

证券研究报告

行业评级：

计算机 强于大市（维持）

中国平安 PINGAN  
专业·价值

# 计算机行业深度报告： ChatGPT火爆出圈，AI算力和行业应用潜力巨大

平安证券研究所TMT团队

2023年2月28日

请务必阅读正文后免责条款

## 投资要点

- **焦点：ChatGPT带火AIGC，Open AI引领技术和生态。**2022年11月，由OpenAI开发的聊天机器人ChatGPT推出并席卷了整个行业，其引入人类反馈的强化学习和监督学习训练方法，具备优秀的聊天对话、文案创作、代码编写等功能，且得到了微软的倾力支持，成为史上用户增长最快的消费应用。聚焦国内，百度在大模型领域积累深厚，总体水平处于行业前端，其创新性引入大规模知识图谱，模型性能得到大幅提升，公司近期宣布其“文心一言”产品将于2023年3月完成内测并面向公众开放，下游关注度极高，目前已经得到新闻传媒、互联网、金融等多个行业客户的广泛支持。
- **变革：AIGC与人更为神似，模型和数据是主要助力。**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。在模型方面，Transformer预训练大模型为生成式AI带来了里程碑式飞跃，其中OPENAI发布的GPT系列大模型掀起AIGC热潮，尤其是ChatGPT火爆出圈。在数据集方面，数据量、多样性、数据质量是训练数据集的关键要素。
- **市场：大模型需要大算力，推动AI服务器市场增长。**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。ChatGPT的训练成本和推理成本高昂。我们认为，随着模型的迭代和AI芯片厂商产品的迭代，ChatGPT的训练成本和推理成本未来必将呈现下降趋势。但一次训练百万美元量级的训练成本和每天百万美元量级的推理成本，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服务器市场带来约910.44亿美元和345.50亿美元的市场空间。市场空间巨大，相关芯片和服务器厂商将深度受益此次ChatGPT浪潮。
- **应用：行业将逐步回归理性，能否突破需要看B端。**未来2-3年是行业应用落地的关键时段，AIGC在经历了近期的炒作热潮结束之后，预计将经历一段下沉期，市场趋向理性。期间，AIGC需要加快教育和融入市场，培育产品和应用。其中，B端是AIGC含金量最大的市场，AIGC需将技术转化成工具和解决方案，为企业和行业赋能。其中，文本、图像生成以及类聊天机器人等产品在B端都有较大应用潜力，游戏、金融和快消等市场有望率先取得进展。
- **投资建议：**近年来AIGC的算法、数据质量等要素有着显著提升，文本、图片/视频、代码等技术能力快速发展，在客服、营销、游戏等赛道都显示出非常强的应用潜力。伴随着chatGPT的催化，国内AIGC产业链也将迎来空前的市场机会。算法方面，预计今年国内龙头AI厂商有望推出能力接近chatGPT的产品；算力是短板，由于供应链的问题，我国高端训练芯片供给可能面对较大的压力，AI算力不足的问题可能显现，相关设备研发和基础设施建设企业可能继续受益。同时，AIGC新技术也会带来新的安全挑战，国产安全厂商也将受益。强烈推荐启明星辰，推荐科大讯飞、金山办公、中科曙光、浪潮信息、紫光股份等；关注AI厂商后续发展机会，如云从科技、海天瑞声、云天励飞（待上市，已同意注册）。
- **风险提示：**1) 技术落地可能不及预期。2) 来自监管和版权方面的风险。3) 国内AI供应链断裂的风险。

# 目录

- 1 焦点：ChatGPT带火AIGC，Open AI引领技术和生态
- 2 变革：AIGC与人更为神似，模型和数据是主要助力
- 3 市场：大模型需要大算力，推动AI服务器市场增长
- 4 应用：行业将逐步回归理性，能否突破需要看B端
- 5 投资建议及风险提示



## 焦点：凭借AIGC爆款应用ChatGPT，Open AI强势出圈

- ✓ 2022年11月30日，由OpenAI开发的聊天机器人ChatGPT推出并席卷了整个行业，短短5天，注册用户数就超过100万，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。ChatGPT拥有强大的语言理解和文本生成能力，能够很好地支持聊天对话、代码撰写、文案创作等任务。
- ✓ 与之前的AI聊天机器人相比，ChatGPT能够理解上下文并给出合理的回复，对话连续性大幅提升，可以称为人工智能发展史上的划时代产品。凭借着ChatGPT的成功，Open AI成为全球AI行业关注的焦点。

### ■ 各软件用户数达到100万所用的时间



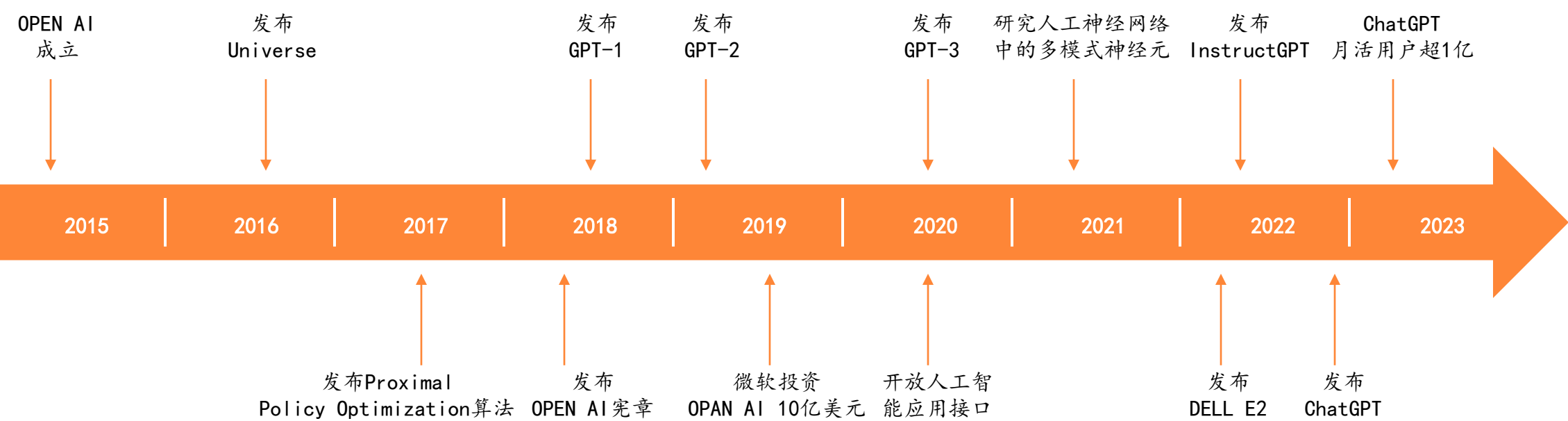
### ■ ChatGPT写诗、评论和代码生成等样例



# OPEN AI：GPT模型技术领先，ChatGPT发布引发变革

- ✓ AIGC (AI Generated Content) 是人工智能生成内容，也称为生成式AI，涉及无监督和半监督学习算法，能够使计算机用先前创建的内容（诸如文本、音频、视频、图像和代码等）来生成新的内容，以此响应用户输入的提示语。
- ✓ OPEN AI是全球领先的AIGC公司。OPEN AI成立于2015年，其最初定位为“非盈利性研究机构”，2018年，OPAN AI发布了GPT-1模型，2019年，公司改制为“有上限的盈利性机构”，并于同年发布GPT-2模型；2022年3月，OPAN AI发布InstructGPT模型；2022年11月，OPEN AI发布ChatGPT。此外，公司还拥有自然语言转图片的应用DELL E2。

## ■ OPEN AI发展历程

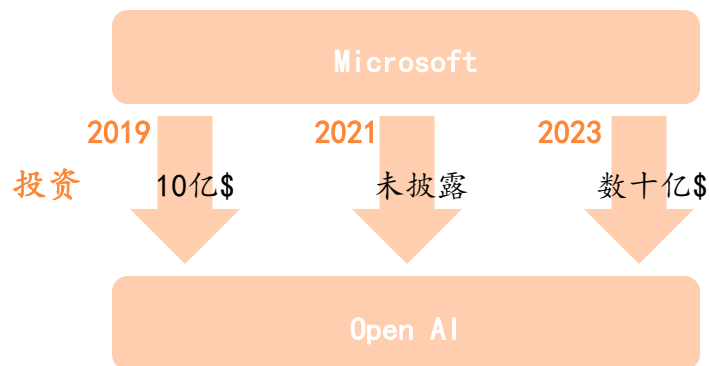


资料来源：百度百科，平安证券研究所

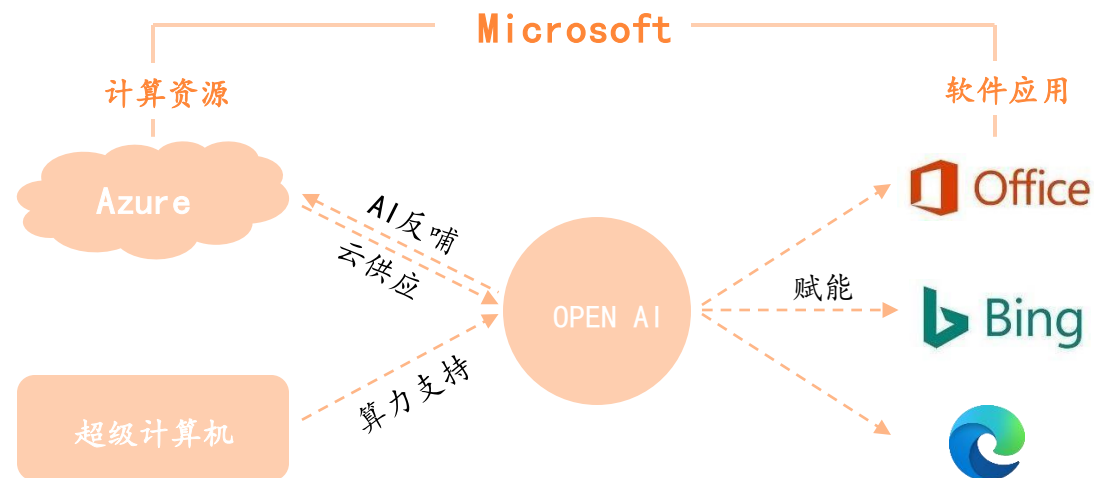
# OPEN AI：微软持续提供资金、算力和生态支持，助力成就爆款

- ✓ 微软高度重视人工智能领域，在资金投入、计算资源、应用开发等方面与OPEN AI形成了密切的合作关系，两者优势互补、合作共赢。
- ✓ 资金投入：2019年，微软向OPEN AI投资10亿美元；2021年，微软再次向OPEN AI投资，数额未披露；2023年1月，微软宣布将通过一项花费数年、数十亿美金的投资深化与OPEN AI的合作，根据美国财富杂志信息，该投资可能高达100亿美元。
- ✓ 计算资源：微软Azure是OPEN AI的独家云供应商，且微软将加大在超级计算机方面的投资以支持OPEN AI的发展，同时OPEN AI也可以反哺增强微软Azure的AI能力，两者能够很好的合作共赢。
- ✓ 应用开发：2月8日，微软推出基于OPEN AI模型的Edge浏览器和Bing搜索引擎，且微软表示未来所有产品将全线整合ChatGPT，包括Office三件套、Azure云服务、企业定制版产品等。

■ 微软投资OPEN AI历史



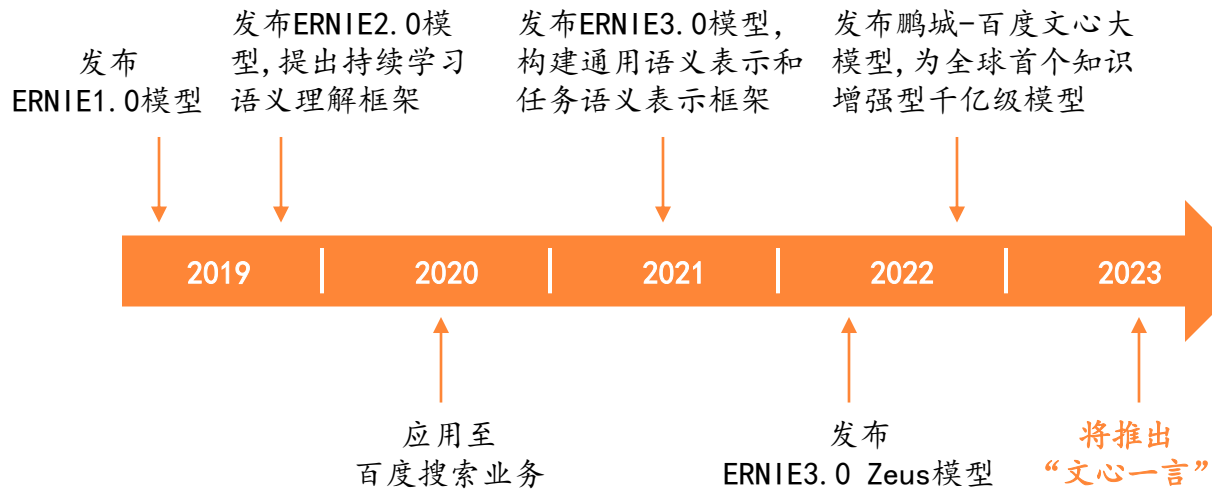
■ OPEN AI与微软在计算资源、应用开发方面的合作关系



# 百度：文心大模型国内全面领先，有望打造成国产ChatGPT

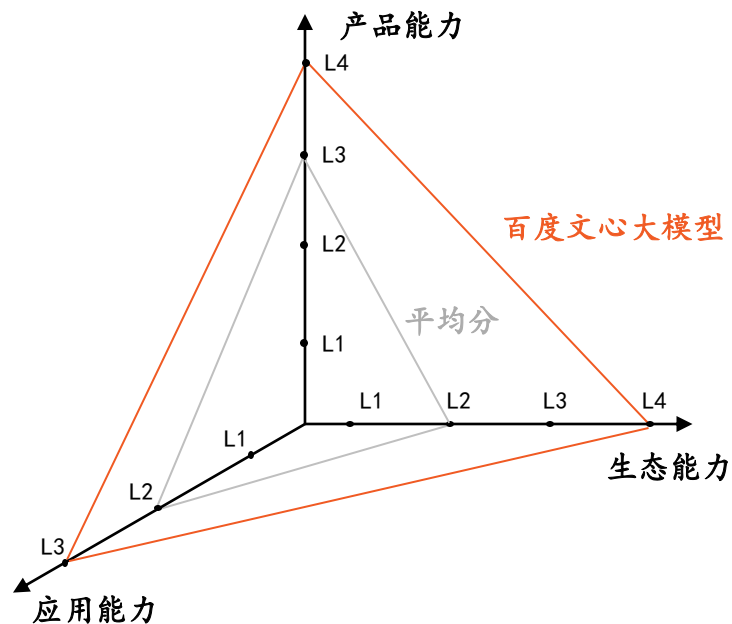
- ✓ 百度在大模型领域拥有较强的技术实力和平台积累，其文心大模型总体位于行业前列。2019年3月，百度首次发布预训练模型ERNIE 1.0，之后针对大模型的技术创新与产业应用持续发力，目前已经实陆续实现了一定程度的突破：2021年，百度发布最新版大模型ERNIE 3.0；2022年，百度发布鹏城-百度文心大模型，其参数规模达到2600亿，较GPT-3参数量高50%；2023年2月，百度宣布其最新的大模型“文心一言”将于三月份完成内测，并面向公众开放。
- ✓ 根据IDC对国内大模型市场的评估，百度文心大模型在产品能力、应用能力、生态能力等方面处于全面领先的业界地位。

