官 Sundar Pichai 也认为, DeepSeek 是一个了不起的团队,“回顾过去三年来的发展,会发现用于 AI 推理的支出比例相比 AI 训练一直在增加。这是件好事,因为推理显然可以支持企业获得良好的投资回报率(即加速推动应用落地)。”他指出,推理使用成本将不断下降,使更多用例变得可行,“机会空间要多大就有多大(as big as it comes),这就是为什么我们在持续投资以期迎接这一时刻。”

(4) 苹果

在苹果财报会议中,库克特别提到了中国 AI 公司 DeepSeek,称其推出的 AI 模型表现出了出色的效率提升和创新能力。尽管分析师询问是否 DeepSeek 会对苹果的利润率构成威胁,库克则表示,任何能够推动效率提升的创新,苹果都会表示肯定。对于苹果来说,DeepSeek 展现出的技术能力显然符合苹果对创新和高效的追求。

(5) 微软

微软 CEO 萨提亚·纳德拉对 DeepSeek 给予了高度评价,认为其展示出的创新能力和技术突破标志着人工智能领域的一个重要进展,还认为随着人工智能技术成本的逐步下降,像 DeepSeek 这样的企业将极大地推动整个行业的发展。

2. 美国联邦政府

（1）美议员推动立法促中美 AI 全面脱钩

日前，美国国会参议员乔什·霍利（Josh Hawley）提出《2025 年美国人工智能能力与中国脱钩法案》（Decoupling America's Artificial Intelligence Capabilities from China Act of 2025）。这是美国国会议员首次提出旨在推动中美人工智能领域“全面脱钩”的法案。该法案提议通过立法手段限制美国与中国的 AI 技术交流，以防止先进技术流入中国。霍利参议员称，流入中国人工智能的每一美元和每一千兆的数据，最终都将被用来对付美国。美国不能以牺牲自身实力为代价来增强我们最大的对手的力量。确保美国的经济优势意味着切断中国与美国创新的联系，并停止对中国创新的补贴。

法案指出：禁止将在中国开发或生产的 AI 或生成式 AI 技术或知识产权进口到美国。禁止向中国出口、再出口或在中国境内转移 AI 或生成式 AI 技术或知识产权。美国个人不得故意在中国境内或为受关注实体进行 AI 或生成式 AI 的研究或开发。美国个人不得故意将 AI 或生成式 AI 研究信息转移到中国境内或受关注实体。违反研发禁令的实体可能面临高达 1 亿美元的罚款。

（2）美国海军禁止官兵使用 DeepSeek 大模型

从美国国会、五角大楼、NASA 到海军，都考虑或已开始禁止使用 DeepSeek，德克萨斯州则成为美国第一个禁止在政府设备上使用 DeepSeek 的州。近日，美国国会更是变本加

厉，提出新法案规定下载 DeepSeek 将构成犯罪，最高可判处 20 年监禁。

3. 美国地方政府

(1) 得克萨斯州宣布禁止政府设备使用 DeepSeek

得克萨斯州则成为美国第一个禁止在政府设备上使用 DeepSeek 的州。

(二) 英国及欧洲

1. 龙头企业单位

(1) 英国阿兰·图灵研究所

英国阿兰·图灵研究所高级 AI 研究员尼尔·劳伦斯 (Neil Lawrence) 指出，英国在 AI 技术的开发和部署方面落后于美国和中国。按照他的说法，虽然德国、法国、芬兰和瑞士有一些“令人鼓舞的迹象”，但“大多数欧洲国家的情况类似”。

(2) 德国墨卡托中国研究中心

德国墨卡托中国研究中心 (MERICS) 科技与创新项目负责人杰伦·格鲁内韦根-劳 (Jeroen Groenewegen-Lau) 也认为，欧洲若想要与中美主要 AI 企业竞争将非常困难。他分析，欧洲在通用人工智能 (AGI) 的竞争也许已经“不现实”，“但欧洲仍然可以通过在 AI 技术的实施和特定领域上深耕，找到自身的价值”。

（3）英国利兹大学

英国利兹大学自动推理教授安东尼·科恩（Anthony Cohn）则提到，从 OpenAI 下一个大型 AI 模型 GPT-5 的发布似乎被推迟来看，当前有迹象表明，基础 AI 模型的技术进步正在放缓，这也意味着其他国家和公司将有机会赶上。

（4）荷兰阿斯麦

阿斯麦总裁兼首席执行官富凯在财报发布会上表示，DeepSeek 推出高效 AI 模型有助于降低 AI 应用成本，为阿斯麦带来更多商机，因为更低成本意味着 AI 应用场景增加，会带动芯片需求增长。

2. 政府及公共机构

（1）意大利

意大利数据保护机构 Garante 下令屏蔽 DeepSeek；1 月 29 日，DeepSeek 应用在意大利的谷歌应用商店（Google Play）和苹果应用商店（App Store）已无法下载。

（2）爱尔兰

爱尔兰数据保护委员会（DPC）要求 DeepSeek 提供更多信息。

（3）法国

法国监管机构国家信息与自由委员会 30 日表示，将对杭州深度求索人工智能基础技术研究有限公司（DeepSeek）

进行问询，以便了解这家中国初创企业的人工智能系统是如何运行的，以及可能存在的隐私风险。

（三）韩国

1. 龙头企业单位

（1）韩国

韩国个人信息保护委员会 1 月 31 日表示，个人信息委将向 DeepSeek 总部发送质询函，内容包括了解人工智能利用个人信息的学习过程等，以确认其个人信息收集项目和流程，以及处理和保存方法。

个人信息委相关人士表示，DeepSeek 在韩国的用户剧增，引发各种忧虑，因此采取上述措施。委员会将根据回信内容等进行综合考量。

（2）LG

LG AI 研究院等机构已启动对 DeepSeek AI 模型的测试和分析。

（3）三星

1 月 31 日，三星电子存储事业部副社长金宰俊（音）在业绩说明会上表示：“我们正在密切关注业界动向，并针对不同情况制定多种应对方案。”他进一步指出，目前可获得的信息有限，因此尚难以准确评估 DeepSeek 的影响，但预计市场上既存在长期机遇，也存在短期风险。

2. 政府及公共机构

(1) 韩国代理总统崔相穆将 DeepSeek 称之为“新的冲击”

韩联社 2 月 5 日报道称，多家韩国公共机构和民营企业以担忧中国产人工智能(AI)大模型“DeepSeek”(深度求索)引发所谓“过度采集用户信息”为由，发布 DeepSeek 禁令。其中，韩国 IT 巨头 Kakao 内部公告称，禁止员工在工作中使用 DeepSeek; 韩移动运营商 LG U+当天也禁止在内联网使用 DeepSeek，还建议员工尽量避免在个人电脑上使用 DeepSeek，并提醒利用该工具时勿输入“敏感或重要的个人信息”; 此外，韩国两大能源国企宣布禁止使用 DeepSeek，韩国代理总统崔相穆将 DeepSeek 称之为“新的冲击”，并直接端出 34 万亿韩元（约合人民币 1710 亿元）的新基金用于支持 AI 和半导体技术发展。他提出韩国的目标是成为世界三大 AI 领先国家之一。

(2) 韩国祥明大学

韩国祥明大学半导体系统工程学科教授李钟焕(音)表示，“DeepSeek 提出的并非全新的概念，但它与未来 AI 半导体的发展方向是密切相关的。随着内存与非内存半导体需求的增加，国内企业也应迅速完成商业化进程，抢占市场先机”。

(四) 其他国家

1. 政府及公共机构

（1）澳大利亚

2月4日，澳大利亚宣布禁止中国人工智能 DeepSeek 在政府系统和设备上使用，声称其构成所谓“国家安全”风险。

（2）日本

日本数字大臣平将明称，希望通过内阁网络安全中心提醒各省厅注意，从保护数据的角度说，在不能消除担忧之前，希望各省厅公务员谨慎使用中国初创企业 DeepSeek 开发的人工智能模型。

他在冲绳县名护市发表演讲时表示，“个人信息能否得到切实保护是问题所在。从保护用户个人信息的角度看，如有必要，个人信息保护委员会应采取相应措施”。

（3）印度

印度电子和信息技术部长阿什维尼·瓦伊什瑙表示，印度计划在未来 10 个月内开发本土大语言模型，并将 DeepSeek 部署在印度的服务器上。

八、DeepSeek 大模型的经验启示

（一）技术

1. 中科院为何难现 DeepSeek 式突破

（1）制度基因的先天性分野

1、目标函数错位：国家使命与市场生存的本质差异

中科院自 1949 年成立之初就被赋予“国家战略科技力量”的定位，其核心 KPI 始终围绕重大基础研究、关键核心技术

攻关展开。在 2016-2020 年承担的 173 项国家重大专项中，92%属于追赶型技术，仅有 8%涉及前沿探索。这种“补短板”导向，使其创新路径天然倾向于渐进式改良。

反观 DeepSeek 这类民营企业，其生存法则要求必须找到市场空白点实现突破。当 OpenAI 在 2020 年发布 GPT-3 时，中科院团队尚在论证“中文大模型必要性”的立项报告，而 DeepSeek 已用抵押房产获得的 2000 万启动资金开始架构设计。市场机制的灵敏嗅觉与决策速度，在此形成鲜明对比。

2、资源分配机制的代际落差

中科院的“课题制”科研模式，至今仍保留着明显的计划经济特征。以某重点实验室为例，其年度预算中 72%用于设备维护与人员工资，真正投入前沿探索的经费不足 15%。更关键的是，科研经费使用需严格遵循“三公经费”管理规定，购买一张 GPU 显卡需要经历 6 层审批流程。

民营企业则展现出完全不同的资源配置效率。DeepSeek 在模型训练阶段，曾创下 72 小时内调用 3000 块 A100 芯片的纪录。这种资源调度的敏捷性，源于其构建的“云算力交易所”创新模式——通过区块链技术将分散的算力资源证券化，实现算力资产的实时交易与组合优化。

3、人才激励的范式冲突

中科院实行的“职称-项目-奖励”三位一体评价体系，导致科研人员深陷“四唯”（唯论文、唯职称、唯学历、唯奖项）

困局。某人工智能团队负责人透露，其团队每年需完成 12 篇 SCI 论文的硬性指标，直接挤压了 30% 的工程化时间。这种机制下，科研人员更倾向于选择风险小、易出成果的改良型课题。

而 DeepSeek 等企业构建的“技术期权池”制度，让核心研发人员可分享产品商业化收益的 15%-25%。其首席科学家年仅 32 岁，却掌握着数亿元研发资金的自主支配权。这种“高风险高回报”的激励机制，吸引了大批顶尖人才从体制内出走——2022 年清华计算机系博士毕业生中，选择加入科技创业公司的比例首次突破 60%。

（2）创新效率的量子级差距

1、技术迭代速度的维度差异

在深度学习领域，中科院某团队 2018 年启动的“智源”大模型项目，历时三年完成千亿参数模型训练。而 DeepSeek 在 2021-2023 年间，实现了从百亿参数到万亿参数的六级跳。这种速度差源自两者完全不同的技术进化路径：前者遵循“立项-论证-采购-开发”的线性流程，后者采用“开发-测试-迭代”的敏捷模式。

更值得关注的是技术路线的选择机制。中科院团队在模型架构设计阶段，需组织三次专家评审会，耗时近五个月完成技术方案论证。而 DeepSeek 通过 A/B 测试平台，可在 24

小时内完成 20 种架构方案的并行验证，利用实时数据反馈快速决策。

2、工程化能力的代际鸿沟

中科院某自然语言处理实验室的成果转化率长期徘徊在 8%左右，其开发的语义分析系统在真实场景中的错误率是商业产品的 3-5 倍。究其原因，在于缺乏工程化思维——研究人员更关注算法在标准数据集上的表现，而非实际应用中的鲁棒性。

DeepSeek 则构建了独特的“场景反哺”机制。在开发医疗对话系统时，其工程师团队直接入驻三甲医院急诊科，通过 2000 小时的真实医患对话记录优化模型。这种“浸泡式开发”使其产品在特定场景的准确率达到 97.3%，远超学术界的 benchmark 水平。

3、生态系统的降维打击

中科院的创新生态仍停留在“实验室-期刊-评奖”的闭合循环，其技术成果转化主要依赖技术转让办公室（TTO）。统计显示，中科院近五年专利转化平均周期为 4.7 年，且 76% 的转化发生在关联企业。这种封闭性导致创新要素难以流动。

而 DeepSeek 搭建的开源社区已吸引全球 23 万开发者参与，形成了“技术突破-开源共享-商业变现”的飞轮效应。其开发的 MoE（混合专家）架构通过开源快速迭代，在 GitHub 上

获得 8.4 万星标，衍生出 300 多个行业解决方案。这种开放创新模式带来的网络效应，是传统科研机构难以企及的。

（3）体制突围的艰难探索

1、中科院的自我革新实验

面对困局，中科院并非毫无作为。其 2018 年启动的“率先行动”计划，试图在管理体制上突破：成立深圳先进院等新型研发机构，实行“预算包干制”；在合肥建设“科学岛”，探索“沿途下蛋”式成果转化。但这些改革仍受制于事业单位体制——某新型研究院在尝试员工持股时，因触及国有资产管理规定被迫中止。

2、新型研发机构的第三条道路

苏州生物医学工程研究所的“事业单位+市场化运营”双轨制试验颇具启示。该所通过设立产业基金控股产业化公司，科研人员可保留编制参与创业。这种模式下诞生的心擎医疗，仅用三年就研发出国产首个体外人工心脏。但这种模式推广面临制度障碍：涉及编制、社保、国资监管等多重政策壁垒。

3、国家实验室的体制嫁接尝试

合肥量子信息科学国家实验室引入民营资本参与建设，形成“国家所有、混合运营”模式。实验室将基础研究设施向国盾量子等企业开放，企业则反馈工程化经验。这种协同创新使中国量子通信技术保持领先，但知识产权归属等深层矛盾仍未完全破解。

(4) 破局之道：构建科技创新共同体

1、制度层面的“破壁工程”

- 推行“创新特区”政策，在特定领域允许科研机构试行企业化运营

- 建立“技术移民”通道，打破事业单位编制对人才流动的束缚

- 改革科研经费管理制度，引入“负面清单”和“里程碑拨款”机制

2、评价体系的范式重构

- 建立“三维评价”体系：学术价值（30%）、产业影响（40%）、社会效益（30%）

- 试点“终身追认制”，对基础研究成果实行长期价值评估

- 将技术转移成效纳入机构考核核心指标

3、创新要素的量子纠缠

- 构建“国家实验室-民营企业-风险投资”创新联合体

- 建立科研设施共享平台，将 1500 亿元存量设备资产证券化

- 发展技术经理行业，培育专业化的科技成果转化服务商

（5）超越二元对立的创新哲学

中科院与 DeepSeek 的对比，本质上是两种创新范式的时代对话。前者代表着重型化、组织化的“大科学”传统，后者彰显着敏捷化、网络化的“新科学”特征。在 AI 制药领域出现的“中科院基础研究+Biotech 公司临床开发”新模式，预示着未来科技创新既非单纯的体制内突破，也不是完全的市场化奇迹，而是需要构建更具弹性的创新生态系统。

当中国科技创新进入“无人区”，或许真正的破局点在于：如何让中科院的战略定力与 DeepSeek 的市场锐度发生化学反应，孕育出兼具国家使命担当与市场生存能力的“新物种”。这场静悄悄的科技体制改革，或将决定中国能否在下一个创新周期掌握定义未来的权力。

2. 王凡、刘少山：中国如何引领全球具身智能？

DeepSeek 的战略启示 | AI 观察

在人工智能领域的迅猛发展浪潮中，一个名为 DeepSeek 的新兴 AI 大模型应用异军突起，以超乎想象的速度跃居全球科技舞台的中心。在极短的时间内，其每日活跃用户数量便超越了先前的热门平台 ChatGPT，并在此后的日子里持续激增，成功触及 2000 万日活跃用户的里程碑，彰显了惊人的成长动能。

这一非凡成就不仅令 OpenAI 与 NVIDIA 等业界巨头深感惊讶，同时也激起了国内 AI 企业的热烈反响。阿里云、腾

讯云等国内科技领军企业迅速表态，宣布对 DeepSeek 提供支持，此举标志着中国 AI 企业在全球范围内的竞争力迈上了一个崭新的台阶。

本文通过借鉴 DeepSeek 的发展经验，探讨中国如何构建自主创新、开源、可落地的具身智能技术路线，并引领全球发展。DeepSeek 通过自主创新和开源，不仅提升了中国在 AI 领域的技术独立性，也推动了产业升级，增强了中国在国际 AI 竞争中的发展信心。沿着这一思路，未来我国具身智能的发展需要具备几个关键特征：开源、可落地、适应多个应用场景，并能够成为国际标准的基础模型。最近发布的具身智能开源项目 OmniRL 突破了上下文强化学习的上限，展示了上下文强化学习这一技术路线的可行性。作为开源项目，OmniRL 展现了可赋能多个产业的巨大潜力。与 DeepSeek 类似，OmniRL 为中国在全球具身智能标准的制定中提供了技术支持，帮助中国在国际 AI 竞争中占据有利位置。

