人工智能专题：小米AI布局

行业研究 · 行业专题

计算机 · 人工智能

投资评级：优于大市（维持）

证券分析师：熊莉
021-61761067
xiongli1@guosen.com.cn
S0980519030002

联系人：云梦泽
021-60933155
yunmengze@guosen.com.cn

摘要

- **小米科技通过聚焦底层技术与AI赋能，正式推出自研大模型。**2023年8月14日，雷军宣布小米将深耕底层技术，长期持续投入，推动软硬深度融合，并全面赋能AI，提出公式（软件×硬件）^{AI}。小米的自研大模型主要突破方向为“轻量化、本地部署”，既保障用户的数据安全，又提升生产力。自2016年成立AI实验室以来，小米逐步布局了包括视觉、语音、NLP等12个技术领域，预计在2022-2026年间将投入超过1000亿元的研发经费，沉淀技术积累。

- **小米MiLM2升级发布，性能与技术全面提升。**2024年11月，MiLM2模型完成了从一代到二代的升级，增强了模型的参数矩阵，支持云边端结合，提升了推理速度和量化性能，且长文本处理能力处于行业领先地位。团队在预训练、量化、推理加速等领域也进行了一系列技术创新，包括SUBLLM、TransAct、INTRADoc等新结构，进一步提升了训练和推理效率。小米的「人车家全生态」战略，旨在构建涵盖人、车、家等场景的智能生态系统，在实时交互需求日益增长的背景下，对大模型的生成、闲聊、翻译等能力提出了更高要求，MiLM2模型表现出色。

- **MiLM2大模型矩阵灵活扩展，充分适配多元化场景。**小米自研大模型团队灵活扩展了模型参数规模，涵盖0.3B至30B多个量级，以适应不同的业务场景和资源需求。针对终端场景，0.3B到6B的模型可完成具体、低成本任务，并在微调后与百亿参数开源模型相媲美；6B至13B的模型支持多任务微调，达到几百亿开源模型效果；30B模型则专为云端设计，具备强大的zero-shot学习和复杂任务处理能力。端侧方面，4B模型在设备端成功部署，通过创新的“TransAct大模型结构化剪枝方法”和端侧量化技术，小米显著提升了训练和推理效率；MiLM2-30B作为云端模型，在指令遵循、常识推理和阅读理解等方面表现优异。

- **小米大模型应用落地，全面赋能各类设备。**提升用户体验。在手机端，支持AI图片编辑和智能视频剪辑；平板用于自动生成会议纪要和行业报告，提高工作效率；电视端提供影视问答、健身和家庭计划等功能；汽车端具备语音控车、智能导航等功能，提升出行便捷性。这些成果已应用于澎湃OS、小爱同学、智能座舱、智能客服等产品，帮助解决多样化业务需求。

- **2023年，小米正式宣布将集团战略升级为“人车家全生态”。**“超级小爱”借助 AI 算法与大模型技术，大幅提升自然语言处理与决策能力；小米澎湃 OS 2 内置的 Xiaomi HyperAI 带来多重创新，在 AI 写作、识音、字幕等方面提升，还与“超级小爱”紧密融合，将 AI 能力深入系统层，为全生态设备赋能；小米智能家庭平台作为生态链的控制与电商中枢，开放接入第三方产品，成功打造出涵盖产品接入、众筹孵化、电商销售、用户触达与控制分享的完整生态闭环。

- 风险提示：技术迭代不及预期、市场竞争加剧、个股梳理仅基于产业链结构，不涉及主观投资建议。

小米AI生态全景

2023年，小米正式宣布将集团战略升级为“人车家全生态”。“超级小爱”借助 AI 算法与大模型技术，大幅提升自然语言处理与决策能力；小米澎湃 OS 2 内置的 Xiaomi HyperAI 带来多重创新，在 AI 写作、识音、字幕等方面提升，还与“超级小爱”紧密融合，将 AI 能力深入系统层，为全生态设备赋能；小米智能家庭平台作为生态链的控制与电商中枢，连接米家生态链公司，开放接入第三方产品，成功打造出涵盖产品接入、众筹孵化、电商销售、用户触达与控制分享的完整生态闭环。

