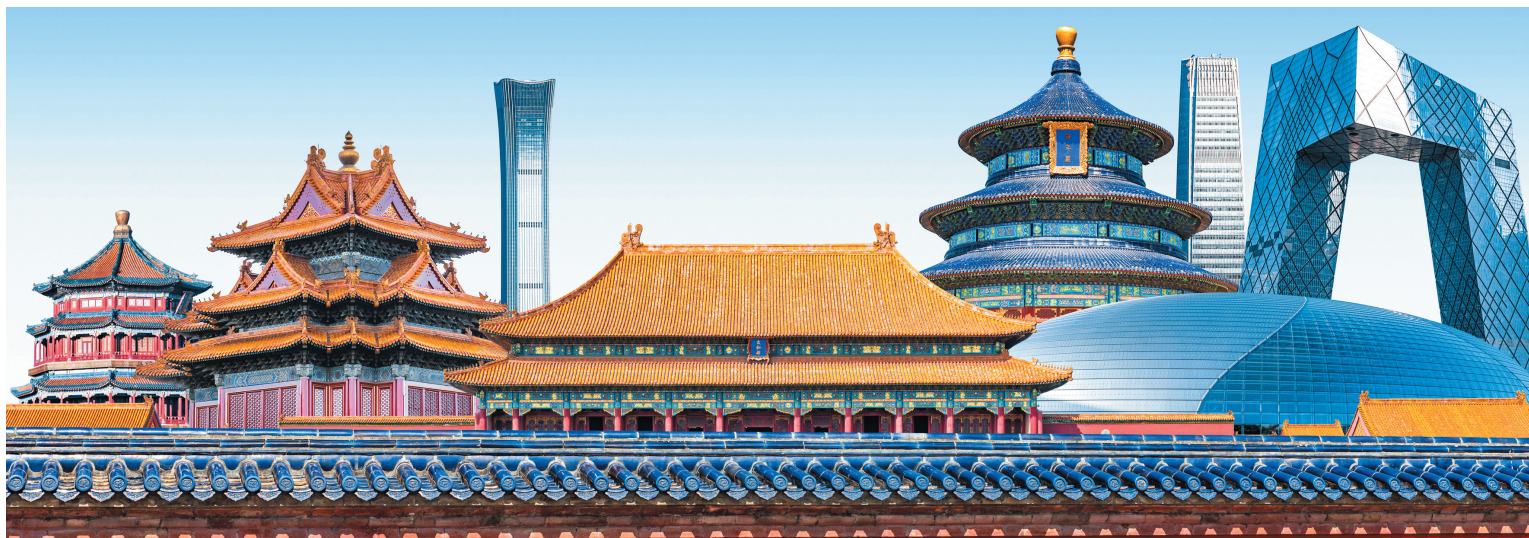


北京：人工智能+消费大主场



北京市海淀区的鼎好大厦，具身智能机器人小盖正在便利店上岗，为顾客夹取烤肠、取货。20公里外的国贸，一位白领在下班前5分钟对手机上的“小团”说：“公司附近，6个人，晚上7点，川菜，能订上位。”等她走出电梯，餐厅已经订好，打车券发到账户。

AI第一城的底座能力，正在北京无数个细微的消费现场兑现。

2026年6月，商务部等8部门印发关于加快“人工智能+消费”发展的实施意见，7月国务院关于《扩大消费“十五五”规划》的批复中明确，深入推进“人工智能+消费”。

政策红利之下，谁能更好承接？答案藏在AI底座里。北京不是政策出台后才开始布局AI。截至2026年7月，北京累计备案大模型259款，日均调用量超14亿次。在这里，智谱GLM5.2、Kimi K3等一批来自北京的大模型在全球开源榜登顶，智源研究院牵头打造的最新版众智FlagOS支持九成以上主流开源大模型的多芯片部署。

早在2024年，北京在全国率先发布“人工智能+”行动计划，是一座把算力+算法+场景凑齐的AI城市。2025年北京AI核心产业规模预计为4500亿元。2026年上半年，北京可穿戴智能设备、高能效等级家用电器和音像器材类商品零售额均增长六成以上，服务机器人、工业机器人等产品产量分别增长230%、75.5%。北京正在证明：AI第一城的含金量，最终会在数字世界和物理世界的千行百业直接或间接被兑现。

人工智能“第一城”的底盘

2026年6月，在中关村国际创新中心召开的智源大会上，一位参观者在清微智能的展台停下脚步，台上展示的是清微AI芯片，采用3.5D异构堆叠与Chiplet架构，让可重构计算芯粒与DRAM存储芯粒实现了三维垂直堆叠，从传统2D平面上的“单车道”，变成立体贯通的“四车道”。

同月，清微智能Day0适配智谱GLM-5.2模型，次月，依托FlagOS适配腾讯混元Hy3。摩尔线程和大模型的联动同样迅速，7月21日—8月3日，摩尔线程全功能GPU（图形处理器）先与极佳视界的具身世界模型完成训练与推理验证，随后宣布Day0完成Kimi K3和MiniMax H3适配。

芯片是“基座”，算力是“能源”，2026年上半年北京新增智能算力2.2万P，供给规模达到8.2万P。按照当前算力水平，3万P算力可支撑上千个大模型并行训练。基础设施对模型的支持已经实实在在体现在性能上。

7月17日凌晨，月之暗面发布Kimi K3，2.8万亿参数、100万Token上下文，这意味着Kimi K3可以一次性完整载入整本《封神演义》，并且还有富余空间继续附带额外资料、提问历史。这款大模型发布当天，即登顶由全球用户共同投票的“大模型竞技场”Arena榜。特斯拉CEO马斯克在相关内容下留言“Impressive（令人印象深刻）”。

3月，月之暗面发布注意力残差技术报告后，马斯克就曾转发称赞。当月，在海淀区举办的中关村论坛年会全体会议的演讲台上，月之暗面创始人兼CEO杨植麟分享了这份报告，7月下旬，《北京市关于支持智能体引领发展的若干措施》新闻发布会上，北京市发展和改革委员会党组副书记、副主任林剑华也提到注意力残差。

那场新闻发布会上，基础模型智谱GLM-5.2，通用智能体开发平台字节扣子、百度文心，垂直领域智能体平台和框

架范式智能、面壁等以及推理引擎Tile RT、中科加禾SigInfer、清程极智赤兔等均被提及。

透过这份名单看，北京AI底座、产业链分工被串联起来。

2026福布斯中国人工智能科技企业TOP50榜单中，北京有21家企业上榜，居全国之首，覆盖大模型、AI芯片、企业级AI等细分产业赛道。

其中“原点社区”所在地的海淀，有10家企业在列，这块总面积约3平方公里的区域是北京首批4个人工智能创新街区之一，集聚了1000余位AI科学家、1.3万余名AI开发者，还有10万余名AI青年学子。

作为全国人工智能产业创新高地，北京累计完成备案的259款大模型，覆盖通用大模型、行业垂直模型、手机端侧模型等多元品类，头部模型注册用户超20亿，日均调用量超14亿次。