证券研究报告

政策与技术共振，AIGC赋能千行百业释放价值潜能

——计算机行业2023年中期策略报告

行业评级

计算机 强于大市（维持）

证券分析师

闫磊 投资咨询资格编号：S1060517070006

付强 投资咨询资格编号：S1060520070001

2023年6月16日

请务必阅读正文后免责条款

投资逻辑图

ChatGPT火爆出圈

- 2022年11月，由OpenAI开发的大模型聊天机器人ChatGPT火爆出圈，短短5天，注册用户就超过100万，仅仅两个月月活用户数已经破亿
- 2023年3月，OpenAI发布GPT-4，GPT-4可以接受图片作为输入。相比ChatGPT，GPT-4在多模态处理能力、高端推理能力等方面提升明显
- 2023年5月，OpenAI在官网宣布，在美国推出苹果手机版本的ChatGPT应用程序

我国重视和支持AIGC产业发展

- 2023年4月，国家互联网信息办公室起草了《生成式人工智能服务管理办法（征求意见稿）》
- 2023年4月，中央政治局召开会议，指出要重视通用人工智能发展，营造创新生态，重视防范风险。
- 2023年5月，北京市发布《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）》和《北京市促进通用人工智能创新发展的若干措施》。上海市发布《上海市加大力度支持民间投资发展若干政策措施》。深圳市发布《深圳市加快推动人工智能高质量发展高水平应用行动方案（2023-2024年）》
- 2023年6月，国务院办公厅发布《国务院2023年度立法工作计划》，《人工智能法》已列入立法计划

我国大模型技术快速发展

算力底座支撑

- ✓ 无论从全球市场还是国内市场来看，浪潮信息AI服务器市场占有率都稳居第一
- ✓ 国内企业字节跳动、腾讯、阿里巴巴、百度等是全球AI服务器的主要采购方
- ✓ 海光信息等国内AI芯片企业的主打或在研产品性能已经可以对标全球主流AI芯片产品（如英伟达A100芯片）

大模型百花齐放

- ✓ 智谱AI、百度、阿里、科大讯飞等公司陆续推出自研大模型
- ✓ 智谱AI推出ChatGLM-130B和ChatGLM-6B，已获得较为广泛的应用
- ✓ 百度文心一言，在文学创作、商业文案创作、数理推算、中文理解、多模态生成等五大应用场景展现了强大能力
- ✓ 阿里通义千问是一个超大规模的语言模型，功能包括多轮对话、文案创作、逻辑推理、多模态理解、多语言支持等
- ✓ 科大讯飞星火大模型具有七大核心能力，其中部分能力已超越ChatGPT

C端和B端应用前景广阔

- ✓ **AIGC+办公：**生成会议纪要、对话问答、总结分析、智能改写等
- ✓ **AIGC+传媒：**客服问答、新闻撰写、营销文本生成、文案生成、内容续写等
- ✓ **AIGC+金融：**风险识别、客户信用能力评估、知识检索等
- ✓ **AIGC+教育：**精准教学、自动评阅、科学管理、个性化学习等
- ✓ **AIGC+医疗：**医疗器械、医疗服务、医药研发、医疗管理等
- ✓ **AIGC+汽车：**“大模型+智能座舱”提供不同车型的智能语音交互服务，能够回答与智能汽车相关的多种问题，让汽车驾驶更加智能

投资要点

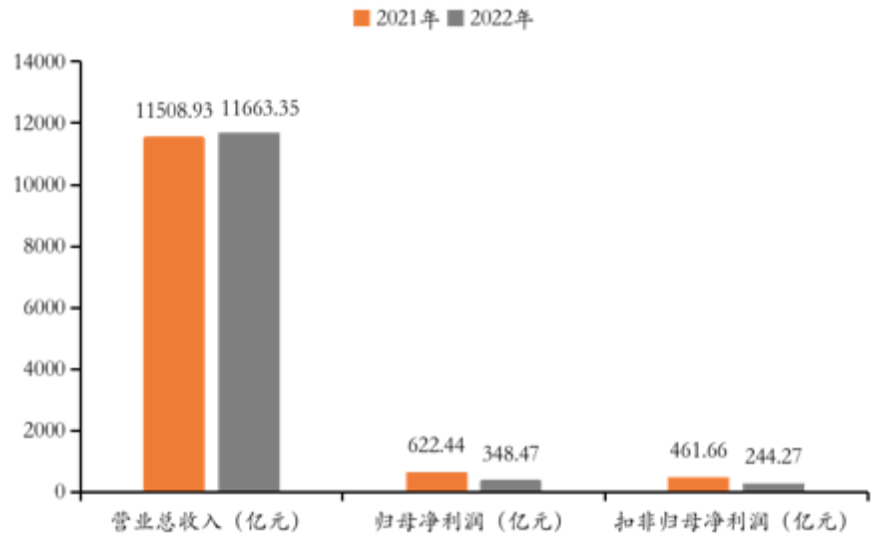
- **行业回顾及投资逻辑：**受项目实施进度放缓、验收延迟以及2021年上市公司员工人数增长较快等因素的影响，计算机行业上市公司2022年归母净利润承压明显。2023年一季度，行业上市公司业绩表现依然不佳，营收同比微幅下降1.24%，扣非归母净利润大幅下降99.85%。市场表现来看，计算机指数表现良好，涨幅跑赢沪深300指数，且在31个申万一级行业排名靠前，估值已处于历史较高水平。展望下半年，我们判断，在政策和技术共振的推动下，我国AIGC产业未来发展前景广阔。我们坚定看好AIGC产业链的投资机会。维持对计算机行业的“强于大市”评级。
- **算力：大模型需要大算力，AI芯片和服务器市场迎来发展机遇。**ChatGPT的火爆出圈，带动了整个AIGC产业的发展。我国十分重视和支持AIGC产业的发展，近期相关利好政策相继出台，为AIGC的后续发展护航。大模型的发展需要大算力，根据OPENAI数据，训练GPT-3 175B的模型，需要的算力高达3640 PF-days。我们认为，随着全球和中国人工智能厂商布局大模型，大模型将为全球和中国AI芯片和AI服务器市场的增长提供强劲动力。根据我们的估算，大模型将为全球和中国AI服务器带来约891.2亿美元和338.2亿美元的市场空间。市场空间巨大，相关芯片和服务器厂商将深度受益大模型的发展浪潮。
- **算法：我国大模型快速发展，已初步具备商用能力。**当前，以OpenAI、Google、Meta等为代表的国外厂商凭借先发优势、算力优势、以及数据集等方面的优势，在全球GPT大模型领域具有领先优势。但我国大模型产品也在快速发展，正奋起直追，尤其2023年3月以来，多家厂商推出了自研的通用大模型，某些功能已可比肩ChatGPT。同时，国产大模型在各行业的应用以及生态建设也取得积极进展。我们认为，我国大模型虽然相比GPT-4或仍有一定差距，但在短期内达到或接近ChatGPT的水平是可以预期的。我国大模型的小型化技术已经比较成熟，可助力实现大模型在各细分场景的应用落地。我国大模型产品已经初步具备商用能力。
- **应用：大模型赋能千行百业，AIGC未来发展前景广阔。**当前，人工智能在我国各行业已经得到广泛应用。随着大模型时代的到来，我国多家互联网企业的掌门人均表示，大模型时代，所有的产品都值得用大模型重做一次。根据中国互联网络信息中心（CNNIC）数据，截至2022年12月，我国网民规模达10.67亿，互联网普及率达75.6%。我们认为，随着国产大模型的逐步成熟，在政策与技术的共振下，我国大模型产品面向我国庞大的互联网C端用户群和丰富的行业应用场景，将与产品和应用场景深度融合，赋能我国数字经济的发展。参考我国数字经济的巨大体量，我国AIGC产业未来发展前景广阔。
- **投资建议：**展望下半年，我国大模型产品已经初步具备商用能力。我国北上深三地利好通用人工智能发展政策的发布，彰显了我国对于AIGC发展的重视和支持，同时将为我国其他城市发布类似政策带来示范效应。随着《人工智能法》列入《国务院2023年度立法工作计划》，我们判断，后续政策的出台将为我国AIGC产业的发展护航。在政策与技术的共振下，我国AIGC产业未来发展前景广阔。AIGC产业的发展需要大算力，我国AI芯片和AI服务器市场迎来发展机遇，相关芯片和服务器厂商将深度受益。我们坚定看好AIGC产业链的投资机会。维持对计算机行业的“强于大市”评级。在标的方面：1）算力方面，推荐浪潮信息、中科曙光、紫光股份、海光信息、龙芯中科，建议关注工业富联、寒武纪、景嘉微；2）算法方面：推荐科大讯飞，建议关注三六零；3）应用方面，推荐金山办公，建议关注拓尔思、彩讯股份、航天宏图；4）网络安全方面，强烈推荐启明星辰，推荐深信服、绿盟科技。
- **风险提示：**1）合规风险上升。2）技术创新不及预期。3）供应链风险。

CONTENT
目录

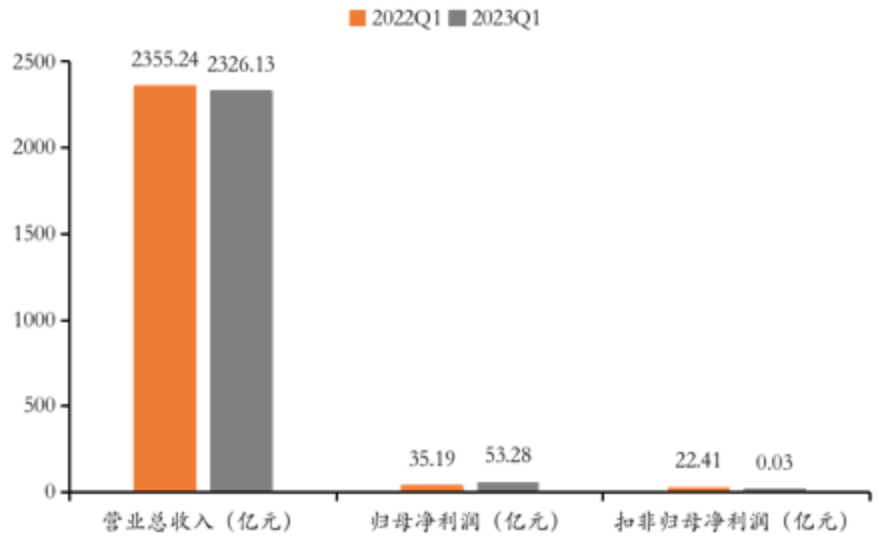
基本面：2022年整体业绩承压，2023Q1扣非利润表现不佳

- 受项目实施进度放缓、验收延迟以及2021年上市公司员工人数增长较快等因素的影响，计算机行业上市公司2022年归母净利润承压明显。在剔除ST等表现异常的个股之后，2022年，行业营业总收入合计为11663.35亿元，同比增长1.34%；归母净利润合计为348.47亿元，同比大幅下降44.02%。2023Q1，行业营业总收入合计为2326.13亿元，同比下降1.24%；归母净利润合计为53.28亿元，同比增长51.43%，主要是因为部分公司非经常性损益较大；扣非归母净利润合计为0.03亿元，同比大幅下降99.85%，扣非归母净利润表现不佳。

2021-2022年计算机行业上市公司业绩情况



2022Q1-2023Q1计算机行业上市公司业绩情况



数据来源：Wind，平安证券研究所

行情回顾：行业行情虽有震荡，但整体处于上行态势

- 年初以来，受信创、数字中国、数据要素、AIGC等主题的推动，计算机行业市场关注度高，行业行情整体处于上行态势。1月至4月初，在行业市盈率处于历史低位的背景下，受信创、数字中国、数据要素、AIGC等主题的推动，行业指数在4月初到达年初至今的高点。之后，随着AIGC合规风险的上升，叠加行业上市公司整体2022年业绩和2023年一季度业绩表现不佳的影响，行业指数高位回落。5月25日，全球AI芯片龙头公司英伟达发布2024财年第一财季财报，其对2024财年第二财季营收的乐观预期，加强了市场对于AI算力需求增长的确定性，提升了市场对于AI算力乃至整个AIGC主题的信心与关注度。5月30-31日，北京市、上海市、深圳市相继发布政策利好通用人工智能发展，彰显我国对于AIGC发展的重视与支持。受益市场对AIGC信心的提振，行业行情开启反弹。

