

2026

AI原生基础设施白皮书

面向Token产效提升与价值创造的代理式AI
基础设施新范式

出品机构：甲子光年智库

研究主笔分析师：黄琪丹

发布时间：2026年10月

目录

前言

PREFACE

01

摘要

ABSTRACT

03

新趋势

PART 01 NEW TRENDS

1.1 趋势洞察：AI主导权争夺与AI产业阶段切换同步开启

05

1.2 产业判断：词元本位时代开启

07

1.3 范式落点：词元的产效、价值转化与运营治理呼唤新一代基础设施

14

新范式与新定义

PART 02 NEW PARADIGM & DEFINITION

2.1 AI原生基础设施 = Token Factory + Agent OS + Token Ops

15

2.2 典型架构解析：蓝耘科技·AI原生基础设施架构

18

新分工与新模式

PART 03 NEW INDUSTRY DIVISION & MODELS

3.1 产业新分工：Token Factory专业化分工与价值再分配

29



目录



3.2 商业新模式：建立效果计价与成本计价的桥梁	31
3.3 组织新模式：以任务驱动为核心的AI原生组织	33
3.4 治理新模式：面向Agent行为的约束机制	34

新实践

PART 04 NEW PRACTICES

4.1 Token需求分化为生产、调用与价值转化三类	35
4.2 Token生产需求典型案例	36
4.3 Token调用需求典型案例	38
4.4 Token价值转化需求典型案例	40

结语

CONCLUSION

42

定义者系列
THE DEFINER SERIES

定义者

DEFINER

定义未被定义之物

AI原生基础设施

AI NATIVE INFRASTRUCTURE



蓝耘科技&甲子光年智库

2026

前言

P R E F A C E

AI原生基础设施

面向Token产效提升与价值创造的代理式AI基础设施新范式

从电力到算力，人类每一次基础设施的跃迁，都重塑着文明的运行方式。百余年前，电力普及推动工厂由依靠人力转向机器生产，奠定了现代工业的基础；数十年前，互联网让信息跨越时空自由流动，重构了商业与社会的连接方式。如今，智能能力开始进入生产经营与社会生活，它正成为继电力、网络之后支撑新一轮产业变革的关键基础能力。

沿着这一历史进程，全球竞争的焦点也在向AI基础设施深处转移。科技竞争正由地缘资源争夺转向AI主导权竞争，芯片、算力、模型和数据逐渐成为影响国家科技实力与产业竞争力的战略要素。推理需求快速增长，推动算力产业由设备供给转向产能与效率竞争；国产芯片、异构算力和模型生态加速发展，也对软硬件适配、资源调度及运营治理提出更高要求。AI基础设施由此成为承接技术自主、产业升级与智能能力规模化供给的重要支撑。

宏观竞争的深化，与AI产业自身的发展阶段相互叠加。生成式AI正由以模型能力为中心的发

展阶段，进入以Agent持续执行任务为特征的代理式AI阶段。企业需要的不再只是一次模型回答，而是对业务任务进行理解、规划和拆解，并通过模型推理、工具调用与多方协同完成结果交付。在此过程中，Token的重要性由模型调用计量单位进一步延伸至智能生产与任务运行效率的核心资源。如何稳定、高效、规模化地生产高质量Token，并将其转化为可交付的业务价值，正在成为AI产业落地的新命题。

面对战略竞争与产业演进提出的双重要求，传统算力供给体系的能力边界需要扩展。以芯片、服务器和智算中心为核心的资源底座仍是AI发展的基础，但模型、数据、知识与工具等多元资源的组织与协同需求日益凸显，AI基础设施的核心功能也由提供计算资源，升级为组织Token的生产、运行与治理。

AI基础设施由此需要突破单一资源供给模式，形成覆盖Token生产规模与效率、价值转化与全链路运营治理的系统能力，推动智能能力由底层资源投入走向可持续的业务价值创造。

这意味着，AI基础设施的演进并非简单增加算力、模型或开发工具，而是围绕智能能力的形成与交付，对底层资源、运行机制和治理体系进行系统重组：向下组织能源、算力、数据与模型，在扩大Token产能的同时优化生产效率；向中间优化推理、调度、知识与工具调用，提升Token在Agent任务链路中的调用效率、输出质量与交付稳定性；向上支撑Agent进入业务流程，实现Token的价值转化；同时以全链路运营治理持续优化Token的成本、质量、用量与业务效果。

基于上述判断，本白皮书提出，AI原生基础设施是面向Token产效提升与业务价值创造的代理式AI新型基础设施体系。白皮书将从产业趋势出发，对其架构边界与核心构成进行定义，并进一步讨论由此带来的产业分工演变、商业模式、组织架构与治理机制升级。

在此基础上，白皮书以蓝耘科技为实践样本，分析其围绕Token Factory、AI Infra OS、Token Ops与Agent应用形成的布局，呈现国内企业在智算底座、智能调度、运营治理与业务场景落地等方面的探索。这些实践并非AI原生基础设施的唯一形态，而是观察产业由资源供给走向Token生产与运营的一扇窗口。

基础设施的价值，最终体现为推动一项新技术从局部突破走向规模化运行与广泛应用。电力与互联网的发展历史反复表明，率先构建支撑新技术稳定运行的基础设施，往往意味着在新一轮产业周期中获得更大的主动权。面向以

Token为核心运行资源、以Agent为主要应用载体的AI新阶段，如何将分散的资源模块与技术能力组织为稳定、高效且可持续创造价值的Token服务，正是本白皮书希望与产业界共同探讨的核心命题。

本白皮书是蓝耘科技与甲子光年基于产业观察和企业实践形成的阶段性研究。鉴于相关技术、产品与产业分工仍在快速演进，难免存在疏漏与不足，敬请业界同仁批评指正，并期待与各方共同完善对AI原生基础设施边界、路径与价值的认识。

甲子光年智库

摘要

ABSTRACT

Part 01:

聚焦产业新趋势。全球科技竞争正由地缘资源争夺进一步转向AI主导权竞争，推动Token成为新的战略生产要素。

在产业层面，Token逐渐成为代理式AI体系中可标准化计量的生产与运行单位；Agent则打破数据、模型与应用串行衔接的传统方式，使Token贯穿任务理解、推理、工具调用与结果交付。生产对象和生产方式的变化，进一步推动算力产业由设备供给转向产能与效率竞争，并围绕Token的生产、调度、流通与消费形成新的运行链路。

在此基础上，引入“ $Y=O \cdot F(T)$ ”的类比分析框架，解释了智能能力如何由模型输出转化为业务价值。随着Token进入规模化生产与消费阶段，质量、成本、权限、安全与业务效果也需要纳入统一治理。

上述变化共同推动AI基础设施由算力资源供给，升级为支撑Token生产、价值转化与全链路治理的系统能力。

Part 02:

AI原生基础设施的定义与架构解析。AI原生基础设施是“面向Token产效提升与业务价值创造的代理式AI新型基础设施体系”，由Token Factory、Agent OS与Token Ops三部分构成。

Token Factory组织能源、算力、数据和模型资源，扩大Token产能并优化生产效率；Agent OS围绕Agent任务执行，组织模型、路由、知识、上下文、记忆与工具等能力，推动Token转化为可交付的智能能力和业务价值；Token Ops贯穿生产与运行全链路，通过数据汇聚、监测分析、安全治理和持续反馈，使各环节状态可观测、问题可定位、能力可优化。

在这一通用框架下，蓝耘科技将相关基础设施能力具体化为Token Factory、AI Infra OS、Token Ops，并围绕行业约束展开解题。Token Factory通过国产芯片适配、单机扩卡、集群扩展和智算中心建设扩大有效产能；AI Infra OS通过算力供应、推理调度、模型服务和知识工程提升运行效率与业务适配；Token Ops汇聚全链路数据，统筹用量、成本、质量、安全与智能运维。

Part 03:

聚焦AI原生基础设施催生的产业新格局与新模式。从产业新分工看，云计算厂商需要以弹性产能应对需求峰谷，模型厂商需要稳定、高性价比的算力支持，国产芯片厂商则需要加快软硬件适配与规模化落地。三方需求汇聚，推动独立智算服务商兴起，并承担异构算力生产组织者、算力资产运营者与跨生态连接者等角色。

从商业新模式看，随着Token成本结构逐步可拆解，任务质量与业务效果逐步可计量，资源投入与价值产出之间得以建立映射。AI服务由单一资源计费，走向用量收费、能力收费与效果增值收费相结合的组合模式，使供给方能够依据交付能力和业务结果定价，也使客户有望更清晰地衡量投入产出。

从组织新模式看，Agent进入企业业务流程，将推动组织由人力层级协作走向人机协同。Agent逐渐成为承接具体任务的数字化角色，人与Agent、Agent与Agent之间形成新的分工与协作关系。

从治理新模式看，Agent自主执行任务也带来新的风险边界。治理对象由模型输出扩展至任务理解、工具调用、过程执行与结果交付，需要建立覆盖身份权限、过程审计、质量评估和责任追溯的机制，使智能能力在扩大应用范围的同时保持安全、可靠与可控。

Part 04:

聚焦Token需求分化下的企业实践。随着AI能力进入不同业务环节，客户需求沿Token价值链分化为生产、调用与价值转化三类，蓝耘科技分别以算力服务、模型服务和Agent应用承接相关需求。

面向大模型企业所代表的Token生产需求，核心问题是获得高性能、稳定且可弹性扩展的算力。蓝耘科技通过智算中心、异构硬件适配、集群扩展与算力调度，提升算力向有效Token转化的能力，降低客户自建和运维成本。

Token调用需求以模型能力消费企业为代表。此类企业不直接参与Token生产，而是通过外部模型服务获取所需能力，面临模型接入门槛高、版本迭代难跟进、服务稳定性难保障以及行业适配复杂等问题。蓝耘科技通过MaaS平台提供统一API、主流SOTA模型、弹性资源调度与持续技术支持，降低模型调用和运维成本。

Token价值转化需求的关注点已由获得模型能力进一步转向业务结果，面临Agent建设、系统集成、专业知识适配与持续运营等难题。蓝耘科技分别构建“数字人力官”和营销获客Agent，将模型能力嵌入人力资源与民宿营销流程，推动Token消耗转化为服务提效、客户增长和经营收益。

Part 01

新趋势

词元时代正在到来

注：为贴近词元经济表述，本章节统一采用词元描述Token，采用智能体描述Agent

1.1 趋势洞察：AI主导权争夺与AI产业阶段切换同步开启

趋势1：大国博弈的焦点从地缘军事转向科技主导权，而AI正站在这场争夺的中心

大国博弈的底层逻辑正在发生结构性转变。传统地缘竞争以领土、军备与能源通道为核心，投入高、对抗性强，难以转化为发展动能。2018年芯片出口管制标志着科技成为大国博弈的首要切入点，AI正处于这一切入点的中心。

AI主导权的争夺沿三个层级展开：叙事与规则层，即概念、标准与治理框架的制定权；算法能力层，取决于大模型与智能体的技术水平；智算底座层，涵盖高端芯片、智算中心与能源

保障。越往底层替代难度越高——规则可以谈判，模型可以追赶，而智算底座建设周期长、资本密度高，构成最难逾越的壁垒。三层各有分工：智算底座决定产出规模，模型能力决定产出质量，叙事与规则决定产出标准。

这一迁移已从趋势演变为现实：2018年芯片管制开启科技博弈；2022年ChatGPT发布后，AI上升为各国国家战略；2026年中国词元日消耗量达140万亿，词元消耗量首次成为国力对比指标；同年8月，CSIS¹提出Tokenpolitik²，将生成与分发词元的能力定义为国家竞争力。

图1：大国博弈焦点转移



事件佐证，2018起进入焦点转折期



大国博弈的计量单位，正从航母吨位转变为Token产量

智算底座决定“产出规模”，模型能力决定“产出质量”，叙事与规则决定“产出标准”

