信达证券 CINDA SECURITIES

Research and Development Center

AI 大模型的长期垄断形成与竞争要素

2023 年 4 月 13 日

证券研究报告

行业点评

行业专题研究（普通）

电子

投资评级 看好

上次评级 看好

莫文宇 电子行业首席分析师
执业编号：S1500522090001
联系电话：13437172818
邮 箱：mowenyu@cindasc.com

韩宇杰 联系人
邮 箱：hanzijie@cindasc.com

信达证券股份有限公司
CINDASECURITIESCO.,LTD
北京市西城区闹市口大街9号院1号楼
邮编：100031

AI 大模型的长期垄断形成与竞争要素

2023 年 4 月 13 日

本期内容提要：

- **GPT 模型基于 Transformer，它的本质即全局特征提取器。**将词向量、位置向量和分段向量相加，便得到了 GPT 模型的输入表示。在模型的训练过程中，这些向量将通过多层 Transformer 结构进行处理，以捕捉词汇之间的复杂关系。词向量（Token Embeddings）：每个词片段都被映射到一个固定长度的向量，捕捉该词片段的语义信息。这些词向量在模型的预训练过程中学习得到。位置向量（Positional Embeddings）：GPT 使用固定长度的位置向量，用于捕捉词片段在输入序列中的位置信息。这些位置向量与词向量相加，生成包含位置信息的输入表示。分段向量（Segment Embeddings）：GPT-2 不使用分段向量，但在 GPT-3 及 BERT 等其他模型中，它们用于区分不同的输入段。模型的训练就是寻找这些向量之间存在的位置关系，以发现语言作为知识的载体，其本身所蕴含何种数学相关性。

- **提升参数量=提升性能、提升泛化能力，长期垄断局面可能形成：**从论文研究来看，参数量的提升有助于构建语言预测模型的精确度，同时提高泛化能力。泛化能力的提升意味着一个参数量超级庞大的大模型，其在垂直细分领域的预测能力可超过针对垂直领域开发的中等参数量模型，这意味着 AI 的发展长期也是强者恒强的垄断过程，即参数量超级庞大的模型在任何垂直领域都具备优势，垂直细分的小模型难有生产空间。

- **数据标注的地位被弱化，AI 产业的经济竞争也是文化竞争：**无论是 GPT 还是 SAM，其在训练过程中，大量依靠互联网原生内容训练，因此一种语言的高质量文本内容的丰富程度，将决定基于该语言的大模型能力强弱，中文互联网文本内容生态亟待加强。

- **AI 大模型至少是一次中等规模的产业革命：**仅从时间节点 ChatGPT 的表现来看，AI 的能力边界取决于过去人类产生的知识，它对于工业来说就是极大地降低了知识获取难度。将人类知识的海洋汇聚于一个语言的入口，它更像是 windows 之于电脑，开启了普通人接触高性能设备的通道，也开启了电子产品/AI 从企业端进入消费端的大门。

- 针对电子行业，我们认为在大模型格局未完全形成之前，参数量的无上限堆砌是各家企业发力的焦点，故算力的“军备竞赛”无可避免，若以动态视角来看 AI 芯片及服务器相关上游的弹性存在超预期可能。建议关注：兴森科技、兆威机电、芯原股份、寒武纪、通富微电。