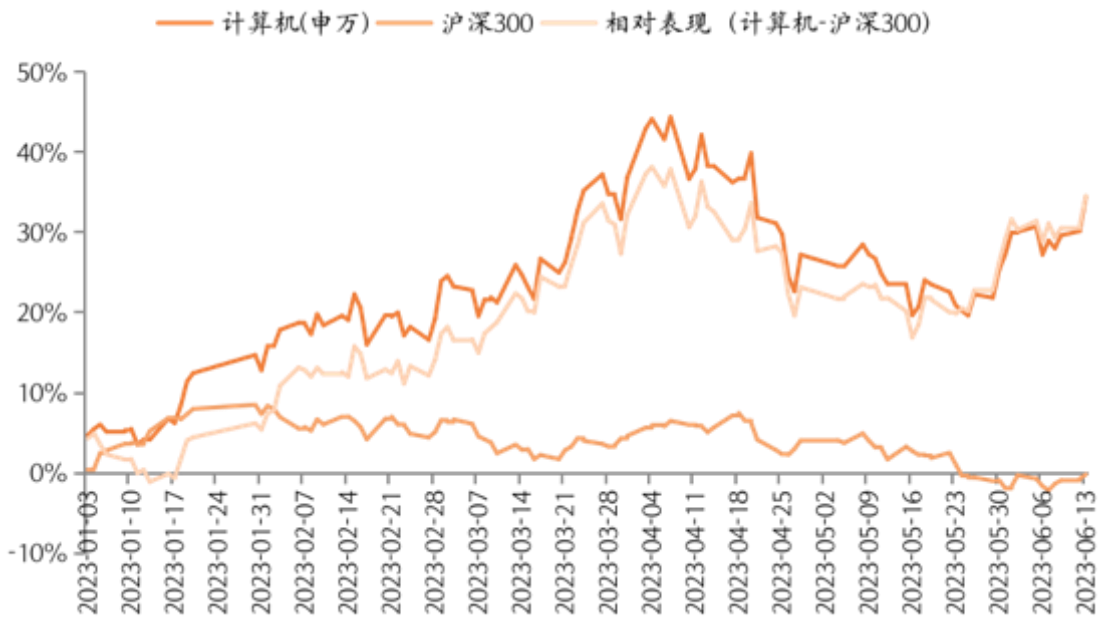
年初以来计算机行业指数表现

数据来源：Wind，平安证券研究所

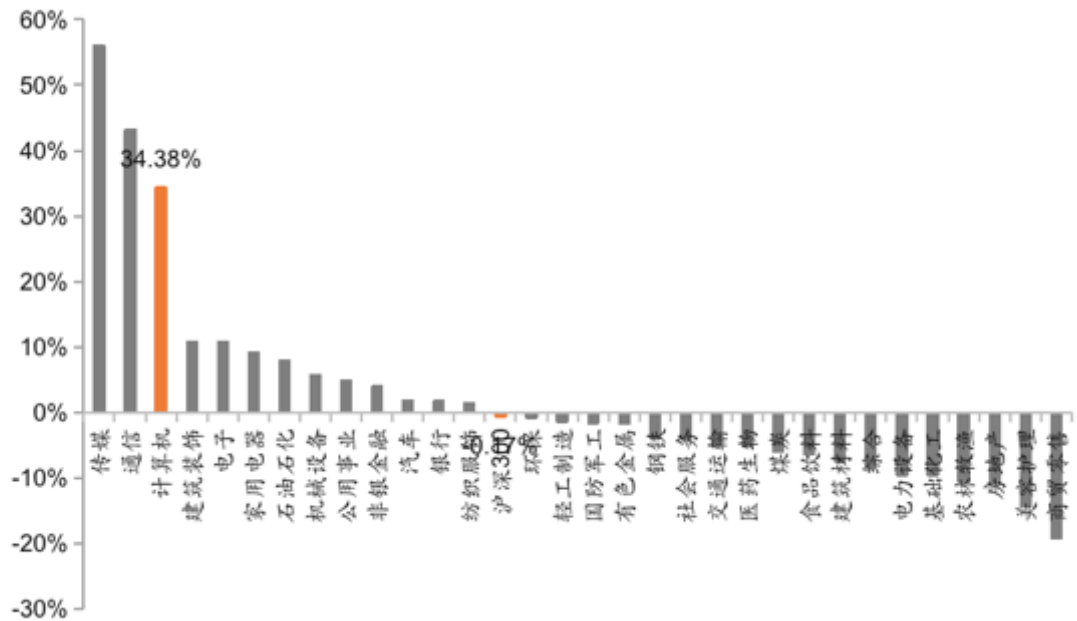
行情回顾：行业指数年累计涨幅跑赢沪深300，位列第3位

- 截至到2023年6月13日，申万计算机指数上涨了34.38%，跑赢沪深300指数34.55个百分点，在31个申万一级行业中排名第3位，排名靠前。

行业指数相比沪深300指数表现



行业指数相比沪深300指数表现



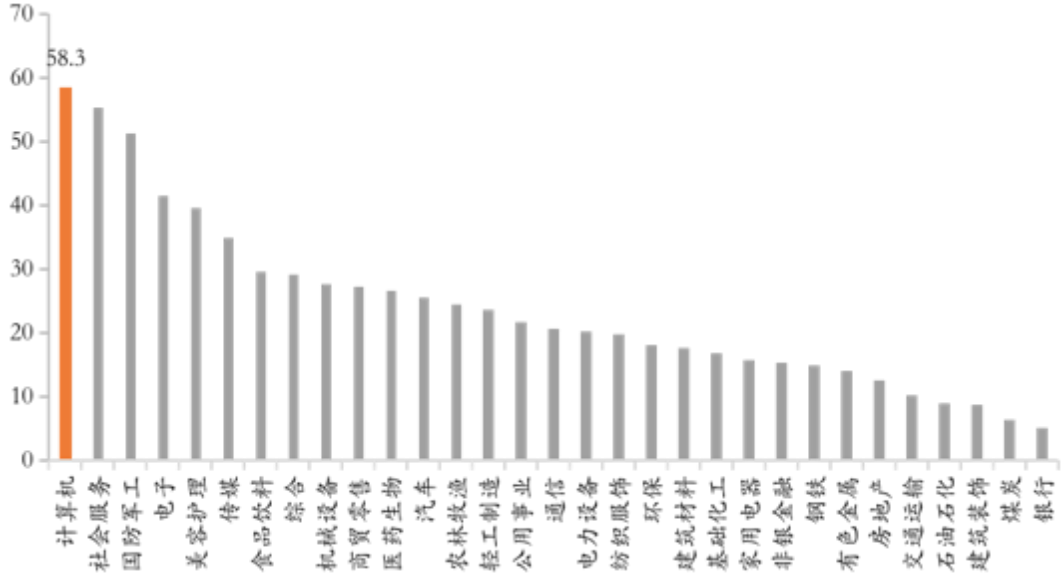
数据来源：Wind，平安证券研究所

估值水平：估值处于历史较高水平，我们坚定看好AIGC产业链的投资机会

- 截至6月13日，计算机行业估值处于历史较高水平。根据我们的统计，2015年以来，申万计算机行业历史市盈率（TTM整体法，剔除负值）中位数为47.9倍。计算机行业6月13日市盈率（TTM整体法，剔除负值）为58.3倍，在历史市盈率中位数水平之上，处于历史82%分位。
- 展望下半年，我们判断，我国北上深三地利好通用人工智能发展政策的发布，彰显了我国对于AIGC发展的重视和支持，同时将为我国其他城市发布类似政策带来示范效应。随着《人工智能法》列入《国务院2023年度立法工作计划》，后续政策的出台将为我国AIGC产业的发展护航。在政策与技术的共振下，我国AIGC产业未来发展前景广阔。我们坚定看好AIGC产业链的投资机会。维持对计算机行业的“强于大市”评级。

计算机行业当前估值高于历史中位数水平

行业市盈率在31个申万一级行业排名第1位



数据来源：Wind，平安证券研究所

CONTENT
目录

ChatGPT引领AI发展新浪潮，国内政策加码助推通用人工智能发展

- 2022年11月，由OpenAI开发的大模型聊天机器人ChatGPT火爆出圈，短短5天，注册用户就超过100万，仅仅两个月月活用户数已经破亿。ChatGPT的火爆出圈，带动了整个AIGC产业的发展。AIGC(AI Generated Content)即人工智能生成内容，也称为生成式AI，AIGC实现了从分析内容到创造生成新内容的跨越，而模型、数据集、算力、应用是催生AI技术新范式的重要因素。
- 我国十分重视和支持AIGC产业的发展，近期相关利好政策相继出台。2023年4月，国家网信办起草了《生成式人工智能服务管理办法（征求意见稿）》。2023年5月，北上深三地相继发文，针对算力、算法、应用、监管等产业发展核心要素和关键环节提出具体措施以支持大模型的发展，国内通用人工智能发展的技术和商业环境将进一步优化。2023年6月，国务院办公厅发布《国务院2023年度立法工作计划》，《人工智能法》已列入立法计划。AIGC正处于发展初期，监管合规与新技术发展相辅相成，有效引导“技术向善”，为AIGC的后续发展护航。

我国AIGC发展相关政策（部分）

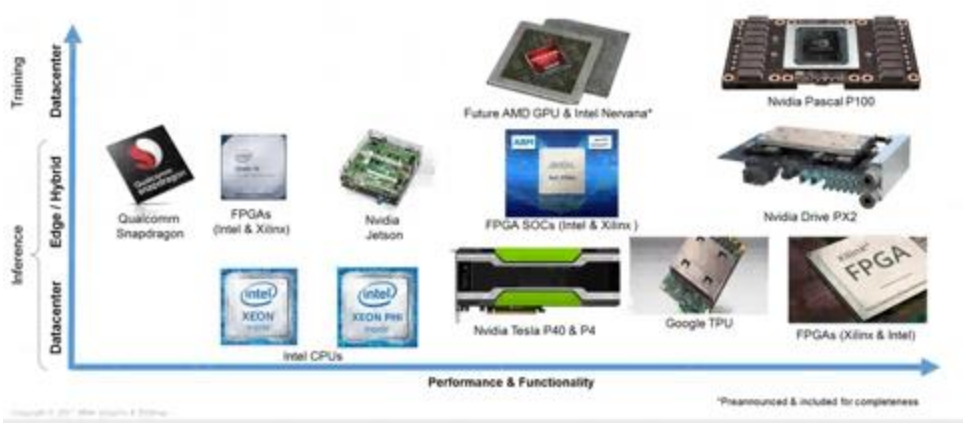
时间	部门	政策	相关内容
2023.2	中共中央、国务院	《数字中国建设整体布局规划》	系统优化算力基础设施布局，促进东西部算力高效互补和协同联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局。
2023.4	网信办	《生成式人工智能服务管理办法（征求意见稿）》	对生成内容、主体责任、数据源和数据处理等方面均做出了规定。
2023.4	中央政治局	中央政治局会议	要重视通用人工智能发展，营造创新生态，重视防范风险。
2023.5	北京市政府	《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）》	2025年，人工智能算力布局初步形成，国产人工智能芯片和深度学习框架等基础软硬件产品市场占比显著提升，算力芯片等基本实现自主可控
2023.5	北京市人民政府办公厅	《北京市促进通用人工智能创新发展的若干措施》	《若干措施》针对算力资源和数据要素供给能力的提升提出了具体的措施，如加强与云厂商的合作、加快算力中心和数据训练基地的规划建设等；提出要“系统构建大模型人工智能技术体系”、“推动通用人工智能技术创新场景应用”。
2023.5	上海市发改委	《上海市加大力度支持民间投资发展若干政策措施》	发挥人工智能创新发展专项等引导作用，支持民营企业广泛参与数据、算力等人工智能基础设施建设。
2023.5	中共深圳市委办公厅	《深圳市加快推动人工智能高质量发展高水平应用行动方案（2023-2024年）》	要打造全域全时场景应用，推进“千行百业+AI”，孵化高度智能化的生产机器人。
2023.6	国务院办公厅	《国务院2023年度立法工作计划》	人工智能法草案等预备提请全国人大常委会审议。

数据来源：政府网站，平安证券研究所

AI芯片进入舞台中央，广泛应用于训练或推理

- AI芯片是指所有能够用于人工智能的芯片，主要包括GPU、ASIC、FPGA三大类。AI芯片按应用场景可以分为训练芯片和推理芯片：训练芯片用于算法模型开发、训练，利用标记的数据，通过该芯片“学习”出具备特定功能的模型；推理芯片用于应用层，利用训练出来的模型加载数据，通过芯片计算“推理”出各种结论。按照部署的位置可以分为云端芯片和边缘端芯片：云端芯片部署在公有云、私有云或者混合云上，不但可用于训练，也可用于推理，算力强劲；边缘端芯片主要应用于嵌入式、移动终端等领域，此类芯片一般体积小、耗电低，性能要求也相对不高，一般只需具备一两种AI能力，用于推理。

人工智能芯片产品图谱



三类AI芯片简介

AI芯片	定义	优势	典型厂商
GPU	通用图形处理器	高并行结构，生态体系成熟，跨平台支持。易于编程，成为主流的并行数据处理加速器	英伟达、AMD、海光信息、摩尔线程
ASIC	专用集成电路	专门为深度学习计算定制的芯片，如神经网络处理器NPU、张量处理器TPU，效率高、功耗低、体积小	谷歌TPU、寒武纪、海思昇腾
FPGA	现场可编程逻辑阵列	高度并行的结构和低延迟；可编程和灵活性强，能够适应模型算法迭代	赛灵思、英特尔Altera

AI芯片分类

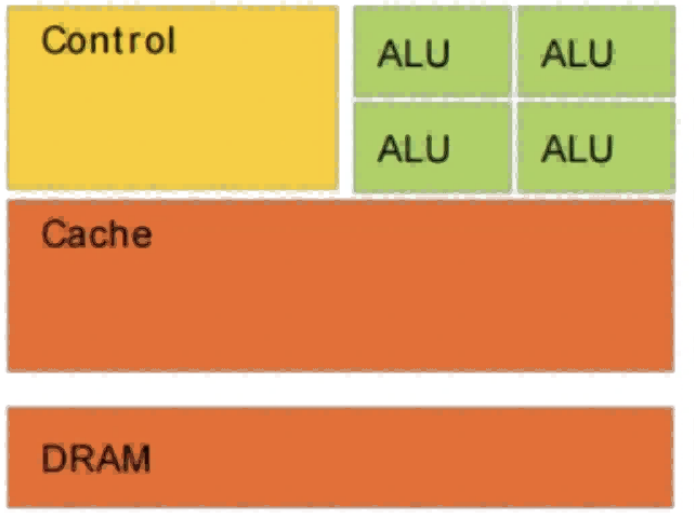
芯片	训练	推理
云端	GPU FPGA ASCI	CPU+GPU异构 FPGA ASIC（TPU等）
边缘端	无法完成训练工作	GPU FPGA ASIC

数据来源：IDC，平安证券研究所

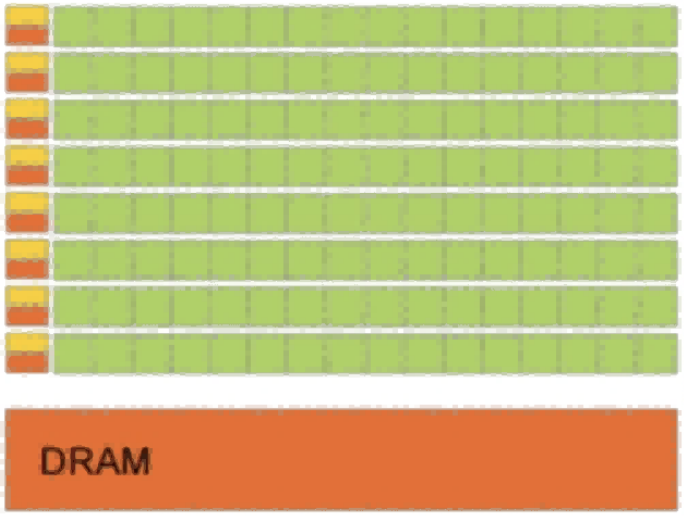
GPU大规模并行运算优势明显，是AI芯片市场的首选

- GPU即图形处理器，主要分为传统GPU和GPGPU，分别用于图形渲染和通用计算，用于AI服务器的GPU一般指后者。GPU中超过80%部分为运算单元（ALU），擅长大规模并行运算，主要应用于PC、服务器、数据中心、自动驾驶等领域，在数据中心被广泛应用于AI的训练、推理高性能计算等场景，是国内数据中心加速服务器市场的首选。据IDC统计，2021年，GPU在我国AI服务器加速芯片市场占有90%以上的份额。但未来随着非GPU芯片逐渐增多，IDC预计，2025年其他非GPU芯片占比将超过20%。
- GPU市场目前仍由英伟达、AMD等国外厂商主导，国内正处于发展起步阶段，在AI芯片市场的竞争力较弱，未来在大模型技术发展的催化下，叠加美国限制向中国出口高端GPU芯片等因素，国产GPU芯片将迎来发展机遇。