趋势2：AI技术价值重心迈向任务成果交付，词元产效成为产业新关注点

高质量数据与模型始终贯穿AI产业发展的主线，但伴随AI技术形态演进，各阶段的技术能力重点有所迁移。

感知式AI阶段，产业以标注数据训练专用模型，使机器获得视觉识别、语音识别等感知能力，并通过NLP实现基础语言理解，但该阶段的理解深度有限。

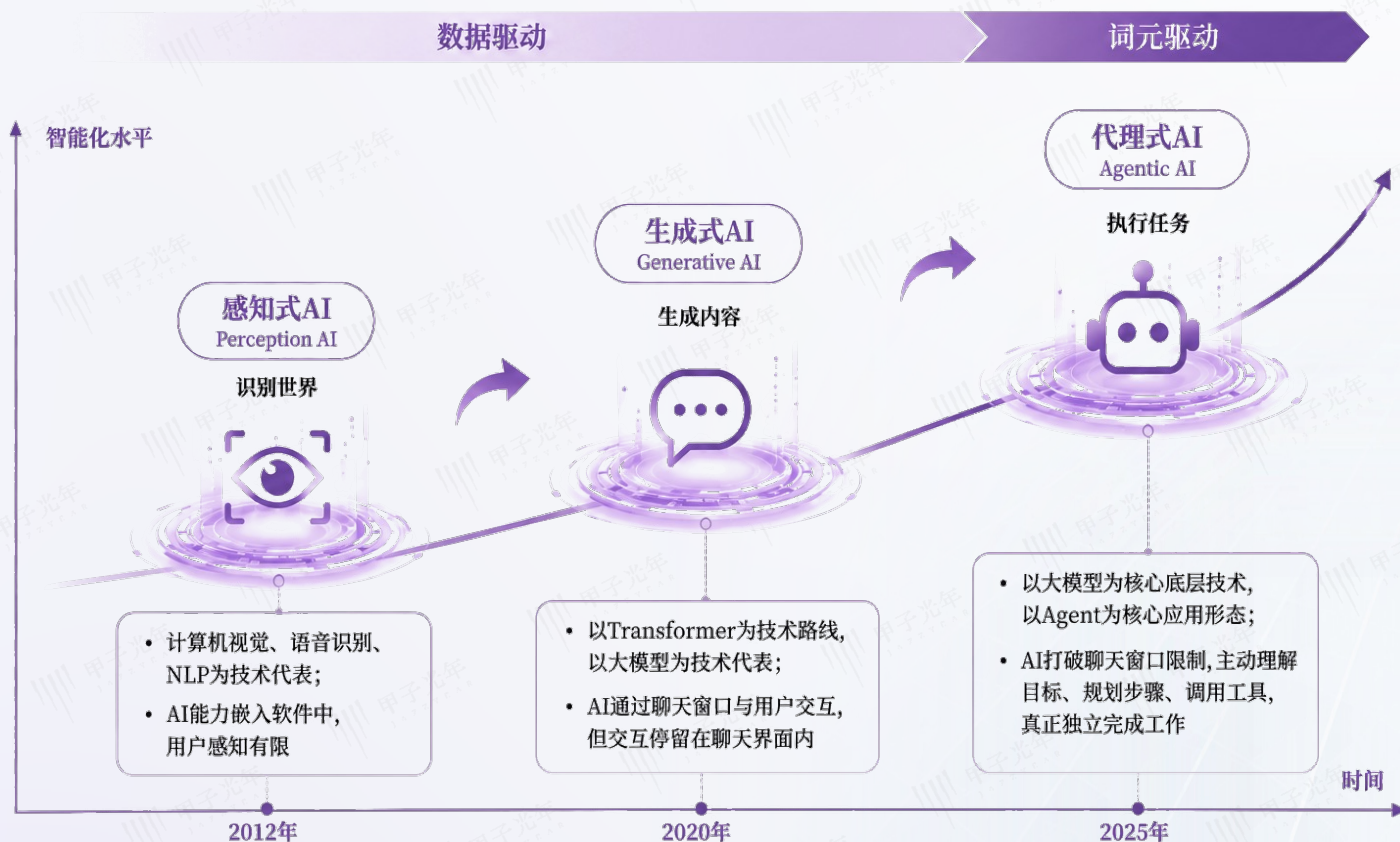
生成式AI阶段，海量语料、多模态数据与模型规模共同驱动内容生成能力突破，AI由识别判断走向知识理解与内容创造。

进入代理式AI阶段，AI技术价值由内容生成拓展至任务执行。数据与模型仍是智能能力底座，

智能体本质上是以大模型为推理与决策核心、在Harness工程机制支持下运行的任务执行框架。词元并非代理式AI阶段的新产物，生成式AI阶段已普遍采用输入与输出词元计量模型调用。但在代理式AI中，多轮推理、上下文更新、工具交互与结果反馈等使词元消耗随任务执行动态累积。为降低无效消耗、提升资源利用效率，产业通过智能路由、缓存复用、算网协同等方式优化词元产效，在控制成本的同时提升任务完成质量。词元的产业意义从模型调用的计量单位，延伸为衡量智能供给效率、连接基础资源投入与业务价值创造的重要尺度。

由此，围绕词元的生产、调度与消耗的工程体系逐步深化，正在深入影响整个AI产业链的分工格局与价值分配。

图2：AI技术能力演进



1.2 产业判断：词元本位时代开启

判断1：词元经济是生产对象、生产要素与消费协同升级的结果

经济模式的演进，体现为生产对象、生产要素与消费特征的同步升级。

生产对象的演进本质上是交付形态由实向虚、由物向智的升维。农业经济直接产出自然资源，工业经济将资源转化为商品，服务经济围绕商品形成专业分工，数字经济对分工流程进行数字化提效，词元经济则在数字化基础上嵌入智能能力，推动生产对象从“效率工具”进一步走向“代劳能力”。

生产要素围绕生产能力的来源持续演进，每一轮的经济形态都在把更高层次的能力嵌入生产过程。从土地、劳动力、资本、人才，到数据这一无形资产，再到词元，生产体系的驱动基础正由传统要素转向智能要素。

消费特征反映了人类需求重心的持续跨越。从满足基本的生存需要到物质功能，到服务体验，再到追求便捷实时的个性化体验，最后走向按需调用智能完成思考与执行，用户不再满足于标准化产品或服务，而是直接获取结果与能力。

图3：经济模式演进过程



注释：不同经济形态并非线性替代，上述时间表示相应生产范式集中兴起的阶段。

判断2：数据不等同于词元，词元构成代理式AI的标准化生产单位

在传统AI产业叙事中，数据往往被置于生产要素的核心位置，被视为智能能力形成的基础来源。但在代理式AI生产体系中，词元才是真正的标准化生产单位。

数据是词元的根，它记录客观世界，类似自然资源，是模型训练的原材料，影响模型参数与知识边界。但数据非标准化程度较高，难以形成统一计量口径，价值定价也高度依赖具体场景。数据成本主要集中在采集、清洗、标注与

治理等前置环节；沉淀为语料或知识资产后，后续复用的边际成本相对不明显。

词元是大模型处理信息的最小标准化单元，类似工业产品，标准化程度高。模型运行过程以词元为单位记录输入、输出与上下文消耗，因此具备清晰的计量口径。此外，词元是驱动Agent运行的直接燃料。Agent的调用、推理、生成与交互都会转化为词元消耗，并影响运行成本、响应速度与交付质量。任务复杂度提升会推动词元消耗增长，并增加推理计算需求，而实际运行成本会受推理过程与缓存机制共同影响。

图4：数据与词元在代理式AI生产体系中的差异

数据 (Data)	词元 (Token)
客观世界的数字化记录	大模型处理信息的最小标准化单元
类似自然资源（如原油）	类似工业产品（如燃油）
非标准化程度高	标准化程度高
模型训练的原材料	驱动Agent运行的燃料
影响模型的参数和知识边界	影响Agent的运行成本、响应速度与交付质量
计量口径复杂，定价依赖场景	消耗可计量、服务可计价
边际成本不明显，复制成本趋于零	边际成本明显，受推理计算与缓存机制影响

VS

判断3：代理式AI形成“数据造模型、模型生产词元、词元驱动任务”的三阶段生产链条

在数据造模型阶段，需要经过数据加工与模型训练两个环节。原始数据通过清洗、标注、离线编码预处理后投入训练，将海量知识沉淀为模型参数，通过预训练形成基础模型，再经后训练或微调适配任务与交互要求。

模型推理对应模型生产词元的阶段。输入内容首

先完成切分与编码，转化为模型可读取的词元序列，即形成输入词元。模型基于上下文完成推理，通过映射转化为向量送入网络，经过自回归循环生成输出词元，词元经过解码后形成文本、代码等可读内容。

Agent运行对应词元驱动任务的阶段。基于基础大模型的理解与任务拆解能力，Agent围绕任务目标进行工具调度、状态维护、协作、多

轮迭代执行等，直到达到任务的交付条件。

图5：代理式AI的生产过程，以自回归语言模型及相应Agent为例

	代理式AI生产环节	运行机制	阶段产出
数据造模型	数据加工 	对原始数据进行清洗、标注、去重、编码与质量控制	训练数据集，用于模型预训练、微调、测评
	模型训练 	从数据中学习模式与知识，形成模型参数，并通过预训练与微调形成具备生成能力的模型	基础大模型，含文本类、视频类、多模态类
模型产词元	模型推理 	输入内容经切分、编码与模型映射，被编码为输入词元，并基于上下文输出词元	输出词元序列，解码后形成文本、代码等内容
词元驱动任务	Agent运行 	基于基础大模型能力，Agent围绕任务进行工具调度、状态维护、协作、多轮迭代执行等，每轮模型调用处理输入词元，并输出词元，根据反馈持续执行任务直到完成	多轮推理形成的中间结果，直至具备任务交付条件

判断4：算力产业正从设备主导走向生产和效率主导，词元生产效率成为竞争核心

2023—2025Q1，AI算力整体供给偏紧。模型训练需求集中释放，叠加以英伟达高性能GPU在产能、交付周期与区域供应上的多重限制，GPU保有规模成为市场发展的关键约束。

2025Q1以来，以DeepSeek-R1为代表的国产模型加速成熟，Agent等AI应用逐步发展，算力需求由集中式训练向持续、高频的推理扩展。2026年以来，这一趋势进一步深化，推理成为算力需求的重要增量。

推理阶段，核心挑战在于：同等规模的算力资

源如何以更低成本、更快速度持续稳定地产出词元（Token）。算力不再仅是静态的硬件资源，而是一种以词元为产出的生产能力。

这一转变推动算力产业进入结构性拐点：市场从“算力稀缺、囤卡盈利”的设备主导时代，迈向“推理爆发、效率为王”的生产和效率主导时代。结构性转型主要体现在三个层面：

设备产能与价格层面，受英伟达GPU供给与价格约束，算力基础设施由单一架构转向异构架构，以更低成本的硬件形成同等规模或超过原有规模的可用算力。

生产效率层面，随着异构算力卡数量增加和集

群规模扩大，无NVLink/NVSwitch条件下的跨主机调度损耗成为核心瓶颈，算力优化从单卡性能提升转向集群系统效率优化。

词元生产成本与价值层面，算力经营从产出规

模转向单位词元的综合成本与产出价值。这既要统筹固定投入、运行支出与效率损耗，又要提升高价值场景中的有效利用，避免低效浪费，实现从“产得多”向“产得值”的转变。

图6：算力产业需求与供给运营逻辑演进

		2023-2025Q1	2025Q1-2026	2026至今
需求变化	推理如何超过训练	- 国产模型处于追赶阶段 - 训练需求集中释放 集中式训练主导	- DeepSeek-R1模型发布，标志国产模型走向成熟 - Agent开始快速发展，带动推理需求快速增长 训推分化日益明显	- 以Agent为代表的AI应用加速扩散，推理超过训练 - 算力需求转向低成本、快速、持续稳定产出词元
	设备产能与价格	GPU供给紧张，GPU保有规模决定产能	异构架构兴起，国产算力供给扩大，竞争转向产能规模与设备价格	
供给运营变化	生产效率	以单卡性能为主，效率未构成核心矛盾	异构集群扩大，异构卡数和集群规模扩大，通信、调度与软件适配成为关键，优化重点转向集群系统效率优化	
	词元生产与价值	关注设备获取成本，词元成本尚未成为核心指标	推理调用量增长，开始综合核算固定投入、运行支出和效率损耗，并兼顾单位有效词元成本与价值产出	

判断5：词元经济是围绕词元的生产、调度、流通、消费的新型经济活动模式

词元经济是以“词元”这一大模型处理信息的“最小标准化单元”为核心载体，围绕词元的生产、调度、流通、消费构建的新经济活动模式，关注词元如何被持续生产、动态组织、进入任务链路，并在运行过程中被消耗和转化。

从运行链路看，词元由电力、数据、算力和模型等要素共同生产；进入Agent任务链路后，在底层供给、中间运行与上层应用之间贯穿流动，并在被组织、分配、计量与持续消耗的过程中，支撑智能能力形成。

词元调用与消耗量快速增长，验证了词元经济正在进入规模化运行阶段。

国家数据局数据显示，2025年，全国日均词元调用量从年初的超万亿增长到年末的100万亿，全年词元调用量约21100万亿，而在2026年日均词元调用量已突破140万亿。

图7：词元经济运行链路

01 词元生产

词元的供给端，通过电力、数据、算力、模型等要素的共同作用，生成可被调用的词元

02 词元调度

词元在Agent任务链路中，受任务目标牵引和算力驱动，被组织、分配与计量

03 词元流通

词元进入模型、工具、知识等不同能力模块，并连接底层供给、中间运行与上层应用

04 词元消费

词元在Agent运行过程中被持续消耗，支撑理解、推理、生成与执行等智能能力的形成

判断6：词元经济的价值转化分析框架——编排能力（O）×词元（T）

在工业经济和服务经济中，生产函数通常描述资本（K）、劳动（L）等要素投入与产出之间的关系，通过资源投入推动价值增长。而在词元经济中，核心生产对象正在从物理资源转向智能资源——词元，为更好理解词元、编排能