（1）DeepSeek 的突破对具身智能的启示

最近，DeepSeek 的横空出世让人们看到了中国在通用人工智能（AGI）领域突破技术壁垒、摆脱国际依赖、构建本土生态的巨大潜力。作为一家专注于 AGI 研发的公司，DeepSeek 在自然语言的复杂推理训练上取得关键性突破，打破了长期以来西方科技巨头在大模型领域的垄断格局。DeepSeek 的成功不仅推动了中国 AI 产业的自主创新，还在

国家安全、国际竞争、数据主权等方面提供了战略支撑，使中国在全球 AI 竞赛中占据了重要地位。然而，DeepSeek 的变革局限于语言智能领域，人工智能的终极目标不仅仅是处理文本、生成语言，而是具备理解、推理、决策和执行的能力，能够在复杂的物理世界中自主学习并适应变化。真正的通用智能需要超越语言模型，迈向具身智能(Embodied AI)。具身智能不仅需要理解世界，还要能够在物理世界中感知、决策并行动，具备自适应能力，在不断变化的环境中自主学习。如何构建具备通用学习能力的具身智能，将成为下一代人工智能发展的关键。

(2) DeepSeek: 推动中国 AGI 领域自主创新

DeepSeek 的崛起代表了中国在大语言模型领域的一次重要突破。它不仅和技术层面挑战了国际科技巨头，也在战略层面打破了中国对西方技术的依赖。DeepSeek 推出的大模型（如 DeepSeek-R1）在自然语言的复杂推理中展现了卓越的性能，特别是在中文语境下，部分表现甚至优于国际竞品。这一技术突破标志着中国在 AGI 领域的崛起，打破了过去在关键技术领域的依赖局面。

DeepSeek 在产业升级方面也作出了重要贡献。通过自主研发底层框架和训练方法，DeepSeek 不仅降低了中国在人工智能技术上的外部依赖，还推动了中国 AI 产业的全面升级。现在 DeepSeek 的技术已经深度渗透到金融、医疗、教育、制

造等多个行业，赋能这些行业的数字化转型并提升效率。以金融领域为例，DeepSeek 的模型在金融风控中提供了更精准的数据分析，帮助企业更好地预测风险并做出决策。

在国际竞争中，DeepSeek 增强了中国在全球 AI 话语权争夺中的筹码。随着中国在人工智能领域的崛起，DeepSeek 让中国能够与美欧等科技大国平起平坐。通过技术输出、开源模型等方式，DeepSeek 为中国争取到了在 AI 伦理、安全等领域的更多话语权，逐步扭转了国际 AI 标准长期由西方主导的局面。

（3）具身智能：人工智能从虚拟世界走向物理世界

尽管 DeepSeek 在语言智能领域取得了突破，但人工智能的终极目标远不止于此。具身智能是下一阶段发展的关键，它的目标是让人工智能不仅局限于基于共性知识的信息处理和语言推理，还能够像人类一样感知和适应陌生环境，理解新任务，并做出自主决策。具身智能的挑战在于如何让 AI 具备理解、感知、推理、决策和执行的能力，同时能够在复杂、多变和非共性的物理环境中自主学习并适应新的任务。

例如，工业机器人不仅需要理解不断调整的生产任务，还需要在变化环境中持续改进技能；自动驾驶系统不仅要处理一些常识范围内的问题，还需要处理长尾的，超越共性知识的复杂突发问题，这也正是目前自动驾驶落地最大瓶颈之一。因此，具身智能要求 AI 系统具备比语言模型更高的复

杂度，包括本体异构、长时记忆、实时决策和实时自主学习。当前的强化学习方法虽然在多个领域取得了成功，但在具身智能的实现过程中仍然面临许多挑战。传统强化学习方法依赖大量的训练数据，并且通常只能在固定任务环境下进行学习，缺乏对未知任务的泛化能力。同时，传统模型往往依赖梯度优化，导致它们在任务切换时容易发生“灾难性遗忘”，无法像人类一样进行终身学习。要突破具身智能的瓶颈，必须构建出既能适应当前任务又能在新环境中进行自我进化的智能体。

（4）上下文强化学习：具身智能的发展方向

随着具身智能的不断发展，上下文强化学习（**In-Context Reinforcement Learning, ICRL**）为这一领域提供了新的研究方向。传统的强化学习方法通常依赖于特定环境中训练好的策略，这些策略对于任务和环境变化的适应能力较差，限制了智能体的泛化能力。在面对新任务或变化环境时，传统强化学习方法往往需要重新进行大量训练，而不能像人类一样快速适应。

上下文强化学习正是为了解决这一问题。ICRL 通过结合上下文信息（如环境的动态变化或外部输入）来调整学习策略，使得智能体能够更加灵活地应对任务变化和环境的不确定性。这种方法不仅可以在已知环境中优化策略，还能在新的、未知的环境中迅速调整和学习，而无需从头开始训练。

在具身智能的背景下，上下文强化学习能够帮助智能体在多样的物理环境中做出适应性决策，增强了智能体的灵活性和自主学习能力。

OmniRL 进一步突破了上下文强化学习效果 and 泛化能力的上限。OmniRL 基于马尔可夫决策链假设生成产生了百万级别的环境和任务，以及百亿时间步的强化学习过程数据。利用大规模元训练（Meta-training）方法，OmniRL 使智能体不再只针对特定任务或一类任务，而是学会在广泛任务中“如何学习”的能力，突破了传统强化学习方法的限制。这使得 OmniRL 能够以很好的表现适应完全没有训练过的强化学习任务。

OmniRL 另外一个核心贡献是证明了任务多样性而不是数据量对于智能体学习能力的重要性。OmniRL 证明随着任务数量增加，模型能够从掌握任务本身，转变为掌握更“如何学习”的能力，这种能力对比前者泛化性更强。同时还证明了，这种自主学习能力，对于上下文长度会提出更高要求，从根源上证明了长序列推理的重要性。为了区分以掌握任务本身为目标的“预训练”方法，OmniRL 凸显了以“如何学习”为目标而不以掌握具体任务为目标的“元训练”的潜力。

OmniRL 采用的大规模高效线性注意力机制，提升了模型在长序列任务中的表现和长期记忆能力，使得智能体能够

像人类一样在长期交互过程中进行学习，并具备更强的环境适应性。

尽管 OmniRL 在研究领域展现出较高的潜力，但它仍处于探索阶段，尚未实现大规模应用。我们应当保持谨慎态度，认识到具身智能的成熟仍需要克服诸多技术挑战，特别是在复杂高维系统、本体异构、实时决策等方面。

（5）具身智能的应用前景与行业变革

具身智能的广泛应用将带来深刻的行业变革，特别是在智能制造、自动驾驶、智能医疗等领域。具身智能能够使工业机器人在生产线自主适应不同任务，提升生产线的柔性化程度，减少对人工干预的依赖。例如，在自动驾驶领域，OmniRL 的自主学习能力将有助于系统适应全新的交通规则，陌生的室外和停车场环境，以及完全不同的硬件条件。在智能医疗领域，具身智能可以辅助医生针对个体病患提供长期跟踪的，精准的个性化健康管理服务。

此外，具身智能将在教育、物流、农业等多个行业带来革命性的变革。它不仅提升了生产力和效率，还能够改变传统行业的运作模式，推动产业升级和新经济模式的形成。通过具身智能的应用，企业能够实现更加智能的自动化生产，提供更加高效的个性化医疗方案，以及更灵活的城市交通管理。

（6）政策建议：开源发展、标准化和产业化落地

随着全球对具身智能技术的日益关注，中国在这一领域的加速发展需要政策层面的强力支持。为确保中国在全球 AI 竞争中的领先地位，政策可以从以下三个方面着手，推动具身智能的技术突破、产业化应用和社会融合。

一是支持开源发展，促进技术创新和全球合作：开源是加速具身智能技术创新的关键。政府应鼓励科研机构和企业积极参与具身智能的开源项目，推动核心技术、算法和平台的共享与合作。通过资助开源平台建设和资源共享，政府能促进技术创新、降低研发成本，并增强中国在全球 AI 开源社区的影响力。

二是制定具身智能标准，确保技术可控性和安全性：具身智能的快速发展离不开标准化的引导。政府应主导制定符合中国需求的具身智能技术标准，以涵盖数据安全、伦理规范和行业应用等方面。这不仅能保证技术的可控性、可靠性和安全性，还能增强中国在国际标准化工作中的话语权，推动全球技术竞争中占据主动地位。

三是推动具身智能的产业化应用：政府应通过专项资金支持和税收优惠政策，推动具身智能技术在智能制造、医疗健康、自动驾驶等行业的应用落地。通过设立示范基地和产业园区，加速技术从实验室到市场的转化。同时，应鼓励跨

行业合作，推动实际应用的试点和商业化，帮助企业降低技术应用的风险和成本。

(7) 结论：把握技术机遇，领导未来的全球 AI 竞争推动各行业发展的核心驱动力。政策制定者应当高度关注具身智能的发展趋势，抓住这一战略机遇，加速技术的进一步突破和产业化应用，确保中国在全球人工智能竞赛中的长期领先地位。

未来，具身智能的广泛应用不仅将改变人类的工作和生活方式，还会成为全球科技竞争中的新方向，推动人工智能技术发展进入一个全新的时代。

(8) 本文作者

王凡：现任深圳市人工智能与机器人研究院（AIRS）具身智能中心研究员，在人工智能产业和研究方向上有超过 15 年经验，曾获得吴文俊人工智能科技进步奖特等奖。

刘少山：现任深圳市人工智能与机器人研究院（AIRS）具身智能中心主任，ACM 科技政策委员会成员、IEEE 国际设备和系统路线图（IRDS）机器人计算方向主席。研究方向为具身智能、计算系统、科技政策。

3. 专家观点 | 鲁传颖：人工智能重塑国家安全的范式和逻辑

文 | 同济大学政治与国际关系学院教授，网络空间国际治理研究基地常务副主任 鲁传颖

人工智能（AI）作为一项具有战略性、通用性和军民两用性特点的技术，其与国家安全之间的复杂关联愈发引人关注。历史经验表明，战略性技术如核武器、太空技术和互联网，往往与国家安全密不可分，它们不仅推动了技术的进步，也在军事、情报和政治领域发挥了巨大作用。乌克兰危机中，人工智能技术在认知战、情报战、无人机、无人作战车辆、巡飞弹、卫星图像识别、决策辅助和网络对抗等领域的广泛应用，凸显了其在国家安全中的重要地位。同时，人工智能的通用性特征使其成为类似电力般的革命性技术，能够在医疗、金融、制造、教育、交通等多个领域带来创新和优化。这种广泛的应用潜力使得人工智能不仅在民用领域大有作为，而且对军事应用也产生了深远影响。与此同时，人工智能的军民两用性也导致其在国家安全中出现泛安全化问题。

随着人工智能技术的不断发展，国家安全进入了数字化和智能化时代，技术安全风险也在深刻塑造和改变国家安全的性质与特点。因此，我们需要深入思考人工智能将如何塑造未来的国家安全，并探讨国家安全治理体系的变革趋向。

（1）人工智能赋能国家安全

人工智能作为一项革命性的通用技术，正以前所未有的速度和广度渗透到国家安全的各个领域，对传统的国家安全逻辑和政策版图带来了全面而深刻的影响。人工智能不仅通过其强大的技术赋能，改变了国家安全的内容和政策实践，

使得传统的安全观念与治理策略面临前所未有的挑战；同时，它还通过重塑战争形态，颠覆了人们对国家安全认知的传统框架，使国家安全的内涵和外延发生了重大变化。

随着私营部门在人工智能技术领域的崛起，越来越多的尖端科研成果出自企业之手，而非传统的国家实验室。这一变化不仅为国家安全带来了新的机遇和活力，也造成了诸多复杂而棘手的问题。因此，深入探讨人工智能对国家安全的影响，不仅有助于我们更好地理解这一技术革命带来的挑战和机遇，更有助于我们为国家安全战略和政策制定提供科学依据和有力支撑。

1) 人工智能重塑公民国家安全认知：战争形态与信息战

随着人工智能技术的飞速进步，未来的战争模式正酝酿着一场根本性的变革。传统的人力密集型军事任务，如前线侦察、物资运输及直接火力支援等，正逐步被高度智能化的机器人与无人机系统取代。这一转变不仅显著降低了军事行动中的人员伤亡风险，还极大地提升了作战的效率与精确打击能力，使得战争形态向更高效、更精准的方向发展。掌握先进人工智能技术的国家在信息战中占据显著的优势地位。人工智能技术凭借其强大的数据处理与分析能力，能够迅速识别敌方的战略弱点与作战意图，为决策层提供精确、及时的情报支持，从而在战争中赢得先机。此外，人工智能技术还通过模拟与预测战争进程，为军事战略与战术的制定提供

科学依据。通过构建复杂的战争模型，人工智能可以模拟不同战略选择下的战争发展态势，帮助指挥员作出更为明智的决策。这种基于数据的战争规划方式，无疑将极大地提升战争的胜算。2022年乌克兰危机作为近年来全球范围内爆发的重大冲突之一，被普遍认为是首次大规模应用人工智能技术参与作战的现代战争。在这场冲突中，双方不仅利用社交媒体平台传播虚假信息，以干扰对方的士气与决策，还借助先进的卫星图像分析软件，实时追踪敌军动向，为精确打击提供有力支持。这些实例充分展示了人工智能技术对未来战争形态的深刻影响，预示着在未来的战争中，人工智能技术将发挥越来越重要的作用，成为决定战争胜负的关键因素之一。因此，应高度重视人工智能技术在军事领域的应用与发展，加强相关研究与投入力度，以确保国家安全。

2)人工智能全面赋能国家安全能力：情报分析能力跃升

在国家安全这一关键领域内，人工智能技术的应用正逐步深化，其影响力从最初的简单分类任务扩展至复杂的决策支持过程。在情报分析的前沿阵地，人工智能技术以其卓越的数据处理能力，对源自社交媒体、新闻报道、卫星图像等多维度的海量数据进行深度扫描与智能分析，能够精准捕捉异常活动迹象，及时预警潜在的安全威胁。这一能力的跃升，不仅极大提高了情报分析的精确度与时效性，更为决策层提供了全面、准确的信息。随着生成式人工智能的突破，通过

高效整合文本、图像、音频等多模态信息，实现了内容信息生产方式的变革，为国家安全战略的实施提供了更为灵活多变、全面深入的支持框架。

3)人工智能打破国家安全资源垄断：私营部门的角色与国际合作

在全球范围，人工智能不仅引发了军事和情报领域的革命，更打破了政府对于国家安全的垄断。长期以来，政府作为国家安全事务的主导者乃至唯一合法参与者，掌握着安全领域的先进技术、威胁情报、高端人才等资源。然而，随着人工智能技术的迅猛发展，这一传统格局正经历着前所未有的变革。科技企业在人工智能技术领域展现出强大的研发实力与创新能力，成为推动人工智能技术进步不可或缺的力量。大型国际科技企业的崛起，无疑对国家安全格局产生了深远且复杂的影响。从积极层面看，科技企业的参与为国家安全注入了新鲜血液，带来了更多的技术创新与支持，有助于提升国家安全的整体效能。然而，私营企业的利益诉求与价值取向往往与国家安全目标存在潜在的冲突与矛盾，如何在保障国家安全的同时兼顾私营企业的合理利益，已成为亟待解决的重要议题。面对这一挑战，国际社会开始积极探索建立更为开放、包容的合作框架，以期在人工智能技术与国家安全之间找到平衡点。与此同时，各国政府也在积极寻求与私营企业的深度合作，通过政策引导、资金扶持等多种手段，

鼓励私营企业积极参与国家安全领域的研发与应用。这种政府与企业携手并进的合作模式，不仅加速了人工智能技术的迭代升级，还提升了国家安全的整体水平与应对能力，为构建更加稳固的国家安全体系奠定了基础。

（2）人工智能技术安全对国家安全的深层次影响

赋能国家安全仅为人工智能的影响之一，其更为深刻且广泛的效应在于人工智能自身所蕴含的技术安全性对国家安全本质与特性的重塑。随着人工智能与国家安全的全面深度融合，人工智能技术中潜藏的安全风险亦随之渗透至国家安全的各个维度与环节，形成了一种复杂的映射关系。若缺乏对人工智能技术安全性的深刻理解，便难以准确把握国家安全所面临的新态势与新特征。