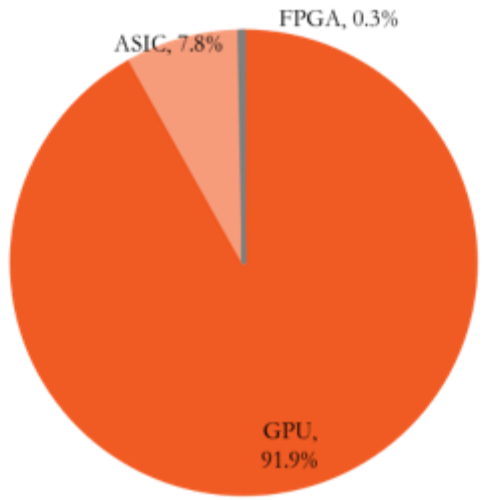
CPU架构



GPU架构



2021年我国AI服务器加速芯片市场份额



数据来源：IDC，平安证券研究所

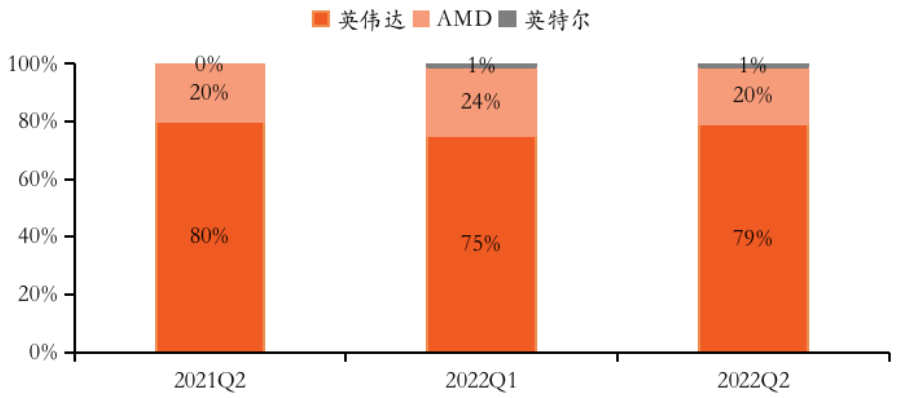
全球GPU市场竞争格局：英伟达独占鳌头，AMD跟随其后

- 英伟达是全球人工智能计算领域的领导者。A100和H100是英伟达在AI芯片方面的拳头产品。H100芯片于2022年初发布，并于当年9月量产，H100集成了800亿个晶体管，性能较上一代提升了一个数量级。英伟达是全球GPU芯片市场的绝对龙头。根据市场研究机构Jon Peddie Research（简称JPR）的数据，2022年二季度，英伟达在全球独立GPU芯片市场占有率为79%，AMD以20%的市占率跟随其后。
- 作为全球GPU市场份额第二的厂商，AMD近期推出了最新款GPU芯片。2023年6月14日，AMD推出了针对AI的最新款处理器芯片MI300A和MI300X，后者为针对大语言模型进行优化的版本。相比上一代产品MIX250X，MI300在AI的算力/能耗方面预计实现8倍/5倍的优化；相比H100，MI300可以提供2.4倍的HBM（高带宽内存）密度，以及1.6倍的HBM带宽。相对公司自身的前期产品以及竞品，此次发布的芯片或平台较大幅度的提升了产品性能和能效。面对最大的竞争对手英伟达，AMD除了硬件性能追赶之外，还需要在软件平台上下更多的功夫，毕竟生态强健、功能完善和易用的软件平台也是用户的重要关注点。

英伟达、AMD主要GPU产品性能参数对比

性能参数	V100 PCIe	A100 80GB PCIe	H100 80GB PCIe	MI250X	MI300X
FP16（TFLOPS）		312	756.5	383	
FP32（TFLOPS）	14	19.5	51	47.9	-
FP64（TFLOPS）	7	9.7	26	47.9	
GPU 显存	32/16GB HBM2	80GB HBM2e	80GB HBM3	128GB HBM2e	192GB HBM3
GPU 显存带宽	900 GB/s	1935GB/s	2TB/s	3.2TB/s	5.2TB/s
最大热设计功耗（TDP）	250W	300W	300-350W	500W	-

全球独立GPU出货量市占率



数据来源：英伟达官网，AMD官网，JPR，平安证券研究所

国内GPU产品发展势头良好，产品性能已可对标英伟达主流产品

- 国内目前部署GPU赛道的厂商主要有海光信息、景嘉微、沐曦、壁仞科技、芯动科技等。当前，国内GPU产品发展势头良好，产品性能已可对标英伟达主流产品。以海光信息为例，海光信息DCU（Deep Computing Unit 深度计算器，是GPGPU通用图形处理的一种）产品具备强大的计算能力和高速并行数据处理能力，深算一号产品指标基本达到国际上同类型高端产品的水平，已成功实现商业化应用。对标NVIDIA A100产品，海光信息DCU单芯片产品基本能达到其70%的性能水平，同时，海光DCU产品的片间互联性能还有较大的提升空间。目前，海光信息第二代DCU产品-深算二号处于研发阶段。

海光信息DCU产品-深算一号规格特点

项目	深算一号（海光8100）
典型功耗	260-350W
典型运算类型	双精度、单精度、半精度浮点数据和常见整型数据
计算	（1）60-64个计算单元（最多4096个计算核心） （2）支持FP64、FP32、FP16、INT8、INT4
内存	（1）4个HBM2内存通道 （2）最高内存带宽1TB/s （3）最大内存容量为32GB
I/O	（1）16 Lane PCIe Gen4 （2）DCU芯片之间高速互联

海光信息DCU产品与行业典型可比产品参数对比

项目	海光	NVIDIA	AMD
品牌	深算一号	Ampere 100	MI100
生产工艺	7nm FinFET	7nm FinFET	7nm FinFET
核心数量	4096（64 CUs）	2560 CUDA processors	120 CUs
内核频率	Up to 1.5GHz（FP64） Up to1.7GHz（FP32）	Up to 1.53Hz	Up to 1.5GHz（FP64） Up to1.7GHz（FP32）
显存容量	32GB HBM2	80GB HBM2e	32GB HBM2
显存位宽	4096 bit	5120 bit	4096 bit
显存频率	2.0 GHz	3.2 GHz	2.4 GHz
显存带宽	1024 GB/s	2039 GB/s	1228 GB/s
TDP	350 W	400 W	300 W
CPU to GPU互联	PCIe Gen4×16	PCIe Gen4×16	PCIe Gen4×16
GPU to GPU互联	xGMI×2， Up to 184GB/s	NVLink， Up to 600GB/s	Infinity Fabric×3, Up to 276GB/s

数据来源：各公司公告，平安证券研究所

ASIC是一种定制芯片，可提供更高能效表现和计算效率

- ASIC（专用芯片）是一种为特定目的、面向特定用户需求设计的定制芯片，具备性能更强、体积小、功耗低、可靠性更高等优点。在大规模量产的情况下，还具备成本低的特点。ASIC芯片主要应用于深度学习加速，在推理侧，相较其他AI芯片在效率和速度方面具有明显优势。其中表现最为突出的ASIC就是谷歌2015年发布的TPU（张量处理芯片）和英特尔2022年发布的Gaudi 2。我国的AISC行业发展迅速，主要企业有寒武纪、澜起科技、黑芝麻、地平线、华为海思、阿里巴巴等，部分国产AISC技术已经达到国际领先，如在BF16浮点算力方面，华为海思的昇腾910已超过谷歌的最新一代产品TPUv4。

华为海思训练、推理芯片性能对比

芯片	昇腾Ascend910	昇腾Ascend310
功能	训练	推理
工艺	7nm	12nm
算力	INT8 640TOPS FP16 320TFLOPS	INT8 22TOPS FP16 11TFLOPS
功耗	310W	8W
内存	HBM2E	2*LPDDR4x

ASIC芯片在推理侧性能优势明显

	训练		推理		通用型	推理准确率
	销量	速度	效率	速度		
CPU	1x ba seline				很高	~98-99.7%
GPU	~10-100x	~10-1000x	~1-10x	~1-100x	高	~98-99.7%
FPGA	-	-	~10-100x	~10-100x	中	~95-99%
ASIC	~100-1000x	~10-1000x	~100-1000x	~10-1000x	低	~90-98%

数据来源：华为官网，CSET，平安证券研究所

寒武纪ASIC产品不断迭代，最新一代产品有望承接国内AI算力需求

- 寒武纪的第三代云端推训一体芯片思元370，最大算力高达256TOPS（INT8），是第二代产品思元270算力的2倍。此外，与市场主流同尺寸芯片相比，思元370系列加速卡在实测性能和能效方面表现出一定优势。公司的思元370芯片及加速卡与数家头部互联网企业完成适配工作，已经进入了批量销售环节；与金融、运营商等众多行业领域中的头部公司实现了批量销售或达成合作意向。思元590是寒武纪最新一代云端智能训练芯片，目前尚未正式发布。思元590采用MLUarch05全新架构，实测训练性能较在售旗舰产品有大幅提升，有望承接国内逐渐升级的AI算力需求。

MLU370系列加速卡规格

性能参数	MLU370-S4	MLU370-X4	MLU370-X8
制程工艺	7nm		
INT8	192 TOPS	256TOPS	
INT16	96 TOPS	256 TOPS	128 TOPS
FP16	72 TFLOPS	96 TFLOPS	
BF16	72 TFLOPS	96 TFLOPS	
FP32	18 TFLOPS	24 TFLOPS	
内存容量	24GB		48GB
内存带宽	307.2GB/s		614.4 GB/s
系统接口	x16 PCIe Gen4		
最大热设计功耗	75W	150W	250W

思元370芯片

MLU370-S4智能加速卡

MLU370-X4 智能加速卡

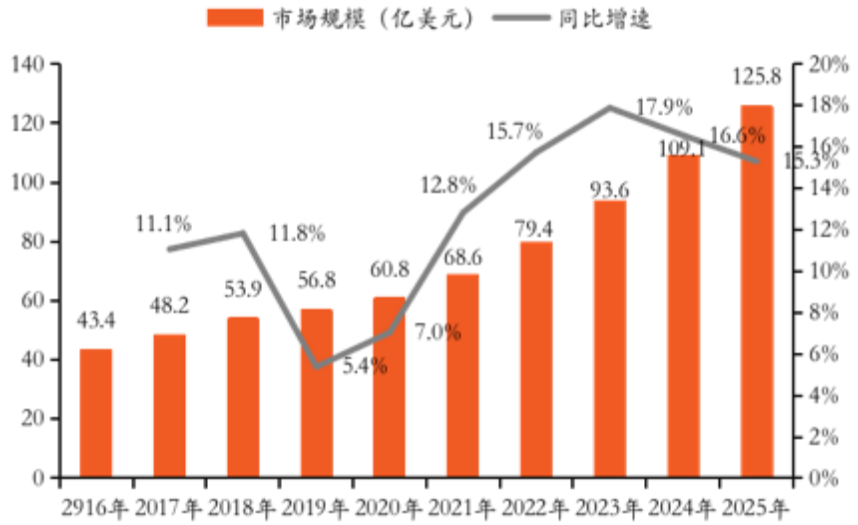
MLU370-X8 智能加速卡

数据来源：寒武纪官网，平安证券研究所

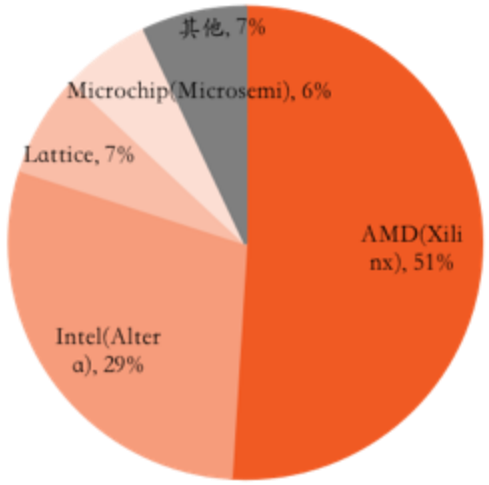
FPGA算力强、灵活度高，但技术难度大、国内差距较为明显

- FPGA（现场可编程门阵列）芯片集成了大量的基本门电路以及存储器，灵活性介于CPU、GPU和ASIC之间，在硬件固定之前，允许使用者灵活使用软件进行编程。FPGA相比与其他加速芯片，能更好得实现算力、成本与功耗之间的平衡。FPGA相比CPU和GPU，没有取指和译码操作，能耗比指标更优秀；相比ASIC开发周期更短、易用性更强。
- Frost&Sullivan数据显示，2021年全球FPGA市场规模已达68.6亿美元，2025年有望增长至125.8亿美元，年复合增长率约为16.4%。当前，全球FPGA市场被赛灵思（被AMD收购）、英特尔、莱迪思、微芯科技等海外巨头垄断，四家合计占据90%以上的市场份额。国内FPGA市场起步较晚，市场份额较小，但是安路科技、复旦微电子等国产厂商正在持续加速在FPGA的布局。

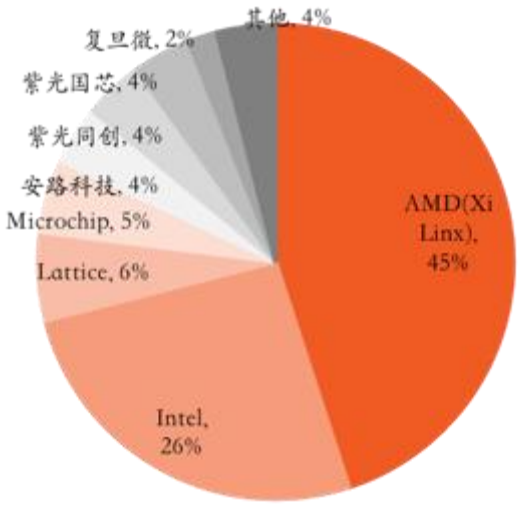
2016-2025年全球FPGA市场规模及预测



2021年全球FPGA市场竞争格局



2021年我国FPGA市场竞争格局

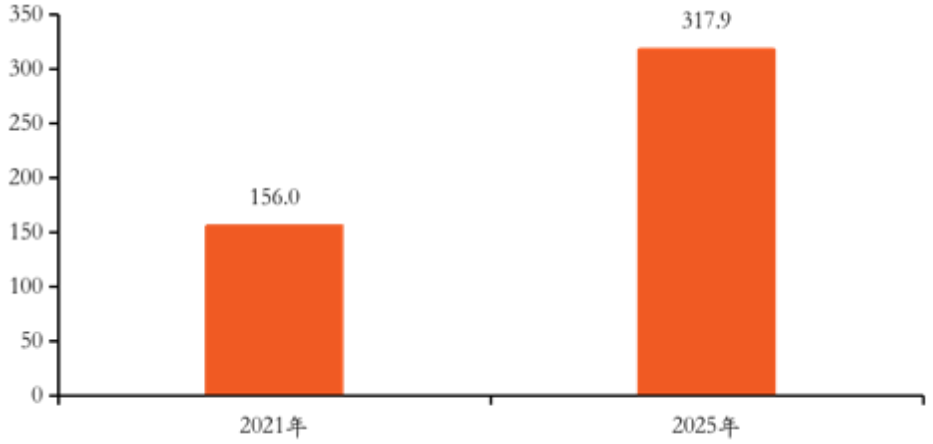


数据来源：Frost&Sullivan，安路科技公告，平安证券研究所

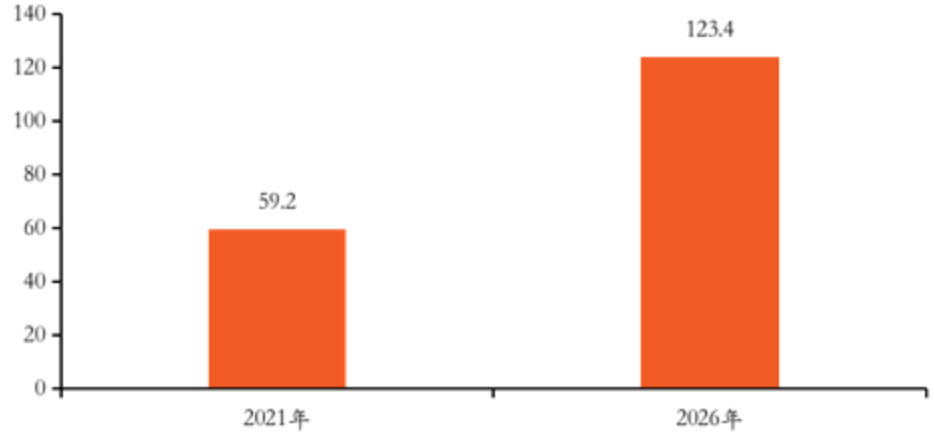
AI服务器借助加速卡获取强大算力，市场需求快速增长

- AI服务器是指能够提供人工智能（AI）的数据服务器，具有强大的图形处理和高性能计算能力，既可以用来支持本地应用程序和网页，也可以为云和本地服务器提供复杂的AI模型和服务，能支持多种常用的AI技术，如机器学习、自然语言处理、计算机视觉、生物信息分析等。AI服务器与普通服务器的区别主要在于计算架构的不同，AI服务器通常根据应用场景的不通，采用CPU+GPU/ASIC/FPGA或其他加速卡的异构式计算架构。
- 当前国内外AI服务器市场规模快速增长。据IDC统计，2021年全球AI服务器市场规模为156亿美元，预计2025年将达到317.9亿美元，2020-2025年的GAGR为19.5%。2021年我国AI服务器的市场规模为59.2亿美元，预计2026年将达到123.4亿美元，2021-2026年的CAGR为15.8%。