高盛预测，到2030年4月，全球消费级智能体和企业智能体每月将分别消耗约60千万亿和56千万亿词元，合计消耗量（116千万亿）为2026年4月的24倍。

图8：2024-2026年3月中国日均词元调用量

单位：万亿

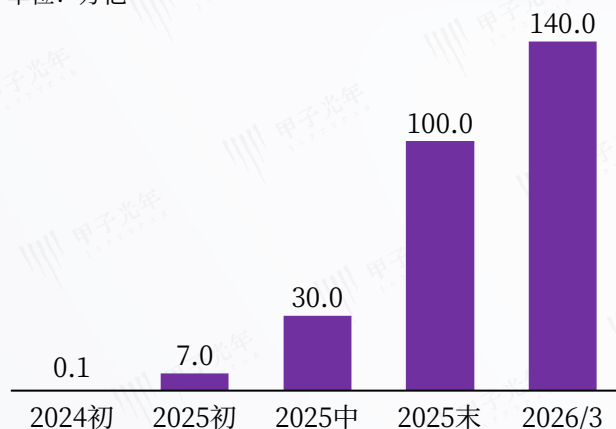
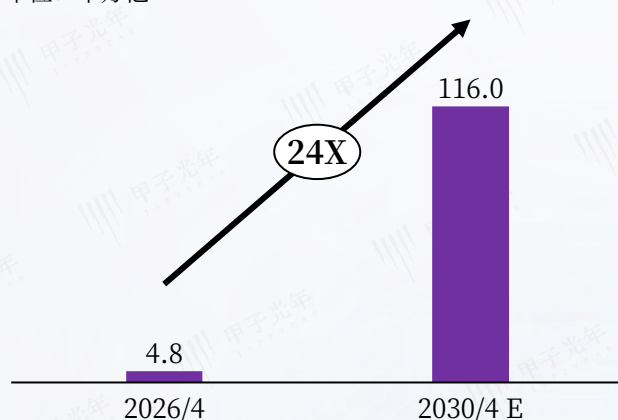


图9：2026年4月-2030年4月全球词元消耗量

单位：千万亿



力与业务产出之间的关系，可引入“ $Y=O \cdot F(T)$ ”的类比分析框架。

其中，T（词元）是模型智能生成的基础颗粒，每个词元都是一次预测，其质量受模型能力与上下文影响；F(T) 是持续产生的词元流，即智能供给的具体形态，其效能可近似视为单位预测质量与有效生成规模的乘积——质量由模型

能力决定，规模受任务需求与调用强度影响。词元本身不直接产生业务价值，而是承载模型输出的判断、计划与指令。

O（编排能力）则基于Agent框架与Harness工程，组织模型、工具与外部资源，将词元流

中的智能导入任务执行流程，从而影响智能供给的转化效率，最终形成业务价值Y。业务结果还会反向校验词元的成本与收益，驱动模型与编排策略持续优化，形成“生成—编排—兑现—反馈”的完整闭环。

图10：生产函数变化

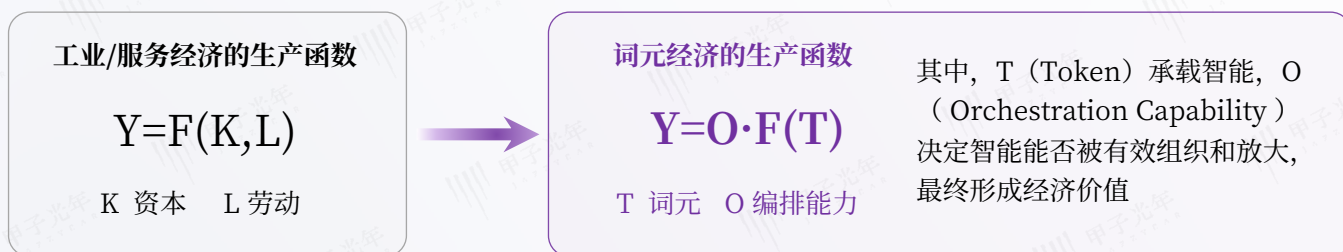


图11：词元到经济价值的转化过程



判断7：词元规模化消耗，倒逼运营治理体系建立，推动治理重心转向有效词元

当词元进入高频生产与规模化应用，消耗会在不同任务链路与场景中持续累积，无效调用、低质输出、资源空转、成本失控与安全隐患等问题也被同步放大。词元计量只能回答“消耗了多少”，无法判断“是否有效”。大量词元沉积在冗余上下文、错误路由、低质生成或无法落地的任务链路中，最终只会形成成本，难以沉淀为业务价值。词元治理的关键，是让词元消耗从可计量走向可观测、可评估、可优化。词元治理不是单点管理，而是贯穿生产、运行与价值沉淀全过程的运营治理框架。生产层关

注成本、效率与稳定性，通过成本精细化管理、生产效率优化、资源能效优化、安全部署与服务保障等措施，解决算力密集调用、产能吞吐约束、能耗散热压力和供给不稳定等问题。

运行层关注质量与可靠性，通过用量计量、智能路由、质量评估与反馈迭代、权限控制与审计等措施，解决用量冗余、路由失准、输出不稳定和运行安全隐患等问题。

价值层关注价值与收益，通过有效词元识别、交付效果评估和收益归因等措施，把词元消耗映射到任务结果与业务收益，避免词元只形成成本，难以沉淀为价值。

图12：词元生产—运行—价值沉淀的运营治理框架



1.3 范式落点：词元的产效、价值转化与运营治理呼唤新一代基础设施

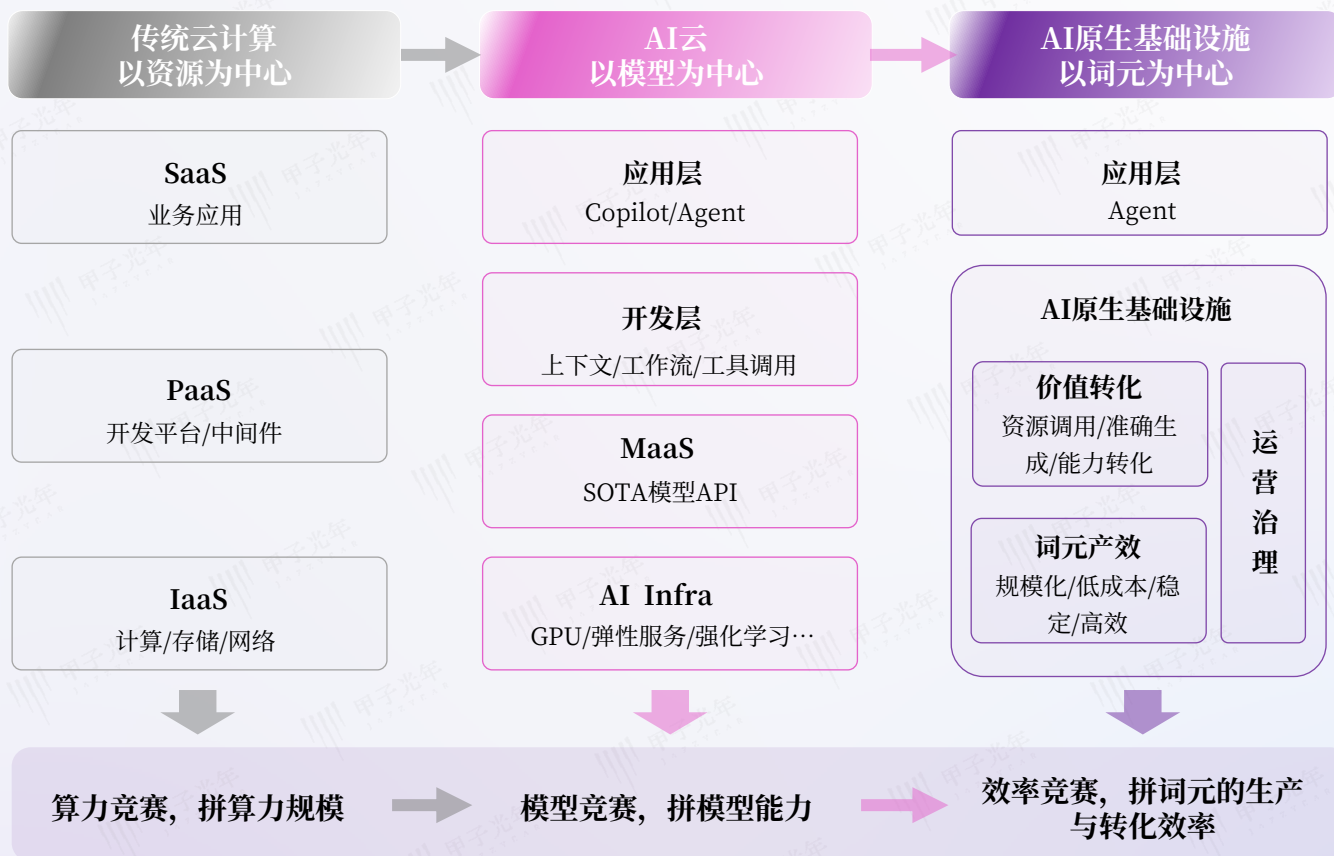
AI基础设施演进正在发生两重变化：一是供给形态由分层资源调用，转向生产、运行与治理的一体化支撑；二是能力重心由算力规模和模型能力，转向词元生产与转化效率。

从形态看，传统云计算以 IaaS、PaaS、SaaS 为主要形式，企业按层调用算力、存储、网络开发平台和软件能力，整体呈现分层供给结构。AI云阶段，GPU、弹性服务、MaaS与开发工具链被纳入同一云服务体系，基础设施开始围绕模型训练、推理和应用开发形成融合化供给，算力、模型与开发能力之间的边界逐步收敛。进入代理式AI阶段，复杂任务运行进一步要求

词元生产、编排转化与运营治理形成连续闭环，基础设施走向更深层的一体化生产与运行。

从能力重心看，传统云计算侧重资源规模，AI云阶段转向模型能力，AI原生基础设施阶段，模型能力逐渐成为基础配置，价值交付取决于三类能力：一是词元能否被规模化、低成本、稳定、高效地生产；二是词元能否在任务链路中经过能力组织、执行生成、产出封装与反馈迭代，转化为可交付成果；三是词元生产、消耗与交付过程能否被持续观测、评估和优化。基础设施建设由此从资源供给与模型能力，转向词元的产效、价值转化与运营治理。

图13：从资源中心到词元中心，AI基础设施范式演进



Part 02

新范式与新定义

AI原生基础设施

注：为贴近架构理念，本章节的词元统一称为Token，智能体称为Agent



2.1 AI原生基础设施 = Token Factory + Agent OS + Token Ops

定义：以Token产能、运行效率与价值创造为核心的代理式AI基础体系

AI原生基础设施（AI-Native Infrastructure）是面向Token产效提升与业务价值创造的代理式AI新型基础设施体系，其向下支撑Token稳定、规模化与高效生产，向上支撑Agent持续执行任务，贯穿全链路优化Token效率、质量与成本，推动智能能力转化为业务价值。

从结构上看，AI原生基础设施由Token Factory、Agent OS和Token Ops三部分构成。Token Factory组织底层生产资源，扩大Token产能并优化生产效率、质量与成本；Agent OS通过任务编排组织Token在模型、知识与工具间持续流动，优化Token的运行效率、质量与成本，并将智能能力转化为可交付的业务价值；Token Ops贯穿生产与运行全链路，通过监测、分析、维护与反馈，使各环节状态可观测、问题可定位、能力可优化，共同推动Token产效提升与价值创造。

Token Factory：组织电力、算力、数据与模型等资源，支撑Token产能并优化产出效率

Token Factory解决Token如何持续、规模化产出的问题，由能源、生产底座、生产资源与生产车间构成。它通过组织电力、算力、数据与模型等生产要素，推进智算生产底座建设，整合数据、算卡、模型资源并提供算力服务，从而提升Token产能；同时以资源调度、能源管理和软硬件协同降低闲置与损耗，使Token生产兼顾规模、效率、质量与成本。Token Factory的核心使命，是将分散的生产资源（以算卡为核心）组织为稳定、弹性、可交付的生产能力。

Agent OS：组织Token高效运行，推动智能能力向业务价值转化

Agent OS解决Agent如何组织各类能力、持续完成任务的问题。Agent接收自然语言输入后，对目标进行理解与拆解，并由Runtime提供任务执行环境，按需调用路由、模型、上下文、记忆、知识与工具等功能模块，推动任务由指令进入实际执行过程。

