新一代人工智能基础设施

白皮书

中国智能算力产业联盟

人工智能算力产业生态联盟

商汤科技智能产业研究院

编委会

指导单位

中国信息通信研究院云计算与大数据研究所、中国科学院计算技术研究所、中国智能算力产业联盟、人工智能算力产业生态联盟

指导委员会

何宝宏　中国信息通信研究院云计算与大数据研究所所长

张云泉　中国科学院计算技术研究所研究员

安静　中国智能算力产业联盟秘书长

杨帆　商汤科技联合创始人、大装置事业群总裁

陈宇恒　商汤科技联合创始人、大装置事业群副总裁

鸣谢

王婉秋、李兆松、成功、杨松、宾佳丽、孙振邦、苏立宇、刘武、贾海刚、刘雅婧、代继、何聪辉、曹阳、张雯、许伟军、杨学燕、蒋慧、王进、谭覃、何茜、刘沛、张琛

编写组

王月　中国信息通信研究院云计算与大数据研究所数据中心部副主任

周彩红　中国信息通信研究院云计算与大数据研究所数据中心部工程师

程大宁　中国科学院计算技术研究所特别研究助理

刘亮　商汤科技智能产业研究院战略研究主任

杨燕　商汤科技智能产业研究院战略研究主任

田丰　商汤科技智能产业研究院院长

“商汤成立之初，我们认为 AI 产业在未来一定会形成分化。在分化的过程中，整个 AI 基础设施上的各个要素，以一种更加高效、低成本的方式，被提供给更多的人使用，从而降低 AI 基础设施的构建成本和使用门槛。”

— 杨帆，商汤联合创始人、大装置事业群总裁

目录

关键发现

1. 2023 年是人工智能产业发展的分水岭，以大模型、生成式 AI 为发展里程碑的技术革新，推动着人工智能发展进入全新的 2.0 时代，人工智能由之前点状、创新应用，逐步规模化发展赋能企业业务流程各个环节，并逐步向产业深水区发展，推动产业链分化成熟，需要全新的基础设施来实现更好的支撑。

2. 大模型和生成式的发展对算力、算法平台、数据提出全新要求，传统以 CPU 为中心的云计算基础设施已无法满足。不仅需要大规模、高性能、高稳定性算力资源，智能化数据管理流程，以及高效普惠 AI 开发平台；还要打造体系化工程系统保证基础设施面向大模型训练、生成式 AI 应用落地的新目标。

3. Model as a Service（MaaS）成为新一代 AI 基础设施的核心，其本质是通过云服务向开发者和企业提供更高效的大模型服务。MaaS 加速了 AI 应用部署的周期，提升了创新的迭代速度，降低了企业应用大模型服务的多方面成本，推动了 AI 与各行业的深度整合。通过纳入开源和闭源大模型，MaaS 还助力于构建成熟的生态系统，促进生成式 AI 应用的规模化落地。

4. 新一代 AI 基础设施不是传统云的 AI 化，两者具有明显定位和发展路径的差别。新一代 AI 基础设施主要面向产业用户，为大模型训练、区域行业及应用孵化创新提供 AI 基座。新一代 AI 基础设施跟随产业布局，采用“大中心+节点”模式，构建起覆盖整个区域的算力网络，并通过建（设）运（营）联动促进区域经济的一体化和智能化发展。

5. 新一代 AI 基础设施为政务服务、产业升级和科研创新等领域带来了前所未有的社会价值。将原本分散、碎片化的政务应用，通过“一模通办” 为政务服务提质增效。将加快推进传统产业上下游各个环节的智能化转型，催生新业态、新模式的不断涌现。加速科学实验的自动化和智能化，激发人工智能驱动科学研究（AI for Science）的新范式。

6. 本白皮书提出业界首个“新一代人工智能基础设施评估体系”，通过产品技术、战略愿景、市场生态三大维度、十二个评估指标，对 AI 基础设施厂商进行定性和定量的全面评估。SenseCore 商汤大装置，成为市场领导者，在各个评估指标的得分超过厂商平均分，并在市场响应、市场认知、产品战略、工程化建设四个评估指标拿到满分。

7. SenseCore 商汤大装置在产品服务能力呈现出较强的产品实力和技术积累，不仅超前布局了算力基础设施，还通过布局 MaaS 平台，在自身大模型业务的加持下，形成了整套 AI 基础设施产品架构，满足客户大模型训练、生成式 AI 应用的大规模落地需求。

8. 新一代人工智能基础设施将会通过支持大模型的爆发式发展，带来知识工程的生产力变革，重构软件生态，颠覆原有数字经济霸主，并随着本身的技术革新和突破，实现边际成本持续下降，边际效益持续增长等特征，进而实现 AI 算力成本的持续下降，真正带来普惠 AI。

导语：AI 新基建开启“三浪变革”

第一浪是“知识生产力变革”，大模型是知识工程的生产力变革，天然具有跨领域知识的连接性。上一次知识革命是 11 世纪的毕昇发明的泥活字印刷术、15 世纪的古登堡发明的铅活字印刷术，让人类千年历史中积累的庞大知识工程通过印刷书籍形式推广传承，知识从手工抄写到活字印刷速度提升了 118 倍，自此浩瀚的知识源源不断地从印刷作坊以令人惊叹的速度向全球传播，堪称中世纪的“知识互联网”。在比尔盖茨的《未来之路》中提到，在谷登堡印刷革命之前，整个欧洲大陆大约只有 3 万册书，几乎都是圣经或圣经评注性著作，而到了 1500 年，各类题材的图书猛增到 900 多万册。各种传单和其他印刷物影响了政府、宗教、科学以及文学。宗教精英圈子以外的人士第一次有机会接触到书面信息。据多方研究数据表明，大型语言模型显著提高知识学习速度、知识检索速度、知识传播速度、知识推荐准确性，具有跨语言、跨学科领域、跨信源的独特优势。在人机协同模式下，大型语言模型将人类科学论文的阅读时间缩短 40%，知识搜索时间缩短 20%，而这仅仅是 ChatGPT 出现一周年的“起点”，鉴于大型语言模型远超人类的超高速学习能力，预计将在 2026 年学习完所有人类历史上的高质量文本数据[1]。**人类的知识革命大幕刚刚开启，高新科研、三大类产业、公共服务的知识型工作范式正在遵循“计算->数据->模型->服务”链条重构。**

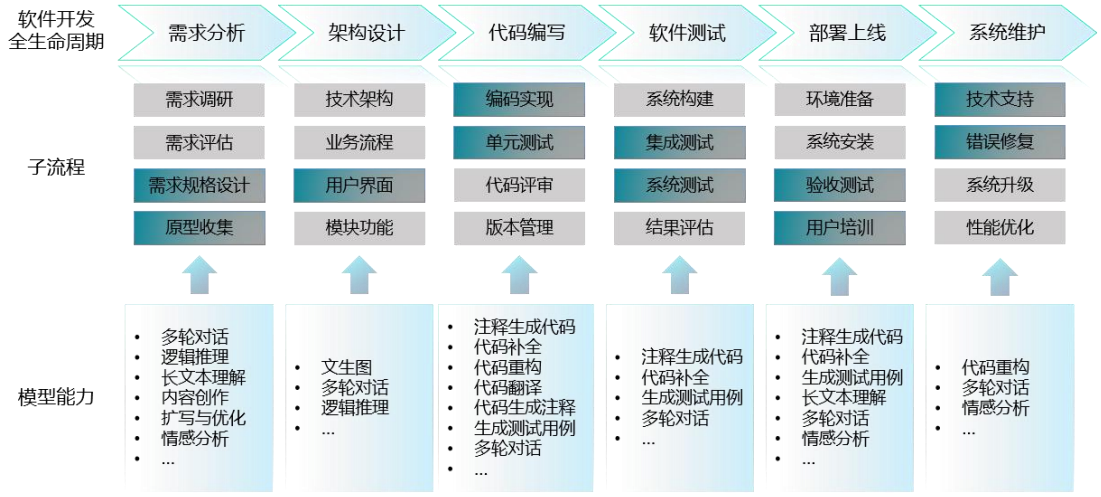
第二浪是“软件变革”，每次软件大革新，都会诞生新的超级平台，颠覆原数字经济霸主，从 Windows、AppStore 到 GPTs 都不例外，当前智能编程助手改变代码生产流程，大语言模型成为新一代 AGI 服务入口、软件调度枢纽。

[1] Epoch AI Research 研究机构预测，大模型对数据的需求正在飞速增加，人类历史上可用于训练的高质量文本将在 2026 年“耗尽”。

20 世纪 90 年代，未来学家雷·库兹韦尔发现指数级发展的规律：“一旦技术变得数字化，即被编辑为 0 和 1 表示的计算机代码，它就能够脱离摩尔定律的舒服，开始呈指数级加速发展。”所以数字经济中每一代超级平台企业都是软件创新型企业。中国程序员人数位居全球第二，世界上最好的开发语言应是中文，例如商汤科技发布的“代码小浣熊”Raccoon 智能编程助手，覆盖软件需求分析、架构设计、代码编写、软件测试等环节，支持中文、英文注释生成代码、跨编程语言翻译、单元测试用力生成、代码修正（改 Bug）、代码重构、编程技术知识问答，在 Python、Java、C、C++、Go、SQL 等 30 多种主流编程语言，以及 VS Code、IntelliJ IDEA 等主流集成开发环境(IED)上，提升开发者编程效率超过 50%，并在以 71%的一次通过率刷新 HumanEval 测试集成绩（GPT-4 一次通过率 67%）。从此人类程序员将 80%的代码量交由语言大模型编写，人类开发专家的时间和精力逐步转移到更具创新性和高价值的工作中，商汤称其为软件 2.0 时代的“新二八定律” ***（见图 1）***。

另一方面，多篇权威论文显示，大型语言模型能够面对复杂任务，灵活自动实现多软件串行、多模型协同组合，例如 AI Agent、MoE 架构（Mixture-of-Experts）、综合型智能客服、GitHub Copilot 等，能在日常使用中跨模型共享成果、快速学习迭代、增强安全性与伦理性保障。**在庞大 AI 算力规模、训练数据集基础上，新一代 AI 原生软件应用，导致“传统软件智能化，智能软件枢纽化”全面普及，尤其是那些能满足目前还难以预知需求的新工具，新一代青少年将在新兴 AI 软件与 MaaS 模型化创新思维逻辑上成长起来，并将新型生产力软件带入办公室与家庭。**

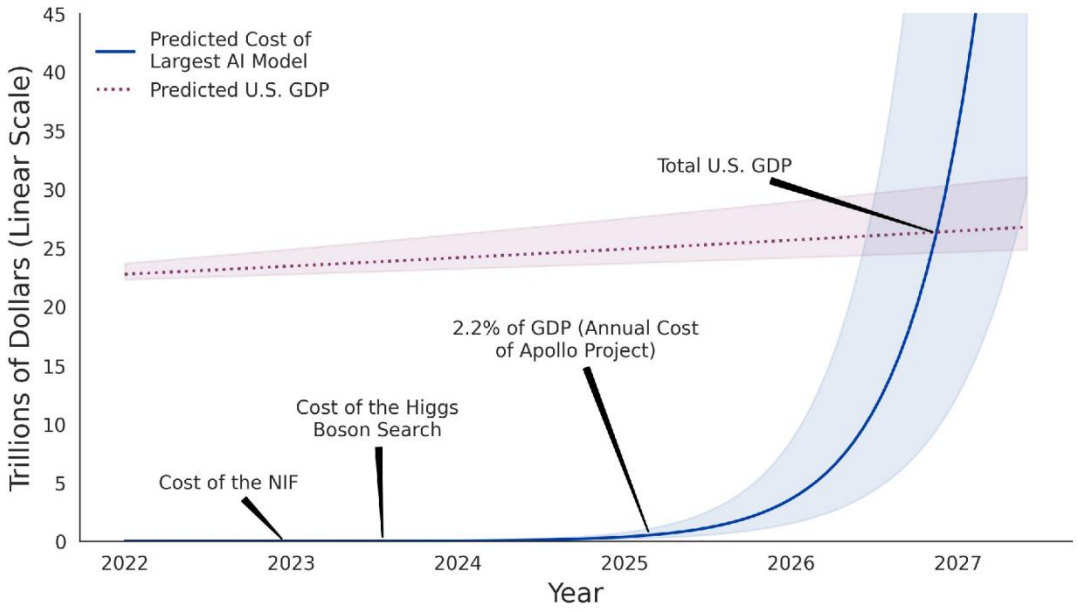
图 1：大语言模型智能编程助手，赋能软件开发提效降本



第三浪是“AI 计算变革”，在大型语言模型的 Scaling Law（规模定律）指数级算力需求，与线性增长的区域基建投入矛盾下，AI 算力基础设施将迎来大量技术工程创新，持续降本增效，普惠优势让 AI 真正成为赋能千行百业的通用型基础设施，同时“百模大战”变为 AI 产业专业化分工。据 AI Now《计算能力和人工智能》报告指出，早期 AI 模型算力需求是每 21.3 个月翻一番，而 2010 年深度学习后 (小模型时代)，模型对 AI 算力需求缩短至 5.7 个月翻一番，而 2023 年，大模型需要的 AI 算力需求每 1-2 个月就翻一番，摩尔定律的增速显著落后于社会对 AI 算力的指数级需求增长速度，即“AI 超级需求曲线”遥遥领先传统架构的 AI 算力供给，带来了 AI 芯片产能瓶颈、涨价等短期市场现象。CSET (Center for Security and Emerging Technology) 在《AI and Compute》报告中预测：“在计算价格没有任何变化的情况下，尖端模型成本预计将在 2026 年 6-11 月超过美国 GDP ***（见图 2）***。”未来学家雷·库兹韦尔认为，从 1890 年到现在，人类计算设备的（单位时间）的运算能力一直在成倍增强，每当一项指数型技术（例如符合摩尔定律的芯片技术）的实用性达到极限时，就会有另一项

技术取而代之。所以，针对大模型高昂的训练成本、有限的 GPU 供应量、芯片间通讯瓶颈的核心挑战，各国均采用大规模智能基建资源投入，并在 AI 芯片、智能算力集群、大模型架构、专用模型加速等技术栈环节创新突破，相信在未来 3 年通过一系列基础设施的技术革新，持续降低 AI 计算整体成本（采购、建设与运营），释放出各行各业的生成智能全民应用创新能力，尤其是推理算力成本下降，对中国 AI 2.0 的大市场、大用户量至关重要。**同水电煤等平价公共服务一样，人人用得起 AI 算力，人人训得起 AI 数据，人人做得好 AI 模型。**

图 2：大模型算力的成本压力（来源：CSET）



Note: The blue line represents growing costs assuming compute per dollar doubles every four years, with error shading representing no change in compute costs or a doubling time as fast as every two years. The red line represents expected GDP at a growth of 3 percent per year from 2019 levels with error shading representing growth between 2 and 5 percent.