2021-2025年全球AI服务器市场规模（亿美元）



2021-2026年中国AI服务器市场规模（亿美元）

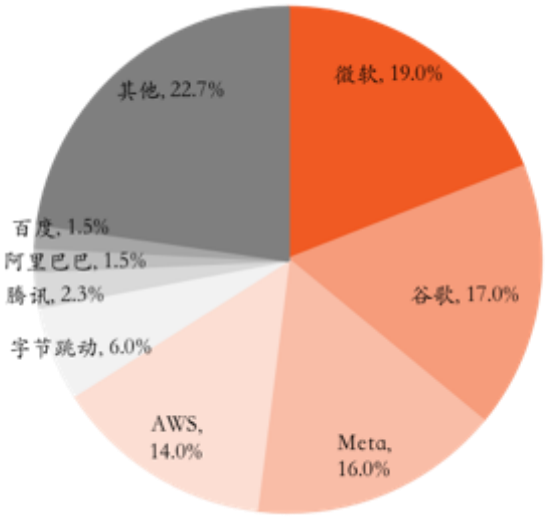


数据来源：IDC，平安证券研究所

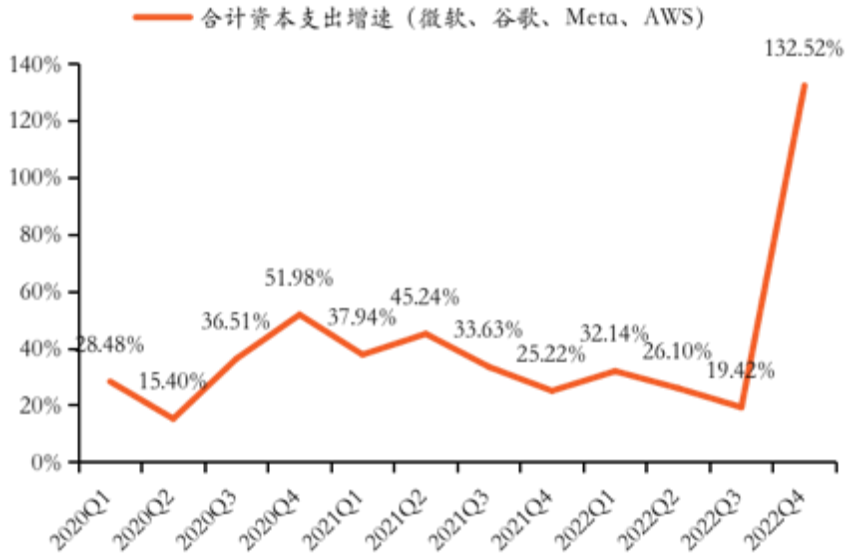
互联网云巨头贡献AI服务器主要需求，资本投入力度有望维持增长

- AI服务器市场的下游主要是以大型云计算厂商为主。TrendForce统计数据显示，2022年AI服务器采购量中，美国四家云厂商，微软、谷歌、Meta、AWS的采购量位居前四，合计占比约66%。国内企业字节跳动、腾讯、阿里巴巴、百度紧随其后，在AI基础设施方面的建设步伐较为领先，采购量合计占比约为11.3%。未来，随着AIGC、边缘计算、自动驾驶等新兴技术和应用的不断普及，各大云厂商有望持续加大在AI相关基础设施方面的投入，持续为AI基础设施市场注入发展动力。

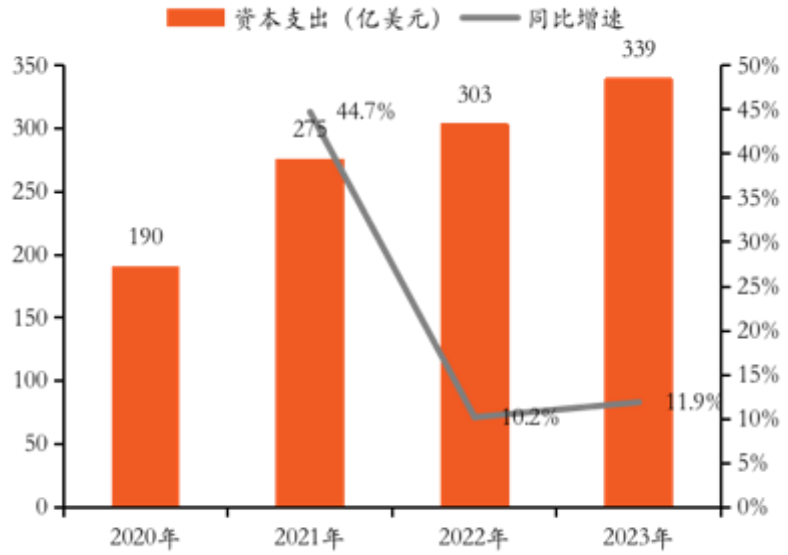
2022年AI服务器下游采购情况



2020年海外主要云厂商合计资本支出增速



2020-2023年我国云计算支出规模及增速

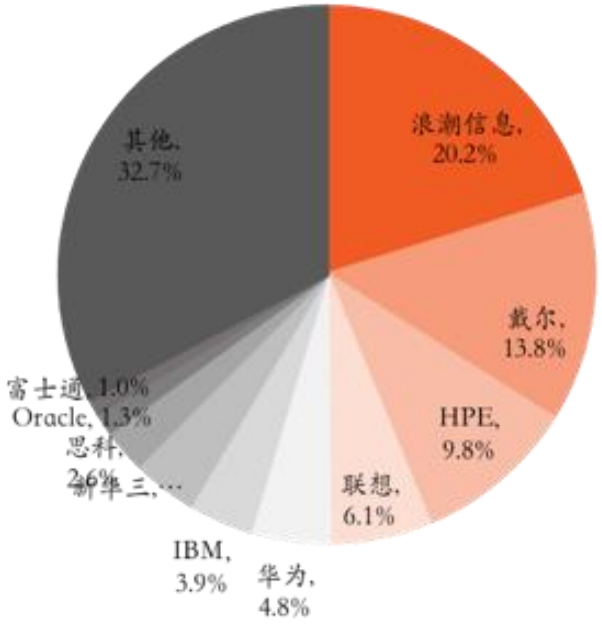


数据来源：TrendForce，平安证券研究所

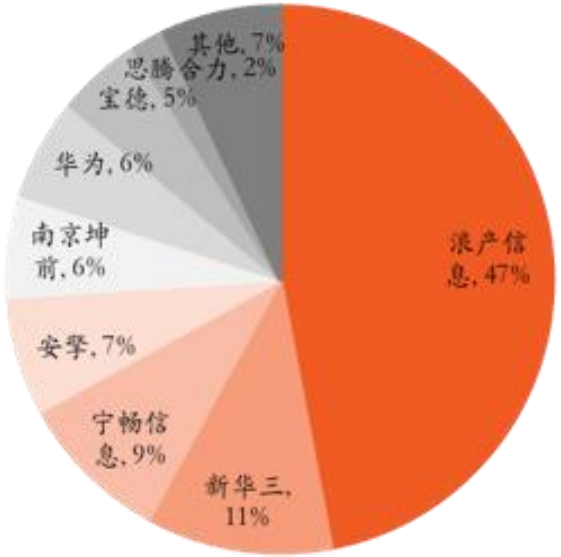
浪潮信息全球领跑，国产厂商在AI服务器领域大有可为

- 无论从全球市场还是国内市场来看，浪潮信息市场占有率都稳居第一，此外，新华三、宁畅、安擎等诸多国产厂商也正在加速推进人工智能基础设施产品的优化升级，并积极探索AI赋能传统产业的应用落地方向。国产厂商在AI服务器领域大有可为。

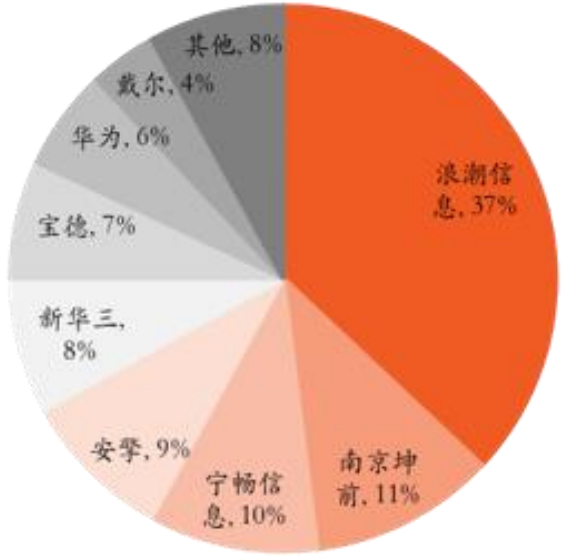
2021H1全球AI服务器市场竞争格局（按销售额）



2022年我国AI服务器市场竞争格局（按销售额）



2022年我国AI服务器市场竞争格局（按出货量）

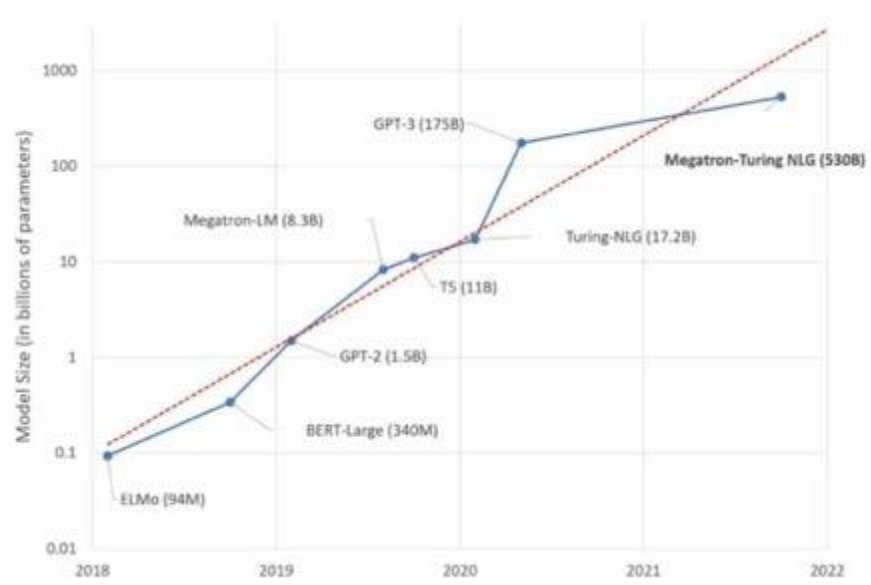


数据来源：IDC，TrendForce，腾讯调研云，平安证券研究所

大模型的实现需要十分强大的算力来支持训练过程和推理过程

- 大模型的实现需要十分强大的算力来支持训练过程和推理过程。根据OPENAI数据，训练GPT-3 175B的模型，需要的算力高达3640 PF-days（即以1PetaFLOP/s的效率要跑3640天）。2018年以来，大模型的参数量级已达到数千亿参数的量级规模，对算力的需求将呈现指数级增长。

◉ 2018-2022年大模型参数增长变化趋势



◉ 各个模型所需计算量及参数量

	模型	总计算量(PF-days)	总计算量(flops)	参数量(百万个)	训练token量(十亿个)
T5模型	T5-Small	2.08	1.80E+20	60	1,000
	T5-Base	7.64	6.60E+20	220	1,000
	T5-Large	26.7	2.31E+21	770	1,000
	T5-3B	104	9.00E+21	3,000	1,000
	T5-11B	382	3.30E+22	11,000	1,000
BERT模型	BERT-Base	1.89	1.64E+20	109	250
	BERT-Large	6.16	5.33E+20	355	250
	RoBERTa-Base	17.4	1.50E+21	125	2,000
	RoBERTa-Large	49.3	4.26E+21	355	2,000
GPT模型	GPT-3 Small	2.60	2.25E+20	125	300
	GPT-3 Medium	7.42	6.41E+20	356	300
	GPT-3 Large	15.8	1.37E+21	760	300
	GPT-3 XL	27.5	2.38E+21	1,320	300
	GPT-3 2.7B	55.2	4.77E+21	2,650	300
	GPT-3 6.7B	139	1.20E+22	6,660	300
	GPT-3 13B	268	2.31E+22	12,850	300
	GPT-3 175B	3640	3.14E+23	174,600	300

◉ 人工智能不同时代对算力翻倍的需求时间

数据来源：澎湃新闻，OPENAI，《COMPUTE TRENDS ACROSS THREE ERAS OF MACHINE LEARNING》，平安证券研究所

大模型的训练成本和推理成本高昂

- 大模型的训练成本和推理成本高昂。以ChatGPT为例，在训练端：根据澎湃新闻信息，2020年，微软宣布与OpenAI合作，建成了一台超级计算机，专门用来在Azure公有云上训练超大规模的人工智能模型。这台为OpenAI开发的超级计算机拥有超过28.5万个CPU核心，拥有超过1万个GPU（V100 GPU芯片）。以此规格，如果自建IDC，以A100 GPU芯片替代V100 GPU芯片，依照A100和V100的性能换算，需要约3000个A100 GPU芯片。根据英伟达网站信息，NVIDIA DGX A100服务器搭载8块A100芯片，估算需要375台NVIDIA DGX A100服务器，每台NVIDIA DGX A100服务器的价格为19.9万美元，则自建IDC的训练服务器的算力成本为7462.5万美元。若在云端训练，据Lambda Labs首席科学官Chuan li介绍，拥有1750亿个参数的GPT-3单次训练成本达到460万美元。

自建IDC的训练成本估算

A100 GPU 芯片（个）	1台NVIDIA DGX A100服务器搭载A100芯片个数	NVIDIA DGX A100服务器（台）	NVIDIA DGX A100服务器价格（万美元）	训练成本（万美元）
3000	8	375	19.9	7462.5

