2024 年 04 月 03 日

计算机行业研究

买入（维持评级）

行业深度研究

证券研究报告

计算机组
分析师：孟灿（执业 S1130522050001）　联系人：赵彤
mengcan@gjzq.com.cn　zhaotong3@gjzq.com.cn

如何实现 AGI：大模型现状及发展路径展望

投资逻辑：

目前大模型能力仍处于 Emerging AGI 水平，就模型成熟度而言，语言大模型>多模态大模型>具身智能大模型。根据 DeepMind 的定义，AGI 应能够广泛学习、执行复杂多步骤的任务。模型的 AGI 水平可分为 Level-0 至 Level-5 共 6 个等级，现阶段大模型在处理任务的广泛性上还有很大提升空间，即使是国际顶尖的大模型也仍处于 Level-1 Emerging AGI 阶段。不同类型大模型成熟度差异较大，目前大语言模型能力相对完善，落地应用场景丰富，底层技术路线较为成熟；多模态大模型已经能够面向 B\C 端推出商业化产品，但细节优化空间较大；具身智能类大模型还在探索阶段，技术路线尚不清晰。

现阶段讨论 AGI 能力提升仍需聚焦于多模态大模型的训练和应用。目前学界和业界重点关注 Scaling Law 的有效性，以及模型算法的可能改进方向。

- Scaling Law 仍有深入空间。根据 OpenAI 研究，随模型参数量、数据集规模、训练使用的计算量增加，模型性能能够稳步提高，即 Scaling Law。从训练样本效率、训练时长、各类资源对模型的贡献维度来看，目前 Scaling Law 仍是提高模型性能的最优方法。OpenAI 测算在模型参数量扩展到 88 万亿及之前，Scaling Law 依旧有效，则中短期仍可延续此路线进行训练。
- 模型骨干网络架构尚未演变至终局，微调及稀疏结构成为提升模型性能的重要方法。目前主流大模型均采用 Transformer 作为底层骨干网络，但针对编码器\解码器选择、多模态融合、自注意力机制等方面的探索仍在持续推进。微调使用更小的数据量、更短的训练时间，让模型能够适应下游任务，以降低边际落地成本。以 MoE 为代表的稀疏结构通过分割输入任务并匹配专家模型，能够提高模型的整体性能。

开源模型性能优化速度快于闭源模型。我们认为，目前第一梯队 AI 大模型纷纷进军万亿参数，且不远的将来大模型将逐步逼近十万亿参数收敛值，对于本轮 AI 浪潮而言，找场景或优于做模型。在场景选择方面，对“幻觉”容忍度高且能够替代人工的场景可实现应用率先落地，如聊天机器人、文本/图像/视频创作等领域；而对“幻觉”容忍度较低的行业需要等待大模型能力提升或使用更多场景数据训练。

投资建议

算法、数据、算力是影响模型性能的关键因素，相关企业能够直接受益于大模型训练的持续推进，推荐国内 AI 算法龙头科大讯飞等，建议关注数据工程供应商以及算力产业链相关公司。对于行业类公司而言，寻找通过 AI 赋能带来效率提升的场景更为重要，建议关注 AI+办公领域的金山办公、万兴科技，AI+安防领域的海康威视，AI+金融领域的同花顺等公司。

风险提示

底层大模型迭代发展不及预期；国际关系风险；应用落地不及预期；行业竞争加剧风险。

内容目录

图表目录

2022 年 11 月 ChatGPT 推出后，自然语言处理领域取得重大突破，正式进入大模型时代，2023 年被称为“大模型元年”；2023 年 3 月，具备多模态能力的 GPT-4 惊艳发布，海内外科技巨头、研究机构等纷纷跟进；至 2024 年 2 月 Sora 面世，大模型在视频生成领域实现代际跃迁，虚拟现实成为可能。在此背景下，学界和业界对于大模型终局，即是否能够实现 AGI（Artificial general Intelligence，通用人工智能）的讨论热度日益提升。

本文主要盘点目前各类主流大模型性能情况，试图讨论大模型性能提升并最终实现 AGI 的可能路径，并分析在实现 AGI 过程中的相关产业链投资机会。

1. 距离 AGI 还有多远：语言大模型较为成熟，处于 Emerging AGI 水平

根据 DeedMind 的创始人兼首席 AGI 科学家 Shane Legg 的定义，AGI 能够执行一般人类可完成的认知任务、甚至超越这个范围。具体而言，AGI 应能够学习广泛任务，能够执行复杂、多步骤的任务。DeepMind 根据 AI 模型性能和学习处理任务的广泛性对 AGI 水平进行分类，从 Level-0 无人工智能，到 Level-5 超越人类共 6 个等级。

图表 1：AGI 可以根据性能和广泛性划分为 6 个等级

等级	主要特征
Level-0 无人工智能（Narrow Non-AI）	·只能完成明确定义的任务，比如计算器软件或编译器
Level-1 初现（Emerging AGI）	·性能相当于或略优于一个不熟练的人类。比如一些前沿语言模型在某些任务上已经达到了初现 AGI 的水平
Level-2 熟练（Competent AGI）	·至少能够在大多数任务上达到熟练人类的水平。目前的前沿语言模型在某些任务上已经接近熟练 AGI 的水平
Level-3 专家（Expert AGI）	·在大多数任务上能够达到专家人类的水平
Level-4 大师（Virtuoso AGI）	·在大多数任务上能够达到顶尖人类的水平
Level-5 超越人类（Superhuman AGI）	·在所有任务上都能超过 100% 的人类

来源：《Levels of AGI: Operationalizing Progress on the Path to AGI》，国金证券研究所

现阶段大模型在处理任务的广泛性上还有很大提升空间，虽然 GPT-4、Gemini 1.5、Claude 3 等模型已经能够处理文本、图像、视频等多模态输入，但尚未具备独立决策和执行行动的能力。此外，现阶段更多的模型仍聚焦在某单一领域进行性能提升，比如 Kimi 在处理长文本输入领域表现突出，但尚不能进行图片生成；Sora 能够高质量完成文生视频任务，但不具备问答功能。因此，现阶段评价大模型性能情况、分析模型演进方向，仍需根据模型专长领域进行分类。

图表 2：大模型可根据功能进行分类

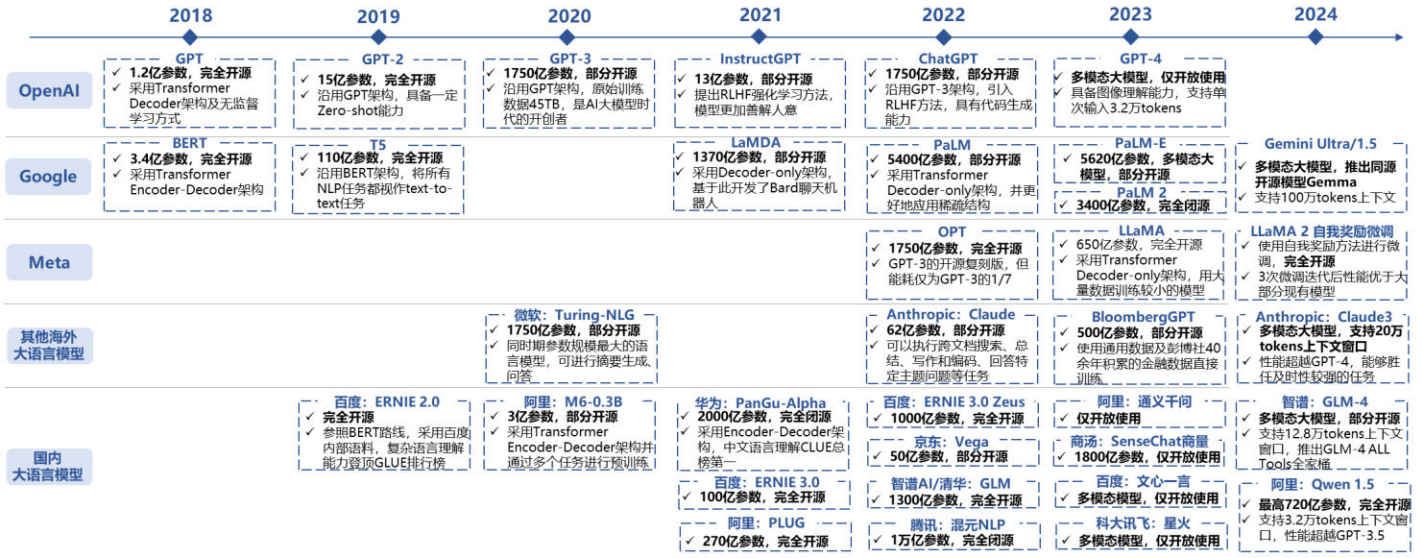
模型分类	主要内容	代表模型
语言大模型	·专注于处理自然语言，能够理解、生成和处理大规模文本数据 ·用于机器翻译、文本生成、对话系统等任务	ChatGPT、Llama
视觉大模型	·专注于计算机视觉任务，如图像分类、目标检测、图像生成等 ·能够从图像中提取有关对象、场景和结构信息	ViT、SAM
多模态大模型	·能够处理多种不同类型的数据，如文本、图像、音频等，并在这些数据之间建立关联 ·多模态大模型能够处理文图融合、图像描述、文生视频等任务	GPT-4、Claude3
策略大模型	·专注于进行决策和规划，能够在面对不确定性和复杂环境时做出智能决策，可用于机器人控制	AlphaGo、RT-1/2/H