操作系统与大模型

人车家全生态操作系统：小米澎湃OS
基于AI的操作系统：Xiaomi HyperOS 2
AI大模型矩阵：包括MiLM - 6B等模型

智能电动汽车

8月：小米**声音大模型**在SU7首次上车
10月底：**城市 NOA** 全国都能开，开启全量推送
12月底：**端到端全场景智能驾驶**，开启先锋版推送

软件应用与服务

智能语音助手小爱同学：全面搭载大模型，多模态能力升级
涵盖**小爱通话、小爱翻译、小爱视觉、家庭传声、听声识人**等多种功能

硬件产品

智能手机：小米15系列搭载骁龙8至尊版
平板：Pad7 Pro搭载第三代骁龙8s旗舰处理器
智能家居：空调、冰箱、洗衣机等
可穿戴设备：手环、手表、耳机等

金山办公

WPS 365：包含了WPS Office、WPS AI企业版和WPS协作，打通了文档、AI、协作三大能力
WPS Office：内置在线智能文档、智能表格、智能表单等
WPS云文档（金山文档）：提供多样化的云办公体验
WPS AI：升级至 2.0，推出WPS AI办公助手、企业版、政务版

金山云

国内第一梯队游戏云服务商，提供全套云产品，包括统一的 IaaS 基础设施、PaaS 层和 SaaS 应用软件

资料来源：小米官网，金山办公官网，金山云官网，国信证券经济研究所整理

小米科技战略升级，小米自研大模型正式亮相

国信证券 GUOSEN SECURITIES

聚焦底层技术与AI赋能，自研大模型正式推出。2023年8月14日，雷军在年度演讲中宣布小米科技战略升级：深耕底层技术、长期持续投入，软硬深度融合，AI全面赋能，总结为公式（软件×硬件）^{AI}。**小米自研大模型正式亮相，主力突破方向为“轻量化、本地部署”**，让用户在享受安心的数据保护的同时，拥有大模型带来的先进生产力。小米自研大模型在当时权威中文评测榜单C-EVAL和CMMLU中，取得同参数量级第一的好成绩；小米自研手机端侧大模型初步跑通，部分场景效果媲美云端。

以AI为基石，沉淀技术积累。小米很早就对人工智能进行布局，2016年小米AI实验室成立，并组建了第一支视觉AI团队，2023年4月成立专职大模型团队，截至2023年8月小米人工智能团队已经有3000多人，逐步建立了视觉、语音、声学、知识图谱、NLP、机器学习、多模态等AI技术能力，并已布局了12个技术领域，99个细分赛道，2022-2026年至少会投1000亿以上的研发经费。

图1：小米AI发展历程

时间	团队	能力发展	平台发展
2016.7	AI 视觉团队	视觉、机器学习	MACE、MAFE、CloudML、多媒体
2017.9	小米 AI 实验室	视觉、声学、语音、NLP、知识图谱、机器学习、大模型	MACE、CloudML、MiNLP、图谱平台、多媒体、搜索推荐平台
2018.12	AI 影像算法团队	视觉、机器学习	MACE、MAGE、CloudML
2021.3	自动驾驶团队	视觉、机器学习	MACE、CloudML
2021.8	小米机器人实验室	视觉、声学、语音、NLP、知识图谱、机器学习、大模型	CloudML
2023.4	大模型团队	NLP、视觉	CloudML、MiNLP

图2：小米自研手机端侧大模型初步跑通，部分场景效果媲美云端

资料来源：小米公司公众号，国信证券经济研究所整理

小米大语言模型 MiLM 正式备案，多场景赋能提升效率

国信证券 GUOSEN SECURITIES

2024年5月16日，小米大语言模型 MiLM 正式通过大模型备案。历经一年多的打磨，小米大语言模型已形成包括多种参数规模和形态的模型矩阵，既通过小米澎湃OS系统和人工智能助手小爱同学落地C端产品，也在集团内进行开源，为生产、销售、员工工作等各环节赋能。

自有数据驱动深度优化，隐私安全与效率兼顾。

- 数据上，小米自身挖掘整理的训练数据占比达到了80%，其中小米自有的产品和业务数据量达到3TB（截至2023.8）；
- 效率和效果上，小米根据对Transformer结构的理解进行改良，充分考虑设备端芯片的特色要求，合理设置模型的宽度和深度；
- 训练策略上，采用小米提出的ScaledAdam优化器和Eden学习率调度器，显著提升收敛速度的同时减少了优化器中显存的浪费，实现“轻量化”；
- 隐私上，模型部署到端侧后，信息不用上传到云端，所有计算都在本地进行，保证用户隐私不被泄露。