- 在推理（用户访问）端：ChatGPT推出仅两个月月活用户数已经破亿，2023年1月，全球每天约有1300万独立访问者使用ChatGPT。以ChatGPT日活用户2000万估算，假设每天每用户提10个问题，则每天有2亿的访问量。若自建IDC，假设每个问题平均20个字，ChatGPT在A100 GPU芯片上对每个字的响应时间是350毫秒，则2亿的访问量需要A100芯片运行388889个芯片小时，即每天需要16204（388889/24）个A100芯片同时工作，需要2026(16204/8)台NVIDIA DGX A100服务器同时工作，则自建IDC的推理服务器的算力成本为4.03亿美元。若在云端推理，据《Fortune》杂志数据,每次用户与ChatGPT互动,产生的算力云服务成本约0.01美元，则每天2亿的访问量，对应的云端成本为每天200万美元。

自建IDC的推理成本估算

日活用户数（万）	单用户每日提问次数	每个问题平均字数（个）	A100 GPU对每个字的响应时间（毫秒）	每日消耗GPU计算时间（小时）	每天需要A100 GPU芯片（个）	NVIDIA DGX A100服务器（台）	NVIDIA DGX A100服务器价格（万美元）	推理成本（亿美元）
2000	10	20	350	388889	16204	2026	19.9	4.03

数据来源：澎湃新闻，环球零碳，英伟达官网，平安证券研究所

大模型将为全球和中国AI芯片和AI服务器市场的增长提供强劲动力

- 根据《2022年北京人工智能产业发展白皮书》数据，截至2022年10月，北京拥有人工智能核心企业1048家，占我国人工智能核心企业总量的29%。以此计算，我国人工智能核心企业总数约为3614家。初步来看，我国参与大模型的企业大致可以分为两类：即C端应用的企业和B端应用的企业。在C端应用方面，假设其中有20家企业自研或与合作方共同研发通用大模型，自建IDC训练和推理面向庞大C端月活的千亿量级参数的大模型，算力需求可参照ChatGPT，即单一企业自建IDC推理和训练大模型的算力成本约为4.78（4.03+0.75）亿美元，按20家企业估算，算力需求为95.6亿美元。
- B端应用方面，假设在3614家人工智能核心企业中，有10%的企业即361家企业使用合作伙伴的大模型（包括开源大模型），使用垂直行业数据进行进一步的训练得到垂类大模型，并将垂类大模型部署到客户的数据中心，则训练成本为垂类大模型的训练成本，推理成本为客户运营垂类大模型的成本。假设垂类大模型的训练成本为ChatGPT训练成本的1/10即0.075亿美元，则合计训练成本为27.08亿美元，假设每个垂类大模型厂商平均部署100个客户，每个客户运营大模型平均需要3台DGX A100服务器，则每个客户平均推理成本为59.7万美元，合计推理成本为215.52亿美元，算力需求为242.6（27.08+215.52）亿美元。
- 以上，根据我们的初步估算，大模型的应用将为我国AI服务器市场带来338.2亿美元的市场需求。以2021年我国AI服务器市场规模占全球AI服务器市场规模的份额估算，则将为全球AI服务器市场带来约891.2亿美元的市场空间。市场空间巨大，相关芯片和服务器厂商将深度受益大模型的发展浪潮。

大模型将为国内带来的AI服务器市场增量测算

	企业数量（家）	企业训练算力成本	企业（对应客户）推理算力成本	算力总成本（亿美元）
C端应用的企业	20	15亿美元（0.75亿美元*20）	80.6亿美元（4.03亿美元*20）	95.6
B端应用的企业	361（3614*10%）	27.08亿美元（0.075亿美元*361）	215.52亿美元（59.7万美元*100*361）	242.6
合计	-	-	-	338.2

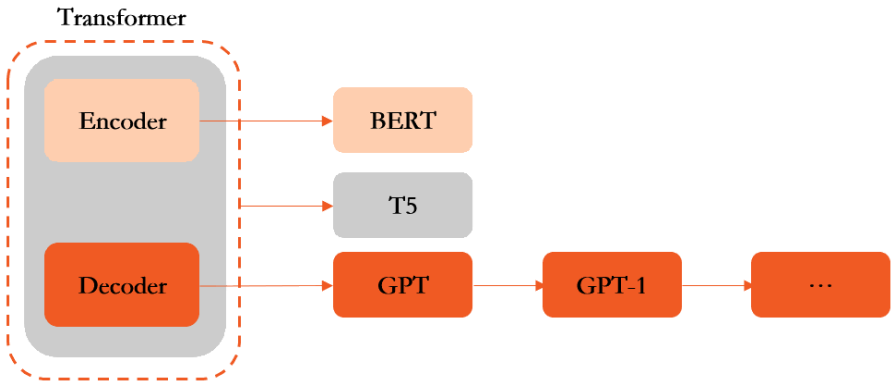
数据来源：《2022年北京人工智能产业发展白皮书》，IDC，平安证券研究所

CONTENT
目录

Transformer架构衍生三类模型，生成式预训练模型正成为主流

- Transformer模型是由Google团队的Ashish Vaswani 等人发表论文《Attention Is All You Need》提出的模型概念，随后迅速成为各类预训练大模型的基础架构。该模型是一个深度学习模型，其标志性特征是采用了self-attention机制，可为输入数据的各部分分配不同权重，核心是从关注全部到关注重点，从而节省资源，快速获得最有效的信息。
- 以Transformer架构为基础衍生出的典型预训练语言模型大致可以分为三类：1）Encoder模型（以Google的BERT模型为代表）：又称自编码模型，适用于内容理解任务，例如情感分析等；2）Decoder模型（以OpenAI的GPT系列模型为代表）：又称自回归模型，适用于生成式任务，例如文本生成；3）Encoder-Decoder模型（以Google的T5模型为代表）：又称 Seq2Seq模型，通常用于需要内容理解和生成的任务，例如翻译。
- 随着ChatGPT的火爆，基于Decoder模型的生成式预训练Transformer模型（ Generative Pre-Trained Transformer，简称GPT）正成为主流。GPT具有可以接受大量无标注的文本数据进行预训练等优势，随着模型的参数量逐渐增大，训练数据逐渐增多，GPT在文本生成、问答系统展现出强大能力。

Transformer模型系列分类



Decoder模型与Encoder模型的区别

	基于语言模型的生成式模型（Decoder）	基于双向编码的预训练模型（Encoder）
输入顺序	从左到右的单向模型	双向模型
训练数据	大量的网络文本数据	两个大型语料库（包括Wikipedia和BooksCorpus）
预训练方式	通过预测下一个词来学习语言模型	通过预测下一个词来学习语言模型
微调方式	需要指定输入输出的语言模型任务	可以应用在多种任务上，如文本分类、命名实体识别
应用场景	主要用于自然语言生成任务	主要用于自然语言理解任务

数据来源：CSDN，arxiv.org，《Transformers in Vision: A Survey》NASA/ADS，平安证券研究所

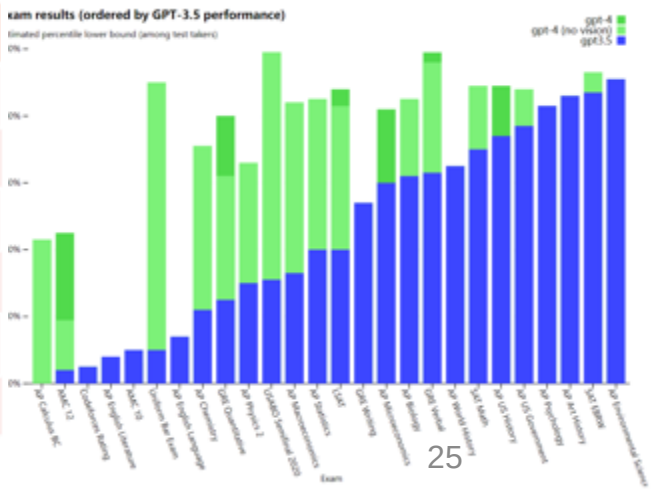
OpenAI的聊天机器人ChatGPT火爆出圈，推出两个月后月活用户突破1亿

- 2018年，OpenAI研发团队发布《Improving Language Understanding by Generative Pre-Training》，介绍了基于生成式预训练的语言模型GPT（Generative Pre-Trained Transformer）的原理和性能。该论文的核心内容是通过基于Transformer架构的生成式预训练方法，将无标签的大规模文本数据用于预训练，从而在多个NLP任务上取得了优异的效果。
- 2022年11月，OpenAI公司推出的聊天机器人ChatGPT火爆出圈。它能够理解和生成自然语言，为用户提供各种应用场景的解决方案，例如撰写文案、生成代码、回答问题等，在面对多种多样的问题时能够对答如流，似乎打破了机器和人的边界。在ChatGPT推出两个月后，月活用户已突破1亿。2023年3月，OpenAI公司发布了GPT-4。相较于ChatGPT，GPT-4的多模态处理能力、高级推理能力等多种能力得到巨大提升。2023年5月，OpenAI公司推出苹果手机版本的ChatGPT应用程序，实现ChatGPT在移动端的应用。

GPT系列模型发展进程

GPT-4与ChatGPT的对比

	GPT-4	ChatGPT
模型种类	多模态模型	自然语言处理模型
输入模态	文本、图像	文本
输出模态	文本、图像	文本
文字输入长度	25000字	2000字
英语准确度	85.5%	70.1%
普通话准确度	80.1%	



数据来源：OpenAI官网，平安证券研究所

Google发布PaLM2大模型，能力对标GPT-4

- 2023年2月，Google推出基于LaMDA大模型的聊天机器人Bard，对标ChatGPT。Bard具有文本生成、对话问答、编写代码等能力，是一款可以与用户进行自然、流畅和有趣的对话的应用程序。相较于ChatGPT，它能够访问互联网并获取实时信息，提供最新的问题答案，但目前仅支持英文、日文和韩文版本。
- 2023年5月，Google发布对标GPT-4能力的PaLM2大模型，并将其赋能于聊天机器人Bard身上。进化后的Bard将具备强大的逻辑推理、数学运算以及代码编写的能力。Google宣称未来Bard会将谷歌地图、文档和Gmail的信息带入对话中，目前已经实现导入生成表格至Sheets应用的功能；还会在Adobe等第三方工具的帮助下回复用户，例如根据文本生成图像。由于PaLM 2包含的模型中最轻量版本Gecko小到可以在手机上运行，AIGC应用在移动端部署的进程有望加速。

不同模型在推理任务上的评估情况

	SOTA	GPT-4	PaLM	PaLM 2
WinoGrande	87.5[a]	87.5[a] (5)	85.1[b] (5)	**90.9** (5)
ARC-C	**96.3**[a]	**96.3**[a] (25)	88.7[c] (4)	95.1 (4)
DROP	**88.4**[d]	80.9[a] (3)	70.8[b] (1)	85.0 (3)
StrategyQA	81.6[c]	-	81.6[c] (6)	**90.4** (6)
CSQA	**91.2**[e]	-	80.7[c] (7)	90.4 (7)
XCOPA	89.9[g]	-	89.9[g] (4)	**94.4** (4)
BB Hard	65.2[f]	-	65.2[f] (3)	**78.1** (3)

不同模型在数学能力上的评估情况

Task	SOTA	PaLM	Minerva	GPT-4	PaLM 2	Flan-PaLM 2
MATH	**50.3**[a]	8.8	33.6 / **50.3**	42.5	34.3 / 48.8	33.2 / 45.2
GSM8K	92.0[b]	56.5 / 74.4	58.8 / 78.5	92.0	80.7 / 91.0	84.7 / **92.2**
MGSM	72.0[c]	45.9 / 57.9	-	-	72.2 / **87.0**	75.9 / 85.8

不同模型在编码能力上的评估情况

	HumanEval		MBPP		ARCADE	
	pass@1	pass@100	pass@1	pass@80	pass@1	pass@30
PaLM 2-S*	**37.6**	**88.4**	**50.0**	**86.6**	**16.2**	**43.6**
PaLM-Coder-540B	35.9[a]	**88.4**[a]	47.0[a]	80.8[a]	7.9[a]	33.6[a]

数据来源：《PaLM 2 Technical Report》，平安证券研究所

Meta推出130亿参数的开源模型LLaMA，性能可匹配1750亿参数的GPT-3

- 2023年2月，Meta公司推出了一款AI大型语言模型LLaMA，宣称仅用约1/10的参数规模，便实现了匹配OpenAI GPT-3、谷歌PaLM等主流大模型的性能表现。LLaMA大模型具备常识推理、闭卷问答、阅读理解、数学推理、代码生成、大规模多任务语言理解、训练期间的能力进化等多种能力，并且LLaMA大模型共有四个不同的参数量版本，分别为70亿、130亿、330亿和650亿四种参数规模。

- 根据Meta研究团队发布的数据，130亿参数的LLaMA模型在单个GPU上运行时，性能表现可能超过1750亿参数GPT-3。2023年3月，Meta宣布将LLaMA基础大型语言模型“开源”，不作商用目的免费供给研究人员，并且在GitHub上提供了精简版LLaMA。无论是更少参数规模的模型实现优于更大参数规模模型的研究成果，还是Meta公司将LLaMA模型开源，都意味着AI大模型在性能和广泛应用的发展上不断加速。

- 2023年4月，Meta公司发布了一个计算机视觉基础模型Segment Anything Model（SAM）。该模型是一种可提示的分割模型，可以在不需要额外训练的情况下对不熟悉的对象和图像进行零样本泛化，从而“分割”任何图像中的任何对象。相较于以前交互式分割（需要人为引导方法）和自动分割（需要预先定义特定对象）这两种分割方法，SAM实现了两种方法的综合，是一种同时具有通用性和自动分割能力的分割方法。

不同模型在常识推理任务的zero-shot性能

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88.0**	82.3	-	83.4	**81.1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80.0**	**57.8**	58.6
	65B	85.3	**82.8**	**52.3**	**84.2**	77.0	78.9	56.0	**60.2**

SAM的分割效果

数据来源：《LLaMA: Open and Efficient Foundation Language Models》，SAM官网，平安证券研究所

国内厂商纷纷布局大模型，国产大模型快速发展

➢ Transformer架构发布以来，我国多家科技公司和科研机构纷纷开始布局大模型。尤其2023年3月以来，我国国产大模型快速发展，多家厂商推出了自研的通用大模型，某些功能已可比肩ChatGPT，国产大模型发展势头良好。