来源：金科应用研究院公众号，国金证券研究所

- 在语言大模型以及偏重问答能力的多模态模型领域，自 2020 年 GPT-3 发布后进入爆发期，各主流玩家加速模型迭代，包括 OpenAI 的 GPT 系列、Google 的 Gemini 系列、Meta 的开源 LLaMA 系列等。目前定量测评分数最高的为 Anthropic 旗下的 Claude 3 Opus，在 MMLU（Undergraduate Level Knowledge）、GSM8K（Grade School Math）、MGSM（Multilingual Math）等多个测试项目中准确率超过 85%；模型参数量最高的为 23 年 3 月谷歌发布的 PaLM-E，参数量达到 5,620 亿，是 ChatGPT 的 3.2 倍，模型能够理解自然语言及图像，还可以处理复杂的机器人指令；谷歌于

24 年 2 月发布的 Gemini 1.5 能够处理的上下文长度高达 100 万 tokens（相当于 70 万单词，或 3 万行代码，或 11 小时音频，或 1 小时视频），为目前长文本处理能力的上限。

图表3：海内外语言及多模态大模型进展概览



来源：《Large Language Models: A Survey》，《A Survey of Large Language Models》，洞见学堂公众号，机器之心公众号，级市平台公众号，新智元公众号，阿里云开发者社区，京东技术公众号，中国科学基金公众号，数据派 THU 公众号，浙江省软件行业协会公众号，深圳大学可视计算研究中心公众号，量子位公众号，钛媒体 AGI 公众号，彭博 Bloomberg 公众号，腾讯科技公众号，百度 AI 公众号，鹏城实验室公众号，CSDN 公众号，文心大模型公众号，中国人工智能学会公众号，腾讯开发者公众号，阿里云公众号，商汤智能产业研究院公众号，36 氪，科大讯飞公众号，科大讯飞开发者平台，GLM 大模型公众号，阿里通义千问公众号，国金证券研究所

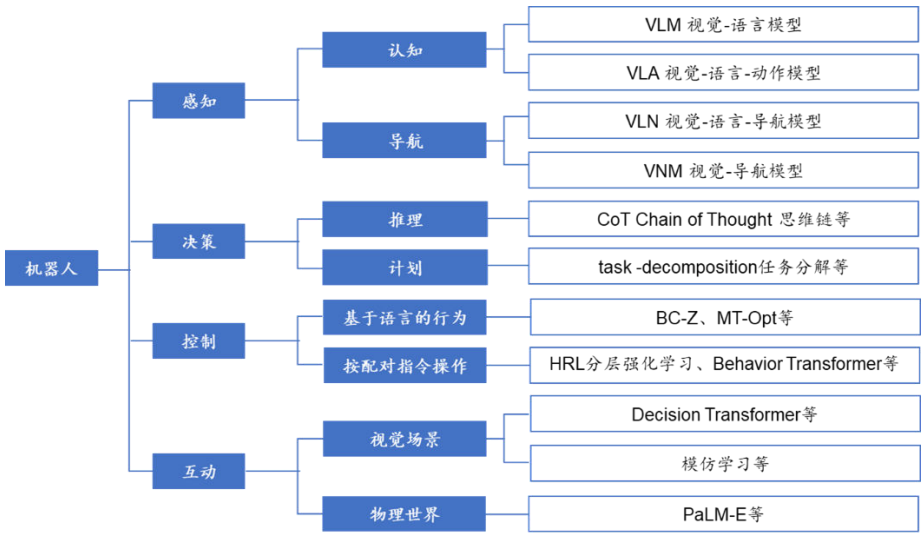
- 文生图、文生视频类模型可追溯至 2014 年的 GAN 框架，2021 年 OpenAI 发布 DALL-E 后图像生成类模型开始爆发，包括谷歌的 Imagen、OpenAI 的 DALL-E 2、Stability 旗下的 Stable Diffusion；至 2023 年文生图功能与大语言模型相结合，并出现文生视频技术，24 年 2 月 OpenAI 发布文生视频模型 Sora，在生成视频长度和质量上均为目前最优水平。

图表4：海内视觉及其他大模型进展概览

来源：《Large Language Models: A Survey》，《Improved protein structure prediction using potentials from deep learning》，《High-Resolution Image Synthesis with Latent Diffusion Models》，机器之心公众号，新智元公众号，信息与电子工程前沿公众号，级市平台公众号，AI 科技评论公众号，AIGC 开放社区公众号，腾讯研究院公众号，中国生物技术网公众号，数据派 THU 公众号，阿里云公众号，智源社区公众号，百度 AI 公众号，中国企业家俱乐部公众号，商汤科技 SenseTime 公众号，商汤智能产业研究院公众号，AIGC 视界公众号，飞书公众号，搜狐科技公众号，AIGC Research 公众号，智东西公众号，国金证券研究所

- 机器人模型包括感知、决策、控制、交互 4 个部分，涉及视觉、图像、声音、导航、动作等多个模态，在实际应用中需要根据特定的环境、动作、障碍、反馈等数据进行决策，因此，机器人对算法的跨模态、泛用性要求更高。

图表5：机器人涉及到的模型种类较多



来源：《Large Language Models for Robotics: A Survey》，国金证券研究所

将语言大模型的底层框架和训练方式应用于机器人的感知、决策、控制成为现阶段重要趋势。2021 年 OpenAI 推出基于 Transformer 架构和对比学习方法的 VLM（视觉-语言模型）CLIP；2022 年起，谷歌先后推出 RT-1/RT-2/RT-X/RT-H 系列模型，同样采用 Transformer 架构，能够将语言描述的任务映射为机器人行动策略；24 年 3 月，初创公司 Figure 与 OpenAI 合作推出机器人 Figure01，由 OpenAI 提供视觉推理和语言理解能力，Figure01 能够描述看到的一切情况、规划未来的行动、语音输出推理结果等。