## ■ 百度文心大模型发展历程



资料来源：百度，IDC，平安证券研究所

## ■ 中国大模型市场2022年评估结果-百度文心



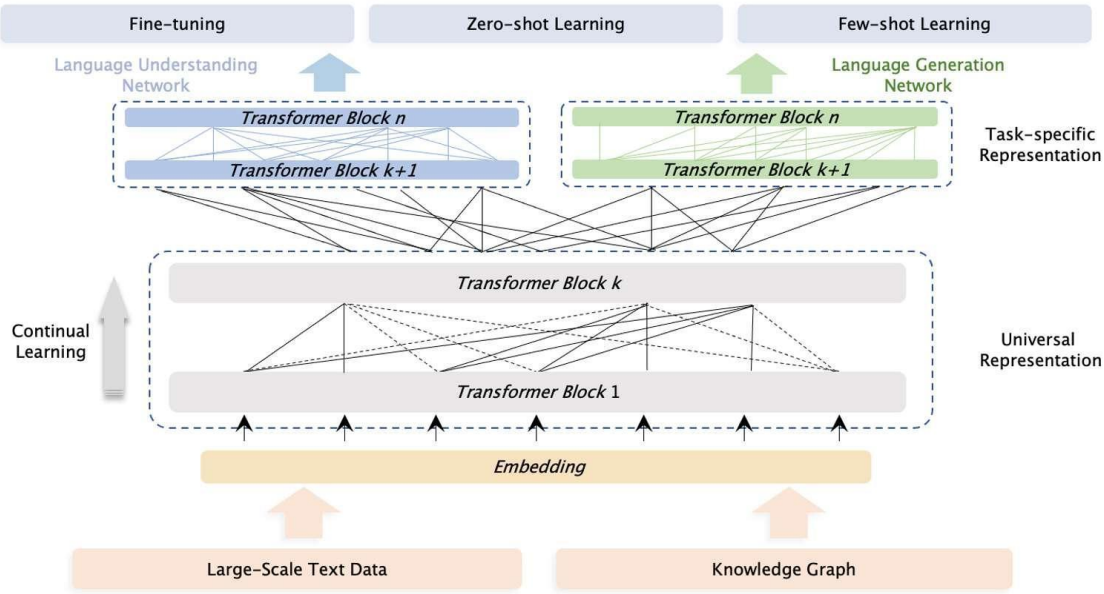
# 百度：创新性引入大规模知识，文心大模型性能大幅提升

- ✓ 百度文心大模型拥有NLP、CV、跨模态、生物计算等大模型组合，具有产业级和知识增强两大特点。文心大模型源于产业、服务于产业，可以满足真实场景中的诸多应用需求，真正发挥大模型驱动AI规模化应用的产业价值；文心大模型在海量文本数据的基础上引入大规模知识图谱，促进了结构化知识和无结构化文本之间的融合与共享，模型能力大幅提升。
- ✓ 2021年百度发布了ERNIE3.0模型，首次在百亿级预训练中引入了大规模知识，模型学习效率和可解释性大幅增强。框架方面，ERNIE 3.0模型框架分为“通用表示”和“任务表示”两层，同时具备语言理解和语言生成两种功能；性能方面，ERNIE 3.0刷新了54个中文NLP任务基准，且其英文模型在复杂语言理解任务评测中超过了人类水平0.8个百分点。

■ 百度文心大模型全景图



■ ERNIE3.0框架结构



资料来源：百度，ERNIE 3.0: LARGE-SCALE KNOWLEDGE ENHANCED PRE-TRAINING FOR LANGUAGE UNDERSTANDING AND GENERATION，消费日报网，平安证券研究所



# 百度：“文心一言”生态建设取得进展，具备商用拓展潜力

- ✓ 文心NLP大模型面向语言理解、语言生成等场景，具备超强的语言理解、对话生成、文学创作等能力。文心NLP模型将大数据预训练与多源丰富知识相结合，通过持续学习，不断吸收海量文本数据中词汇、结构、语义等方面的新知识，实现模型效果不断进化。
- ✓ “文心一言（英文名：ERNIE Bot）”是百度基于文心大模型技术推出的生成式对话产品。2023年2月7日，百度宣布将于2023年3月份发布“文心一言”，目前已经有大量下游客户参与文心一言的生态建设，行业覆盖面广泛，包括新闻、传媒、互联网、家装、汽车、金融等，其中新闻传媒业积极性最高，反映出“文心一言”这类AIGC产品将给媒体创作类行业带来巨大影响。在2023 AI+工业互联网高峰论坛上，百度智能云宣布“文心一言”将通过百度智能云对外提供服务，率先在内容和信息相关的行业和场景落地。

## ■ 文心NLP大模型



ERNIE 3.0 Zeus



ERNIE 3.0



鹏城-百度文心



ERNIE 3.0-Tiny



ERNIE-Health



ERNIE-Finance



PLATO



ERNIE-UIE



ERNIE-Search



ERNIE-M



ERNIE-Code



ERNIE-Sage

## ■ “文心一言”生态合作伙伴



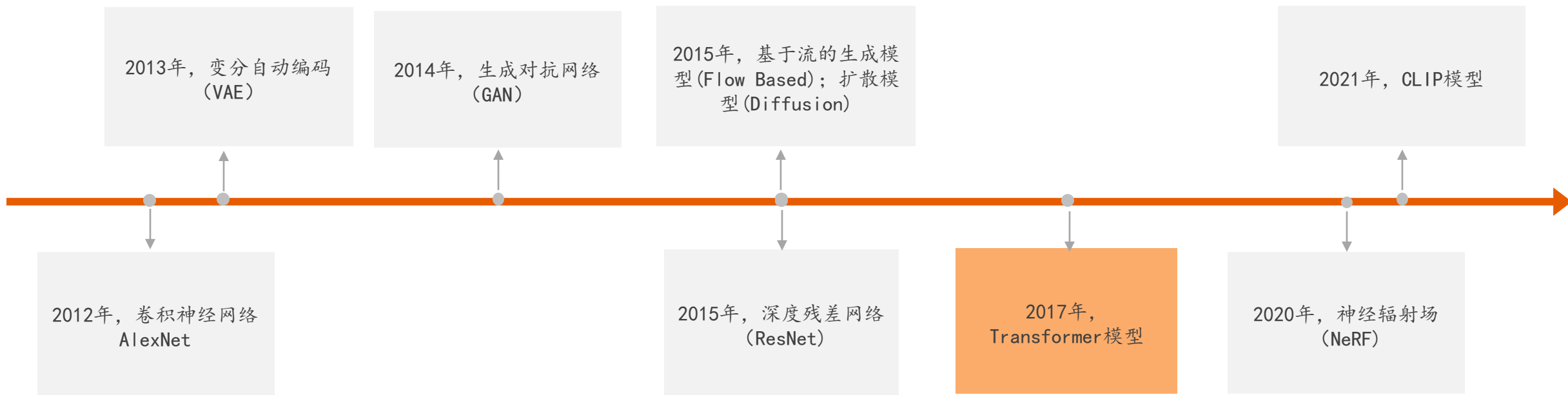
# 目录

- 1 焦点：ChatGPT带火AIGC，OpenAI引领技术和生态
- 2 变革：AIGC与人更为神似，模型和数据是主要助力
- 3 市场：大模型需要大算力，推动AI服务器市场增长
- 4 应用：行业将逐步回归理性，能否突破需要看B端
- 5 投资建议及风险提示

## 改变：AIGC实现了从分析预测到生成创造的跨越，AI开始更像人

- ✓ **AIGC为人工智能技术带来技术变革。**相比于之前的分析式AI，AIGC不再局限于分析已有数据的规律，而是实现了从分析内容到创造生成新内容的跨越，让AI更像人。事实上，早在2014年，随着一种深度学习架构生成对抗网络（GAN）的出现，生成式AI就开始流行。但近年来，生成式AI在模型、数据、算力都大幅度提升和改进，行业进入爆发期，其中模型的演进影响最为深远。

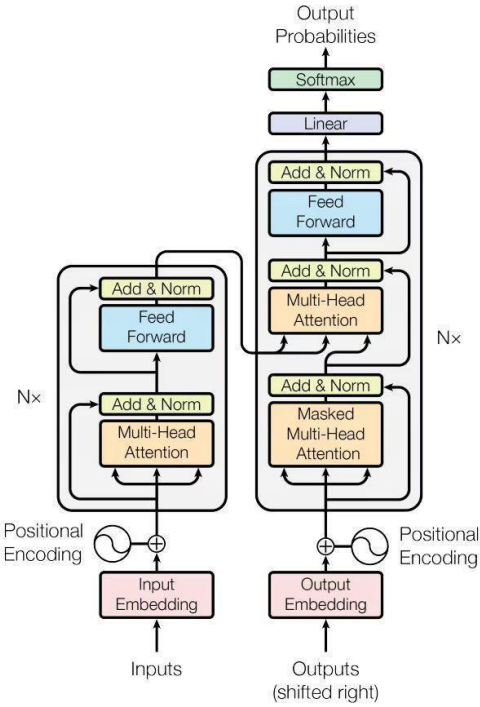
### ■ NLP发展历程



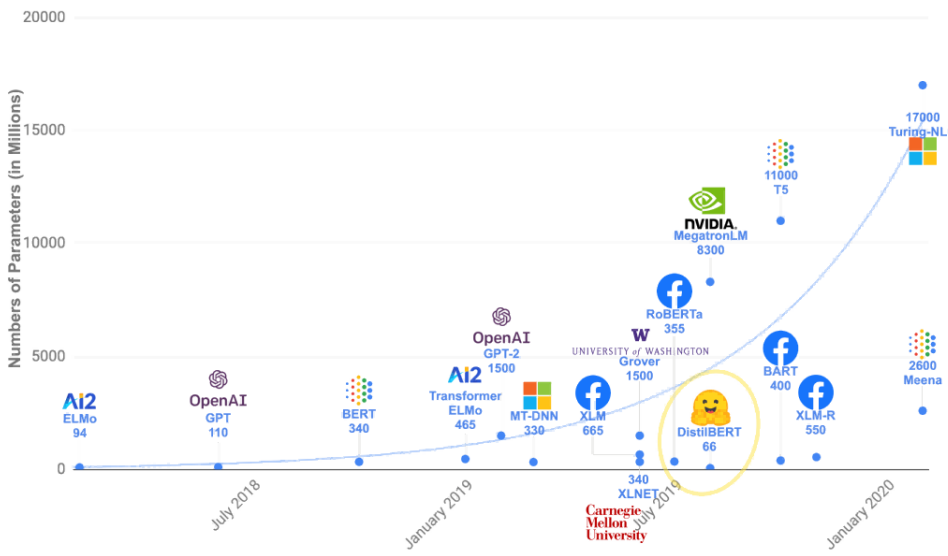
# 突破|模型：预训练大模型Transformer助力，AIGC实现里程碑式飞跃

- ✓ 预训练大模型为生成式AI带来里程碑式飞跃。2017年Transformer架构的出现改善了循环神经网络(RNN)的局限性，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，支持GPT-3和LaMDA等大型语言模型(LLM)。Transformer模型可并行训练，使GPU性能改善所带来的模型训练效果显著提升，同时可以通过不断增加模型参数量以及预训练使用的数据量来提升模型性能。
- ✓ 大模型（又称基础模型）可经过海量、多样化数据（通常无标注文本）的预训练，再针对广泛的下游任务进行微调或适应。针对不同的任务和应用场景，只需要将大模型迁移学习到下游任务即可实现，避免了传统NLP技术需要从头开始训练下游任务的痛点。