人工智能技术安全（AI Safety），其核心在于确保人工智能系统自身的安全性与可信赖性。基于模型算法的人工智能系统，面临着诸如“黑箱操作”、算法偏见、恶意诱导、认知幻觉、系统操纵等多重风险，直接关乎人工智能是“向善”还是“向善”，是稳定运行还是走向失控。从技术治理的视角出发，增强算法模型的可解释性、透明度及鲁棒性，是构建人工智能安全可信框架的关键路径。然而，技术治理并非“万能钥匙”，无法全然消解人工智能系统面临的所有风险，原因在于人工智能技术与风险之间存在着一种共生共存的复杂关系。

更深层次地，人工智能在全面、颠覆性、革命性地赋能国家安全的同时，也驱动了国家安全范式的深刻转型——从追求确定性因果机制的传统模式，转向应对不确定性相关性机制的新范式。在这一范式转换的过程中，国家安全正面临着风险来源多样化、系统安全脆弱性加剧、治理边界模糊化、威胁认知不确定性增加等一系列挑战。这些挑战要求我们在享受人工智能带来的安全与效率提升的同时，必须深刻洞察并妥善应对由人工智能技术发展所引发的国家安全新变局。

一是风险来源多样化。在传统国家安全领域，能够威胁到国家安全的风险来源相对明确。这使得国家只需要将有限的资源投入相对重要的领域就可以在最大程度上维护国家安全。人工智能技术广泛应用所带来的影响之一便是加速了国家安全风险来源的多样化。在传统国家安全框架中，军事威胁占据主导地位，而随着人工智能技术的渗透与融合，网络攻击、信息战、经济间谍活动等新型威胁日益凸显，构成了对国家安全的复杂挑战。

人工智能技术为网络攻击提供了更为强大的工具与手段。相较于传统的人工操作，人工智能技术能够实现自动化、智能化的网络攻击，极大地提高了攻击的效率与规模。此外，人工智能技术在信息战中的应用，使得虚假信息的制造与传播变得更加容易与逼真。深度伪造技术（**deepfake**）便是其中的典型代表。通过先进的机器学习算法，深度伪造技术能够

生成高度逼真的虚假视频与音频，从而误导公众，破坏社会稳定。不仅如此，人工智能技术还可能被用于经济间谍活动，窃取国家重要经济数据与商业机密。攻击者可以利用人工智能算法对海量数据进行深度挖掘与分析，从中获取有价值的信息与情报，进而对国家经济利益造成损害。

二是威胁认知不确定性。人工智能系统的“黑箱”特性增加了威胁认知的难度。许多先进的人工智能模型，尤其是深度学习模型，其内部决策机制复杂且难以解释，导致国家安全机构在理解人工智能系统行为、预测其潜在风险时面临巨大挑战。这种不透明性不仅限制了国家安全机构对人工智能系统安全性的全面评估，还可能为恶意攻击者提供隐蔽的操作空间，进一步加剧威胁的不确定性。对于关键的国家安全设施如核控制系统，在与人工智能系统相结合后不透明性增加，使得核控制系统在关键时刻的决策过程难以预测和解释，这可能导致在紧急情况下引发对系统决策的质疑，进而影响整体的国家安全。

面对人工智能技术带来的威胁认知不确定性，国家安全策略的制定与实施需进行相应调整与优化。一方面，国家安全机构应加强对人工智能技术的深入研究与持续监控，建立快速响应机制，确保能够及时识别并应对新出现的威胁。这包括投资研发先进的检测工具、建立跨部门协作机制以及提升人员的技术能力等。另一方面，国家安全策略应更加注重

灵活性与适应性，以应对人工智能技术带来的不确定性。这意味着需保持对人工智能技术发展趋势的敏锐洞察，不断调整策略方向与重点，确保国家安全体系能够有效应对不断变化的威胁环境。

三是系统安全脆弱性。人工智能系统的复杂性与高度互联性如同一把“双刃剑”，在赋予系统强大处理能力的同时，也无形中扩大了安全漏洞的暴露面，为潜在的安全威胁提供了可乘之机。人工智能系统的复杂性、快速迭代性、对数据的高度依赖以及决策过程的不透明性，共同构成了其安全脆弱性的核心要素。面对这一挑战，需加强人工智能系统安全标准的制定与执行，推动安全技术创新，提升系统的鲁棒性与自恢复能力，同时加强数据保护与治理，确保人工智能技术的健康发展，为维护国家安全与社会稳定奠定坚实基础。

人工智能系统的“机器幻觉”进一步增加了安全的脆弱性。人工智能系统往往由多层算法、模型及海量数据交织而成，不仅容易出现稳定性差的问题，还经常会出现“幻觉”，包括一些情况下“一本正经的胡说八道”。对此，探索“机器幻觉”出现的原因以及如何对其矫正需要耗费大量的时间，并且难以保证其不会再次出现类似情况。在系统自身易出现幻觉的情况下，如何预防人工智能应用过程中出现的风险成为了高难度的挑战。

人工智能系统对数据的高度依赖性，构成了其安全脆弱性的另一重要维度。数据作为人工智能系统的“血液”，其质量与完整性直接关系到系统决策的准确性与可靠性。然而，数据在收集、存储、处理及传输过程中，面临着被篡改、污染或窃取的风险。一旦训练数据遭到恶意操纵，人工智能系统便可能基于错误的信息作出决策，从而引发严重的安全问题。例如，在自动驾驶汽车领域，若训练数据被故意篡改，车辆可能无法正确识别道路标志、行人或其他车辆，进而在实际行驶过程中发生碰撞事故，不仅危及乘客生命安全，还可能对公共安全构成重大威胁。

四是治理边界模糊性。随着人工智能技术的迅猛发展，其跨领域、跨行业的广泛应用正逐步重塑着国家安全的治理框架，其中最为显著的影响之一便是治理边界的模糊性。这一模糊性不仅体现在国家安全治理的传统领域与新兴领域的交织上，还深刻影响着国际法与国内法应对人工智能相关安全挑战的适用与协调。

在传统意义上，国家安全治理主要聚焦于军事防御与情报收集，这些领域有着明确的组织架构、法律基础及操作流程。然而，人工智能技术的广泛应用打破了这一界限，使得国家安全议题不再局限于传统的硬实力范畴，而是渗透到经济、社会、科技等多个维度。例如，在金融领域，人工智能算法被用于风险评估、市场预测等，其决策过程与结果直接

影响到国家经济的稳定与安全；在医疗领域，人工智能辅助诊断、药物研发等虽提高了医疗服务效率，但也带来了患者隐私泄露、医疗数据安全等新的公共卫生安全问题。这些跨领域的应用使得国家安全治理的边界变得模糊，传统建立在科层制基础之上的治理机制，由于不同部门负责专门事务，难以有效覆盖所有潜在风险，甚至出现治理失灵。

（3）人工智能时代的国家安全治理体系

各国政府普遍对人工智能所带来的广泛机遇与复杂挑战给予了高度重视，并据此制定了一系列相关的国家战略、政策举措，同时设立了专门的治理机构以应对这一新兴领域的独特需求。在国际安全层面，无论是联合国等多边国际组织，还是诸如北大西洋公约组织、上海合作组织等区域性组织，均建立了相应的治理机制，旨在规范和引导人工智能的健康发展。然而，现有的治理机制主要基于国际社会在核军控、大规模杀伤性武器军控等传统安全领域的知识和经验，在应对人工智能所带来的风险来源多样化、系统安全脆弱性、治理边界模糊性、威胁认知不确定性等新型挑战时，显现出较强的不适应性。这种不适应性若未及时关注和有效解决，极易导致泛安全化和过度安全化等负面现象的出现，不仅会大幅增加社会的整体成本，还可能对人工智能技术的创新与发展产生抑制作用。

因此，国际社会亟需从理念层面、机制构建、方法创新以及议程设置等多个维度出发，重新构建一套适用于人工智能技术风险的国家安全治理体系。这一体系应充分考虑到人工智能技术的独特性与复杂性，以及其对国家安全、社会稳定和全球经济发展的深远影响，从而确保在保障国家安全的同时，也能促进人工智能技术的健康、可持续发展。

首先，在治理理念层面，应构建一个基于人工智能安全风险全面认知框架。人工智能为国家和社会带来的安全风险具有长期性、持续性特征，且不存在一劳永逸的解决方案。这些风险主要源自三个维度：技术本身的缺陷与不完善性、技术的滥用及不负责任的使用，以及社会应对机制的缺失。针对前两者，技术治理与监管机制的逐步建立已成为趋势；然而，关于社会如何有效应对人工智能时代风险的探讨仍显不足。从风险韧性社会的视角出发，尽管人类无法根除自然灾害，但可通过多种措施减轻其影响。同理，技术缺陷、技术滥用与韧性社会之间应建立一种良性互动关系。过度聚焦于前两者可能导致追求绝对安全与过度治理的困境。因此，构建人工智能风险韧性社会需从提升公众与组织的技术素养、建立风险共担的社会保障体系等方面着手，以实现技术进步与社会发展的和谐共生。

其次，在治理机制层面，应着重强调技术标准、法规政策与战略举措之间的协同与整合。人工智能作为一项前沿技

术，其治理体系必须具备一定的前瞻性与适应性，以确保在技术生命周期的各个阶段都能实施有效的监管和引导。技术标准作为治理的基础，应确保其具有科学性、合理性和可操作性，为人工智能技术的研发和应用提供明确的指导和规范。同时，法规政策应与技术标准紧密衔接，通过立法手段明确人工智能技术的合法边界，保护个人隐私和数据安全，防止技术滥用和恶意攻击。

战略举措则是实现技术标准与法规政策协同的关键。政府应制定长远的人工智能发展战略，明确技术发展的方向和目标，为技术标准和政策法规的制定提供宏观指导。此外，战略举措还应包括建立跨部门、跨领域的协作机制，加强政府、企业、科研机构和社会组织之间的沟通与合作，共同推动人工智能技术的健康发展。在此过程中，敏捷治理成为一种重要的方式。面对人工智能技术的快速迭代和广泛应用，治理机制应具备灵活性和快速响应能力，能够及时调整和优化治理策略，以适应技术发展的新趋势和新挑战。通过敏捷治理，可以确保治理机制与技术发展的同步性，提高治理效能，促进人工智能技术的可持续创新与发展。

再次，在治理方法维度，应构建政府、产业界及研发机构高效协同机制。人工智能的迅猛发展已促使国家安全领域从传统安全及非传统安全范畴，逐步迈向数字安全。在此背景下，企业与技术社群在国家安全体系中的地位日益凸显，

亟需建立更为高效、灵活的沟通协调机制。传统的自上而下、等级化及中心化的国家安全技术、知识与信息传递模式已难以适应当前形势，如何更有效地与大型科技企业及技术社群加强联系，促进政府与企业间人员的流动与交流，成为政府重塑人工智能安全领域权威性的关键所在。以美国为例，其人工智能安全治理主要依托人工智能安全研究所进行。与传统政府机构不同，该研究所人员精简，主要工作通过与企业合作完成，其核心职能在于搭建政府与企业、企业与企业之间的沟通桥梁。这种定位使研究所更注重成为人工智能安全治理的核心节点，虽不具备传统政府机构的命令发布权，却能占据信息与知识的中心位置，进而提升政府在该领域的权威性与影响力。

最后，在议程设置层面，需深入关注人工智能时代国家安全所蕴含的丰富内涵，这包括但不限于生存性风险、算法模型风险、应用型风险以及广义上的国家安全风险。作为人工智能时代最为严峻的挑战之一，生存性风险涉及技术失控、系统崩溃等极端情况，可能对人类社会造成不可逆转的损害。因此，议程设置应优先考虑如何建立有效的风险预警与应对机制，确保人工智能技术的健康发展。算法模型风险则源于算法设计的不完善、数据偏差以及训练过程中的不确定性，这些因素可能导致算法决策出现偏差，甚至引发歧视、不公平等问题。在议程设置中，应加强对算法模型的监管与评估，

确保其公正性、透明度和可解释性。应用型风险主要关注人工智能技术在具体应用领域可能带来的安全隐患，如自动驾驶、医疗诊断等。议程设置需针对这些领域的特点，制定具有针对性的安全标准和规范，确保技术的安全应用。此外，广义上的国家安全风险也是议程设置不可忽视的一部分。人工智能技术的快速发展可能对国家主权、安全和发展利益产生深远影响。因此，议程设置应充分考虑如何平衡技术发展与国家安全的关系，确保国家在人工智能时代的战略利益。

4. 巴黎 AI 峰会，李飞飞教授开幕演讲：当前才是“首个真正的 AI 时代”

在全球瞩目的巴黎 AI 峰会开幕式上，斯坦福大学李飞飞教授发表生动的演讲。作为 ImageNet 创始人，她的每一次公开发声都格外引人注目。这一次，她带来了一个跨越五亿年的宏大叙事：从生命最初获得感知能力的进化史，讲到如今 AI 的突飞猛进。“毫无疑问，历史学家今后一定会把这段时间称作‘真正的第一个 AI 时代’。”李飞飞在演讲中如此断言。在她看来，从任何维度来看——无论是计算机专业中人工智能方向的学生人数、投资金额、新创企业数量，还是技术突破——AI 都正在掀起一场深度和广度空前的革命。

以下是演讲原文：

各位早上好，我非常荣幸且激动能够在本周与大家共聚一堂，并带来此次开幕演讲。我也非常期待接下来最充满活力的讨论，甚至可能做出一些重要决定。

（1）人工智能的历史与起源

对我们许多人而言，人工智能往往被视为一个关于技术、关于现代世界的故事，也毋庸置疑是一个关于未来的故事——这正是我们大家今天齐聚于此的原因。然而对我而言，它还是一个可以追溯到所有生命起源之初的故事，早在五亿年前就已经开始。那是一个极其久远的时代，当时连“视觉”这一概念都尚未出现，实际上连眼睛都尚未进化，没有任何生物曾看过这个世界，所有生命都被置于你现在在屏幕上所看到的那片黑暗当中。当然，要在十分钟左右的时间内讲完五亿年的历史确实有点勉为其难，所以我给大家带来一个简要版：当进化赋予了你们在屏幕上所见的这些简单生物以感知周围世界并对这些感知（即便十分微弱）做出反应的能力时，一场进化军备竞赛由此展开。最初只是一种被动的体验——让光线进入体内，但不久之后就变得更加丰富并富有主动性。神经系统开始进化，“看见”开始转变为“洞察”，看见进而变为理解，而理解则推动了行动，所有这些都孕育了“智能”，从而永久地重塑了地球上生命的本质。时至今日，五亿年之后，人类的智能已经让我们的物种可以想象并塑造工作和生活的方方面面。而我们也不再满足于仅仅拥有自然

赋予的这份智慧。我们怀着的好奇心如今驱使我们想要创造出与我们同样聪明、甚至更胜一筹的机器。这也是为什么我们本周要讨论的这项重要技术，其起源可以追溯到 20 世纪中叶。英国的伟大数学家艾伦·图灵（Alan Turing）极具前瞻性，早在计算机尚未真正出现之前，他就已经开始思考如何让它们拥有可与人类比肩的认知能力。对我而言，他的著作一直在向人类发出挑战：要大胆设想能够思考的机器，就像他自己所想象的那样。这股好奇心与雄心同样也感染了美国的早期计算机科学家们，他们不仅举办了首个研究项目来探索“会思考的机器”的可能性，还在 1956 年那个炎热的夏天正式提出了“人工智能”这一术语——比这个概念进入公众视野要早几十年。这张图片正是当时他们为那次研讨会撰写研究论文截图。说实话，令人觉得有趣的是，他们当时以为只需要两个月就能解决这一重大课题的很大一部分，甚至能破解智能的奥秘。也许他们太过乐观，但不得不佩服他们的胆识。如今，我们已经在那个“两个月项目”里走过了 820 个月，但确实也取得了不少实质性进展。

（2）现代 AI 三大关键要素

另一个对 AI 的误解在于，认为它只是计算机和工程领域的事情，然而事实上，它一直以来都是一门高度跨学科、富有活力的追求。我们当今所处的现代 AI 时代，正是三项极其重要又各自独立的技术与科学进步汇聚而成的成果。第