图表6：将 Transformer 架构应用于机器人决策、控制等成为现阶段重要趋势

模型名称	发布时间	发布机构	功能类别	主要内容
CLIP	2021	OpenAI	感知-VLM	· 网络结构主要包含 Text Encoder 和 Image Encoder 两个模块，分别提取文本和图像特征，然后基于比对学习让模型学习到文本-图像的匹配关系； · CLIP 使用大规模数据(4 亿文本-图像对)进行训练，基于海量数据，CLIP 模型可以学习到更多通用的视觉语义信息，可应用于图像文本匹配、图像文本检索等任务。
LM-Nav	2022	谷歌	计划	· LLM\VLM\VNM 三个模型的结合，LLM 用于提取指令中的地标，VLM 用于将文本地标与图像关联，而 VNM 用于执行导航任务； · 系统以目的地环境的初始观察结果、以及用户给的文本指令作为输入，通过系统中的三个预训练模型得出执行计划。
RT-1	2022	谷歌	决策、控制	· 建立在一个 transformer 架构上，该架构从机器人相机中获取瞬时图像以及以自然语言表达的任务描述作为输入，并直接输出 tokenized 动作； · RT-1 可以以 97% 的成功率执行 700 多个训练指令，并且可以泛化到新的任务、干扰因素和背景。
PaLM-E	2023.3	谷歌	感知-VLM、控制	· 通过 PaLM-540B 语言模型与 ViT-22B 视觉 Transformer 模型相结合，PaLM-E 最终的参数量高达 5620 亿，其训练数据为包含视觉、连续状态估计和文本输入编码的多模式语句； · PaLM-E 不仅可以指导机器人完成各种复杂的任务，还能生成描述图像的语言。
RT-2	2023.7	谷歌	感知、决策、控制	· 使用 Transformer 架构的视觉-语言-动作模型，能够从网络和机器人数据中进行学习，并将这些知识转化为机器人可以控制的通用指令 · 在机器人训练中未见过的场景中，准确性由 RT-1 的 32%提高到 62%
RT-X	2023.10	谷歌	感知、决策、控制	· 由基于 Transformer 的 RT-1-X 模型和视觉语言动作模型 RT-2-X 组成。RT-1-X 模型在特定任务上的平均性能比 RT-1 模型和原始模型提升 50%。RT-2-X 的涌现能力约为 RT-2 的 3 倍，动作指令也可从传统的绝对位置拓展至相对位置
RT-H	2024.3	谷歌	感知、决策、控制	· 能通过将复杂任务分解成简单的语言指令，再将这些指令转化为机器人行动，来提高任务执行的准确性和学习效率； · RT-H 的 MSE 比 RT-2 低大约 20%，这表明行动层级有助于改进大型多任务数据集中的离线行动预测

来源：极市平台公众号，DeepTech 深科技公众号，机器之心公众号，OSC 开源社区公众号，国金证券研究所

按照 DeepMind 的 6 级 AGI 水平分类，目前国际顶尖大模型仍处于 Level-1 Emerging AGI

阶段。具体而言，各类大模型成熟度：语言大模型>多模态大模型>具身智能类大模型。

- 语言大模型能力相对完备，底层技术路线大多选择 Transformer Decoder-only 架构，结合 MOE 和多模态 embedding，算法细节优化方向区别较小。以 GPT-4、Gemini 1.5、Claude 3 为例，语言大模型在推理、长文本、代码生成领域已经能够完成初级任务，但距复杂、专业水平仍有差距；
- 多模态大模已经能够面向 B\C 端提供商业化产品，底层技术路线主要采用 Diffusion Transformer，但细节优化空间较大，高质量和成规模的数据集仍在发展初期；
- 具身智能类大模还在探索阶段，底层技术路线尚不清晰，数据收集、训练方法、测评方法等都处于发展初期。在实际应用场景中准确率较低。

图表 7：各类大模型能力现状

模型分类	主要内容
语言大模型	· **常规测试：**超越入门级人类，距离特定领域专家还有一定差距； · **推理：**常识 /入门数学/基础科学正确率高，面对复杂任务（如研究生级别、竞赛类问题等）还有差距； · **长文本：**比较完善，输入长度可拓展至 10M tokens，能够满足绝大部分场景需求 · **代码生成：**简单任务正确率高，复杂任务（工业级、竞赛级等）仍有差距 · **多模态理解：**定性分析能力较高，但定量分析错误率较高
多模态大模型	· **基本生成：**风格不稳定较难控制； · **语义理解：**已有大幅改善但仍需要提供准确的 prompt； · **清晰度：**分辨率可达 4K 以上，基本满足商业应用但生成速度较慢； · **一致性/连贯性：**最新文生视频模型 Sora 可生成 60s，但仍不够稳定
具身智能类大模型	· **任务感知：**技术路径多，在物体距离、材质关键信息的提取和识别等任务中表现不稳定，需要依靠执行过程中动态调整； · **行动规划：**在简单任务（如家庭环境中的简单操作）中仍然准确率较低，在复杂多步骤任务中不可用，新任务泛化能力弱，并且延迟较为严重； · **运控算法：**深度学习方法尝试初期，大部分仍然依靠 hard-coding

来源：彬复资本公众号，国金证券研究所

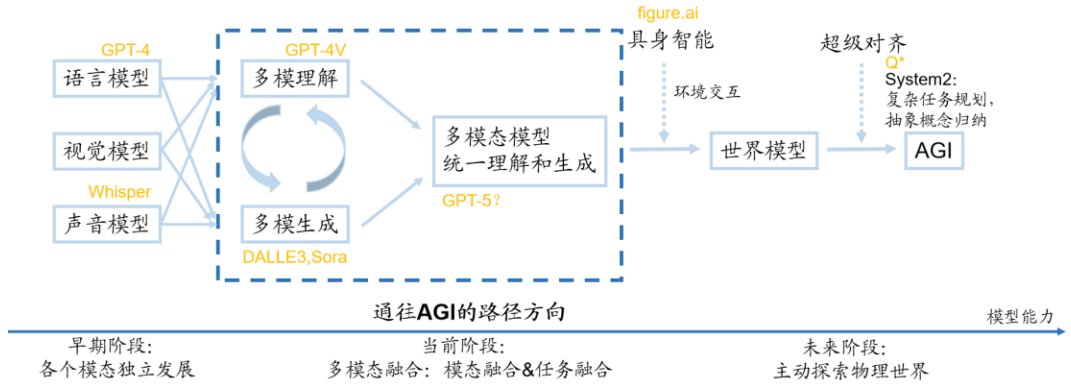
2. 如何实现 AGI：Scaling Law 仍有深入空间，底层算法框架有待升级

参考 OpenAI 的大模型研发布局，实现 AGI 的过程可以分为 3 个阶段，目前处于多模融合的第 2 阶段。AGI 发展路径与当下各类模型水平相对应，即在语言大模型、视觉大模型相对成熟的基础上发展多模态大模型，而后探索具身智能类应用。

- 第一阶段为单模态系统，包括语言模型、视觉模型、声音模型等，各个模态独立发展；
- 第二阶段为多种模态、多种任务模型相融合。但根据模型的能力侧重点不同仍可分为两类：一是以 GPT-4 等为代表的多模态理解模型，二是更强调生成性能的多模态生成模型，如 Sora 等。预计这两种能力会在后续的大模型发展中进一步融合。
- 第三阶段将进一步强调模型与外部环境的交互，以及面对复杂任务的处理能力，将以机器人或者一个设备的大脑为载体，进一步靠近乃至实现 AGI。

因此，现阶段讨论 AGI 能力提升仍需聚焦于多模态大模型的训练和应用，在多模理解和多模生成能力较好融合后，再推演具身智能的模型框架、训练方法会更加清晰。

图表8：以 OpenAI 布局为例，看 AGI 发展路径



来源：阶跃星辰公众号，国金证券研究所

多模态大模型与语言大模型的训练流程相似，包括数据工程和模型工程两部分。其中数工程包括数据清洗、分词、位置编码等，模型工程包括模型框架选择、训练方法选择、算法选择、模型预训练、微调等。模型的预训练、微调、推理等环节均需要算力支持。