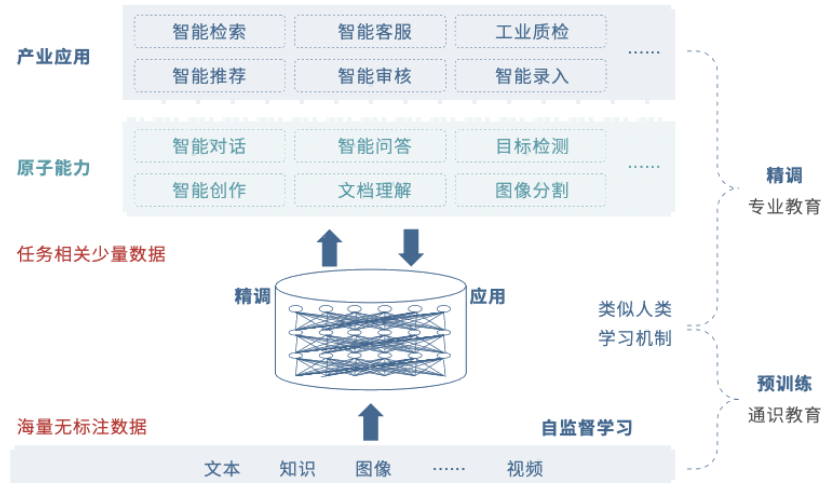
■ Transformer模型架构



■ 模型所需参数量变化趋势



■ 训练大模型“预训练+精调”模式

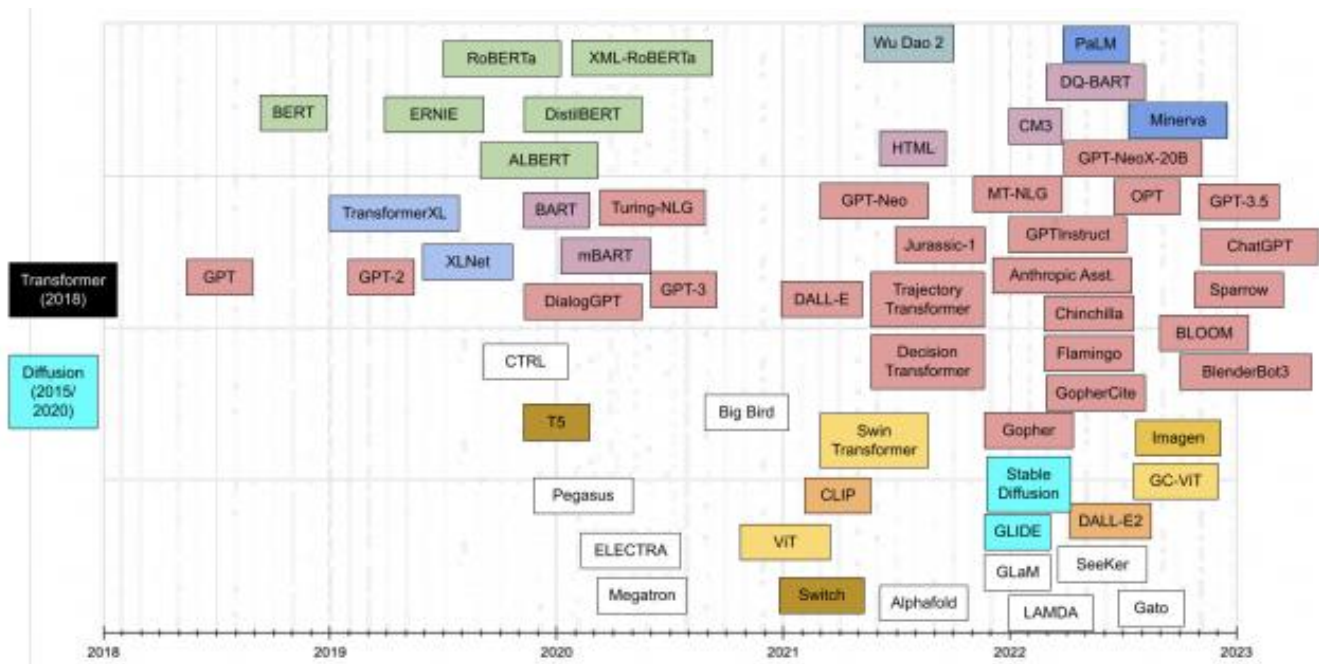




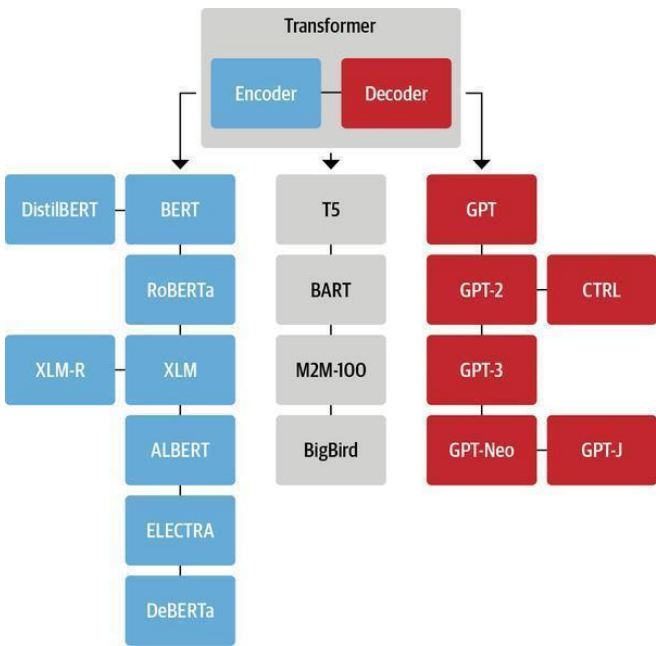
# 突破|模型：Transformer衍生出三类模型，打下生成式AI算法的基础

- ✓ 深度神经网络结构升级引发AIGC技术开发新范式，各类预训练模型层出不穷，可用于NLP（诸如GTP-3）、CV(诸如Florence)或多模态领域（诸如Stable Diffusion）。
- ✓ 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BERT为代表）：又称自编码模型，适用于内容理解任务，例如需要理解输入语义的任务，例如情感分析；2) Decoder模型（以GPT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T5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。

■ Transformer模型家族时间轴



■ Transformer模型系列分类



# 突破|模型：GPT系列大模型演进和应用较快，最终掀起AIGC热潮

✓ **Transformer**三条路线中，**GPT模型**进展较快：2018年，由OpenAI提出的生成式预训练模型GPT诞生，提出了半监督学习方法，即通过“预训练+微调”的方式让模型通过大量无标注数据学习，从而缓解人工标注数据有限的问题；2019年，GPT-2去除finetune，参数量增加，采用zero-shot（零样本）学习，模型的泛化能力提升；2020年，GPT-3采用few-shot（小样本），并将训练参数又提升两个数量级，模型准确率和性能再次提升；2022年，InstructGPT在GPT-3基础上采用奖励机制，通过人为标注和强化学习的方法提升模型输出结果真实性、无害性和有用性，13亿参数版本实现了比1750亿参数版本GPT-3更好的模型性能。

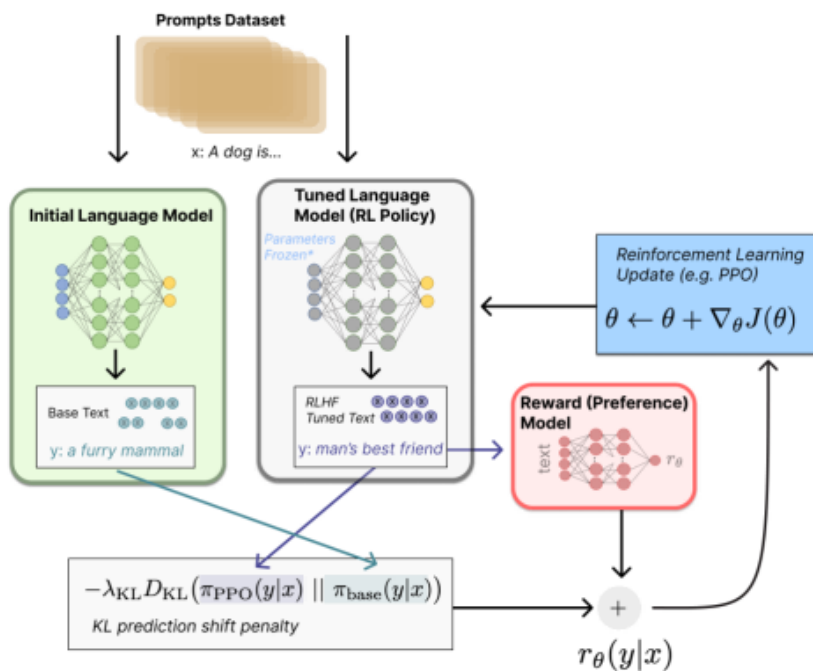
■ GPT系列模型比较

AI模型	发布时间	模型层数	词向量维度	数据量	参数量	说明
GPT	2018.6	12	768	约5GB	1.17亿	无监督预训练模型+有监督微调
GPT-2	2019.2	48	1600	40GB	15亿	无监督预训练模型，去除GPT的微调,架构上改进（将层归一化移动到每个子层的输入并在最终的自注意力块后增加一层归一化，将上下文窗口大小从512增加到1024）
GPT-3	2020.5	96	12888	45TB	1750亿	无监督预训练模型，参数量相比GPT-2提升两个数量级，架构上优化（将上下文窗口大小从1024增加到2048，采用交替密度和局部带状稀疏注意模式）
InstructGPT	2022.3				13亿	从预训练的GPT3模型开始，并在有监督微调后通过强化学习加入奖励模型机制
GPT-3.5						GPT3.5系列包括一系列诸如Davinci-003等模型，它们是在InstructGPT的基础上构建的
ChatGPT	2022.11					ChatGPT采用GPT3.5（又名GPT3 Davinci-003）预训练模型，并使用RLHF微调，这与InstructGPT中大致相同，但在数据收集方面略有不同。ChatGPT也不仅只是个模型，因为它包括类似于BlenderBot3的内存存储和检索的扩展。

# 突破|模型：文本、图像生成是起点，多模态生成也正在兴起

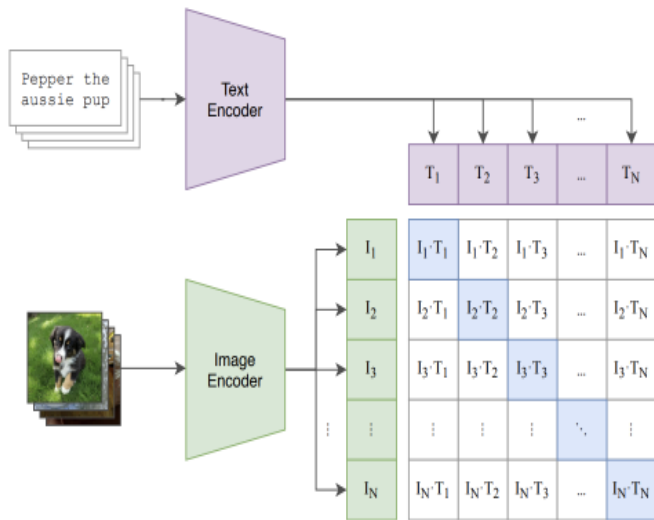
- ✓ ChatGPT建立在GPT-3.5大语言模型（LLM）系列之上，并通过监督学习和强化学习技术进行了微调，引入基于人类反馈的强化学习机制（Reinforcement Learning from Human Feedback, RLHF）。ChatGPT能完成问答、诗歌创作、代码编写、文案写作、翻译等，接近真实人类的思维方式，在生成效果上取得重大突破。
- ✓ 除了文本、代码生成之外，图片、视频甚至是图文多模态模型也都开始应用，而且应用规模将大幅增加。多模态大模型能够实现文本、语音、图像、视频等各种模态之间的融合和内容形式生成。例如，OPENAI开发的CLIP模型可同时将文字和图片相匹配。

## ■ RLHF强化学习原理步骤

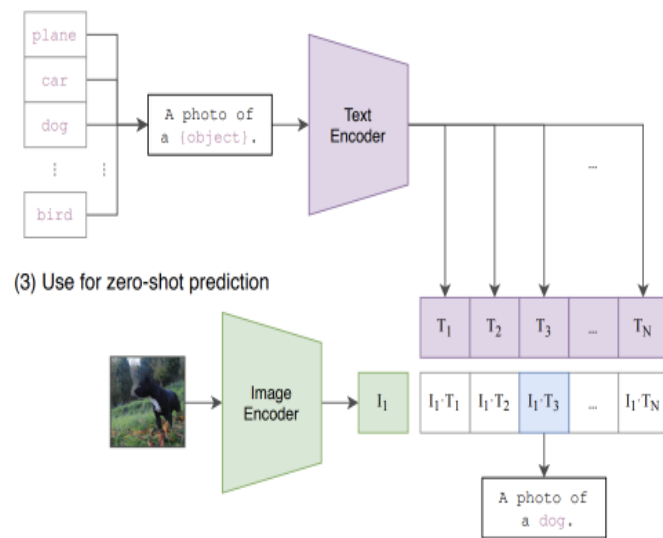


## ■ 基于CLIP的对比图文多模态预训练

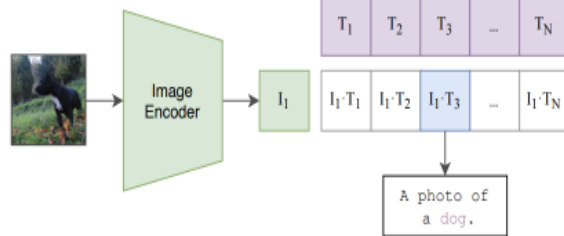
### (1) Contrastive pre-training



### (2) Create dataset classifier from label text



### (3) Use for zero-shot prediction



## 突破 | 数据集：数据量、多样性、数据质量是关键要素

- ✓ 语料库的数据量、多样性、数据质量成为训练数据集的关键要素。GPT使用了包含7000本书的BookCorpus数据集；GPT-2则收集了更加广泛、数量更多的语料组成数据集，包含800万篇Reddit上高赞的文章网页，大小为40GB，Reddit上的数据会包括各个领域，所以既保证了数据质量、数量又保证了数据的多样性；GPT-3则采用了5种语料库（5000亿tokens），大小增至45TB；InstructGPT的预训练和GPT3相同，但用标注数据和提示语进行了微调和优化，OPENAI雇佣了40名标注人员（labeler）且进行了培训，通过指示学习构建训练样本来训练奖励模型，最后通过奖励模型的打分排序来指导强化学习模型的训练。

### ■ GPT系列模型数据集概览



#### GPT

- 语料库：BookCorpus数据集（7000本书）上的无监督预训练，在几个特定任务数据集上有监督微调，包括SNLI、RACE、Quora等
- 数据量：5GB



#### GPT-2

- 语料库：WebTex（包含800万个Reddit上高赞的文章网页）
- 数据量：40GB



#### GPT-3

- 语料库：500B tokens 包含 Common Crawl (410B), WebText2 (19B), Books1 (12B), Books2 (55B), Wikipedia (3B)
- 数据量：45TB



#### InstructGPT

- 语料库：预训练同GPT3语料库，但用标注数据和提示语进行了微调和优化
- SFT 数据集包含 13K 的 prompts; RM数据集包含33K的 prompts; PPO数据集包含 31K 的 prompts。总共77K的数据中涉及人工标注的有46K左右。



# 突破|数据集：增加数据集采样频率且加入人工标注，数据质量提升

- ✓ OPENAI发现Common Crawl虽然包含一万亿个词，但未经过滤或轻微过滤的版本数据质量偏低。鉴于此，他们参照高质量语料将Common Crawl的45TB数据压缩过滤到570GB，并且添加了几个高质量的精选语料库（如Wikipedia）。
- ✓ 值得注意的是，为了得到更高质量的训练数据，数据集不是按其大小成比例采样，而是质量更高的数据集被更频繁地采样。而InstructGPT与GPT-3最大的区别在于引入RLHF，训练SFT、RM、PPO模型的数据集既有来自API用户的，也有来自标注人员的。其中SFT、RM的prompts来自于在线API上的用户使用数据以及标注人员标注的数据，PPO则全都是从API数据中采样的。虽然人类标注只有46K，但是InstructGPT对于没见过的prompt依然有较好的泛化能力。