我国国产大模型情况（部分）

公司	发布时间	模型名称	模型特点
北京智源人工智能研究院	2021.6	悟道2.0大模型	该模型是中国首个参数量达到1.75万亿的双语多模态预训练模型，在中英双语共4.9T的高质量大规模清洗数据上进行的训练。在研发过程中智源研究院建设了全球最大的语料数据库WuDaoCorpora2.0保证行业内大规模智能模型的研发有丰富的数据支撑。此外，智源研究院将提供多种形态的模型能力服务，赋能AI技术研发，未来将与更多合作伙伴以模型研发和应用落地为导向促进产业集聚。
智谱AI	2022.8	GLM-130B大模型	该模型是双语千亿级超大规模预训练模型，在准确性等关键指标方面与OpenAI、谷歌大脑、微软等公司的大模型接近或持平，支持在华为昇腾、海光和申威等国产芯片上进行训练和推理，目前已有全球1000余家机构申请使用。公司未来将形成AIGC产品矩阵，提供智能API服务，目标通过认知大模型链接物理世界的亿级用户，赋予机器像人一样“思考”的能力。
百度	2023.3.16	文心一言大模型	该模型是参数量为2600亿的大语言模型，具有文学创作、商业文案创作、数理推算、中文理解、多模态生成等能力。目前，文心一言大模型相较于3月份发布的版本实现了推理性能提升10倍，在企业应用的高频、核心场景下高性能版推理性能提升50倍，模型效果得到巨大提升。此外，文心一言大模型相继嵌入到百度地图、小度音箱等百度产品，实现业务和场景的应用落地。未来，百度将用 AI 原生的思维把全部产品重新做一遍，并实现AI时代的新IT技术栈（芯片层、框架层、模型层、应用层）的全栈产品优化，大幅提升效率。
华为	——	盘古大模型	盘古系列大模型目前共分为L0、L1、L2三个层级，分别指基础大模型、指行业大模型和面向更加细分场景的推理模型。目前华为已推出的L1模型有盘古金融大模型、盘古矿山大模型、盘古气象大模型、盘古电力大模型等，推出的L2模型有类似基于气象大模型的短临气象预报、台风预测等多种场景模型。未来盘古大模型将继续坚持“AI for Industries”理念，实现各行业细分场景的应用落地。
阿里云	2023.4.11	通义千问大模型	该模型是超大规模语言模型，具有多轮对话、文案创作、逻辑推理、多模态理解、多语言支持等功能，钉钉、天猫精灵等阿里旗下产品在接入通义千问大模型后变得更加智能化和人性化。该模型还可以通过API插件实现AI能力泛化，从而帮助各行业企业拥有自己领域的垂直大模型，实现具体场景的应用落地。未来阿里将会把所有产品都接入通义千问实现产品全面升级，并通过通义千问为各行业企业打造专属大模型。
科大讯飞	2023.5.6	星火认知大模型	该模型为认知智能大模型，具有文本生成、语言理解，知识问答、逻辑推理、数学能力、代码能力、多模态能力七大核心能力，其中部分能力已超越ChatGPT。此外，星火认知大模型已实现在教育、办公、汽车、数字员工、AI虚拟人等领域的应用落地，未来将赋能更多领域。公司将在6月9日推出“讯飞星火认知大模型”V1.5，并努力在10月24日前实现星火大模型在通用认知大模型能力上对标ChatGPT，在中文能力上超越ChatGPT，在英文能力上达到与ChatGPT相当的水平。
中科创达	2023.5.18	魔方Rubik大模型	该模型为算法全部自研的通用大模型，未来将聚焦于赋能边缘端，结合操作系统技术形成AI原生的操作系统产品，并为客户做好私有化部署。目前RUBIK基础平台全面覆盖了边缘端、语言大模型、多模态、机器人等大模型系列，魔方Rubik大模型已在边缘AI、智能助理领域得以应用。

数据来源：各公司官网，平安证券研究所

智源推出完整悟道3.0大模型体系，并宣布将系列大模型全面开源

- 2021年6月，北京智源研究院发布了参数规模达到1.75万亿的超大规模智能模型悟道2.0。该模型是首个在100%国产超算平台上训练的万亿模型，是具有国产属性的超大模型，它可以同时处理中英文和图片数据，并且在问答、作诗、配文案、视频、绘画、菜谱等多项任务中正在逼近图灵测试标准。此外，首个中国原创虚拟学生“华智冰”基于该模型诞生。
- 2023年6月，智源发布了完整的悟道3.0大模型系列，包括悟道·天鹰（Aquila）语言大模型系列、FlagEval（天秤）大模型语言评测体系以及悟道·视界视觉大模型系列。其中悟道·天鹰（Aquila）语言大模型系列包含了Aquila基础模型（7B、33B）以及AquilaChat对话模型和AquilaCode文本-代码生成模型。类ChatGPT模型AquilaChat目前综合能力可以达到GPT-4能力的70%左右，文本代码生成大模型AquilaCode-7B与OpenAI编码大模型Codex-12B相比，两者在HumanEval pass@1上的结果较为接近。此外，智源宣布悟道3.0大模型系列将全面开源，展现出其坚持开源开放的决心。

虚拟学生“华智冰”

Aquila基础模型特点

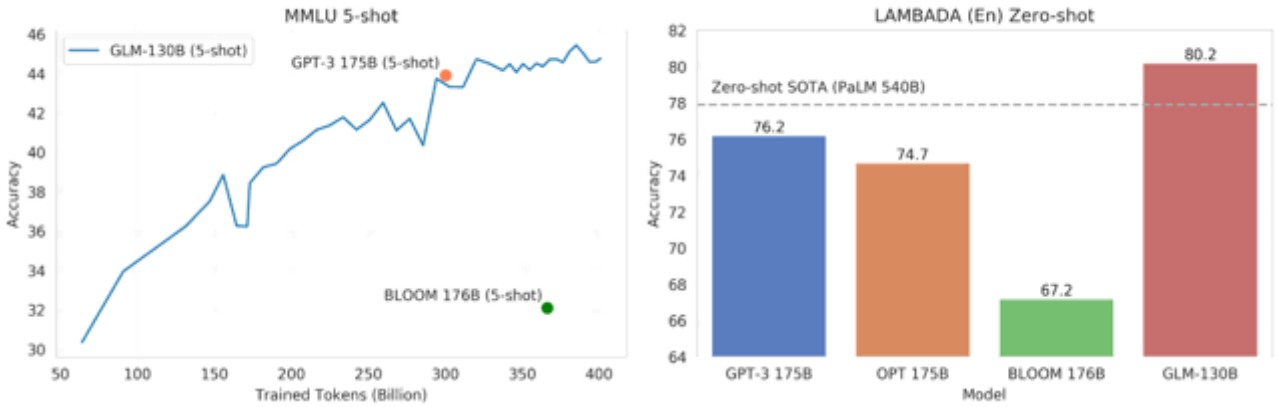
AquilaCode代码生成样例

数据来源：智源研究所，国际科技创新中心，清华大学，平安证券研究所

智谱AI推出ChatGLM-130B和ChatGLM-6B，已获得较为广泛的应用

- 2022年8月，智谱AI推出千亿级开源大模型GLM-130B。该模型是双语预训练语言模型，它的少样本学习性能在多任务语言理解基准上达到并超过了GPT-3的水平，在被广泛用于大规模语言模型性能评估的LAMBADA基准上达到了80.2%的准确率，而GPT-3 175B为76.2%。2022年11月，斯坦福大学大模型中心对全球30个主流大模型进行了全方位的评测，GLM-130B是亚洲唯一入选的大模型。后续智谱以GLM-130B为基座开发出对话机器人ChatGLM-130B，它是一个初具问答和对话功能的千亿中英语言模型，并针对中文进行了优化。
- 2023年3月，智谱发布了开源的中英双语对话GLM模型ChatGLM-6B，它的参数量为62亿。它的优势在于模型的轻量化使得它具有较低的部署门槛，模型的开源也让不少人尝试部署在本地体验，并且模型已能生成相当符合人类偏好的回答。目前，ChatGLM-6B全球下载达到200万，数百垂直领域模型和国内外应用基于该模型开发。

GLM-130B的任务表现



ChatGLM-130B已获得较为广泛的应用

时间	公司	事件
2023.05.25	联想	联想接入ChatGLM-130B API开发智能打印产品。
2023.05.15	中国民航信息网络公司	中国民航信息网络公司基于接入ChatGLM-130B API开发航旅智能产品。
2023.04.25	清华研究生会	清华研究生会基于ChatGLM-130B开发的“水木ChatGLM”上线，服务全校同学。
2023.04.24	360	360基于ChatGLM-130B联合研发千亿级大模型“360GLM”。
2023.04.15	值得买	值得买部署ChatGLM-130B私有化实例用于电商平台产品。
2023.04.14	美团	美团私有化部署ChatGLM-130B，联合研发“美团GLM”。
2023.03.10	竹间智能科技	竹间智能科技接入ChatGLM-130B API开发智能客服产品。

数据来源：清华网站，智谱AI官网，ChatGLM官网，平安证券研究所

百度推出文心一言大模型，与多家企业合作共建文心一言生态

- 2023年3月，百度宣布正式推出大模型文心一言，该模型具有文学创作、商业文案创作、数理推算、中文理解、多模态生成五种能力。2023年6月，百度宣布时隔约3个月后文心一言实现了推理性能提升10倍，高性能版文心一言-Turbo的推理性能提升50倍。文心一言-Turbo可以实现在保证与文心一言相同效果的同时降低参数规模，降低企业在使用大模型的算力成本。
- 截止2023年6月，已有超15万客户申请接入文心一言，超300家企业成为文心一言生态伙伴，实现超400个具体场景落地。目前百度已初步搭建文心一言应用矩阵，文心一言已落地百度搜索、百度地图、小度智能音箱等多个百度自有业务，同时也在不同行业的企业场景中实现落地，如长安汽车、地平线、知乎等。此外，基于文心大模型的AI辅助编程工具Comate开始内测，文心一言大模型的应用产品线不断丰富。
- 2023年3月，百度推出的“文心千帆”大模型平台开启内测。该平台为一站式企业级大模型平台，为行业伙伴提供SFT大模型效果调优、Prompt工程等服务，努力在智能办公、旅行服务、电商直播、政务服务、金融服务五大行业打造行业样板间。未来平台的迭代方向将向着：1）效果，保证与业务的结合能够更好地解决问题。2）性能，通过在技术架构等方面的性能提升实现企业应用的降本增效。

文心一言问答界面

文心千帆功能模块

数据来源：文心一言官网，百度官网，百度发布会，平安证券研究所

华为发展盘古大模型三大层级模型体系，持续推进大模型为各行业赋能

- 2023年4月，田奇在“人工智能大模型技术高峰论坛”分享华为对于大模型的发展方向。盘古大模型目前发展出L0、L1、L2三大层级的模型体系，L0是指NLP大模型等基础大模型；L1指电力、气象、矿山等行业大模型；L2指气象中的台风预测等面向各行业中细分场景的模型，目前盘古大模型已在100多个行业场景中完成验证。
- 盘古大模型的核心定位是为各行业赋能，目前已发布了关于矿山、气象、药物、分子、电力、海浪、金融等行业的大模型，并在行业中拓展场景模型，实现应用效率的提升。在技术层面，华为盘古大模型具有优秀的泛化能力、高效的样本筛选能力、小样本/零样本能力以及低门槛的AI开发等特点，实现在为行业企业赋能的同时更好地实现企业的降本增效。
- 盘古大模型主要基于一站式AI开发平台ModelArts，实现大模型计算、通信、存储以及算法的优化等。 ModelArts支持全流程MLOps开发，提供全流程工具栈，以及提供开放架构，通过联合伙伴以第三方工具集成的方式，打造包括自动驾驶等全流程工具链赋能各类AI开发场景。平台目前已实现在互联网、自动驾驶等多领域的降本增效。

盘古大模型分层应用

盘古NLP大模型：文本生成、内容理解

盘古CV大模型：分类、分割、检测

盘古科学计算大模型：气象预报、药物分子优化、海浪预测

盘古系列	行业&领域	场景
盘古CV大模型	工业质检	偏光片质检、煤矿质检
	物流仓库监控	物的银行
	时尚辅助设计	门店半定制设计
盘古NLP大模型	智能文档检索	类案检索
	智能ERP	企业财务异常检测
	小语种大模型	阿拉伯语大模型
盘古科学计算大模型	气象预报、海浪预测	盘古气象大模型

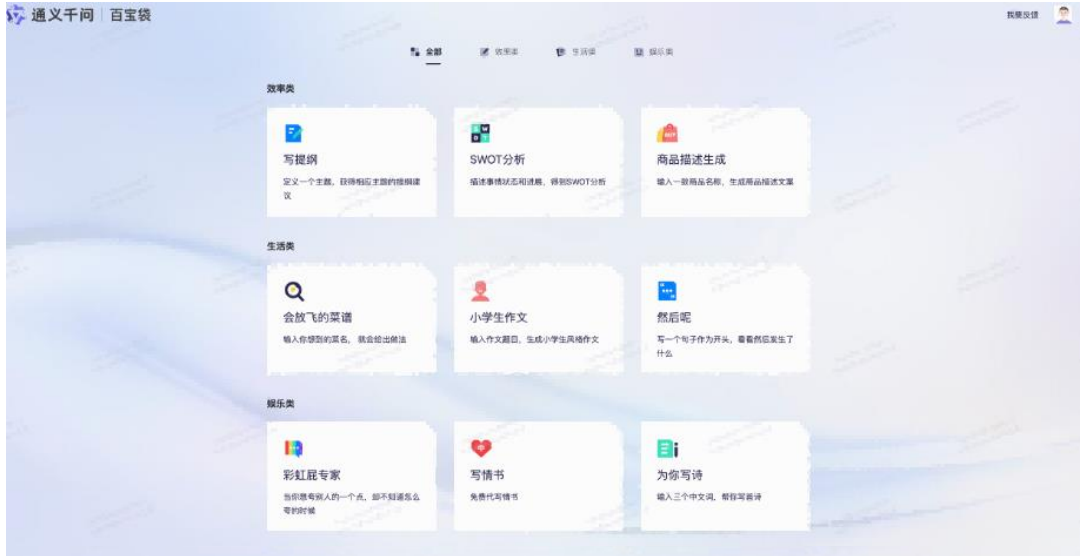
ModelArts架构

数据来源：华为官网，华为发布会，华为云公众号，平安证券研究所

阿里云推出通义千问大模型，将帮助更多企业拥有专属大模型

- 2023年4月，阿里云推出了超十万亿参数量的自研大语言模型通义千问。通义千问具有多轮对话、文案创作、逻辑推理、多模态理解、多语言支持等能力，目前的输入文本上限为1000字。此外，通义千问自身搭载了9种应用，引导用户更好地了解通义千问能够实现的各种能力。
- 通义千问将陆续嵌入阿里电商、搜索、导航、文娱等场景，后续通义千问大模型将接入阿里所有的产品中，实现产品的全面升级。目前通义千问已接入天猫精灵、钉钉等产品中，使产品具有了多项AI新功能。并且通义千问可以通过API插件实现AI能力泛化，为更多企业打造具备行业能力的专属大模型。

