



Bruce

Founding Organizer, OpenClaw Asia

Former **Microsoft** and **Alibaba**. M.Sc. in Computer Science, Simon Fraser University.

Early OpenClaw user. Organizer of Asia's first OpenClaw builder meetup in Hong Kong. Now building the OpenClaw Asia community across **Hong Kong, Shenzhen, Beijing, Shanghai, and Tokyo**.



OpenClaw Asia — Build AI Agents Across Asia

OpenClaw Asia connects builders, creators, and users across Asia.



Events

We organize AI events, workshops, and demo days across Asia.



Builders Network

Connecting developers, founders, and AI builders.



Real Use Cases

Sharing real use cases, not just ideas.

Cities

BEIJING

SHANGHAI

SHENZHEN

HONG KONG

MACAU

JAPAN

VIETNAM

THAILAND

Whether you're a developer, creator, or AI explorer, you're welcome to join. Let's build the future of AI together.

openclawasia.ai

Inference Inflection Arrives

OpenClaw与 Token工厂

开源Agent运动如何重塑AI价值分配结构



GitHub Activity





OpenClaw HK

Feb 13 | Hong Kong 2月13日 | 香港



Feb 13 | Hong Kong 2月13日 | 香港



OpenClaw Builders Meetup | 香港开发者聚会

Feb 13 | Hong Kong 2月13日 | 香港







ry. So I kind of with the uh whole high cycle and
是今年一月开始的。所以我有点像整个

s China: From Core Code to Real Workflows



Vincent Koc

OpenClaw 核心贡献者
Core Contributor to OpenClaw



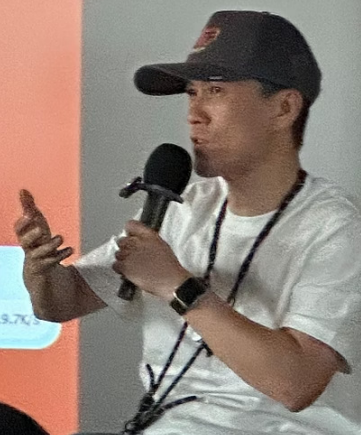
王盟
Wang Meng

ArkClaw 研发架构师
ArkClaw R&D Architect at
Volcano Engine



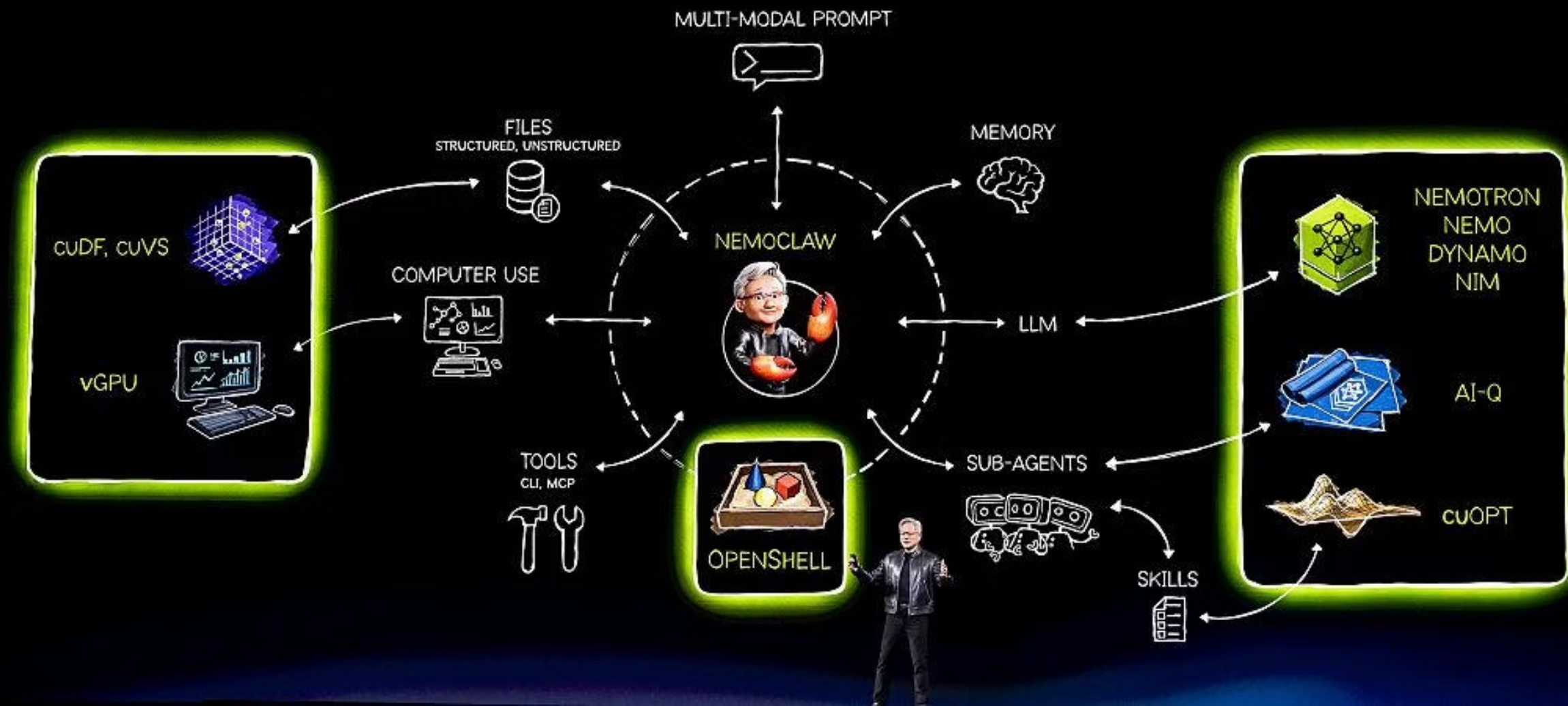
Bruce Huang

opencldwasia



Announcing NVIDIA NemoClaw Reference OpenClaw

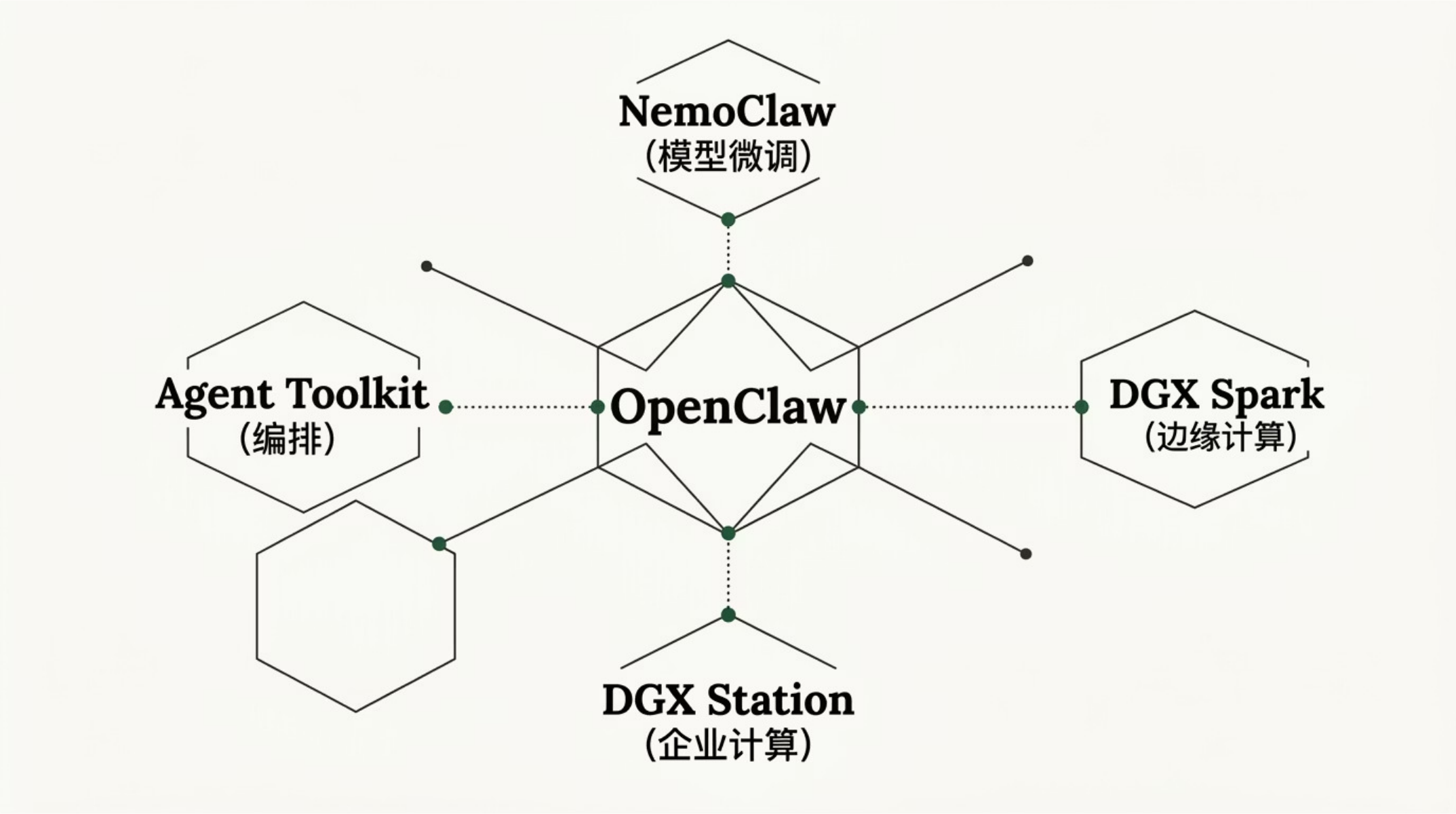
NVIDIA Agent Toolkit for Building Specialized Agents