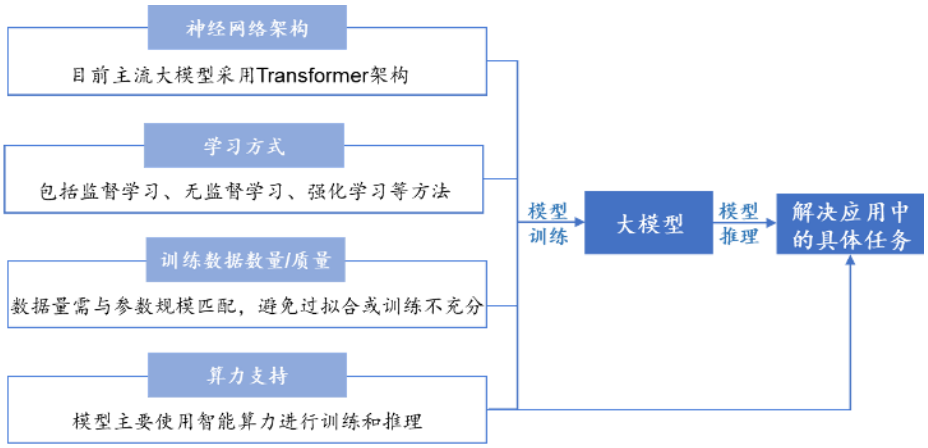
图表9：大模型训练主要环节

来源：《Large Language Models: A Survey》，国金证券研究所

模型架构及神经网络层数决定模型参数量，通常将参数规模千万量级及以上的深度学习模型称为“大模型”；训练使用的数据集大小需要与模型参数规模相匹配，避免产生过拟合或训练不充分等问题；算力需求与模型算法结构、参数规模等紧密相关。因此当我们讨论模型性能提升时，可以重点从神经网络架构和训练方法、数据量、算力等维度入手。本文后续章节将就目前学界和产业界重点关注的问题进行讨论：

- 在保持现有模型架构不变的情况下，增加神经网络层数，进而扩大参数规模、训练数据集规模的 Scaling Law 的天花板在哪里？
- 当仅凭 Scaling Law 不能进一步提升模型性能时，算法层面有哪些可以改进的方向？

图表10：多重因素决定模型性能

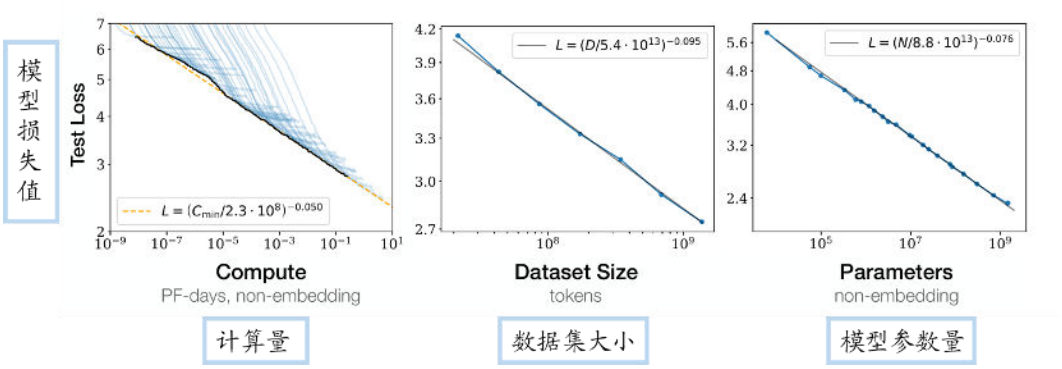


来源：国金证券研究所

2.1 Scaling Law：中短期内，持续扩大参数量仍能改善模型表现

OpenAI 通过研究证明，随着模型大小、数据集大小和训练所用计算量的增加，语言模型的性能也会稳步提高。为了获得最佳性能，这三个因素必须同时放大：1）当不被其他两个因素瓶颈限制时，模型性能表现与每个单独的因素之间存在幂律关系；2）在其他两个因素充足的前提下，模型表现和第三个因素成幂方关系。

图表11：模型性能随着模型大小、数据集大小和训练所用计算量的增加呈现幂律提升

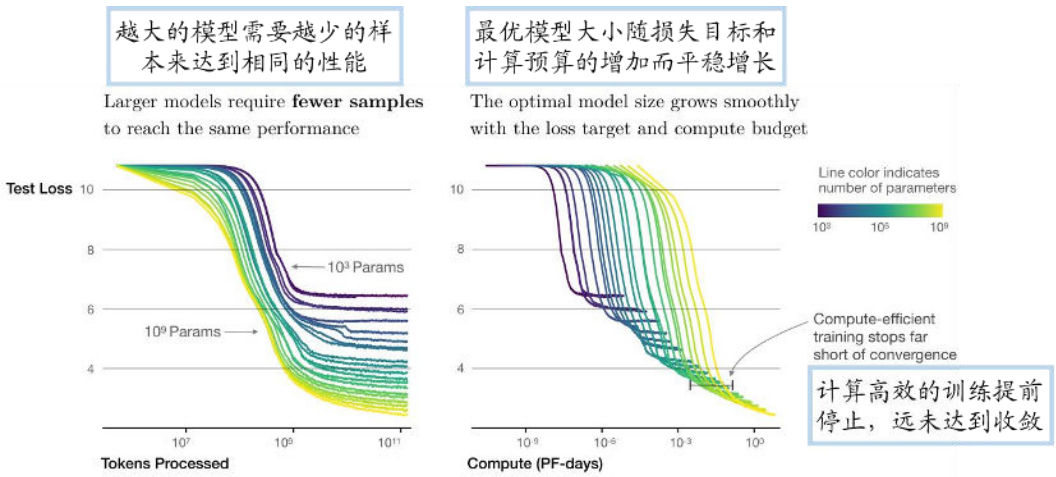


来源：《Scaling Lawsfor Neural Language》，国金证券研究所
说明：Test Loss 是指在机器学习或深度学习中，在测试数据集上计算的模型损失值。它用来衡量模型在未见过的数据上的性能表现，通常用于评估模型的泛化能力。Test Loss 的数值越低，表示模型在测试数据上的预测结果与真实结果的差距越小，模型的性能越好。

从训练样本效率、训练时长、各类资源对模型的贡献维度来看，目前 Scaling Law 仍是提高模型性能的最优方法：

- 参数规模更大的模型在训练过程中的样本效率更高、性能提升更快。当计算量固定（比如固定要进行 n 次浮点计算）而数据集大小 D 和模型参数量 N 不固定时，OpenAI 发现训练大模型的性价比更高。主要由于随着模型规模的增加，每个优化步骤所需的数据点数量减少，从而提高了样本效率。

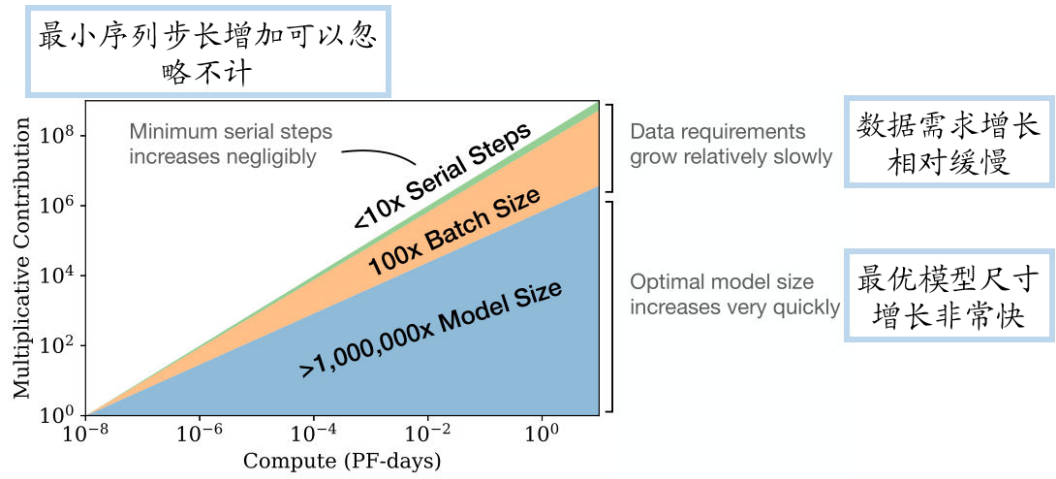
图表 12：参数规模更大的语言模型在训练过程中的样本效率更高且性能提升更快



来源：《Scaling Laws for Neural Language》，国金证券研究所

- 模型参数规模对于性能提升的贡献度更高。研究表明，在有限的资源下，为了达到最佳的训练效果，应当优先考虑扩大模型参数量 N，同时合理调整其他训练参数以保持训练效率和避免过拟合。