■ GPT3模型数据集

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

■ InstructGPT各模型数据集大小（单位:prompts的数量）

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

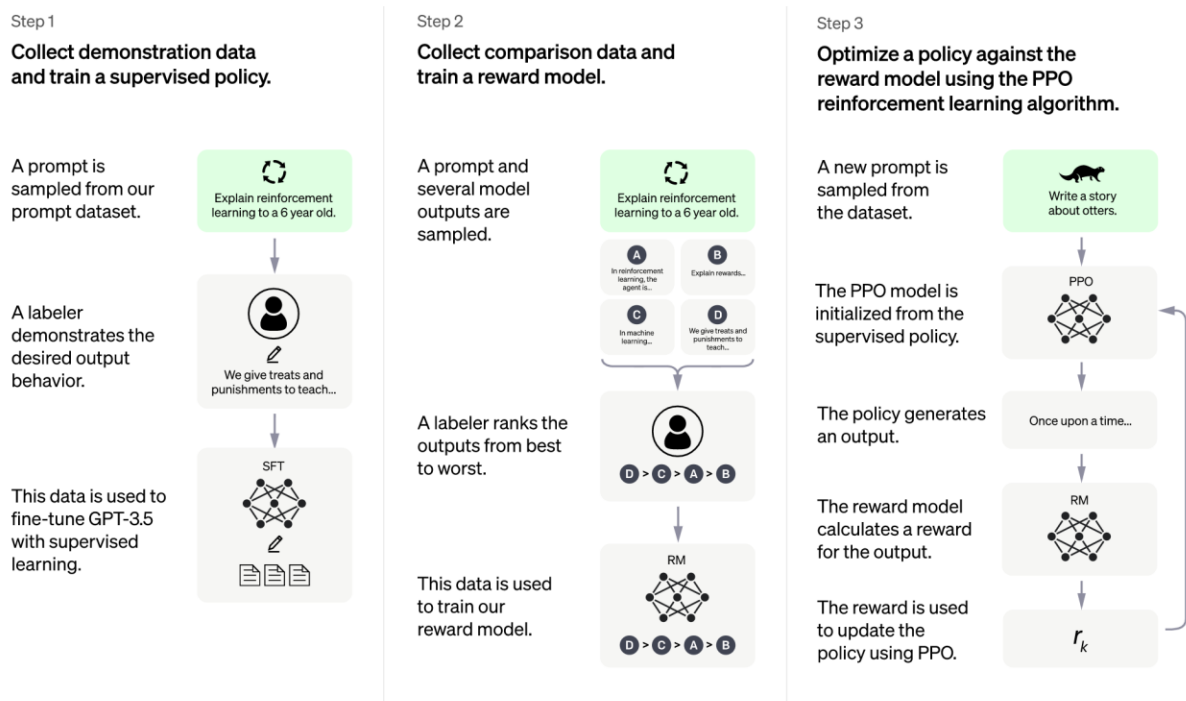
# 目录

- 1 焦点：ChatGPT带火AIGC，OpenAI引领技术和生态
- 2 变革：AIGC与人更为神似，模型和数据是主要助力
- 3 市场：大模型需要大算力，推动AI服务器市场增长
- 4 应用：行业将逐步回归理性，能否突破需要看B端
- 5 投资建议及风险提示

# 算力：ChatGPT是基于InstructGPT模型，大幅扩大数据量级而得到

- ✓ ChatGPT模型训练结合了监督学习和强化学习，具体的训练方法分为三步：1) 收集数据集并进行人工标注，输入到GPT-3.5模型中进行GPT-3.5模型微调，训练输出一个监督模型（SFT）；2) 收集SFT模型生成的多输出比较数据集，对诸多答案进行排序打分（人工标注），训练输出一个奖励模型（RM）；3) 利用RM模型作为奖励函数，通过PPO算法对SFT模型进行强化学习，持续迭代生成模型，此过程无需人工标注。

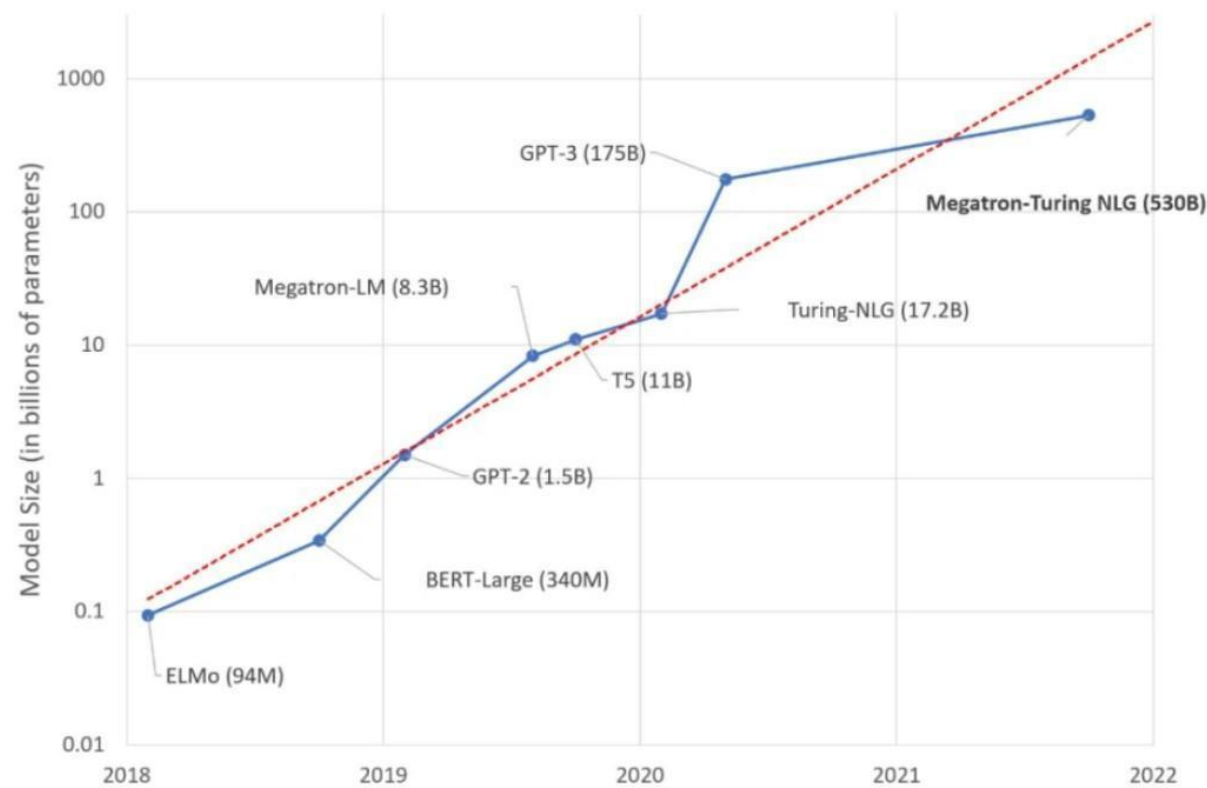
## ■ ChatGPT模型训练过程



# 算力：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



资料来源：澎湃新闻，OPENAI，平安证券研究所

各个模型所需计算量及参数量

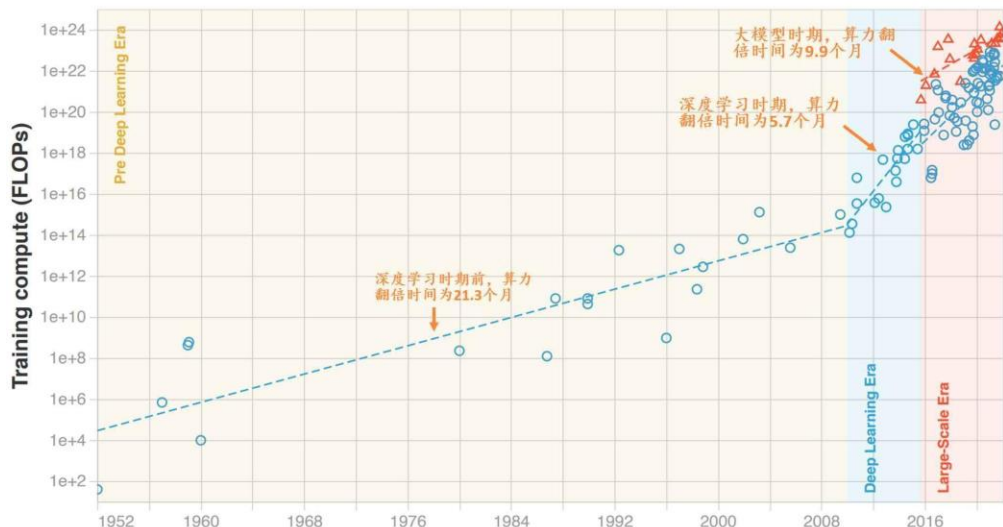
	模型	总计算量(PF-days)	总计算量(flops)	参数量(百万个)	训练令牌量(十亿个)
T5模型	T5-Small	2.08	1.80E+20	60	1,000
	T5-Base	7.64	6.60E+20	220	1,000
	T5-Large	26.7	2.31E+21	770	1,000
	T5-3B	104	9.00E+21	3,000	1,000
	T5-11B	382	3.30E+22	11,000	1,000
BERT模型	BERT-Base	1.89	1.64E+20	109	250
	BERT-Large	6.16	5.33E+20	355	250
	RoBERTa-Base	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



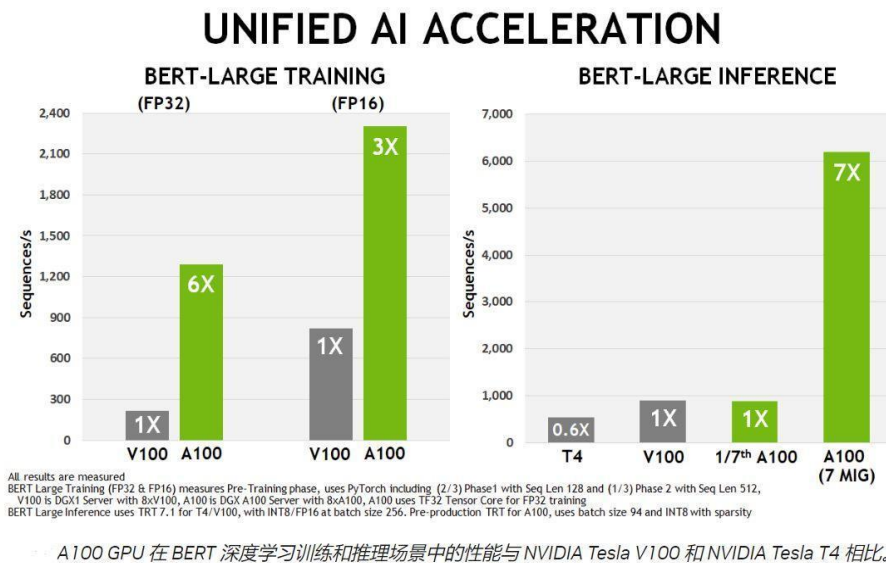
# 算力：AI芯片的性能提升成为决定大模型从理论实践到大规模应用的关键要素

- ✓ 根据《COMPUTE TRENDS ACROSS THREE ERAS OF MACHINE LEARNING》研究结果，大模型时代，算力翻倍的需求时间是9.9个月。AI芯片的性能提升成为决定大模型从理论实践到大规模应用的关键要素。
- ✓ 在供给端，AI芯片厂商推出了性能越来越强的新产品来应对人工智能发展的算力需求。以英伟达为例，2020年，英伟达推出了A100 GPU芯片，相比上一代V100 GPU芯片，A100 GPU芯片性能大幅提升。针对大模型，A100 GPU芯片相比V100 GPU芯片可提供高达3倍的训练速度。

## ■ 人工智能不同时代对算力翻倍的需求时间



## ■ 英伟达A100 GPU的性能远超V100 GPU



A100 GPU 在 BERT 深度学习训练和推理场景中的性能与 NVIDIA Tesla V100 和 NVIDIA Tesla T4 相比。

# 算力：ChatGPT的训练成本和推理成本高昂

- ✓ 在训练端：据报道，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

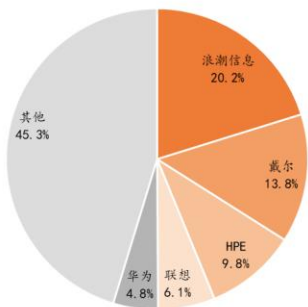
日活用户数（万）	单用户每日提问次数	每个问题平均字数（个）	A100 GPU对每个字的响应时间（毫秒）	每日消耗GPU计算时间（小时）	每天需要A100 GPU芯片（个）	NVIDIA DGX A100服务器（台）	NVIDIA DGX A100服务器价格（万美元）	推理成本（亿美元）
2000	10	20	350	388889	16204	2026	19.9	4.03

资料来源：澎湃新闻，环球零碳，英伟达官网，平安证券研究所

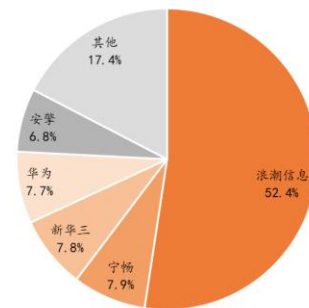
# 算力：大模型将为全球和中国AI芯片和AI服务器市场的增长提供强劲动力

- ✓ 我们认为，随着模型的迭代和AI芯片厂商产品的迭代，ChatGPT的训练成本和推理成本未来必将呈现下降趋势。但一次训练百万美元量级的训练成本和每天百万美元量级的推理成本，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。假设其中有2%的企业自建IDC训练和推理大模型，如前所述，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，以此估算，大模型将为我国AI服务器市场带来约345.50亿美元的市场空间。以2021年我国AI服务器市场规模占全球AI服务器市场规模的占比估算，则将为全球AI服务器市场带来约910.44亿美元的市场空间。市场空间巨大，相关芯片和服务器厂商将深度受益此次ChatGPT浪潮。