GTC 完成了平台级提级 - OpenClaw的”成人礼“

OPENCLAW 已进基础设施层

一次开发者大会完成的不只是产品发布，而是一次身份跃迁——OpenClaw 从社区现象跃升为平台入口，再固化为基础设施。平台开始围绕它重组。



5个

配套基础层

NemoClaw、OpenShell、DGX Spark、DGX Station、Agent Toolkit 同步就位

1次

身份跃迁

社区项目 → 平台入口 → 基础设施，三级跃迁在单次 GTC 完成定调

“回答问题” => “调用工具并执行任务”

市场误判了 AI 需求口径

旧范式：Chat 问答

用户主动发起 → 单轮响应 → 对话结束
需求以「提问频次」计量
消耗随会话终止而归零

Chat 量尺

已失效

旧误判

1 个核心

3类

高密商用场景

代码生成、流程自动化、数据分析——均已进入后台常驻形态

新范式：后台任务循环

任务常驻后台 → 持续执行 → 自循环消耗
需求以「任务密度 × 时长」计量
消耗不随人离线而停止

→ Chat 提问

→ 任务执行

→ 持续消耗

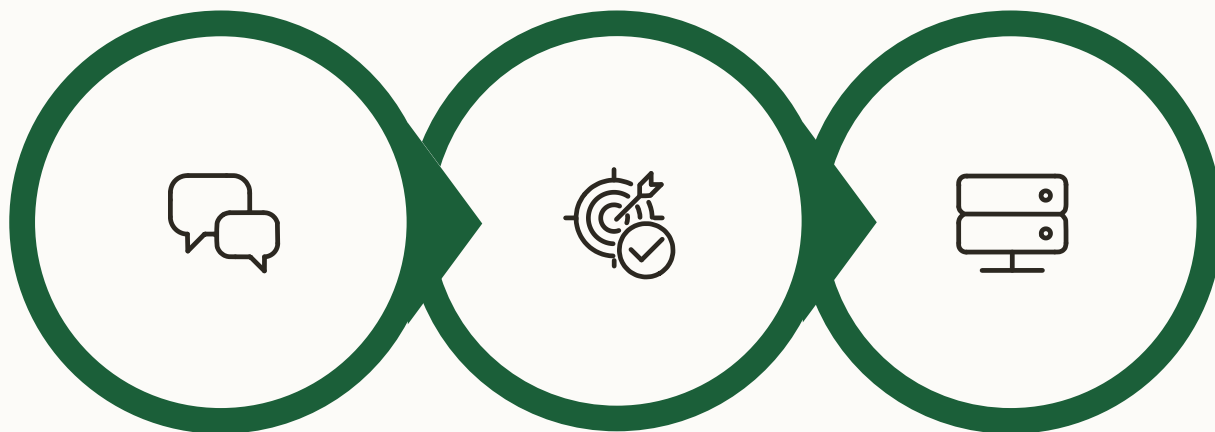
1个

旧误判待修正

以 Chat 会话量衡量 AI 需求的框架，已无法反映真实消耗规模

Agent 正在成为常驻进程

软件结构将随之改写



Chat响应

Goal设定

常驻进程

Agent 将成默认工作负载

平台设计假设已切换

旧平台架构假设

以 Chat 为主负载

单轮请求 / 短会话 / 人机交替

同步响应模型

用户等待 → 返回结果 → 结束

单点优化

延迟、质量、成本三选二

4层

同步架构升级

推理调度 / 内存管理 / 权限控制 / 计费模型——上游整栈跟改

新平台架构假设

以 Agent 为主负载

长任务 / 多步推理 / 异步并发

异步执行模型

任务派发 → 后台运行 → 结果回流