图表 13：模型参数规模对于性能提升的贡献度更高



来源：《Scaling Laws for Neural Language》，国金证券研究所

OpenAI 对 Scaling Law 的适用空间进行测算，认为模型参数量在扩展到 88 万亿及之前 Scaling Law 仍会发挥作用。目前业界预测 OpenAI 下一代大模型 GPT-5 参数量预计达到 10 万亿级别，神经网络层数或达 1,300 层，相较 88 万亿的“天花板”仍有扩充空间。因此，中短期来看持续扩大模型参数量仍有望改善模型表现。

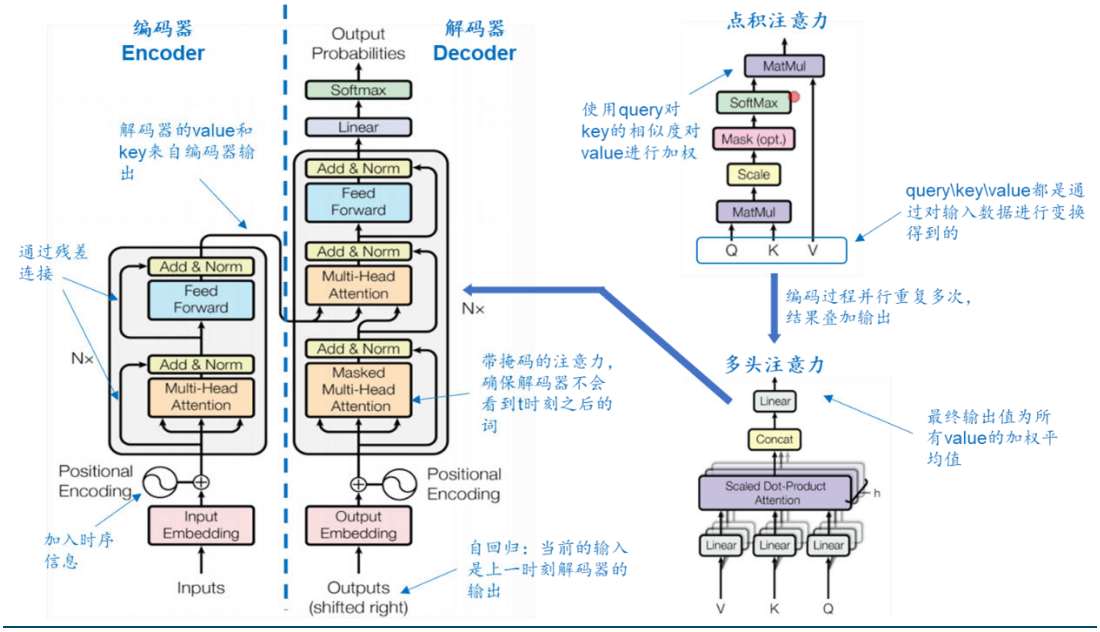
2.2 算法改进：骨干网络架构仍有创新空间，微调及稀疏结构能够提升性价比

我们曾经在 2023 年 3 月发布报告《大模型时代，AI 技术向效率提升演进》，对大模型训练方法、数据效率、开发效率、算力效率、工程化效率的发展趋势进行梳理，本节将结合过去一年的大模型迭代情况，进一步讨论可能的算法演进方向。

2.2.1 基于 Transformer，在架构选择、多模态融合、自注意力机制方面进行创新

2017 年谷歌将注意力机制引入神经网络，提出了新一代深度学习底层算法 Transformer。由于其在物体分类、语义理解等多项任务中准确率超过 CNN、RNN 等传统算法，且能应用于 CV、NLP 等多个模态，Transformer 的提出使得多任务、多模态的底层算法得到统一目前主流大模型均采用 Transformer 作为底层骨干网络，但在编码器\解码器选择、多模态融合、自注意力机制等方面有所创新。

图表14：Transformer 模型结构及自注意力机制原理



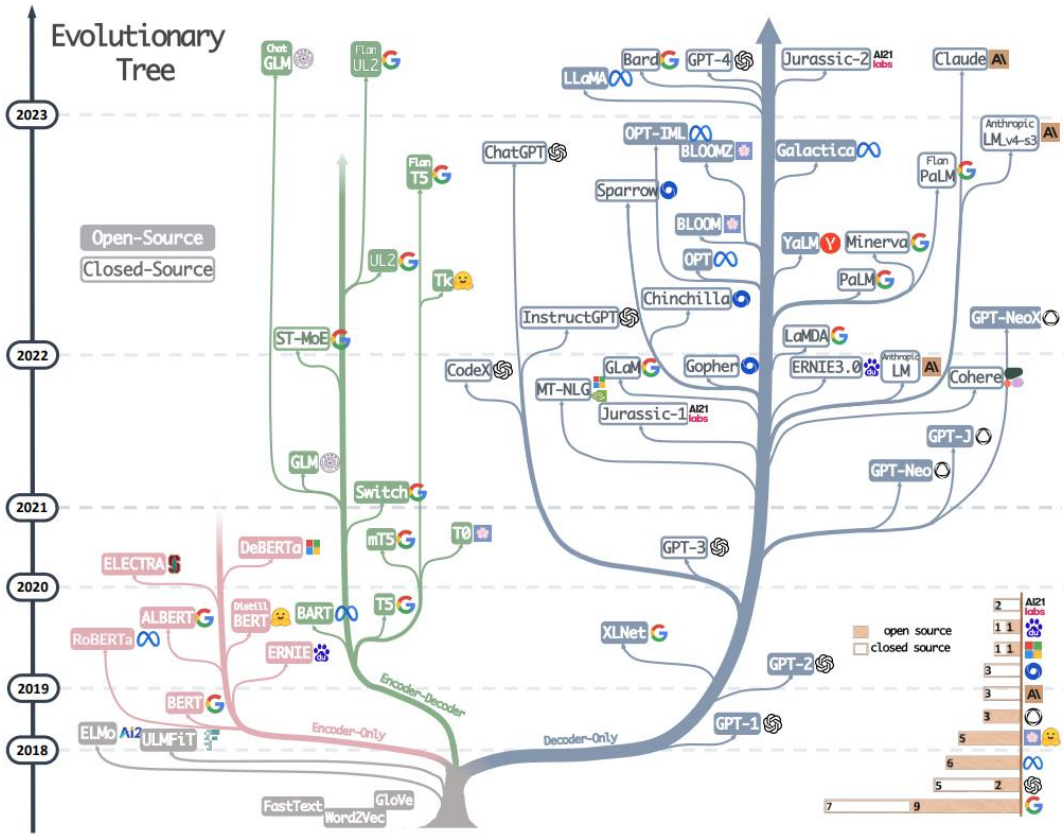
来源：《Attention Is All You Need》，国金证券研究所

- **针对 Transformer 的编码器-解码器（Encoder-Decoder）结构进行拆分选择：**

  Transformer 模型采用编码器-解码器结构，其中编码器负责从输入内容中提取全部有用信息，并使用一种可以被模型处理的格式表示（通常为高维向量）；而解码器的任务是根据从编码器处接收到的内容以及先前生成的部分序列，生成翻译后的文本或目标语言。