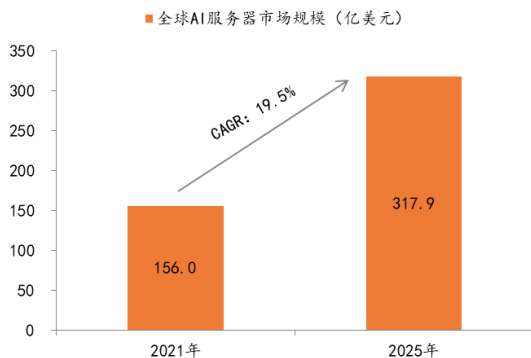
## ■ 2021H1全球AI服务器市场竞争格局



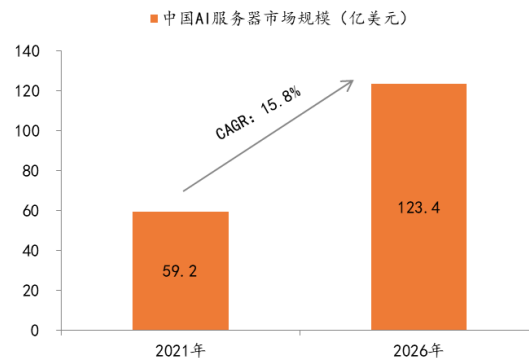
## ■ 2021年中国AI服务器市场竞争格局



## ■ 2021-2025年全球AI服务器市场规模CAGR为19.5%



## ■ 2021-2026年中国AI服务器市场规模CAGR为15.8%



资料来源：IDC，浪潮信息，平安证券研究所

# 目录

1 焦点：ChatGPT带火AIGC，OpenAI引领技术和生态

2 变革：AIGC与人更为神似，模型和数据是主要助力

3 市场：大模型需要大算力，推动AI服务器市场增长

4 应用：行业将逐步回归理性，能否突破需要看B端

5 投资建议及风险提示

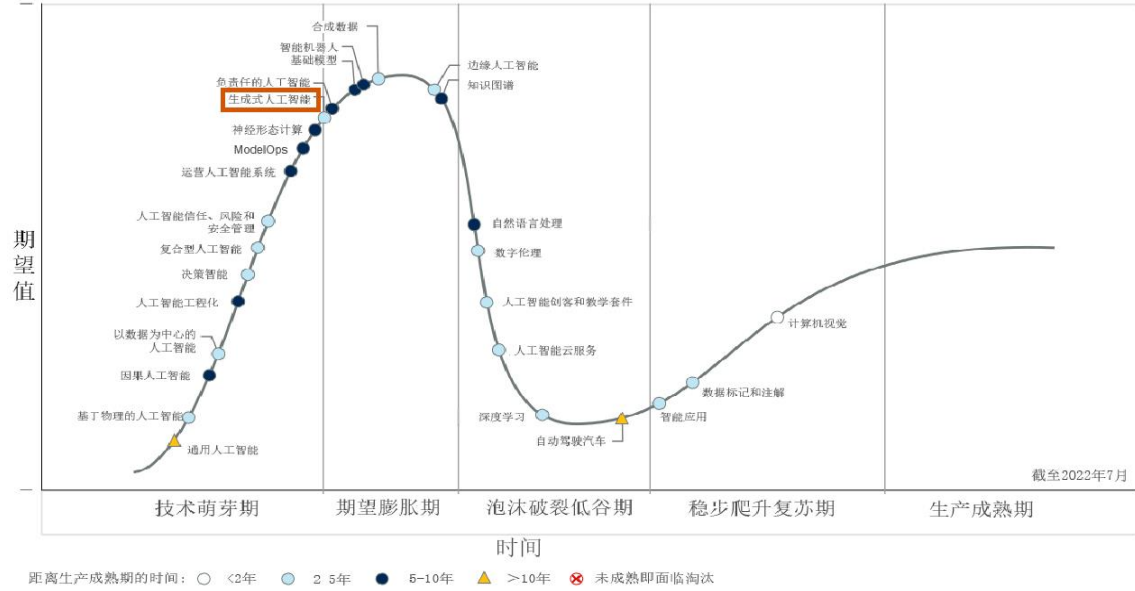


# AIGC将逐步回归理性，未来2-3年重点培育应用和教育市场

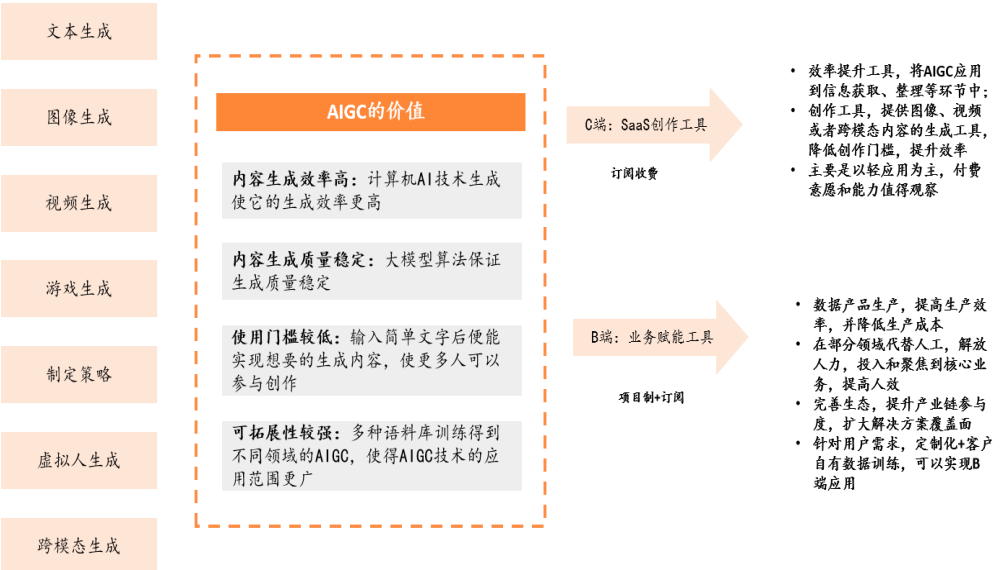
- ✓ 未来2-3年是AIGC走向成熟的关键时期。类似于计算机视觉等相对成熟的技术，AIGC在经历了近期的炒作热潮结束之后，预计行业将经历一段下沉期，市场趋向理性。AIGC则需要加快开始教育和融入市场，培育产品和应用。如果能够像计算机视觉一样，找到合适的商业化场景，行业后续将逐步得到市场认可，并持续做大；如果不能，这项技术也可能持续停留在低谷期，甚至被市场淘汰。
- ✓ ChatGPT的快速推进，标志着AIGC在C端有着巨大潜力，但B端、G端相关产品应该将是公司的主要收入来源。C端看，主要是来自于内容生成等工具性的需求，一般都是轻应用，尝鲜之后的付费意愿值得观察，主流的产品收费模式将是SaaS应用订阅；B端看，是含金量最大的市场，需要将技术转化成工具和解决方案，为企业和行业赋能，收费模式可能包括项目定制+后续订阅收入。

## 生成式AI走向炒作顶峰

2022年人工智能技术成熟度曲线



## AIGC潜在商业模式

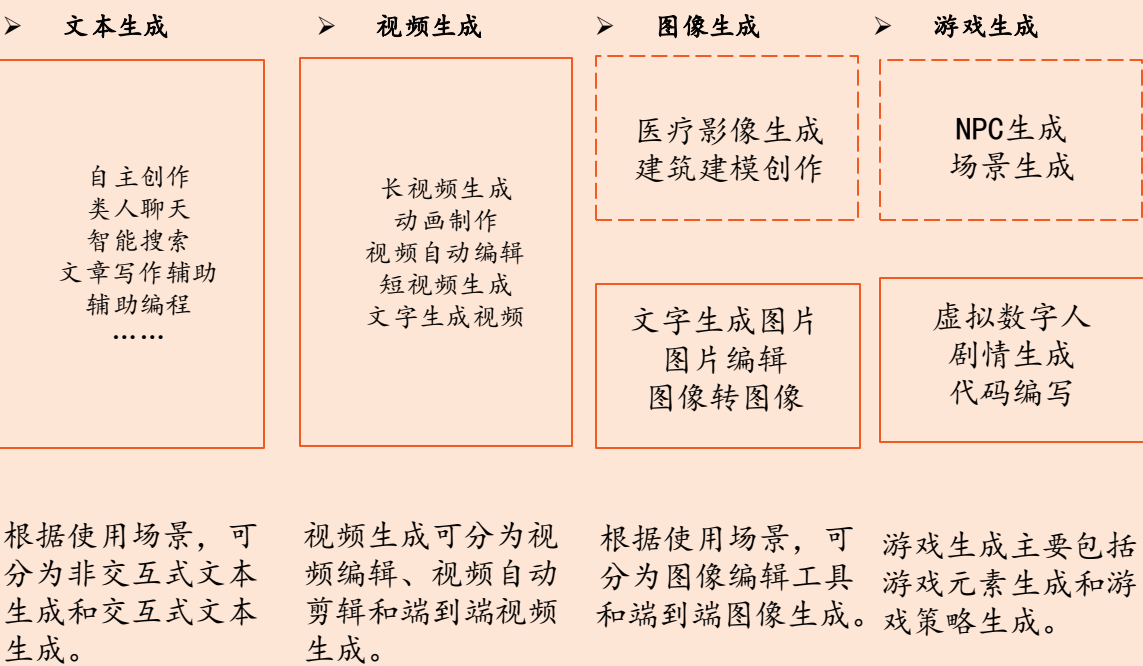


资料来源：Gartner（2022.7），平安证券研究所

# 文本、代码等应用有望近两年进入快速发展期，图像、视频生成成熟需时日

- ✓ 文本生成：应用相对成熟，在垃圾邮件的检测、基础的问答等领域都得到了应用，并在长篇写作等方面有所建树，快速发展阶段预计很快到来；后续预期在专业论文或者文学创作上实现落地；
- ✓ 图像生成：已经开始应用于产品设计、建筑模型，而且在医疗领域也实现应用，但达到设计师等专业人士的水平，还有差距；
- ✓ 视频、游戏生成：目前还处在早期尝试阶段，如实现电子游戏和电影的个性化创作，可能需要等待较长时间。

## AIGC的应用场景



## AIGC在不同场景中的发展预期

	2020年之前	2020年	2022年	预计2023年	预计2025年	预计2030年
文本	垃圾邮件的检测和翻译 基础问答回复	基础文案写作 生成初稿	长篇写作 完善初稿	对论文等专业文本内容垂直微调	终稿水平优于人类平均水平	终稿水平优于专业作家
代码	单行代码自动补齐	多行代码生成	更长篇幅 更高的准确度	更多种语言 更符合垂直行业的需求	文本到产品（终稿）	文本到产品终稿的效果优于全职开发人员
图像			美术作品 商标 摄影作品	产品设计、建筑等实物模型	产品设计、建筑等终稿	终稿优于专业艺术家、设计师、摄影师
视频/3D/游戏			3D、视频模型的初次尝试	视频和3D文件的初稿	完善初稿	AI Roblox实现电子游戏和电影个性化

大模型的实用情况：● 初次尝试    ● 基本完成    ● 准备就绪

# 国内AIGC垂直应用起步较晚，文本、图像等领域都还是蓝海赛道

- ✓ 相较美国市场，国内AIGC在相对成熟的赛道上参与者也不是很多，这和之前的预测性AI趋于红海的市场有着很大的差异。一方面，国内在算法和应用开发上存在差距，同时公有云SaaS发展不足带来的训练数据质量和应用通用性较差，限制了国内应用的大面积铺开；另一方面，国内B端为应用付费的意愿，尤其是AIGC这类新应用付费的意愿并不够强，反而是国内一些To C的赛道发展的更好一点。
- ✓ 从趋势上看，AIGC最终的市场依然是B端，但是考虑到中美市场的差异，国内企业依靠单一通用应用去打所有市场的可能性比较小，需要通过垂直化、行业定制化的开发AIGC方案，早期可能只面向1个或者几个行业赛道，有积累之后持续拓展其他行业。

■ 美国AIGC主要赛道参与企业（不完全统计）

文本生成		图像生成		视频生成		游戏生成	
业务类别	公司	业务类别	公司	业务类别	公司	业务类别	公司
文案编辑/ 市场营销	Jasper	图像生成	openAI	视频创作/ 剪辑	runway	NPC/化身	REPLICA
	Sudo write		Jasper		wombo		rct.ai
	Writesonic		OpenArt		descript		circle labs
	wordtune		KREA		ROSEBUD.AI		alethea
	Rytr		artbreeder		synthesia		coqui
	Regie.ai		play ground	数字人类	Rephrase.ai	游戏平台	latitude
	Anyword		Aragon		Colossyan		REGRESSION
	LAIKA		Picsart		Movio	游戏资产	mirage
	LEX		Stability.ai		elai.		KAEDIM
	Moombeam		PhotoRoom		Kinetix	游戏情节 创作	hidden door
聊天机器人	Convai	设计	VIZCOM	头像/动作 捕捉	DEEPMOTION		AI DUNGEN
	openAI		Diagram		plask	背景音乐	Soundful
	Google		alpaca	个性化视频	tavus		AIVA
	Inworld	自拍/头像	AVATAR		Rephrase.ai		Mubert
	Quickchat		PRISMA LABS				
客户服务	Outplay	3D图像	ASIS				
	Cohere		KAEDIM				
知识整合	Rewind	图像数据库	mirage				
	metaphor		Lexica				
搜索引擎	Andi	时尚	Nyx.gallery				
法律服务	Patentpal		CALA				

■ 国内AIGC主要赛道参与企业

垂直赛道初创公司

综合赛道初创公司

其他机构

文本生成

图像生成

视频生成

游戏生成

澜舟科技 langboat

聆心智能

彩云小梦

Giiso 写作机器人

META SOTA 秘塔写作猫

SHANNON.AI 火龙果写作

字语智能 用AI创造更有价值的信息

MPS智能写作

ZMO.AI

诗云科技

图宇宙

SIX PEN Art

拟仁智能 HAIHUMAN TECHNOLOGY

跳悦智能 JUMPY.AI

灵感世界 inspir.ai 超参数

小冰

阿里妈妈·创意中心

倒映有声 DAO VING

Baidu 百度

Alibaba 阿里巴巴集团

Tencent AI Lab

网易伏羲

meitu

腾讯优图

3A AI 北京智源人工智能研究院

BYTEDANCE AI LAB

Microsoft 微软亚洲研究院

# 文本生成：通过营销文案、摘要、故事生成引擎，赋能文创、金融等行业

- ✓ 功能：输入结构化的数据、图片或者文本来生成一段新的文本。例如输入结构化的数据、输入一张图片，或者输入若干关键词来生成文本。目前主要功能是生成可控文本，即文本中带有期望得到的属性，比如情感、关键词、事实以及主题等。
- ✓ 应用：营销文案引擎、文章摘要、故事生成、散文小说、文本复述、研报生成。
- ✓ 挑战：对篇章结构、多样性的可控；长文本生成面临的前后逻辑不一致、主题不一致以及事实不一致等问题。