- **风险提示：**1.技术迭代不及预期；2.地缘政治风险；3.技术路径、产业趋势发生重大变化。

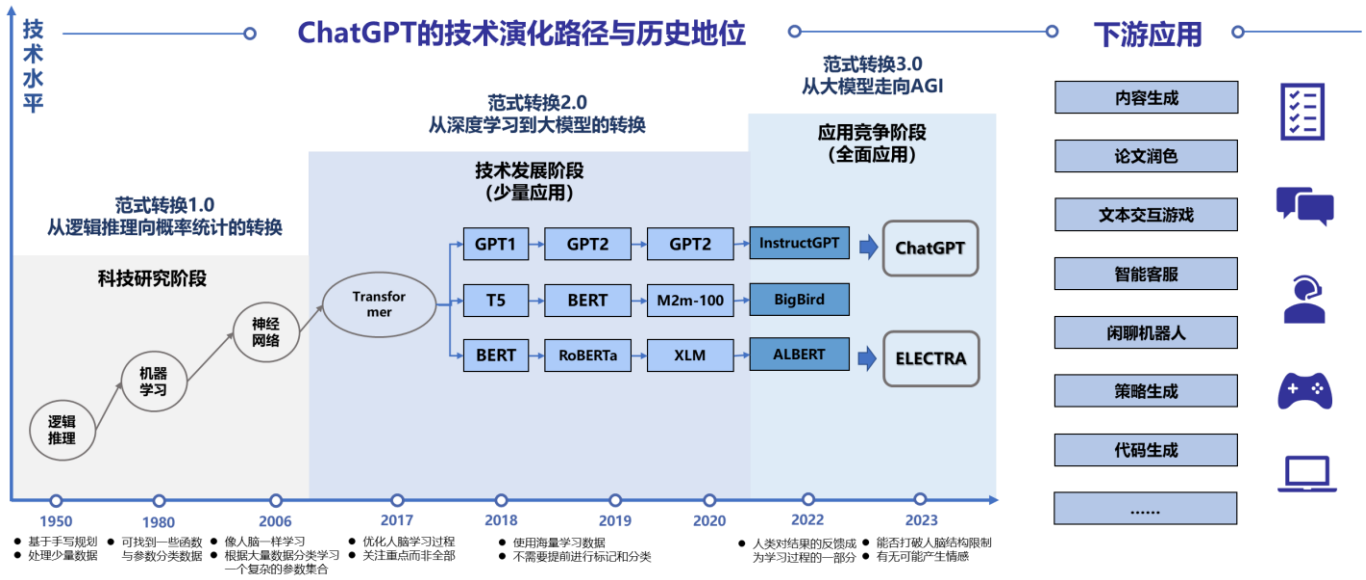
目录

图目录

回归 AI 基本面：技术视角的分析

人工智能经历了漫长的研究过程，近年来在范式上的转变奠定了 ChatGPT 的基础。基于通用类模型构建 AI 系统的模型被称为基础模型，即在大规模数据上训练、微调并适配到各种下游任务的模型。基础模型基于深度神经网络和自监督学习，已经存在了几十年。随着 Transformer 的诞生，基础模型的规模迅速壮大，应用范围突飞猛进。大模型在此基础上通过“暴力美学”实现大算力、大数据、大参数下的通用模型能力，可根据具体垂直应用进行微调。ChatGPT 是在 GPT-3.5 系列模型的基础上形成的，于 2022 年初完成训练，**其出现代表 AI 技术的第三次范式升级，即从大模型走向 AGI（通用人工智能）。**