  目前主流大模型可以根据骨干网络架构的差异分 Encoder-only、Encoder-Decoder、Decoder-only 共 3 类，其中 Encoder-only 主要为谷歌的 Bert 及其衍生优化版本；使用 Encoder-Decoder 架构的模型有谷歌的 T5 以及清华智谱的 GLM 等；OpenAI 的 GPT 系列、Anthropic 的 Claude 系列、Meta 的 LLaMA 系列等均采用 Decoder-Only 架构。

图表15：根据底层骨干网络差异可以将大模型分为三类



来源：《Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond》，国金证券研究所

Decoder-Only 架构更适合生成类任务且推理效率更高，为大模型厂商所青睐：1）功能方面，Encoder-Only 架构更擅长理解类而非生成类任务，以采用 Encoder-Only 架构的 Bert 为例，其学习目标包括 Masked LM(随机遮盖句子中若干 token 让模型恢复)和 Next Sentence Prediction(让模型判断句对是否前后相邻关系)，训练目标与文本生成不直接对应；2）推理效率方面，Encoder-Decoder 和 Decoder-Only 架构均能够用于文本生成，但在模型效果接近的情况下，后者的参数量和占用的计算资源更少，且具有更好的泛化能力。

图表16：三种骨干网络特点对比

骨干架构	主要特点
Encoder-Only	· 适用于不需要生成序列、只需要对输入进行编码和处理的单项任务，如文本分类、情感分析等； · 核心思想是利用神经网络对输入文本进行编码，提取其特征和语义信息，并将编码结果传递给后续处理模块
Encoder-Decoder	· 通常用于序列到序列任务，如机器翻译、对话生成等； · 优点是能够更好地处理输入序列和输出序列之间的关系，从而提高机器翻译和对话生成等任务的准确性；缺点是模型复杂度高、训练时间和计算资源消耗较大
Decoder-Only	· 常用于序列生成任务，如文本生成、机器翻译等，可以进行无监督与训练； · 能够从已有的信息扩展出新的内容，但需要大量的训练数据来提高生成文本的质量和多样性

来源：《Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond》，Paperweekly 公众号，极市平台公众号，国金证券研究所

使用 Encoder-Decoder 亦可训练出成熟的多模态大模型，或将凭借架构优势在多任务融合领域实现反超。清华大学和智谱 AI 公司共同研发的 GLM 系列模型采用 Encoder-Decoder 架构，于 24 年 1 月发布 GLM-4 最新版本，模型在基础能力（英文）、指令跟随能力（中英）方面能够达到 GPT-4 90%以上的水平，在对齐能力（中文）、长文本能力方面超过 GPT-4，在文生图方面能力逼近 DALLE-3。目前 GLM4 支持根据用户意图，自由调用网页浏览器、代码解释器和文生图模型，并上线个性化

智能体定制功能。

图表17：智谱 GLM-4 在多项任务中能力比肩 GPT-4

基础能力（英文）

	MMLU (5-shot)	GSM8K (5-shot)	MATH (4-shot)	BBH (3-shot)	HellaSwag (10-shot)	HumanEval (0-shot)
GPT-4	86.4	92.0	52.9	83.1	95.3	67.0
Gemini-Ultra	83.7	94.4	53.2	83.6	87.8	74.4
GLM-4	81.5	87.6	47.9	82.3	85.4	72.0
GLM-4 / GPT-4	94%	95%	91%	99%	90%	100%

指令跟随能力（中英）

	IFEval Prompt级别、中文	IFEval Instruction级别、中文	IFEval Prompt级别、英文	IFEval Instruction级别、英文
GPT-4	72.4	80.0	79.5	85.4
GLM-4	63.4	71.9	67.7	76.4
GLM-4 / GPT-4	88%	90%	85%	89%

对齐能力（中文）

	专业能力	中文理解	基本任务	数学计算	文本写作	综合问答	角色扮演	逻辑推理	中文推理	中文语言	总分
GPT-4	7.94	6.93	7.81	7.65	7.93	7.42	7.51	7.37	7.47	7.59	7.53
GPT-4 Turbo	8.65	7.33	7.99	7.80	8.67	8.61	8.47	7.66	7.73	8.29	8.01
GLM-4	8.91	8.07	7.87	7.75	8.44	8.42	8.58	7.01	7.38	8.38	7.88
GLM-4 / GPT-4	112%	116%	101%	101%	106%	113%	114%	95%	99%	110%	105%

AlignBench 2023

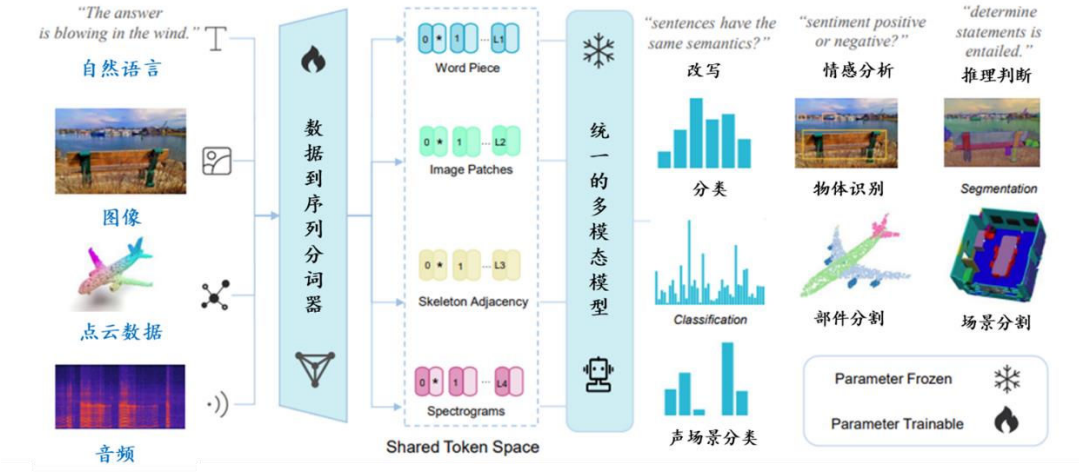
来源：GLM 大模型公众号，国金证券研究所

- **Transformer+其他现有算法，推进多模态性能提升：**

自注意力机制（Self-Attention）使得 Transformer 架构能够处理多模态任务。自注意力机制将输入数据进行线性映射创建三个新向量，分别为 Q/K/V，其中 Q 向量可以看作是某个人的关注点，V 向量可以看作是具体的事物，而 K 向量可以看作是人对不同事物的关注程度。通过计算 Q 向量和 K 向量的点乘，可以得出一个值，表示这个人对某个事物的关注程度，然后将这个关注程度与 V 向量相乘，以表示事物在这个人眼中的表现形式。这种方式使得模型能够更好地捕捉长序列中不同部分的关联性和重要性，而各种模态的信息均可以通过一定方式转化为一维长序列，因而 Transformer 具备处理多模态问题的能力。

以上海 AI Lab 和香港大学联合推出的 Meta-Transformer 为例，该模型通过一个多模态共享的分词器，将不同模态的输入映射到共享的数据空间中，进而实现了处理 12 种非成对的模态数据，包括文本、图像、点云、音频、视频、X 光、红外等。

