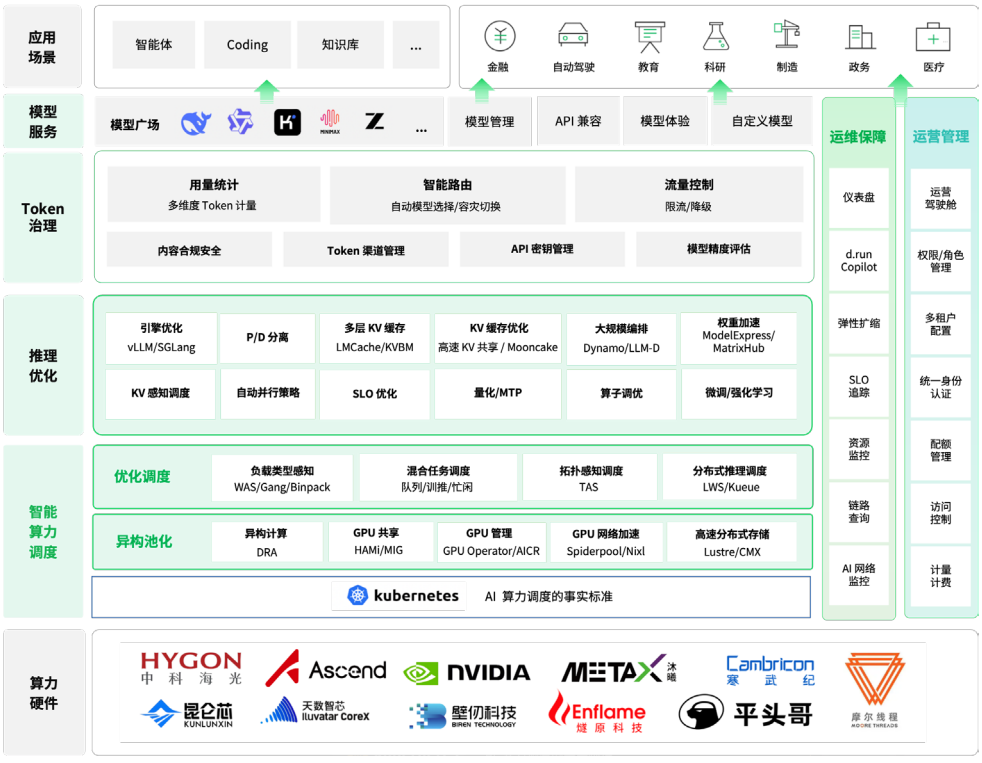
d.run AI 操作系统

打通 **算力 - 模型 - Token** 的企业级服务平台

依托全球前三的 Kubernetes 调度技术与 vLLM 等主流开源推理引擎核心贡献积淀，统一纳管多元异构算力，实现颗粒化调度、全栈推理优化与全链路 Token 治理，算力利用率超 80%，将算力高效转化为可管可控的 Token 化 AI 生产力；平台汇聚全球主流大模型生态，配备可视化运营驾驶舱与 d.run Copilot 智能助手，面向企业全部门输出稳定高效的 AI 服务，全方位支撑业务智能化长期升级。



01

智能算力调度
资源利用率超 80%

- 全球前三、亚洲第一的 Kubernetes 技术领导者，定义全球算力智能调度技术标准。
- GPU 显存与计算单元颗粒化调度，支持单卡多模型并行和多卡分布式推理，及空闲资源池化复用，算力利用率超 80%。
- 国内外主流 AI 硬件统一纳管，精准适配企业异构计算需求，助力算力高效转化为企业 AI 生产力。

02

推理深度优化
Token 吞吐提升超 24 倍

- 依托在 vLLM（全球 Top 5）、SGLang 及 NVIDIA Dynamo 的核心贡献，实现对主流推理引擎的深度掌控与全栈协同优化。
- 基于 PD 分离、精细化 KV Cache 存储协同及智能调优方案，在同等硬件条件下大幅提升 Token 产出及承载量，攻克长上下文高并发难题。