一项是对感知算法的研究——即生物（包括动物乃至人类）如何理解周围世界，并最终创造出让机器也能具备类似能力的数学模型。在达特茅斯 AI 夏季研讨会上延续了图灵的大胆假设之后，神经生理学家 David Hubel 和 Torsten Wiesel 率先揭示了哺乳动物视觉皮层中神经处理的层次结构，为他们赢得了诺贝尔奖，也彻底改变了我们对视觉处理的认知。大约在同一时期，心理学家 Frank Rosenblatt 搭建了最早的神经网络原型之一——感知器（Perceptron）。这项工作启发了数十年后无数计算机科学家，尤其是早期先驱者，如 Kunihiro Fukushima、Jeff Hinton、Yann LeCun、Yoshua Bengio 等，去设计愈发先进的模型，最终形成了我们今天所熟知的深度学习神经网络算法，让机器拥有了自身近乎神奇的感知和分析能力。与此同时，第二条研究脉络也开始出现——认知科学家转而研究人类自身的大脑，揭示我们对周围环境进行感知时所拥有的惊人深度和复杂性。他们的研究明确指出，无法将我们的脑与其所处的进化环境分割开来。大脑不过是一个被禁锢在颅骨黑暗之中的机器，却又从生命最初时刻开始，就急切地向外部学习，通过每一次光线的照射、每一次触觉刺激、每一声细语来获取信息。对我个人而言，作为在 21 世纪初成长起来的科学家，当时感知算法几乎是我所在领域的唯一关注点，而认知科学带给我的启示在于“规模”。演化与发育能如此有效地塑造人类智慧，是因为它们能利用大量数

据来驱动我们的学习。我和我的合作者、学生们推测，同样的规律也适用于机器，只不过这一次，数据来自现代数字设备和互联网，而不再是依靠生物感官。由此激发了我所在实验室发起的 ImageNet 项目——它是第一个针对人工智能的大规模互联网训练与评测数据集。我们提出了一个关键假设：数据是让神经网络等高容量算法真正“活起来”的关键。这个理念让神经网络焕发新生，开启了利用海量数据来推动 AI 进步的全球趋势，也成为如今所说的 AI “规模定律” (scaling law) 的一部分。最后，如果没有极其强大的运算能力来支撑，以上所有成就都不会发生，或者根本连起点都无法跨越。最初的里程碑包括冯·诺依曼 (von Neumann) 在上世纪 40 年代提出的计算机体系结构理论，它一直沿用至今，并在 70 年代初催生了第一批微处理器。可有趣的是，真正让硅芯片原始运算速度获得飞跃的，竟是电子游戏。为追求更逼真的游戏画面，90 年代初诞生了一个小型产业，但在短短二十年内迅速膨胀为全球巨头——以英伟达 (NVIDIA) 等公司为代表，开发出了越来越强大的图形处理器 (GPU)。结果证明，这恰恰是让神经网络算法得以利用互联网规模大数据进行学习的最后一块拼图。所以，如果现场有游戏玩家，那我们要感谢你们。

（3）2012 年之后的 AI 突破

当然，后面的故事不仅仅是“历史”，更是建构未来的“配方”。2012 年，在 ImageNet 挑战赛上，我的实验室率先将算法、数据和计算这三大要素首次大规模结合在一起，几乎在一夜之间就改变了我们整个领域。那是机器首次能够理解并准确地描述海量图像——成百上千万张之多。如今我们对此已经习以为常，但在当时，这是一项前所未有的壮举，甚至有些像科幻小说里的场景。它就像推倒了第一块多米诺骨牌，随后的一系列突破接连不断，并且来得越来越快。如今，十多年过去了，我们依然在探索这一切的意义。当初只是学术界的一个好奇心驱动的研究，如今却年复一年地吸引着商界领袖、创业者、行业分析师，甚至政治家们的热切关注。如今，十多年过去了，我们依然看不到尽头。毫无疑问，历史学家今后一定会把这段时间称作“真正的第一个 AI 时代”。从任何一个你能想到的指标来看——无论是计算机专业中人工智能方向的学生人数、投资金额、新创企业数量，还是其他方面——AI 都是一场在深度和广度上不断扩大的革命。我想在座的各位都不需要我来重复，过去几年里，人工智能又迎来了更惊人的跃进——也就是大型语言模型的出现。这些大型语言模型把现代 AI 所依赖的三大要素——算法、数据和计算力——都推向了更极致的规模。它们采用了一种名为“Transformer”的新型模型架构，训练数据几乎覆盖了整

个互联网，并在数量惊人的最新、最强大芯片上运行。正如我们所见，其结果是在机器能力上比过去十年来的任何一次突破都更为迅猛。现在我们几乎将“AI 可以用自然语言流畅地跟我们对话”视为理所当然，AI 能回答我们提出的几乎任何领域的问题，甚至能够生成各种复杂形式的内容——从图像、声音、音乐到视频，无所不包。ChatGPT 所取得的惊人成就就是这项创新如何影响我们日常生活的最佳例证，因其创造了用户采用速度的历史新纪录。而且，这些能力并不只是表面上看着“会说话”而已。若你看看这张图表，会发现 AI 模型在各种基准测试（从手写识别到博士水平的科学问答）上的表现近几年都在飞速攀升；其中一些难度极高、对人类来说具有挑战性的任务，AI 的成绩简直像坐火箭一样往上蹿，几乎呈垂直上升的趋势。

（4）大模型与近期 AI 进展

不仅如此，从能够流畅使用语言开始，大型语言模型还在朝更具“主动性”的方向发展——它们学会了将复杂任务分解成若干步骤，并逐步规划如何实现目标。人们目前将这种趋势称作“具备代理能力（*agentic*）的 AI”。而在 2025 年，这似乎正成为这场深远技术演进的新篇章。对许多用户和企业而言，这些能力已经非常实用，而它们还远不止于此，后续的发展潜力更是不可估量。我们目前所讨论的还大多停留在“语言智能”层面，但人类之所以为人，是因为我们拥有

更全面、更完整的智能。从“被动感知”到“主动行动”的新阶段正在到来。在我自己的研究领域——涉及摄像头和机器人时，我看到 AI 正逐渐具备“创造、理解、推理和交互”这些维度，进而在数字或物理的三维空间里与人和环境相互作用。我们称之为“空间智能 (spatial intelligence)”和“具身智能 (embodied intelligence)”。举个例子：假设你看到这样一张图片，视觉智能很容易让我们识别出里面的所有物品：那只猫、那盆植物、那张桌子，以及那杯牛奶。但是，这就代表我们感知系统的全部功能了吗？我敢打赌，其中不少人不仅仅“看懂了”那张图，还会产生“哎呀，快伸手把那杯牛奶抓住，别让它掉到地上打碎了！”的冲动。这只是个小例子，但它清楚展示了“从观察到行动”之间多么巨大的差别，也说明了我认为我们如今所处的拐点：AI 正从观察者变成与我们并肩行动的“执行者”。实际上，就连我一直崇拜的图灵先生，他对于未来的预见在这一点上也可能有些“过于狭隘”或“向内看”了。因为进化本身已经一再告诉我们：智能的真正强大之处，不仅仅在于“思考”，而在于让思考能驱动行动。想想看，人类的空间智能让我们得以建立灿烂的文明：从古老的金字塔到工业革命，从科学发现到艺术创作。如今，当 AI 逐渐拓展我们的空间和具身智能，这将如何进一步改变我们与周遭世界的关系？它又能帮助我们创造和发现些什么？可以预见，我们已经在打造某种面向未来的新形态了。

（5）空间与具身智能的兴起

接下来，我想给各位展示一些“空间智能”技术迅猛发展的例子。这四个案例来自我在斯坦福的学生和与 World Labs 同事所做的工作：左上角是对复杂日常视频的语义标注，右上角是艺术风格迁移，左下角是通过文本提示生成视频的算法，右下角则是让任何图像都能变成一个 3D 世界。没错，你现在看到的就是梵高画作中那家法国咖啡馆，被我们用生成式 AI 模型带进了一个想象中的三维世界。同样令人激动的还有机器人技术，也就是“具身 AI”。这两项出自我实验室的研究成果，将大型语言模型与视觉模型整合到机器人的学习中，使得机器人能在更开放、更贴近现实的环境下执行人类日常任务，而不仅仅是过去那种在高度编程和设定好动作序列的场景里活动。这些进展都蕴藏着巨大潜能。不过，如果 AI 真的不再只是“会思考的机器”，而是同时成为“会行动的机器”，那么我们整个社会在引导这项技术发展方向时，就承担着更加紧迫而重大的责任。我认为毫不夸张地说，这一切正将我们带到一个对整个人类文明至关重要的历史关头。我们要如何应对？我想，这也正是我们这周聚在这里的主要原因。

（6）以人为本的 AI 与三大核心价值

这个问题其实一直在指引我的研究工作。虽然我肯定，想要找到简单答案几乎不可能，但有一个主线在我多年的探

索中始终贯穿，那就是我所说的“以人为本的 AI (Human-Centered AI)”，并且它包含三个核心价值：尊严 (dignity)、能动性 (agency) 和社区 (community)。首先是尊严。当面对越来越强大的技术时，我们人类常常会反思：到底是什么在定义“我们是谁”？如果把我们能做的所有事情都一一剥离，剩下的、那份能够自主决策并亲自行动的尊严感，依然是我们作为人的本质之一。如果 AI 这项技术能帮助我们守护并且“找回”这份尊严，尤其是能帮助到最脆弱的人群，那将是让我最激动的事情。在这一点上，我想给各位看一个例子：如何让机器人和 AI 技术帮助那些严重瘫痪患者重新获得自主行动的可能。在这项斯坦福大学的研究中，我的学生利用完全“非侵入式”脑电波 (EEG) 记录获取使用者的思维信号，通过 AI 算法对这些信号进行解码并发送指令给机械手臂。最终，机械手臂在患者“意念”的控制下做出包括烹制日式寿喜烧在内的一整套烹饪动作。这便是 AI 帮助人们重新赢得自主与尊严的一个示例。其次是能动性 (agency)。我所在实验室多年来的核心理念，就是探索如何让 AI “增强 (augment)” 人类的能力，而不是取代人类。毫无疑问，每一次重大的技术变革都会对劳动力市场产生冲击，AI 也不例外。然而，我相信 AI 可以帮助我们从小创造力到医疗保健、从科研到制造业，在方方面面获得新的能力。因为在许多方面，AI 的技能与人类的技能其实是互补关系。我们

完全可以利用这个“数字或物理上的合作者”来强化自身能力。过去十年里，我的实验室一直在研究 AI 如何改进医疗服务，这让我见识到许多能让 AI 帮助提高医疗质量、同时减轻医护人员负担的机会。这里给大家展示三个例子：利用搭载 AI 的智能摄像头来协助医院监督医护人员的洗手规范，辅助记录病人复健时的移动训练情况，以及跟踪手术器械的使用状况。最后是社区（community）。过去十年里，科技也常常被诟病“把人们越拉越远”，催生了各种“信息茧房”和“狂热话题诱饵”等。如今，AI 又一次走到了十字路口：一条路是让 AI 主导很多社交体验，甚至为每个人极度定制并强化他们原本的偏见；另一条路则可能帮助我们构建更健康、更强大并具有包容性的社区，比如借助 AI 教育助手，让更多人（包括孩子和成人）获得学习的机会。这里我举两个简单案例：左图是利用 AI 和虚拟现实技术，为有阅读障碍的大学生提供个性化辅助工具；右图是美国水牛城大学开发的一个 AI 专家系统，用来缓解语言治疗师（speech-language pathologist）短缺的问题，帮助对 3 至 10 岁儿童进行早期语言干预。

（7）AI 治理：理性、务实与多方协作

以上这些例子引出了我今天想和大家分享的最后一个想法：在这样一个关乎人类文明走向的关键时刻，我们如何通力合作，以一种理性、务实且负责任的方式去治理 AI，从

而让它继续释放惊人的潜力？最起码，我们需要从以下几个方面开始：首先，务必要基于科学而不是科幻来进行治理。无论是大街小巷还是华尔街，关于 AI 的话题中都充斥着夸张炒作和耸人听闻的描述，往往导致对 AI 的治理政策被各种误导。相反，我们应该用更科学的方法来评估和测量 AI 的实际能力与局限，从而制定更加精准、可落实的政策，并且基于现实而非幻想。其次，我们需要采取务实而非意识形态化的态度来看待 AI 治理。AI 有望成为一种强大技术，如果使用得当，它确实能让我们更好地工作和生活。因此，我们不应该只顾在上游阻碍这项仍然非常年轻的技术进一步探索与研究；相反，我们应该更多关注它的具体应用场景，确保它能被善加利用，并对可能产生的不良后果加以防范。最后，我们必须投入更多资源，构建更健康、更蓬勃的 AI 生态系统，让学术界、创业者、开源社群以及公共部门都能与大企业一道参与进来、发挥关键作用。如果 AI 真的要改变世界，我们就需要各行各业、社会各个群体都能在这个变革过程中发挥作用、共同塑造未来。我在演讲开头提到过，现代 AI 之所以兴起，主要归功于算法、数据和计算力这三大要素的结合。但若这三种关键资源仅仅掌握在少数几家公司手中，AI 生态系统就会面临好奇心驱动的研究乏力、顶尖教育人才受限、开源项目缺少资源、多学科交叉探索不足等诸多问题。75 年前，艾伦·图灵对未来一瞥，发出了“人类是否有胆量

去构造能够思考的机器”的挑战。今天，我们已经把这个挑战推进到一个恐怕是他当年难以想象的高度。AI 时代的技术飞跃令人叹为观止，我相信，如今是时候提出一种新的挑战了：与其再问“我们能否创造 AI”，不如问“我们能否创造一股向善的 AI 力量”。简而言之，今天我想向各位发出邀请：让我们共同努力，去构建“以人为本”的 AI。

谢谢大家。

5. 人工智能驱动信息技术体系重构与再造——兼评《中国至 2050 年信息科技发展路线图》

洪学海 中国科学院计算技术研究所

《中国至 2050 年信息科技发展路线图》和《信息科技：加速人一机一物三元融合》，对我国信息科技领域的长期发展进行了预判。文章回顾了近 15 年来信息科技领域的发展变化，并与当初的预判进行了比较，指出了当初的一些重要判断符合预期。近年来，全球信息技术重大创新的整体节奏在放缓，而新一代人工智能快速演进，将为未来 10 年信息技术体系重构和再造带来新机遇，并将加速信息技术体系创新进程。文章对计算技术、数据空间、网络空间和智能空间的技术体系重构与再造进行了系统分析，指出了挑战性问题，并提出了发展对策。

2009 年，以李国杰院士为组长的中国科学院信息科技路线图专家组，出版了《中国至 2050 年信息科技发展路线图》

(以下简称《路线图》),并于2013年出版的《科技发展新态势与面向2020的战略选择》中发表《信息科技:加速人一机一物三元融合》(以下简称《三元融合》)。《路线图》和《三元融合》前瞻性、战略性分析了我国信息科技未来几十年发展的大趋势和战略重点,为当时我国信息科技发展方向和战略重点圈定了相关重点领域,对我国信息科技领域的重大科研活动起到了一定的指引作用。近15年来,全球信息科技发展已经发生了一系列重大的变化,尤其是近几年来,以大模型为代表的新一代人工智能技术极速发展,为信息技术体系重构与再造创新带来了重大的机遇,将加速驱动信息技术体系创新进程。因此,评估十几年前出版的《路线图》和《三元融合》中对信息科技相关战略预判的一些结论与影响,并在新形势下,进一步前瞻未来10年信息科技变化趋势,为抢占信息领域科技制高点,到2035年实现科技强国等战略目标,具有重要的现实意义。

(1)《路线图》和《三元融合》的重要预判与近15年发展的比较

回顾近15年来,信息科技发展变化与《路线图》和《三元融合》给出的预判,现在看来,很多当初的预判至今依然适用。

1)关于信息科学理论发展的长期预判符合预期

《路线图》认为：信息技术不会变成以增量改进为主的传统产业技术，而是面临一次新的信息科学革命。信息技术的基础理论大部分是在 20 世纪 60 年代以前完成的，近 40 年信息科学没有取得重大突破。上一次基本创新（即基于科学突破的重大发明）的高峰期是在 20 世纪 40 年代，现在已有大量的知识积累，按照经济与技术发展长波规律的推测，21 世纪 20—30 年代可能出现基本创新的高峰。2020 年以后什么技术将成为新的主流技术就会逐步明朗；2020—2035 年将是信息技术改天换地的大变革期。预计 21 世纪上半叶将兴起一场以高性能计算和仿真、网络科学、智能科学、计算思维为特征的信息科学革命。在网络科学和智能科学取得重大突破以后，21 世纪下半叶，基于信息科学的新的信息技术将取得比 20 世纪下半叶更大的发展。