全栈重写

调度、内存、安全、计费同步改

1次

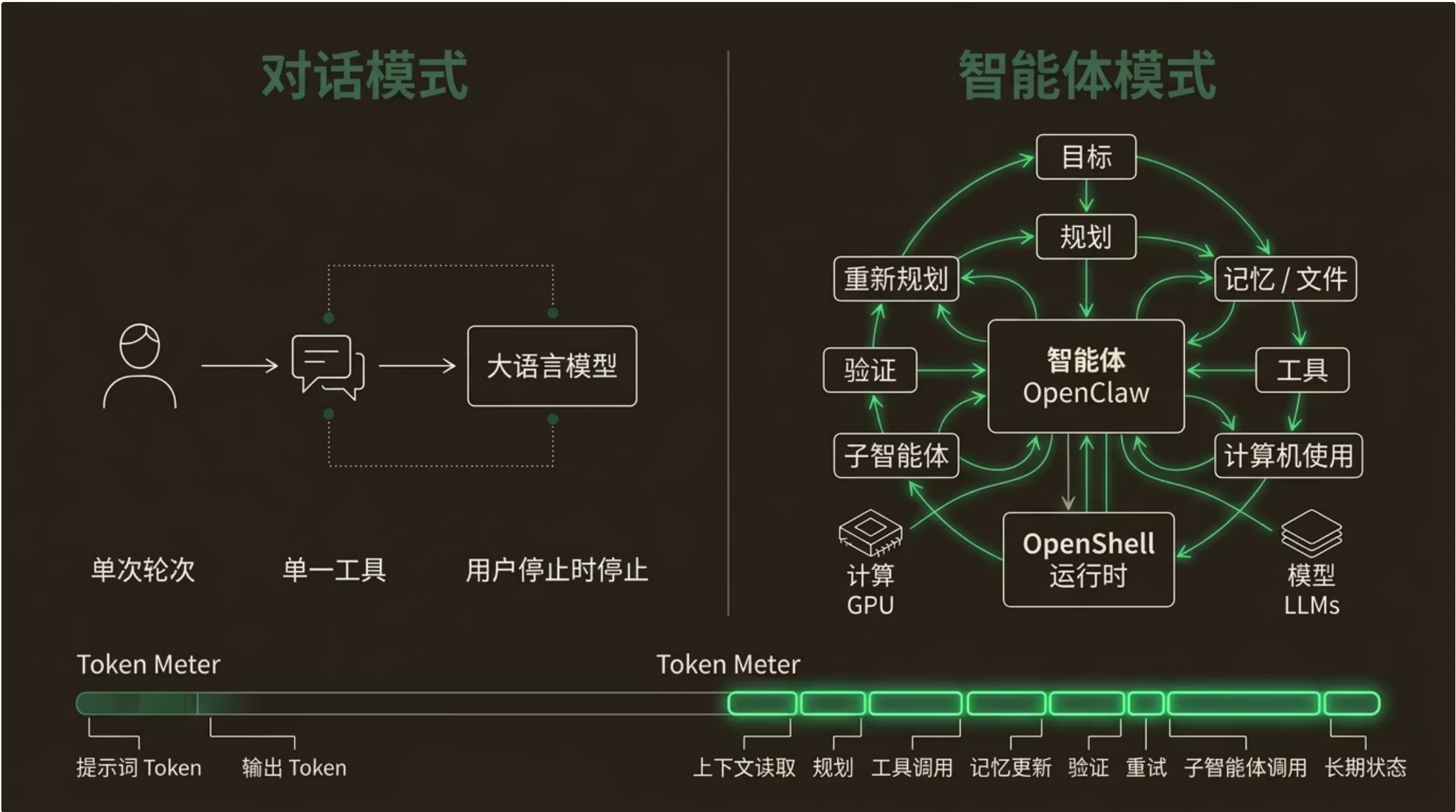
平台切换

渗透仍在前中期，先行布局者将享受结构红利，窗口期有限

Token爆炸的根因：任务从问答转向持续执行

CHAT模式 VS AGENT模式

Agent模式的核心特征——目标驱动、自主循环、持续在线——将Token消耗从线性问答转变为指数级增长。



Chat 模式: 单次成本

Chat模型按问题消耗Token，用户停止即停止。

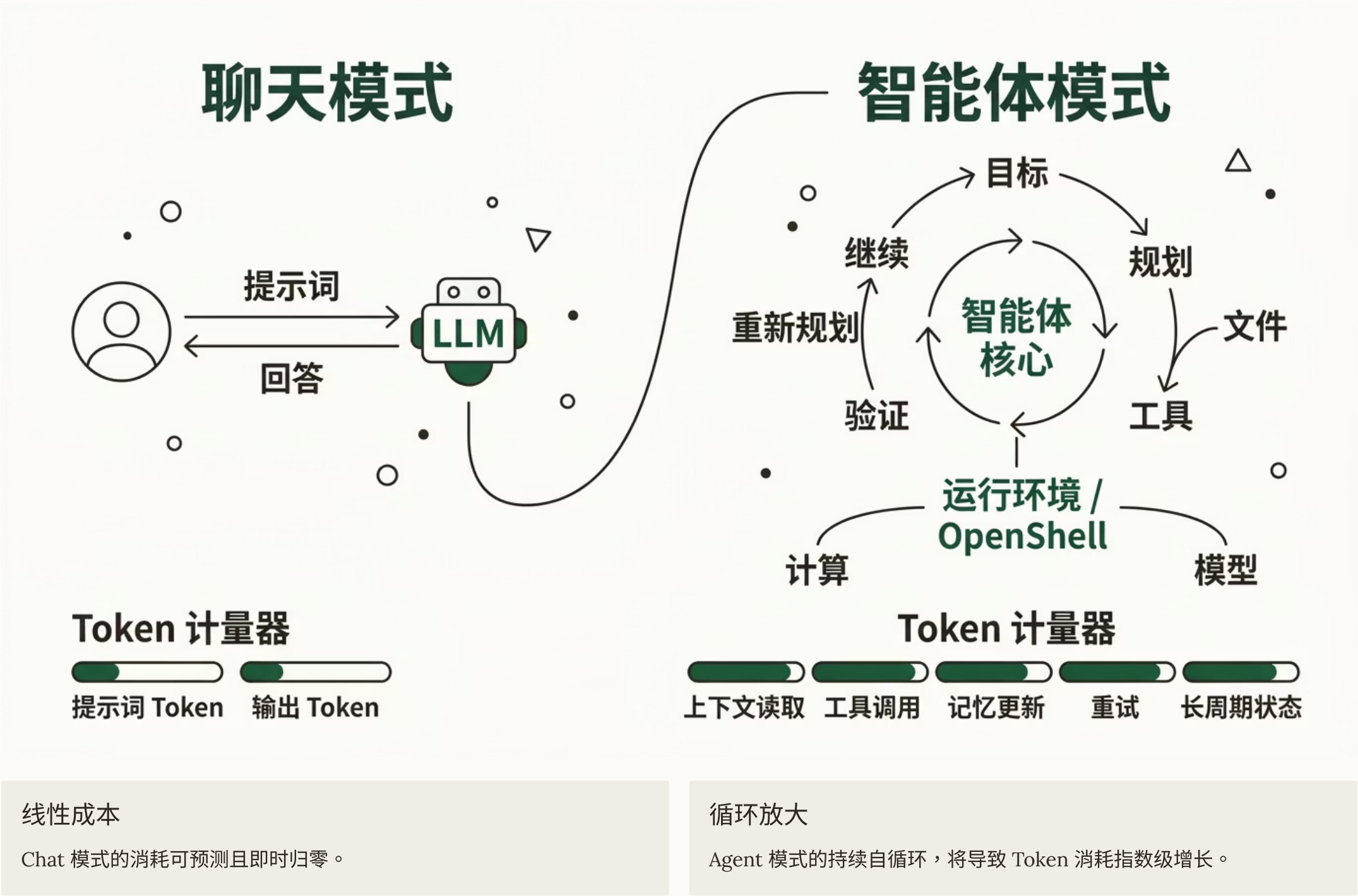
Agent 模式: 持续成本

Agent模型按 workflow 消耗Token，任务未完成则持续消耗。

推理 Token 将黑洞式增长

TOKEN 成本函数已被改写

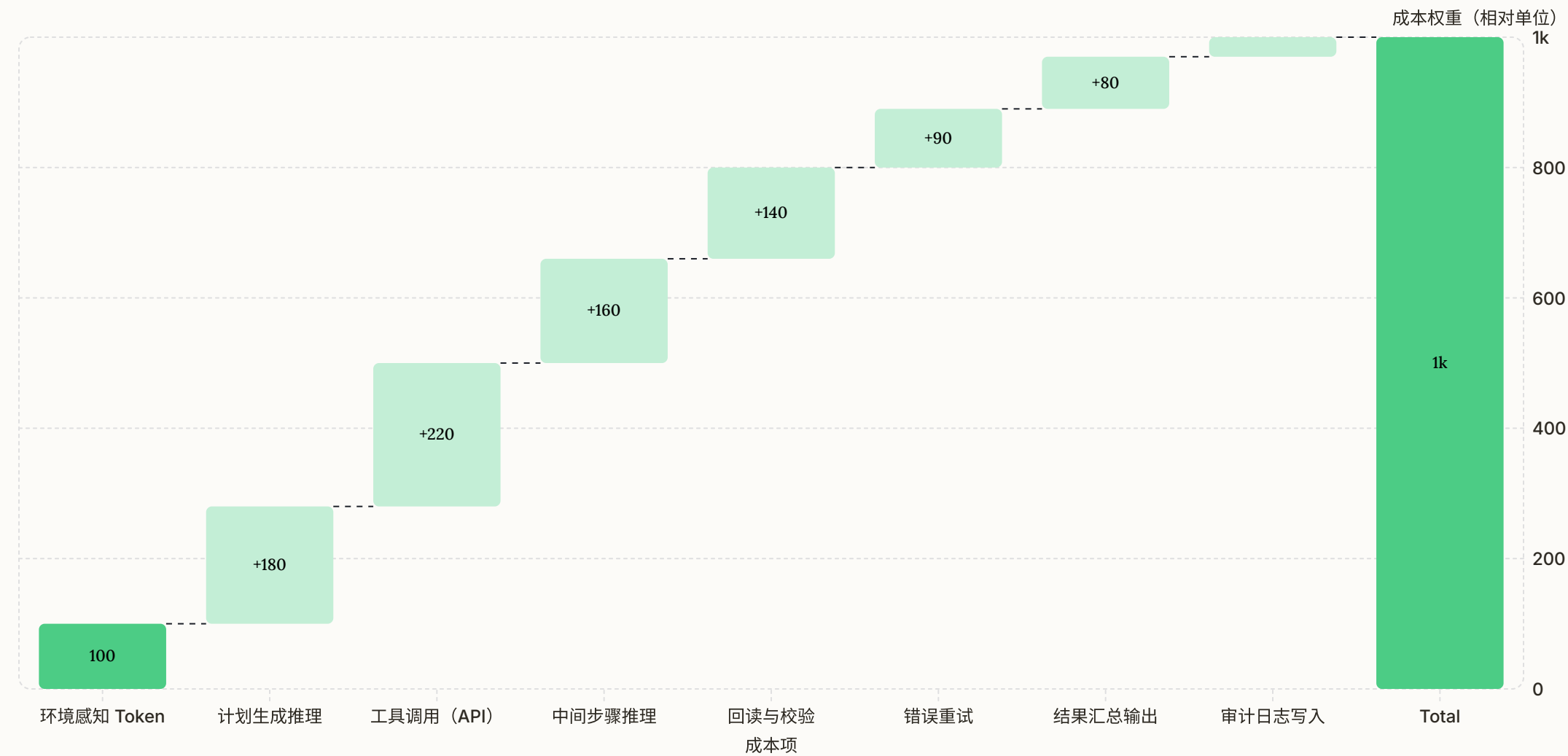
AI 需求的范式转变，将 Token 消耗模式从单次线性推向持续循环放大，导致推理成本呈黑洞式增长。



Token 成本函数已被改写

单轮问答模型已失真

Agent 执行一个任务，调用链路全程计费。**执行链全计费、工具调用烧钱、总成本才是核心**——旧的「每次问答成本」已无法反映真实经济结构。



8类

成本项全链路

每一个执行步骤均产生 Token 消耗，总成本是各项叠加，非单次推理

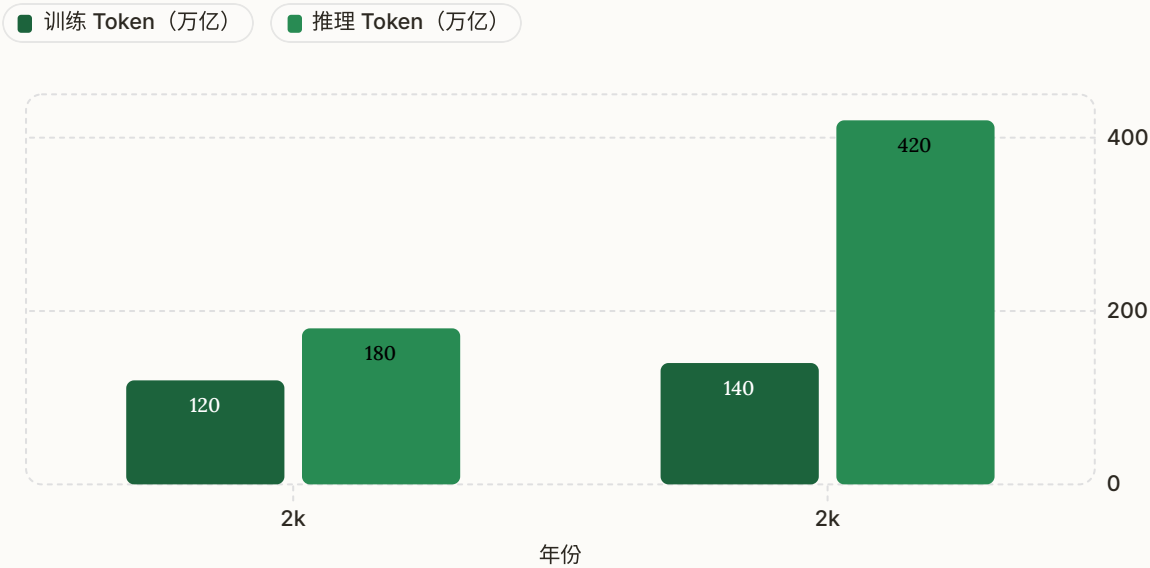
1条

闭环计费链路

环境读取 → 工具调用 → 回读修正，三段构成最小闭环，缺一不可

推理 Token 将黑洞式增长

INFERENCE 正吞下主战场



需求曲线已改道

- 人停手已失效
Agent 常驻，消耗不随人类离线归零
- 后台持续消耗
并发 Agent 数以万计，同时在线运行
- 需求曲线改道
训练 Token 增速平缓，推理 Token 垂直拉升