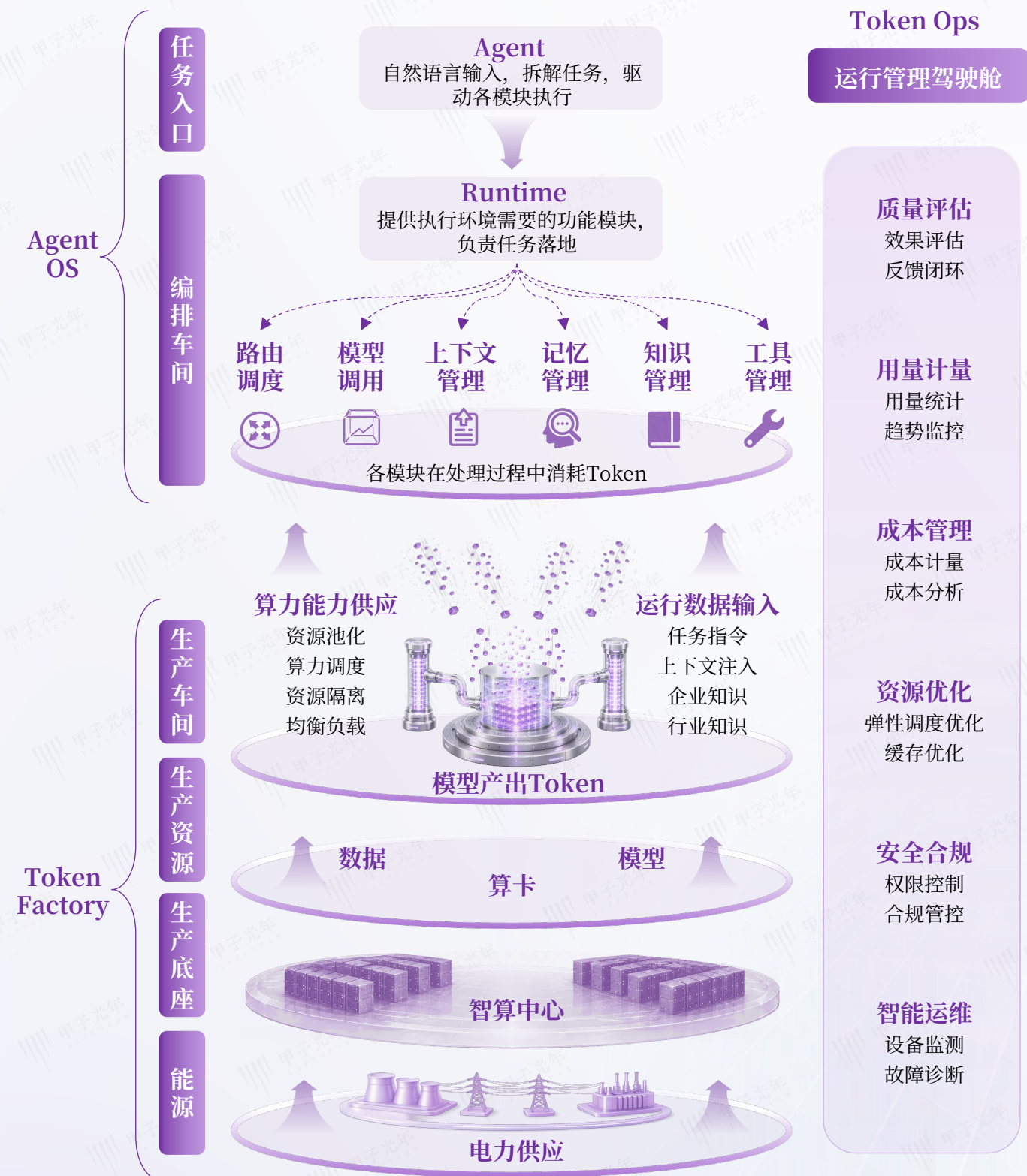
面对跨步骤、长周期的复杂任务，Agent OS还需维持任务状态与信息连续性，协调各模块分工协作，使Agent能够围绕目标持续推进，而非止于单次模型响应。其核心在于组织模型调用、上下文与工具执行，使模型输出进入可执行的任务流程，其中模型调用按明确口径计量Token，让分散的智能能力形成连贯的执行链路，最终将Token转化为可交付的任务结果与业务价值。

Token Ops：贯穿Token全生命周期，以运营治理持续优化产能、效率、价值

Token Ops解决Token生产与运行过程如何被持续监测、治理和优化的问题。这一运营体系汇聚生产、调用、任务结果与设施运行数据，对Token用量、质量、成本、收益、安全及设备状态进行监测分析，识别各环节的损耗、异常与风险，并将结果反馈至生产和运行模块。

同时，通过计量计费、权限合规、调用审计和智能运维明确运营边界，使Token产能、运行效率与业务价值可观测、可审计、可优化，形成安全可信、贯穿全链路的持续改进机制。

图14：AI原生基础设施概念图



特征：能够持续稳定生产Token，并以Token驱动Agent运行时编排，支撑任务完成

技术范式的每一次跃迁，都会推动基础设施的形态、核心生产资源、关键任务与对外价值产出发生变化。

互联网时代，基础设施以网络为核心，承接流量入口，完成终端连接，形成信息触达能力；云计算时代，基础设施升级为云，承接企业工作负载，通过资源调度支撑应用运行；生成式AI时代，基础设施进一步演进为AI云，承接用户意图，以大模型进行Token生产，形成内容生成能力。

进入代理式AI时代，核心入口从意图升级为任务。Agent不再只是响应一次Prompt，而是需要持续判断目标、拆解步骤、调用模型、检

索知识、选择工具并处理反馈。任务执行往往跨越多个模型、工具、知识源和业务流程，基础设施需要在运行时对各能力模块进行动态组织与调度，这构成了Agent运行时编排的核心。

在Agent运行时编排期间，每一次理解目标、拆解任务、选择工具、读取上下文、修正路径与生成结果，都会伴随Token的持续消耗与新的Token生成。Token既是AI原生基础设施持续生产的对象，也是Agent运行过程中被持续消耗的运行燃料，支撑任务执行链路不断推进。

AI原生基础设施的不同之处在于，它不仅需要持续、稳定地生产Token，更需要围绕模型、工具、知识、上下文与Agent，形成统一的运行时编排体系，实现Token的高效供给、动态调度与价值转化，支撑复杂任务稳定完成。

图15：不同基础设施形态的核心差异

比较维度	互联网	云计算	生成式AI	代理式AI
核心入口	流量	工作负载	意图	任务
基础设施形态	网络	云	AI云	AI原生基础设施
核心生产资源	网络带宽	算力	大模型	Token
关键任务	连接终端	调用资源	Token生产	Token生产+Agent运行时编排
Token角色	/	/	计量单位	生产对象，运行燃料
对外价值产出	信息触达能力	应用运行能力	内容生成能力	代劳能力

2.2 典型架构解析：蓝耘科技·AI原生基础设施架构

整体架构

Token Factory

AI Infra OS

Token Ops

AI Native Agent

围绕Token生产、调度、应用与治理构建一体化AI原生基础设施

随着生产要素向智能要素演进，Token成为代理式AI体系中可标准化计量、生产与调用的核心生产单位，贯穿模型推理、资源调用与任务执行，连接智能与业务价值。

AI基础设施的核心诉求也由算力产能交付，转向对Token成本、效率、质量与价值的优化。打通Token的规模化生产、智能调度、运营治理与场景化消费，成为落地的关键。

基于此，蓝耘科技构建以Token Factory和AI Infra OS为核心的一体化AI原生基础设施，向上赋能Agent应用，并以Token Ops贯穿Token生产、调度与应用全链路。

Token Factory：智算底座，提供规模化、稳定、高质量、低成本的Token生产能力

底层能源：通过稳定供配电、绿色能源接入与能耗管理，保障电力持续供应，降低散热成本，满足低碳要求，为智算产能扩张提供能源支撑。

生产设施：以“自建+合作租赁”方式构建智算中心，通过高密度机柜、液冷系统及智能运维提升单位空间承载能力，并整合国产异构芯片、数据与模型资源，扩大可用产能，降低对单一GPU的依赖。

算力服务供应：通过裸金属与Serverless容器云提供独享或弹性算力服务，兼顾性能与资源复用，减少算力空置；通过AI数据生成平台提

供数据采集与质检服务，降低数据准备成本。

AI Infra OS*：连接Token生产与应用需求的智能调度中枢，提升算力转化为Token的效率

L1算力供应：统一纳管异构算力，支持弹性伸缩、私有化部署与监控运维，形成稳定算力池，提升资源利用率。

L2推理引擎与调度：以引擎优化、统一网关和智能路由、成本工程与模型镜像仓库匹配模型与算力，同时依托分布式KV Cache，提升吞吐及缓存利用率，使同等算力生成更多有效Token。

L3模型与质量：通过场景选型、MaaS服务和质量评测，形成适配不同任务的模型，提升Token生成质量。

L4知识与上下文：通过RAG、上下文编排、行业知识包与Agent模板，注入企业知识和业务语境，使Token更懂业务。

Token服务模式：面向不同客户需求，提供标准化、专属化与私有化三类服务

公共Token服务主要面向中小企业与个人开发者，以多租户共享、标准API、按量或套餐计费，提供弹性、低门槛的Token调用服务。

专属Token服务面向大模型厂商、规模化AI应用企业及中大型行业客户，通过独享资源池、独立实例和约定产能保障，满足大规模、高并发Token生产需求，提升Token供给的稳定性、

整体架构

Token Factory

AI Infra OS

Token Ops

AI Native Agent

性能及成本可控性。面向政务、高校、金融等对数据安全、合规及自主可控要求较高的行业客户，通过Token生产组件私有化部署，支持企业定制化服务。

Token Ops：贯穿全栈，以运营治理推动Token增产、降本与价值兑现

以计量计费将Token使用转化为可核算收益，通过成本核算与收益分析，实现Token生产与消费的精细化运营；以权限合规和安全防护保障可信交付，通过多租户隔离提升资源复用与服务产出，并依托调用留痕与审计追踪，增强服务可追溯性与运行稳定性，借助智能运维减少停机损耗与运维成本。

AI Native Agent: Token的消费终端，将高质量Token转化为实际业务价值

蓝耘科技已形成天衍桌面端Agent、星河营销Agent等应用，在真实场景中消费Token并转化为业务价值。应用侧产生的任务需求、执行反馈与效果数据进一步反馈至智算底座和AI Infra OS，形成“生产与调度—消费—优化生产与调度”的业务飞轮。

基于上述整体架构，下文将重点围绕Token Factory、AI Infra OS、Token Ops及AI Native Agent四大核心模块展开，分别解析Token生产、智能调度、运营治理与应用价值转化的关键能力。Token服务模式作为连接AI原生基础设施与上层应用需求的服务交付层，主要承担Token能力的标准化封装与差异化供给，其公共、专属及私有化服务模式已在整体架构中予以说明，后文不再单独展开。

