

Token 经济崛起： 从技术单元到万亿级产业



北京湘汉有恒科技有限公司（湘汉智库）制作

2026年10月

目录

CONTENT

01

从技术层面理解 Token

02

回顾 Token 历史使用量轨迹

03

预测未来 Token 使用量变化趋势

04

总结

Token的本质：AI的“字数统计单位”。



人类看书：按“字”算

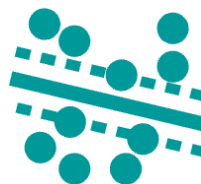


AI看东西：按“Token”算



学术定义：

Token 是AI（特别是大语言模型）处理文本时的基本意义单元。



来龙去脉：

计算机不认识完整的字，它只认识数字。数字化靠比特化（先），智能化靠数字向量化（后），一个Token 对应一个数字向量。

一般来讲，0.75个英文单词 $\approx$ 1 Token;
1个汉字 $\approx$ 1 Token。

文本	Token 数	切割方式示意
Hello	1	Hello
Transformer	3	Trans form er
I love you	3	I love you
Hello, world!	4	Hello , world !

文本	Token 数	拆分方式
你	1	你
你好	1 或 2	你 好 或者 你好，取决于词表优化
龔龔 (dá生僻字)	3	退化为 3 个 UTF-8 字节 Token

“你好厉害！” “你好自为之”

Token计量单元由一套复杂的子词 (Subword) 切分算法决定

切分算法目标：

在词表大小
(Vocabulary Size)
与表示效率之间取得平衡

切分太细 → 词表大 → 效率低；
切分过大 → 无法精准表示

(1) BBPE (Byte-level BPE)

采用模型：GPT 系列、DeepSeek、Claude

(2) Unigram

采用模型：谷歌T5全系列模型

(3) WordPiece

采用模型：BERT, Claude(早期)

三大主流切分算法

不同模型采用不同切分算法，导致各自词表存在差异，也直接影响中文的压缩效率（英文影响较小）

Token消耗量=输入Token+（模型推理复杂度×任务难度）思考Token+输出Token

表. Token 使用量的影响因素

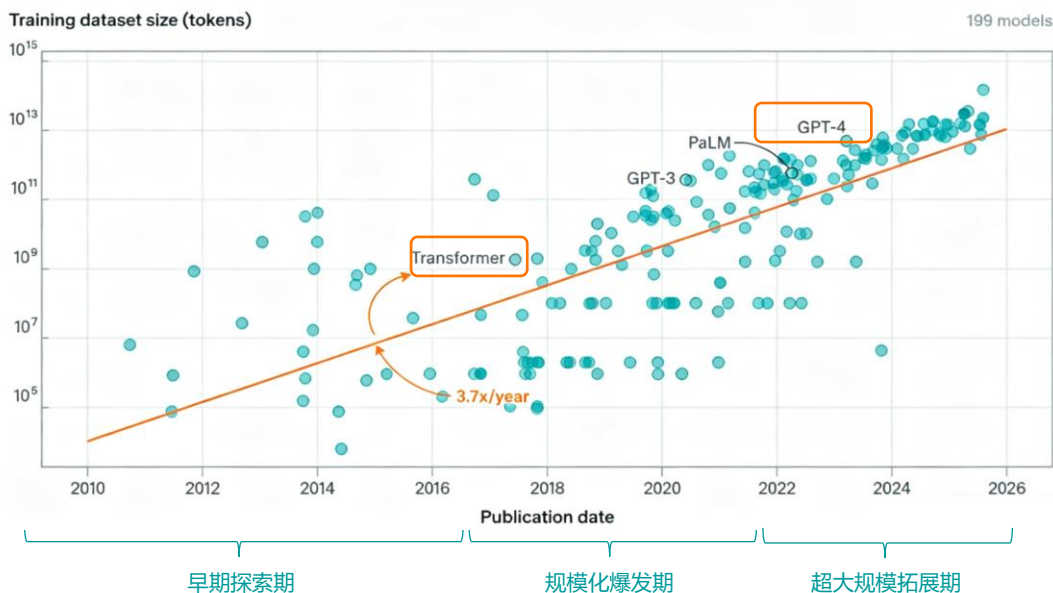
序号	影响因素	举例说明	对Token使用的影响
1	输入文本长度	文档越长，Token越多	正相关
2	模态类型	多模态引入图像、视频、音频Token	大幅提升Token消耗
3	应用场景	搜索、聊天、生成音乐、图像描述	激增（更长上下文）
4	质量/准确性需求	高精度推理，可能更长输出	推理Token增加