1万亿

2027 前推理规模

保守估算，届时全球推理 Token 消耗将突破万亿级年度规模

万级

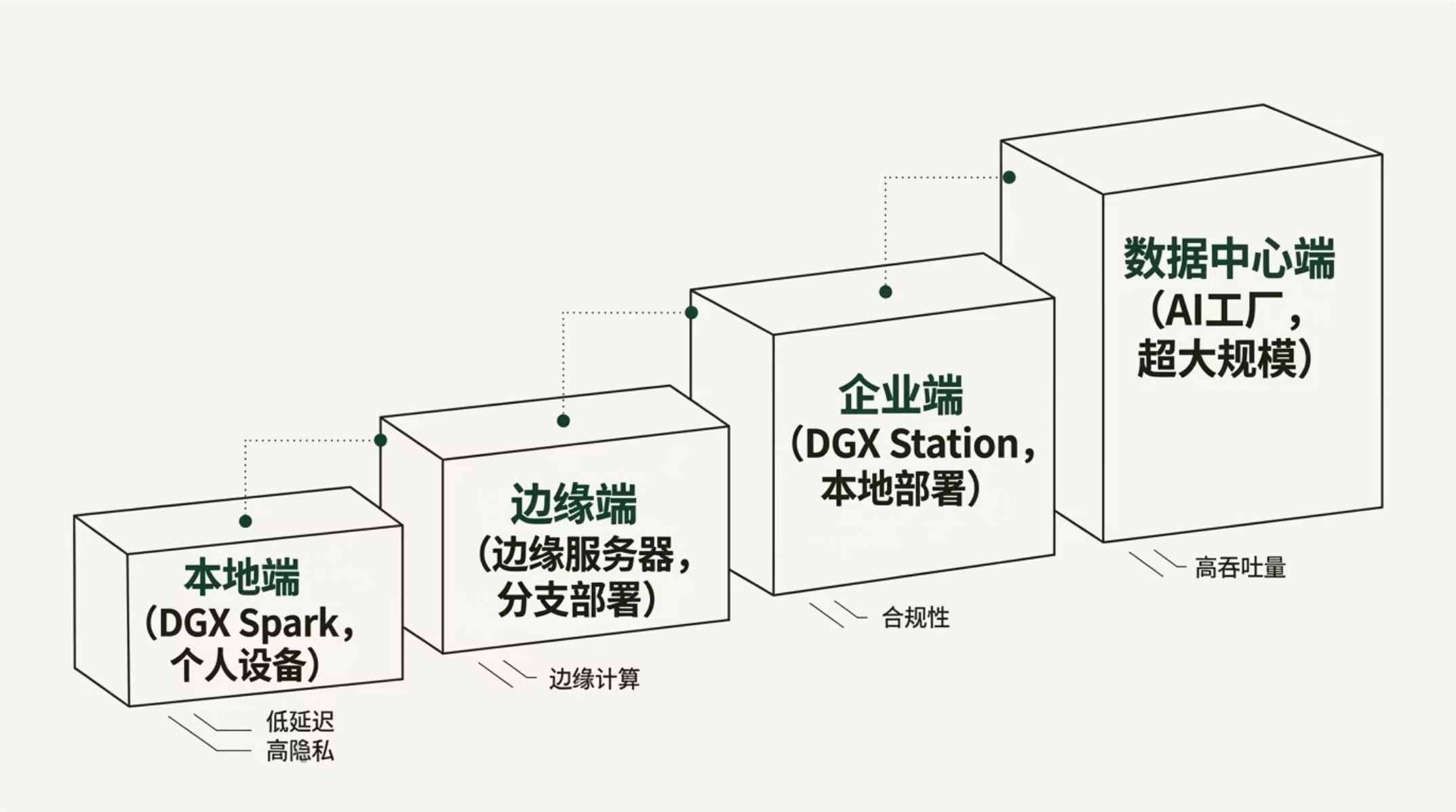
并发请求量级

常驻 Agent 同时在线并发请求，数以万计，实时争抢算力资源

Compute 已成 Agent 供给梯子

供给侧开始多层分布

算力供给不再只有云端一条路。本地端、边缘端、企业端、数据中心端**四层同步被激活**，Agent 可以按任务优先级和成本约束自动选择执行层。



本地端被激活

DGX Spark 落地个人与小团队，低延迟高隐私场景优先

企业端被激活

DGX Station 满足合规要求，私有化部署成企业采购主选

机房端被激活

AI Factory 级数据中心承接高并发推理，Token 产出工厂化

4层

供给面全覆盖

从 DGX Spark 到 AI Factory，供给梯度覆盖个人到超大规模

1套

统一调度逻辑

Agent Toolkit 跨层调度，任务密度与成本约束决定落点

NVIDIA AI Infrastructure for Enterprise Computing

DGX
Spark
1 PFLOPS



RTX PRO
Workstation

DGX
Station
20 PFLOPS

RTX PRO
Enterprise Server

DGX B200

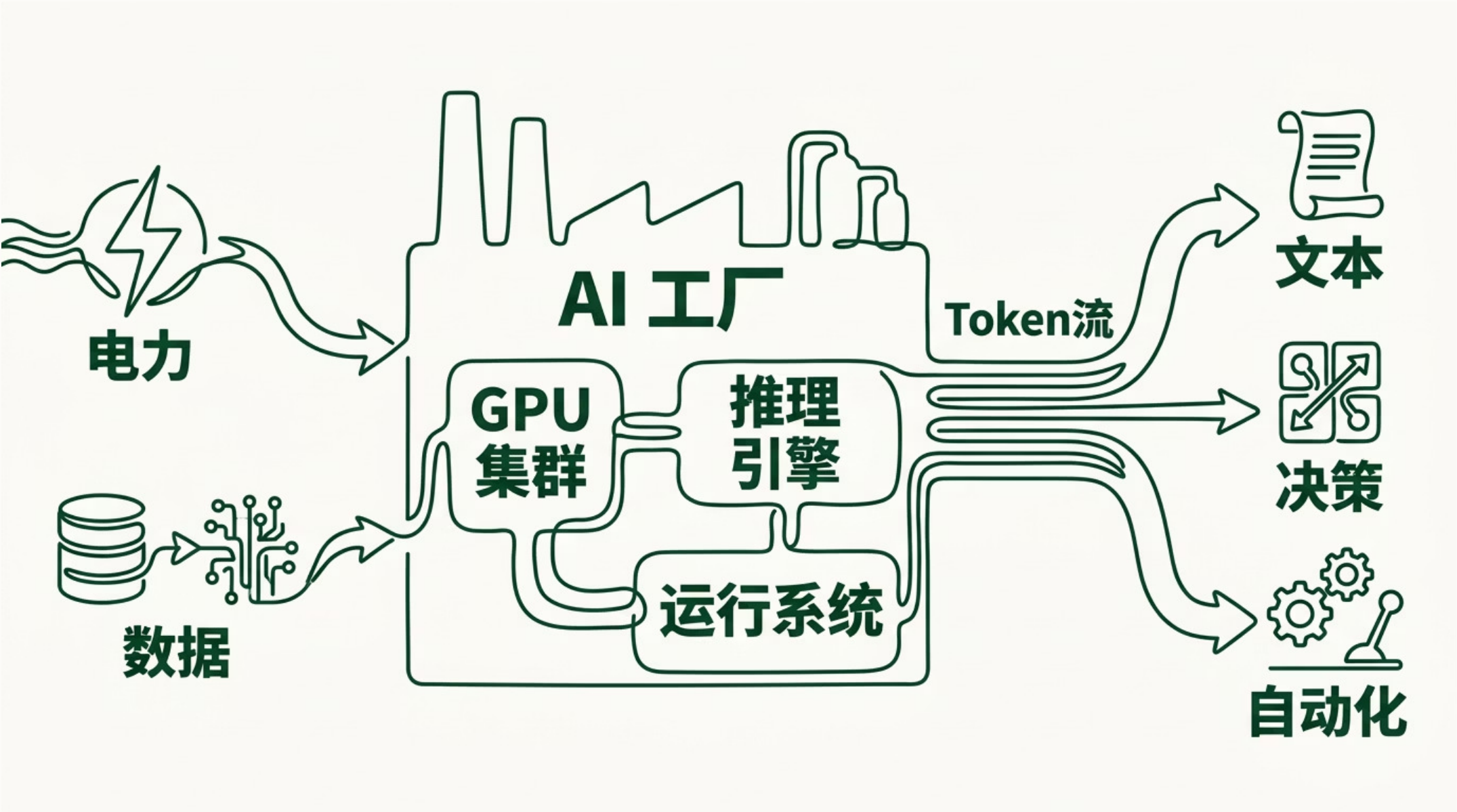
DGX GB300
15 EFLOPS



AI 已从模型变成工厂

TOKEN 是 AI 时代的生产单位

AI 时代的生产模式已发生本质性变革。输入不再是单一的 Prompt，而是 **数据与电力** 等基础资源；输出则是源源不断的 **Token 流**，驱动各类智能工作负载。