近几年，人工智能（AI）突飞猛进，验证了《路线图》的基本预判。AI 现有重大技术发明，如 Transformer 深度学习框架，是基于 Geoffrey Hinton 等科学家长期对神经网络模型的基础理论研究成果。深度学习的黑盒模型为 AI 科学研究提出了急需解决的科学问题，需求的牵引必将引发科学的突破。当前 AI 技术的大发展，预示人类已经处在进入智能时代的前夜，目前的技术离实现真正的通用人工智能（AGI）还有一定的距离，再经过 10—20 年的努力，大概率有基于重

大科学突破的基础发明，如同 20 世纪 40—60 年代一样，发明计算机和集成电路以后，人类就进入了信息时代。

2) 关于信息技术发展的长期预判基本符合预期

人一机一物“三元融合”预判正在成为现实，《路线图》和《三元融合》指出，人一机一物三元融合强调的是物理空间、信息空间和社会空间的有机融合，物理空间分别与信息空间、社会空间源源不断地进行信息交互，而信息空间与社会空间则进行着认知属性和计算属性的智能融合。

近 15 年来，人一机一物三元融合正在加速。快速发展的移动互联网、物联网、4G/5G 高速接入网及边缘智能等，为实现人一机一物三元融合准备好了物质条件，数据智能化为智能融合提供了纽带，计算机系统的基本模式正在从人机共生向人一机一物三元融合世界发展。移动互联网实现了人与人的互联、融合，物联网（传感网）实现了人与环境的互联与融合，工业互联网实现通过网络连接各种工业设备和系统，实现工业数据的实时传输、共享与智能化处理，并通过新一代智能模型，改善、提升工业生产效率和质量，以及成本控制等。人一机一物三元融合最显著的是脑机融合，脑机神经连接是重要的科技突破。埃隆·马斯克的“神经连接”公司，继首例人脑设备植入手术顺利完成后，2024 年 7 月迎来第 2 名人类植入者。通过在人脑皮层植入 NeuroLink 相关芯片并采用大约 10 bits/s 的通信速率，使得瘫痪病人能够通过思考

来控制他们的手机或电脑。这真正实现了人一机一物三元融合。NeuroLink 成为现实，也预示着《路线图》指出的，“今后几十年内模拟计算可能又会成为受人重视的研究方向。我们在重视数字技术的同时，还要探索模拟量处理的新途径以及数模混合处理的新方法”，这一预判得到一定程度的验证。

2) 计算与服务平台的总体结构依然呈现出“集中与分散交替为主”的发展规律

《路线图》指出，计算与服务平台的总体结构几乎是每隔 20 年左右有一个重大的变革，呈现出计算资源集中与分散交替为主的“三国定律”。目前，在“AI 大模型+Agent”

“AI 大模型+RAG”等模式的加持下，未来这种集中与分散交替为主的发展规律将继续延续。装载有轻量化 AI 模型+Agent 或+RAG 或+Copilot 的 AI PC，将进一步改变人们使用计算机和互联网的模式。一方面，计算与服务模式分散化的 AI PC 将进一步解耦人们对集中化的计算资源的需求，以及紧耦合的服务模式；另一方面，集成式的多云系统、联邦云系统、超算互联网等，将进一步提升计算资源的最大化利用（图 1）。因此，《路线图》所指出的计算与服务平台的集中一分散交替的“三国定律”，在未来还会继续演进。这种周期性的变化，使得信息技术领域始终保持着创新的活力。

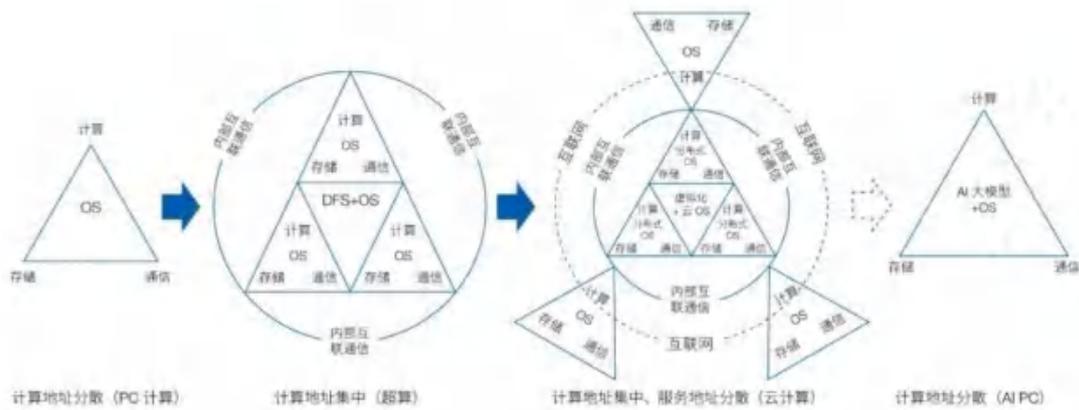


图 1 计算与服务平台总体结构“集中与分散”简略图

2) 集成电路领域延续创新变革发展态势

《路线图》指出，集成电路领域将延续摩尔定律（more Mooer）、扩展摩尔定律（more than More）和超越 CMOS（beyond CMOS）3 个路径发展。

在延续摩尔定律方面，晶圆级的大芯片已经成为芯片结构横向扩展发展的重要成果。2012 年后登拉德缩放比定律受限，摩尔定律发展速度虽然放缓，但围绕缩小 COMS 工艺特征尺寸、提高集成度，以及通过新材料的应用和器件结构的创新来改善电路性能的努力仍然在继续，FinFET 结构和工艺技术在半导体制程发展到 22—5 nm 过程中，发挥了重要的作用。2020 年，在同行都在将晶圆分割成数百颗独立芯片之时，美国 Cerebras Systems 公司则是选择将整个晶圆做成一颗芯片，其发布的 WSE-2 二代晶圆级芯片，采用 CMOS 7 nm 工艺制程，集成了 1.2 万亿个晶体管、40 万个核心。这也是 FinFET 结构和工艺发展的重要成果之一。近年来，我国由于

在集成电路领域受到“卡脖子”，与 FinFET 同时代发明的 FD-SOI 技术也被重新捡起来，成为破解之道之一。

在扩展摩尔定律方面，近些年来，基于先进封装技术的 Chiplet 发展起来。Chiplet 技术允许将整个芯片拆分成多个较小的、可以用不同工艺制造的不同模块，然后通过高速互连方式集成到一个封装中，实现全功能的芯片系统，从而优化性能、功耗和成本。因此，Chiplet 技术被视为一种可以拓展摩尔定律的方式，延续了集成电路行业提高性能和降低成本的趋势。咨询机构 Markets.us 研究报告称，在 2023 年，CPU Chiplet 占据了主导市场地位，CPU Chiplet 市场份额超过 41%。虽然 GPU Chiplet 市场份额低于 CPU Chiplet，但在专业应用领域发挥了关键作用。此外，通过传统微电子工艺，实现光电子器件和微电子器件的单片集成的硅基光电子集成技术，有效解决了集成电路芯片目前金属互联的带宽、功耗和延时等问题，也实现了扩展摩尔定律。

在超越 CMOS 方面，碳纳米管晶体管已经展现出超越商用硅基晶体管的性能和功耗潜力，碳基集成电路技术成为重要发展对象。2019 年 8 月，美国麻省理工学院的 Gage Hills 等在 Nature 发表论文，报告了碳纳米管芯片制造领域的一项重大进展：一个利用 14702 个碳纳米管晶体管构成的 16 位 RISC-V 指令集微处理器 RV16X-NANO，该处理器采用兼容 CMOS 工艺制造，证明可以完全由 CNFET 打造超越硅的微

处理器，为先进微电子装置中的硅带来一种高效能的替代品。2024年7月，北京大学基于碳纳米管晶体管新型器件技术，结合高效的脉动阵列架构设计，成功制备了世界首个碳纳米管基的张量处理器芯片，可实现高能效的卷积神经网络运算。

此外，近15年来，集成电路芯片发展的“摩尔周期”还在延续，能够实现更好性价比的领域专用架构（DSA）处理器，如NPU、TPU、DPU等各种“XPU”，应运而生。国内研发的深度学习系列处理器寒武纪、类脑天机芯、天眸芯和市场上自动驾驶芯片等属于DSA范畴。同时，核心数增多的通用架构的多核和众核处理器仍然在发展，如X86系列CPU处理器和NVIDIA的系列GPU处理器。

3) 软件工程的摩尔定律日益显现

《路线图》指出，软件工程的发展走势将类似于摩尔定律，今后几十年内如果能够使得软件业和服务业也产生摩尔定律现象，无疑将会引发一场革命。目前，基于LLM大模型的AI编码生成助手，为软件工程的摩尔定律提供了依据。如谷歌首席执行官曾透露，该公司通过在代码自增长工具中集成大模型，生成了这家科技公司超过1/4的新代码，包括自动导入包、自动生成构造函数等。目前AI辅助编码方面，出现了两大发展方向：

1. AI 编码助手或者 AI 代码生成器大量涌现，如美国 GitHub 与 Open AI 合作推出的 GitHub Copilot、亚马逊的 CodeWhisperer 等；

2. 传统的低代码/无代码工具，大量引入 AI 辅助功能，如低代码开发旗舰公司 OutSystems 通过 Mentor 新型生成式 AI 驱动的“数字工作者”改变了整个软件开发生命

统计显示，目前软件工程师基于大模型开发应用软件，时间上可以节约 20%—30%。随着面向软件开发的专用大模型能力日益增强，在软件工程领域的摩尔定律也将有望成为现实。

(2) 未来 10 年信息技术体系重构与再造的创新机遇与挑战

《路线图》预测，2020—2035 年将是信息技术改天换地的大变革期，将可能出现基本创新的高峰。自 2019 年以来，AI 大模型的大发展表明，信息科技已经进入到基本创新突破期的前夜。在全球信息技术创新进入缓慢期的背景下，AI 将加速信息技术体系的创新进程。因此，未来 10 年将是信息技术体系重构与再造的创新机遇期。一方面，对于原理还不太清楚的 AI 科学，一定还会有大的突破；另一方面，信息领域的科学与技术融合发展将成为大趋势，并且信息技术将成为信息科学发展的主要推动力。更重要的是，新一代 AI 将加

速驱动计算技术体系、数据空间技术体系、网络空间技术体系和智能空间技术体系的重构与再造。

1) 信息技术体系重构与再造创新的历史演进

人类对信息技术体系重构与再造创新一直没有停滞。在处理器方面，从 4 位微处理器到 64 位处理器，从复杂指令集（CISC）到精简指令集（RISC），从一级缓存到二级、三级，执行从顺序指令执行到乱序指令执行，从单一核心架构到多核心架构、众核心架构，从通用功能架构到功能专用架构并存，从单个芯片到异构集成封装多核心、多功能的芯片等。目前，在 AI 等领域对算力提出更大需求的牵引下，晶圆级大芯片也被开发出来。

在计算机体系结构方面，20 世纪 50—60 年代，为实现资源的最大化利用，通过分时操作系统提高指令执行效率的批处理计算机得到快速发展。此后，为提高指令并行性进一步提高计算机运行速度，指令流水线技术取得重大突破；为了进一步突破计算性能瓶颈，超标量与超长指令字两种计算机体系结构应运而生。再后来，多任务、多个处理器并行执行的并行计算体系在一系列超级计算机中得以广泛应用。近年来，面向大数据、云计算和 AI 等发展需求，按照数据流动规律进行组织和管理的数据流体系结构重新得到重视，成为计算机体系结构发展的“老树新芽”技术。

在信息技术产品生产方面，从早期的由一家厂商包揽了一个计算机的软件、硬件设计与制造的垂直体系，向多厂商分散、配合生产各自优势产品的扁平化体系方向发展。例如，早期的美国 IBM 公司生产的大型计算机，其操作系统、CPU、存储器等都是由 IBM 自己设计、生产和制造。随着微软 Windows 操作系统、Intel CPU 等的成熟商用，计算机生产商不再做垂直化产品研发的工作，而是将其他厂商的产品进行扁平化分工、整合，由此形成了“Windows+Intel”（Wintel）计算产品体系，并形成了牢固的产品迭代节奏，至今形成了“Wintel”、“ARM+Andrio”和“ARM+iOS”体系。目前，由于 GPU 和 Transformer 架构在 AI 大模型领域的成功，“NVIDIA+Transformer”体系已经基本形成。

2) 计算技术体系重构与再造

近年来，全球信息技术重大创新的整体节奏在放缓。未来 10 年，在人工智能发展驱动下，计算技术体系创新将快速推进。一方面，进入后摩尔时代，为追求算力增长和功能丰富，计算处理器芯片将继续向低成本、高良率的 Chiplet 堆叠集成芯片和 3D 芯片方向发展，异构集成创新成为产业技术主要发展方向，晶圆级大芯片将成为重点发展方向之一；同时，功能丰富的低世代工艺智能芯片、新型硅光电子芯片、碳基芯片，以及非冯·诺依曼体系结构量子计算芯片，也将快速向前发展（图 2）；另一方面，算力空间将向算力聚合与