通义千问的应用百宝袋



接入通义千问后的钉钉斜杠

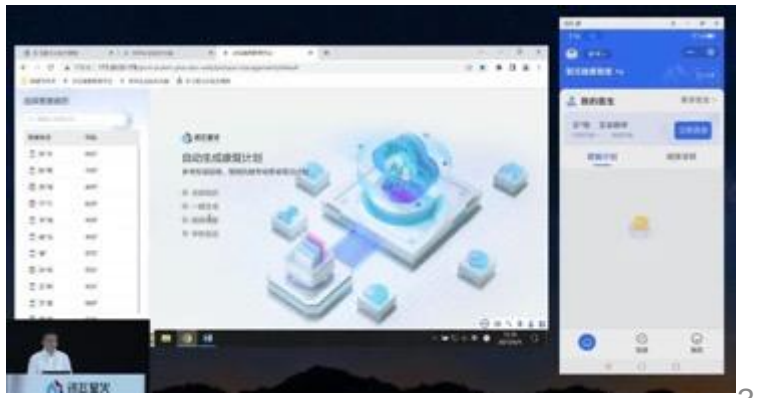
数据来源：通义千问官网，阿里云发布会，平安证券研究所

科大讯飞打造“1+N”星火认知大模型，开放合作共建人工智能“星火”生态

- 2023年5月，科大讯飞发布“1+N”星火认知大模型，该模型具有七大核心能力包括文本生成、语言理解，知识问答、逻辑推理、数学能力、代码能力、多模态能力等，它的数理能力、中文长文本生成和通识知识回答方面已超越ChatGPT。2023年6月，科大讯飞发布星火认知大模型V1.5，星火大模型V1.5在文本生成、语音理解、知识问答、逻辑推理、数字能力以及代码能力等方面都有所提升。公司宣布开放星火认知大模型开发接口，将七大维度能力、200多个小助手对应能力全面开放给开发者，并且支持多端接入快速集成，支持私有化部署，与广大开发者共建“星火”生态。
- 公司高度重视大模型的应用落地，“1+N”中的“1”指星火认知大模型，“N”则指大模型在教育、办公、汽车、人机交互等各个领域的落地。2023年5月，公司发布多款“大模型+教育、办公、汽车、数字员工行业”产品并实现应用，商业模式闭环正在逐渐形成；2023年6月，星火大模型为教育、办公、医疗、工业等更多行业赋能，已发布的产品包括星火语伴 APP、讯飞听见智慧屏、智医助理、羚羊工业互联网平台等。
- 公司计划将在2023年8月实现星火大模型代码能力的升级以及多模态交互能力提升；**在2023年10月24日前实现星火大模型能够全面对标ChatGPT**，在中文能力方面超越ChatGPT，在英文能力方面做到持平，并实现在医疗和教育等领域的业界领先。

星火APP中的星火助手中心

医疗领域的应用落地



数据来源：科大讯飞官网，科大讯飞发布会，平安证券研究所

中科创达发布魔方Rubik大模型，赋能边缘端驱动硬件智能化

- 2023年5月，中科创达发布自研的魔方Rubik大模型，目前参数量级为几十亿，后续将推出千亿参数量级的大模型。该模型的具有三大特点：1）赋能边缘端，主要聚焦赋能汽车、手机以及物联网设备等领域。2）结合操作系统技术，形成AI原生的操作系统产品。3）利用模型能力做好私有化部署服务。

- 魔方Rubik大模型在应用端赋能多款产品。目前中科创达已发布Rubik Solutions、Rubik Enterprise、Rubik OS、Rubik Models、Rubik Device等系列产品及解决方案，涵盖汽车、物联网、手机等多个领域。具体产品包括：1） Rubik GeniusCanvas：面向设计行业工作者的工具，可以实现3D工业模型自动优化，数字人动作自动生成等功能。2） CodePilot：面向安卓和通用开发者的工具，可一键生成部分代码，目前仍处于内部研发阶段。3） Rubik Enterprise Suite：面向企业管理者的工具，可以通过大模型强化内部管理和交互流程，利用私有化领域数据针对性训练大模型。

- 魔方Rubik大模型在边缘计算、智能物联网、智能汽车等领域具有一定优势,驱动硬件智能化。在边缘AI领域，公司具有先进产品EBX边缘智能站满足用户不同网络需求，并且通过中科创达魔方大模型Rubik Studio云平台实现数据标注、算法部署等全流程服务，同时通过模型的先进压缩技术实现边缘AI的良好结果。在智能物联网和智能汽车领域，公司拥有丰富多样的业务场景和产品矩阵，对行业的深度了解以及产品和技术的积累使得魔方Rubik大模型能够在物联网和智能汽车等行业内的应用实现快速落地，为公司产品赋能。

魔方Rubik大模型+Kanzi

魔方Rubik大模型系列产品

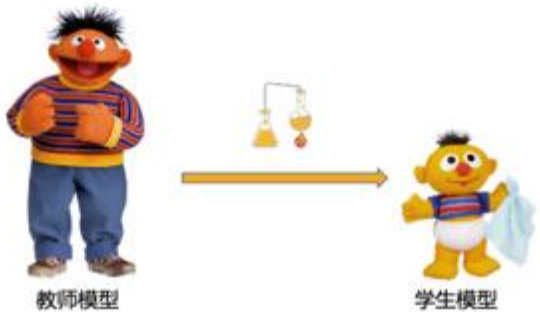
魔方Rubik大模型赋能智能硬件

数据来源：中科创达官网，中科创达发布会，平安证券研究所

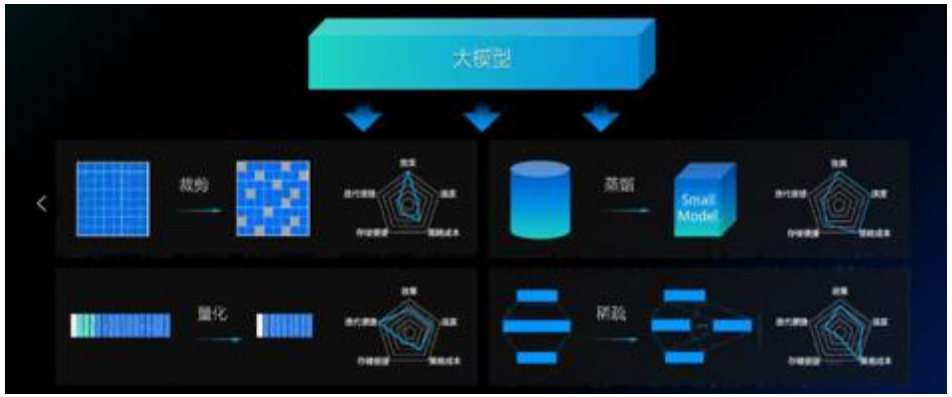
大模型的小型化降低使用门槛，助力实现大模型在各细分场景的应用落地

- 百度利用文心ERNIE-Tiny在线蒸馏方案，结合飞桨模型压缩工具PaddleSlim等工具和技术，实现通过蒸馏、裁剪、量化等方式将大模型（可以理解为“教师模型”）的知识传递给小模型（可以理解为“学生模型”），并且文心ERNIE-Tiny在线蒸馏方案的效果十分显著，模型参数压缩率可达99.98%，压缩版模型仅保留0.02%参数规模就能与原有模型效果相当，解决了大参数模型的成本高、部署难等问题，推动大模型的行业应用落地。
- Meta公司自研大语言模型LLaMA的开源，使LLaMA模型成为了大语言模型领域层出不穷的创新来源。LLaMA系列模型包括7B、13B、33B和65B的参数版本，并且130亿参数量版本的LLaMA模型在大多数基准测试下超越GPT-3。小参数版本LLaMA模型也可以很好的轻量化训练并针对性的用于解决特定业务场景下的问题。由于LLaMA大模型具有高度灵活性、可配置性、高度泛化能力以及强大性能等特点，使它成为了垂直AI模型通用基座的选择之一。
- 大模型小型化技术的逐步成熟，使更多企业跨过使用大模型的门槛。技术的逐步成熟使大模型在参数量降低的同时性能并未发生太大改变，参数量较小的垂类模型依旧拥有较为强大的性能，可赋能各垂类企业大模型能力，助力实现大模型在各细分场景的应用落地。

大模型蒸馏示意图



大模型的小型化方法



LLaMA模型与其他模型的性能对比

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88.0**	82.3	-	83.4	**81.1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80.0**	**57.8**	58.6
	65B	85.3	**82.8**	**52.3**	**84.2**	77.0	78.9	56.0	**60.2**

数据来源：《LLaMA: Open and Efficient Foundation Language Models》，百度官网，百度飞桨官网，平安证券研究所

OpenAI等国外厂商全球领先，国内厂商快速追赶

- 当前，以OpenAI、Google、Meta等为代表的国外厂商凭借先发优势、算力优势、以及数据集等方面的优势，在全球GPT大模型领域具有领先优势。但我国大模型产品也在快速发展，正奋起直追，尤其2023年3月以来，多家厂商推出了自研的通用大模型，某些功能已可比肩ChatGPT。同时，国产大模型在各行业的应用以及生态建设也取得积极进展。

- 以科大讯飞为例，2023年5月6日，公司发布星火大模型。公司董事长刘庆峰在发布会上展示了星火大模型的七大核心能力，并宣布星火大模型在数理能力、中文长文本生成和通识知识回答方面已超越ChatGPT。2023年6月9日，公司发布星火大模型V1.5，V1.5在文本生成、语音理解、知识问答、逻辑推理、数字能力以及代码能力等方面都有所提升，开放式知识问答、逻辑推理和数学能力以及多轮对话能力这三大综合能力方面进一步升级。目前，星火大模型已落地讯飞学习机等C端产品以及教育、办公、汽车、医疗、工业等场景。公司计划将在2023年8月实现星火大模型代码能力的升级以及多模态交互能力提升；在2023年10月24日前实现星火大模型能够全面对标ChatGPT，在中文能力方面超越ChatGPT，在英文能力方面做到持平，并实现在医疗和教育等领域的业界领先。

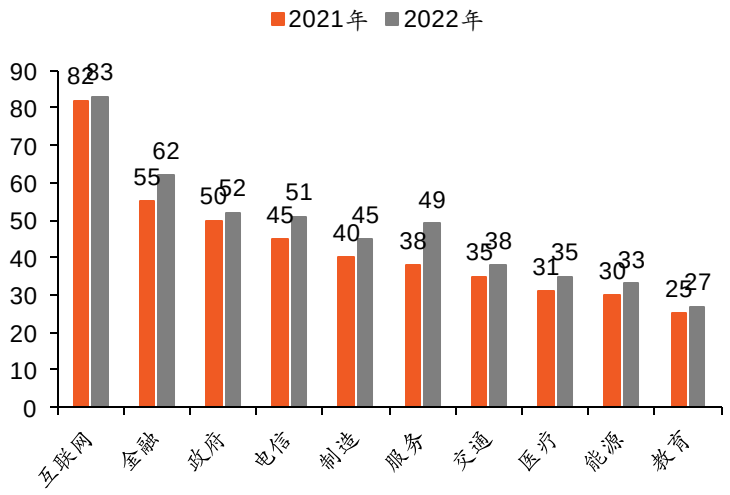
- 以智谱AI为例，2022年8月，智谱AI推出千亿级开源大模型GLM-130B。该模型是双语预训练语言模型，它的少样本学习性能在多任务语言理解基准上达到并超过了 GPT-3 的水平，在被广泛用于大规模语言模型性能评估的 LAMBADA 基准上达到了 80.2% 的准确率，而 GPT-3 175B 为 76.2%。美团、360等企业纷纷基于GLM-130B研发自己的大模型。2023年3月，智谱发布了类ChatGPT的国产大模型ChatGLM-6B，它是一个开源的、支持中英双语的对话语言模型，参数为62亿。目前，ChatGLM-6B 全球下载达到200万，数百垂直领域模型和国内外应用基于该模型开发。我们认为，我国大模型虽然相比GPT-4或仍有一定差距，但在短期内达到或接近ChatGPT的水平是可以预期的。并且从ChatGLM-6B模型的例子来看，我国大模型的小型化技术已经比较成熟，可助力实现大模型在各细分场景的应用落地。我国大模型产品已经初步具备商用能力。

CONTENT
目录

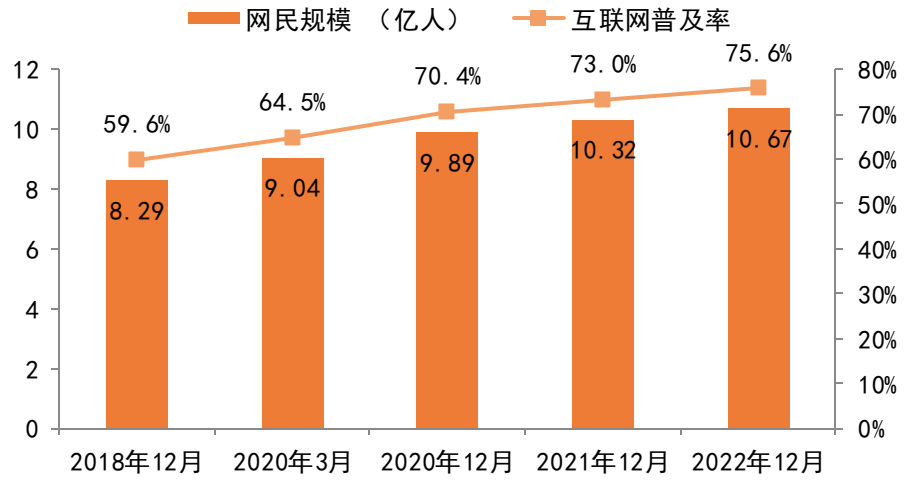
大模型赋能千行百业，AIGC未来发展前景广阔

- 当前，人工智能在我国各行业已经得到广泛应用。根据IDC数据，2022年，人工智能在我国互联网、金融、政府、电信、制造、服务、交通、医疗、能源、教育等行业的渗透率分别为83%、62%、52%、51%、45%、49%、38%、35%、33%、27%。随着大模型时代的到来，我国多家互联网企业的掌门人均表示，大模型时代，所有的产品都值得用大模型重做一次。根据中国互联网络信息中心（CNNIC）数据，截至2022年12月，我国网民规模达10.67亿，互联网普及率达75.6%。
- 我们认为，随着国产大模型的逐步成熟，在政策与技术的共振下，我国大模型产品面向我国庞大的互联网C端用户群和丰富的行业应用场景，将与产品和应用场景深度融合，赋能我国数字经济的发展。根据中国信通院数据，2022年，我国数字经济规模达到50.2万亿元，占GDP比重达到41.5%。参考我国数字经济的巨大体量，我国AIGC产业未来发展前景广阔。

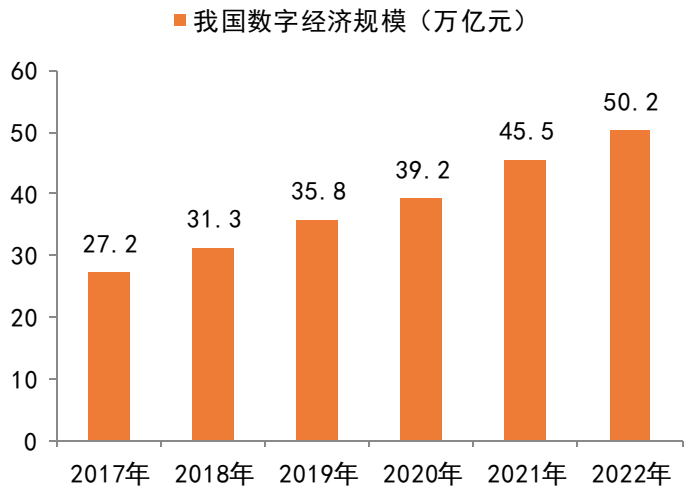
2021-2022年我国各行业人工智能渗透率（%）



我国网民规模和互联网普及率



2017-2022年我国数字经济规模



数据来源：IDC，工信部，国家统计局，中国信通院，《第51次中国互联网络发展状况统计报告》，平安证券研究所

AIGC+办公：Microsoft 365 Copilot 将AIGC与办公软件结合，助力实现降本增效

微软推出Microsoft 365 Copilot 把GPT-4技术引入Office办公软件，将AIGC与办公软件结合。AI办公软件将大型语言模型的力量与业务数据和办公软件应用相结合，实现了发挥创造力、释放生产力、提高办公技能和改变生产方式等目的。Copilot有望带动各行业AI助手市场的发展，产生广泛的行业应用。