国内文本生成典型解决方案



国内文本生成主要应用场景及功能

场景	功能	描述	应用实例
文化创作、快消	基于关键词生成	输入关键词句，将围绕关键词生成具有指定风格特色的原创内容	中文在线：利用文本生成平台进行辅助创作
	自定义模板生成	用户输入文本并标注出待补全部分，模型基于用户输入进行内容补全	
	续写	根据用户选择的题材和输入的前序文本进行续写	
	营销文本生成	输入品牌、核心卖点、应用场景等关键词，生成营销文案	美妆、汽车等行业已经开始有应用
金融行业	研报观点分析	自动抽取研报中的观点内容，并输出观点标签及观点极性。	华夏基金：舆情监测。重点解决传统工具关键字判断准确度低、情绪判断和业务需求不一致、大量重复信息干扰、无法识别舆情对象等问题。
	金融文本情感分析	对金融、财经领域中文本的情感倾向进行分析判断，有效赋能金融财经领域信息分析以及处理。	
	事件检测	从新闻中检测金融相关事件，用于分析和决策。	



# 图片/视频生成：能力提升明显，C端认知需提升但B端潜力较大

- ✓功能：通过AI技术来自动或辅助生成内容的生产方式，通过输入指令，让AI去完成冗杂的代码、绘图、建模等任务，实现“图生图”或者“文生图”的模式，生成新的图片/视频。在开源公司StabilityAI发布了Vincent的图形模型StableDiffusion之后，AIGC图片生成能力显著提高。
- ✓商业模式：C端客户目前需求更为复杂，AIGC在技术上还存在差距，目前很难有大的商业机会，业内企业6Pen调研数据显示，60%用户不曾付费，付费超过100元的比重也只有10%，市场拓展还需要时日；B端客户变现的可能性会更高，垂直客户需求相对聚焦，变现路径也更为多元和成熟，比如广告、营销、文旅等行业，预计都会有较大的空间。

国内图片/视频生成典型应用场景

行业	应用方式	应用实例
广告行业	利用AI制作宣传材料	百度利用文心平台制作电视剧宣传海报，腾讯、美图也在进入该赛道
设计行业	建筑、包装、服饰等设计出图	ZMO.AI通过生成高质量的模特儿照片图片，降低拍照成本和等待时间
营销定制	生成图片或者视频材料	百家号、剪映开放图文转视频功能，可以进行个性化的创作
漫画/动漫行业	配文生成漫画、插画和视频	国内艺术家基于小说《2086元宇宙》制作漫画版，形成可视版本
游戏行业	辅助角色、道具、场景设计	网易推出首个游戏版ChatGPT将会在《逆水寒》手游

利用AIGC生成的绘画、海报和手游





# 类人聊天机器人：AIGC应用集大成者，已从形似开始走向神似

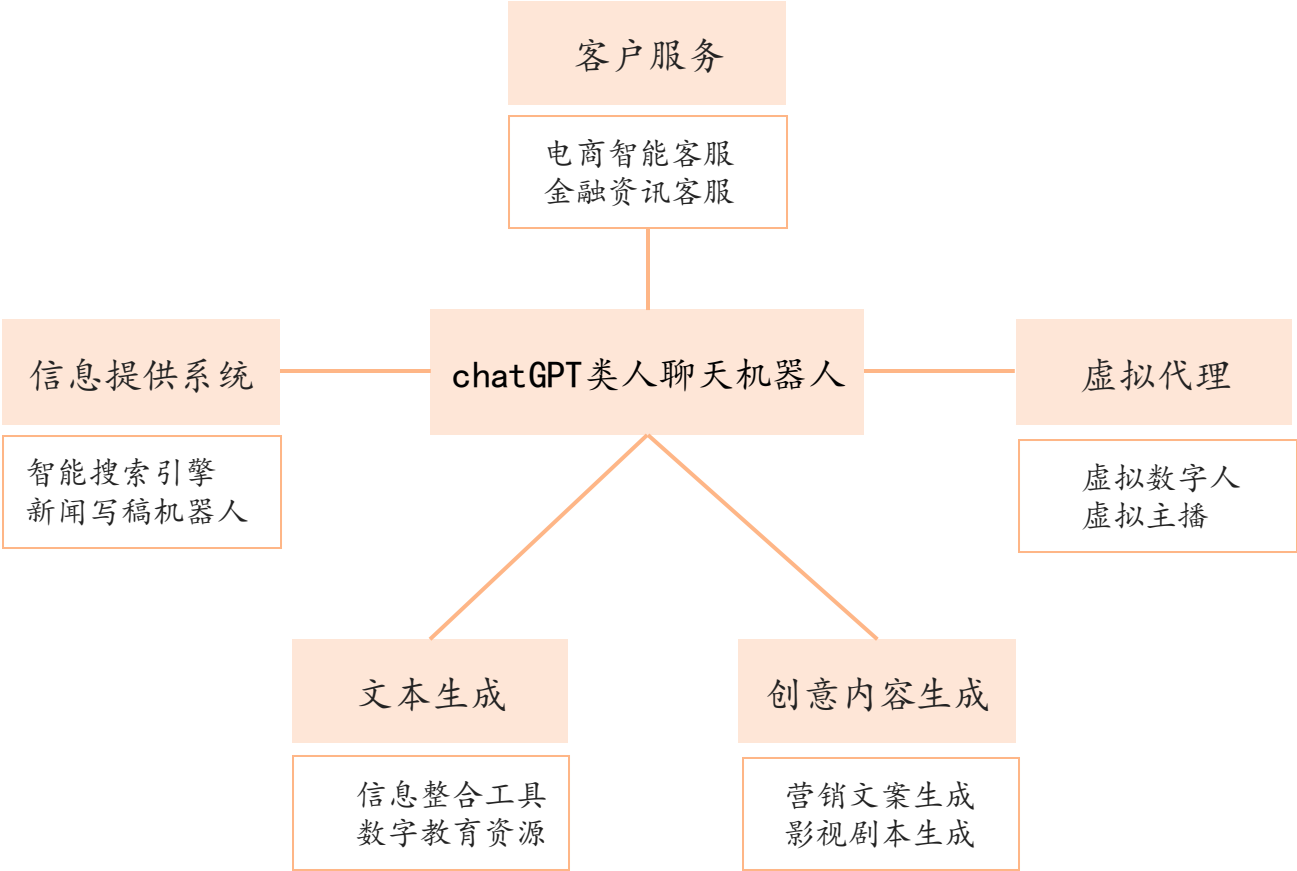
## 类人聊天机器人及头部厂商

类人聊天机器人是一种旨在模拟与用户进行类人对话的电脑程序。它们已被应用于多个行业，包括客户服务、电子商务和娱乐等。

chatGPT带来的变革：此前的聊天机器人只能根据设定问题回复固定内容，同chatGPT融合之后，将文本生成、图像生成、多模态转换技术融合起来，环境感知、理解能力在大幅增强，实现和人的形似到神似。基于chatGPT的类人聊天机器人可以提供客户服务、虚拟代理、内容创意和文本生成、以及信息知识提供等服务。

- 国外重点企业：
  - openAI：公司核心宗旨在于“实现安全的通用人工智能(AGI)”，目前推出的“chatGPT类人聊天机器人”火爆全球，其公司在图片生成和视频生成的应用上也有很深的造诣。
  - Google：公司在AIGC领域布局很久，目前也有“bard”类人聊天机器人的诞生试用，在AIGC的多种技术领域也处于前列。
- 国内重点企业：
  - 百度：是国内布局AI赛道最早的公司之一，目前已宣布“文心一言”类人聊天机器人将很快进入公测环节，图片生成方面也有突破。
  - 小冰：从微软独立拆分出来的人工智能公司，推出的“小冰”机器人是目前全球范围内承载交互量最大的人工智能系统。

## 类人聊天机器人的多种应用场景—以chatGPT为例

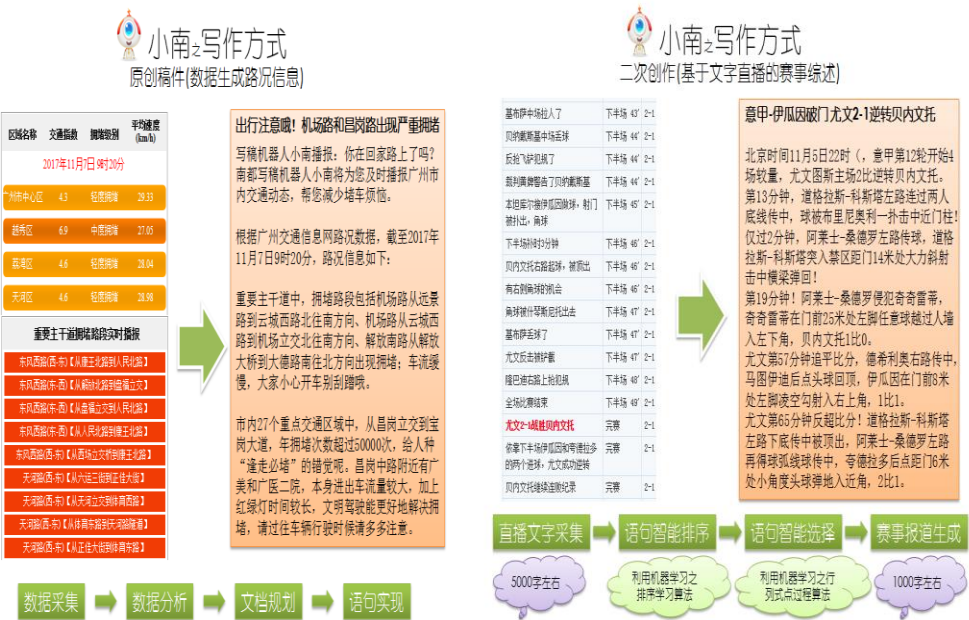


# 类人聊天机器人|新闻写稿：应用起步较早，大幅提升写作效率

- ✓ 利用AIGC技术参与新闻写稿辅助工作，各家新闻机构起步均较早，2015年以来国外包括纽约时报、彭博都已经开始应用，国内随后也在做相关投入，比如新华社的“快笔小新”、第一财经与阿里巴巴合作开发了“DT稿王”、南方都市报与北京大学团队合作研发的写稿机器人“小南”等。
- AI机器人可以帮助精确检索二手信息，收集素材。在高性能的AIGC工具如ChatGPT出现后，就可以像常人对话一样直接提问并获得答案。另外AIGC转写工具可以帮助记者实时生成文稿，自动撰写提纲、精简语句等，进而提高工作效率，保证最终产出的时效性。
- 目前AIGC撰稿工具能在一分钟内生成两千条新闻，且单条质量可比拟人类半小时的作品质量。尤其是针对股市、重大体育赛事等新闻，会在关键时间点自动撰写新闻，编辑只需要做最终审校就可以发出，大幅提升了新闻的时效性，降低了编辑的工作量。
- ✓ 目前AI机器人主要在财经、体育和生活类等新闻中出场，深度、调查报告等还需要人类记者完成，后续随着技术的进步，应用面将拓宽。

## 国内代表性的机器人记者方案

## 新闻写稿机器人重点功能应用

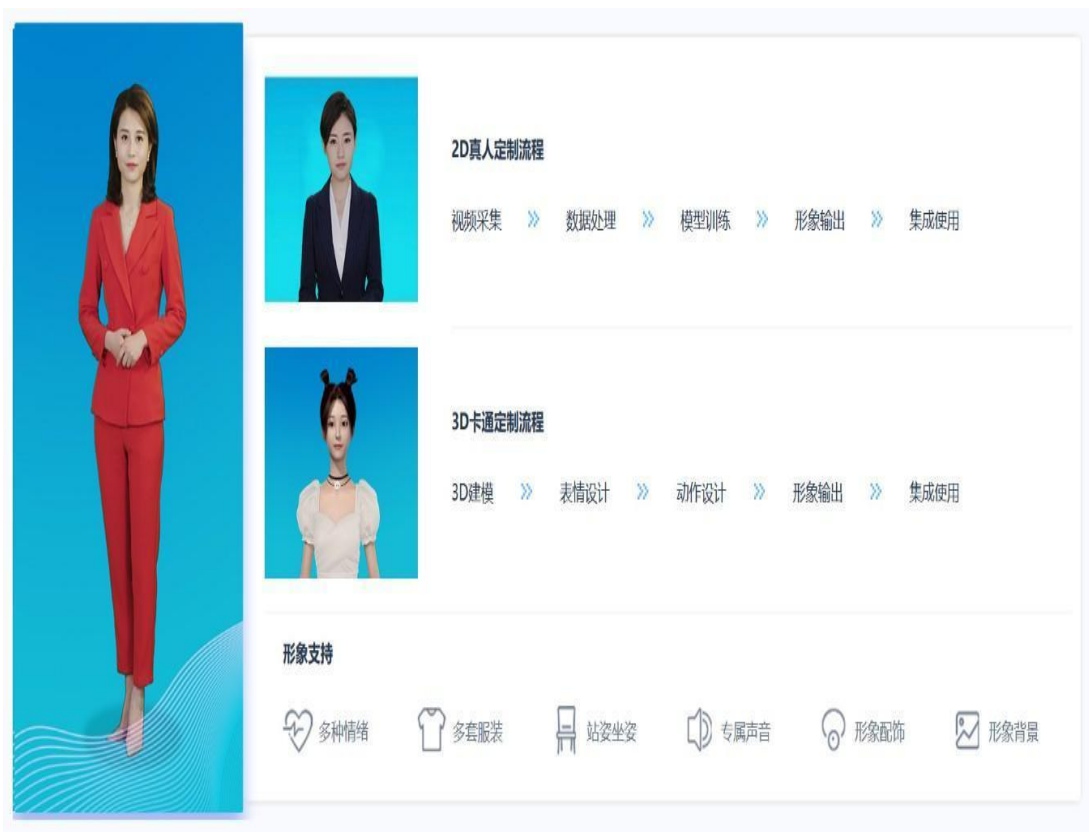


功能	具体应用
文本复述/改写	对原有文本进行改写，观点一致，但写作风格和表达多样
语音交互	通过口述查询，自动生成稿件，通过ChatGPT技术可以实现智能问答
看图写话	图片、视频的自动生成。根据给定的图片或者视频，自动生成可以描述图片或者视频内容的稿件，类似于“看图写话”或是“看视频写话”。
视频写稿	根据现有文字提取摘要和关键词，快速生成对应的短视频稿件。对编辑填写的文本信息进行语义分析，得到摘要与关键词，并自动搜索资源库匹配关联性最强的图片、视频、音频素材，自动制作成一段视频
智能模板生成	通过持续训练，根据不同的内容要求，自动生成模板