图3：小米大语言模型 MiLM 通过大模型备案

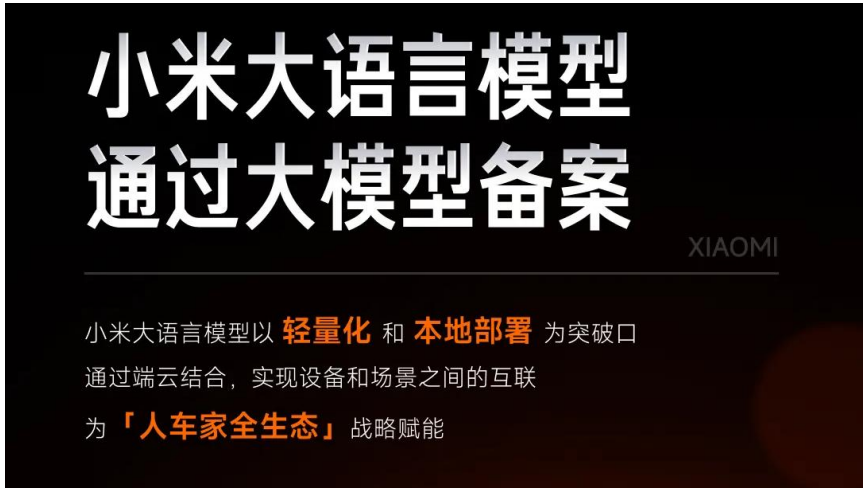


图4：小爱同学落地C端产品

资料来源：小米公司公众号，小米技术公众号，国信证券经济研究所整理

小米MiLM2升级发布：性能与技术全面提升，云边端协同优化

国信证券 GUOSEN SECURITIES

2024年11月，小米大模型已经实现了从一代到二代（MiLM2）的升级迭代。主要升级包括：

- 丰富了模型的参数矩阵，参数规模同时向下和向上扩充，实现了云边端结合，参数尺寸最小为0.3B，最大为30B；
- 在10大能力维度上，相比于第一代模型平均提升超过45%，其中指令跟随、翻译、闲聊等关键能力处于业界前列；
- 在端侧部署上支持3种推理加速方案，包括大小模型投机、BiTA、Medusa，并且量化损失降低78%；
- 支持的最长窗口为200k（第一代为4k），在长文本评测中，效果处于业界前列。

小米大模型团队在预训练、后训练、量化、推理加速等方向做了大量的技术探索和创新。

- SUBLLM模型结构：模型能够区分重要token和不重要token，训练和推理速度分别提升34%和52%。
- TransAct大模型结构化剪枝方法：目的是同时实现高度压缩和较小损失，KV Cache下降了50%，推理速度提升了20%（小米14手机测试）。
- INTRADoc注意力机制：通过屏蔽无关文档，让每个token的概率仅取决于同一文档中的上文信息。
- Mixture of Diverse Size Experts：每一层中设计大小不同的专家结构，并引入了一种专家对分配策略，以在多个 GPU 之间均匀分配工作负载。

图5：SUBLLM架构，提高训练和推理速度

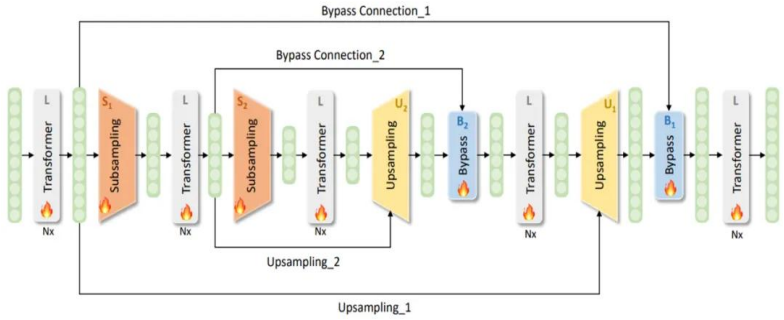
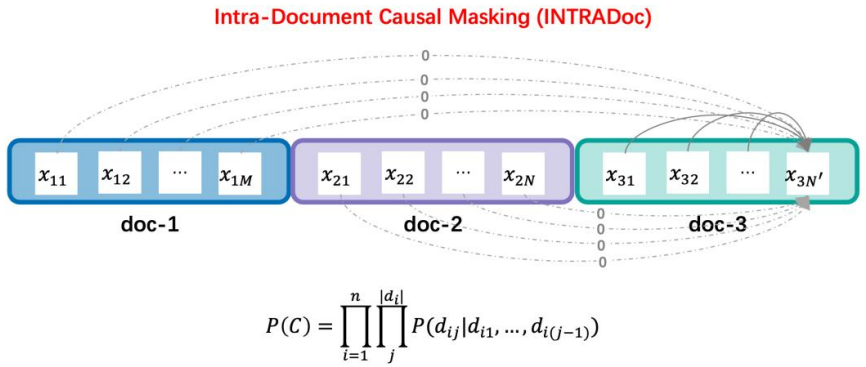