整体架构

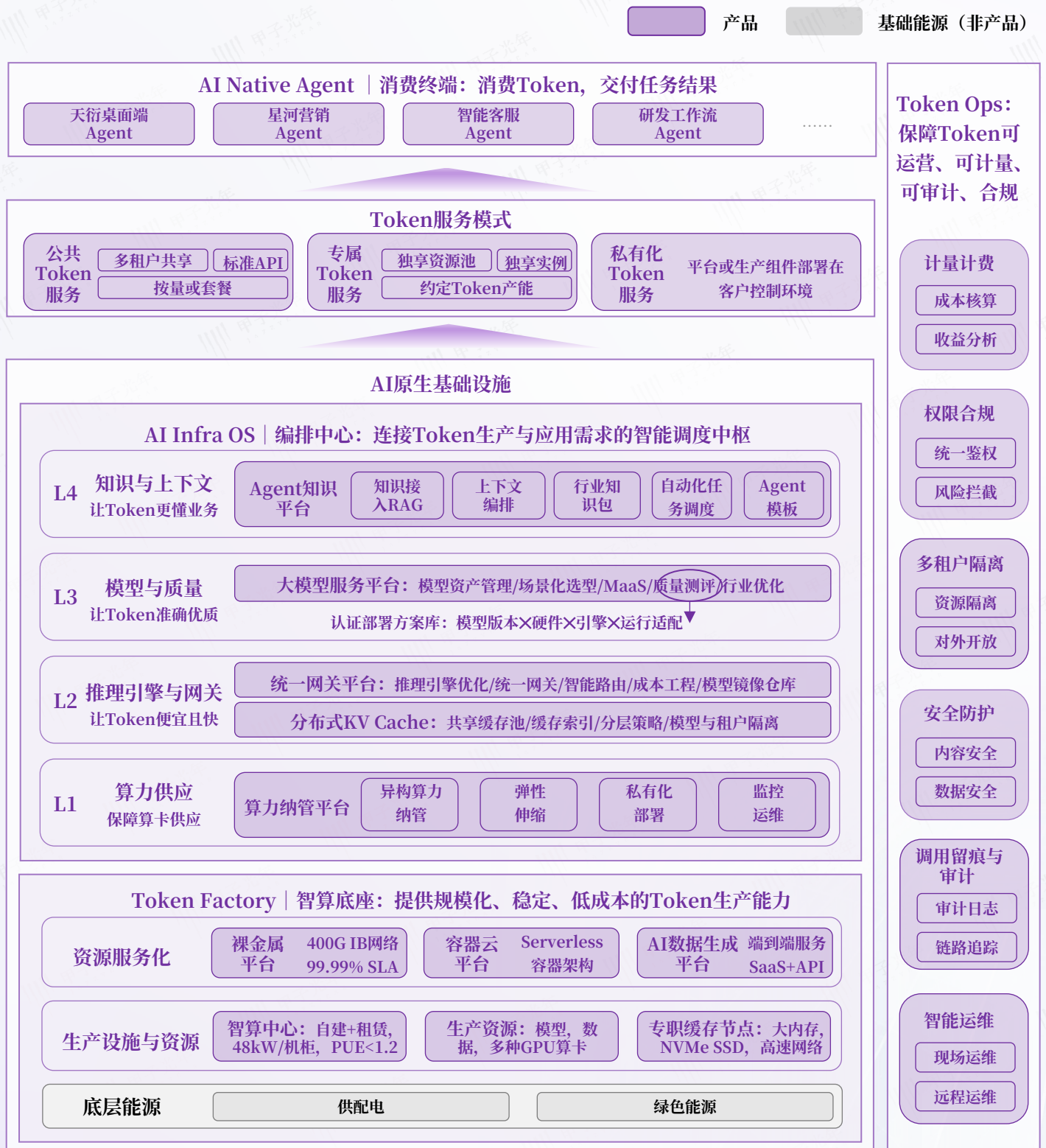
Token Factory

AI Infra OS

Token Ops

AI Native Agent

图16：蓝耘科技·AI原生基础设施架构



整体架构

Token Factory

AI Infra OS

Token Ops

AI Native Agent

智算中心建设：以“风改液”与“绿电技术”突破散热与能源约束，保障Token稳定生产

随着GPU功率密度和推理需求持续增长，智算中心面临散热压力上升、传统机房承载受限以及电力供给难以支撑算力扩张等问题。Token稳定生产不仅取决于算力设备规模，也取决于机房能否承载更高功率设备，以及能源能否持续、稳定供给。

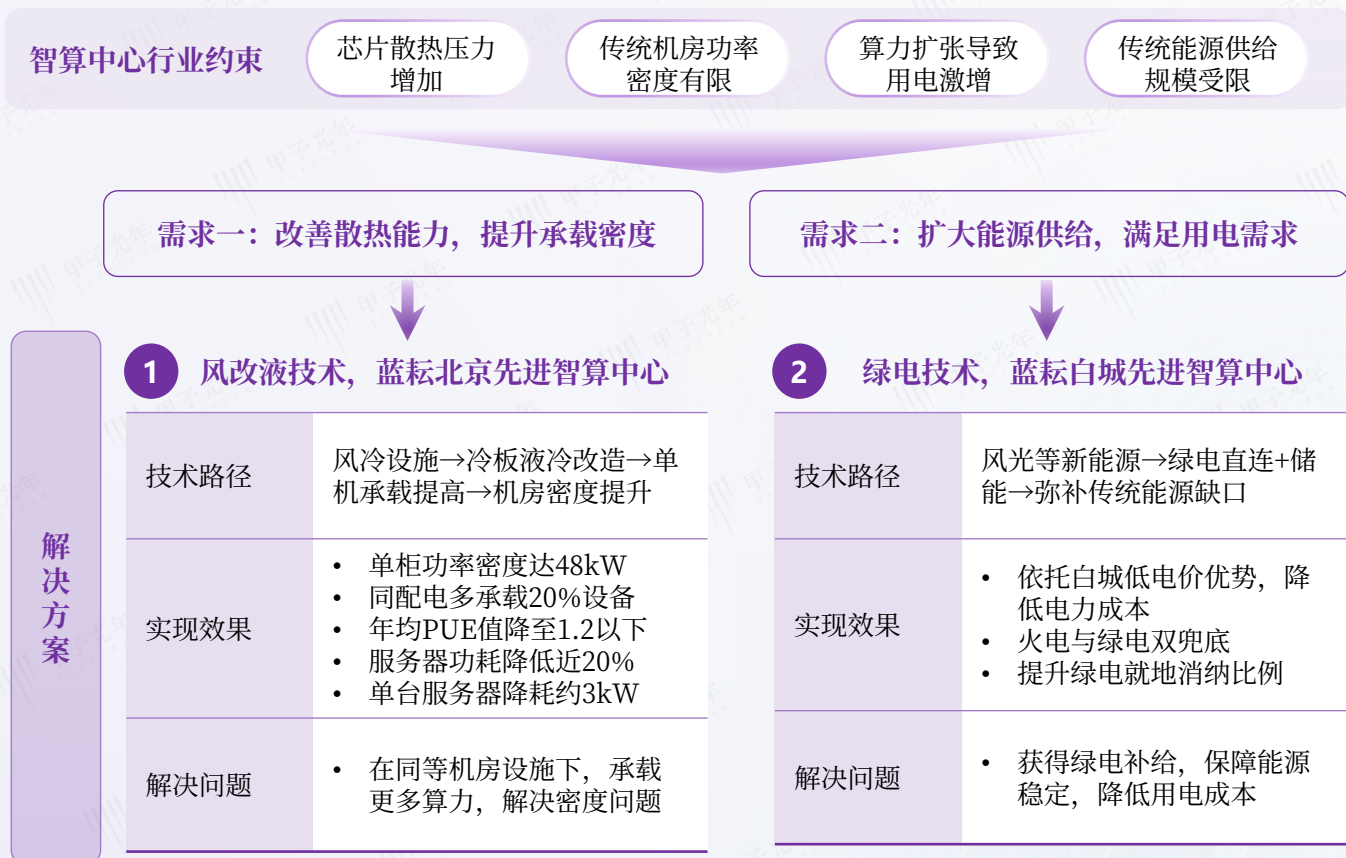
针对上述约束，蓝耘科技围绕提升设备设施承载密度与拓展能源供给能力两条路径，推进智算中心建设。

在提升承载密度方面，蓝耘科技采用“风改液”

技术路径，将风冷设施改造为冷板式液冷系统，突破风冷散热上限。以蓝耘北京先进智算中心为例，改造后的智算中心单柜功率密度达48kW，在总配电不变的前提下可多承载20%的IT设备，年均PUE降至1.2以下，同等机房设施下承载更多算力，缓解了机房密度瓶颈。

在拓展能源供给方面，蓝耘科技采用绿电技术路径，通过风光等新能源，弥补传统能源供给缺口。在蓝耘白城先进智算中心中，该模式依托当地低电价优势，实现火电与绿电双兜底，提升绿电就地消纳比例，降低了用电成本。

图17：蓝耘科技“风改液”与“绿电技术”的智算中心建设实践



整体架构

Token Factory

AI Infra OS

Token Ops

AI Native Agent

图18：风改液方案实景：蓝耘北京先进智算中心

北京先进智算中心外景



机柜液冷管路部署



液冷循环动力系统



冷却液供水系统



图19：绿电方案实景：蓝耘白城先进智算中心

白城先进智算中心外景



智能算力机柜集群



算力机柜阵列



配电系统



整体架构

Token Factory

AI Infra OS

Token Ops

AI Native Agent

算力软件层优化：以调度框架与推理引擎释放集群算力，推动Token生产降本增效

硬件解决了集群“搭得起”的问题，但国产芯片缺乏NVLink高速互联，多卡及跨主机协同更多依赖PCIe与网络传输，易产生通信与任务等待损耗，且随集群规模扩大而加剧。开源推理组件生产化落地门槛高，新增算卡难以等比例转化为有效算力，直接推高Token生产成本，拉低Token产出效率。

针对上述问题，蓝耘科技从调度框架与推理引擎两个软件层面规划了相应的技术路径，致力于让集群“连得通、跑得快”。

在调度框架层面，蓝耘正自研TT分发调度框架，

开源组件已支持多节点推理，但在特定异构硬件、网络条件与业务负载下，仍需完成适配、调优及生产化治理，以减少通信与调度开销。框架沿“同构集群调度—异构混合调度—固件级ARM总线控制”演进，目标是将调度时延降至微秒级，提升多卡级集群有效算力利用率，降低跨主机性能损耗。

在推理引擎层面，蓝耘规划以缓存为核心的推理系统，围绕KV Cache优化、跨主机显存动态共享与自适应批次调度三项能力，提升缓存命中率、缓解显存占用、减少任务等待，推动算力转化为Token，降低单位Token成本。

图20：蓝耘科技·AI Infra OS L1-L2层调度框架+推理引擎集群效率优化方案



整体架构

Token Factory

AI Infra OS

Token Ops

AI Native Agent

推理高效化：以模型供给、智能路由与PD分离打通模型接入、选型与算力部署全链路

推理任务从发起到完成，需经过模型接入、模型选型与算力部署三个环节，但各环节均存在效率问题，并推高Token生产成本。接入环节，多模型接口、能力、价格标准不一，抬高接入管理成本；选型环节，任务与模型错配，高成本模型低效调用、低成本模型无法胜任；部署环节，Prefill与Decode资源需求不同，统一部署易导致算力利用不足。

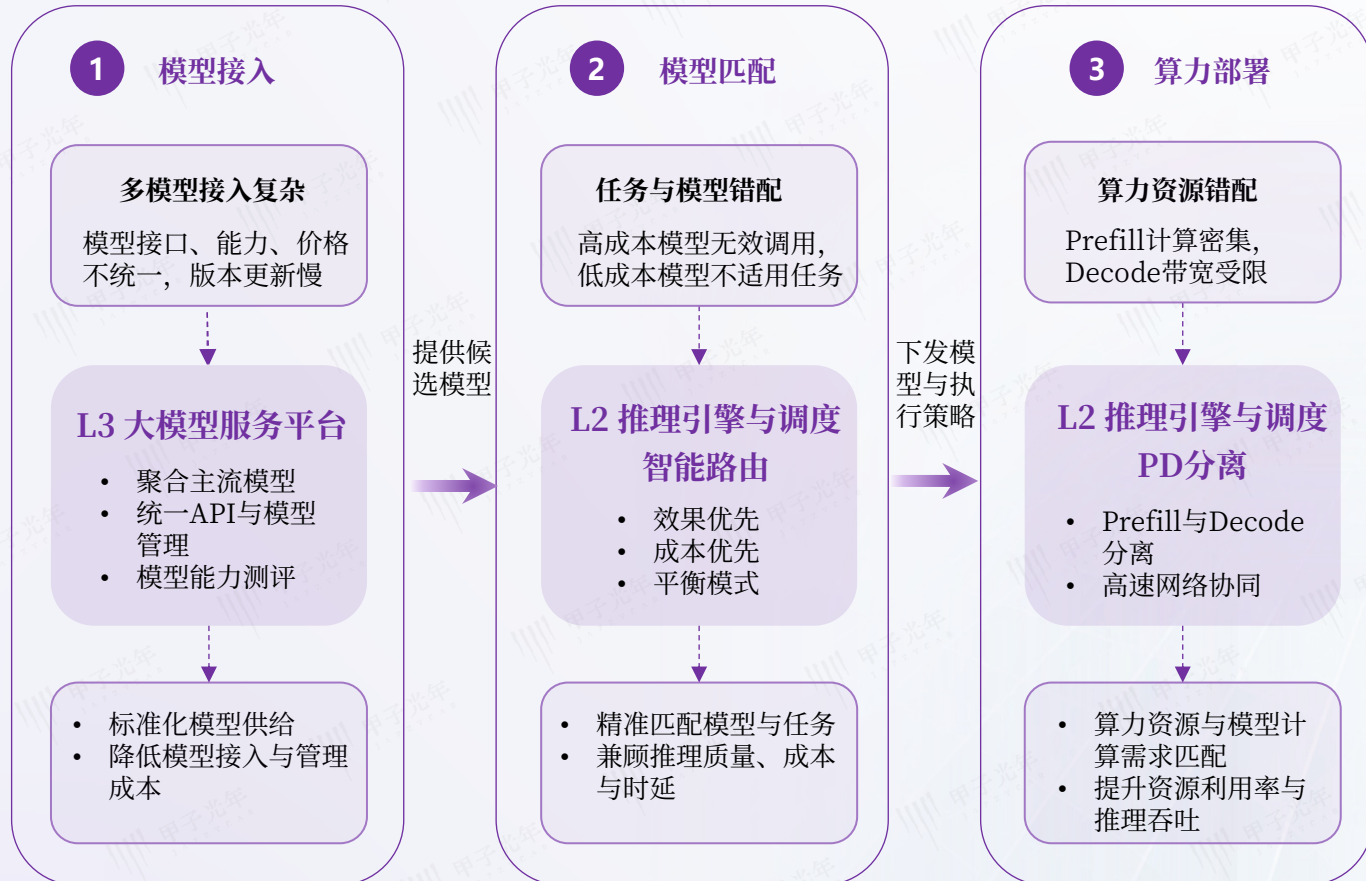
针对接入环节，蓝耘科技提供大模型服务平台，

统一接口聚合多类模型，并持续接入主流SOTA模型，为企业提供标准化API与弹性服务。

针对选型环节，蓝耘科技已经构建智能路由，根据任务对质量、成本与时延的要求，采用效果优先、成本优先或平衡模式，动态选择模型资源与调用链路，减少高成本模型无效调用。

针对部署环节，蓝耘科技实施PD分离方案，将Prefill与Decode两类任务分别部署于适配的算力资源，并采用高速网络协同。该策略可减少算力资源错配，在保障推理质量的同时提升吞吐与算力利用率，降低Token生产成本。