万级

Agent 推理请求

AI Agent 全天候运行，产生巨量的推理请求，持续消耗 Token。

1万亿+

AI 基础设施投资

全球 AI 基础设施投资规模已突破万亿级，支撑 Token 工厂的运作。

☐ 能源 → 算力 → Token → 工作流：这是 AI 时代新的生产关系，每个环节都至关重要。

数据中心已升级成 Token 工厂

物理约束成稀缺底座

工厂类比成立

电力 = 原料

GW 级电力直接决定 Token 产能上限

散热 = 瓶颈

液冷基础设施已成卡脖子变量

吞吐 = 产量

Token/s 是工厂产出的核心计量单位

AI 工厂三要素



1GW+

单体数据中心电力

头部 AI Factory 单体电力消耗已进入 GW 量级，物理稀缺性显现

400亿

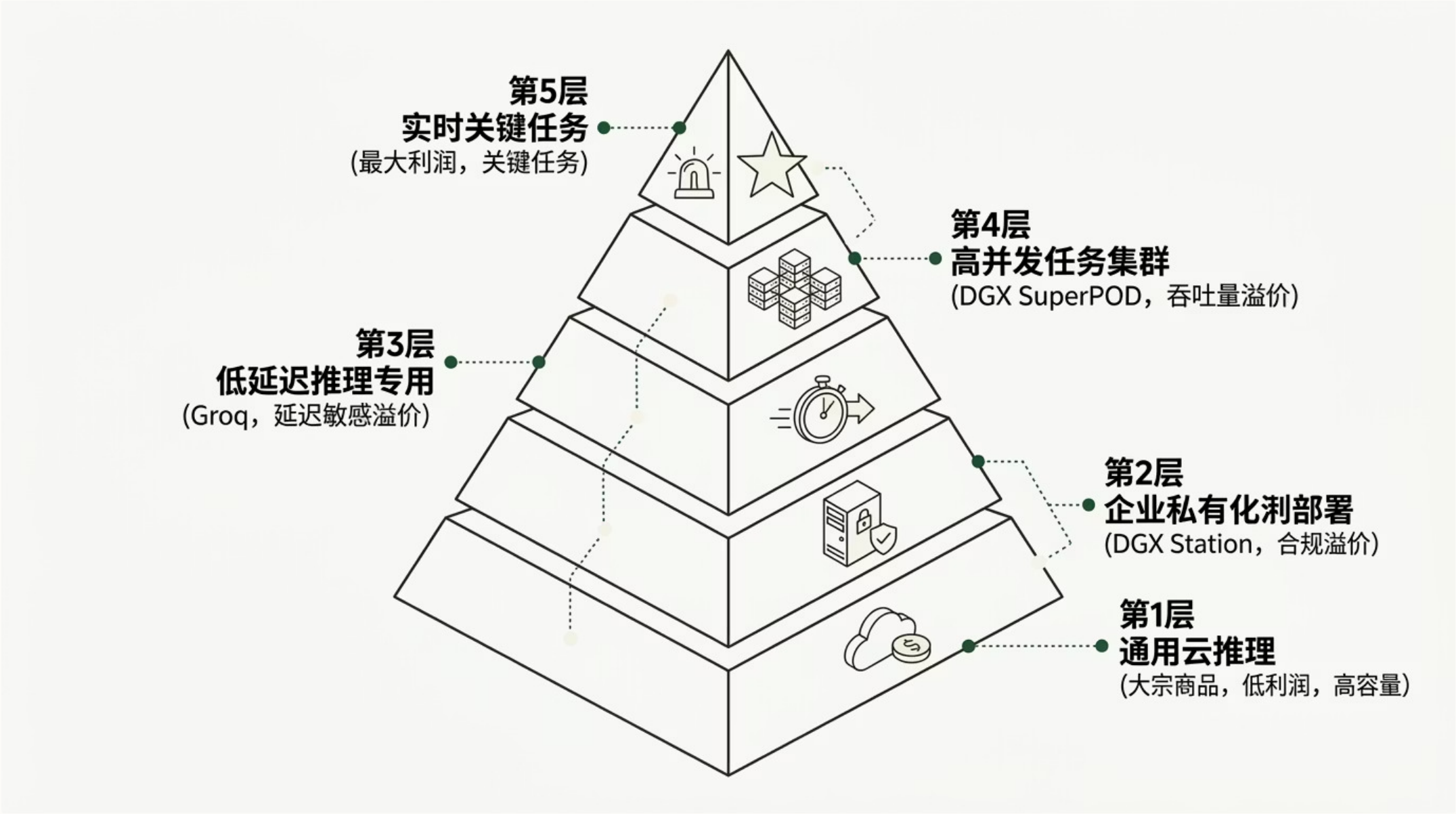
15年投资规模（美元）

单座大型 AI 数据中心全生命周期资本支出，供给侧护城河极深

Token 分层定价重写硬件利润

硬件开始卖价值梯度

Token 开始分层、架构跟随价值、利润表被重写——不同任务密度匹配不同硬件层，高价值任务支撑高硬件溢价，定价逻辑从「算力规格」切换为「Token 单价 × 任务密度」。



5层

Token 定价梯度

从通用云端到关键任务专用，溢价倍数随任务密度与时延要求递增

DGX→Groq

硬件利润重分配

延迟敏感型与高并发型需求将重新定义哪类硬件享有最高毛利

Alibaba Token Hub

各位阿里人：

当下正处于AGI 爆发前夜。大量数字化工作将由数以百亿计的 AI Agent来支撑，而这些AI Agent将由模型产生的Token支撑运行，成为人类与数字世界交互的主要载体。

面向这一历史性机遇，集团决定正式成立Alibaba Token Hub（ATH）事业群，建立以“创造 Token、输送 Token、应用Token”为核心目标的新组织。

Alibaba Token Hub事业群由我直接负责，强化各AI业务战略协同，以AI重塑工作方式，保持敏捷组织。具体业务包括：

- 通义实验室：创造领先的多模态模型，不断追求基础模型能力上限，为集团和业界提供最领先模型。
- MaaS业务线：构建高效开放的模型服务平台和技术体系，支撑全行业AI生态。
- 千问事业部：打造最好的个人AI助手。
- 悟空事业部：打造B端AI原生工作平台，将模型能力深度融入企业工作流。
- AI创新事业部：探索各类AI创新应用，快速验证新模式、新市场。

以上调整即日生效。一起全力以赴！

吴泳铭

2026年3月16日

高层变动同步出现



ATH 证明中国按 Token 重组

中国变量落在组织速度

ATH 是结构动作，不是产品发布

阿里通义 / 百炼 / 千问悟空的整合，本质是**将 Token 确立为经营对象**，围绕 Token 的产生、输送与应用重建组织结构。这是中国头部大厂第一次将 AI 能力链路纳入统一事业群。

→ ATH 是结构动作

组织重排先于产品落地，结构决定速度

→ Token 成经营对象

从「模型能力」切换为「Token 经营」的战略锚点

→ 全链路经营开始

创造 → 输送 → 应用，三段被纳入统一管理视角

中国变量的特殊性

组织速度是中国大厂的核心竞争变量。ATH 的意义不仅是技术整合，更是**以组织重构锁定 Token 经济主导权**的战略卡位动作。

1个

Token Hub

ATH 构建统一 Token 枢纽，覆盖全链路

百亿级

Agent 规模目标

统一补给线支撑数以百亿计 Agent 部署

ATH 把三段压成经营闭环

模型平台应用已被串联

通义负责**创造 Token**，百炼负责**输送 Token**，千问 / 悟空负责**应用 Token**——三段被压入同一事业群，形成 Token 经营的完整闭环，避免能力在组织边界处衰减。

通义：创造
Token

百炼：输送
Token

反馈与改进

千问/悟空：应用
Token



国内“百虾大战”：万物皆可 Claw!



这年头，不搞个“Claw”都不好意思说在做大模型...

微信官方OpenClaw插件 - ClawBot



OpenClaw 产品化竞争的底层逻辑



竞争焦点在封装、入口、生态。

微信接入 OpenClaw 的当前解读

微信是核心变量

入口与关系链

消息流与高频场景

接入仍克制

当前偏战略卡位

微信价值已显现，入口优势待释放。

微信 Agent 生态竞争与入口争夺

阿里悟空 - 偏办公协作

以企业级工作流为核心切入点，深耕办公协作场景，构建专业工具链。

腾讯QClaw - 偏轻入口与消费场景

依托微信生态，覆盖高频消费与社交场景，强化轻量化接入路径。

底层同为 OpenClaw

胜负取决于高频入口

微信具备分发放大效应

微信入口稀缺，腾讯杠杆最大。

后续判断与观察框架

入口决定触达

场景决定频次

体验决定留存

边界决定托付

📌 竞争仍在早期

先跑通入口、场景、体验、边界，才有长期价值。

百 Claw 源自组织工业排产

⚠ 结构推论，非 ATH 原文结论

推论逻辑链

- 组织重排完成
ATH 统一补给线为规模化排产提供基础条件
- 多入口同步铺开
消费端、企业端、开发者端同步部署 Agent 入口
- 多场景密集覆盖
Token sink 密集化，覆盖办公 / 电商 / 金融 / 政务等场景