# 类人聊天机器人|虚拟数字人：形似到神似，活跃于客服、新闻制播等多个场景

- ✓ 虚拟数字人是AIGC的重要应用之一，利用人像、声音生成等技术，可以定制2D、3D的虚拟数字人，可以实现文本到视频的输出生。目前，AIGC在该领域应用更为成熟，智能化水平更高，输出结果更为灵活和应景，交流更为顺畅，实现了从形似到神似的跨越。
- ✓ 目前市场上的虚拟数字人可以充当虚拟客服、数字员工、虚拟老师、虚拟主播、数字导游等角色，应用于金融、传媒、旅游、电子政务等多个场景，提供客户服务和交互、课程制作、新闻制播、政务服务和查询等多方面的功能，大幅提升了内容制作输出效率，保证了用户的服务体验。目前，科大讯飞等人工智能企业，均提供了虚拟数字人的解决方案。

## ■ 虚拟数字人制作流程



## ■ 虚拟数字人当前面向的场景





# 落地案例 | 游戏：AIGC与游戏赛道天然亲和，数据资产生产等应用将加速

- ✓ 游戏是典型的资产复杂度非常高的娱乐形式，人力密集，研发人员持续膨胀，质量、成本、效率不可能三角问题难解。AIGC与游戏具有天然的适配性，在图片/视频生成、虚拟玩家、音乐制作、剧情/策略开发等方面都有非常大的潜力，而且娱乐场景来自监管、伦理、安全性等方面的压力较小，落地较快。
- ✓ 目前，AIGC正在为游戏中的数据资产建立模型，开发智能对话、智能陪玩、智能QA，并能够通过剧情和策略生成等方式，丰富和延展游戏内容，延展游戏寿命。通过AI生成，可以降低生产成本，提高效率，还能保证游戏质量不下降，完美解决不可能三角问题。

## ■ AIGC在游戏资产生成过程中的应用



## ■ AIGC在国内外落地案例

厂商	产品	应用情况
腾讯	“绝悟”陪玩机器人	通过强化学习的方法来模仿真实玩家，包括发育、运营、协作等指标类别，以及每分钟手速、技能释放频率、命中率、击杀数等具体参数，让AI更接近正式服玩家真实表现，将测试的总体准确性提升到95%。
rct A	智能NPC生成、剧情延展、智能Q&A、智能陪玩等	智能NPC为例，其NPC能够分析玩家的实时输入，并动态地生成交互反应，从而构建几乎无限目不重复的剧情，增强自户体验并延长游戏生命周期。
网易游戏	智能NPC、玩家智能对话	武侠手游《逆水寒》宣布实装国内首个游戏版ChatGPT，让智能NPC能和玩家自由生成对话，并基于对话内容，自主给出有逻辑的行为反馈。该手游已经获得版号，预计将于2023年年内上线。
行者AI	游戏资产生成	为游戏生成2D图片、3D 环境、单体建筑、人物角色模型、虚拟数字人等复杂内容生成。

# 落地案例 | 银行：通过数字员工等形式试水AIGC，监管、安全性问题需关注

- ✓ 银行是作为服务业的重点领域，数据丰富且质量优秀，也是与AI融合的比较快的领域。此前，AI营销文案生成、数字员工（客服）等解决方案已经得到应用，承担问题应答、业务操作指引、查询响应和确权等职能。
- ✓ 随着ChatGPT的火爆，各家银行正在探索AIGC进一步为虚拟数字员工赋能，提升智能化水平，在内外部沟通等环节中发挥更重要的作用。在银行线下网点，数字员工就充当“服务型数字人”的角色，提升大堂的服务效率和服务质量；线上业务可能将是各大银行未来利用数字员工争夺的重点，会强调“数字人”IP的打造，树立起银行在虚拟世界的品牌形象，实现客户的转化，很多银行已经将数字员工应用到线上的营销推广中。
- ✓ 后续，AIGC处理复杂问题和创造性问题的能力也将持续提高，银行业应用面临的监管、安全性等问题也将逐步得到解决，AIGC在银行领域的应用范围将扩大。除了当前正在应用的客户服务、营销运营和数字员工之外，未来在风险管控、经营态势分析等领域也会有较大的应用潜力。

■ 主要银行机构AIGC应用情况

■ 银行机构数字员工典型应用

机构	AIGC技术应用进展
宁波银行	开发了数字员工小宁，用于智慧大堂服务和线上营销等场景。
工商银行	开发了数字员工工小智，提供线下支行网点大堂服务，进行产品介绍和营销。
邮储银行	将开始内测文心一言，在智能客服、数字员工、虚拟营业厅等场景进行应用，进一步提升客户体验。
江苏银行	尝试运用ChatGPT技术提升软件开发生产力。
招商银行	在智能客服、流程智能化等方面，广泛应用了大模型技术用于意图识别、信息提取等具体工作。
百信银行	把百度文心一言的智能对话技术成果应用在数字金融、AI数字人、数字营业厅等领域。
中信银行	将接入文心一言生态，在财富管理等场景尝试应用。
兴业银行	接入文心一言，用于智慧网点、智能服务、智能风控、智能运营、智能营销、智能投研等。

## 对外

对外场景包括存、贷、汇、监等一系列金融业务服务：

- 网点客户服务、交流
- 结合用户特征定制专属方案，成为客户的“专属顾问”，并在交互过程中识别用户情感，通过话术激发用户兴趣、引导用户选择产品
- 在用户不感兴趣时进行挽留、适时推荐其他产品，充分给予用户优质体验。



对内场景，数字员工可以充当智慧助手：

- 解答员工问题，及时提醒待办日程，快速办理日常事务，提升员工幸福感
- 为员工制定专属的岗位培训计划，提供专业的职业发展建议，加速员工成长
- HR的“分析助手”，对员工数据进行整合挖掘，形成员工画像和关系图谱，协助HR评估员工能力、调优人员配置

## 对内

资料来源：《金融电子化》、工商银行等、平安证券研究所



# 目录

- 1 焦点：ChatGPT带火AIGC，Open AI引领技术和生态
- 2 变革：AIGC与人更为神似，模型和数据是主要助力
- 3 市场：大模型需要大算力，推动AI服务器市场增长
- 4 应用：行业将逐步回归理性，能否突破需要看B端
- 5 投资建议及风险提示

## 投资建议及风险提示

---

- ✓ **投资建议：**近年来AIGC的算法、数据质量等要素有着显著提升，文本、图片/视频、代码等技术能力快速发展，在客服、营销、游戏等赛道都显示出非常强的应用潜力。伴随着chatGPT的催化，国内AIGC产业链也将迎来空前的市场机会。算法方面，预计今年国内龙头AI厂商有希望推出能力接近chatGPT的产品；算力是短板，由于供应链的问题，我国高端训练芯片供给可能面对较大的压力，AI算力不足的问题可能显现，相关设备研发和基础设施建设企业可能继续受益。同时，AIGC新技术也会带来新的安全挑战，国产安全厂商也将受益。强烈推荐启明星辰，推荐科大讯飞、金山办公、中科曙光、浪潮信息、紫光股份等；关注AI厂商后续发展机会，如云从科技、海天瑞声、云天励飞（待上市，已同意注册）。
- ✓ **风险提示：**1) 技术落地可能不及预期。目前chatGPT处于市场炒作高点，未来将面临着市场落地的考验，由于技术本身尚需持续完善，能否跨越商业化的“死亡之谷”仍待观察。2) 来自监管和版权方面的风险。AIGC生成内容还主要依赖现有素材，可能产生版权纠纷，同时生成内容版权的认定还存在争议，商业化可能面临障碍；在一些特定场景，AIGC可能受到监管的限制；3) 国内AI供应链断裂的风险。可能出现芯片、软件和算法等供给限制更为严格的情况，国内研发和商业化可能受到冲击。

# 重点推荐标的

股票简称	股票代码	2月27日	EPS（元）				PE（倍）				评级
		收盘价（元）	2021A	2022E	2023E	2024E	2021A	2022E	2023E	2024E	
启明星辰	002439.SZ	28.68	0.90	0.92	1.23	1.57	31.9	31.2	23.3	18.3	强烈推荐
科大讯飞	002230.SZ	47.79	0.67	0.25	0.71	0.95	71.3	191.2	67.3	50.3	推荐
金山办公	688111.SH	288.71	2.26	2.42	3.42	4.73	127.7	119.3	84.4	61.0	推荐
浪潮信息	000977.SZ	37.24	1.37	1.63	2.00	2.47	27.2	22.8	18.6	15.1	推荐
紫光股份	000938.SZ	24.48	0.75	0.86	1.07	1.33	32.6	28.5	22.9	18.4	推荐
中科曙光	603019.SH	31.41	0.79	1.02	1.31	1.64	39.8	30.8	24.0	19.2	推荐

# 重点标的 | 科大讯飞——智能语音领军企业

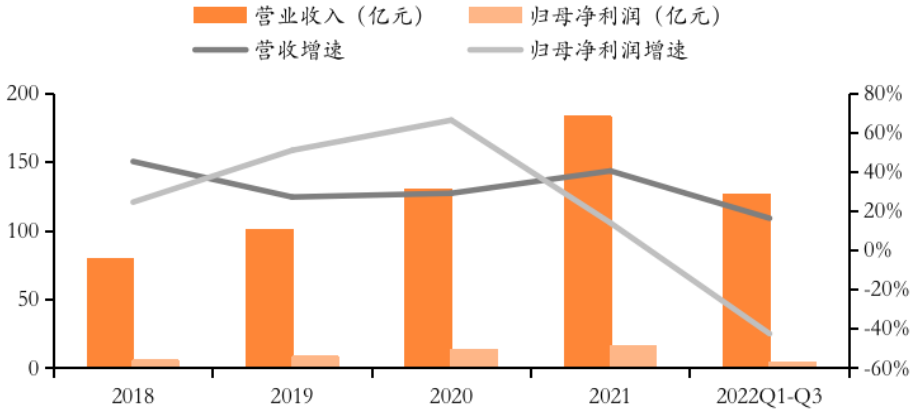
## 公司简介及AI赛道相关布局

- ✓ 科大讯飞成立于1999年，是亚太地区知名的智能语音和人工智能企业。公司作为技术创新型企业，其智能语音、自然语言理解、计算机视觉等核心技术处于国际领先水平，已入驻“中国人工智能国家队”。
- ✓ 公司以“平台+赛道”战略为指引，2B+2C双轮驱动产品落地。公司构建了以智能语音和人机交互为核心的“讯飞开放平台”，并在此基础上推动产品及服务在消费者、智慧教育、智慧城市、智慧司法、智能汽车、智慧医疗、运营商等领域的深度应用。截至2022年5月31日，讯飞开放平台已开放493项AI产品及能力，聚集超过337.3万开发者团队，总应用数超过150.1万，累计覆盖终端设备数35.1亿+。

## 公司财务状况

- ✓ 2022年前三季度，公司实现营业收入126.61亿元（+16.50%）；实现归母净利润4.20亿元（-42.34%）；EPS为0.18元。预计2022年实现营业收入183.14-201.45亿元，同比增长0%-10%；实现归母净利润4.67-6.23亿元，同比下降60%-70%；EPS为0.20-0.27元。“推荐”评级。

## 公司营收及归母净利润情况



## 公司AI产品全栈布局



资料来源：公司官网，WIND，平安证券研究所

# 重点标的 | 金山办公——办公领域AI先行者

## 公司简介及AI赛道相关布局

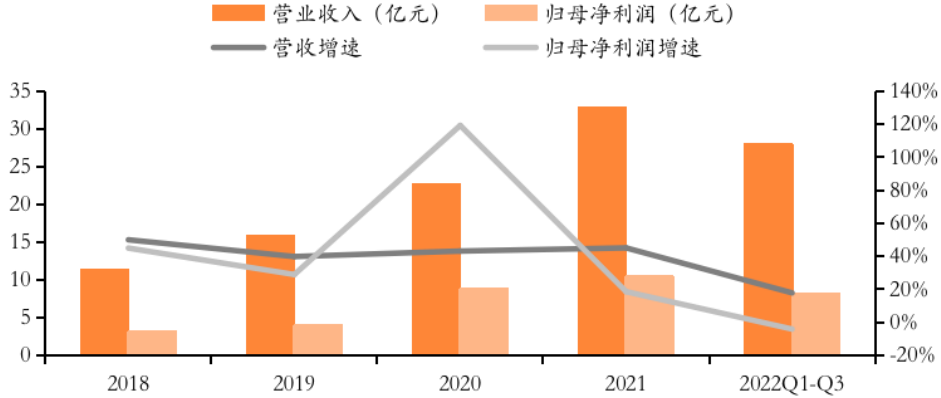
- ✓ 金山办公成立于2011年，是国内领先的办公软件和服务提供商。公司主要从事WPS Office办公软件产品及服务的设计研发及销售推广。公司自2017年开启AI战略部署至今，已经在OCR、文档翻译、图文识别检测等多领域实现产品落地。
- ✓ 公司基于AI中台，打造智能化办公场景。公司AI中台面向计算机视觉、自然语言处理、语音处理等算法研究方向，围绕办公领域，已开发出了近100项AI能力。公司推出的WPS智能写作可以实现文本自动生成、辅助成稿写作、句子智能补写、文本智能校对等功能。
- ✓ 公司开源全球首个办公领域DL框架——KSAI-lite。公司于2021年7月开源业界首个面向办公领域的深度学习框架KSAI-lite，支持离线AI计算，并且适配国内外主流软硬件平台，在OCR、机器翻译、智能校对等场景具有显著优势。