一般大模型费用计算公式：总费用=（输入Token×输入单价）+(思考Token+输出Token)×输出单价。

模型训练端Token 使用量显著增长

图. 2010–2026年间模型训练数据规模（Token数）增长关系

Training data of notable LLMs



1.早期探索期（2010–2017）：小规模、碎片化

该阶段是NLP技术的“史前时代”，以小模型、任务型模型为主，模型训练Token量普遍在 $10^5 \sim 10^{11}$ （十万到千亿）之间，规模小且分散。

2.规模化爆发期（2018–2022）：指数级增长、标杆模型涌现

2017年Transformer架构发布，成为大模型爆发的技术基石。GPT-3（2020）将训练数据量推至数百亿Token级别，奠定“大模型=大数据”的行业共识。谷歌PaLM（2022）规模扩展至千亿Token，验证了“数据scaling law”的有效性。

3.超大规模拓展期（2023–2026）：逼近万亿级、持续加速

模型训练Token量突破 10^{12} ，向 $10^{13} \sim 10^{14}$ （十万亿）迈进，GPT-4（2023）代表了这一阶段的顶峰。行业进入“超大规模预训练”阶段，数据量成为核心壁垒。

推理端的Token消耗，与上下文窗口*以及Transformer 计算结构深度绑定

早期

- **上下文窗口非常有限**

- BERT为512 Tokens，第一代 GPT 模型为1024 Tokens。处理单文本的任务。

中期

- **上下文窗口开始快速扩大**

- GPT-3为2048 Tokens，GPT-4从8k扩展到32k，而 Claude 2达到100k级别。一次性处理一篇论文内容。

当前

- **进入超长上下文阶段**

- 支持200k+ Tokens甚至更高，一次性处理一本书的内容。

上下文窗口：大模型当前能“记住”并一起思考的文字总长度。

自注意力计算复杂度 $=O(N^2)$

其中N是输入Token数量，当上下文窗口Token扩张时，计算复杂度出现平方级放大效应。因此通常需要模型生成越长的推理过程，从而消耗更多的Token。

在使用端，多模态带来 Token 消耗量暴增

图. 多模态数据 Token 化映射标准与数量级对比表

模态	单位	Token 替代定义	Token 数量
文本	1个句子 (~20字)	字/短语 subword	~20个
图像	1张普通照片	Patch Token(按像素块)	196 ~ 1,600个
音频	1秒语音	声学Token(按照音频帧率)	50 ~ 300个
视频	1秒视频 (标准)	空+时Token(音频+图像)	800 ~ 8,000个

同一请求可能产生：**文本Tokens+视频Tokens+跨模态注意力Tokens***。这意味着单条请求的Token数爆发式增长。如果模型是联合预训练和联合推理，那么Token需求量更高。

跨模态注意力Tokens：在联合推理架构中，模型需要计算文本Token与视觉Token之间的注意力矩阵（Attention Matrix）。如果文本长度为 N ，视觉Tokens长度为 M ，跨模态注意力的计算复杂度通常与 $N \times M$ 相关。

AI 向物理世界延伸，是未来 Token 消耗的主要来源

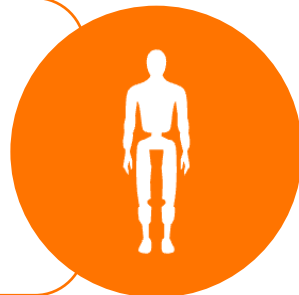
未来，AI向物理世界延伸（具身智能），将促使Token的应用场景从单纯的内容生成扩展到复杂的任务规划，期间会带来巨大的Token消耗。自动驾驶和机器人是目前AI在物理世界的典型应用。



自动驾驶

自动驾驶系统每秒可产生数千到数万个Token的输入需求（视觉语言模型，VLM）。若联网端云协同架构普及（边缘推理+云端感知融合），全球数亿辆联网车辆将形成极为庞大的持续性Token消耗。