3类

应用面（推论）

消费者端 / 企业端 / 开发者端——三类应用面同步铺开，Token sink 密集化

百 Claw 矩阵（推论）



1套

统一补给线

ATH 闭环作为统一 Token 补给线，是百 Claw 规模化的组织前提

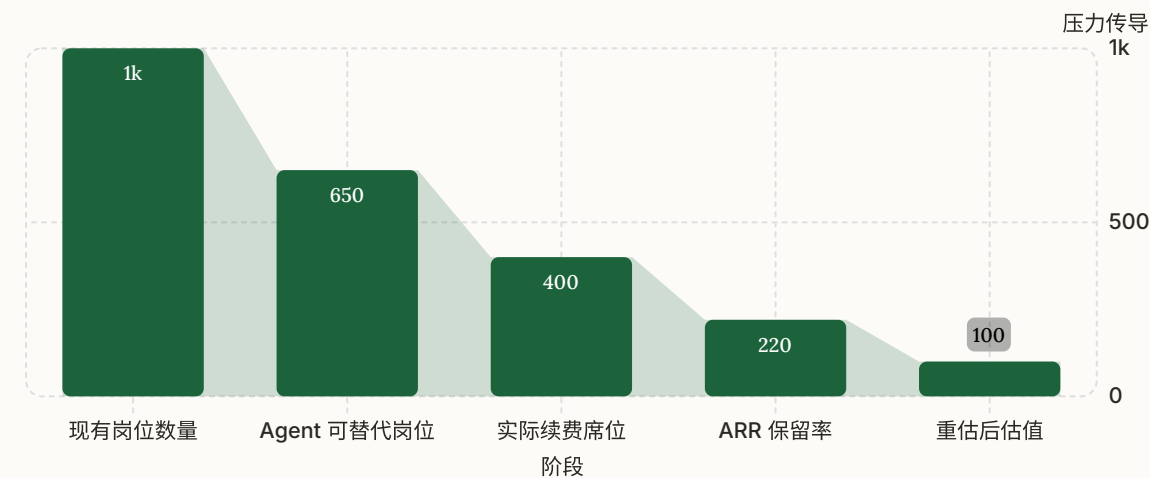


公司	1/2 P/E	3/31 P/E	P/E 变化
Salesforce	33.9x	23.9x	-29.4%
ServiceNow	89.3x	62.6x	-29.9%
Workday	86.5x	50.2x	-42.0%

席位制 SaaS 会先失血

旧收费锚点开始松动

10%
Total conversion rate



断层逻辑

- 岗位会被压缩
Agent 替代可量化岗位，用人需求下降
- 席位会被压缩
岗位减少直接带动 SaaS 席位续费下降
- ARR 先被重估
二级市场将率先对席位制 SaaS 重新定价

2026Q1

压力兑现节点

大规模企业 SaaS 续签周期到来，Agent 替代效应将首次反映在 ARR 数据中

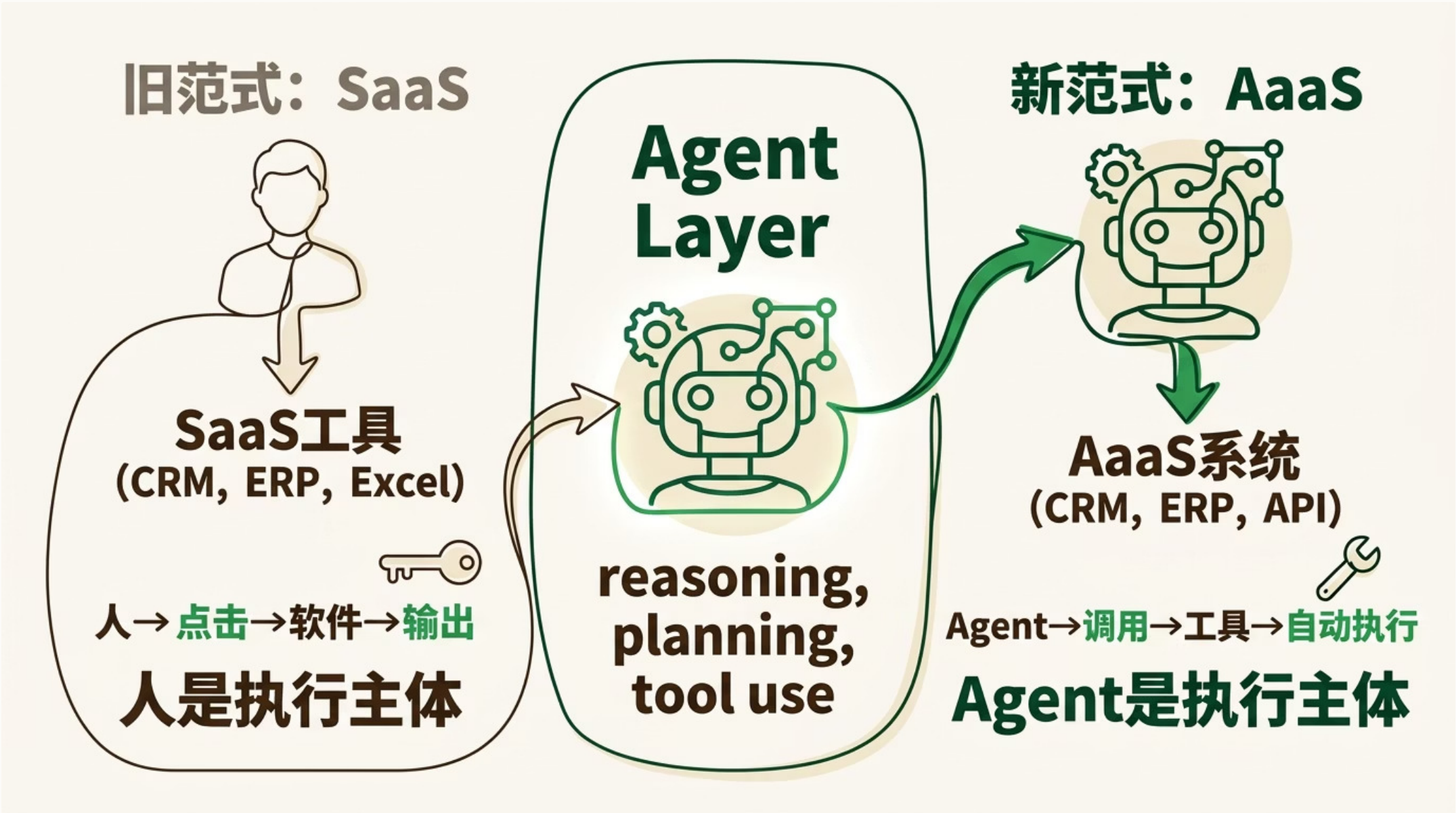
>1万亿

席位制 SaaS 市值压力（美元）

全球席位制 SaaS 估值面临系统性重估，失血规模超万亿量级

SaaS正在被重写为AaaS

软件不再是工具，而是执行系统



AI 时代，软件的本质正在发生根本性转变：从传统的“人用工具”SaaS 模式，进化为由 AI Agent 驱动的“Agent as a Service”(AaaS) 执行系统。这意味着软件不再只是等待人类指令的被动工具，而是具备自主推理、规划和执行能力的智能劳动单元。