图6：INTRADoc注意力机制，消除潜在干扰信息



资料来源：小米技术公众号，国信证券经济研究所整理

MiLM2实力进阶，二代效果全方位提升

小米大模型团队采用自主构建的通用能力评测集Mi-LLMBM2.0，对最新一代的MiLM2模型进行了全方位评估。该评测集涵盖了广泛的应用场景，包括生成、脑暴、对话、问答、改写、摘要、分类、提取、代码处理以及安全回复等10个大类，共计170个细分测试项。以MiLM2-1.3B模型和MiLM2-6B模型为例，对比去年发布的一代模型，在十大能力上的效果均有大幅提升，平均提升幅度超过45%。

小米的「人车家全生态」战略，旨在构建一个涵盖人、车、家等多元化生活场景的超级智能生态系统。在这个系统内，实时交互成为常态，每时每刻都需要精确对接用户千差万别的个性化需求，这对于大模型的生成、闲聊、翻译等能力提出了更高的要求。在这些关键能力上，MiLM2-6B模型的评测成绩十分优异，对比业内同参数规模模型也有较优的效果。

图7：MiLM2模型效果显著提升

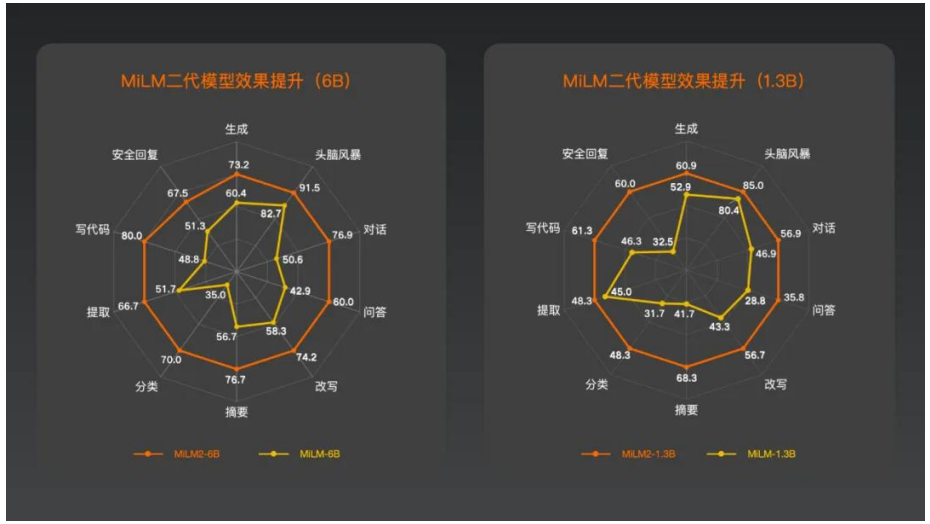
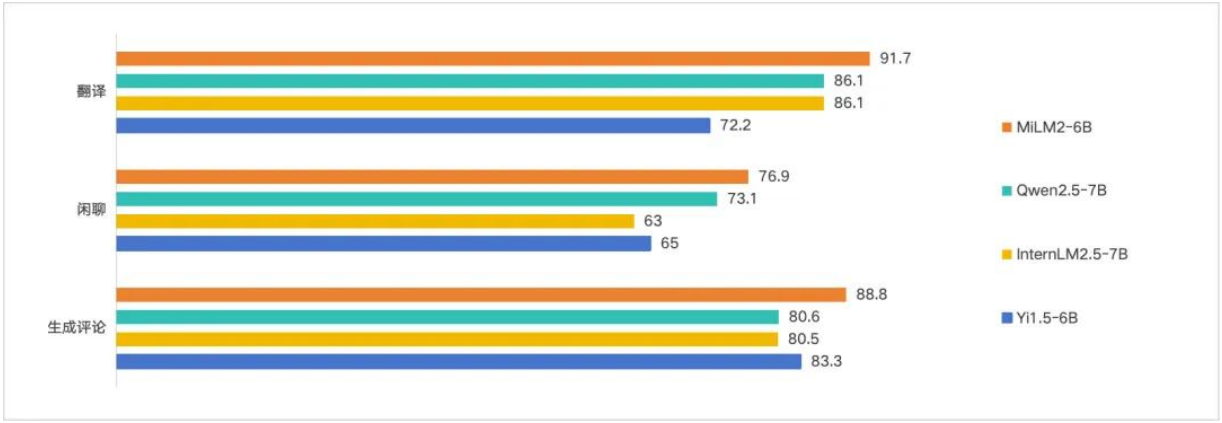