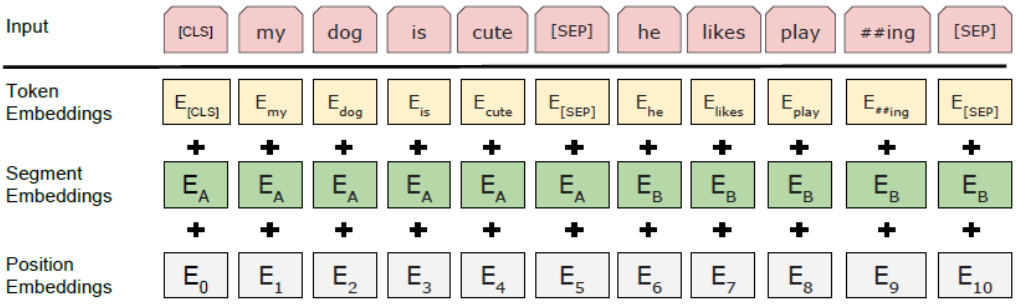
图 1：ChatGPT 的技术演进路径、历史地位与下游应用



资料来源：甲子光年，真格基金，信达证券研发中心

GPT 模型基于 Transformer，它的本质即全局特征提取器。将词向量、位置向量和分段向量相加，便得到了 GPT 模型的输入表示。在模型的训练过程中，这些向量将通过多层 Transformer 结构进行处理，以捕捉词汇之间的复杂关系。词向量（Token Embeddings）：每个词片段都被映射到一个固定长度的向量，捕捉该词片段的语义信息，这些词向量在模型的预训练过程中学习得到。位置向量（Positional Embeddings）：GPT 使用固定长度的位置向量，用于捕捉词片段在输入序列中的位置信息。这些位置向量与词向量相加，生成包含位置信息的输入表示。分段向量（Segment Embeddings）：GPT-2 不使用分段向量，但在 GPT-3 及 BERT 等其他模型中，它们用于区分不同的输入段，分段向量在这些模型中有助于捕获多个句子之间的关系。

图 2：Transformer 模型如何理解语义



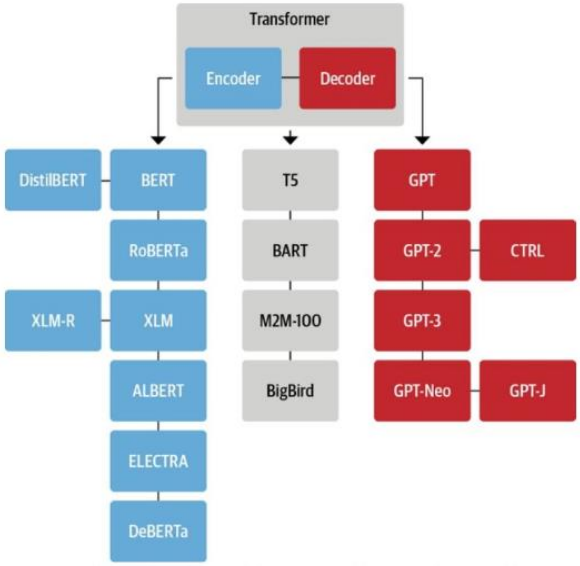
资料来源：BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding（Jacob Devlin 等），信达证券研发中心

Transformer 模型是参数量可以无限增长的通用模型，可以处理长序列的输入、输出，泛化能力强。Transformer 模型是一种基于自注意力机制的深度学习模型，相较于传统 AI 模型如循环神经网络（RNN）和卷积神经网络（CNN），它在处理序列数据时具有更高的并行性和可扩展性。其中，自注意力机制使得模型能够捕捉序列中长距离依赖关系，同时避免了 RNN 中的梯度消失或爆炸问题。Transformer 模型的参数量之所以会随着数据量和任务复杂度无限增长，是因为它可以通过堆叠更多的层或增加隐藏层宽度来提高模型性能，从而适应更复杂的数据和任务；在传统 CNN/RNN 模型中，增加网络参数量会提高模型的拟合能力，但过多的参数容易导致过拟合现象。这意味着模型可能在训练集上表现良好，但在测试集或实际应用中的泛化能力较差。市面上主流的语言模型分为四类：

- 自回归模型（Autoregressive Language Models）：利用因果注意力来预测下一个标记，擅长生成式 NLP 任务。如：GPT（OpenAI）。
- 自编码模型（Autoencoder Models）：更适合文本理解和分析任务，如情感分析、文本分类、实体识别等。如：BERT 模型（Google）。
- 编码/解码模型（Encoder-Decoder）：可以更好地捕捉输入序列的上下文信息，非常适合处理如机器翻译、文本摘要等任务。如：BART 模型（Facebook）。
- 广义自回归模型（Generalized Autoregressive）：XLNet 通过对输入序列的所有可能排列进行建模，能够捕捉到双向的上下文信息；擅长文本分类、命名实体识别和情感分析等。如：XLNet（Google+卡内基梅隆大学）。