一、大模型、生成式 AI 推动 AI 2.0 时代到来

2023 年是人工智能发展的分水岭，大模型、生成式 AI 的发展带动了人工智能领域的范式转换，AI 2.0 时代已经来临。在此之前，人工智能通过模式检测或遵循规则来帮助分析数据和做出预测，更像是一种“分类器”，而 AI 2.0 时代则开启了新阶段：基于大模型的生成式 AI。生成式 AI 可以通过数据训练进而模仿人类的创造过程，将人工智能从传统的“分类器”进化成“生成器”。这样本质上的变化，让 AI 发展到了一个全新的时代***(见图 3)***。Gartner 预测，到 2027 年，高速增长的生成式 AI 将会贡献全球人工智能支出的 42%，规模将超过 1800 亿美元，2023 年到 2027 年的复合增长率高达 169.7%[2]。

另外，作为生成式 AI 发展的基础，大模型也在高速发展。IDC 数据显示，截止 2023 年 11 月底，中国市场发布的大模型已经超过 300 个。生成式 AI 的颠覆性潜能得到越来越多的企业认可，企业不再追问何为生成式 AI，而是希望了解生成式 AI 的投入能带来哪些具体业务价值。Gartner 预测，到 2026 年，超过 80%的企业将使用生成式 AI 的 API 或模型，或在生产环境中部署支持生成式 AI 的应用，而在 2023 年初这一比例不到 5%[3]。

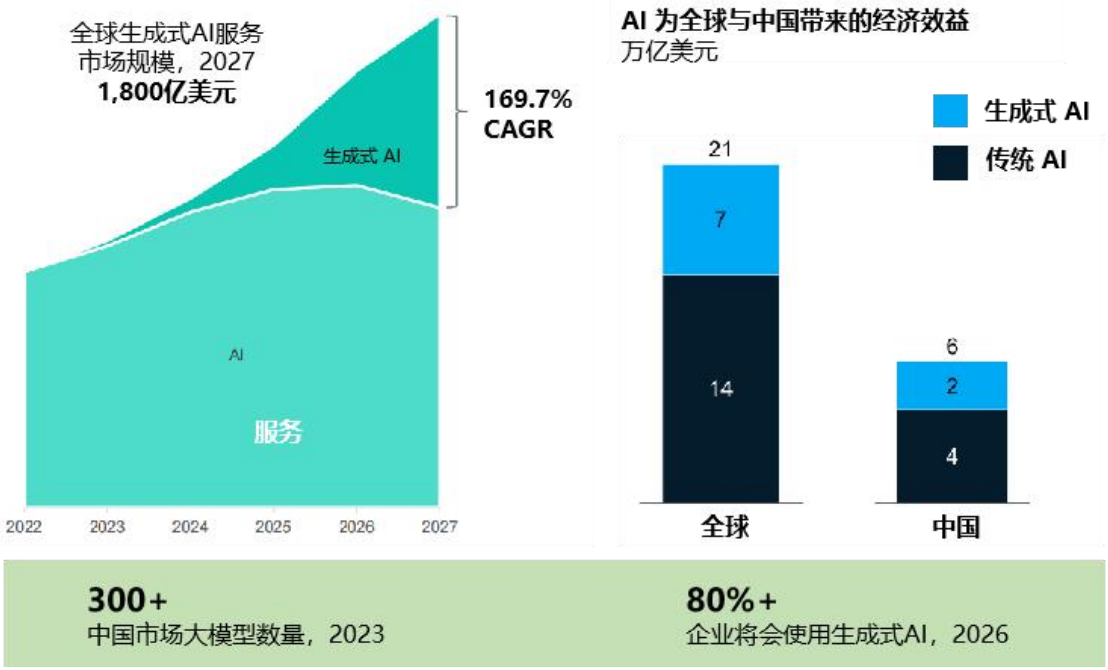
技术变革带动场景拓展，生成式 AI 正在从热烈讨论走向应用落地，其价值创造潜力极为惊人，麦肯锡预测，生成式 AI 有望为全球经济贡献约 7 万亿美元的价值，并将 AI 的总体经济效益提高 50%左右；中国则有望贡献其中约 2 万亿美元，将近全球总量的 1/3[4]。

---

[2] “Forecast Analysis: Artificial Intelligence Services, 2023-2027, Worldwide”, Gartner, October 2023
[3] “Top Strategic Technology Trends for 2024”, Gartner, October 2023
[4] 生成式 AI 在中国：2 万亿美元的经济价值，麦肯锡，2023 年 9 月

图 3：生成式 AI 驱动 AI 市场规模化发展，并带来全新经济效益

（来源：Gartner、麦肯锡、IDC）



1. 生成式 AI 推进产业规模化，AI 无处不在愿景加速实现

生成式 AI 呈爆炸式增长，使 AI 由之前的点状、创新应用，逐步开始在业务流程的各个环节应用部署，企业在积极探索相关价值，以期增强自身业务的竞争优势。麦肯锡调研显示，已有 1/3 的企业受访者表示，其所在组织会在至少一项业务职能中经常使用生成式 AI 应用***（见图 4）***。企业通过以下一系列举措，不断推动 AI 无处不在的愿景实现：

- **加强生成式 AI 领域的投资，应用部署获得持续动力。**自从 ChatGPT 发布以来，企业在 ICT 领域的投资发生了调整与变化，为了更好的跟上此轮技术变革所带来的潜在红利，企业将更多的 ICT 预算投入到生成式 AI 领域，并将会从中获得客观的收益。IDC 调研显示，已有 24% 的中国企业在生成式 AI

上投入资金，69%企业正筛选潜在应用场景或开始测试和概念验证，到 2026 年，中国 40%的企业将掌握生成式 AI 的使用，共同开发数字产品和服务，从而实现与竞争对手相比两倍的收入增长。

- **改变现有 AI 战略，驱动生成式 AI 覆盖公司业务全流程。**企业组织正在改变自身的人工智能战略，围绕人工智能战略的愿景、路线图、用例、治理、以及相应的人才都发生了全面的变化。AI 1.0 时代，企业组织在制定一个典型的人工智能战略更多考虑的是一个长远的规划，并且碎片化的布局，这些随着 AI 2.0 时代生成式 AI 爆发增长所带来的日新月异而发生彻底改变，短期目标、快速行动并逐渐覆盖关键业务成为人工智能战略的新内核，更关键的转变则是用例方面，从之前的预测分析、自动化应用场景，转向内容生成和创造。同时，由于生成式 AI 将会成为不可或缺的生产力工具，培训每个员工如何负责任地使用生成式 AI 工具也成为重点**（见图 5）**。
- **拥抱生成式 AI，促使 AI 与员工实现协同创新。**生成式 AI 扩大了人类的专业知识、创造力和知识范围，提高了人类工作的效率。更关键的是，生成式 AI 使得新洞察、新模式、新能力的创造变得更为清晰，创新的本质是可能性的不断组合，确定最有前景的组合项后，对其进行改进直到实现。人类团队只能探索创新解决方案的一小部分，而生成式 AI 可以帮助人类能够利用更多变量在短时间内探索更多解决方案可能性，并且能够以最小化成本撬动更多价值的产出。Gartner 预测，到 2026 年，将会有超过 1 亿人将与“机器人同事（合成虚拟同事）”协同工作[5]。

[5] “The Future of AI: Reshaping Society”, Gartner, July 2023

图 4：各地区、行业和资历级别的受访者表示，他们已经在使用生成式 AI（来源：麦肯锡）

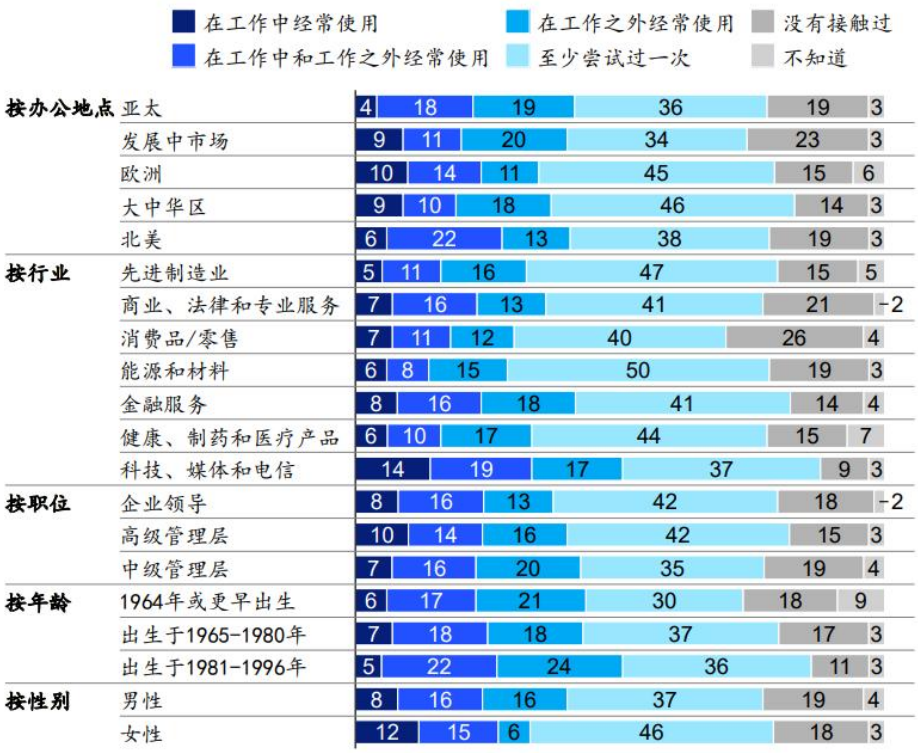


图 5：AI 2.0 时代，企业需要重新制定 AI 战略

2. 产业链成熟分化，基础设施成为 AI 产业发展基座和保障

企业积极拥抱大模型、生成式 AI 的态度，加速了 AI 应用逐步向产业深水区发展，面临千变万化的业务需求和标准，为了更好的应对不同的业务诉求，AI 产业链将会一步成熟分化，上下游的产业角色和环节不断增多，开始需要全新的基础设施来实现更好的支撑，其带来的影响如下：

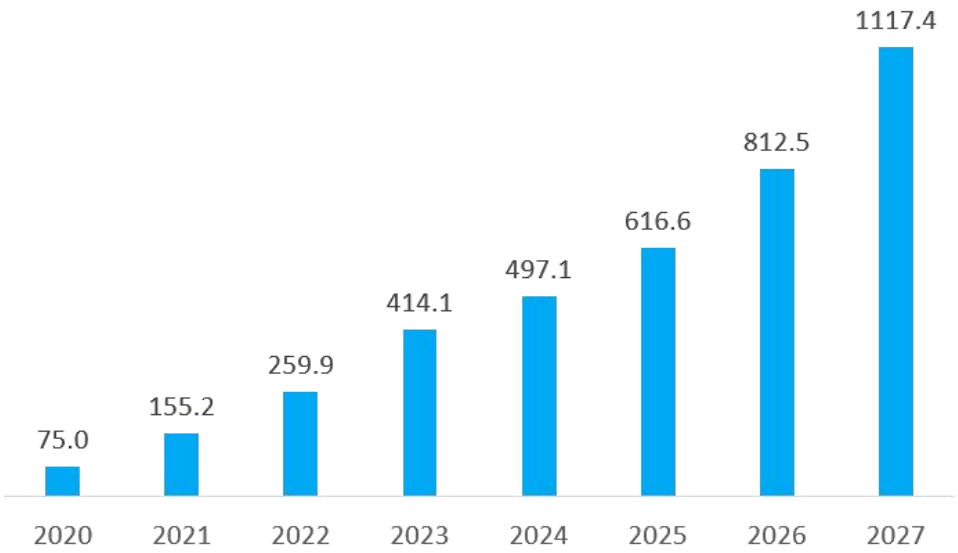
- **智能算力成为 AI 产业发展的关键支撑要素。**大模型训练趋势企业将更多地使用 AI 就绪的数据中心设施或 GPU 集群，从而缩短部署时间，降低设施的长期投资成本。适合大模型训练的智能算力已经成为算力增长的主要动力。IDC 预测，2022 年中国智能算力规模达 259.9 EFLOPS，2023 年将达到 414.1 EFLOPS，预计到 2027 年将达到 1117.4 EFLOPS ***（见图 6）***。2022 - 2027 年期间，中国智能算力规模年复合增长率达 33.9%，同期通用算力规模年复合增长率为 16.6%[6]。
- **人工智能生产范式转向以大模型为核心的开发路径。**在 AI 1.0 时代，AI 应用的开发处于“精耕细作”的阶段，开发人员基于场景化的数据集用明确的代码去表达程序执行的逻辑，并不断基于业务发展而进行迭代，而随着业务场景从通用场景发展到长尾、碎片场景，该模式则逐渐显现出开发成本高，精确度不佳等一系列挑战，在某种程度上，反而限制了 AI 产业的进一步发展。而在 AI 2.0 时代，在基模型+人工反馈的强化学习相结合的加持下，人工智能应用的开发进入“规模化”阶段，体现出“大力出奇迹”的特色。面向业务逻辑对基模型进行微调，辅助提示词工程来开发相应的生成式 AI 应用，进而更快速、低成本、高精度的覆盖更多业务场景，这使得 AI 产业进

[6] 《中国人工智能计算力发展评估报告，2023-2024》，IDC，2023 年 12 月

入了一个高速发展且无处不在的全新时代***（见图 7）***。

- **作为新的生产力工具，生成式 AI 应用发展进入大航海时代。**伴随基模型的高速成熟化发展，生成式 AI 应用也迎来爆发式增长***（见图 8）***。最早，以 ChatGPT、Midjourney 为代表的文生文、文生图应用推向市场并获得高速增长的用户群体。随后，音频生成、视频生成、多模态生成类的应用，以及面向不同行业领域或用户群体的工具类应用，如代码生成、Copilot、数字人、营销工具、聊天助手等，不断推向市场。2023 年 11 月，OpenAI 推出 GPTs 并计划打造 GPT Store，让用户无需代码，结合自己的指令、外部知识和能力创建自定义版本的应用，这种客制化的模式和清晰的商业化模式，让生成式 AI 应用的开发主体由数量不多的 AI 厂商走向海量 AI 开发者[7]。