## 公司财务状况

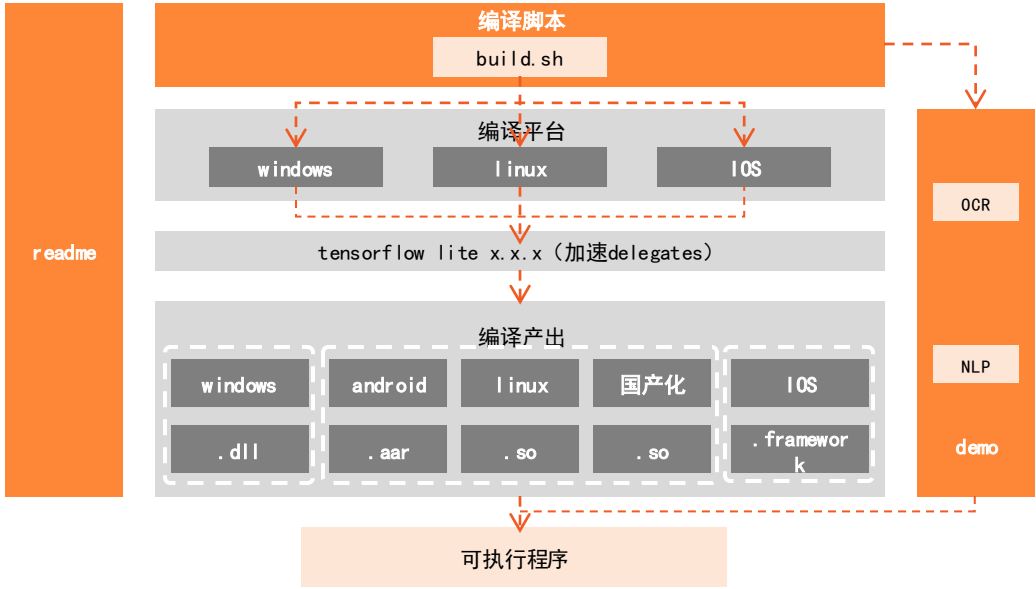
- ✓ 2022年前三季度，公司实现营业收入27.95亿元（+17.83%）；实现归母净利润8.13亿元（-4.10%）；EPS为1.76元。“推荐”评级。

资料来源：公司公众号，WIND，平安证券研究所

## ■ 公司营收及归母净利润情况



## ■ 公司办公领域深度学习框架KSAI-lite





# 重点标的 | 浪潮信息——全球AI服务器龙头

## 公司简介及AI赛道相关布局

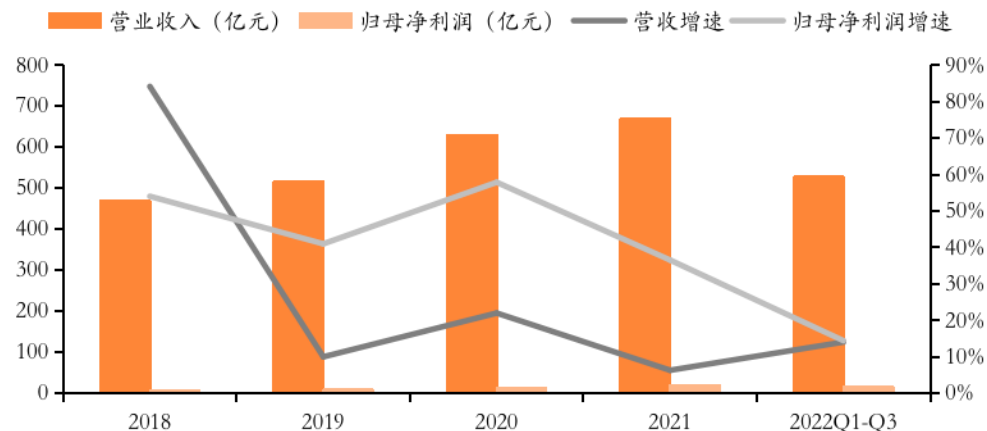
- ✓ 浪潮信息是全球人工智能计算领导企业，AI服务器市占率世界第一。公司目前在售AI服务器产品10余款，在研产品6款，分为AI训练服务器、AI推理服务器、元宇宙服务器等类型，主要应用场景包括自然语言处理、计算机视觉、大型推荐系统等。据IDC统计，2022年H1，公司AI服务器产品市场占有率达到50.3%。
- ✓ 公司立足算力、算法、数据，构建全栈AI能力。算力方面，公司于2022年发布国内首款元宇宙服务器MetaEngine，将为元宇宙数字空间的创建和运行提供强大算力。算法方面，公司于2021年发布“源1.0”大模型，公司已基于此研发出对话、问答、翻译和古文四个技能模型，同时，“源1.0”大模型落地智算中心，通过开源模型API、高质量数据集、模型代码、技能模型等形式对外提供算法基础服务。数据基础设施方面，公司于2022年发布了新一代SSD高速存储介质，具备ZNS数据分流等技术，性能大幅提升。

## 公司财务状况

- ✓ 2022年前三季度，公司实现营业收入527.67亿元（+14.10%）；实现归母净利润15.46亿元（+14.47%）；EPS为1.06元。“推荐”评级。

资料来源：公司公告，公司公众号，WIND，IDC《2022 H1中国加速计算服务器市场报告》，平安证券研究所

## ■ 公司营收及归母净利润情况



## ■ 浪潮人工智能服务器系列



# 重点标的 | 紫光股份——国内领先ICT基础设施供应商

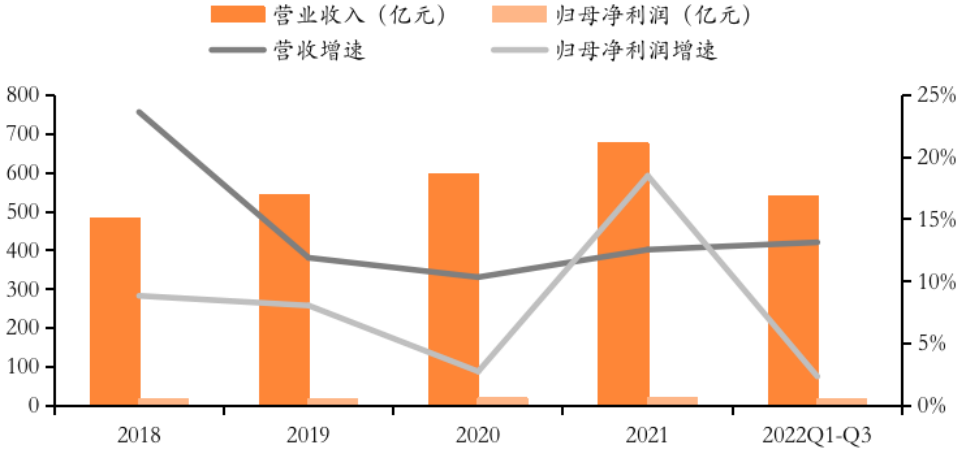
## 公司简介及AI赛道相关布局

- ✓ 紫光股份旗下的新华三集团是国内领先的ICT基础设施服务提供商。新华三提供一系列包括数字中心交换机、AI服务器等在内的高端AI算力基建产品，未来将基于“云智原生”战略，加速AI创新步伐。
- ✓ 新华三以软硬件全站能力赋能AI产业发展。在底层硬件算力支撑上，新华三服务器产品可覆盖从AI训练到推理的全场景。此前，新华三GPU服务器曾登顶MLPerf™国际权威AI基准性能竞赛。在2023年上半年，新华三还将发布基于Intel Eagle Stream平台的GPU服务器R5300 G6。此外，面对AI作业调度、集群管理等难题，新华三推出智能算力中枢，为人工智能提供全场景的算力调度，实现了多维智能的AI开发全流程支持。除复杂的大模型训练场景，新华三集团也面向广大AI开发者，免费开放H3C小智AI实验室，让用户体验真实的AI训练全流程。

## 公司财务状况

- ✓ 2022年前三季度，公司实现营业收入538.90亿元（+13.16%）；实现归母净利润16.40亿元（+2.38%）；EPS为0.57元。“推荐”评级。

## ■ 公司营收及归母净利润情况



## ■ 新华三智能数字平台



# 重点标的 | 中科曙光——智能计算领军企业

## 公司简介及AI赛道相关布局

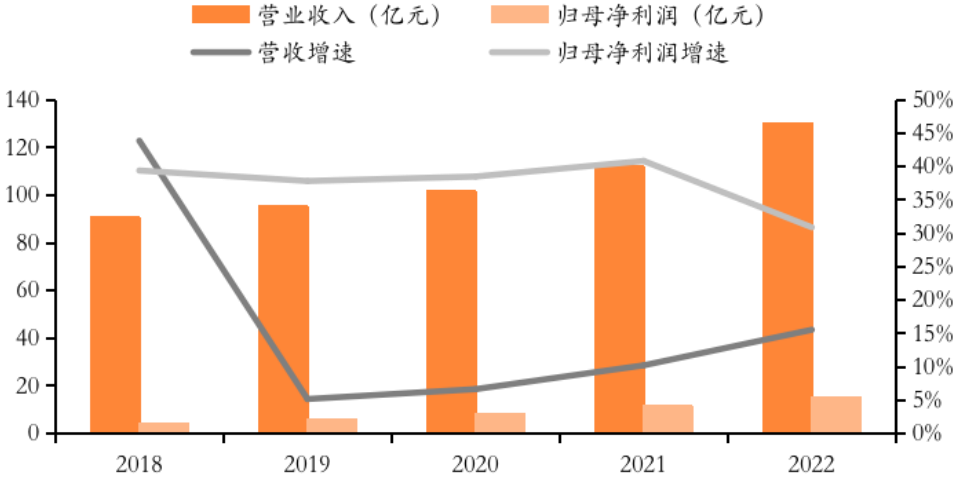
- ✓ 中科曙光是我国核心信息基础设施领军企业。公司在高端计算、存储、安全、数据中心等领域拥有深厚的技术积淀和领先的市场份额，并充分发挥高端计算优势，布局智能计算、云计算、大数据等领域的技术研发。
- ✓ 公司在AI算力领域具有先发优势。公司在2019年就开始布局AI算力服务，将遍布全国的海量异构计算资源链接，形成一体化算力服务网络。该算力服务网实现了对算力、存储、网络以及数据等分布资源的整合，并通过统一的算力服务门户，根据科学计算、工程计算和智能计算等场景需求，向用户输出算力、算法、数据、应用高度协同的一体化资源，降低了AI大模型的训练与推理成本，让用户无需投入大量成本购买硬件设施也可完成AI大模型训练等复杂计算工作。

## 公司财务状况

- ✓ 根据公司2022年度业绩快报数据，2022年公司实现营业收入130.09亿元（+15.56%）；实现归母净利润15.15亿元（+30.92%）；EPS为1.04元。“推荐”评级。

资料来源：公司官网，公司公众号，WIND，平安证券研究所

### ■ 公司营收及归母净利润情况



### ■ 公司产品矩阵



# 重点标的 | 启明星辰——人工智能安全领域实践者

## 公司简介及AI赛道相关布局

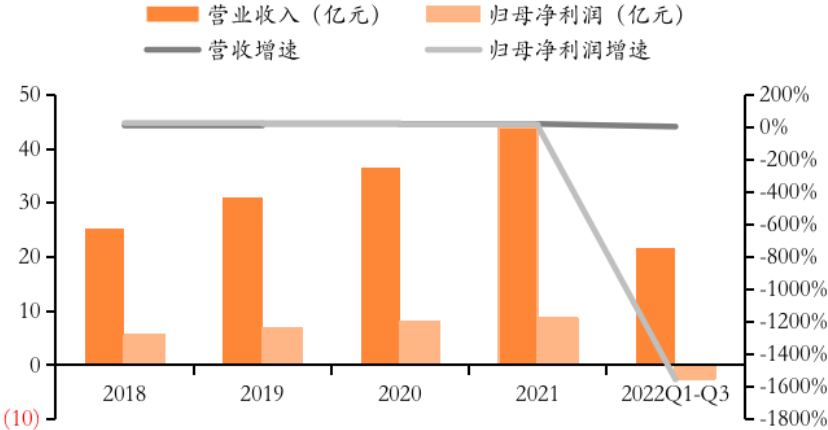
- ✓ 中科启明星辰是国内极具实力的、拥有完全自主知识产权的网络安全供应商，积极推进网络安全领域人工智能应用的发展和工程化落地。
- ✓ **PanguBot——安全运营的“ChatGPT”实战。**启明星辰基于多年对于自然语言理解、自然语言处理、大数据分析、智能化处理方面的智算能力研究基础上，叠加二十余年在安全领域的技术、数据、运营和研究，于2022年发布了“PanguBot(盘古小古)”安全智慧生命体。
- ✓ **盘古人工智能平台——搭建人工智能安全应用的技术基石。**启明星辰盘古人工智能平台（人工智能安全建模和赋能平台）作为安全运营的人工智能底座，为安全领域的人工智能应用提供了研发运营的一体化环境，能够实现基于ModelOps和AIOps的人工智能应用快速搭建、模型全生命周期管理和多重赋能。

## 公司财务状况

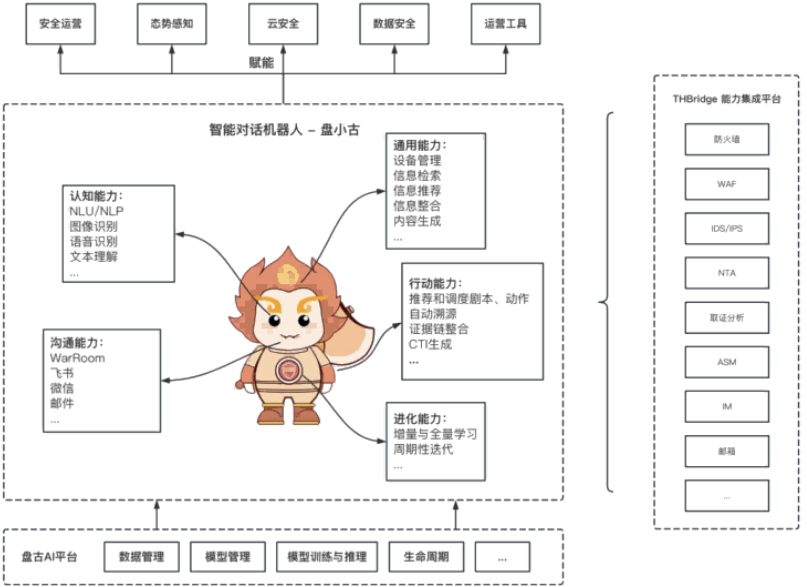
- ✓ 2022年前三季度，公司实现营业收入21.68亿元（+3.39%）；实现归母净利润-2.50亿元（-1554.73%）；EPS为-0.27元。“强烈推荐”评级。

资料来源：公司公告，WIND，平安证券研究所

■ 公司营收及归母净利润情况



■ PanguBot 盘古小古



附：团队主要成员

TMT团队				
行业	分析师/研究助理	邮箱	资格类型	资格编号
电子	徐勇	XUYONG318@pingan.com.cn	投资咨询资格	S1060519090004
半导体	付强	fuqiang021@pingan.com.cn	投资咨询资格	S1060520070001
	徐碧云	XUBIYUN372@pingan.com.cn	一般证券资格	S1060121070070
计算机	闫磊	YANLEI511@pingan.com.cn	投资咨询资格	S1060517070006