10-15人

Agent 可替代劳动量

一个 Agent 在某些业务场景下可替代10-15人的工作量，大幅提升效率。

重构

SaaS 估值模型

传统 SaaS 估值面临重压，AaaS 将以“成果”而非“席位”计费，重塑利润结构。

❏ 核心变革：谁在做事变了，钱往哪里流变了。SaaS 不是被替代，而是被重构为更高效、更自主的 AaaS。

软件入口会收缩到调度器

任务入口将压过 APP 入口

前台：个人 Agent 调度器

用户只与一个入口交互——个人 Agent 前移，成为唯一的任务意图接收层。调度权决定流量分发，谁掌握调度器谁掌握生态入口。

1个前台入口

调度器统一接收用户意图

后台：多个 App 组件化

App 退到后台，成为被调度的功能组件，不再直接与用户交互。单一 App 的品牌入口价值将被系统性削弱。

多个后台组件

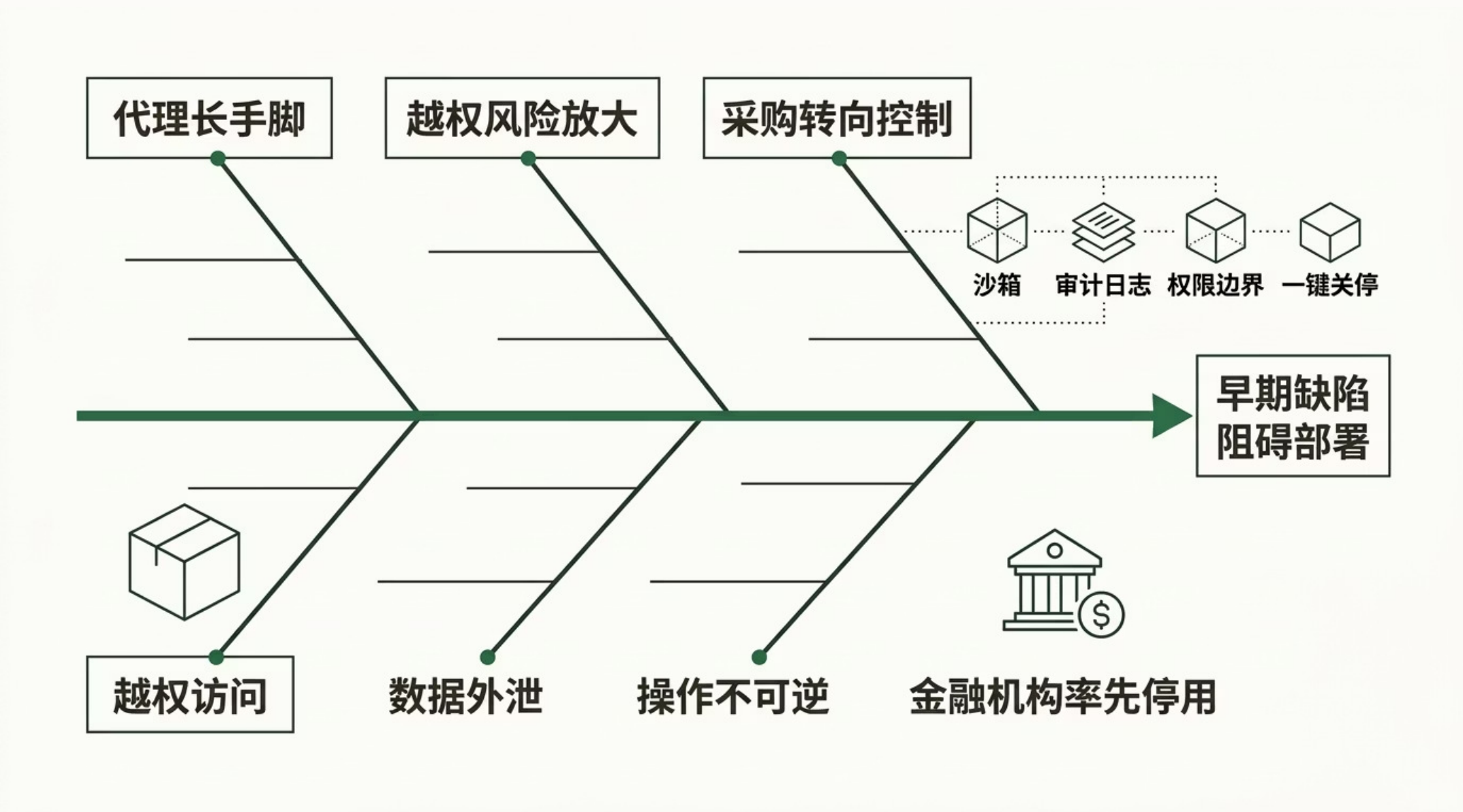
CRM / ERP / 文档 / 数据 — 均被调度



企业采购先审查可控执行

边界控制已成第一门槛

功能不再优先、越权会挡放量、审计回放必要——企业买 Agent，首先问的不是「能做什么」，而是「敢不敢放量」。可控执行已超越功能完整性，成为企业采购的第一卡口。



功能不再优先

能力强但不可控的 Agent，企业宁可不买，安全优先于效率

越权会挡放量

Agent 一旦发生越权事件，采购流程即时冻结，金融机构最严

审计回放必要

每个执行步骤可审计、可回放，已成监管合规的底线要求

3类

早期高频缺陷

越权访问 / 数据外泄 / 操作不可逆——三类缺陷直接触发企业采购否决

金融优先

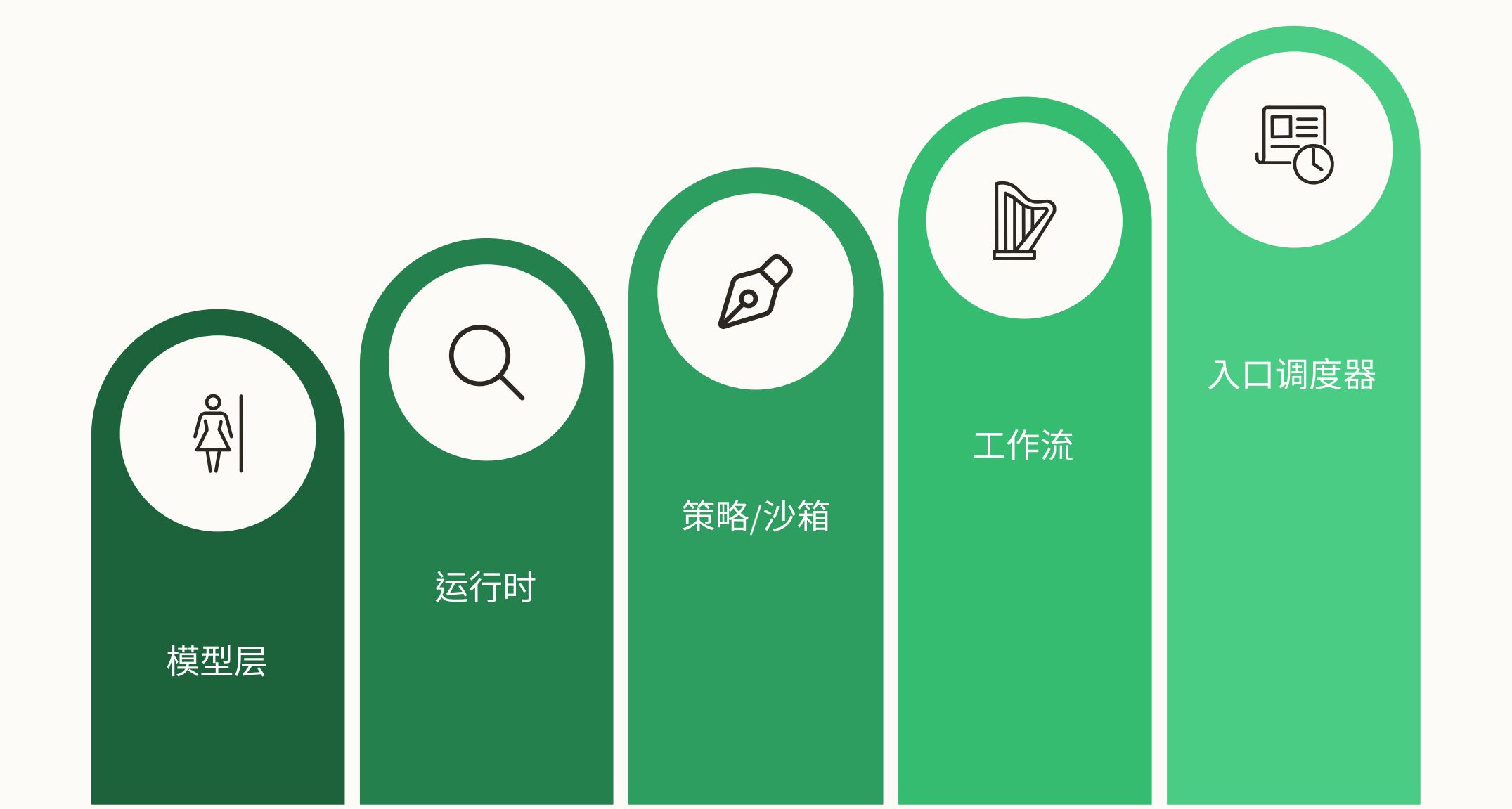
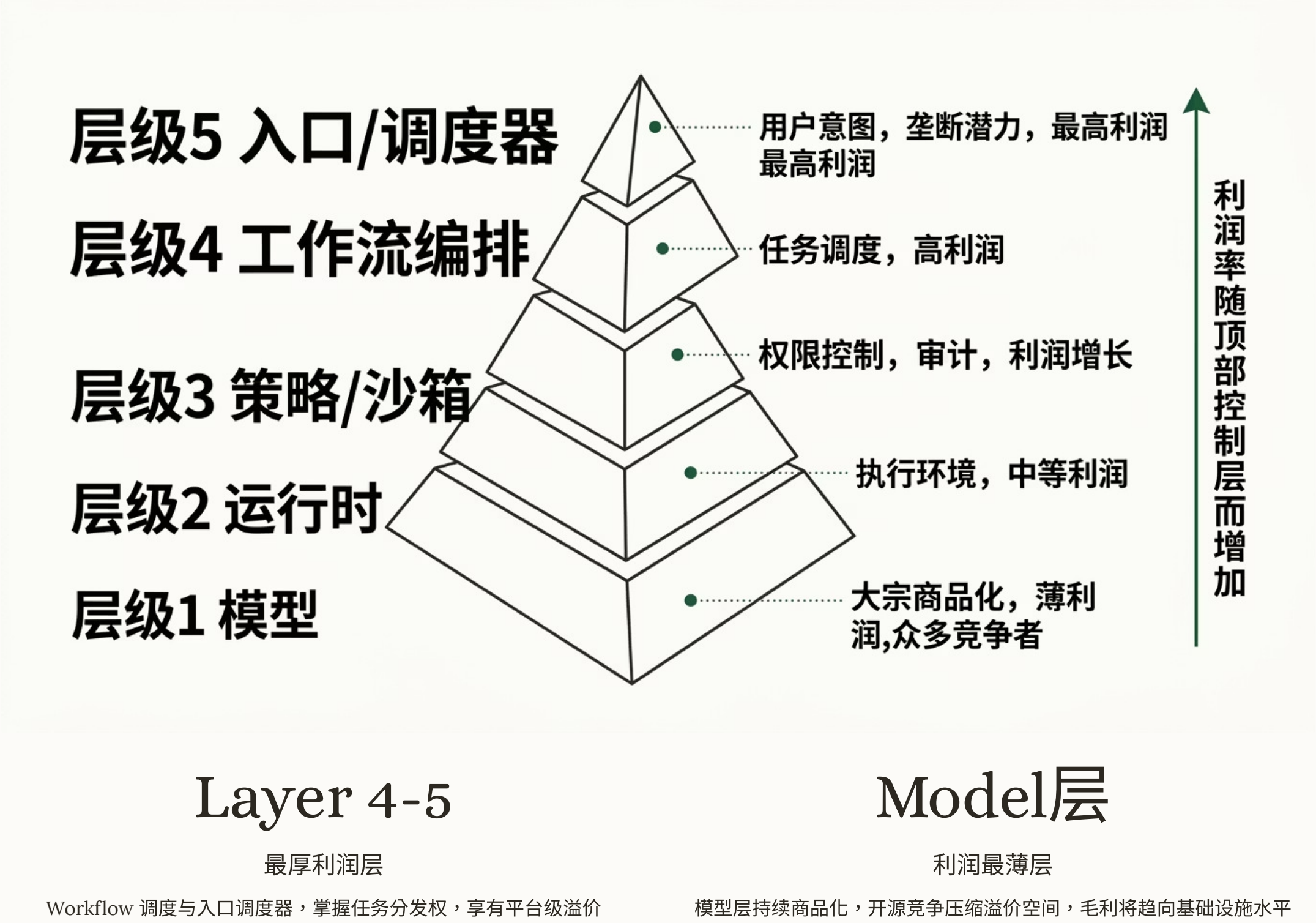
最严审查行业