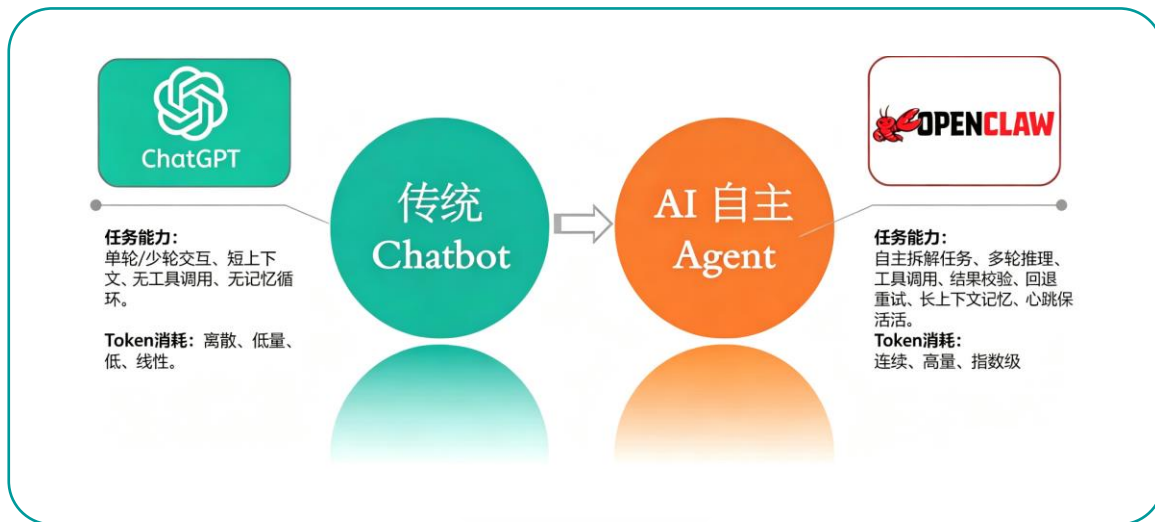
机器人的场景更复杂、决策频率更高（多机协作）、物理反馈循环更紧密，单机Token消耗量可能更高。随着工厂自动化、仓储机器人、家庭服务机器人的规模扩张，这一领域将成为Token消耗增量的重要来源之一。



机器人

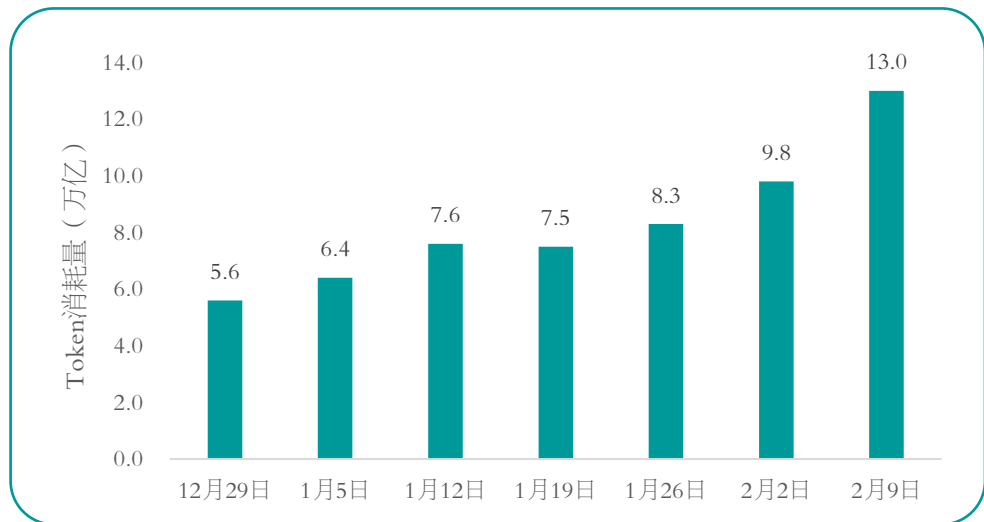
AI 自主 Agent 给 Token 消耗带来指数级增长

表. 传统 Chatbot 与 AI 自主Agent任务能力和 Token 消耗比较



OpenClaw类自主Agent的Token消耗本质是，从“对话”到“执行”的跃迁。OpenClaw（及国产类OpenClaw）是可自主执行闭环任务的Agent框架，核心是从“被动问答”升级为“主动执行”。

表. 2025年12月底-2026年2月初 OpenClaw 在 OpenRouter 平台 Token 消耗量变化



据OpenRouter发布的数据，OpenClaw在OpenRouter平台Token占比从2026年2月初10%升至3月初17%，2月第一周全球Token环比增长33%。