应用场景

- 内容生成：生成会议纪要、对话问答、总结分析、智能改写等
- 理解能力：语言表达判断修改、文字润色、内容分析等
- 多模态：图片生成、语音生成、视频生成等
- 数理计算：数据分析、公式推理等

具体案例

- 金山办公WPS AI：WPS AI是具备大语言模型能力的生成式AI应用，搭载WPS AI后的组件可以实现智能问答、文本生成、文本分析、PPT生成、PDF翻译等功能。
- 万兴科技：万兴科技的PDFelement、万兴喵影（Filmora）和亿图图示（EdrawMax）已经推出AI相关功能，通过赋能多模态创业业务场景，实现创意软件智能化的进一步提升。目前系列软件可提供文档智能处理、文案辅助创作、数字人营销、图片生成等功能。
- 福昕软件：公司在海外市场研究产品集成ChatGPT，在国内市场与百度“文心一言”达成合作。公司将AI赋能完整PDF解决方案全线产品，提供多模态智能识别、电子签名、文字翻译等功能。

数据来源：各公司官网，各公司年报，平安证券研究所

AIGC+传媒：AIGC将成为未来传媒行业发展的重要驱动力

AIGC虽然处于发展早期，但成长曲线十分陡峭，随着越来越多由AI驱动的生产力工具被整合到工作流中，各产业生态势必将被重塑。人工智能作为新的传媒行业发展引擎，未来将全面激活行业各业务，助推传媒行业业务发展进入新纪元。从应用方向上看，当前AIGC赋能工具主要集中在文本、代码、图像、音频等几大领域，与传媒行业密切相关，AIGC将成为未来传媒行业发展的重要驱动力。

应用场景

- 文本方面：客服问答、新闻撰写、营销文本生成、文案生成、内容续写等
- 图片方面：虚拟人创作、平面设计、AI营销等
- 视频方面：动画片生成、广告生成等
- 游戏方面：剧情生成、地图生成等
- 电商方面：智能购物客服、智能推送等

具体案例

- 昆仑万维：公司未来坚定选择AGI和AIGC赛道，目前已推出天工巧绘SkyPaint、天工乐府SkyMusic、天工妙笔SkyText、天工智码SkyCode等系列产品构建产品体系。公司还与奇点智源合作自研国产大语言模型“天工”3.5，但目前还未实现营收。
- 蓝色光标：公司陆续推出了苏小妹、K、虚拟国风音乐女团吾音等虚拟IP，MEME数藏平台以及蓝宇宙等虚拟空间和虚拟产品。AIGC技术产品方面，公司已发布包括分身有术、蓝标智播、创意画廊、销博特“创策图文”以及“萧助理”等多款产品。
- 中文在线：公司的海外产品Chapters和My Escape已在做接入ChatGPT测试，并推出AI主播、AI绘画和AI文字辅助创作三款AIGC相关产品。

数据来源：各公司官网，各公司年报，昆仑万维发布会，中国信通院《人工智能生产内容白皮书》，平安证券研究所

AIGC+金融：大模型可以用于股票预测、信用评估、风险管理等业务

金融机构数字化转型需求的不断提升，以及AIGC技术对金融领域相关信息化产品的强大赋能，使得AIGC+金融领域的应用发展十分迅速，BloombergGPT的发布让人们看到未来金融行业AIGC应用的强大。目前在金融领域，大模型可以用于股票预测、信用评估、风险管理等业务，也可用于建设数字人系统、短视频生成平台、自动研报生成系统等。

应用场景

- 机器学习&知识图谱：风险识别、客户信用能力评估、知识检索等
- 智能语音&对话式AI：智能客服、智能营销、智能培训等
- 自然语言处理：文案生成、投研分析、智能风控、投顾助手等
- 计算机视觉：智能安防、智能内控管理等

具体案例

- 同花顺：公司具有AI虚拟人、同花顺AI开放平台i问财等一系列人工智能产品，为金融行业持续提供数字虚拟人、智能客服机器人、智能质检机器人等40余项AI产品服务。
- 东方财富：公司陆续研发东方财富金融数据AI智能化生产平台、多媒体智能资讯及互动平台系统等多个人工智能相关项目，提供AI投资决策支持、AI智能定投等业务服务。
- 财富趋势：公司持续研发小达智能写手、问小达等产品，形成了一套系统的金融数据解决方案，并研发了智能金融问答、公司图谱等一系列特色AI功能，持续为公司金融产品赋能。

数据来源：各公司官网，各公司年报，平安证券研究所

AIGC+教育：助力精准教学和个性化学习

AIGC 赋予教育材料新活力，将对传统教育的全流程环节范式产生一定影响。相对于阅读和讲座等传统方式，AIGC 为教育工作者提供了新的工具，使原本抽象、平面的课本具体化、立体化，以更加生动、更加令人信服的方式向学生传递知识。例如制作历史人物，为教学注入新的活力；合成逼真的虚拟教师，让数字教学更具互动性和趣味性。

应用场景	具体案例
• 精准教学：智慧体育、智能语言学习、早教、智能批改、VR\AR教学等 • 自动评阅：口语考评、试卷分析、智能阅卷等 • 科学管理：智慧校园、智能排课、智慧安保、智慧招考等 • 个性化学习：自适应学习、拍照搜题、教育机器人、AI学习助手等	• 科大讯飞：公司发布的讯飞星火认知大模型已嵌入科大讯飞AI学习机，使学习机获得充分结合学生刚需、充分考虑产品教育属性、过滤与学习无关的信息的类ChatGPT体验。此外，公司还拥有智能评卷技术、个性化学习手册等产品。 • 佳发教育：公司拥有教考统筹方案、智慧校园整体方案设计、智慧教室、智慧体育等业务，推出AI摄像机、智能黑板、智能设备管理平台、AI体育智能测评系统、智能操场等产品。 • 好未来（学而思）：公司自研数学大模型MathGPT，以数学领域的解题和讲题算法为核心打造而成，后续将推出基于该自研大模型的产品级应用。此外，公司还拥有“AI老师监课系统”、“AI老师记单词”、“AI老师语言学习系统”以及“VR沉浸式课堂”等产品，为教学核心环节赋能。

数据来源：各公司官网，各公司年报，平安证券研究所

AIGC+医疗：赋能诊疗全流程，大幅提升诊后康复管理效率

AIGC赋能诊疗全流程，与医学影像诊断、慢性病管理和生活方式指导、疾病排查和病理研究、药物开发等多个医学领域具有较高契合度。目前，科大讯飞、卫宁健康等公司纷纷推出医疗大模型产品，AIGC+医疗开始落地。

应用场景

- 医疗器械：AI心电图、AI影响、AI放疗器械、医疗机器人等
- 医疗服务：智能检验、智能病理判断、电子病历、AI医疗助理等
- 医药研发：AI制药、AI合同研发服务（CRO）等
- 医疗管理：智慧医院、健康管理、专家系统管理等

具体案例

- 科大讯飞：公司拥有星火认知大模型，结合医疗场景推出诊后康复管理平台，应用AI技术分析病案并生成康复计划，诊后康复管理效率提升10倍以上，目前已与北京协和医院、四川华西医院、武汉同济医院等多家医院开展密切合作。
- 卫宁健康：公司拥有自研大语言模型WiNGPT，利用WiNGPT的大模型能力打造互联网问诊Copilot、医疗报告生成Copilot、PACS Copilot三大组件，基本完全囊括了医疗服务领域全流程布局。
- 创业慧康：公司在2022年推出AI医学影像分析平台、智能影像产品系列和健康管理系统三款AI产品，未来将在医疗影像识别、自然语言处理、医学知识图谱、智能辅助决策等领域实现产品的人工智能化，搭建临床科研一体化平台，实现临床科研一体化。

数据来源：各公司官网，各公司年报，中商产业研究院，平安证券研究所

AIGC+汽车：智能汽车的人机交互和用户体验将迎来新一轮变革

随着AIGC时代来临，借助于大模型的决策和推理能力，智能汽车的人机交互和用户体验将迎来新一轮变革。目前，各厂商通过AIGC技术努力提升智能汽车的智能座舱、自动驾驶等领域的产品体验，AIGC+汽车的应用场景空间广阔。

应用场景

- 智能座舱技术：
  （1）语音交互
  （2）车联网
  （3）OTA

- 自动驾驶技术：
  （1）图像识别与感知
  ①雷达传感器
  ②视觉传感器
  ③定位及位姿传感器
  ④车身传感器
  （2）深度学习
  ①无人驾驶软件系统
  （3）信息共享
  ①实时数据分享
  ①实时数据处理

- 智能车云服务技术

具体案例

- 科大讯飞：公司深度聚焦智能座舱领域，通过“大模型+智能座舱”提供不同车型的智能语音交互服务，能够回答与智能汽车相关的多种问题，让汽车驾驶更加智能。目前科大讯飞的汽车智能化产品合作已覆盖90%以上的中国主流自主品牌和合资品牌车厂。
- 中科创达：公司拥有自主研发的魔方Rubik大模型。在智能汽车领域，公司已在大模型智能助手、车机主题自动调节等领域实现了较为成熟的方案，未来公司将与多家头部客户合作将大模型技术应用于智能座舱、自动驾驶领域。公司推出基于大模型的Rubik GeniusCanvas（天才画布）辅助制作工业参考模型，提升汽车座舱HMI的设计效率与质量；公司还拥有提供软硬全栈的能力，Rubik Server和Rubik Box助力形成智能汽车等智能终端产品。

数据来源：各公司官网，各公司年报，平安证券研究所

CONTENT
目录

投资建议

- 展望下半年，我国大模型产品已经初步具备商用能力。我国北上深三地利好通用人工智能发展政策的发布，彰显了我国对于AIGC发展的重视和支持，同时将为我国其他城市发布类似政策带来示范效应。随着《人工智能法》列入《国务院2023年度立法工作计划》，我们判断，后续政策的出台将为我国AIGC产业的发展护航。在政策与技术的共振下，我国AIGC产业未来发展前景广阔。AIGC产业的发展需要大算力，我国AI芯片和AI服务器市场迎来发展机遇，相关芯片和服务器厂商将深度受益。我们坚定看好AIGC产业链的投资机会。维持对计算机行业的“强于大市”评级。在标的方面：1）算力方面，推荐浪潮信息、中科曙光、紫光股份、海光信息、龙芯中科，建议关注工业富联、寒武纪、景嘉微；2）算法方面：推荐科大讯飞，建议关注三六零；3）应用方面，推荐金山办公，建议关注拓尔思、彩讯股份、航天宏图；4）网络安全方面，强烈推荐启明星辰，推荐深信服、绿盟科技。

风险提示

- 1）合规风险上升。AIGC处在发展的早期，技术自身的漏洞以及不当应用都可能带来巨大的安全风险，如果监管趋严，可能对技术后续落地带来负面影响。
- 2）技术创新不及预期。AIGC是跨学科的产物，国内技术积累和研发同国际水平尚有差距，如果后续研发投入不足或者方向出现偏差，可能对国内人工智能发展造成迟滞影响。
- 3）供应链风险。AIGC未来有希望成为经济社会重要的基础设施，也是各国科技竞争的焦点。但是我国算力、平台和算法都不同程度依赖国际产业链，如果后续西方国家管制趋严，供应链风险将加剧。。

建议标的盈利预测及估值

股票简称	股票代码	6月14日	EPS（元）				PE（倍）				评级
		收盘价（元）	2022A	2023E	2024E	2025E	2022A	2023E	2024E	2025E	
浪潮信息	000977.SZ	52.50	1.41	1.72	2.07	2.50	37.2	30.5	25.4	21.0	推荐
中科曙光	603019.SH	55.41	1.05	1.36	1.74	2.21	52.8	40.7	31.8	25.1	推荐
紫光股份	000938.SZ	34.40	0.75	0.96	1.20	1.49	45.87	35.8	28.7	23.1	推荐
海光信息	688041.SH	81.48	0.35	0.50	0.67	0.84	232.8	163.0	121.6	97.0	推荐
龙芯中科	688047.SH	142.55	0.13	0.57	0.73	0.87	1096.5	250.1	195.3	163.9	推荐
科大讯飞	002230.SZ	75.13	0.24	0.68	0.89	1.18	313.0	110.5	84.4	63.7	推荐
金山办公	688111.SH	471.21	2.43	3.39	4.64	6.38	193.9	139.0	101.6	73.9	推荐
启明星辰	002439.SZ	31.91	0.66	0.91	1.22	1.62	48.3	35.1	26.2	19.7	强烈推荐
深信服	300454.SZ	126.95	0.47	0.88	1.12	1.38	270.1	144.3	113.3	92.0	推荐
绿盟科技	300369.SZ	13.75	0.04	0.35	0.50	0.57	343.8	39.3	27.5	24.1	推荐

股票投资评级：

强烈推荐（预计6个月内，股价表现强于沪深300指数20%以上）

推　　荐（预计6个月内，股价表现强于沪深300指数10%至20%之间）

中　　性（预计6个月内，股价表现相对沪深300指数在±10%之间）

回　　避（预计6个月内，股价表现弱于沪深300指数10%以上）

行业投资评级：

强于大市（预计6个月内，行业指数表现强于沪深300指数5%以上）

中　　性（预计6个月内，行业指数表现相对沪深300指数在±5%之间）

弱于大市（预计6个月内，行业指数表现弱于沪深300指数5%以上）

公司声明及风险提示：

负责撰写此报告的分析师（一人或多人）就本研究报告确认：本人具有中国证券业协会授予的证券投资咨询执业资格。

本公司研究报告是针对与公司签署服务协议的签约客户的专属研究产品，为该类客户进行投资决策时提供辅助和参考，双方对权利与义务均有严格约定。本公司研究报告仅提供给上述特定客户，并不面向公众发布。未经书面授权刊载或者转发的，本公司将采取维权措施追究其侵权责任。

证券市场是一个风险无时不在的市场。您在进行证券交易时存在赢利的可能，也存在亏损的风险。请您务必对此有清醒的认识，认真考虑是否进行证券交易。市场有风险，投资需谨慎。

免责条款：

此报告旨为发给平安证券股份有限公司（以下简称“平安证券”）的特定客户及其他专业人士。未经平安证券事先书面明文批准，不得更改或以任何方式传送、复印或派发此报告的材料、内容及其复印本予任何其他人。

此报告所载资料的来源及观点的出处皆被平安证券认为可靠，但平安证券不能担保其准确性或完整性，报告中的信息或所表达观点不构成所述证券买卖的出价或询价，报告内容仅供参考。平安证券不对因使用此报告的材料而引致的损失而负上任何责任，除非法律法规有明确规定。客户并不能仅依靠此报告而取代行使独立判断。

平安证券可发出其它与本报告所载资料不一致及有不同结论的报告。本报告及该等报告反映编写分析员的不同设想、见解及分析方法。报告所载资料、意见及推测仅反映分析员于发出此报告日期当日的判断，可随时更改。此报告所指的证券价格、价值及收入可跌可升。为免生疑问，此报告所载观点并不代表平安证券的立场。

平安证券在法律许可的情况下可能参与此报告所提及的发行商的投资银行业务或投资其发行的证券。

平安证券股份有限公司2023版权所有。保留一切权利。