金融机构率先封杀不可控 Agent，监管传导至其他行业只是时间问题

控制平面将先截留高毛利

利润开始向控制层沉淀

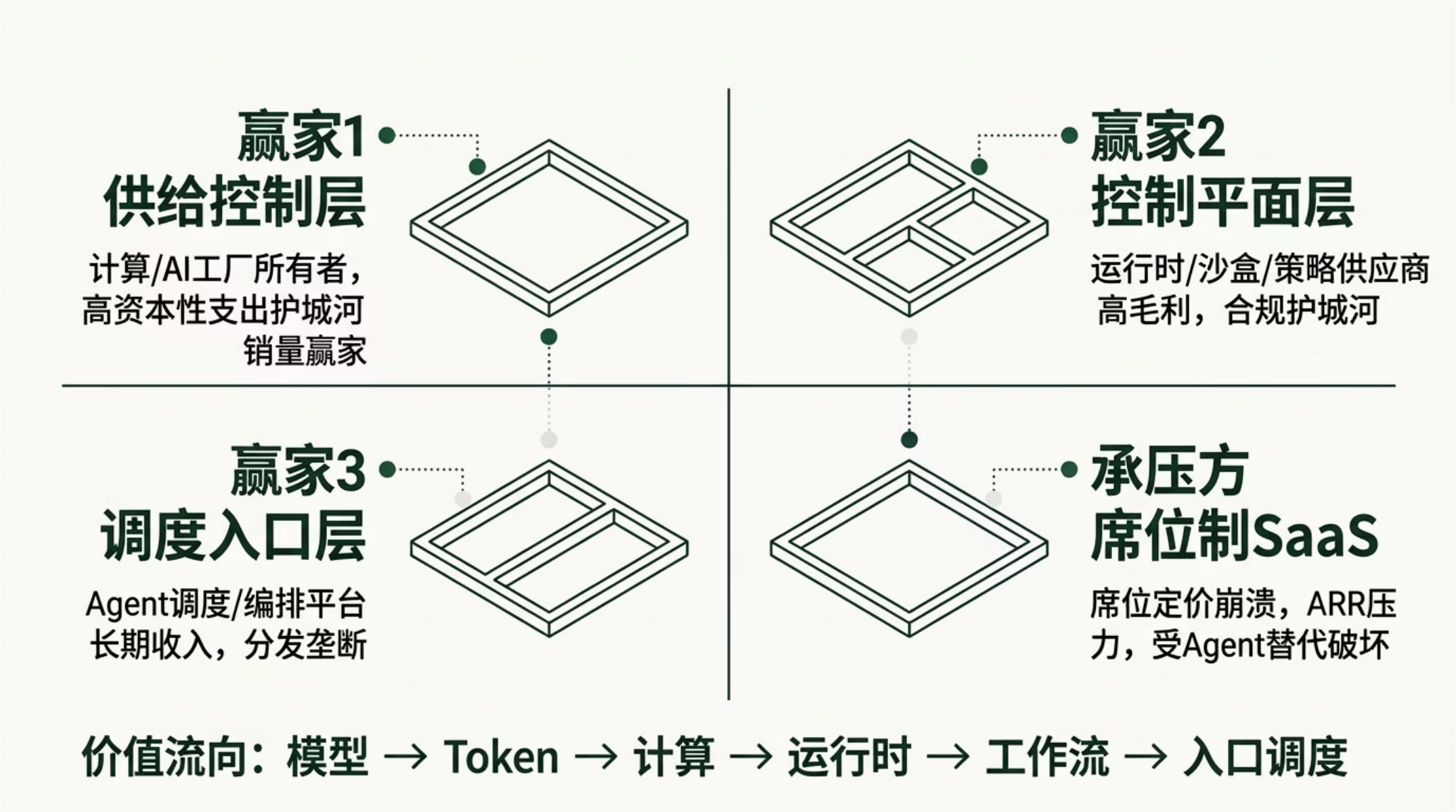
Runtime 值钱、Sandbox 值钱、入口层更值钱——在 Agent 全栈中，控制权越靠近执行层，利润越厚。模型层将持续商品化，控制平面将持续溢价。



AI 利润将流向供给控制入口

价值分配表已被重写

供给侧吃大头、控制层吃高毛利、入口层拿长期收入——这是 Agent 时代价值分配的终局判断。三类赢家格局清晰，一类承压方已现身。



供给侧吃大头	控制层吃高毛利	入口层拿长期收入
AI Factory / Compute 供给方，物理稀缺性构筑最深护城河，量级红利最大	Runtime / Sandbox / Policy 层，合规溢价持续，毛利率最高	调度器 / 编排平台掌握分发权，一旦占位形成长期用户锁定

3类

结构性赢家

供给控制 / 控制平面 / 调度入口——三类玩家在 Agent 时代享有结构性优势

1类

承压方

席位制 SaaS——旧收费锚点失效，ARR 面临系统性重估，先失血先出局

核心心智模型

Token Marketplace = 新一代电商

OpenRouter 不是 API 聚合器，它是“智能”的 Amazon

用“电商”理解 Token 中转站，比用“MaaS / API 网关”更有预测力。

核心问题不再是“谁的模型最好”，而是“谁掌握了智能被发现、购买、组合的交易界面”。

电商组件 → Token Marketplace

电商 → Token Marketplace SKU → Model	电商 → Token Marketplace Seller → Inference Provider	电商 → Token Marketplace Buy Box → Routing Engine
电商 → Token Marketplace Checkout → Unified Billing	电商 → Token Marketplace 卖家评分 → Uptime / 吞吐 / 工具调用	电商 → Token Marketplace 跨境电商 → 跨区域模型分发

三大结构性相似



差异化供给

模型各有所长，如同不同品牌 SKU



差异化需求

按“使用场景”选购，而非追求单一最优

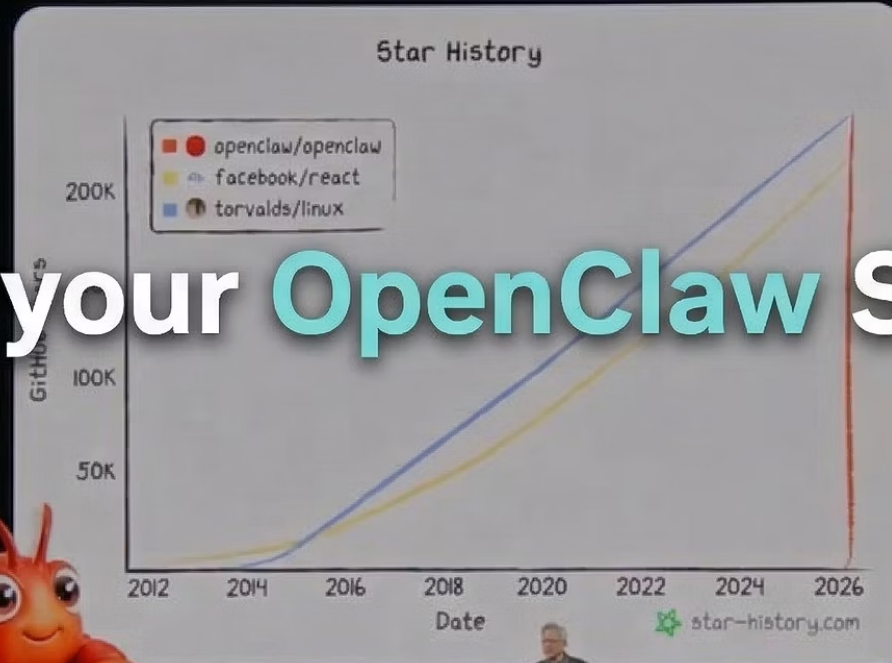


高比较成本

价格 / 性能 / 合规错综复杂，需要聚合层

Inference Inflection Arrives

What is your OpenClaw Strategy?



Jensen Huang at GTC 2026

“Every single company in the world today has to have an OpenClaw strategy.”