以上各类模型各有优劣，我们认为自回归模型、编码/解码模型、广义自回归模型都有较高的商业化前景。

图 3：几类 Transformer 模型的举例



资料来源：Aman.ai，信达证券研发中心

AI 潜在垄断的形成过程、AI 的竞争要素、AI 的历史地位

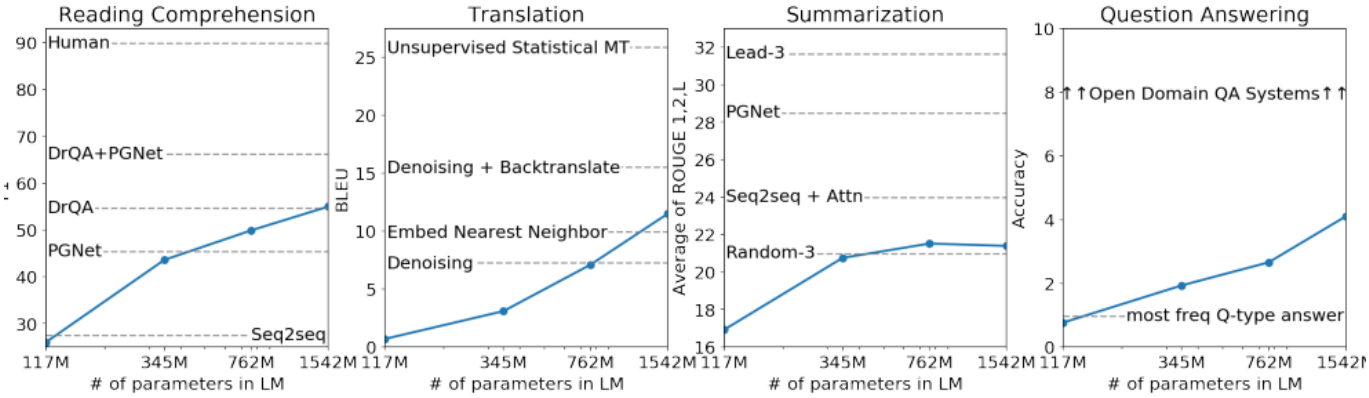
以下，我们从 GPT 及 SAM 论文出发以分析 AI 行业的长期发展趋势：

- **Language Models are Unsupervised Multitask Learners(GPT2)**

中文译为《语言模型是无监督多任务学习器》，指用于模型训练的数据集无需特别标注，只需训练模型通过前文预测后文的能力就能够产生很好的问答效果。文章指出，通过训练语言模型来预测下一个单词或字符的概率，模型可以学习自然语言处理任务，例如问答、机器翻译、阅读理解和摘要生成等。此外，论文还展示了如何将条件信息与语言模型相结合来执行特定任务，这种方法称为零样本学习，因为它不需要任何特定于任务的标记数

据。因此，我们可以说语言模型是无监督多任务学习器。同时，训练 GPT 的数据来源并不需要特别标注，只需要筛选互联网中的高质量的文本内容就可以形成优秀的模型。

图 4：模型能力从参数体量的增加中受益

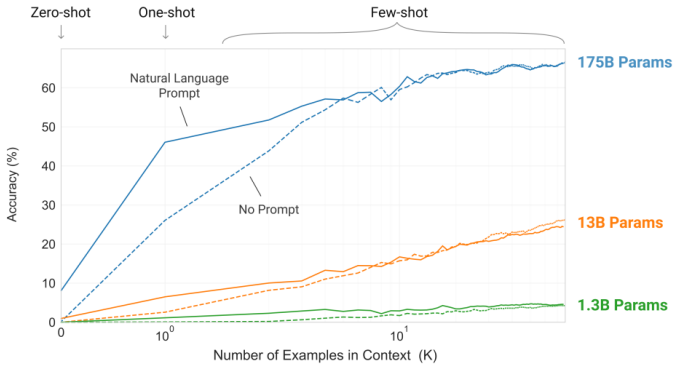


数据来源：Language Models are Unsupervised Multitask Learners（Alec Radford 等），信达证券研发中心