图8：MiLM2翻译、闲聊、生成评论效果与同行业对比



资料来源：小米技术公众号，国信证券经济研究所整理

MiLM2大模型矩阵灵活扩展，充分适配多元化场景

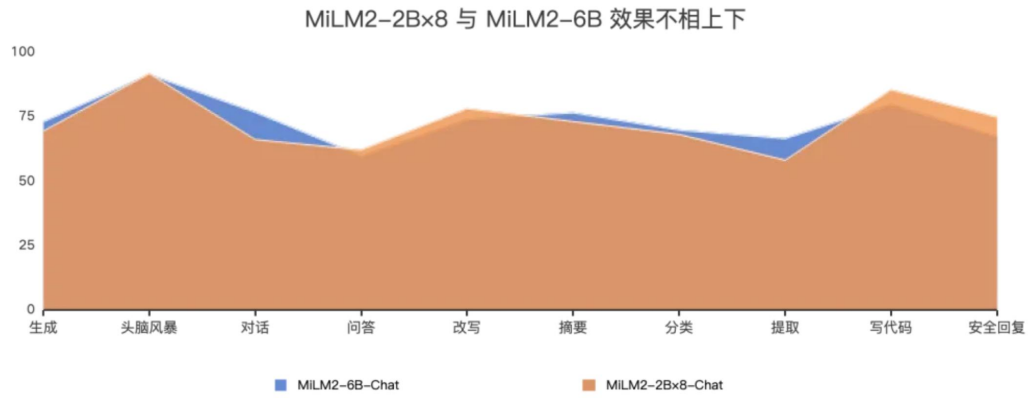
在坚持轻量化部署的大原则下，小米自研大模型团队构建并不断扩充了自研大模型的模型矩阵，将大模型的参数规模灵活扩展至0.3B、0.7B、1.3B、2.4B、4B、6B、13B、30B等多个量级，充分考虑多元化的业务场景及资源限制。

- 0.3B~6B：终端（on-device）场景，应用时通常是一项非常具体的、低成本的任务，微调后可以达到百亿参数内开源模型效果。
- 6B、13B：支持多任务微调，微调后可以达到几百亿开源模型的效果。
- 30B：云端场景，具备坚实的zero-shot/上下文学习或一些泛化能力，模型推理能力较好，能够完成复杂的多任务。

小米自研大模型矩阵不仅包含多样的参数量级，同时也纳入了各种不同的模型结构。在二代模型系列中，大模型团队加入了两个MoE结构的模型：MiLM2-0.7B×8和MiLM2-2B×8。以MiLM2-2B×8为例，根据评测结果，该模型在整体性能上与MiLM2-6B不相上下，而解码速度实现了50%的提升，提升了其运行效率。

图9：MiLM2模型矩阵，参数规模灵活扩展

图10：MiLM2-2B×8与MiLM2-6B效果不相上下



资料来源：小米技术公众号，国信证券经济研究所整理

端云并重：4B模型端侧落地，30B模型云端部署

国信证券 GUOSEN SECURITIES

端侧新增4B模型：小米的大模型团队在端侧部署方面取得了显著进展，成为业界首个在移动设备上成功运行1.3B和6B大模型的公司。小米大模型团队创新性地提出了“TransAct 大模型结构化剪枝方法”，仅用8%的训练计算量即从6B模型剪枝了4B模型，训练效率大大提升；同时小米大模型团队自研了“基于权重转移的端侧量化方法”和“基于Outliers分离的端侧量化方法”，大幅降低了端侧量化的精度损失。MiLM2-4B模型总共40 层，实际总参数量为3.5B，目前已经实现在端侧部署落地。