图21：蓝耘科技·AI Infra OS L2-L3层模型供给、智能路由与PD分离方案



整体架构

Token Factory

AI Infra OS

Token Ops

AI Native Agent

知识与上下文工程：沉淀企业知识与成熟能力，推动Agent持续进化，提升Token业务价值

Agent具备基础任务执行能力，并不意味着能够稳定完成复杂业务。专业知识不足容易造成输出泛化，任务上下文不连续会导致理解偏移，成熟能力未能沉淀则会形成重复开发，进而造成Token浪费，制约Token向业务价值转化。

蓝耘科技从三方面完善L4知识与上下文能力：一是通过知识接入RAG，将多源数据与行业知识转化为可检索、可引用的企业知识资产，增强Token的专业性与业务相关性；二是规划建

设上下文编排，统一组织会话记忆、任务状态和多Agent协作信息，提高复杂任务执行的连续性与准确性；三是计划将经过场景验证的Skills、工具和工作流封装为可复用的Agent能力，减少重复开发并缩短构建周期。

知识、上下文与Agent模板在场景运行中根据效果反馈持续迭代，共同推动Agent能力进化。一方面促进与业务紧密结合的高价值Token形成，减少错误生成、重复调用与人工校正；另一方面降低Agent开发与使用成本，促进价值提升与成本优化。

图22：蓝耘科技·AI Infra OS L4 知识与上下文解决方案



整体架构

Token Factory

AI Infra OS

Token Ops

AI Native Agent

前瞻布局：以MaaS应用发布平台承接多云融合，以自动发现协议构建Agent小程序生态

Agent是包含用户交互、业务处理、工具调用与模型推理的完整AI应用。通用云计算承载应用服务、数据处理和网络连接，云推理提供GPU算力与模型能力。当前两类资源分立部署，开发者需自行完成资源配置、网络连接、服务发布与运维，增加Agent交付复杂度。

与此同时，未来AI应用将形成“中心Agent+专业Agent”的小程序生态，由中心Agent拆解需求，发现并调用专业Agent协同执行。Agent由独立应用走向网络化协作后，需要建立统一的注册、发现与互联机制，使分散能力能够被识别、组合和流通。

顺应上述演进，蓝耘科技从平台化供给与生态互联两方面延伸AI Infra OS：规划以MaaS应用发布平台统一封装云计算与云推理能力，并以Agent自动发现协议*支撑Agent跨平台互联。

多云融合：云计算与云推理一键封装，降低Agent应用部署门槛

针对云计算与云推理资源独立部署、增加Agent开发难度的问题，蓝耘科技规划通过MaaS应用发布平台，统一接入和封装云计算资源与云推理能力。

在该方案下，开发者按标准提交Agent定义，含前端页面、Prompt配置、工具链声明、知识库关联及算力规格，平台自动完成资源分配、服务部署、网络配置与弹性伸缩策略，实现从代码到上线服务的一键发布。

通过屏蔽两类云资源的配置与协同复杂性，开发者可聚焦Agent业务逻辑，将基础设施、模

型调度、部署运维、能力分发等复杂工作交由平台承载。

生态互联：以Agent自动发现协议连接分散能力，促进Agent发现与协作

针对Agent分散在不同平台、能力难以识别和复用的问题，蓝耘科技正在自研Agent自动发现协议，通过标准化描述Agent的身份、能力与服务接口，使已发布Agent能够注册至全网资源目录，其他开发者、企业用户及合作平台可在各自生态渠道中搜索、发现并调用已注册的Agent能力。

另外，MaaS应用发布层的Agent一键发布平台提供主动分发机制——Agent发布方可选择将能力推送至指定合作渠道，实现“发布即发现、发行即上线、推送即授权”。该机制推动Agent能力由孤立应用走向生态流通，为跨平台发现与协作奠定基础。

同时，蓝耘规划了统一的Agent发布标准格式，包含前端页面定义、Agent逻辑配置、资源声明三部分。开发者按标准格式开发Agent，即可在平台发布并被兼容生态识别。标准化的封装格式不仅是“一键发布”的技术前提，也是未来跨平台Agent互操作的基础。

通过MaaS应用发布平台与Agent自动发现协议，蓝耘科技推动AI Infra OS由内部运行支撑体系向开发者和合作生态延展，形成“Token生产—Agent运行—能力发布—生态互联”的完整链路。

整体架构

Token Factory

AI Infra OS

Token Ops

AI Native Agent

Token Ops：以全链路数据汇聚支撑精细经营，让Token从消耗指标变为运营对象

Token贯穿推理运行、Agent任务执行与业务结果生成，但各环节的运营数据分散、难以统一关联，导致成本难归集、消耗与业务价值脱节、流转不可追溯、状态波动无保障等难题。只有将每次Token消耗还原到具体资源、任务与业务结果，才能判断投入是否形成有效产出，并为后续优化提供依据。

蓝耘科技通过Token Ops 汇聚推理运行、Token计量、Agent任务、身份权限及生产设

施数据，建立覆盖用量、成本、质量与业务价值的运营账目，支持成本核算、效果评估与收益分析；同时通过权限管理、多租户隔离、安全防护和调用审计保障Token可信使用，并以智能运维监测设备状态、识别异常和优化能耗，保障生产环境稳定运行。

Token Ops成为连接底层生产与上层价值创造的运营枢纽，将Token调用转化为可计量、可分析、可审计、可结算的经营活动，使向下的资源成本与向上的业务价值能被量化和持续优化，推动基础设施由技术供给走向价值经营。

图23：蓝耘科技·Token Ops全链路运营治理体系



整体架构

Token Factory

AI Infra OS

Token Ops

AI Native Agent

以场景落地验证价值，推动成熟能力向更多场景与行业复制

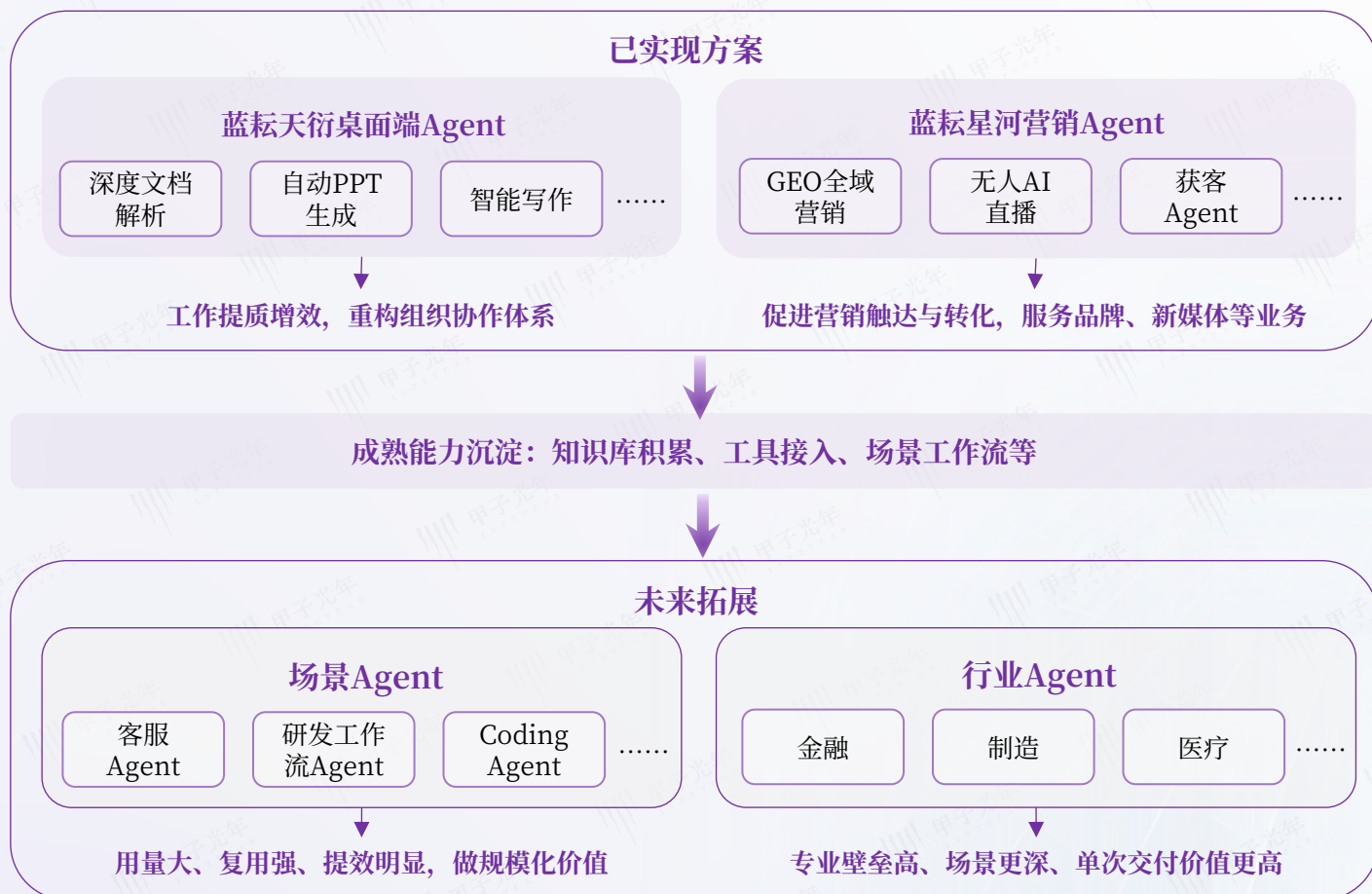
Agent应用是Token进入业务流程并形成可交付结果的关键载体，Agent将Token消耗转化为任务执行、效率提升与业务增长，使智能能力的价值能够在真实场景中被验证。

目前，蓝耘科技已形成天衍桌面端Agent与星河营销Agent两类应用。前者围绕深度文档解析、PPT生成和智能写作等办公任务，提升工作效率并重构人机协作方式；后者通过全域营销、无人AI直播与获客Agent等能力，服务营

销触达、客户转化及品牌与新媒体业务。真实场景的运行也推动知识库、工具接入与场景 workflow 等成熟能力持续沉淀。

未来，蓝耘科技将沿两条路径拓展Agent应用：一是将成熟能力横向复制至客服、研发、Coding等通用业务场景，以高复用和规模化使用创造提效价值；二是结合金融、制造、医疗等领域的专业知识、业务流程与合规要求，向行业Agent纵向深化，形成更高的专业壁垒与单次交付价值。

图24：蓝耘科技AI Native Agent应用落地与价值拓展路径



Part 03

新分工与新模式

产业分工、商业、组织与治理重构

注：为贴近产业理念，本章节的词元统一称为Token，智能体称为Agent



3.1 产业新分工：Token Factory专业化分工与价值再分配

Token产能与效率的双重需求重塑分工，独立智算服务商因承接算力生产调节价值而兴起

Token产能扩张与生产效率提升的双重需求，推动Token Factory进一步分工。云计算厂商拥有客户和云资源基础，但推理需求峰谷波动明显，自建扩容周期长、跨地域调度复杂，需要借助外部供给弹性补充产能、提高资产利用率；模型厂商擅长模型研发，却面临自建算力投入重、异构适配与持续运维成本高等问题，需要稳定、高效且更具性价比的算力支撑。

国产芯片厂商则面临规模采购不足、适配成本

高与生态成熟缓慢等制约，需要通过项目部署、运行验证与生态磨合，实现规模化落地。

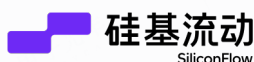
三方诉求共同催生独立智算服务商。其整合多源异构算力，提供设施运营、弹性供给、跨域调度及软硬件适配，承接波动性需求并促进国产算力应用，成为算力生产的专业调节者，推动产业从重资产自建走向分工与合作供给，产业价值由单一的算力资源销售，转向围绕Token产能与效率的专业化协作与分配。