中国大模型 Token 调用量乘势崛起， 自主Agent 给中国AI产业创造巨大机遇

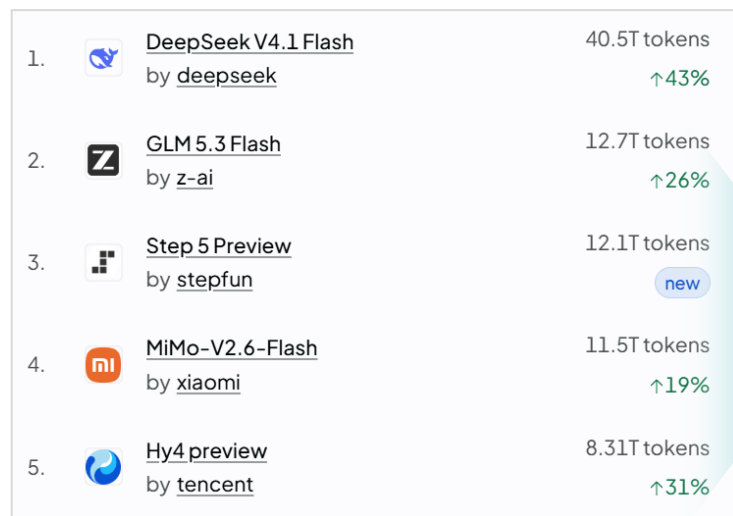
图. 过去一年OpenRouter平台中国、美国及全球Token调用量变化态势



数据来源: OpenRouter 平台

随着OpenClaw推动全球Token调用量暴增，2026年2月中国AI的模型调用量出现历史性拐点，首次超过美国。3月上旬保持超越美国的记录。

图. 2026年10月第一周OpenRouter平台Token调用量Top 5模型 (如下图)



数据来源: OpenRouter 平台

2026年10月第一周，全球调用量排名前五的大模型全部来自中国厂商。

中国大模型Token调用量崛起的主要原因

(1) 价格远低于美国模型：中国大模型价格只有美国的1/3 ~ 1/10。

(2) 性能比肩全球顶尖模型：中国头部模型在代码生成、长文本处理和逻辑推理能力上已追平甚至在部分基准测试中超越美国模型。

(3) 快速迭代与开放生态：面对自主Agent带来的新需求，敏捷迭代（工程化能力）。另外，中国厂商不强制绑定自有生态，降低了全球开发者的接入门槛。

短期内产生替代 Transformer 架构导致 Token 使用量下降的概率极低

（一）学界工业界关注的主流替代方案的局限

1

Mamba（选择性状态空间模型，SSM）

在需精确全局信息召回的任务上，表现不稳定，且大规模训练验证尚不充分。

2

RWKV（线性复杂度 RNN 变体）

顺序递推限制了对长距离依赖关系的捕获能力，在复杂推理任务上与 Transformer 仍有差距。

3

Hyena / H3 / RetNet 等线性Attention变体

在需要精确 Token 间交互的任务中，近似注意力的信息损失导致效果上限明显低于标准 Attention。

（二）生态的迁移成本极高

数年积累的训练基础设施、CUDA 内核优化、模型微调与部署工具链，全部深度绑定在 Transformer 架构上，任何替代方案都需要从零重建这套体系，工程代价极高。

（三）替代架构：缺失万亿级实证支撑与模糊的多模态路径

Transformer 已在万亿参数规模上完成了稳定验证，而替代架构普遍缺乏同等量级的实验支撑，多模态扩展路径也尚不明确。

短期内发生完全替代的概率极低，主流算力与工程投入仍将集中在 Transformer 体系。

主流 Attention 架构与推理优化方案不能直接降低 Token 使用量

主流 Attention 变体：

(1) MHA (Multi-Head Attention, 多头注意力)

每个注意力头独立维护Q、K、V三套参数。计算与内存开销随头数和序列长度线性增加。在长上下文下KV Cache 体积巨大，是显存瓶颈的主要来源。

(2) GQA (Grouped Query Attention, 分组查询注意力)

多个 Query 头共享同一组 Key/Value 头，相当于 MHA 和 MQA (Multi-Query Attention) 的折中方案。以较小的精度代价换取 KV Cache 体积大幅压缩。

(3) MLA (Multi-head Latent Attention, 多头潜在注意力)