服务能力提升方向前进，将由超级计算、云计算，向云联邦、超算智算融合、算网融合等方向发展（图3）。

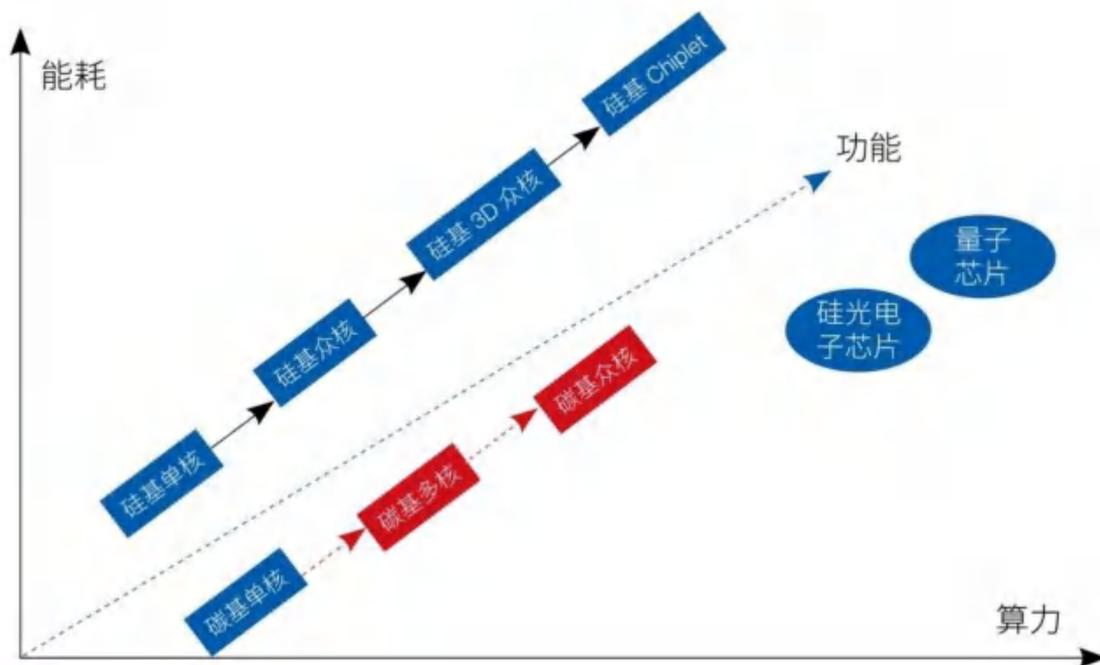


图2 算力芯片重构与再造演进路线图简图

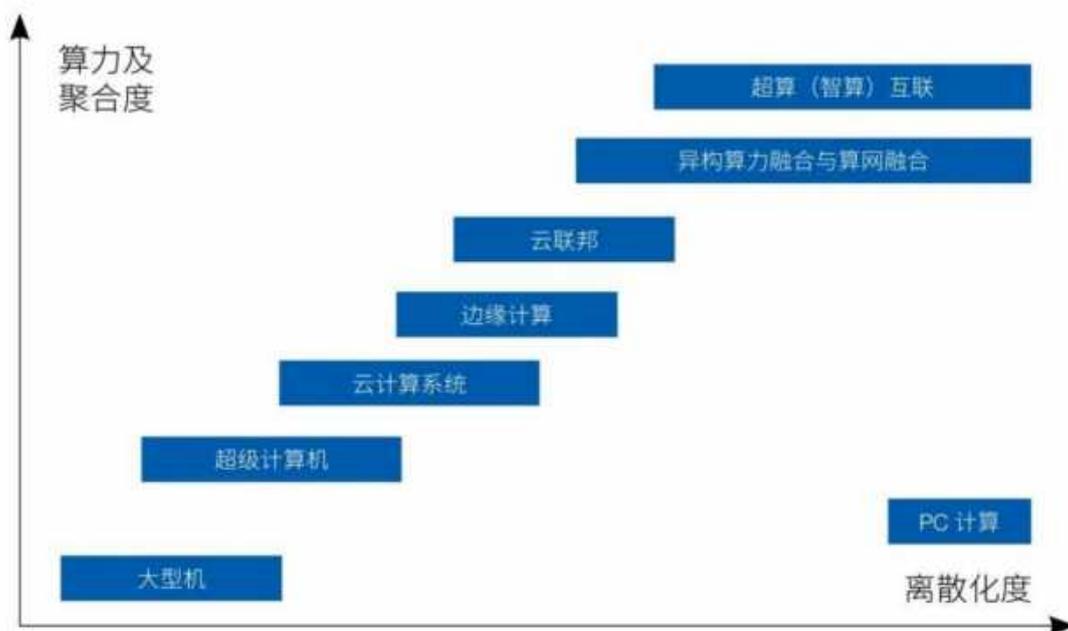


图3 算力空间重构与再造演进路线图简图

3) 处理器技术体系重构与再造

在处理器设计环节，芯片设计的复杂性日益呈指数级增长，导致设计周期长、成本高，传统的设计方法难以应对这种复杂性。未来 10 年，AI 技术将成为优化处理器芯片设计过程的关键工具，提高设计效率和质量，并生产出性能更高的芯片。2024 年 3 月，美国白宫科技政策办公室发布《国家微电子研究战略》，明确提出加强将 AI 和机器学习及基于物理的方法集成到 EDA 工具中。在 2024 年全球芯片领域顶会 Hotchips 年会上，多个报告展示了 AI 辅助芯片设计的重要成果。其中，美国 Synopsys 公司报告了强化学习（reinforcement learning）在芯片设计中的应用，其 Synopsys.ai 套件，在大语言模型支持下，在整个 EDA 堆栈中充分利用生成式人工智能（AIGC），用于 RTL（register-transfer level）设计、验证及其他辅助资料创建的生成功能等。在处理器制造环节，比利时微电子研究中心（IMEC）发布的工艺路线图指出，FinFET 晶体管结构将在 3 nm 走到尽头，然后过渡到新的 GAA（gate all around）结构，集成电路工艺尺度将进入埃米阶段。此路线图提出的 CMOS 2.0 范式愿景指出，CMOS 2.0 是通往真正 3D 芯片的道路。但这种方法面临的挑战在于将严重依赖后端供电网络（BPDN），会将所有供电改到晶体管的背面，并需通过系统技术协同优化（STCO），重新思考设计过程，改变设计方法。相对于硅基集成电路计算芯片，碳基芯片具有良好的低功耗、与硅基工

艺兼容良好等特点。如前文所述，目前碳基芯片已经取得积极进展，将成为未来与硅基芯片并行发展的主要方向之一。

从更长期来看，处理器芯片重构与再造创新发展趋势体现在如下 5 个方面：

1. 电路理论方面，从电子电荷向分子、极化、强电子相关态、自旋方向等方向发展；

2. 材料方面，从硅基向碳基、宏观分子材料、纳米结构、复合金属氧化物等方向发展；

3. 器件结构方面，从 CMOS 器件向分子器件、自旋器件、铁磁性器件、量子器件等方向发展；

4. 数据载体方面，从模拟量、数字量向量子位等方向发展；

5. 系统结构方面，由冯·诺依曼架构、多核众核架构，向可重构、量子、神经形态学计算架构等方向发展。

目前，量子计算还处在探索阶段，主流方案包含超导、离子阱、光量子、超冷原子、硅基量子点和拓扑等多条技术路线，基本都沿着量子计算优越性由专用量子计算向通用量子计算的路线图发展。

对我国而言，在 FinFET 时代，集成电路制造正在面临美国联合其盟友的打压、封锁，向更先进制程工艺芯片制造发展受阻，同时 X86、ARM 指令架构也存在需要授权的制约。随着集成电路制造工艺向 3D 方向发展，前道工艺光刻

机的作用将会减弱，刻蚀机、薄膜机和电子束检测设备等作用将会增加，我国发展集成电路技术与产业有了新机遇。因此，我国一方面要积极发展集成电路先进制程制造工艺技术，研发先进制程的前道工艺极紫外光刻（EUV）设备；另一方面，要积极推动发展刻蚀机、薄膜机、电子束检测设备等，向满足 3D 芯片制造工艺方向发展。此外，基于先进封装的 Chiplet 技术，以及与 FinFET 同时代产生的 FD-SOI 技术，也可能为我国计算芯片发展提供新的选择。FD-SOI 是不同于 FinFET 的技术与工艺路线，是一种平面工艺技术，具有减少硅几何尺寸同时简化制造工艺的优点，在 22/12/10 纳米 FD-SOI 技术性能与 14/7/5 纳米 FinFET 技术相当；并且，FD-SOI 相比 FinFET，具有 PPA 平衡性好等特点。但目前主要集成电路制造设备与工艺是围绕 FinFET 生态的；FD-SOI 面临的挑战是还没有建立完整的产业链生态，扩大市场难。

4) 算力空间技术体系重构与再造

为适应人工智能等领域对更大算力和多种资源协同服务的需求，未来 10 年，面向应用的计算将与智能计算深度融合，表现为临近互联、封装集成、大规模向量化等，并期待 RISC-V+AI Core 指令接口的标准化。在 AI 发展驱动下，算力空间重构与再造的一个重要方向是优化冯·诺依曼架构，即在一个计算机系统内部区域做工作，想办法通过减少内存和处理单元之间的路径，实现高内存带宽及较低的访问开销，

打破冯·诺依曼架构“存储墙”和“功耗墙”等问题，实现更高效能的计算。

在 AI 计算方面，由于不同模型对算力和带宽的需求不同，如基于 Transformer 的 LLM 计算中，性能瓶颈常常是在带宽而非计算，即带宽跑满、算力闲置。为此，存算一体、存内计算(PIM/PNM)技术路线被提出，利用片内的高带宽，处理常见的矩阵运算和部分向量运算。具体是通过 TensorCore (张量计算核心) 的乘累加单元阵列提供算力，通过 HBM 的高带宽使数据能源源不断到达 TensorCore。在 AI 计算中，虽然 PIM/PNM 可以减少数据从内存读取的时间，提高计算效率，从而加速模型的训练和推理过程，但其面临众多技术挑战，包括将计算单元集成到内存芯片中的复杂性、功耗和散热问题，以及数据一致性和可靠性问题等。

在片上和集群的算力互联方面，使用片上光网络(ONoC)连接多个同构的处理单元，如互联 CPU、AI Core 等，这是将光集成电路(PIC)作为 NoC 与 AI Core 的硅片集成，封装成一个 AI 芯片，直接在芯片上做光电转换输出到光缆，实现了芯片之间的互联。基于光网络的互联方案与动态调度方面，如 Google TPU4 AI 训练集群的光互联方案，是将 64 颗 TPUv4 以 $4 \times 4 \times 4$ 的方式构成一个三维立体结构 Cube，并且这个 AI 训练集群的拓扑互联方案的重构，可根据需求实时对 AI 计算资源缩容和扩容。

面向人一机一物三元融合的广泛应用场景，AI 等算力与服务需求以及高性能计算机（超算）从 E 级向 Z 级发展面临诸多难题，未来将跨网域、多异构的算力与服务聚合是一条重要的发展途径。将超算、智算和量子计算等多种异构算力资源融合，并结合算力网实现算网融合是一个大胆的创意，但多种异构算力资源的融合是一个难题。例如，AI 算力主要由基于 GPU、NPU 等芯片，但 AI 的应用通常会同时用到 CPU、GPU、TPU 等，怎么解决 CPU 运算与 GPU、TPU 等运算衔接是个难题。因此，需要解决多种算力资源虚拟化、接口标准统一、高效协同计算、应用任务分布与调度、编程模型等难题。同时，异构算力如何与算力网进行融合也是个难题，需要突破原有的并行计算、云计算等思想，进行颠覆性的创新。为此，需要进行原理性创新和技术实验的重大科技基础设施来支撑。自 2023 年以来，美国 IBM、微软、英伟达、谷歌等公司分别合作，在构建量子—超算异构融合算力平台方面取得进展，如 IBM 利用 127 比特量子云平台与“富岳”超算的结合，实现了包含 28 个原子的 FeS 团簇分子计算。这为多种算力融合技术发展路径提供了很好的借鉴。

5) 数据空间技术体系重构与再造

数据已经成为 AI 大模型发展的三大核心要素之一。中国工程院发布的《数据空间发展战略蓝皮书（2024）》（以下简称《蓝皮书》）定义的未来数据空间是：人一机一物互联，

产生大量数据，通过社会再生产，数据又作用于人一机一物，这一实践活动最终形成了人类活动的新空间。即：在 AI 发展驱动下，数据空间将由单一计算机系统的数据空间、基于互联网的数据中心空间，发展到人一机一物融合的人类智能活动空间。因此，未来 10 年，构建面向人类智能活动的数据空间技术体系，是一项重要任务。

在 AI 发展需求驱动下，构建面向人一机一物三元融合的人类智能活动的数据空间，还面临众多挑战，主要包括：

1.理论范式问题。基于什么理论来支撑数据空间构建，是基于自组织理论还是他组织理论？基于自组织理论要研究数据空间构建是怎样从混沌无序的状态向稳定有序的终态的演化，以及系统内部各要素之间的协同机制。基于他组织理论，要研究数据空间构建需要设计怎样的制度、政策及机制等外部力量，推动数据空间的建立。此外，对于数据空间建立的复杂度，是基于机械论还是耗散结构理论（即是强调将数据空间系统是由数据孤岛构成，通过系统组织形成“整装数据”结构），还是强调引入负熵来形成有序数据空间结构？等等。

2.技术问题。数据空间的架构是什么？各个数据孤岛是通过什么方式进行连接？联接协议是什么？接口是什么？有什么标准？如何保护各个数据源的数据隐私与安全？数据空间中的数据如何进行统一标识？如何封装？数据空间

如何与算力空间进行耦合？数据空间的各类数据如何统一标识和表示，以便于进行综合挖掘和智能应用？等等。目前，《蓝皮书》为此提供了一些参考，包括把数据空间看作一个数据要素场、构建数联网根服务体系，以数联网、数据标识、数据向量化、深度神经网络学习及 AI 大模型等作为技术应用体系等。

6) 网络空间技术体系重构与再造

几十年来互联网体系结构保持相对稳定，但依然存在两大问题：鲁棒性差，脆弱，不安全；适应性弱。为此，国内外对互联网体系结构的创新努力一直在持续，基本上采取渐进式和变革性两条技术路线，即：采用“自下而上”打补丁的方法，升级更新具体技术适应新业务和新终端；采用“自上而下”完善体系结构，从根本性解决问题和克服挑战。变革性的技术路线是采用“推倒重来”的思路设计全新网络，国内外先后开展了 New Arch、Clean Slate、GENI、FIND、FIA、FIRE、新一代互联网体系结构理论、面向服务的未来互联网体系结构与机制等研究计划或项目，先后提出了 Open Flow、SDN、NDN、CCN 等互联网体系结构的新设想和新技术。

未来 10 年，AI 将驱动网络空间技术体系从硬件链路到网络协议，乃至网络应用的全体系的创新。除了在新型网络架构、寻址路由、内生安全等领域继续突破创新外，互联网

体系结构还将向智能化、敏捷化、网存算一体化不断发展，从而支撑超大规模、人一机一物融合、跨时空壁垒的智能连接与服务。随着 AI 技术的发展，互联网将围绕 4 个方面进行创新：

1.网络设备方面，将 AI 能力下沉到网络设备。通过感知业务质量，感知信道性能等，让网络设备具有更强的感知能力，以便更好地优化网络，提升网络为特定业务的服务能力。

2.路由协议方面，通过增强数据面网络层的能力，简化控制面和管理面，满足应用需求的多样性，并提升边缘能力来降低对中间节点的要求。通过协议的简化和优化，使得网络可感知、可溯源、可定位，实现更好的性能及内生安全要求。

3.网络管理方面，首先是网络本身的智能化，随着 AI 大模型在网络管理中的应用，提升网络智能化水平，让网络能够具有自动驾驶、自动发现、自动配置、自动维护的能力；其次，网络需要能够适应更多业务发展，具有更好的开放性和安全性，为云网融合、算网融合实现资源一体化调度。

4.安全方面，改变过去网络安全被动式响应和防御的方式，未来要从网络设备的内生安全，到端到端的云、网、边、端协同防护体系的构建，持续提升网络安全防护能力。

7) 智能空间技术体系重构与再造

人类大脑是一个既能处理专用任务又能处理通用任务，并具有“自觉意识”的低能耗智能空间。信息领域的智能空间是人工智能融合的空间，是能处理各种模态信息的智能体的集合，其愿景是构建像人类大脑的智能体，使得处理各种专用任务的智能与处理通用任务的智能进行深度融合，解决智能碎片化的问题。

2024 年诺贝尔物理学奖和化学奖都颁发给机器学习领域的科学家，表明以机器深度学习为代表的 AI，成为未来 10 年最有前景的技术。近年来，以大模型为代表的新一代人工智能的研究和应用取得突破性进展，催生 AI 新的理论和应用范式，推动 AI 理论和应用进入新的发展阶段。美国 Open AI 公司的 ChatGPT 成为新一代人工智能发展的分水岭。AI 大语言模型技术不断成熟，率先在文本 AIGC 产生大规模应用，并催生图像生成模型蓬勃发展，如 DALL-E2、Stable Diffusion 等。大模型从“一专多能”向“多专多能”前进，带动新一轮应用范式创新。AI 大模型技术也使信息领域的科技创新和产业生态发生巨变，促进了从底层芯片到应用的穿透式的技术与产业生态重构，并将逐渐重构数据空间、算力空间、开发框架、算子库，乃至基础芯片、系统及开源和应用服务模式等。AI for science 正在重构科学研究范式(AI4R)，加速驱动科学研究进程，产生颠覆式突破。当前，AI for science 已经在生物领域取得显著成效。例如，2024 年 5 月，

美国 Google 旗下 DeepMind 公司发布的新一代 AI 模型 AlphaFold3，能够预测蛋白质、DNA、RNA、小分子等的几乎所有生物分子结构和相互作用；AI for Science 已经在内嵌物理模型的神经网络 (PINNs)、约束条件下组合优化问题求解、偏微分方程求解等领域取得初步成效，证明其具备解决科学研究与计算问题的强大潜力。未来，AI for science 将继续驱动生命科学、医药研发、物理装置控制、数学发现、材料发现与合成等领域创新发展，并将在跨界交叉领域进一步驱动产业创新和更广泛的落地应用。同时，多模态大模型开启了通用人工智能 (AGI) 发展道路，使 AGI 进展到“临界点”。

未来 10 年，智能空间技术体系重构与再造面临诸多挑战。

1. 目前的智能感知可适应性差、认知机理不明、泛化能力弱等问题已经开始制约 AI 的更广泛应用，亟待深入探索类人多模态感知、人机混合智能理论与方法，以及突破多源信息复杂场景和自主无人系统等决策、行为智能等。

2. 当前 AI 研究已经从深度学习时代走向“自监督+深度学习+强化学习”的大模型时代，以深度学习为基础的人工智能理论已经遇到天花板，急需探索 AI 新的理论。Minyoung Huh 等最近发表的“柏拉图表征假说”，证明了神经网络训练，正趋向于在它的表征空间中收敛成一个共享的现实“世界统

计模型”。Max Tegmark 团队发表的一项颇具颠覆性的研究，揭示了 LLM 中竟存在类似于人类大脑的脑叶分区结构，显示出模型内部的几何结构与人类大脑的功能分区有着惊人的相似性。这一发现表明，AI 在某种程度上模仿了人类大脑的信息处理方式。这类研究不仅提供了理解 LLM 内部运作的新视角，也对 AI 的潜力有了更为深刻的认识，为未来 AI 的发展提供了新的理论支持。随着研究的深入，未来或许能在大模型的基础上，开发出更智能、更类人的 AI。

3. 当前的多模态大模型面临数据资源耗尽、能耗极大、成本极高和商业模式等问题。更重要的是，GPT-5 研发受阻，文本域里规模定律（Scaling Law）可能已遇到天花板，而探索新型 Scaling Law 的预训练多模态大模型、空间智能是不是就是通用人工智能（AGI）的发展技术路径，还有待进一步的验证。