图 6：中国智能算力规模及预测，2020-2027，基于 FP16 计算，EFLOPS（来源：IDC）



[7] “Introducing GPTs”, OpenAI, November 2023

图 7：AI 2.0 时代，人工智能的生产范式发生了根本性改变

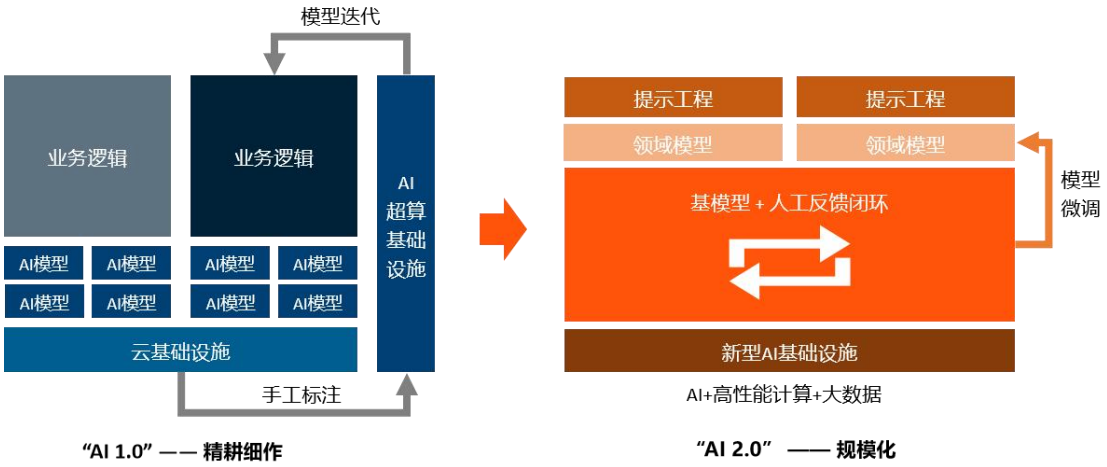
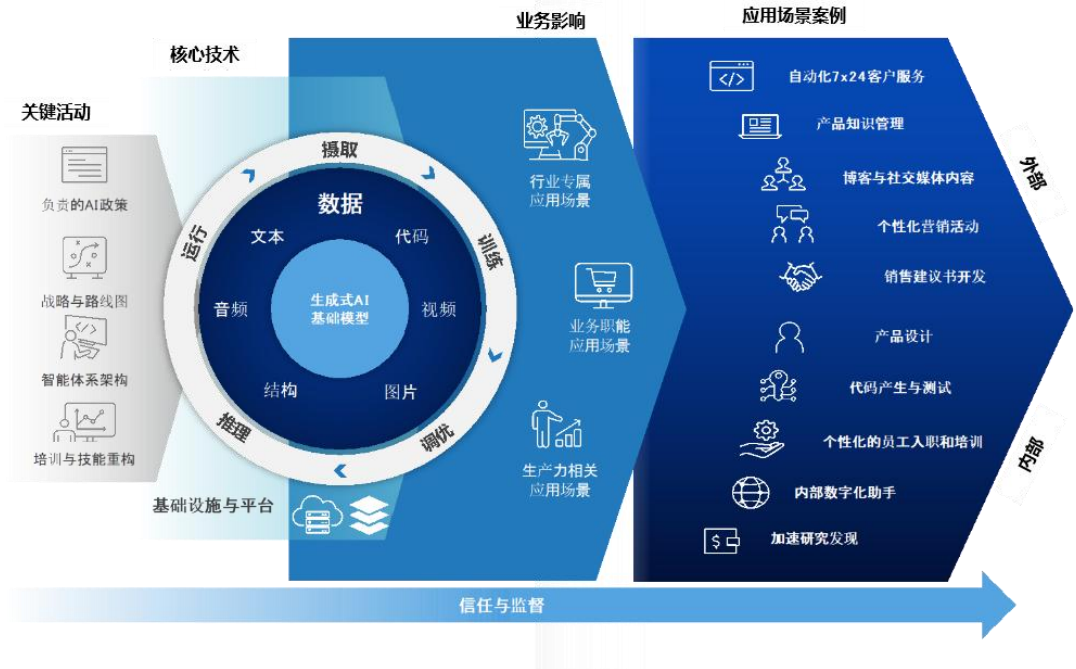


图 8：AI 2.0 时代，人工智能产业迎来更加繁荣的"大航海时代"（来源：IDC）



二、AI 2.0 时代对 AI 基础设施提出了全新要求

进入 AI 2.0 时代，传统针对移动互联时代应用、以 CPU 为中心的云计算基础设施，无法满足大模型训练、生成式 AI 应用爆发所带来的挑战，这些新的挑战对 AI 基础设施的关键环节都提出了全新的要求，包括算力、算法平台、数据，以及围绕三个环节的工程系统建设。

1. 传统计算基础设施无法满足大模型、生成式 AI 的新要求

大模型训练、生成式 AI 应用不仅对 GPU 或异构计算的需求大幅增加，传统 CPU 算力已经无法满足；还对 GPU 集群的计算效率、稳定性等方面的提出诸多要求，算力不是一个简单的堆砌，而是要转为大模型而优化的复杂的系统性工程，再加上巨大的投资，如何保持稳定性和高效率也成为关键，展开来看：

- **以 GPU 为核心的 AI 算力需求爆发性增长。**以 OpenAI 为例，训练一次 1750 亿参数的 GPT-3 模型大概需要的算力约为 3640PFlops-day，共使用了 1024 块 A100(GPU)训练 34 天。随着模型参数量不断升级，AI 算力需求也在持续递增。过去四年，大模型参数量以年均 400%复合增长，AI 算力需求增长也超过 15 万倍，远超摩尔定律[8]。例如，GPT-4 参数量大约是 GPT-3 的 500 倍，用了约 2 万-3 万张 A100,训练 1 个月左右的时间。除大模型训练外，随着生成式 AI 应用爆发，高并发推理也将进一步推高算力需求，未来或将远远超过训练阶段的算力当量***（见图 9）***。
- **高性能和高效率成为算力基础设施的关键。**为了更好支持大模型训练，多机多卡组成大集群分布式训练成为必选。但大集群不等于大算力，在分布式训

[8] 《WOT 全球技术创新大会：创新不止，实战为王》，东方财富网，2023 年 6 月

练下集群中由于网络通信或数据缓存等问题都会造成大模型训练效率降低。例如，一般千亿、万亿参数规模的大模型，训练过程中通信时间占比最高可达 50%[9]。如果通信互联不好，会影响大模型训练效率，也会限制算力集群的进一步扩展，这就要求集群具备高速互联的网络连接。并行训练要求网络基础设施具备高度可靠，一条链路的负载不均导致网络堵塞，就会成为系统短板，影响到数十个甚至全部 GPU 节点信息同步***（见图 10）***。此外，大模型训练过程中会通过 Checkpoint 来保存模型参数（权重），进而实现大模型训练的连续性。但是，传统训练方式下当模型参数量大时，Checkpoint 写入时间会变久，导致导致 GPU 利用率降低。例如，1750 亿参数的 GPT-3 模型，假设文件系统写入速度为 15GB/s，完成一次 Checkpoint 需要 2.5 分钟，也就相应造成 2.5 分钟的资源浪费。因此，支撑大模型训练的算力资源，不仅需要在集群硬件层面提升，还需要结合软件层面进行优化设计。

- **独占式、大规模、长时间训练对 GPU 集群稳定性提出更高要求。**大模型训练需要长时间占据规模庞大的 GPU 集群，这导致单个节点发生故障就使得整个训练中断，且故障原因和位置难以迅速界定。以 Meta 的 OPT-17B 训练为例，理论上在 1,000 个 80G A100 上训练 3,000 亿个单词，需要 33 天，而实际训练却用了 90 天，期间出现了 112 次故障，其中主要是硬件故障，导致手动重启 35 次，自动重启约 70 次[10]。节点故障不仅造成训练时间被拉长，也对算力资源带来了巨大浪费。因此，集群训练稳定性非常重要，对集群建设提出更高要求。例如，集群是否具备故障实时监测、断点续训、故障节点自动隔离等能力，以及在故障发生时能否快速定位、迅速恢复等。

[9] 《大模型需要大算力，但光靠 GPU 也不行》，21 世纪经济报道，2023 年 6 月

[10] 《如果没有 AI 算力，大模型这场战役我们可能胜不了》，量子位，2023 年 12 月

图 9：AI 算力需求呈指数级增长，用以满足大模型开发和实践（来源：Epoch）

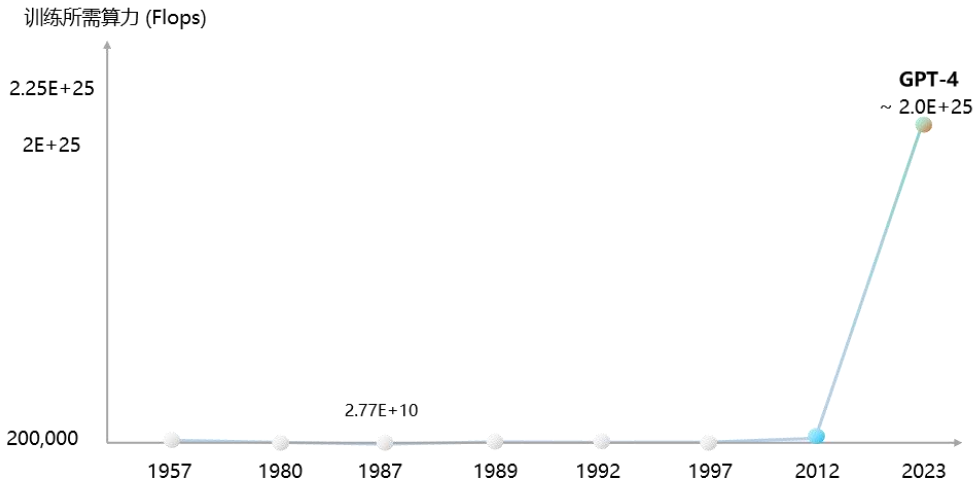
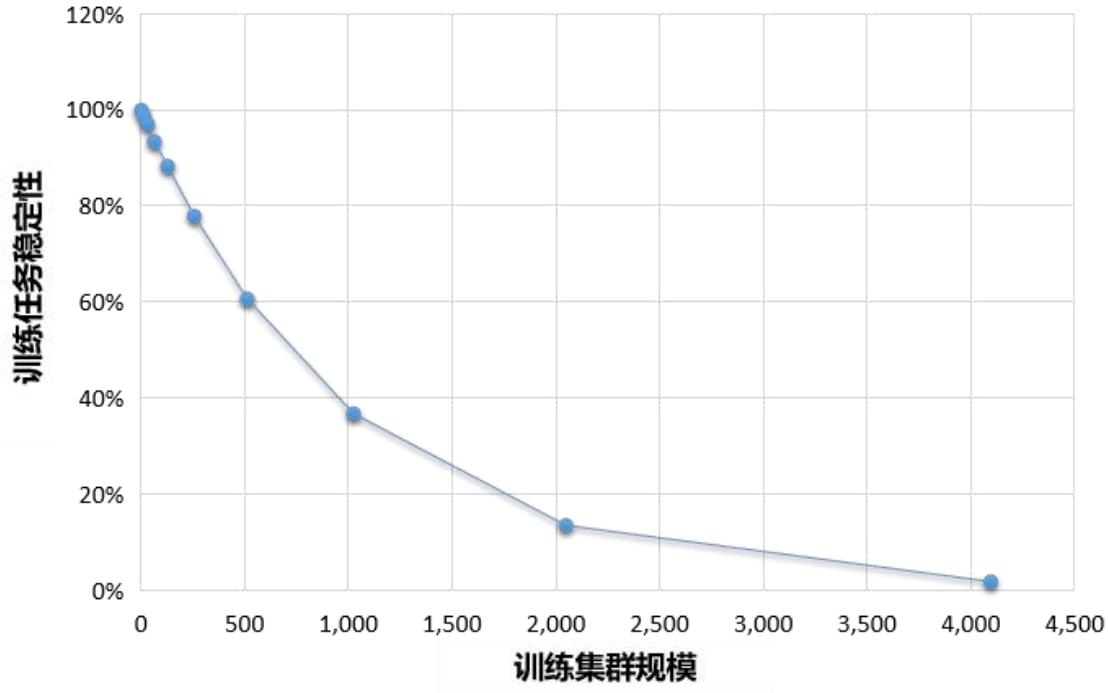


图 10：大模型训练任务的稳定性，随着训练集群规模的扩张而递减



2. 数据质量和效率决定大模型的高质量发展之路

高质量数据决定大模型性能和价值观，对数据的获取、清洗、标注等工作带来了更大挑战，需要更高效的 AI 数据管理流程来匹配大模型时代的新需求。而大模型的训练和应用过程还可能涉及用户隐私和敏感数据等，需要采取有效的数据治理手段来保障隐私和数据安全。

- **构建性能强大和价值对齐的大模型，数据质量和效率是关键。**不同来源数据质量参差不齐，存在重复、无效、虚假或敏感等数据，会直接影响模型性能及价值观。例如，训练数据固有偏见会导致模型产生偏见内容，需要对原始数据进行清洗、标注等预处理过程来保障数据质量和价值对齐。传统数据处理“作坊式”的工作模式，已无法满足大模型训练和迭代激增的“工业化”数据需求。一方面，训练大模型的预处理数据量大，一般可达到 TB 甚至 PB 级别，远多于传统数据规模；另一方面，频繁的模型迭代、再训练也需要加快增量数据的预处理节奏。打造高效的“智能化数据处理流水线”成为关键，弥补传统重人力投入带来的高成本、低效率等问题。
- **保障数据安全和用户隐私，需要更高效的数据治理手段。**企业在使用生成式 AI 将会面临更加突出的用户隐私和数据安全问题。例如，企业开发人员使用 AI 代码辅助生成工具时，一般需要上传企业已有代码库，使大模型给出更精准的代码预测结果；企业营销人员上传过往的营销数据生成高质量的营销内容。这些上传的数据可能关系到用户隐私或涉及企业核心机密，如果保护不当或会造成严重的数据泄露，对用户造成不可逆损害。IDC 全球 2023 年生成式 AI 市场调研数据显示，用户在选择 AI 软件供应商时，强大的数据安全性是最重要的参考指标之一。因此，在大模型训练和交互时，如何将这些