DeepSeek 提出的新型结构，将 K、V 压缩到低秩潜在空间再展开，从根本上降低 KV Cache 存储量，同时保持模型表达能力。

表. Transformer 中 Attention 机制的四大优化策略

优化方案	优点	局限
FlashAttention	速度大幅提升，内存占用显著降低	仅优化计算效率，不改变 $O(N^2)$ 的理论复杂度
稀疏Attention	可处理更长序列，算力消耗线性化	全局信息捕获能力受损，任务质量存在上限
线性Attention	复杂度降至 $O(N)$ ，理论效率极高	实际效果仍落后标准 Attention
KV Cache优化	自回归生成速度大幅提升	长对话或超长上下文导致显存压力指数级上升

Token数量主要由输入提示词（Prompt）长度和模型生成的回复长度决定，属于数据层面的范畴。MHA、GQA、MLA等架构变体，以及FlashAttention、KV Cache优化等技术，核心目标在于降低计算复杂度、减少显存占用（特别是KV Cache体积）并提升推理速度。

未来可能促使Token使用量下降的技术路径仍困难重重

图. 热门技术路线

1

Token 压缩与高效表示:

模型在处理长文本时存在大量冗余Token。Token 压缩技术（如 LLMlingua、AutoCompressor）可在不显著损失语义的情况下将输入Token 压缩 2-10倍。

更高效的多模态Token化

当前图像/视频Token化效率较低（一张图片动辄数百Token），下一代Tokenizer 的优化可能将同等信息量的视觉内容压缩至更少Token。这是多模态场景中降低Token消耗的关键杠杆。

2

3

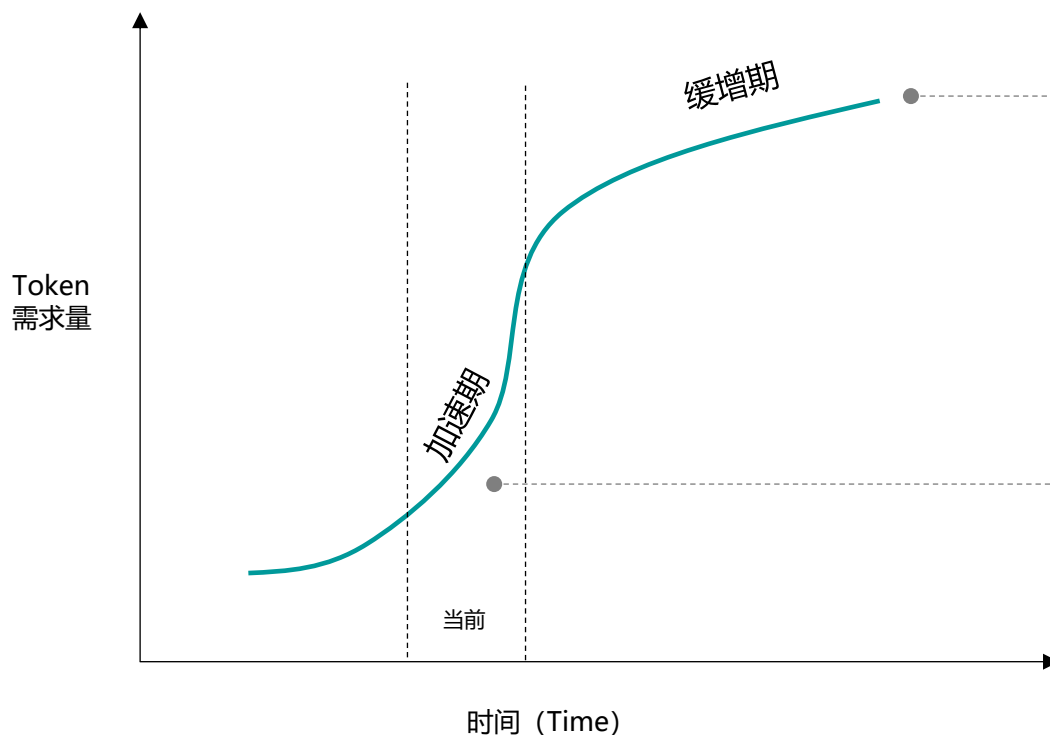
模型蒸馏与专用小模型

对于特定、高频的垂直任务（如客服、代码补全、文档摘要），用大模型蒸馏出的专用小模型，可以用更少Token完成等效任务，同时降低延迟与成本。

左述路径促使Token消耗量下降是困难的，并非因为技术本身无效，而是用户为了追求更聪明、更精准的效果，宁愿多花点Token也不愿冒险压缩丢失信息。结果就是：单次省下的那点Token，瞬间被爆发式增长的使用次数和更奢侈的用法给淹没了，所以总的消耗量很难降下来，甚至还会越用越多。