➢ **Language Models are Few-Shot Learners(GPT3)**

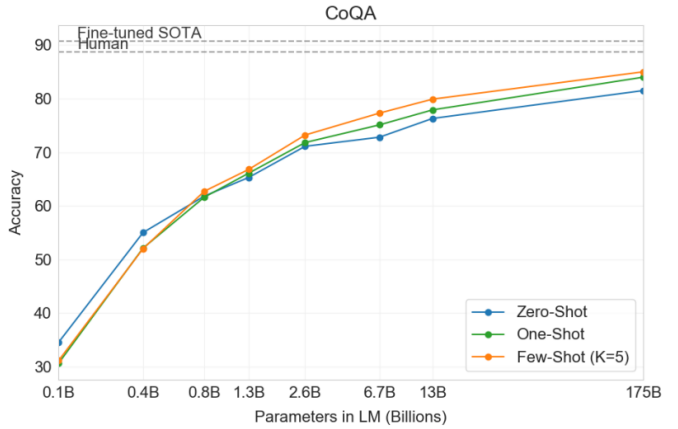
中文译为《语言模型是少样本学习者》，指随着模型参数量的增加，即使针对没有特殊训练的垂直领域，模型的精确度也能快速上升。文章指出，大型语言模型可以通过在大量文本语料库上进行预训练来开发广泛的技能和模式识别能力。这些技能和能力可以在推理时用于快速适应或识别所需的任务。作者使用“上下文学习”一词来描述这个过程的内部循环，该过程发生在每个序列的前向传递中。此外，与传统方法相比，扩大语言模型规模可以显著提高其任务无关、少样本性能，有时甚至可以达到之前最先进的微调方法的竞争水平。总结来看，提升参数体量可以让大型语言模型在没有样本的情况下，也能提高命中率；因此提升模型参数量是各家产品分出胜负的关键手段。

图 5：不同参数体量模型再零样本、少样本情况下命中率比较



数据来源：Language Models are Few-Shot Learners（Tom B. Brown 等），信达证券研发中心

图 6：随着参数增加，零样本/少样本阅读理解任务能力的提高



数据来源：Language Models are Few-Shot Learners（Tom B. Brown 等），信达证券研发中心

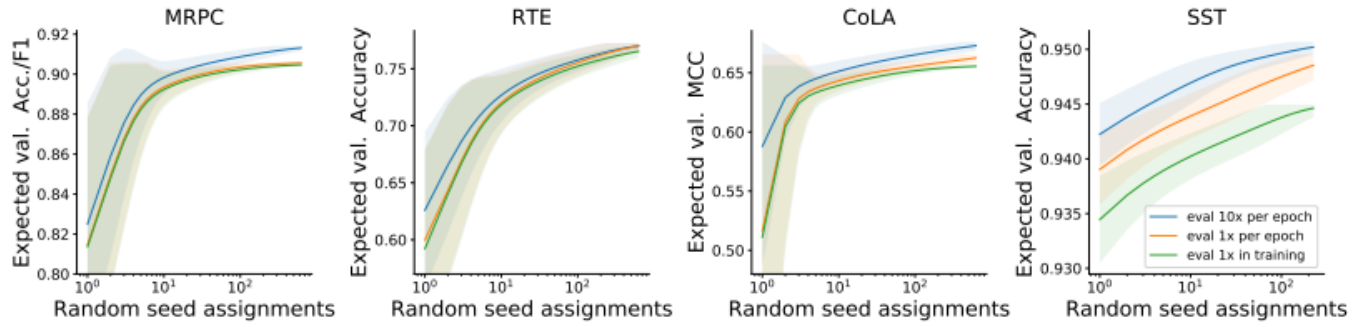
➢ **Fine-Tuning Large Language Models: Weight Initializations, Data Orders, and Early Stopping**