上传数据进行充分隔离、安全保护，这对数据治理提出了很高的要求。

3. 大模型需要全新的 AI 平台服务模式

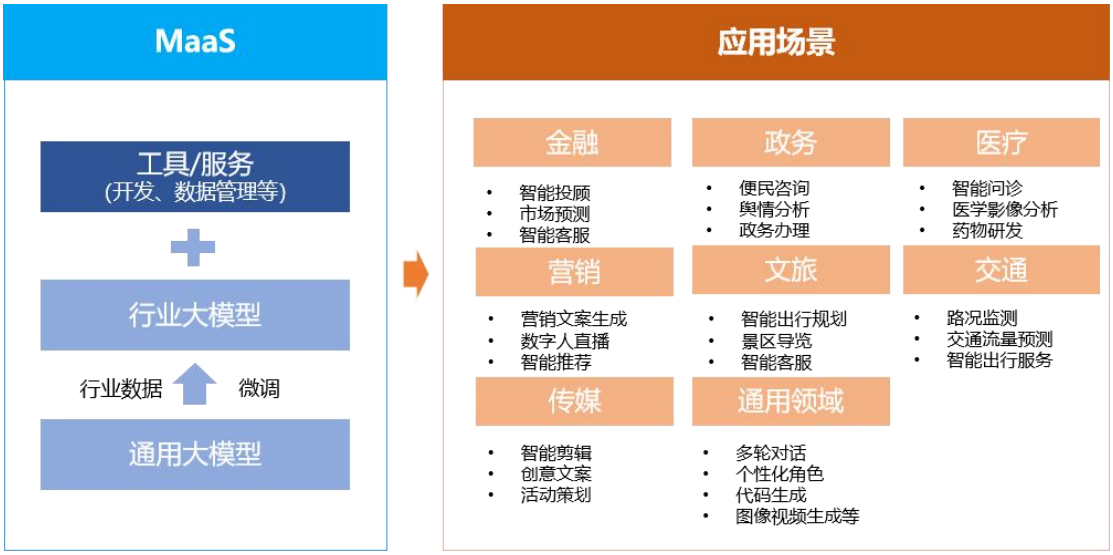
大模型应用能够帮助企业更高效率的实现商业目标，但对于绝大多数企业而言，自研大模型成本高，且模型设计、训练、调优等环节对开发人员专业能力要求较高。MaaS（Model as a Service，大模型即服务）代表着一种全新的 AI 云服务范式，它将大模型作为 AI 基础设施的核心组成，以云服务方式提供给开发人员和企业进行更高效的工业化开发***（见图 11）***。目前，包括微软、华为、百度、商汤等厂商均推出 MaaS 服务。MaaS 降低了企业享受大模型、生成式 AI 红利的门槛，具体来看：

- **MaaS 加快了 AI 应用开发过程，提高了创新迭代速度。**MaaS 平台将预训练好的大模型与开发工具、数据管理一系列等功能封装在一起供开发人员直接调用，大幅节省了企业从零自建大模型及训练调优所耗费的时间和精力，加快了 AI 开发和应用部署速度，使企业能够快速将 AI 功能引入到自身业务场景中，缩短了 AI 新产品、新服务、新模式的上线时间，也加快了创新迭代速度，提升了企业市场竞争力。
- **MaaS 降低了企业成本投入，推动 AI 与各行各业的深度融合。**AI1.0 时代，由于小模型的场景适用性低且开发成本、专业要求都很高，AI 在传统行业的渗透率只有 4%。大模型时代，基于“基础大模型+微调”，不仅大大提升了场景适用性，同时 MaaS 模式也使得企业可以直接调用已训练好的大模型能力，这大大降低了企业 AI 开发成本和 AI 专业门槛，使得企业更愿意在内部更大范围的推进与业务相结合的 AI 创新，促进 AI 与行业的深度融合，行

业 AI 应用的渗透率将全面提速。

- **MaaS 促进大模型生态体系建立，推动大模型应用规模化落地。**MaaS 模式有助于 AI 产业链的高效分工，其中一部分技术实力强和 AI 专家资源丰富的厂商成为 MaaS 主要提供者，将重心侧重在基础大模型能力，以及平台化工具和服务效率上的持续提升，并通过平台开放、开源社区等方式吸引更多的企业和开发者汇集，形成百花齐放的大模型应用开发生态，从而满足更大范围、更多细分场景的 AI 需求，推动应用规模化落地。

图 11：MaaS 平台帮助企业更好的调用大模型能力



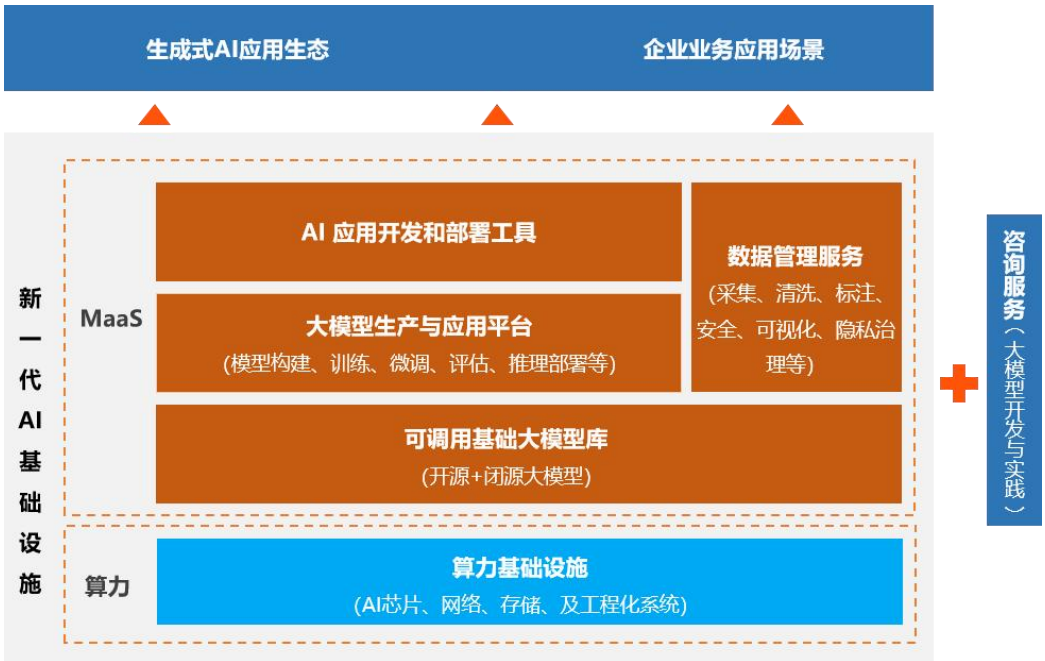
三、新一代 AI 基础设施的定义、特点和价值

AI2.0 时代需要新一代的基础设施来支撑大模型的训练与推理、生成式 AI 应用的规模化落地，其核心要素，如算力、数据服务、大模型服务都需精细化的设计和重构，而非简单的服务器或 GPU 实例的堆砌。新一代 AI 基础设施定义：

以大模型能力输出为核心平台，集成算力资源、数据服务和云服务，专门设计用于最大限度提升大模型和生成式 AI 应用的表现：数据准备与管理、大模型训练、推理、模型能力调用、生成式 AI 应用部署。企业通过新一代 AI 基础设施开发和运行生成式 AI 业务和客户应用程序，以及基模型和行业模型的训练与微调（见图 12）。

在落地实践中，厂商还会基于自身的经验积累，针对用户在训练和使用大模型时面临的 AI 技术问题，为用户提供围绕大模型开发实践的咨询类服务。

图 12：新一代 AI 基础设施主要由算力、MaaS 及相关工具构成



算力基础设施，为大模型训练和推理提供全面的计算、存储等产品及服务，具有“大算力、高协同、强扩展”的基本特性：

- **由高性能异构集群组成强大的算力底座作为算力支撑，**具备高互联的计算网络、高性能的文件存储和大规模的 AI 算力资源。
- **高度的软硬件系统协同为保障，护航大模型任务的高效、稳定运行。**在建构硬件层面的算力集群过程中，融合大模型分布式训练对计算、网络、存储的需求特点，高度集成 AI 软件能力，充分关注数据传输、任务调度、并行优化、资源利用、故障监测等，设计和构建高性能、弹性灵活、高容错的集群系统，保障训练和推理的高效、稳定运行。
- **具备非常强的线性扩展能力，提供弹性灵活的云原生服务。**将 GPU 等 AI 算力资源容器化、资源池化，在高弹性、高可用、高安全性的云原生架构下，使算力的管理能力拓展到整个智算中心，实现对 AI 计算资源的灵活调度、远程共享等目标，可以轻易支持万卡万参的大模型训练迭代。

MaaS 平台层为大模型应用落地提供完整的服务和工具链体系，包括基础大模型库、大模型生产平台、数据管理平台、应用程序开发等主要部分。针对不同用户需求，MaaS 平台层可以提供不同服务类型：

- **提供预构建的基础大模型及 API，**包括开源和闭源的大模型，允许用户调用 API，直接获取大模型相关的能力和服务，降低客户的使用成本，快速满足多个业务场景需求。
- **提供一站式大模型开发工具及服务，**包括模型训练、微调、评估、推理部署等，支持用户训练新的模型，或根据不同行业和业务场景进行微调，快速生成满足自身需求的专属大模型，强化大模型在细分领域的专项能力，推动大

模型在不同行业领域的快速落地。

- **提供 AI 原生应用开发工具，**满足用户基于大模型开发 AI 原生应用需求，赋能和重塑上层 AI 应用生态发展，为终端用户提供更卓越的生成式 AI 体验。
- **提供预构建的高质量数据集及 AI 数据管理服务，**包括数据清洗、标注、安全、合规等，降低用户在数据层面上的投入和成本，保障隐私和数据安全。

1. 新一代 AI 基础设施的主要特点

新一代 AI 基础设施不是传统云的 AI 化，两者具有明显定位和发展路径的差别***（见图 13）***。新一代 AI 基础设施主要面向产业用户，为超大模型研发训练、区域行业及应用孵化创新提供 AI 基座，并跟随产业区域落地向周边辐射，通过可持续运营带动区域经济智能化发展。

- **“建运一体”的智算中心充分发挥基础设施效益，支持区域智能化经济发展。**智算中心不仅是新一代 AI 基础设施的物理载体，还是集公共算力服务、数据开放共享、智能生态建设和产业创新聚集四大功能于一体的综合服务平台。国家信息中心《智能计算中心创新发展指南》测算，在智算中心实现 80%应用水平下，区域对智算中心的投资可带动 AI 核心产业增长 2.9~3.4 倍、带动相关产业增长约 36~42 倍。因此，“建好”智算中心不是目的,只有“用好”才能发挥效益。在建设规划阶段，需要以产业生态为导向，强调对区域产业、科研等应用场景的支撑，选择合理的建设和运营模式，进行集约化建设，并在建成后实现可持续运营，帮助当地更好的消化算力资源，以及促进智能产业生态发展和 AI 人才培养，发挥智算中心普惠高效的赋能效果。
- **“大中心+节点”布局，建设跨地域互补、协同调度的超大规模 AI 算力网络。**

大模型研发及预训练需要低成本、大规模的 AI 算力资源支持，而在应用阶段会更注重满足训练和推理一体需求。如何平衡不同需求下的算力供给，最大效率的使用算力资源非常关键。以“大中心+节点”模式建设算力一张网，实现训推算力协同调度。围绕经济中心布局“大中心”，以低成本大规模算力集群为载体面向万亿参数模型训练与部署；围绕产业基础好的区域落地算力节点，结合长效运营来满足产业训推一体的算力需求，并通过节点布局与大中心联动扩展算力网络辐射范围，跨地域支撑训推算力协同调度。

- **侧重国产化生态建设，增强基础设施的自主可控。**搭建基于国产软硬件的 AI 基础设施，研发全栈国产化大模型，逐步形成自主可控的 AI 大模型产业生态关乎国家安全和战略发展。以芯片国产化适配为例，华为自研昇腾 910 已实现与包括科大讯飞、紫东太初、美团等科技企业的战略合作，基于华为昇腾原生研发、适配的大模型已经超过 30 个，占据中国大模型的数量接近 50%[11]。商汤科技深度学习框架已支持多家国产化芯片训练，适配领域包括图像分类、检测、NLP 等主流领域，推动算法与国产化芯片适配工作。

图 13：新一代 AI 基础设施面对的是不同于传统云的业务要求

[11] 《华为宣布昇腾 AI 集群全面升级 推出首个万卡 AI 集群》，中国新闻网，2023 年 7 月

2. 新一代 AI 基础设施创造社会价值

新一代 AI 基础设施降低了大模型开发和应用的门槛，在政企服务、产业和科研创新等方面创造更大的社会价值（**见图 14）**。具体来看包括三方面：

- **构建政务大模型，“一模通办”为政务服务提质增效。**将原本分散、碎片化的政务应用，用一个性能强大、底座统一的大模型来承载，将大模型能力融入到数字政府的全流程场景中，无需为不同场景重复开发，通过“一模通办”并简化交互入口，全面提升地方政府智能化治理能力，推动各类智能惠企、便民服务的高效、精准实施，让企业、市民都能更好的享受城市公共服务。例如，面对海量的政务数据，依托政务大模型，能够帮助政府快速洞察热点事件、分析惠企政策落地等情况，及时掌握市民的关注点、惠企政策的应用成效等，为后续政策的制定和实施提供支持，提升社会治理水平。例如，基于政务大模型为市民打造统一的便民咨询窗口，可以精准识别企业、群众办事意图等，准确快速地给出最优的办理流程和依据，提高政务服务效率。
- **打造人工智能产业高地，大模型激发区域产业创新活力。**一方面，将加快推进传统产业上下游各个环节的智能化转型。例如，在农业领域，可以结合遥感数据开发出专属的遥感农业大模型，将 AI 技术下沉到水田农地，在种植业监测、耕地用途管理、涉农金融等细分领域助力数字农业技术的升级和推广; AI 基础设施可以赋能工业大模型的研发和应用，实现工业 AI 规模化生产。另一方面，将催生新业态、新模式的不断涌现。例如，MaaS 模式将重塑传统云服务市场格局，将会出现大量行业大模型精调企业，作为通用大模型和企业之间的中间层，助力通用大模型转化为行业大模型。还有海量 AI 原生应用开发企业、云原生安全创新公司等，将打通产业智能化的最后一公里。

- **赋能科学大模型，激发人工智能驱动科学研究（AI for Science）的新范式。**基于大模型对原子运动规律、物质性质等进行预测和模拟，也可对医学图像、天文图像等进行更好的识别和理解，加速科学实验的自动化和智能化，实现自动化合成、自动化表征等。目前，在生物制药、气象预报、地震探测、材料研发等科研领域，大模型技术已带来了巨大的突破。例如，在生物计算领域，Deep-Mind 推出的 AlphaFold2 能够覆盖 98.5%的人类蛋白质组，并对 20 种其他生物蛋白质的结构进行预测[12]。在气象领域，上海人工智能实验室研发的全球中期天气预报大模型“风乌”，首次实现了在高分辨率上对核心大气变量进行超过 10 天的有效预报[13]。