总体趋势：Token需求量呈现S形增长，目前处于加速期，未来增速逐渐放缓

图. Token使用量随时间的变化趋势



- 市场渗透率接近上限，用户基数趋于稳定；
- 新场景拓展难度加大，行业进入存量竞争；
- 需求增长更多来自效率优化，而非规模扩张。

- AI自主Agent高速发展，单次任务的Token消耗量将实现10~100倍的跨越；
- 模态升维，随视频、音频等多模态数据原生接入；
- AI物理世界延伸，具身智能将Token化过程引入现实世界。

中性估计，2030年全球Token消耗量达5000亿亿；中国达2700亿亿，占全球54%

2025年全球基准Token消耗量6.3亿亿，2025年中国基准Token消耗量2.1亿亿*。结合算力供给、技术成速度及落地、政策监管、B端渗透等核心变量，设定基准/高速/低速三类情景，概率赋值贴合产业一致预期。

情景维度	发生概率	分年度同比增速（全球）					全球CAGR (2026-2030)	中国CAGR (2026-2030)	2030年全球总量区间（亿亿）	2030年中国总量区间（亿亿）
		2026	2027	2028	2029	2030				
基准情景 (中性)	60%	900%-1000%	330%-350%	180%-220%	120%-150%	80%-100%	220%-250%	280%-320%	3000~5000	1660~2700
高速情景 (乐观)	25%	1000%-1200%	380%-420%	250%-300%	180%-220%	120%-150%	280%-320%	350%-400%	7200~13600	3880~6560
低速情景 (悲观)	15%	600%-800%	200%-250%	100%-150%	50%-80%	30%-50%	150%-180%	180%-220%	500~1300	360~700

* 2025年中国基准Token消耗量2.1亿亿，数据来源于《全国数据资源调查报告（2025年）》[R]. 北京：国家数据局，2026. 中国占全球全量Token消耗量区间30%~38%计算，2025年全球Token消耗量为5.5~7亿亿Token。本报告中间值为6.3亿亿。

公司简介

北京湘汉有恒科技有限公司（简称“湘汉有恒”）是一家专业的科技情报服务提供商，成立于 2023 年 3 月，总部位于北京市昌平区。公司依托知识图谱、数据挖掘与人工智能等技术手段，从海量科技大数据中挖掘高价值科技情报，致力于让创新型客户的科技创新更有方向、更具力量。公司旗下设立专业科技智库——湘汉智库，为政府部门和企事业单位提供可信的科技循证决策服务。

公司聚焦科技创新情报与科技人才情报两大业务方向，提供科技情报数据服务与科技情报研究服务。自主研发学术论文、全球专利、全球科研项目、企业、科技资讯五大情报系统，其中专利系统收录全球 1.7 亿件专利、覆盖 170 个国家（地区），科研项目系统汇聚 100 多个国家和地区的 680 多万个受资助项目数据，为精准、权威的情报服务提供坚实数据支撑。公司重点关注人工智能、集成电路、量子科技、新一代信息技术、生物医药、新材料、空天技术等前沿领域。

公司创始人长期在中央媒体、知名互联网企业及人工智能公司从事信息情报类研究工作，曾担任国内顶尖人工智能公司科技情报研究咨询部负责人，撰写的科技内参报告多次获中央和北京市领导批示。公司团队专业功底深厚，服务对象涵盖政府部门、事业单位与知名企业，包括北京市科学技术委员会、深圳市南山区人民政府等政府机构，北京理工大学、北京航空航天大学、中国工程院、中国科学技术协会、北京智源人工智能研究院等科研单位，以及阿里巴巴、国家电力投资集团、西门子等知名企业。

公司秉持“让科技情报更有价值”的使命，恪守“兴趣成就、经世致用、进德修业、坚韧有恒”的价值理念，愿与客户携手合作，共同走在科技无尽前沿，成就非凡。



— 让科技情报更有价值 —



— 基于数据和知识研究的科技智库 —



湘汉有恒公众号