中文译为《微调大型语言模型：权重初始化、数据顺序和提前停止》，研究如何对 GPT 模型进行微调，商业化意义斐然。OpenAI 提出了以下几个微调预训练模型的建议：

- **选择合适的预训练模型：**不同的预训练模型可能适用于不同的任务和数据集。因此，需要根据实际情况选择最适合您任务的预训练模型。
- **选择合适的超参数：**超参数包括学习率、批量大小、正则化系数等。这些参数对微调过程和模型性能都有重要影响，需要进行仔细调整。

- **使用早期停止**：早期停止可以防止过拟合，并提高模型泛化能力。通常，可以使用验证集上的性能来确定何时停止训练。
- **数据增强**：数据增强可以帮助提高模型的鲁棒性和泛化能力。例如，在文本分类任务中，可以使用随机删除、替换或插入单词等方法来增加数据样本。
- **多次微调**：由于随机种子现象，即使使用相同的超参数值也可能导致不同的结果。因此，建议多次微调，并记录每次实验的结果。
- **合理评估模型性能**：在评估模型性能时，应该使用多个指标，并根据实际情况进行选择。同时，还应该注意避免过度拟合测试集。

图 7：即使只增加随机种子的数量，微调的预训练模型效果也会增强



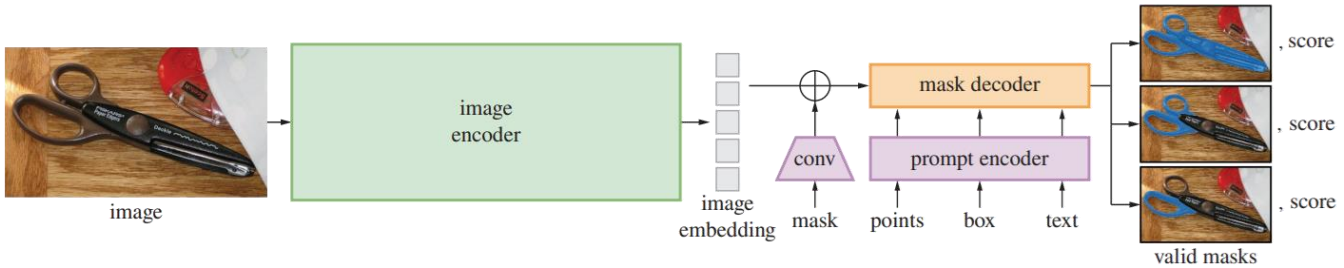
数据来源：Fine-Tuning Large Language Models: Weight Initializations, Data Orders, and Early Stopping（J Dodge 等），信达证券研发中心

➢ Segment Anything Mode

针对**零样本泛化、高能效图像分割模型（SAM，Segment Anything Mode）**，作者提到了传统的图像分割方法存在以下局限性：①需要大量标注数据；②计算资源要求高；③难以泛化到新的任务和场景。相比之下，SAM 模型具有以下技术上的不同之处：

- **可提示性**：SAM 模型是一种可提示模型，可以通过提供适当的提示来零样本迁移到新的图像分布和任务中。这使得它可以在缺乏大量标注数据的情况下进行训练和推理。
- **灵活性**：SAM 模型支持灵活的提示，可以根据不同的任务和场景进行调整。这使得它可以适应各种不同类型的图像分割问题。
- **高效性**：SAM 模型使用了一个轻量级的掩膜解码器来预测分割掩膜，从而实现了实时计算。这使得它可以在低功耗设备上运行，并且比传统方法更加高效。
- **泛化能力**：由于使用了可提示性和灵活性，SAM 模型具有更好的泛化能力，可以适应各种新任务和场景。
-