图 14：新一代 AI 基础设施赋能政务、产业和科研创新价值

[12] 《Nature 重磅：AlphaFold 对人类 98.5%蛋白质进》

[13] 《上海 AI 实验室发布“风乌”大模型，全球气象有效预报时间首破 10 天》，上海人工智能实验室，2023 年 4 月

3. 新一代 AI 基础设施赋能企业享受生成式 AI 红利

基于新一代 AI 基础设施，企业可以高效部署生成式 AI 应用，充分利用生成式 AI 推动各项创新优化***（见图 15）***。具体来看：

- **帮助企业实现业务洞察和流程优化，提高决策和生产效率。**大模型可以提供高效的数据分析和预测功能，帮助企业提升决策效率，还可以帮助企业实现流程自动化，减少重复劳动，从而提高生产效率。一方面，传统企业服务软件主动拥抱大模型技术进行软件智能化升级。例如，微软将大模型引入 ERP 产品组合中，覆盖财务、采购和供应链三大模块来优化预算、运营、财务、采购等业务流程。另一方面，企业根据场景需求，基于大模型能力进行业务升级。微博将商汤大语言模型与其业务数据进行融合，让博主拥有 AI 营销助手，进行 AI 自动选品、AI 生成营销内容、AI 生成带货视频，以及 AI 产品咨询与客服运营等，打通“知识、种草、品牌店铺、下单”的内容电商全流程，弥合了微博广大商家和博主从私域获取流量到商业变现的“数字鸿沟”。
- **推动大模型应用融入日常办公中，改变工作模式的同时提升员工效率。**大模型能够帮助员工快速写文本、写 PPT、写代码、分析报表等，成为员工的办公助手。大模型也能提升员工协作和知识共享的智能化水平。微软将 GPT-4 引入 Office 应用程序中，推出 AI 助手 Microsoft 365 Copilot，用户可以通过自然语言交互的方式来进行文档处理、会议记录，快速进行协同办公，极大地提升了办公效率。宝马利用 Amazon Q 生成式 AI 助手帮助宝马数据分析师在数小时内构建仪表板，为宝马客户期望的精确体验提供快速的数据分析支持，大大提高了分析效率，以往需要数天时间完成。
- **帮助企业在个性化、智能化服务等方面进一步提升客户价值。**企业可以基于

大模型能力，为客户提供个性化产品和服务，也可以结合生成式 AI、数字人等，以更自然、智能化的交互形式，成为客户的个性化助手。金融机构在智能投顾、财富管理、客户服务、理财产品营销等方面尝试大模型。工商银行已在国内同业率先实现百亿级大模型在智能助手、知识运营助手、金融市场投研助手等多个场景落地；平安银行则通过大模型提升数字人功能，改善智能客服体验。某消费金融机构披露数据显示，由大模型赋能的智能客服对客户意图理解准确率达到 91%，相较于传统人工智能有 68%的明显提升。

- **为企业在产品/服务、业务模式、技术等方面提供更大的创新空间。**企业创新投入高，大模型技术可以有效降低创新试错成本，同时加大创新发现的可能。生物医药公司正在尝试将大模型技术与医药研发环节相结合。美国合成生物学公司 AbSci 在零样本的情况下使用大模型设计构建不同于现有抗体数据库的抗体，包括了所有三重链条 CDR 的从头测序版本，也是靶标检测最关键的抗体区域，并且通过 10 万多种抗体进行验证，发现命中率高出生物基线水平检测的 5 到 30 倍，提高了候选新药进入临床的效率以及成功率。

图 15：生成式 AI 为企业带来四方面业务红利

四、新一代 AI 基础设施厂商格局与评估

AI 2.0 时代出现了两个新的市场发展空间，一个是基于大模型的生成式 AI 应用，也是爆发最快，但是落地机会依然处于探索期；另外一个则是为大模型和生成式 AI 提供基础设施，包括算力、MaaS 等一系列服务，进而支撑前者的高速发展。“要想富，先修路”，新一代 AI 基础设施就是 AI 2.0 时代的“路”，是支撑大模型、生成式 AI 繁荣发展的基座。在此背景下，AI 基础设施市场在 2023 年进入高速发展阶段，越来越多的厂商进入，根据自身产品技术的布局和优势，提供 AI 基础设施服务体系。

与此同时，为了更快的享受到 AI 2.0 时代的红利或是更快的开展自身技术的落地，用户更加倾向于选择外部供应商来帮助他们搭建技术基座。IDC 调研显示，超过 54%的被访者计划将生成式 AI 相关的投资预算用于外部供应商提供的 AI 基础设施、AI 软件平台以及相应的专家咨询服务。

1. 云计算、AI 原生、硬件系统三类厂商塑造市场格局

AI 基础设施市场依然处于高速竞争的初级阶段，从进入该市场不同厂商的所属类别来看，AI 基础设施厂商格局基本由云计算、AI 原生、硬件系统三类厂商构成***（见图 16）***。三类厂商的业务出发点和产品规划方向分别具有以下特点：

- **云计算厂商依托其成熟的云平台加强 AI 相关产品体系的建设。**此类别厂商以阿里云、百度智能云等传统云计算厂商为代表，在大模型、生成式 AI 爆发之际，全面升级云计算基础设施以适应 AI 2.0 时代对基础设施提出的新要求。升级手段包括启动 GPU 集群的建设和相应工程化系统建设，加强原机

器学习平台面向大模型的分布式训练能力，高带宽网络的连接，以及培育自身对 AI 技术发展的理解等。2023 年 11 月，阿里云在云栖大会上宣布，面向智能时代，阿里云将通过从底层算力到 AI 平台再到模型服务的全栈技术创新，升级云计算体系14。2023 年 12 月，百度表示，为满足大模型落地需求，正在基于“云智一体”战略重构云计算服务[15]。

- **AI 原生厂商基于领先的 AI 产品体系的积累进一步巩固算力资源投入。**“春江水暖鸭先知”，作为长期浸淫 AI 产业的 AI 原生厂商，以商汤科技、科大讯飞为代表，也在不断加强基础设施产品矩阵的规划。由于已经拥有成熟的 AI 原生产品技术以及领先的经验洞察，此类厂商更加重视的是在算力资源层面的布局，以及逐步发展开拓以云服务为核心的交付能力，同时也在不断加强对于 MaaS 平台的建设。2021 年初，商汤科技正式对外公布其 AI 基础设施战略布局，并建设上海临港 AIDC 作为 AI 基础设施的物理载体，力求打通算力、算法和平台，全面构建面向 AI 时代的基础设施。
- **硬件厂商凭借算力产品的优势布局逐步加强 AI 产品技术的布局。**传统硬件厂商，分别从自身原有的服务器或芯片开始加强布局 AI 基础设施，满足用户的训练与推理需求。IDC 数据显示，作为服务器市场中的关键组成部分，AI 服务器市场增速最高，2023 年中国市场规模将达 91 亿美元，同比增长 82.5%，2027 年将达到 134 美元，五年年复合增长率为 21.8%，此领域的代表厂商有浪潮、新华三等。浪潮作为 AI 服务器出货量领先的厂商之一，还在不断加强对 AI 平台能力的补齐，同时在近期推出了“源 2.0”基础大模型，并宣布全面开源，进一步完善自身的 AI 基础设施体系。

[14] 《大模型时代的阿里云，将云计算进行到底》，36 氪，2023 年 11 月

[15] 《AI 原生与大模型将从三个层面重构云计算》，腾讯网，2023 年 12 月

图 16：新一代 AI 基础设施厂商格局一览

	云计算厂商	AI 原生厂商	硬件系统厂商
主要特征	从云计算进入 AI 基础设施市场，重点提升算力层面的智能化、加强 MaaS 服务能力	从 AI 原生进入 AI 基础设施市场，重点加强算力层面的建设和云交付能力的构建	从传统算力设备进入 AI 基础设施市场，重点布局 AI 相关的产品技术与应用服务
优势	具有丰富节点布局的云数据中心，以及经受过业务发展锤炼的云计算平台及相关产品	领先的技术积累使其能够快速进入并领先于生成式 AI 发展趋势，同时拥有深厚的 AI 落地应用的经验，基础设施从一开始就是面向 AI 原生体系	深度研发算力硬件设施，如 AI 服务器、AI 芯片，不断提升计算能力和处理效率，在算力层拥有领先的技术储备
挑战	面对 AI 2.0 时代，需要加强生成式 AI 技术的研发能力和储备，对现有的云计算基础设施进行全面升级和重构	加强算力资源的投资与建设，尤其要开始布局智算中心网络并提升商业化运营能力	需补齐 AI 相关的技术积累和产品矩阵，还要思考如何从硬件供应商向技术服务提供商的身份转变

代表厂商	阿里云、百度智能云、火山引擎、腾讯云等	第四范式、科大讯飞、旷视、商汤科技等	华为、寒武纪、浪潮、新华三等

注：本次研究聚焦于能够提供比较全面 AI 基础设施产品服务的厂商。对于市场上仅提供 AI 基础设施部分产品服务的厂商暂不纳入此次研究范畴。

2. 评估体系：产品能力、战略愿景与市场生态

新一代 AI 基础设施的市场格局处于高速发展中，越来越多不同出身的技术厂商进入该市场提供相关服务，如何全面了解 AI 基础设施厂商发展水平，为企业用户的 AI 基础设施选型提供可靠性指导，成为当下一个关键课题。为此，出于充分评估 AI 基础设施厂商综合能力的目的，我们特意搭建了 AI 基础设施评估体系，包括三个评估维度、十二个指标***（见图 17）***。

- **产品服务能力。**主要评估的是 AI 基础设施的技术积累情况和产品功能丰富度，以及支撑大模型训练和生成式 AI 应用落地的工程化能力，具体包括四大评估指标：产品技术及服务、工程化、市场响应及发展、客户体验。
- **战略规划愿景。**主要评估的是厂商对 AI 基础设施市场发展的理解，以及是否有明确的围绕产品、销售的战略规划和清晰的执行路线图，当然也包括当下已经积累的创新投资和布局，具体包括四大评估指标：市场认知、产品战略、销售战略、创新规划。
- **市场生态表现。**主要评估的是 AI 基础设施的商业化落地表现，以及围绕基础设施的生态布局，并评估厂商在此领域的业务表现和客户和行业的覆盖情况，具体包括四大评估指标：营收及客户、行业覆盖、生态体系、商业模式。

图 17：新一代 AI 基础设施评估体系

评估维度	评估指标	指标描述	权重
产品服务	产品技术	衡量厂商在 AI 基础设施产品服务在关键领域的竞争力和成功程度。简单来说，围绕大模型训练与推理、生成式 AI 应用落地的产品技术，以及交付部署模式，包括不限于数据管理能力、算力规模、MaaS 服务体系、云交付、安全隐私、以及相应的工具链。	高
	工程化建设	这一指标主要衡量厂商在 AI 基础设施的产品服务矩阵是否针对大模型训练与推理层面提供优化措施，包括不限于训练效率、稳定性、可用性、以及相关经验的沉淀等。	高
	市场响应	该指标将会系统的衡量厂商在市场上的发展势头和成功程度，以及对市场发展趋势的捕捉和理解能力，还包括营销洞察及产品定位和对市场动态变化的应对能力。	中
	客户体验	该指标关注客户在采购厂商服务之后的合作体验，包括不限于落地执行性、相关培训的到位，以及领先于客户的技术洞察力。该指标还会衡量现有客户对厂商的满意度情况，以及未来相关性的看法。鉴于支撑大模型、生成式 AI 依然是新兴的技术领域，所以为客户提供专业的相	中

		关专家服务水平也会被纳入进来衡量。	
战略规划	市场认知	该指标，主要衡量的是厂商对市场买家需求的理解情况，并将其转化为产品和服务的能力，以及厂商与客户不断变化的需求的契合程度，以及其客户对厂商所输出的思想理念和新兴技术的接受程度。	中
	产品战略	衡量厂商支持未来将创造商业价值的关键趋势的能力。根据厂商的产品发展路线图，其中有利于创造商业价值的产品功能将会成为衡量重点并得分，并且能够关注厂商是否能够提出有充分验证逻辑的领先于其它友商的发展路径。	高
	销售战略	这一指标关注厂商的销售策略如何驱动市场上的企业用户了解和信任其相关产品服务，包括不限于 AI 基础设施的产品化输出与销售策略制定，相应的销售组织建设与资源投入，以及是否有一套清晰的叙事逻辑来传达其在市场中的价值和优势，还包括在垂直行业的布局和认知程度等。	高
	创新规划	主要衡量厂商在多大程度上为 AI 基础设施能力投资、扩展相关价值体系和提供独特能力，还包括公司在研发投入和人才储备情况，以及人工智能论文、专利数量的积累和创新等。该	高

		指标评判厂商是否正在树立可供其他供应商效仿的创新标准。	
市场生态	商业模式	重点衡量厂商基于 AI 基础设施构建的商业化模型，以及不同商业化模式的目标客户群体定位情况，以及可信服的发展路径。	中
	行业覆盖	聚焦厂商在不同行业所提供的的解决方案能力，以及覆盖行业的广度和深度。	中
	生态体系	衡量厂商的 AI 开源策略，底层硬件层面的国产化适配能力，还包括高校合作创新的布局情况，以及其它合作伙伴和联盟等合作形式的发展情况等。	中
	业绩规模	厂商在 2023 年的营收数字和同比增长幅度，以及在 2023 年的付费客户数量。	中

3. 商汤科技评估结果：新一代 AI 基础设施市场领导者

基于上述评估体系，选取了十二家最具代表性的 AI 基础设施厂商，从产品技术、战略愿景、市场生态三大维度的十二个评估指标进行定性和定量的全面评估。根据指标描述里的衡量细节，为厂商打分，得出十二个指标的得分情况（满分 10 分）。评估结果显示，商汤科技是 AI 基础设施市场的领导者***（见图 18）***。

商汤科技在当前 AI 基础设施评估体系中得分超过厂商平均分，并在市场响应、市场认知、产品战略、工程化建设四个评估指标拿到满分。商汤在产品服务

能力呈现出较强的产品实力和技术积累，早于友商的布局为其带来了领先性差距。在战略规划方面，商汤拥有领先的市场认知，并在战略规划上面设定了明确的执行路径，捕捉到了 AI 2.0 时代所带来的产业变化，并基于此进化了自身的产品战略和销售战略。在市场生态方面，商汤构建了算法开源生态体系，与业界共享创新成果，并打造了多元化的商业模式，并在垂直行业拥有较全面覆盖。

商汤科技推出的 AI 基础设施产品与解决方案为“SenseCore 商汤大装置”，早在 2021 年，商汤就提出“大装置”的概念，成为中国首个提出新一代 AI 基础设施思考并付诸于行动的厂商，2022 年作为其 AI 基础设施重要载体的人工智能计算中心（AIDC）正式投入运营。SenseCore 商汤大装置搭建了完善的 AI 基础设施架构，包括 AI 原生算力基础设施、大模型生产平台、模型即服务，辅以 AI 专家服务和数据服务助力相关技术举措的落地***（见图 19）***。