图25：Token Factory视角下的三类需求催生独立智算服务商



三类需求力量汇聚

★ 独立智算服务商 应运而生，成为算力生产的调节者，实现规模化算力撮合与弹性调度



独立智算服务商贯通生产组织、资产运营与生态连接，形成三重核心能力

独立智算服务商并非简单采购与转售算力，而是跨越算力生产、资产经营与生态流通三个环节，将分散、异构且波动的算力资源进行运营和盘活，并转化为稳定产能、可持续资产收益与标准化服务供给。其价值既来自技术提效，也来自对资产状态、供需波动和产业协作关系的持续经营。

作为生产组织者，其适配不同芯片、架构与集群，统一纳管跨地域算力资源，并依据任务需求动态编排调度，扩大有效供给、提升Token生产效率。

作为资产运营者，其覆盖智算设施建设、运维、

容量规划与能源管理，通过软硬协同提高设备利用率、优化运行状态和提供弹性保障，降低空置与故障风险，提升资产全生命周期收益。

作为中立生态连接者，其连接芯片、云、模型与应用需求，将异构资源封装为可调用、可计量的标准服务，降低生态锁定、供需匹配与产业协作成本。

三重能力随适配经验、运营数据、资源规模和合作网络持续积累，并相互强化，共同构成独立智算服务商的差异化竞争优势。

图26：独立智算服务商的三重角色与核心能力



3.2 商业新模式：建立效果计价与成本计价的桥梁

行业面临Token效果计价与成本计价之间的断层问题

云服务按GPU卡时、存储和带宽收费，软件服务按账号和模块收费，大模型API则普遍按照Token消耗收费。在模型API时代，用户直接决定输入内容和调用规模，Token消耗与使用量基本对应，按量计费能够较好反映服务成本。

随着Agent开始承接客户服务、营销获客、企业招聘、代码开发等复杂任务，用户购买的对象逐渐从模型调用转向任务结果。从客户视角看，消耗多少Token不是首要问题，任务是否完成、结果质量如何、是否创造业务价值，正

在成为新的产品评价标准。

然而，供应商仍需围绕Token消耗、模型调用和算力投入进行成本核算。Token依然是供应商内部的重要成本单位，却难以直接映射客户感知的业务价值。

虽然当前的商业模式已向Token计量转变，但效果评估缺位制约了新范式确立。产业面临的核心挑战不是统一计量Token，而是在Token成本与业务价值之间建立稳定的映射关系。

图27：Token的业务价值与成本尚未形成统一映射



在供需之间建立统一的Token价值映射机制，实现成本与价值的有效对齐，正成为新商业模式待解决的重要命题

成本可拆、价值可计量，在成本与效果之间建立映射桥梁

Token成本可逐层拆解，价值可按任务计量，成本与效果之间的映射桥梁由此建立，多元计价成为可能。

Token成本可拆解为资产与能源、生产运营、财务三类来源，并分别对应硬件与能源优化、生产运营提效、经营与资金优化三条管控路径，实现分项降本。

在此基础上，以任务为对象归集成本投入、评估完成质量与交付效果，将Token投入与业务产出逐一关联，形成服务成本基准与价值衡量依据。映射关系一旦建立，收费模式便可从单一用量计费走向组合定价：用量计费覆盖生成投入，能力收费体现技术与服务溢价，效果增值收费按任务成果加成，Token的价值创造由此获得可计量的兑现通道。

图28：Token成本与效果映射及组合定价机制



3.3 组织新模式：以任务驱动为核心的AI原生组织

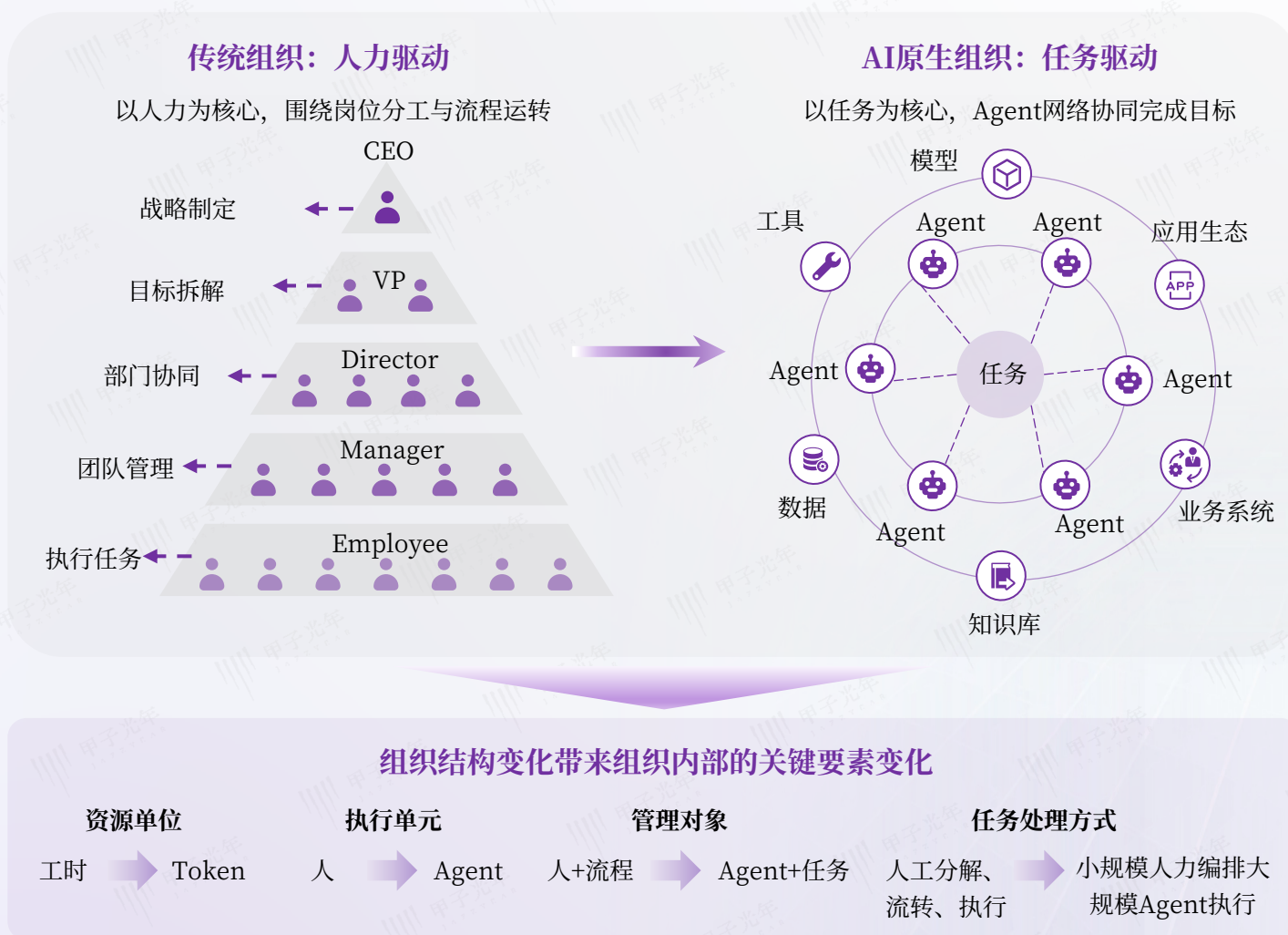
AI原生基础设施的演进，正在推动组织形态发生变化。随着模型、工具、数据和应用接口逐步标准化，Agent得以跨系统调用多类资源，而Token为模型调用部分的统一计量尺度。企业获得了除人之外的新执行主体，组织结构也随之发生调整，从“人力驱动型组织”转向“AI原生组织”。

在组织架构的转变过程中，任务成为新的组织中心，Agent集群由人完成编排，并围绕任务

协同运作，调用模型、工具、数据、知识库、业务系统、应用等资源，任务执行方式更灵活。

组织结构变化带来了组织内部的关键要素变化。资源单位由工时转向Token，执行主体由人扩散为Agent，治理重心由员工与流程转向Agent与任务，任务执行方式则由人工分解、流转和执行逐步转向小规模人力编排大规模Agent协同完成。

图29：从“人力驱动型组织”到“AI原生组织”的结构重构与要素变化



3.4 治理新模式：面向Agent行为的约束机制

随着Agent从内容生成走向任务执行，治理对象正从模型能力扩展至Agent行为。面对恶意提示词注入、身份与权限滥用、供应链风险、记忆投毒、工具与资源滥用等潜在风险，全球监管逐步形成治理的关注重点，行业实践也开始构建身份、权限、行为和执行过程等工程治理能力，Agent治理框架正在形成。

监管层面，各国法律法规及行业标准围绕风险

分级治理、透明义务、人类干预和责任划分等方向，立法标准正在形成共识，但在具体治理路径和实施方式上仍存在差异。

行业层面，企业开始围绕身份认证、权限管理、行为监控、输入输出控制等关键能力构建Agent治理框架，推动治理延伸至Agent执行的全过程。

图30：Agent风险威胁来源及治理体系



Part 04

新实践

AI原生基础设施典型案例

注：为贴近产业理念，本章节的词元统一称为Token，智能体称为Agent



4.1 Token需求分化为生产、调用与价值转化三类

在AI能力落地过程中，客户需求正沿Token价值链分化为生产、调用与转化三类。

Token生产需求：客户以大模型企业为代表，核心诉求是获得高性能、稳定、可扩展的算力资源，支撑模型开发，从而服务Token生产。

Token调用需求：客户以AI应用企业、传统企业为代表，核心诉求是通过外部SOTA模型获

得可用的Token生产能力，避免负担过重的基础设施建设，驱动原有业务升级与创新。

Token价值转化需求：客户以场景型企业客户为代表，核心诉求是将Agent应用嵌入具体业务流程，将Token消费转化为流程优化、效率提升等结果效益。

图31：客户Token需求的三层分化



4.2 Token生产需求典型案例

蓝耘科技×某头部AI大模型公司A：Token生产扩张驱动算力基础设施升级

公司A是一家专注通用大模型研发的AI企业，其产品已覆盖办公协作、内容生成、搜索问答等多个应用场景，并通过持续模型迭代与API能力输出，形成模型能力开放体系。

在Token产业视角下，公司A属于“自主构建大模型能力并持续生成Token”的典型代表，在大模型规模化演进过程中，算力成本、资源调度与系统稳定性，正在成为影响Token生产能力持续扩张的关键约束。

• 挑战一：算力资源供给不足与成本持续上升

随着模型参数规模扩大及多模态能力增强，训练阶段对高性能GPU算力依赖显著提升，算力规模急需扩容。同时，模型迭代速度加快进一步放大算力消耗，使单位Token生产成本持续上升。

应对方法：蓝耘科技GPU算力云服务解决方案

一方面，蓝耘GPU算力云服务搭载千卡级英伟达高端GPU计算资源、针对AI大模型训练、推理场景计算网络量身打造的专用InfiniBand网络以及先进的分布式训练集群；

另一方面，该服务支持按需灵活计费、轻松调整存储容量及扩展容量，帮助客户节省自行扩建资源的成本。

• 挑战二：模型训练故障与运维问题

模型版本持续迭代对训练环境一致性与推理服务稳定性提出更高要求。在大规模训练任务与高频算力调度场景下，GPU故障、物理服务器异常以及任务中断等问题，容易导致训练流程中断与资源浪费，进一步提升了基础设施容错能力与自动化恢复机制的要求。

应对方法：蓝耘科技算力调度解决方案

蓝耘自研的基于Kubernetes平台二次开发的算力调度平台，针对AI大模型训练过程中驱动、GPU硬件、物理服务器硬件等故障导致的训练停止痛点，可释放无人工干预即可自动拉起失败任务的“自愈”功能，降低了大模型训练用户的运维难度。

蓝耘科技×某头部AI大模型公司B：蓝耘智算云平台，破解模型扩张下算力、成本与资源调度挑战

公司B是一家致力于打造认知智能大模型的科技企业，依托清华大学计算机系的技术成果转化，目前已研发并构建多语言大模型与高精度的通用知识图谱，持续推进大模型在政务、企业与科研等场景的应用落地。

与公司A类似，公司B也属于“模型能力驱动型Token生产主体”，需要依赖外部的算力基础设施完成模型训练与推理任务，从而持续生成可调用的Token能力。

• 挑战一：模型规模扩张带来的算力供给压力

公司B拥有自己的云计算平台，但随着模型规模的不断扩张，训练参数规模与数据处理量不断提升，已有算力已经无法满足模型训练需求，尤其在高峰训练任务集中时，易出现算力不足与资源排队问题。

• 挑战二：高频模型迭代带来的成本上升压力

模型体系持续快速迭代，在提升模型能力的同时，也带来了更高的训练频率与推理调用成本。自建算力基础设施需要长期投入GPU集群与运维体系，使整体模型训练与推理成本持续攀升，导致单位Token生产成本增加，控制单位Token生产成本成为核心挑战之一。

• 挑战三：多任务并发下的资源使用效率不足

在模型训练、知识增强与推理服务并行运行的场景下，算力资源需在多任务之间动态调度与分配，由于任务类型与负载差异较大，容易出现资源竞争与调度不均，进而导致资源利用率下降，影响整体Token生产效率与系统稳定性。