图 8：Segment Anything Model (SAM)概览



数据来源：Segment Anything（Alexander Kirillov 等），信达证券研发中心

本文总结三个核心结论：

1. **提升参数量=提升性能、提升泛化能力，长期垄断局面可能形成**：从论文研究来看，参数量的提升有助于构建语言预测模型的精确度，同时提高泛化能力。泛化能力的提升意味着一个参数量超级庞大的大模型，其在垂直细分领域的预测能力可超过针对垂直

领域开发的中等参数量模型，这意味着 AI 的发展长期也是强者恒强的垄断过程，即参数量超级庞大的模型在任何垂直领域都具备优势，垂直细分的小模型难有生产空间。

2. **数据标注的地位被弱化，AI 产业的经济竞争也是文化竞争：**无论是 GPT 还是 SAM，其在训练过程中，大量依靠互联网原生内容训练，因此一种语言的高质量文本内容的丰富程度，将决定基于该语言的大模型能力强弱，中文互联网文本内容生态亟待加强。
3. **AI 大模型至少是一次中等规模的产业革命：**仅从时间节点 ChatGPT 的表现来看，AI 的能力边界取决于过去人类产生的知识，它对于工业来说就是极大地降低了知识获取难度。将人类知识的海洋汇聚于一个语言的入口，它更像是 windows 之于电脑，开启了普通人接触高性能设备的通道，也开启了电子产品/AI 从企业端进入消费端的大门。

针对电子行业，我们认为在大模型格局未完全形成之前，参数量的无上限堆砌是各家企业发力的焦点，故算力的“军备竞赛”无可避免，若以动态视角来看 AI 芯片及服务器相关上游的弹性存在超预期可能。建议关注：兴森科技、兆威机电、芯原股份、寒武纪、通富微电。

风险因素

1.技术迭代不及预期；

2.地缘政治风险；

3.技术路径、产业趋势发生重大变化。

研究团队简介

莫文宇，毕业于美国佛罗里达大学，电子工程硕士，2012-2022 年就职于长江证券研究所，2022 年入职信达证券研发中心，任副总经理、电子行业首席分析师。

韩宇杰，电子行业研究员。华中科技大学计算机科学与技术学士、香港中文大学硕士。研究方向为半导体设备、半导体材料、集成电路设计。

郭一江，电子行业研究员。本科兰州大学，研究生就读于北京大学化学专业。2020 年 8 月入职华创证券电子组，后于 2022 年 11 月加入信达证券电子组，研究方向为光学、消费电子、汽车电子等。

机构销售联系

区域	姓名	手机	邮箱
全国销售总监	韩秋月	13911026534	hanqiuyue@cindasc.com
华北区销售总监	陈明真	15601850398	chenmingzhen@cindasc.com
华北区销售副总监	阙嘉程	18506960410	quejiacheng@cindasc.com
华北区销售	祁丽媛	13051504933	qiliyuan@cindasc.com
华北区销售	陆禹舟	17687659919	luyuzhou@cindasc.com
华北区销售	魏冲	18340820155	weichong@cindasc.com
华北区销售	樊荣	15501091225	fanrong@cindasc.com
华北区销售	秘侨	18513322185	miqiao@cindasc.com
华北区销售	李佳	13552992413	lijia1@cindasc.com
华北区销售	赵岚琦	15690170171	zhaolanqi@cindasc.com
华北区销售	张斓夕	18810718214	zhanglanxi@cindasc.com
华东区销售总监	杨兴	13718803208	yangxing@cindasc.com
华东区销售副总监	吴国	15800476582	wuguo@cindasc.com
华东区销售	国鹏程	15618358383	guopengcheng@cindasc.com
华东区销售	朱尧	18702173656	zhuyao@cindasc.com
华东区销售	戴剑箫	13524484975	daijianxiao@cindasc.com
华东区销售	方威	18721118359	fangwei@cindasc.com
华东区销售	俞晓	18717938223	yuxiao@cindasc.com
华东区销售	李贤哲	15026867872	lixianzhe@cindasc.com
华东区销售	孙僮	18610826885	suntong@cindasc.com
华东区销售	贾力	15957705777	jiali@cindasc.com
华东区销售	石明杰	15261855608	shimingjie@cindasc.com
华东区销售	曹亦兴	13337798928	caoyixing@cindasc.com
华东区销售	王赫然	15942898375	wangheran@cindasc.com
华南区销售总监	王留阳	13530830620	wangliuyang@cindasc.com
华南区销售副总监	陈晨	15986679987	chenchen3@cindasc.com
华南区销售副总监	王雨霏	17727821880	wangyufei@cindasc.com
华南区销售	刘韵	13620005606	liuyun@cindasc.com
华南区销售	胡洁颖	13794480158	hujieying@cindasc.com
华南区销售	郑庆庆	13570594204	zhengqingqing@cindasc.com
华南区销售	刘莹	15152283256	liuying1@cindasc.com
华南区销售	蔡静	18300030194	caijing1@cindasc.com
华南区销售	聂振坤	15521067883	niezhenkun@cindasc.com
华南区销售	宋王飞逸	15308134748	songwangfeiyi@cindasc.com