图 18：新一代 AI 基础设施厂商评估结果 - 商汤科技

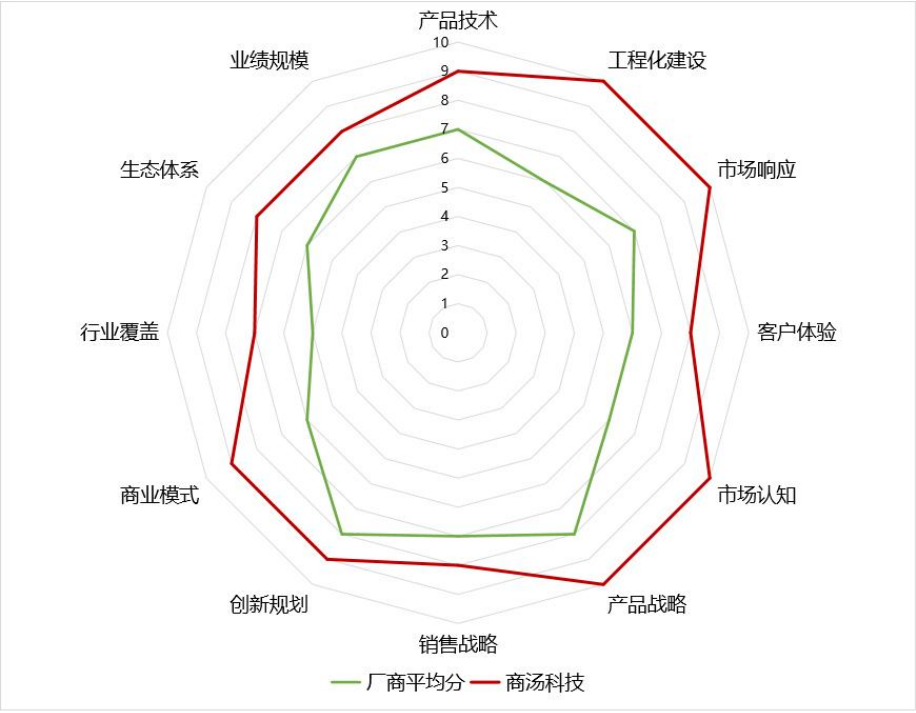


图 19：SenseCore 商汤大装置全景图

4. SenseCore 商汤大装置技术发展优势

SenseCore 商汤大装置致力于建设高效率、低成本、规模化的新一代 AI 基础设施，以人工智能大模型开发、生成、应用为核心，打造一站式、自主研发的 AI 云、AI 平台、AI 服务解决方案，赋能 AI 生产新范式。商汤大装置完善技术储备和创新，打造算力资源、MaaS 平台、数据服务和国产化适配生态，基于全国范围的算力中心和节点，面向大模型、生成式 AI 场景构建产品 ***（见图 20）***：

- **AI 原生云算力基础设施，使能极致大模型开发体验。**商汤大装置在算力层围绕着大模型开发和实践进行了全面打造，包括计算、存储、网络与管理平台 ***（见图 21）***。打造高稳定性的算力池，实现千卡并行训练线性加速比超过 90%，实现 30 天长时间训练不间断，实现分钟级的异常检测和断点续训。优化存储系统，实现 IOPS 缓存系统高于 500 万，存储处理达千亿级别，应对视觉、多模态数据管理需求。高性能无损训练网络，单节点训练网达到 1.6T 的带宽并向 3.2T 发展。提供云管理平台，实现公有、专有、私有、混合等

多云部署模式，并支持 Web 控制台、移动控制台、CLI 命令行、API 调用等多种方式，提供涵盖监控、运维、计费、消费等可视化平台。商汤大装置将会进一步升级算力基础设施，实现 2-3 倍性能提升，提升 50%性价比。

- **结合 SenseNova 商汤日日新大模型体系，MaaS 平台输出大模型时代的 AI 生产工具。**基于商汤自研的日日新大模型体系以及开源开放的第三方大模型，深化打造 MaaS 平台，构建一套完整的大模型生产力工具，包括模型微调、模型推理、内容安全等关键功能，支持不同的应用场景，使得企业客户能够低成本的、快速的接入大模型研发体系。为行业用户提供基模型和开发者工具，帮助行业用户实现针对自身业务场景、领域知识的大模型微调，更好的赋能自身业务场景的智能化转型。AI Studio 则为 AI 开发者提供一系列的 AI 开发工具套件，帮助开发者完成大模型的开发工作。商汤大装置将会探索嵌入模型、搜索增强生成、提示工程等能力，逐步构建 AI 智能体生态，助力各行业重构企业应用***（见图 22）***。
- **提供开放、一站式、低成本的 AI 数据管理与标注平台。**商汤大装置提供面向海量训练数据，开放、易用、高效的 AI 数据管理平台（AIDMP），覆盖数据产生、数据获取、检索分析、可视化、数据使用、合规审核等各个环节，提升数据管理的效率和便利性，严格的访问控制，确保数据安全。提供大规模非结构化数据的检索功能，达到秒级返回，可快速挖掘出新样本。提供高质量业内公开数据集，使用 PythonSDK 工具快速加载数据。面向大模型微调、RLHF、AIGC、自动驾驶等场景，为企业提供一站式的高质量、低成本的数据标注服务，并针对性支持图片、视频、点云等数据格式。
- **国产化芯片适配进入深耕领域，打造 AI 国产化大生态。**商汤自研深度学习

训练框架支持寒武纪、华为、海光、天数、燧原在内的多家国产芯片训练。在算力层面，通过国产化适配，不仅在训练环节通过与国产化芯片进行合作适配，加速落地大模型训练标杆项目，同时，在推理环节，也积极适配进而支持大模型的推理和生成式 AI 应用。在 2023 年，商汤大装置实现多个国产化落地案例，通过与国产化芯片厂商合作，基于领先的算法能力平台和国产硬件资源，打造坚实的新一代 AI 基础设施。2021 年，商汤联合工信部电子标准院及头部算力生态合作企业制定芯片及算力评测标准。

图 20：全面支撑大模型训练与推理、生成式 AI 与应用的基础设施

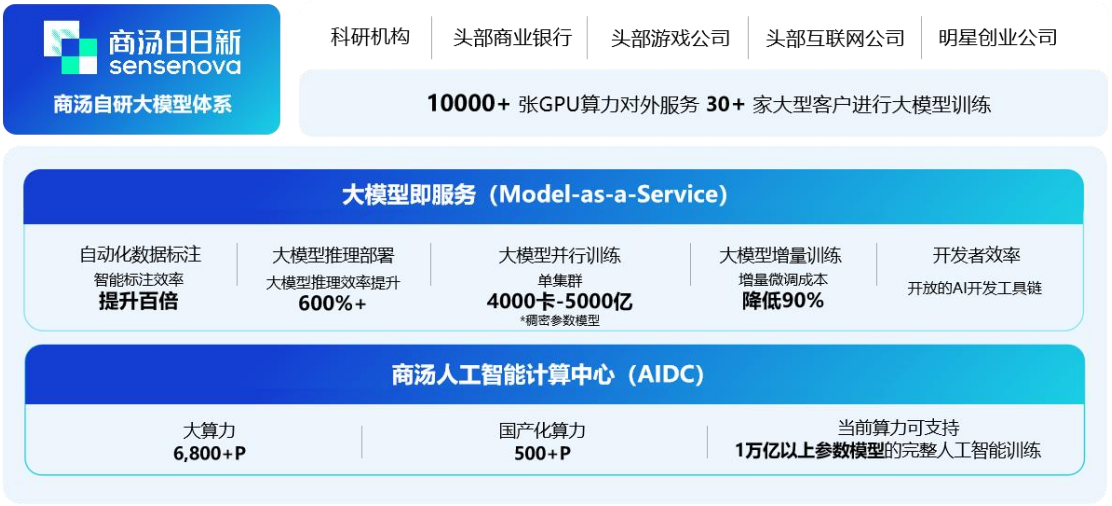


图 21：商汤大装置构建了面向大模型训练与推理的 AI 原生云算力基础设施

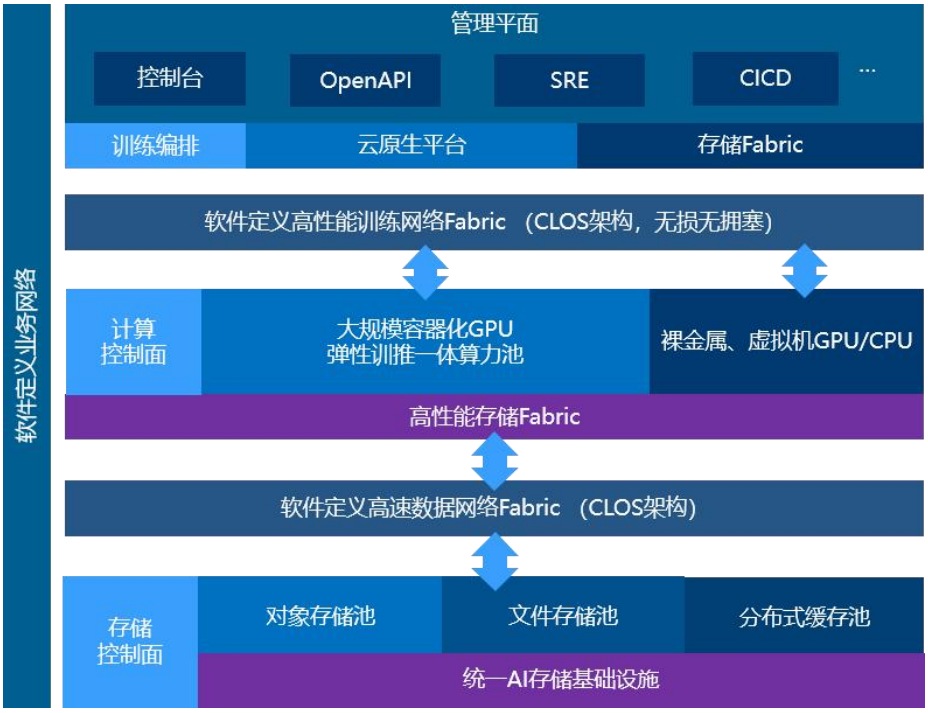
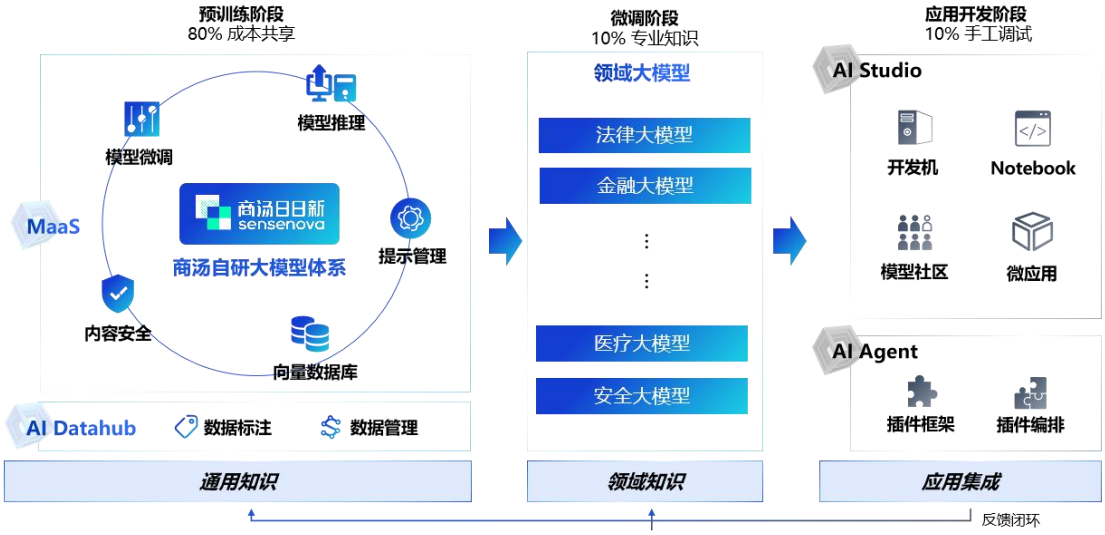


图 22：商汤大装置围绕大模型应用开发搭建了全新的大模型即服务平台



5. SenseCore 商汤大装置业务布局优势

2023 年以来，共有超过 1,000 个参数量数十亿到上千亿的大模型在商汤大装置上完成了训练，实现了领跑行业的技术迭代速度，并且支持了数十款生成式 AI 应用的落地与部署。大装置沉淀了算法和工程上的经验知识，建立了一套工程化的体系来支撑大模型的快速迭代，相关的软件、系统和硬件都以服务大模型快速迭代为目标做研发配合。

- **储备领先的 AI 算力资源，并着眼长远规划。**大模型、生成式 AI 的发展离不开超大规模的算力。商汤建设的上海临港 AIDC 已经是目前全国范围内最大的人工智能算力中心之一。截止 2023 年 Q4，已经上架的 GPU 数量 30,000 张，总体算力规模达到了 6,800P，在支持商汤科技自身大模型研发的同时，支持外部客户训练大模型和应用部署***（见图 23）***。2023 年以来，共有超 1,000 个参数量数十亿至上千亿的大模型在大装置上完成训练，实现了领跑行业的技术迭代进度，并支持了数十款生成式 AI 应用。在服务内部与外部客户的同时，商汤大装置积累了大模型训练和推理的算法、工程层面的优秀经验，以及不断优化完善相应的配套软件系统。商汤还在规划 AIDC 的二期建设，其算力规模也和当前的一期规模类似，将进一步拓展算力储备规模。
- **推动智算中心跨区域多点布局，聚焦区域 AI 产业生态，做实业务运营。**除了自建的上海临港 AIDC，商汤还在积极布局区域算力节点，以需求为导向，立足当地区域的产业生态，共同参与承接地方政府的智算中心建设，并重点加强建成后的运营机制打造，面向当地 AI 相关需求企业进行相关服务的运营，帮助当地政府打造区域智能化产业高地。商汤大装置会结合自身的软件平台能力、专家服务能力，支持区域智能化产业转型/升级，帮助当地企业

闭环落地***（见图 24）***。商汤大装置已经在广州、重庆、深圳和福建成并运营了当地的智算中心，还有更多的区域智算中心在建设中。

- **提供 AI 专家服务，协助客户训练大模型、落地生成式 AI 应用。**基于经验积累，商汤大装置围绕大模型的训练对外提供专家服务，覆盖大模型规划和大模型训练两个阶段。在大模型规划阶段，通过提供大模型开发咨询服务和大模型代训练服务，能够让客户快速理解大模型开发训练的关键节点和潜在痛点。进入大模型训练阶段，商汤大装置提供了全面的，从数据、训练到推理的大模型全生命周期服务，保障大模型开发成果落地。更关键的是，针对不同行业的客户，通过 AI 专家服务，为客户基于大模型端到端设计整个系统，极大降低复杂度，帮助客户提升产品在行业中竞争力***（见图 25）***。