图表18：Meta-Transformer 模型能够处理 12 种非成对的模态数据

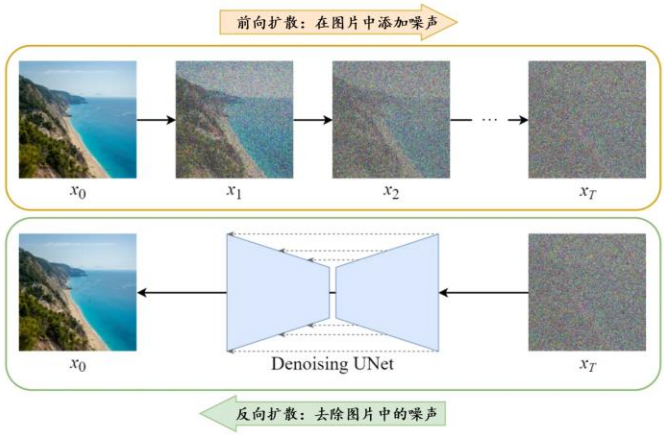


来源：《Meta-Transformer: A Unified Framework for Multimodal Learning》，国金证券研究所

将 Transformer 与其他模态领先算法融合，能够显著提升多模态处理能力，有望加速大模型多模态融合趋势。24 年 2 月 OpenAI 发布文生视频大模型 Sora，主要根据 Diffusion Transformer（DiT）框架设计而成。其中，扩散模型（Diffusion）是一种图

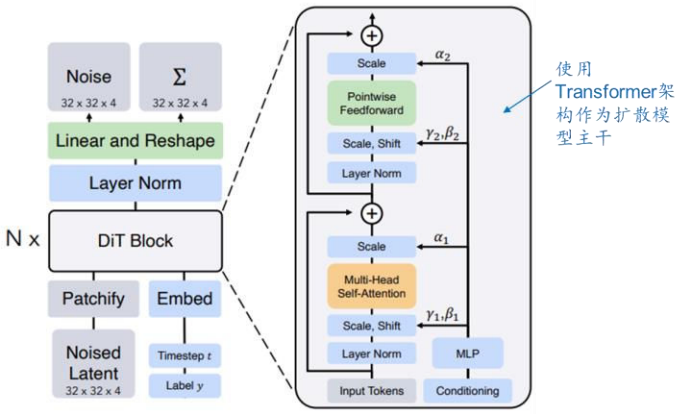
像生成方法，通过逐步向数据集中添加噪声，然后学习如何逆转这一过程。扩散模型能够生成高质量的图像和文本，但仍存在可扩展性低、生成效率低等问题。DiT 模型在扩散模型基础上引入 Transformer 架构，通过将图像分割成小块（patches），并将这些块作为序列输入到 Transformer 中，DiT 能够有效地处理图像数据，同时保持了 Transformer 在处理序列数据时的优势，能够显著改善扩散模型的生成效率。此外，将自动驾驶领域的 BEV（鸟瞰视图）模型与 Transformer 相结合，已经成为目前自动驾驶领域主流感知框架，并在众多辅助驾驶产品中量产应用。

图表19：扩散模型示意图



来源：极市平台公众号，国金证券研究所

图表20：Diffusion Transformer 模型结构



来源：《Scalable Diffusion Models with Transformers》，国金证券研究所

- **基于 Transformer 的细节创新已成为学界重点研究方向，非 Transformer 结构的探索持续推进，有望推动骨干网络升级。**

Transformer 自 2017 年发布后对深度学习领域产生颠覆性影响，学界在持续探索改变框架细节以实现模型性能进一步突破。目前针对 Transformer 的创新尝试包括模块改进、架构改进、效率优化等方面。华为诺亚方舟实验室等联合推出新型大语言模型架构盘古-π，通过增强非线性，在传统 Transformer 架构上做出改进，由此可以显著降低特征塌陷问题。在使用相同数据训练的情况下，盘古-π（7B）在多任务上超越 LLaMA 2 等同规模大模型，并能实现 10%的推理加速。

图表21：针对 Transformer 的创新研究持续推进

改进维度	相关论文	改进方法
自注意力机制	Rethinking Attention: Exploring Shallow Feed-Forward Neural Networks as an Alternative to Attention Layers in Transformers	探索浅层前馈神经网络作为 Transformer 中注意力层的替代方案，通过消融研究和替代网络试验，支持了该方法的可行性，表明浅层前馈网络在简化序列到序列任务的复杂架构方面具有潜力
	FLatten Transformer: Vision Transformer using Focused Linear Attention	使用聚焦线性注意力的视觉 Transformer，该模块适用于多种视觉转换器，并在多个基准测试中实现了性能提升
Transformer 架构	Simplifying Transformer Blocks	研究人员修改了模块，移除了跳过连接、投影或值参数、顺序子块和归一化层，以简化结构。在自回归解码器和 BERT 编码器模型实验中，简化版 Transformer 与标准版速度和性能相当，但训练吞吐量提高 15%，参数减少 15%
	Token Merging: Your ViT But Faster	提出了令牌合并（ToMe）方法，使用准确的匹配算法将相似标记组合在一起，使得图像和视频吞吐量大幅提升，精度下降很小。
	Effcient Long-Range Transformers: You Need to Attend More, but Not Necessarily at Every Layer	提出一种转换器变体 MASFormer，使用混合注意跨度来高效处理远程和短程依赖关系。在自然语言建模和生成任务中，MASFormer 表现出与普通变压器相当的性能，但计算成本显著降低（高达 75%）
精度与效率平衡	EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention	提出了一种称为 Efficient ViT 的高速视觉 Transformer，为了提高现有 transformer 模型的速度，研究人员使用了一种三明治布局的新构建块，使用单个内存绑定的 MHSA，在保证通道通信的同时提高内存效率

来源：CDSN，国金证券研究所

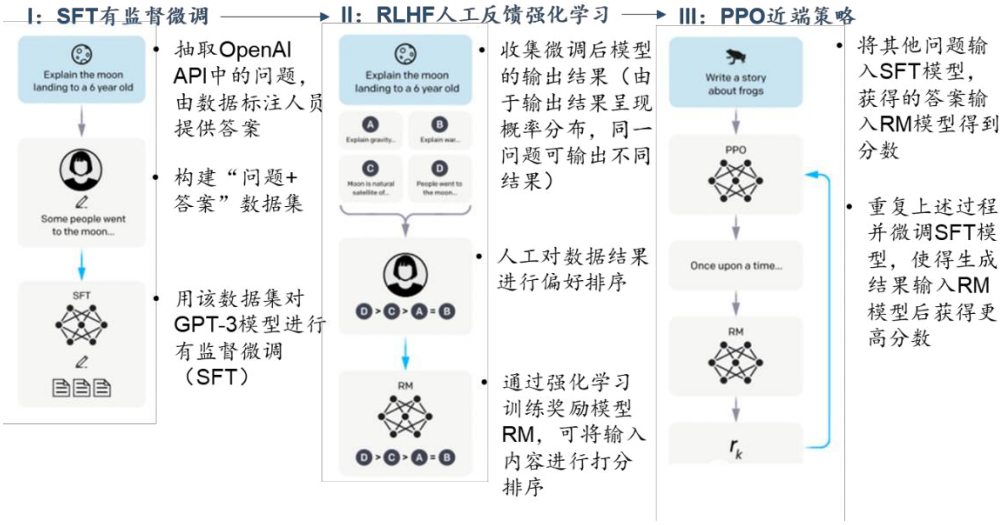
2.2.2 微调方法的改进促进模型性能和落地效率提升

“基础模型+微调”已成为大模型开发范式，通过微调让基础模型针对特定任务类型、应用场景进行二次训练，能够极大提升大模型在实际应用中的智能水平。相较于过去“一场景、一任务、一模型”的训练方式，微调能够是使用更小的数据量、更短的训练时间使模型能够适应下游任务，显著降低了边际落地成本。

目前大模型的微调方法可以分为全量微调（Full Fine-tuning）和参数高效微调（PEFT，Parameter-Efficient Fine-Tuning）两种：