03

精细化 Token 治理
让每一份 Token 都转化为生产力

- 统一纳管公有云、自建集群、第三方等多源 Token，兼具核算、调度、容灾等能力，保障供给成本最优、稳定合规。
- 支持按部门、业务、项目多维度拆解用量与配额管理，全链路可追溯，让每一份 Token 消耗的 ROI 可见、可衡量、可优化。
- 通过智能路由支持动态负载均衡与无感故障转移，结合模型评估反馈，自动选择最优模型组合，构建高韧性、低延迟、可进化的 AI 推理环境。

04

多样化大模型服务
同步全球前沿大模型生态

- 汇聚全球顶尖大模型生态，深度集成 DeepSeek、GLM 等主流开源模型及企业自研 / 第三方 API，确保技术架构始终与前沿保持同步。
- 通过标准化的 API 封装与自助式敏捷部署，大幅降低模型接入门槛，助力开发者高效构建企业级 Agent 与智能应用。

05

高效运维与运营
让合规无感，让决策智能

- Token 内容全程可控，严防企业机密、敏感信息随调用流出，操作人、用量、内容全链路留痕，满足安全合规要求。
- 多维度可视化运营驾驶舱，实时监测 Token 消耗、服务 SLA 等核心指标，自动识别异常并告警，前置化解风险。
- d.run Copilot 智能助手一键生成预测、规划与调优方案，实现智能化自愈与自动化进阶。

AI OS

深入业务场景，赋能企业智能化升级

深入金融风控、智能制造、全球零售等核心业务场景，通过 **d.run AI OS** 支持企业 AI 基础设施的建设、管理和运营，最大化提升算力利用率，并能无缝集成主流模型，提供即插即用的模型服务，让 AI 真正深入企业的业务肌理，实现从“数字赋能”到“智能驱动”的全面跃升，支撑头部企业构筑市场竞争的领先壁垒。

01

中国 500 强的金融资产管理集团

提供从 GPU 选型规划、运维体系搭建到大模型适配的全链路解决方案，助力企业高效构建一体化 AI 算力基础设施，孵化出规章制度、常见问题解答、案例推荐以及行研分析等一系列大模型应用，成功实现数据驱动的业务赋能与智能化运营。

算力供应稳定可靠

多场景 AI 智能

02

某半导体液晶屏制造巨头

基于 AI 一体机为该企业构建高性能、高利用率的 GPU 算力集群，并通过智能调度平台和制造业场景化模型应用，显著提升企业资源使用效率和业务智能化水平。

30% ▲
软件研发效率提升

10,000+
AI 智能体

500 名
“硅基数字员工”

30%
办公流程自动化

03

欧姆龙

全球知名的自动化控制及电子设备制造领军企业

基于云原生底座，帮助企业打造面向“工业制造”与“企业行政”双场景的 AI 知识库，通过智能问答辅助客服高效处理业务，并为员工提供企业制度与日常管理事项的自助查询服务，大幅降低跨部门、跨岗位的重复沟通成本，实现企业全方位降本增效。

65%
知识整理工时节约

60% ▲
客服处理效率提升

52%
行政咨询人力节约

45%
报销前置沟通时长缩短

AI OS

定义 AI 基础设施新标准

基于 **d.run AI OS**，构建可复制、可落地的 AI 基础设施样板工程，通过异构算力统一管理与模型全生命周期运营，深度赋能区域算力基座建设、产业生态服务以及国产化软硬协同，加速行业智能转型。

01

HKGAI

区域算力基座标杆

联合港科大打造支撑首个港产大模型 HKGAI V1 落地的高效能算力集群，并全面赋能面向全港市民的 AI 服务平台 HKGAI Studio 的底层架构，显著提升资源利用率的同时，确保了从算力管控到模型运行的全栈技术自主可控。

90%

GPU 利用率

60%

模型微调效率提升

02

模速空间

产业生态服务标杆

为全国首个大模型创新生态社区打造算力资源服务平台，通过高效的资源池化、弹性调度与成本优化，系统性降低了创新门槛，滋养着社区内大模型及应用团队从研发到落地的全过程。

算力 - 模型 - 应用

一站式服务

03

Shanghai Cube

国产化软硬协同标杆

联合国产算力厂商共同研发国内首款开放架构的智算超节点 Shanghai Cube，以国产化替代为核心，软硬深度协同，为系统性推进国产算力芯片的规模化应用夯实了基础。

80%

性能达到进口高端设备

全栈国产化

“芯片-系统-平台”