图 23：商汤上海临港 AIDC 算力规模领先

图 24：商汤大装置赋能打造区域 AI 产业智能化高地

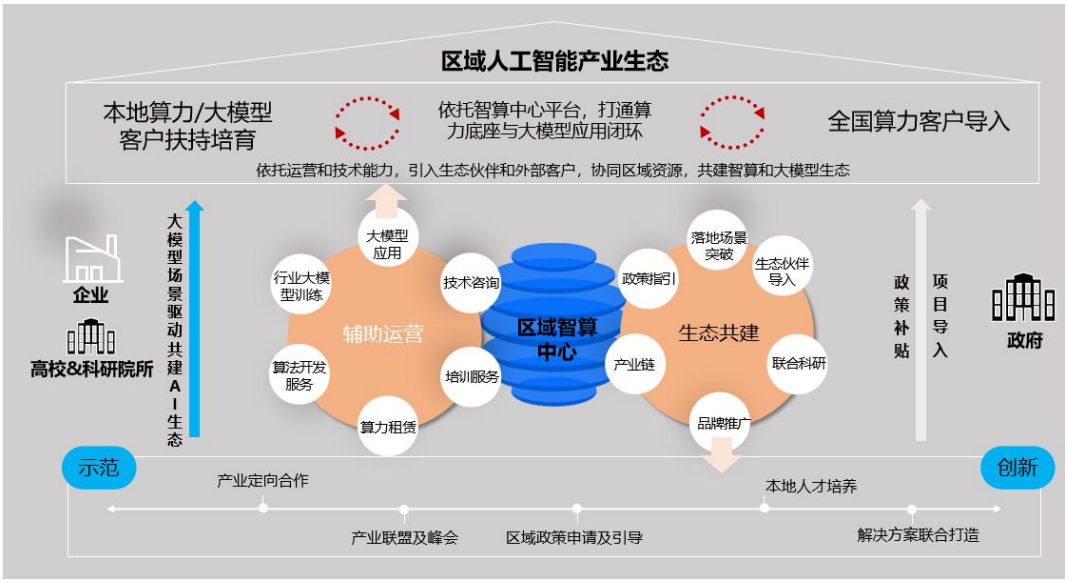
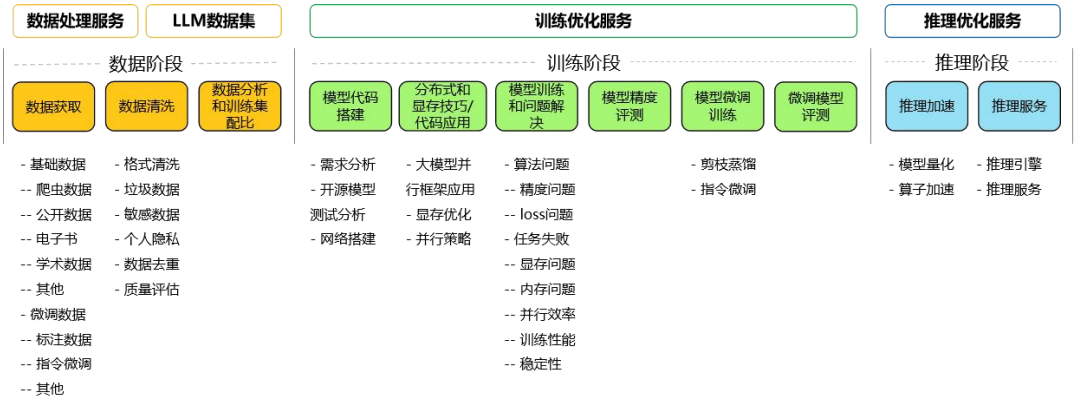


图 25：商汤大装置在大模型训练阶段提供全链条 AI 专家服务



五、新一代 AI 基础设施实践案例

从大模型的训练到生成式 AI 应用的部署，再到区域智算中心的建设与运营，以及 AI 专家服务，新一代 AI 基础设施不断赋能企业、政府更好的利用大模型、生成式 AI 技术来赋能自身的智能化转型，开拓全新的数字化业务。

1. 大模型训练

1.1. 商汤自研日日新大模型体系

商汤科技作为一家专业从事人工智能研究与开发的企业，其在大模型领域的经验已经积累了多年。公司的目标是在这个基础上打造出全球领先的基模型，并通过 AI 基础设施的形式，向各行各业的用户提供高效、低成本的大模型能力和服务。这种举措旨在推动公司通用人工智能（AGI）的发展战略规划。

挑战：

要创建一个性能卓越且具有竞争优势的基模型，需要克服三大难题：首先，基模型的参数量和数据量庞大，因此训练过程通常需要占用大量的计算资源并持续较长时间。为此，算力平台必须能提供稳定、低成本的计算服务。其次，为确保模型输出的性能和价值观满足预期，训练过程需要大量高质量、规模庞大的数据进行喂养。最后，基模型的训练过程复杂，往往需要多轮迭代和优化才能获得最佳效果，因此，需要具备丰富的超大模型开发和训练的经验积累。

方案：

为自主研发的基模型训练迭代提供了数万张 GPU 的算力支撑，以及 1.5 万

亿 token 和 6,000 亿中文高质量数据。2023 年以来，商汤大装置从零开始，已完成了超过 1,000 个参数量从数十亿到上千亿的大模型训练迭代，积累了深厚的超大模型开发和训练的专业经验，能够提供满足超大模型需求的算法开发服务、训练优化服务以及推理优化服务。

价值：

依托商汤大装置推出自研的“日日新”大模型体系，包括商量大语言模型，AIGC 大模型等。新华社研究院《人工智能大模型体验报告 3.0》中，商量大语言模型在定量实测的情商维度上第一，并在定性评估中入选大模型市场未来领袖象限。研究机构弗若斯特沙利文发布《2023 年中国大模型行研能力评测报告》，商汤商量斩获总榜第一，并在报告撰写能力、模型基础能力两个子榜位居第一。今年 9 月份 SuperCLUE 评测中，商汤“商量”位列中文大模型总榜及 AI 智能体子榜双榜榜首，展示出极强的综合竞争力。商汤商量在金融、手机、医疗、汽车、地产、能源、传媒、工业制造等众多垂直行业，已经与超过 500 家客户建立了深度合作，加速行业企业的智能化转型。

1.2. 某大模型厂商

该大模型厂商成立于 2021 年，是一家业界领先的认知智能公司，致力于以自然语言处理技术为基础，为全球企业提供新一代认知智能平台，助力企业数字化转型升级。其主要产品是基于大模型打造的一系列功能引擎（包括搜索、生成、翻译、对话等）和垂直场景应用。

挑战：

在预训练阶段，实现百亿级别参数量的大模型训练需要消耗大量的计算资源。自行构建和管理算力平台的成本非常高昂，对客户负担大。此外，随着业务规模的不断扩大，自建算力系统的可扩展性受到限制，难以满足客户日益增长的计算需求。同时，硬件的并行加速对模型训练的速度和效率产生了一定的制约。因此，为了能够更有效地支持大模型的研发工作，客户对算力平台的性能提出更高要求。

方案：

基于商汤大装置，商汤为客户打造了一个高性能算力平台，总算力规模达到500P，专注于满足客户超大模型训练需求。通过商汤大装置的资源调度机制，确保硬件资源的最大化利用。同时，针对客户业务增长需求，商汤还提供了灵活的硬件扩展方案，以满足不断变化的计算需求。除此之外，商汤还同步提供了全面的技术支持和定期维护服务，以确保平台的持续稳定运行。

价值：

通过使用商汤大装置的高性能 AI 算力池，解决了客户自行建造算力平台在前期投入大、运维成本高等痛点。同时，商汤通过及时的技术支持和稳定的开发环境，进一步提高了客户模型训练的速度和效率，加速了客户的研发周期，让客户能够专注于自身业务需要，训练出更优质的大模型。依托商汤大装置，客户推出了 400 亿参数量的孟子大模型，可处理多语言、多模态数据，同时支持多种理解和生成任务，能快速满足不同领域、不同应用场景的需求。

2. 生成式 AI 应用

2.1. 中公教育 AI 数字人

中公教育创立于 2003 年，是一家职业技能培训学校，业务包括教培图书、面授培训、在线课程，时至今日已发展成为 1000 多个各地分校、超 5000 位专职师资。当基于大模型的 AIGC 技术刚崭露头角时，中公教育在业内率先开启了教育产品革新，加速推动降本增效。

挑战：

对于教育机构，师资是最大的核心资产，也是最大的成本支出，中公教育在线课程部门中公网校采用“双师课堂模式”给万名学员上网课，分配最好的老师在线上讲解核心内容，当地教师进行线下面授与辅导，能够部分缓解全国市场对“名师”的供需矛盾。但带来了新问题，优质师资在线授课也相应挤占了他们的产品教研时间，影响课程开发进度和内容质量，除此之外还有“名师”离职风险。

方案：

依托商汤大装置，结合“日日新”体系下的如影数字人平台与商量语言大模型技术，中公网校与商汤科技经过数月的联合研发，在数字人形象、声音、互动形式，以及课件研发、内容调优等关键环节上持续迭代，上线了首款人工智能课程——“AI 系统班”，并发布虚拟数字讲师“小鹿老师”授课。通过 AI 技术分析优秀师资的教学过程，针对性训练虚拟数字人模拟他们的教学方法和风格，并通过数字化方式还原真实的教学场景，使得虚拟数字人能为学员提供高质量的学习课程。在教学过程中，虚拟数字讲师“小鹿”能依托专业的内容知识库，分析

学员的学习数据，实现与学员的教学互动，为他们提供实时的反馈和建议，帮助他们更好的理解和掌握知识，提升学习效率。

价值：

“AI 系统班”上线一个月，就有超过 7 万名学员报名选购在线课程。首先，通过商汤"如影"app，能实时生成大量"小鹿"老师的视频，满足各种教学需求，且无需拍摄器材和场地，相比传统人工直播，降低了 80%的录课成本。其次，"小鹿"老师节省了真人教师的时间，使他们能专注于课程研发，提升教学质量。与真人讲师不同，"小鹿"的语言表达精准，课程内容丰富，是普通面授课程的 2-3 倍，学习效率也提高 2-3 倍。最后，数字人产品价格的降低，使得可以研发多款"小鹿"老师形象，在抖音等平台运营多个账号，增加曝光度，聚集粉丝。

2.2. 微博智能化内容营销

作为中国最大的社交媒体平台之一，微博拥有着广大的创作用户群体和丰富的社交数据资源。在当前流量红利逐渐减退的大背景下，微博计划将大模型和生成式 AI 技术应用于其内容营销业务，以 AI 原生应用的视角重新审视商业需求。通过充分利用平台的用户和数据优势，微博旨在提升内容电商的效果和质量。

挑战：

在当前市场环境下，尽管微博上的数千万博主和企业品牌商都有强烈的商业需求，但从私人领域获取客户到实现商业价值转化，却存在着一道“数字鸿沟”。一方面，那些在专业知识内容方面表现出色的微博达人，可能不擅长选择产品、撰写营销文案或提供售前售后咨询服务，并且缺乏专业设备或直播间。另一方面，

商家和店铺虽然有供应链优势，但他们可能不熟悉私有域流量运营或者不擅长使用种草内容进行营销。与传统的代运营机构合作的费用过高，这进一步增加了品牌商的运营成本。

方案：

在短短的几个月里，微博增值设计研发中心与商汤科技紧密合作，以商汤商量语言大模型为基础，结合微博多年积累的海量数据资源和精准的行业分类标签，共同训练出一款专门针对微博业务的营销大模型。在这个模型的基础上，每位博主都拥有了一个个性化的 AI 营销助手，该助手采用对话式的需求服务，可以协助博主进行 AI 自动选品、AI 生成营销内容、AI 生成带货视频，以及 AI 产品咨询与客服运营等任务。通过这种方式，帮助微博广大博主和商家成功地打通了“知识、种草、品牌店铺、下单”的内容电商全流程。

价值：

借助 AI 营销助手的强大功能，微博的商家和博主们成功跨越了从私有域获取流量到商业变现的“数字鸿沟”。在客服方面，AI 营销助手不仅能够根据客户特点进行一对一的个性化售前咨询，还能够在同一时间服务百位以上消费者，日平均服务量高达数千人。在内容生成方面，它极大地提升了内容生产效率，使得不懂编程的商家和博主也能够掌握生产力“倍增器”。在营销大模型的支持下，仅仅几个月的时间里，超过 7000 名微博达人就利用 AIGC 生成了 69 万篇文案，覆盖了 9 万多种商品，日平均发布数量超过 5800 多篇。更为关键的是，AIGC 营销文案的采用率达到了惊人的 95%，证明博主们对 AI 营销工具的青睐和认可。

3. AI 专家服务

3.1. 证通股份智能化 SCRM 系统

证通股份成立于 2015 年，是由国内多家证券、基金、期货经营机构，以及互联网企业和金融服务机构以市场化方式共同发起成立的信息技术综合服务企业，以“让金融服务更加安全便捷”为愿景，秉承“让数据说话，用科技赋能”的使命，以建设“最值得信赖的信息技术综合服务平台”为目标，聚焦“数据洞察、智能运营、信创升级”三大核心服务领域，扎根证券业，辐射泛金融行业，提供信息技术综合服务解决方案。

挑战：

在竞争激烈的金融市场中，各大机构都在竭力争取有限的客户资源，这给证券公司的客户营销带来了极大的挑战。证通股份希望能够将大模型技术应用于其基于企业微信的 SCRM 系统中，并开发包括销售话术 AI 辅助、客户对话智能分析和营销文案 AI 生成在内的一系列功能。这些功能旨在帮助金融机构的客户经理更快地锁定目标客户，提高营销活动的转化率，从而增强金融机构的竞争力。

方案：

当前，商汤基于 Embedding 模型和优化 prompt 的策略进行模型微调，与证通股份携手合作，共同研发了适用于金融行业的知识库问答系统大模型。并将该模型能力成功地整合到证通 SCRM 系统中，为用户提供了丰富的功能，包括知识构造、对话处理、以及检索增强对话等。这些功能使得用户能够高效地解析上传的企业文档，进行深度的知识提取，并且通过快速的知识索引，准确地回答

用户的问题。除此之外，商汤还利用重庆智算云为证通股份提供了强大的对外推理服务，进一步增强了其业务能力。

价值：

目前，金融行业知识库问答系统大模型正在进行严格的测试阶段。在包含150 条数据的测试集中，该模型的检索准确度已经超过了 90%，展示出了较高的准确性和可靠性。一旦这个大模型完全融入 SCRM 系统，它将能够显著提升与投资者的互动效率。系统将能够更全面的捕捉投资者信息、交易行为和投资偏好，进行深度的用户画像分析，从而帮助客户经理更准确地把握投资需求，提供更加个性化的投资建议和服务。