未来智能空间重构一个可能的途径是在大模型、空间智能的基础上，构建更多各种功能的智能体 Agent，与大模型等进行深度融合与功能拓展，整合各种人工智能为一个集成的智能环境，如发展具身智能。但正如李国杰院士指出的，具身智能还有许多认知问题需要解决，这预示着未来通过具身智能发展 AGI 还面临众多挑战。另外一个可能的途径是 Open AI 发现的扩展测试时计算技术，即：在大模型推理阶段，扩展思考和计算，进行反向多步推理，让大模型能够自

主学习策略、拆解任务、识别并纠正错误。得益于思维链(CoT)推理的引入，Open AI o1 模型是测试时计算的突破，证明了模型可以进行更深入的推理和解决更复杂的问题，为解决规模定律天花板问题和发展 AGI 提供了全新的思路。当然，还有其他可能的技术路径，这些努力将加速 AGI 的到来。

(3) 发展对策

1) 聚焦 AI 驱动的信息领域重大科技问题、突破关键理论与技术

我国信息科技领域既面临短期破解局部“卡脖子”和建立自主技术体系与生态的长期需求，又要面临后摩尔时代颠覆性技术创新突破、抢占科技制高点的历史使命。因此，要持续加强信息科技领域的中长期战略研究，聚焦新一代 AI 的理论、技术与新模型产品，聚焦发展 AI 之根基的各类高性能 XPU 算力芯片的设计与制造、高质量数据集和新型网络的构建，以及高效能计算基础设施建设等，将技术创新与工程实现和技术产品与市场生态建设等统一起来，组织科研院所与市场龙头企业进行协同创新。一方面，聚焦“卡脖子”真问题，除芯片制造前道工艺 EUV 光刻机外，未来集成电路向 3D 芯片发展，前道工艺光刻机作用在减弱，高精度、高可靠与高性能的刻蚀机、薄膜机及电子束检测设备成为芯片制造关键设备，急需布局研制；同时自主架构高性能及开源 RISC-V 指令集的各种 XPU 芯片也需加强研发，并用相当

长的时间打造我国信息技术与产业自主生态。另一方面，要抓住 AI 驱动全球信息技术领域正在进入重构与再造的创新机遇，系统研究计算技术、数据空间、算力空间、数据空间、网络空间和智能空间的技术体系重构与再造面临的挑战性问题，协同市场龙头企业，组织技术创新研发与工程实现队伍，推进我国信息科技领域的整体创新。

2) 布局信息领域重大科技基础设施

没有大的工程牵引、重大科技基础设施支撑，信息科学不会有大的发展。信息科技发展历史上，重大科技基础设施为信息技术创新奠定了物质基础，产生了一系列重大技术发明。例如，在互联网领域，美国早期支持 ARPANET 的基础设施，为人类创造了互联网。围绕未来网络，美国组织了“从零开始”（“Clean Slate”）的革命性未来网络研究，先后支持下一代互联网研究基础试验设施（GENI）和从零开始设计新的互联网架构（FIND）等项目。在云计算领域，美国也支持了全球信息网格（GIG）和云计算测试床（CloudLab）等项目。重大科技基础设施的建设为美国等在信息科技领域领先，创造了极大的优势。为抓住全球信息技术发展进入到缓慢期以及计算技术、算力空间、数据空间、网络空间和智能空间等技术体系重构与再造的机遇期，我国应加快布局信息领域的重大基础设施，抢占信息科技制高点。

3) 一体化布局信息科技领域“装置群”

在自然科学领域，往往围绕一个或两三个重大科学问题，构建一个大型的基础设施，进行重大科学探索与发现的实验研究。信息领域与之不同，信息领域的重大科技问题特征是：基础原理、基础共性技术和领域应用 3 类问题并存，许多重大科技问题分散在信息技术体系框架的各个部分；并且，信息科技的基本原理、共性技术等局部问题一旦突破，将能够影响整个领域和其他领域的发展。我们不能指望通过一个集中化、实验型的基础设施来解决信息领域各个环节面临的问题。因此，信息领域的重大科技基础设施应该是一个“装置群”，以应对各个子领域的科技问题。同时，信息领域的重大科技基础设施应该是一个柔性可组合的离散、分布式“大装置”，要进行一体化的布局。

（二）企业

1. DeepSeek 的创新三重门

小天才的规模化 + 华为式的军团平推 + 原创（哲学式）思想。

（1）小天才们的胜利

大模型训练是一个相当复杂的软硬一体的工程，而 DeepSeek 几乎重新设计了绝大部分关键组件：MLA, GRPO, DeepSeekMoE, DualPipe, FP8 混合精度, R1-Zero, MTP 等等。范围之广，密度之大，非常震撼。

团队中其中有许多都拥有国际竞赛背景（吴作凡，任之洲，周雨杨，罗煜翔等），好多甚至是实习或者刚开始读博（DS-Math 的作者邵智宏/Peiyi Wang, Zihan Wang 等）。

小天才式的创新，更有名的例子来自于美国：GPT 的最早提出者 Alec Radford，思维链 CoT 的提出者 Jason Wei，Sora 的主力贡献者 Bill Peebles，都是初入职场，甚至没有博士训练的年轻人。

我们把这些人称之为“小天才们”，因为他们不仅有天才般的 learn 的能力，更重要的是，他们没有 unlearn 的负担。如果给“小天才式的创新”一个定义，那就是：给定约束，寻找最优解的能力。我们来举例 DeepSeek 的三个核心算法模块：**MLA**：传统 Attention 内存占用太大，怎么办？改造 Attention 模块，通过低秩压缩，让 KV Cache 的效率达到最优。**DeepSeekMoE**：传统 MoE 专家不够精细，激活参数大，怎么办？改造 MoE，通过增加共享专家和细粒度专家，大幅提升了 MoE 中的专家学习的效果。**GRPO**：传统 PPO 需要训练两个模型，效率低，怎么办？通过去掉 Value Model，引入 Group-Relative 作为 baseline，大大提升了训练效率。

抛开专业术语，用人话来理解背后的模式是：一个非常优秀的工程师，遇到问题会尝试不同的技术选型，选一个最好的方案。而小天才会说：在座的都不够好，为什么不重写一个？然后写出了行业最佳方案。这样的例子，在 DeepSeek

的文章里非常多，而每一个背后，都站着一个小天才。梁文峰自己说过，这样的小天才式的创新，在硅谷每天都在发生。但这不能抹杀 DeepSeek 出现的意义，因为他不仅证明了中国能孕育硅谷型的创新文化，甚至还可以进一步规模化这种模式。毕竟，我们有着全世界最多的小天才们。想象一下，这样的文化，扩散到更多的创新型公司，成为主流，会是怎样的一种盛况？这是创新的第一重门。

（2）华为式的军团平推

如果说，DeepSeek 只是复制了硅谷的小天才模式，为什么会让那些发明这个模式的 OpenAI 们，如此紧张？这就要提到第二个关键要素。

当你深度阅读完 DeepSeek 的论文，尤其是 V3，一种新的感受会慢慢浮现：这是一个从底层硬件到上层算法的复杂大系统，以优雅的顶层设计环环相扣，以大破大立的方式平推完成。

这就是创新的第二重门：军团式的协同创新。而这，恰恰就是中国擅长的模式，也是我借用华为之名的原因。硅谷教父 Peter Thiel，对这种创新有过一个深刻的表述。他认为一个真正的垄断式创新，需要构建 **Complex, Vertically Integrated System**，许多不同层面的创新必须同时发生，并以一种高度协同的方式组合在一起。他认为 Elon Musk 成功的

根本,就来自于对这种协同复杂系统的追求。而 Elon Musk, 也被认为某些方面是最接近中国企业的。

在第一重门,我们拿着放大镜,看 DeepSeek 在三个算法上的极致优化。但这些单点优化,也容易引入超越算法范畴的问题。

所以让我们切换视角,从更全局的视角,来重新理解 DeepSeek 的创新结构,第一条线是模型的迭代节奏(这里忽略了很多支线): 2023 年 6 月, DeepSeek 成立; 2024 年 2 月,发表 DeepSeek-Math, 提出 GRPO; 2024 年 5 月,推出 V2,提出 MLA 和 DeepSeek-MoE; 2024 年 11 月,推出 V3,提出了 MTP 和一整套软硬一体优化方案等; 2025 年 1 月,推出 R1, 提出 R1-Zero。

从 DeepSeek-Math 的第一次推出,到最后走到 R1,不到 12 个月。但速度甚至都不是讨论的重点。而是另一条纵向的主线: 他几乎从基础设施搭建,到底层硬件优化,到模型算法创新,一整个自己重做了一套。更关键的是,他们是高度协同优化,逢山开路,遇水搭桥,充满了整体的逻辑性。

(0) 首先构建自己的集群(萤火),为了更高效的并行训练,那就自研 HAI LLM 训练框架。(1) 发现传统 Attention 的 KV Cache 开销太大,那就用 low-rank 来重新设计 MLA,缓存量下降 90% 以上。(2) 发现传统 MoE 缺乏共享专家和细粒度专家,涌现能力不够,那就设计自己的

DeepSeekMoE 解决专家精细度和共享知识问题。(3)发现引入的细粒度专家在训练中负载均衡有挑战,那就自研专家偏好的路由算法,保证每个专家都训练充分。(4)发现 MoE 在管线并行(大模型不同层次分卡训练)中很多通信带宽和计算带宽无法对齐,造成浪费,那自己设计 DualPipe 算法解决。(5)为了确保上面的带宽完全对齐,直接写 PTX 这种底层语言来绕过 CUDA 限制,来精准控制 GPU SM 的数量和 Warp 数量做通信处理。(6)发现 NVLink 和 IB 的带宽有三倍差距,那就确保 MoE 算法设计中只路由四个节点的专家,每个节点内平均访问三个专家来对齐带宽。(7)发现 Tensor Parallelism 通信开销大,通过在 MLA 上重计算等方法释放显存来直接跳过 TP 步骤。(8)为了进一步提升训练效率,开发了精细化的 FP8 混合精度(细粒度量化),在保证训练效果的情况下大幅减少计算和通信。(9)为了更进一步提升训练表现,在训练过程中引入 MTP 让每次训练密度更高。(10)为了提高 V3 的推理能力,那就拿 R1 来 distill V3 的推理。(11)为了在强化学习中少训练一个 Value Model,开发出 GRPO。(12)为了解决小推理模型 Self-play 推理学习的不足,就把大模型学出来的推理通过 Distill 注入到小模型里。(13) more

在介绍完方案的最后，V3 专门开辟章节，提出了一系列对硬件厂商设计下一代芯片的方案，包含大量的对通信，量化算子的设想。

3.5. Suggestions on Hardware Design

Based on our implementation of the all-to-all communication and FP8 training scheme, we propose the following suggestions on chip design to AI hardware vendors.

在 DeepSeek 的脑海中，是把创造 AGI 这件事情，看成一个大蓝图的，而这个蓝图是没有边界的。算法？通信？数据？硬件？都在需要解决的范畴内，而且高度协同的平推下去。如果给他们足够的精力和钱，他们会做自己的硬件，造自己的电网。这是真正创新者的样子，就像乔布斯说过：真正在乎软件的人，应该去制造属于自己的硬件。

背后的中国元素：和小天才模式不同，这种模式恰恰是硅谷文化中相对忽视的，马斯克除外。2019 年我回国，发现字节和快手的推荐能力比 Meta 强，当时很意外。然后我开始了解到华为，到今天，以华为为代表的中国式组织，在国际竞争中释放了巨大的能量，电动车、内容分发、智能硬件甚至电商。只不过，今天是 AGI 这个会改变文明走向的领域，DeepSeek 以如此迅猛的姿态登场，再一次刷新了全世界的认知。这就是创新的第二重门。可以看到在 DeepSeek，当我们把创新的第一重门（小天才式）和第二重门（体系协同式）融合在一起，威力相当大。

(3) 原创（哲学性）思想

那是否还可以继续往下挖掘？是否存在一种更底层的创新元素？一个类似 o1 或者 R1 这样的颠覆式创新系统，构建于什么基础之上？答案是一个原创的、全新的系统结构。那这个原创的结构，又从何而来？回答这个，可以引用一段很喜欢的 Ilya 的话（视频前 2 分钟）：我喜欢思考非常基本的问题...几乎可以将这些问题视为哲学问题。例如，什么是学习？什么是经验？什么是思考？...我认为技术就像是一种自然力量，但看起来我们可以通过设计算法，来做有用的事情，同时解答这些(哲学问题)。这就像是一种应用哲学。

没错，第三重门的答案，是一种近乎哲学式的，原创的思想。

如果我们把视角拉远，看一下过去十年 AI 的创新源头。分别是：Google 2017 年的 Transformer，DeepMind 2017 年的 AlphaZero，以及 OpenAI 2018 年的自回归模型 GPT。但是这不仅仅是三个模型啊，内核是三个哲学性思想，关于“学习”的本质：（1）用什么结构来理解世界万物，和他们的内在关系（Transformer）。（2）如何通过不断的预测，把世界万物的复杂性压缩到直觉中（GPT）。（3）怎么基于直觉构建深度思考，完成对世界的进一步抽象（Zero）。

我们今天看到的 OpenAI o1，或者 DeepSeek R1，都是这三个哲学性思想，组合后的产物。从思想出发，先搭建了

一个基础的系统结构，然后通过第一重门和第二重门的创新，把这个结构极致的打磨和再组合，显化成今天改变世界的产品。哲学性思想的创造为何如此美妙？因为他们并不是为了找到一个优秀的解法，而是为了问出更本质的问题。

(4) DeepSeek 跨越第三重门了吗？

梁文锋可能也会说，今天的 DeepSeek，还没有展现出这种原创（哲学性）的创新。但我却在读他们的文字中，瞥见了藏在背后的希望。第一个例子，是关于 MLA 的产生，来源于一个研究者“总结了 Attention 架构的演进规律”：

暗涌：这次MLA创新是如何发生的？听说idea最早来自一个年轻研究员的个人兴趣？

梁文锋：在总结出Attention架构的一些主流变迁规律后，他突发奇想去设计一个替代方案。不过从想法到落地，中间是一个漫长的过程。我们为此组了一个team，花了几十个月时间才跑通。

第二个例子，是 DeepSeek-Math 文章中很大篇幅，分享了对所有后训练过程的统一框架的认识。尽管有类似的思想存在过，但你似乎可以感受到那些背后无边界的热烈讨论。

5.2. Insights of Reinforcement Learning

5.2.1. Towards to a Unified Paradigm

In this section, we provide a unified paradigm to analyze different training methods, such as SFT, RFT, DPO, PPO, GRPO, and further conduct experiments to explore the factors of the unified paradigm. Generally, the gradient with respect to the parameter θ of a training method can be written as:

$$\nabla_{\theta} J_{\theta}(\theta) = \mathbb{E}[(q, \theta) \sim \mathcal{D}] \left(\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \underbrace{GC_{\theta}(q, \theta, t, \pi_{-i})}_{\text{gradient w.r.t. } \theta} \nabla_{\theta} \log \pi_{\theta}(o_i | q, o_{-i}) \right) \quad (5)$$

第三个例子，则是他们在 R1 中展开的 R1-Zero 的探索，这个大胆尝试，希望跳过所有的后训练步骤，让大模型在直觉模型的基础上，完全不依赖标注数据，自我演进成一个有思维能力的模型。R1-Zero 没有用到最终的 R1 产品

中，但论文中却浓墨重彩的探讨他的启发意义。对，以及那个 Aha Moment, 原文是这样写的：“这不仅仅是模型的 Aha, 也是观察模型的研究者的 Aha!”