应对方法：蓝耘科技智算云平台解决方案

蓝耘智算云平台方案为用户提供可直接使用的高性能计算服务，实现计算资源统一分配调度、训练数据集中管理、算力资源的高效协同与稳定供给。

依托灵活配置能力，用户可按需选择GPU高端显卡资源，进行多任务并行的大模型训练，有效进行算力扩容，缩短了训练周期，节省了自行扩建资源的成本，并且避免资源限制与不合理配置带来的浪费问题。

灵活配置算力资源

可根据业务需求自主选择、灵活搭建GPU卡数、CPU核数、带宽以及操作系统类型，满足多应用下的算力扩容或效率需求。

专业服务节省运维成本

由专家团队在线服务并提供保障，基于英伟达大规模AI集群实施经验与大模型实施训练部署调优经验，节省运维成本。

4.3 Token调用需求典型案例

蓝耘科技×某AI+教育公司：调用外部Token生产能力驱动教育服务创新

该公司是一家专注于AI+教育的高科技企业。围绕论文评阅、作业批改等场景构建产品体系，核心产品为“AI智评平台”，为高校与个人提供智能学习、智能评阅与学习管理服务。

在Token调用视角下，该公司属于“模型能力消费型企业”，并不直接参与Token生产，而是通过接入领先大模型能力，并将模型能力转化为稳定、可调用的智能能力，从而驱动教育场景中的业务创新。

• 挑战一：模型能力获取门槛高

公司围绕论文评阅、作业批改等教育场景构建AI产品，需要具备文本理解、内容生成与逻辑分析等复杂模型能力。然而，自建大模型不仅需要算力资源投入，还需要算法、工程与运维团队持续支撑，技术门槛与投入成本较高。

• 挑战二：模型能力持续演进难以跟进

目前市场上的大模型能力迭代速度极快，企业若自行维护模型体系，需要持续投入资源进行模型选型、能力评测与版本升级，难以及时跟进领先模型能力的发展节奏。

• 挑战三：模型服务稳定性与弹性供给难保障

论文评阅、学习分析等教育场景对服务连续性与结果可靠性要求较高。企业若自行部署模型，不仅需要持续投入运维资源保障服务稳定运行，还需要处理模型部署、推理优化、故障恢复等复杂问题，增加整体技术负担。

应对方法：蓝耘科技MaaS平台解决方案

蓝耘元生代MaaS平台提供标准化大模型服务能力，通过统一模型接口、持续模型供给与弹性资源调度体系，帮助企业将更多资源聚焦于教育场景创新与产品优化，实现从“建设模型能力”向“调用模型能力”的转变。

快速获取SOTA模型能力

平台持续接入主流SOTA模型，并通过统一API开放服务，帮助该公司快速获得论文分析、内容生成、逻辑检测等教育场景所需的大模型能力，降低接入门槛与开发成本。

弹性算力供给提升模型服务可靠性

依托智算云资源池与弹性调度能力，平台能够为模型服务提供稳定可靠的算力支撑。面对模型版本升级、业务规模增长及多场景调用需求，持续保障模型服务的可用性、稳定性与响应能力。

蓝耘科技×某数字人力公司：模型能力驱动“数字人力官”智能化建设

该公司是一家聚焦企业用工管理、灵活用工与数字履约服务的科技企业，围绕企业组织管理与人才服务场景构建数字化服务体系。随着企业客户对用工咨询、岗位分析与方案规划等智能化服务需求增长，该公司推出“数字人力官”Agent，通过引入大模型能力提升服务效率与服务质量。

该公司同样作为“模型能力消费型企业”，通过调用外部Token生产能力，为Agent产品提供底层能力支撑，并将模型能力转化为用工咨询、岗位分析、方案规划等业务服务能力。

• 挑战一：模型能力接入与系统集成复杂

面向人力资源服务场景，该公司需要将AI能力与现有业务系统、服务流程及客户触点深度融合。但若采用自建模式，需要解决模型调用、接口开发与系统集成等问题，增加整体建设复杂度。

• 挑战二：底层模型能力建设成本高

“数字人力官”等智能产品需要具备知识问答、方案生成、岗位分析等能力。企业若自建大模型体系，需要投入大量的算力资源与人力资源，整体投入成本较高。

• 挑战三：行业能力适配与持续优化难

“数字人力官”不仅需要接入领先的通用大模型能力，还需要结合用工政策、岗位结构、组织管理等专业知识进行持续优化。企业若自建相关能力，需要长期投入资源开展模型选型、知识补充与场景适配，建设与维护难度大。

应对方法：蓝耘科技MaaS平台解决方案

平台化调用，灵活集成

蓝耘元生代MaaS平台提供标准化API接口与模型调用能力，支持与公司现有业务系统、服务流程及客户触点快速集成，将智能问答、方案生成等能力融入“数字人力官”产品。

降低模型建设门槛

平台持续提供领先大模型服务能力，公司无需投入大量资源自建模型与算力基础设施，即可快速获得领先的智能能力。

支持行业能力持续优化

蓝耘持续提供模型能力升级与技术支持，帮助公司将人力资源领域知识沉淀到Agent产品，不断提升其在人力资源业务领域的专业性与服务质量。

4.4 Token价值转化需求典型案例

蓝耘科技×某数字人力公司：Agent驱动Token价值转化，重塑企业用工服务模式

该公司与4.3节所述数字人力官案例为同一家公司。在完成模型能力接入后，该公司基于大模型能力开发“数字人力官”Agent。此时重点已从Token调用转向Token价值转化，即通过Agent将模型能力融入企业用工管理场景，实现服务效率与服务质量提升。

- **业务痛点一：传统服务方式高度依赖人工经验，专业性与规模化响应难以同时保证**

在人力资源服务场景中，企业客户需求差异大，涉及用工咨询、岗位规划、政策匹配等多类业务内容。传统服务模式高度依赖人工经验，方案梳理与服务响应需要大量人工参与，难以同时保证服务的专业性与规模化响应。

- **业务痛点二：人力资源场景Agent建设与业务适配复杂度高**

面向企业客户的人力资源场景Agent不仅需要具备自然交互能力，还需要理解用工政策、岗位结构与企业需求，并与现有业务系统和服务流程深度融合。同时，还需要确保输出内容符合人力资源服务场景的专业表达与业务边界，整体建设与落地难度较高。

应对方法：蓝耘科技“数字人力官”Agent建设方案

蓝耘科技基于大模型API能力与Agent搭建经验，协助该公司构建面向企

业用工场景的“数字人力官”Agent，将通用模型能力与人力资源领域知识、服务规则和业务经验相结合，使其具备需求理解、信息分析与服务建议生成能力。

人力资源Agent能力快速落地

基于前期完成的大模型能力接入与调用体系建设，公司能够快速构建“数字人力官”Agent，并接入现有业务平台，实现智能服务能力从验证到上线的快速落地。

Agent能力与人力资源场景融合

蓝耘科技协助公司完成Agent交互逻辑与业务流程编排设计，使“数字人力官”能够嵌入现有服务体系，围绕企业输入的信息自动完成需求识别、问题拆解与服务建议输出。

蓝耘星河*×某精品连锁民宿运营企业：营销获客Agent驱动文旅运营增长

该企业是一家专注于精品民宿连锁运营、民宿托管及文旅配套服务的综合性企业，在多个热门景区周边布局精品民宿门店，覆盖主题客房、亲子套房、团建院落等多种业态，持续为游客提供沉浸式度假住宿体验。

在Token转化视角下，该企业属于“Agent能力消费企业”。作为旅游生活服务运营主体，随着线上流量成为游客获取住宿信息的重要入口，该企业开始探索营销获客Agent建设，通过Agent将Token能力转化为内容生产、流量获取与客户运营能力，推动线上曝光向订单转化和业务持续增长。

• 业务痛点一：线上内容生产与流量获取能力不足

在文旅消费线上化趋势下，内容平台已成为民宿获客的重要入口。

该企业缺乏专业内容创作与营销运营能力，难以将民宿特色、景区资源及服务体验持续转化为适配不同平台的种草内容、短视频作品与宣传物料，线上曝光与流量获取能力不足，制约品牌影响力与新客增长。

• 业务痛点二：营销运营体系智能化程度不足，制约经营增长能力提升

企业在营销增长过程中缺乏智能化运营体系支撑，难以围绕客户留存、账号增长、品牌推广与经营分析形成持续优化机制。

同时，获客渠道高度依赖OTA平台，自有流量资产沉淀不足，难以通过数据洞察和运营策略优化构建稳定的新客获取与复购增长能力，整体经营增长面临较大压力。

应对方法：蓝耘星河营销获客Agent解决方案

蓝耘星河针对民宿行业营销获客场景构建营销获客Agent解决方案。基于大模型能力、行业知识与业务规则，Agent能够自主完成内容生成、营销辅助、运营分析与增长决策等任务，帮助企业构建“内容生产—流量获取—客户转化—运营优化—收益增长”的智能运营闭环。

内容营销能力

围绕民宿特色、景区资源与目标客群需求，Agent自动完成图文种草、爆款视频生成与复刻、标准化营销海报生成，快速提升线上内容曝光与流量获取能力，统一品牌宣传形象。

运营增长能力

结合经营数据、渠道表现、品牌曝光与GEO监测结果，辅助企业开展运营分析、营销规划与增长决策，持续优化客户留存、品牌影响力与经营收益表现。

结语

CONCLUSION

我们正站在智能时代基础设施体系成型的起点。

未来，Token将成为贯穿智能活动运行的核心资源；Agent将从辅助工具走向业务执行主体，进入研发、生产、营销、服务和经营决策等更多场景。以Token为载体的智能服务，也将由少数机构掌握的稀缺能力，逐步转化为企业和个人均可调用的新型公共能力。由此开启的，不只是一次技术升级，更是一场生产力基础、组织形态与产业秩序的系统变革。

这一进程不会自然发生。Token产能能否持续扩大，异构算力能否高效协同，Agent能否稳定进入业务流程，成本、质量、安全与价值能否得到有效治理，都将决定智能能力规模化落地的速度与边界。AI产业需要建设开放协同、稳定运行并能持续进化的基础设施体系，打通算力、模型与应用之间的协同链路，提升Token生产、调度与应用的整体效率。

面向这一未来，国内已具备形成自主AI基础设施体系的重要条件：丰富的产业场景、持续完善的国产算力生态、广泛分布的智算资源，以及庞大的企业智能化需求，为技术迭代和规模化应用提供了独特土壤。下一阶段，需要产业各方进一步打破芯片、算力、模型与应用之间

的边界，推动多元资源高效组织、Token产能灵活供给与调用，并使Token创造的业务价值能够被衡量。

这既需要持续突破关键软硬件技术，也需要形成专业化的产业分工、可持续的商业模式和与Agent时代相适应的治理规则。只有让资源供给者、基础设施服务商、模型企业、应用开发者与行业客户共同参与，才能把局部技术优势转化为系统能力，把单点创新转化为产业进步。

AI原生基础设施不会是一成不变的终局形态。它将在真实任务中被验证，在产业协作中被完善，在持续运行中不断进化。但它所指向的方向已经逐渐清晰：让每一份算力转化为有效Token，让每一次Token调用产生更可靠的结果，让Token消费创造出更多的实际业务价值。

甲子光年智库

法律声明

版权声明

本报告由甲子光年旗下的甲子光年智库制作完成，报告内容的版权及相关知识产权均归北京甲子光年科技服务有限公司所有。任何单位或个人在引用本报告内容时，须保持内容的原始性，不得进行歪曲、删改或误导性引用，并须注明报告出处为“甲子光年智库”。否则，由此引发的一切后果由引用方自行承担，甲子光年保留追责权利。

免责条款

本报告中的行业观点、行业数据、市场预测和相关分析主要来源于甲子光年智库研究团队通过桌面研究、专家访谈、问卷调查、公开数据整理及甲子光年产品数据等方式获得。我们已尽合理努力确保观点与数据的准确性、完整性与可靠性，但甲子光年不对其作出任何明示或暗示的保证。在任何情况下，本报告中包含的内容、观点或数据均不构成对任何单位或个人的投资建议、法律建议或其他形式的专业意见，相关决策应由读者自行判断并承担风险。甲子光年对因使用本报告内容而产生的任何直接或间接损失不承担任何法律责任。

THANKS

谢谢观看

北京甲子光年科技服务有限公司是一家科技智库，包含智库、媒体、社群、企业服务版块，立足于中国科技创新前沿阵地，动态跟踪头部科技企业发展和传统产业技术升级案例，致力于推动人工智能、大数据、物联网、云计算、AR/VR交互技术、信息安全、金融科技、大健康等科技创新在产业之中的应用与落地



关注甲子光年公众号



扫码关注甲子光年智库公众号

智库管理分析师

黄琪丹
13520577912 (微信同号)

智库院长

宋涛微信
stgg_6406

商业合作负责人

郑爽
18600502376 (微信同号)