4. 智算中心建设与运营

4.1. 重庆智算中心

重庆作为西部大开发的重要战略支点，处在“一带一路”和长江经济带的联结点上，是我国西南地区重要经济中心，同时还肩负着东数西算“承东启西”的战略使命。2022 年，重庆印发的《重庆市软件和信息服务业“满天星”行动计划（2022—2025 年）》（简称“满天星”行动计划）提出大力发展以软件和信息服务业为重点的数字产业，其中重点提到要加快利用场景驱动人工智能产业高质量发展，以及在信创环境下人工智能软硬件产品一体化发展。

挑战：

为了实现"满天星"行动计划的目标，即推动场景驱动的人工智能产业的高质量发展，当地政府面临着三大关键挑战：一是需要有充足的 AI 算力资源储备，

以满足西南区域范围内的 AI 研发和产业 AI 应用落地所需的大规模计算需求。这种海量的算力需求对资源的依赖性极高，因此需要有稳定且充足的资源供应；二是面对海量分散、碎片化的场景应用需求，传统的 AI 生产效率低，为每个场景单独开发算法的成本非常高昂，大部分产业用户难以负担。因此，政府希望能够实施以大模型能力为核心的基础设施建设，提供覆盖建设和运营的一体化算力和 AI 服务；三是考虑到美国对华芯片政策的日益收紧，政府从长远的角度出发，希望能够推动国产化替代，增强基础设施的自主可控性。

方案：

商汤科技（西南）人工智能计算中心（简称 AIDC）是商汤与重庆南岸区携手打造的战略项目。AIDC 的首期建设以国产化算力为核心，其峰值算力高达 100 Petaflops，并计划在中长期内进一步扩大至 1000 Petaflops。AIDC 建成后，由当地的国企与商汤共同组建的合资公司全权负责运营。AIDC 将利用其平台优势，依托商汤的 SenseCore 大装置和“日日新”大模型体系，为西南地区的各行各业、科研机构等产业用户提供涵盖 AI 算力、大模型训练、推理部署和应用开发等方面的软硬件结合的 AI 基础设施云服务。这种全新的 AI 生产方式将更高效地处理海量的开放式任务，以更高效率、更低成本带动当地 AI 产业升级。

价值：

商汤科技（西南）人工智能计算中心已经完成了 30Petaflops 的国产化算力建设，打造了国产化示范点，加大对自主技术、自主产品、自主生态的支持，有利于南岸乃至重庆人工智能产业发展迎来新机遇和新生态。随着后期的运营，它将会连接起算力、算法和平台，显著降低人工智能的生产要素成本，实现高效、

低成本、大规模的 AI 创新和落地，从而打通商业价值的闭环，解决长尾应用的难题。重庆智算中心建成之后，一方面是为大规模模型训练提供最成熟的一站式工厂式生产模式，另一方面是为后期应用落地节约拓展成本，并进一步保障国产平台的替代不影响运营期间的市场化收入。此外，它还将助力西南地区的产业升级，为当地政务、民生、经济等领域的数字化应用，以及智能汽车、生物医药、高端装备制造、新材料等重点产业提供强大的人工智能算力资源和应用支持。

4.2. 福建省大数据算力平台

作为数字中国建设的策源地和实践地，福建省数字经济发展指数全国排名第 9 位。2022 年全省数字经济增加值突破 2.6 万亿元，占全省经济总量近 50%，同比增长 13%。为深化“数字福建”建设，高起点打造国家数字经济创新发展试验区，《福建省新型基础设施建设三年行动计划（2023—2025 年）》提出要响应国家大力推进算力基础设施高质量发展号召，在“提升智算中心”部分，明确设定目标，“建设省算力资源一体化服务平台，构建低成本公共算力服务体系”。

挑战：

为了充分实施“三年行动计划”中的算力平台项目建设，地方政府提出了两个关键诉求：一是省内算力发展主要由需求驱动。本地算力需求旺盛，算力消费应用水平和行业应用水平较高，相对而言，现有智能算力规模较小，且多头建设导致资源分散，难以满足大型计算需求。因此，需要根据未来需求的变化重新规划，集中建设算力平台。二是，省内产业众多，除平台建设外，需要深入挖掘福建的重点行业、代表性高校院所以及广大企业的潜在需求，打通创新链、产业链、价值链和数据链，带动产业生态构建，服务数字福建建设大局。

方案：

由福建省大数据集团的成员单位——福建省星汉智能科技有限公司（简称“星汉智能”）投资的大数据算力平台已经在福州市滨海新城正式启动试运行。该平台的第一阶段已经建成了 125PFLOPS 的算力规模，并且配备了能够每秒传输 400GB 的高速算力网络以及 5500TB 的高性能存储。该平台以国际领先的高性能算力硬件为基础，通过业界领先的 AI 算力平台软件，打造高性能 AI 算力池、云容器实例、AI 云开发机等算力服务产品，为大模型训练推理提供高性能、高质量的算力服务。

价值：

这一平台作为省大数据集团的重要战略投资项目，积极响应了国家关于推进算力基础设施高质量发展的号召，促进省内人工智能产业发展。目前，已经吸引了多家数字生态领域的领军企业入驻。自平台投入使用初期，算力资源的使用率就已经超过了 80%。该平台立足于闽北地市，服务范围覆盖了整个省份，有助于推动生成式 AI 技术在医疗、金融、制造等各个产业中的应用落地，打造出一批新的算力业务、模式和业态，加速产业的数字化和智能化转型升级。此外，平台也会吸引专门的人才聚集，为省内的数字经济高质量发展注入强大的动力。

六、建议

生成式 AI 将引领人工智能领域的重大突破，为政府、各行业以及生成式 AI 初创企业带来前所未有的创新机遇。面对这一变革，企业都应迅速采取行动，以便在 AI2.0 浪潮中抓住发展机遇：

- **不要等待，要有所行动，主动评估和选择合适的新一代 AI 基础设施供应商。**大模型、生成式 AI 正变得越来越复杂，这意味着更多的数据要处理、更大参数的大模型、更频繁的再训练和微调。不符合要求的 AI 基础设施将会无形中为企业的生成式 AI 之旅带来额外成本，高效能的新一代 AI 基础设施可以大大降低大模型的训练和推理时间，让企业的 AI 团队更加有效率。请参考本白皮书提出的“新一代 AI 基础设施厂商评估体系”，重新评估现有的供应商，或者开辟新的供应商，选择合适的供应商伙伴。
- **大模型、生成式 AI 初创企业可选择外部供应商，而非一味的谋求自建。**对于大模型、生成式 AI 创业公司而言，加快大模型训练流程和加速应用推向市场是更高优先级的任务。已经拥有算力储备，并积累了产业经验的新型 AI 基础设施供应商，能够更好的帮助这些创业公司将精力聚焦在自身业务发展，而非底层资源的投资和搭建，规避了自建 AI 基础设施带来的巨额成本投入，加快了大模型训练和部署进程。同时，AI 基础设施供应商往往拥有客户资源和应用场景，也为初创公司的商业化带来便利。OpenAI 利用微软为其专属打造的 AI 基础设施，用以自身大模型的训练和推理服务。
- **区域政府落实“建运一体”的智算中心打造 AI 创新高地。**各地政府近期相继出台了 AGI 相关政策措施来加快人工智能创新高地建设。在这些政策中均

重点提到了算力端发展，加大算力基础设施的投资力度，同时强调了人工智能的高质量发展，拓展生成式 AI 创新应用场景的深度与广度。为了加快区域 AI 高地建设，一方面，由于不同区域差异，各地政府需要明确本地产业优势、算力供给规模和支持措施等，选择产业集聚地区进行人工智能算力集群布局，为创新主体提供普惠算力；另一方面，依托智算中心平台，与具备大模型技术和运营能力的厂商合作，打通算力底座与生成式 AI 应用闭环，共建本地大模型生态。

- **评估自身企业的生成式 AI 就绪情况。**从顶层设计、中间执行、人才和团队建设等，全面自审生成式 AI 应用的能力预备水平，无论这一能力获得是企业内部还是通过战略合作伙伴。企业的人工智能基础现状和获得的技能是成功的关键因素。具体来说，例如，在顶层设计上，企业是否建立了评估和跟踪开源生成式 AI 代码、数据和培训模型使用的指导方法；是否开展了全新的针对生成式 AI 的使用培训，以及是否设立了相关的安全、隐私及伦理的专属团队等。

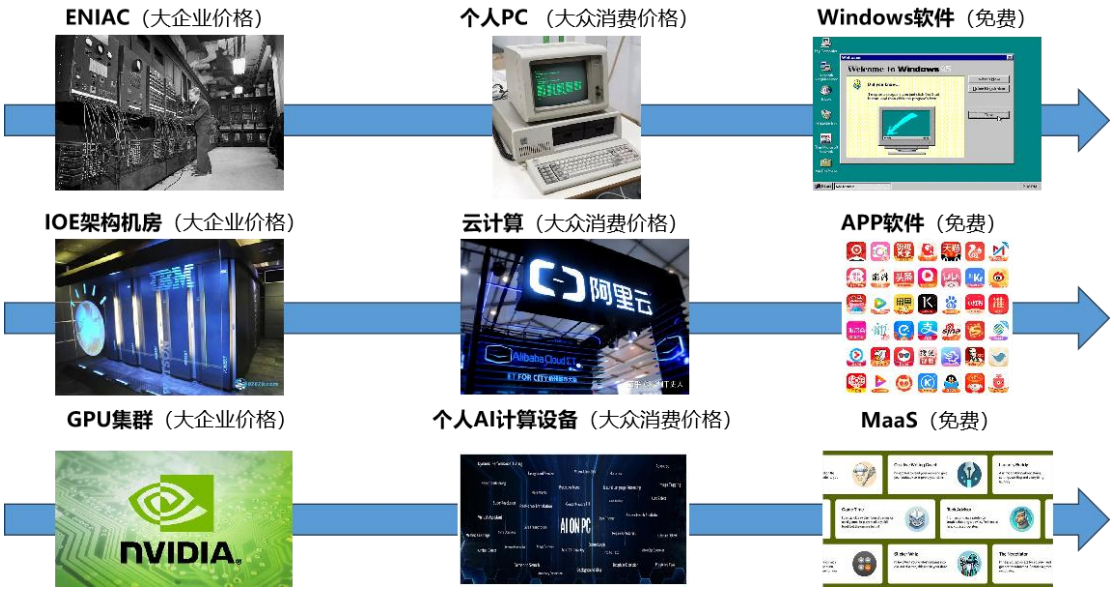
结语：新一代人工智能基础设施的“经济规律”

第一，传统经济学理论中，新型社会基础设施的重大意义在于促进技术进步、提高生产率、加速内生经济增长，具有良好的“正外部性”与“网络效应”。例如在 1955 年出版的《经济增长理论》一书中，经济学家阿瑟·刘易斯认为“基础设施投资对于经济增长至关重要，因为它们为生产要素提供了有效的运作环境。”在 2023 年 2 月，国务院印发的《数字中国建设整体布局规划》中明确：“数字中国建设按照‘2522’的整体框架进行布局，即夯实数字基础设施和数据资源体系‘两大基础’，推进数字技术与经济、政治、文化、社会、生态文明建设‘五位一体’深度融合，强化数字技术创新体系和数字安全屏障“两大能力”，优化数字化发展国内国际‘两个环境’。”**数字中国建设，与“数字基础设施互联互通”、“数据资源规模和质量”紧密相关。**

第二，支撑新一代生产力的智能计算基础设施，在长周期建设过程中，具有“资本密度”、“算力密度”、“数据密度”持续增加的特征，目前投资总额尚具有极大提升空间，将不断增强我国经济的比较优势。支撑生成式智能应用的智能计算中心资产规模日趋庞大，新型 AIDC(AI Data Center)智能基础设施通过高效整合能源、算力、训练数据、大模型等新生产要素资源，为智能经济为核心的“数字经济 3.0”构筑护城河，支撑实体经济生产结构的数字化转型调整。正如卡尔·马克思在《资本论》中所说“基础设施，即固定资本，是生产过程所必须的持久性条件。” 1995 年，美国“信息高速公路计划”总投资额 4000 亿美元，占年度 GDP 的 5.8%，开启了全球互联网“数字经济 2.0”浪潮，近年美国数字经济年均增速超过 6%，占 GDP 比重已超过 10%，年度产值高达 2 万亿美

元，以及 5%的就业岗位。发改委数据显示，2022 年开始，中国“东数西算”工程项目总投资额超过 4000 亿元[16]，在年度 GDP 中占比不到 1%，仍有巨大投入空间。IDC 预测 2022-2027 年，中国智能算力规模复合增长率达 33.9%，是通用算力的 2 倍增速，用于行业应用的推理算力从 2024 年出现“拐点”，首次超越训练算力规模，2024-2027 年推理算力占比从 67.7%提升至 72.6%[17]。2025 年中国数字经济规模将首次超过实体经济（GDP 占比超过 50%）[18]。

图 26 计算基础设施的“三代浪潮”：从昂贵到免费



第三，在中国“智能计算基建化，传统基建智能化”的过程中，科技创新是推动经济增长、社会基础设施高质量发展的源动力，而智能计算基础设施具有边际成本持续下降、边际效益持续增长的特征。创新理论家约瑟夫·熊彼特在《经济发展理论》一书中提出了“创新是推动经济增长和社会基础设施演进的关键动

[16] 《“东数西算”工程总投资额超过 4000 亿元，算力集聚效应初步显现》，金融界，2022 年 9 月
[17] 《中国人工智能计算力发展评估报告，2023-2024》，IDC，2023 年 12 月
[18] 《2023 年中国数字经济发展趋势及市场预测报告》，前瞻产业研究院

力。”纵观科技发展史，过去 100 年的电力价格持续下降带来了电力经济大爆发，过去 20 年移动通讯价格、云计算价格持续下降带来了移动互联网经济大爆发，同样具有信息技术属性的 AI 计算基础设施，通过模型软件革新、AI 芯片计算架构革新，将带来“智能基建革新”、“智能经济大爆发”，社会供给侧的成本下降引发社会需求指数级扩张，最终 AI 带来的商业创新价值将远超 AI 算力成本，“免费 AI 服务”成为“产业终局”，普惠 AI 造福地球上的每一个人。

“在现在这个时间点上，能不能用 AI 来命名时代，取决于它能不能把我们这个时代生产要素的成本规模化下降，从而才能让 AI 走进千家万户。”

— 徐立，商汤科技董事长兼 CEO

联系作者

刘亮：liuliang2@sensetime.com

田丰：tianfeng@sensetime.com

杨燕：yangyan1@sensetime.com