所以，你感受到了吗？在克制的技术语言背后，藏着一种无法掩盖的兴奋和热爱。这是无边界的好奇心的味道，这是迫不及待想和世界分享的味道。是原创的、哲学式的思想，呼之欲出的味道。

（5）最后一块拼图

我不知道梁文锋是不是中国的 Ilya Sutskever, 但我坚信一个跨过创新三重门的组织，需要一个思想性的领袖。说到这里，今天 OpenAI 最缺的，不就是这个吗？（写第一篇时的心情重现）谁都无法预测未来。但也许哪一天，在 DeepSeek 的新论文中，会出现 Transformer 和 AlphaZero 级别的全新思想。更大胆一点，如果我们在 DeepSeek 的启发下，产生了许多全新时代的创新型组织，在各自的领域跨过创新的第一重门和第二重门，并进一步产生了一批颠覆式的思考者，和原创的、哲学性的思想。再大胆一点，如果这些思想和创新成果，以透明和开放的方式分享给全世界。这样的世界，你会更愿意参与吗？这是 DeepSeek, 会真正让历史记住的东西。

2. DeepSeek 颠覆了什么？——大模型“国产之光”破局的启示

一家人工智能初创企业浅浅扇动两下翅膀，即掀起全球科技界一阵“海啸”。短短 30 天，中国初创企业深度求索（DeepSeek）先后发布两款性能比肩 GPT-4o 的大模型，“1/18 的训练成本、1/10 的团队规模、不分伯仲的模型性能”令硅谷大受震撼。最新一期《经济学人》封面文章第一时间让位给这一土生土长的国产大模型：《低成本中国模型的成功动摇美国科技优势》。很快，华尔街也同样感受到了这种被动摇的震感。事实上，这匹黑马的贡献绝非“低成本”这一个标签所能概括。它不仅重新定义了大模型的生产函数，还将重新定义计算。不论开源与闭源未来的优势如何，这股冲击波都将迫使全球科技界重新思考：当“规模定律”与“生态壁垒”不再绝对，什么才是下一赛季 AI 竞争的核心？或许我们能从中获得新的启示。

（1）击穿三大定式

1 月下旬，DeepSeek 在中区、美区苹果 App Store 下载榜单中登顶，超越 ChatGPT、谷歌 Gemini 等全球顶尖科技巨头研发的模型产品。具体而言，它颠覆了什么？

——打破“越强越贵”的成本诅咒

价格感人是让 DeepSeek 快速出圈的第一个标签。DeepSeek-R1 的 API 服务定价为每百万输入 tokens 1 元（缓存命中）/4 元（缓存未命中），每百万输出 tokens 16 元，而 o1 模型上述三项服务的定价分别是 55 元、110 元、438 元。

凡是使用过几款大模型的用户很快就能形成这样一个共识：就推理能力而言，DeepSeek 直逼 OpenAI 的 o1、Meta 的 Llama-3 等一流模型，甚至在回答问题之前还能给出它的推理过程和思考链路。AI 投资机构 Menlo Ventures 负责人 Deedy 对比谷歌 Gemini 和 DeepSeek-R1 后表示，DeepSeek-R1 更便宜、上下文更长、推理性能更佳。低成本比肩 o1 模型，令硅谷的“烧钱模式”一时间遭到猛烈质疑。然而在过去，大模型服务是标准的“一分钱一分货”，想要用上更优性能的产品必须支付更高昂的费用，以覆盖整个模型训练过程中更高算力成本的支出。

——超越“性能-成本-速度”的不可能三角

当硅谷仍在为 GPU 万卡集群豪掷千亿资金时，一群土生土长的中国年轻人用 557.6 万美元证明：AI 大模型的比拼或许并不只靠规模，更重要的是看实际效果。有句话形象地概括出 DeepSeek 的优势：“不是 GPT 用不起，而是 DeepSeek 更具性价比。”传统模型训练，需要在性能、成本、速度之间权衡，其高性能的获得，需要极高的成本投入、更漫长的计算时间。而 DeepSeek 重构了大模型的“成本-性能”曲线，同时压缩了计算周期。根据 DeepSeek 技术报告，DeepSeek-V3 模型的训练成本为 557.6 万美元，训练使用的是算力受限的英伟达 H800 GPU 集群。相比之下，同样是开源模型的 Meta 旗下 Llama-3.1 模型的训练成本超过 6000 万美元，而 OpenAI

的 GPT-4o 模型的训练成本为 1 亿美元，且使用的是性能更加优异的英伟达 H100 GPU 集群。而使用过程中，DeepSeek 给出反馈的时长也大部分控制在 5 秒至 35 秒之间，通过算法轻量化、计算效率最大化、资源利用率优化，成功压缩了计算时间，降低了延迟。

——走出“参数膨胀”陷阱

ChatGPT 横空出世后 700 多天里，全球人工智能巨头不约而同走上了一条“大力出奇迹”的“暴力美学”路线，参数越“炼”越大，给算力、数据、能耗带来了极大压力。很长一段时间，参数几乎成为大模型厂商比拼的最大焦点。而另辟蹊径的 DeepSeek 恰巧处于对角线的另一端：并不盲目追求参数之大，而是选择了一条通过探索更高效训练方法以实现性能提升的“小而精”路线，打破了“参数膨胀”的惯性。例如 DeepSeek-R1（4B 参数）在数学推理、代码生成等任务上具有比肩 70B 参数模型（如 Llama-2）的能力，通过算法优化、数据质量提升，小参数模型一样能实现高性能，甚至能够“四两拨千斤”。

（2）实现三大跃升

“DeepSeek 出圈，很好地证明了我们的竞争优势：通过有限资源的极致高效利用，实现以少胜多。中国与美国在 AI 领域的差距正在缩小。”面壁智能首席科学家刘知远说。算力

封锁下的有力破局，得益于 DeepSeek 技术架构、数据策略、工程实践三方面的关键突破。

——技术架构：重新定义参数效率大模型的千亿参数不应是冰冷的数字堆砌，而应是巧夺天工般地重组整合。

传统大模型 Transformer 架构好比一条承载车辆的高速公路，当车辆（数据）数量足够多的时候，每辆车必须和前后所有车沟通完成才能继续行驶（计算），导致堵车（计算慢、能耗高）。而 DeepSeek 创新的架构则把一条串行的高速路，变成了一个辐射状的快递分拣中心，先把货物（数据）按类型分类打包，再分不同路线同时出发开往不同目的地，每辆货车（计算）只需选择最短路径。因此既能提高速度又能节约能耗。

——数据策略：质量驱动的成本控制

DeepSeek 研发团队相信，用“炼数据”取代“堆数据”，能使训练更具效率。传统的数据策略好比去农场随便采撷，常有价值不高的烂菜叶（低质量数据）。而 DeepSeek 创新的数据蒸馏技术，有针对性地筛选掉质量不高的烂菜叶：一方面自动识别高价值数据片段（如代码逻辑推理链），相比随机采样训练效率提升 3.2 倍，另一方面通过对抗训练生成合成数据，将高质量代码数据获取成本从每 100 个 tokens 的 0.8 元降低至 0.12 元。

——工程实践：架起“超级工厂”流水线

大模型传统的训练方式好比手工造车，一次只能装配一台，效率低下。而 DeepSeek 的 3D 并行相当于一方面通过流水线并行把造车流程拆分为 10 个步骤，同时组装 10 辆车(数据分块处理)，另一方面通过张量并行，把发动机拆成零件，分给 10 个工厂同时生产（模型分片计算）。至于推理过程，传统模型好比现点现做的餐厅，客户等菜时间长，推理过程慢。而 DeepSeek 采用的 INT4 量化，能把复杂菜品提前做成预制菜，加热（计算）时间减半，口味损失不到 5%，实现了大模型的低成本工业化。

（3）超越技术的启示

拆解 DeepSeek 的成功公式不难发现，通过底层架构创新降低 AGI 成本，同时以开源策略构建生态护城河，提供了中小型机构突破算力限制的可行路径。此外，我们还能从中得到一些超越技术的启示。

一直以来，驱动 DeepSeek 的目标并非利润而是使命。

“探索未至之境”的愿景也指向一种与之匹配的极简且清爽的组织架构。一名人工智能科学家表示，在人工智能大模型领域，大厂、高校、传统科研机构、初创公司在资源禀赋上其实各有优势，但结构性的局限很难扭转。DeepSeek 这种类型的初创公司能很好地弥补其中的一个缺位——具有大工程能力、不受制于短视商业逻辑的创新定力、创业团队扁平化组织机制带来的丝滑流畅的协作度。

据不具名人士透露,其 V3 模型的关键训练架构 MLA 就源于一位年轻研究员的个人兴趣,经过研判后 DeepSeek 组建了专项团队开展大规模验证与攻关。而 R1 模型果断调整强化学习路线,领先于其他机构实现了近似 o1 的推理能力,核心原因之一也归功于其青年团队对前沿技术的敏锐嗅觉与大胆尝试。“我们创新缺少的不是资本,而是信心,以及组织高密度人才的能力,调动他们高效地进行创造力与好奇心驱动的创新。” DeepSeek 创始人梁文锋说。

在 DeepSeek 身上,我们看到了皮克斯动画工作室的影子。这个年轻的中国初创企业与那个创作 27 部长片有 19 部获得奥斯卡最佳动画、同样以创新为鲜明标签的组织,有着不谋而合的组织机制与范式选择,以及由此带来的高企的创新成功率与人才留存率。不同于先发者略显封锁的护城堡垒,DeepSeek 赢得赞誉的还有它的开源路线。将代码、模型权重和训练日志全部公开,不仅需要格局,更需要勇气与实力。

“在颠覆性技术面前,闭源形成的护城河是短暂的。即使 OpenAI 闭源,也无法阻止被别人赶超。”梁文锋这样解释选择开源的原因,“开源、发论文,其实并不会失去什么。对于技术人员来说,被 follow(追随模仿)是很有成就感的事。”从技术到愿景,DeepSeek 坚定选择的始终是一条难且正确的路。这也是为什么,即便别国在人工智能领域已坐享先发优势,后发者依然有机会凭借技术创新、成本革命打破大模型

竞争的传统逻辑，打破人工智能行业竞争格局，打破“他国更擅长从 0 到 1 的原始创新，而中国更擅长从 1 到 10 的应用创新”的成见，重塑竞争优势的奥秘。北京时间 2 月 1 日凌晨，在 OpenAI 发布其推理系列最新模型 o3-mini 时，我们从中看到了一个熟悉的字眼：“Deep research”。这是否是国产之光“被 follow”的开端尽管不好断言，但我们至少身体力行走出了一条“大力出奇迹”之外，另一条具有自主创新基因的制胜之道。正如梁文锋此前接受采访时所说，“中国要逐步成为贡献者，而不是一直‘搭便车’。”

（三）地方

1. 为什么 DeepSeek 没有出现在浦东张江

张江的资源向“国家队”倾斜：张江科学城的土地、资金优先支持中科院上海分院、复旦张江校区等“嫡系部队”，民营企业需排队等“残羹剩饭”。

张江是“精英内卷”：名校博士挤破头进中芯国际当“高级操作工”，而真正敢撕掉 SCI 论文、all in 颠覆性技术的人凤毛麟角。

想在张江创业，如同硬科技的“学历军备竞赛”：先亮出 MIT 博士学位、顶会论文或英特尔/ASML 工作经历。某芯片初创公司创始人直言：“在张江，没三篇 ISSCC 都不好意思要融资。”

而杭州是“幸存者偏差”：无数草根创业者沦为直播带货炮灰、跨境电商“刷单难民”，但一个李佳琦的成功就能让年轻人前赴后继。

浙江在政府考核中设立“创新免责条款”，官员在支持新兴技术时决策失误可减免追责，浦东张江则缺乏类似机制，导致官员“不敢为”。

两地的差异本质是市场原教旨主义 vs 国家资本主义的缩影——杭州把政策当“橡皮泥”；张江把政策当“铁栅栏”，确保创新不越雷池半步。

张江是顶层设计主导，从土地规划到产业目录均由政府强力主导，创业者更像“系统内的创新者”，缺乏杭州式的民间野性。张江的官方服务仍困在工业化时代的治理逻辑——重硬件、重流程、重短期可见成果，而杭州已进化到数字经济时代的“企业合伙人”模式——敢放权、敢试错、敢用资源换产业话语权。

以上论断可能与现实有偏差，但这是 DeepSeek 给出的答案；如有雷同，也纯属大模型巧合。

张江高科技园区作为中国科技创新的重要高地，在集成电路、生物医药、人工智能等领域成就斐然，但为何未能孕育出类似深度求索（DeepSeek）这类聚焦大模型与生成式 AI 的企业？

从产业生态、资源禀赋、政策导向等多维度解析，押司将 DeepSeek 给出的答案尽量避开敏感词，并择要整理如下。其中提到张江存在的一些问题，有则改之，无则加勉。

杭州与张江（代表上海浦东）在历史基因、政策逻辑和创业文化上存在结构性差异，这种差异不仅塑造了两地的产业格局，也深刻影响着年轻人的创业选择。

若张江不打破“盖章经济”的路径依赖，其在 AI 时代的竞争力可能进一步被杭州、深圳等“猛兽派”城市甩开。这不是单纯的政策力度问题，而是治理体系代际差距的体现。

张江追求“零风险”的保守心态：官员对 AI 伦理、数据安全等问题高度敏感，常因“怕出事”而限制创新，例如要求企业所有数据本地化存储，导致跨国研发合作受阻。

杭州将 AI 视为“换道超车”的机会，愿意承担试错风险，例如对无人驾驶事故的责任认定，杭州率先出台地方法规明确企业免责条款，而上海仍在“研究”。

张江是技术官僚的“合规优先”逻辑：政策设计围绕**硬件制造与成熟产业（芯片、医药），注重技术安全性和流程合规性，例如芯片流片需层层审批，政策资源倾向于“看得见摸得着”的实体成果。

张江对 AI 等软性创新的支持停留在口号层面，例如大模型企业需要的开放数据接口、跨境数据流动试点等实质性政策难以突破行政壁垒。

典型案例是：某 AI 创业公司申请医疗数据训练权限，因张江“数据安全顾虑”被拒，最终转投杭州余杭区，依托浙江“政务数据开放试验区”获得授权。

相比之下，杭州是野蛮生长的“绿灯思维”：政府以互联网思维运作，主动为 AI 企业“开绿灯”，例如允许企业在监管 sandbox 中试错（如自动驾驶路测、AI 诊疗），甚至协调阿里云等巨头开放算力资源。

杭州的政策工具更灵活：通过“一事一议”定制补贴（如按模型参数量补贴研发费用），而张江的补贴标准仍固守“专利数量、固定资产投资额”等传统指标。

张江尚存在科层制的“公章马拉松”：企业从注册到落地需对接多个部门（科委、经信委、市场监管局），流程复杂耗时，某 AI 初创企业反映“为申请算力补贴，跑了 3 个月才盖章”。

张江的窗口服务僵化：工作人员常以“不符合既有政策”为由拒绝创新诉求，缺乏变通能力，例如对“大模型训练用电”这类新需求，仍按工业用电标准审批，导致成本高企。

而杭州是数字化政府的“极限压缩”：依托“最多跑一次”改革，企业注册、政策申报全流程在线化，杭州某生成式 AI 公司从签约到获得首笔补贴仅用 15 天。

杭州采取主动上门服务：经信部门设立“AI 专班”，针对头部企业直接派驻联络员，协调解决算力采购、人才落户等痛点，甚至为企业对接浙大实验室资源。

张江的产业链“孤岛化”：政府更擅长服务中芯国际、华虹等巨头，但对中小 AI 企业的生态整合能力薄弱。例如，张江 AI 企业难以对接上海超算中心的算力资源（优先服务科研机构），而杭州政府可直接协调之江实验室、阿里云与创业公司“搭桥”。

张江的资本对接滞后：张江的国有资本（如浦东科创集团）偏好中后期项目，而杭州的民资本（如湖畔山南、元璟资本）敢押注早期 AI 项目，政府甚至参与组建“大模型专项基金”。

另外，张江的办公场地、人才成本较高，初创企业或对成本敏感的公司可能倾向选择成本更低的区域（如杭州、苏州等周边城市），或在政策扶持力度更大的园区落户。

1992 年张江高科成立，是国家布局集成电路、生物医药等“卡脖子”产业的战略基地，央企、外企研发中心扎堆（如中芯国际、罗氏制药）。张江的优惠政策长期向集成电路、生物医药等“硬科技”倾斜，例如芯片流片补贴、医药审批绿色通道等。

张江的产业布局以硬件驱动型科技为核心，形成了独特的生态特征：硬件与制造的强基因：张江的集成电路产业全

球领先，生物医药依赖实体研发和生产，人工智能领域更偏向芯片设计（如寒武纪、地平线）和终端应用（如机器人、自动驾驶）。

杭州早期以旅游和轻工业立市，1990年代借力民营经济（如娃哈哈、万向集团）完成原始积累。

杭州有互联网时代的“造富神话”：马云在湖畔花园居民楼创立阿里的故事，让年轻人相信“PPT+胆量=估值过亿”。

阿里巴巴的崛起（1999年）彻底改写城市基因，互联网草根创业文化深入骨髓。企业倒逼了改革：阿里巴巴等巨头崛起后，政府被迫调整规则（如全国首个跨境电商综试区、数据知识产权质押试点），政策本质是为既有企业“开后门”。

杭州对灰色创新（如早期互联网金融 P2P、直播打赏）睁一只眼闭一只眼，等企业做大后再补监管（典型“先污染后治理”）。

杭州从西湖时代来到了钱塘江时代：依托电商、直播、MCN等产业链，草根创业者靠一部手机就能开网店、做网红（如义乌小商品卖家、抖音腰部主播），失败成本低，阶级跃迁幻觉强。

残酷现实是：张江与杭州是两个平行宇宙的样板。