云端新增30B模型：MiLM2-30B 模型是小米二代大模型系列中参数量级最大的模型，专为云端场景设计 。在云端环境中，大模型需要更高效地遵从并执行用户的复杂指令，深入分析多维度任务，并在长上下文中精准定位信息。针对这些重点目标，大模型团队选择了一系列开源的评测集，对 MiLM2-30B 模型的专项能力进行评估。结果表明，MiLM2-30B 模型在指令遵循、常识推理和阅读理解能力方面均有超越主流竞品的出色表现。

图11：MiLM2端侧能力与同行业比较

基座模型	通用			数学		推理			代码
	GPQA	BBH	WinoGrande	GSM8K	MATH	DROP	MULTI-NLI	WorldSense	MBPP-Plus
MiLM2-4B	29.29	60.06	71.27	78.17	43.64	52.79	56.59	39.13	44.61
Qwen2.5-3B	26.30	56.30	71.10	79.10	42.60	51.69	44.34	33.12	49.4
Llama3.2-3B	27.27	47.28	69.22	33.81	7.96	47.6	43.33	29.55	38.1

- Qwen2.5-3B：Qwen2.5-LLM: Extending the boundary of LLMs (GPQA, BBH, Winogrande, GSM8K, MATH, MBPP-Plus) and OpenCompass (DROP, MULTI-NLI, WorldSense)
- Llama3.2-3B：Evaluate with OpenCompass and Llama-3.2-3B model weight

图12：MiLM2云侧能力与同行业比较

模型	FollowBench-zh		IFEval	
	HSR Avg (%)	SSR Avg (%)	EN (Instruction Strict ACC)	ZH (Instruction Strict ACC)
MiLM2-30B-Chat	62.7	67.9	86.0	86.1
Qwen2.5-32B-Instruct	55.8	62.2	84.9	82.5
GPT-3.5	55.2	65.2	66.9	62.2
GPT-4	73.5	78.6	85.4	-

- GPT-3.5和GPT-4：FollowBench，IFEval (Zhou et al., 2023)
- Qwen2.5-32B-Instruct结果采用FollowBench和IFEval官方代码测试

资料来源：小米技术公众号，国信证券经济研究所整理

小米第二代大模型应用落地，赋能设备与提升业务效率

国信证券 GUOSEN SECURITIES

小米大模型全面赋能各类设备，提升用户体验与生活便利性。

- 手机端支持AI图片编辑，用户可进行动漫风转换、背景替换等操作；文档问答与智能成片功能帮助快速提取要点和自动剪辑个性化视频。
- 平板在办公和学习场景中提供强大支持，可自动生成会议纪要框架、行业报告等，提高工作效率。
- 电视端提供影视问答、健身休闲和家庭计划等功能，丰富娱乐体验的同时增强健康和家庭生活的管理。
- 汽车端具备语音深度控车、智能导航、旅行助手等功能，提升驾驶安全和出行便捷性。

目前，小米第二代自研大模型取得的进步和成果，已经开始渗透到真实的业务场景与用户需求中，不仅帮助集团内部解决了多样化的业务需求、实现工作提效，也已经在澎湃OS、小爱同学、智能座舱、智能客服中开始应用落地。

图13：小爱赋能手机

图14：小爱赋能汽车

资料来源：小米公司公众号，国信证券经济研究所整理

相关合作标的梳理

业务领域	相关个股
IDC	世纪互联
训练数据集销售服务	拓尔思
服务器 GPU	英伟达、寒武纪、海光信息（合作待定）
手机芯片	高通、联发科、芯海科技、慧智微、唯捷创芯（射频）、艾为电子（音频/马达等）、圣邦股份（电源/模拟开关等）
IOT SoC	高通、汇顶科技、瑞芯微、恒玄科技（手表/手环芯片）、炬芯科技（手表芯片）、全志科技（音箱、CyberDog机器人等）
手机硬件组件等	比亚迪电子、华勤技术、舜宇光学、瑞声科技、思特威、乐鑫科技
小米汽车	龙旗科技、无锡振华、经纬恒润、模塑科技、蓝思科技
小米家电	TCL、创维数字、晶晨股份、视源股份
AI办公软件及服务	金山办公
测试技术和软件服务	慧博云通
独家供应小米首款机器人所需软板	弘信电子

资料来源：国信证券经济研究所整理

风险提示

- AI应用落地不及预期、市场需求不及预期、行业竞争加剧、宏观经济波动。

资料来源：国信证券研究所整理

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032