分析师声明

负责本报告全部或部分内容的每一位分析师在此申明，本人具有证券投资咨询执业资格，并在中国证券业协会注册登记为证券分析师,以勤勉的职业态度,独立、客观地出具本报告；本报告所表述的所有观点准确反映了分析师本人的研究观点；本人薪酬的任何组成部分不曾与，不与，也将不会与本报告中的具体分析意见或观点直接或间接相关。

免责声明

信达证券股份有限公司(以下简称“信达证券”)具有中国证监会批复的证券投资咨询业务资格。本报告由信达证券制作并发布。

本报告是针对与信达证券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。本报告仅提供给上述特定客户，并不面向公众发布。信达证券不会因接收人收到本报告而视其为本公司的当然客户。客户应当认识到有关本报告的电话、短信、邮件提示仅为研究观点的简要沟通，对本报告的参考使用须以本报告的完整版本为准。

本报告是基于信达证券认为可靠的已公开信息编制，但信达证券不保证所载信息的准确性和完整性。本报告所载的意见、评估及预测仅为本报告最初出具日的观点和判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会出现不同程度的波动，涉及证券或投资标的的历史表现不应作为日后表现的保证。在不同时期，或因使用不同假设和标准，采用不同观点和分析方法，致使信达证券发出与本报告所载意见、评估及预测不一致的研究报告，对此信达证券可不发出特别通知。

在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，也没有考虑到客户特殊的投资目标、财务状况或需求。客户应考虑本报告中的任何意见或建议是否符合其特定状况，若有必要应寻求专家意见。本报告所载的资料、工具、意见及推测仅供参考，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人做出邀请。

在法律允许的情况下，信达证券或其关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能会为这些公司正在提供或争取提供投资银行业务服务。

评级说明

投资建议的比较标准	股票投资评级	行业投资评级
本报告采用的基准指数：沪深 300 指数（以下简称基准）； 时间段：报告发布之日起 6 个月内。	**买入：**股价相对强于基准 20%以上；	**看好：**行业指数超越基准；
	增持：股价相对强于基准 5%～20%；	**中性：**行业指数与基准基本持平；
	持有：股价相对基准波动在±5%之间；	**看淡：**行业指数弱于基准。
	卖出：股价相对弱于基准 5%以下。	

风险提示

证券市场是一个风险无时不在的市场。投资者在进行证券交易时存在赢利的可能，也存在亏损的风险。建议投资者应当充分深入地了解证券市场蕴含的各项风险并谨慎行事。

本报告中所述证券不一定能在所有的国家和地区向所有类型的投资者销售，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专业顾问的意见。在任何情况下，信达证券不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任，投资者需自行承担风险。