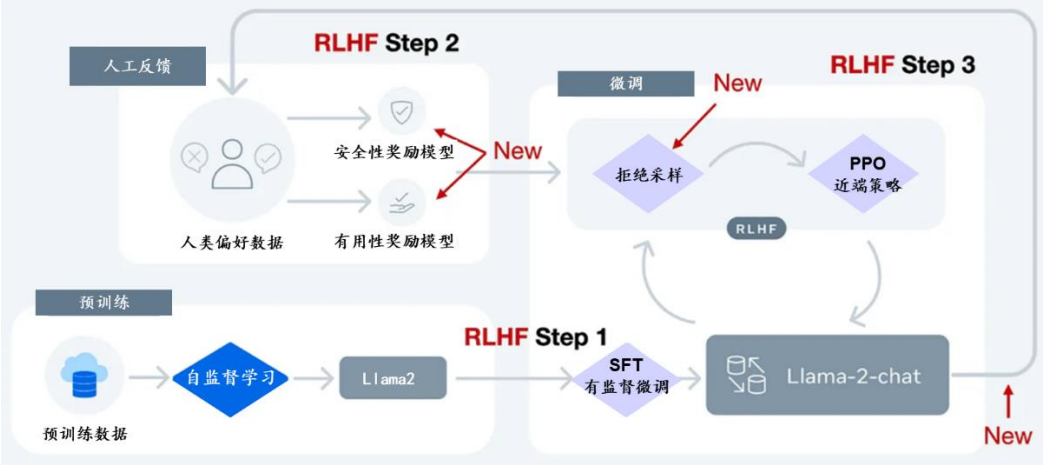
1）全量微调：利用特定任务数据调整预训练模型的所有参数，以充分适应新任务。它依赖大规模计算资源，但能有效利用预训练模型的通用特征。ChatGPT（InstructGPT）使用的基于人类反馈的强化学习微调 RLHF 即为全量微调，通过使用 RLHF 模型输出内容能够更加符合人类语言习惯。23 年 7 月，Meta 旗下的开源模型 Llama-2-chat 对 RHLF 进行了改进，通过创建两个奖励模型、增加额外的拒绝采样步骤，使得生成内容在安全性和有用性方面表现更好。

图表22：InstructGPT 中的 RLHF 技术



来源：《Training language models to follow instructions with human feedback》，国金证券研究所

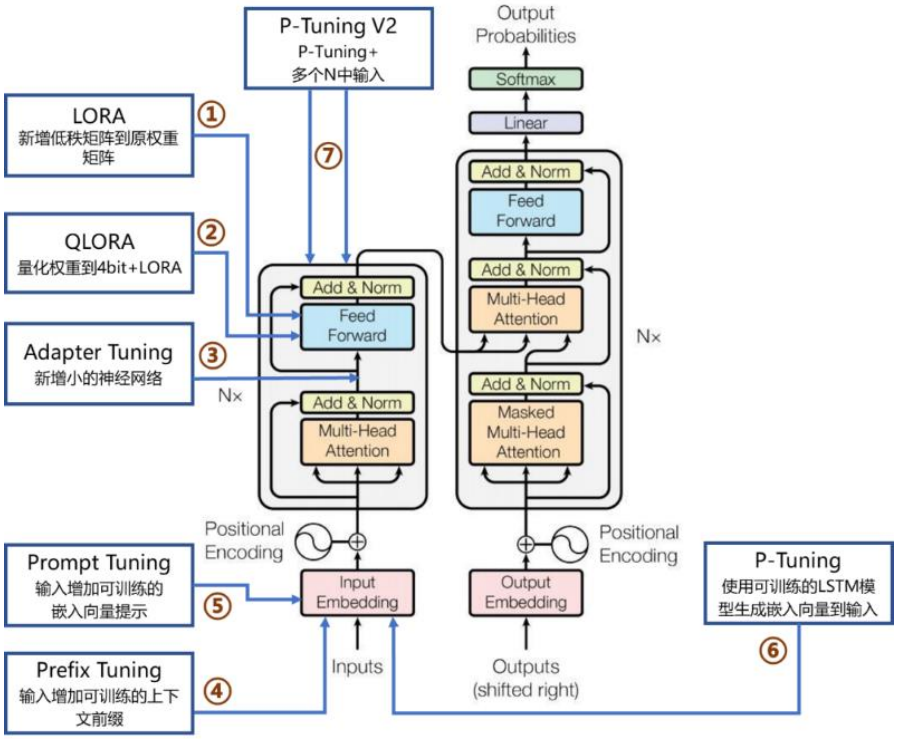
图表23：Llama-2 对 RHLF 的奖励模型进行改进



来源：《Llama 2: Open foundation and fine-tuned chat models》，机器之心公众号，国金证券研究所

2）参数高效微调：旨在通过最小化微调参数数量和计算复杂度，实现高效的迁移学习。它仅更新模型中的部分参数，显著降低训练时间和成本，适用于计算资源有限的情况。常见的 PEFT 技术包括 LoRA、Prefix Tuning、Prompt Tuning、Adapter Tuning 等多种方法。其中 LoRA 是微软推出的低秩自适应技术，它相当于在原有大模型的基础上增加了一个可拆卸的插件，模型主体保持不变，随插随用，轻巧方便，使用 LoRA 时可以节省 33%的 GPU 内存。

图表24：针对 Transformer 架构大模型的 PEFT 微调方法



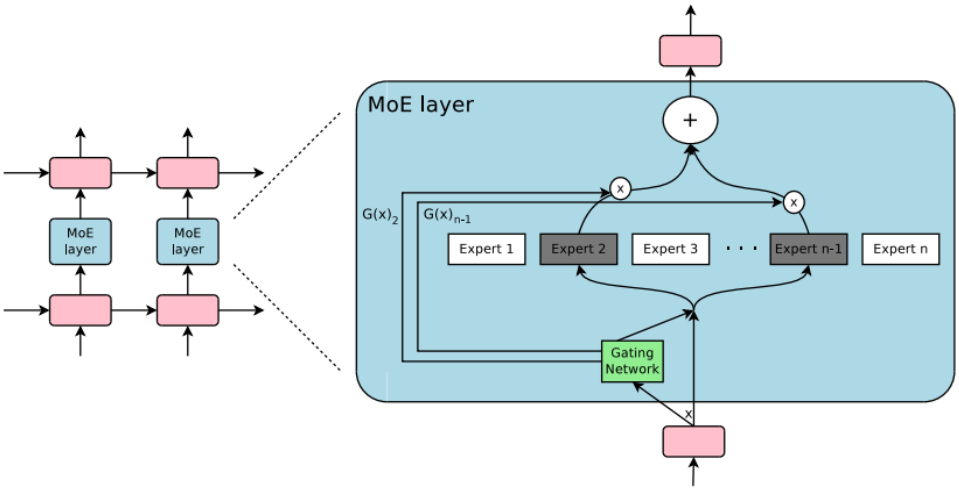
来源：CSDN，国金证券研究所

2.3.3 使用 MoE 进行任务分割，更高效地利用计算资源

AI 大模型的计算架构决定了模型中人工神经网络的各神经元之间相互作用的方式。计算架构可分为稠密结构和稀疏结构 2 种：1）使用稠密结构的大模型在计算时需要激活整个神经网络，算力和内存消耗较大，主要应用于 GPT-3 等早期 AI 大模型中；2）稀疏结构允许系统的某些特定部分单独执行计算，根据输入的特定特征或需求，只有部分参数集合被调用和运行。

稀疏结构的代表是 MoE 混合专家模型，通过将输入数据根据任务类型分割成多个区域，并将每个区域的数据分配一个或多个专家模型。每个专家模型可以专注于处理输入这部分数据，从而提高模型的整体性能。

图表25：MoE 结构中只激活部分网络



来源：《Outrageously Large Neural Network》，国金证券研究所

尽管 MoE 提供了若干显著优势，例如更高效的预训练和与稠密模型相比更快的推理速度，但仍有继续改进的空间：

- 训练挑战：虽然 MoE 能够实现更高效的计算预训练，但它们在微调阶段往往面临泛

化能力不足的问题，长期以来易于引发过拟合现象。

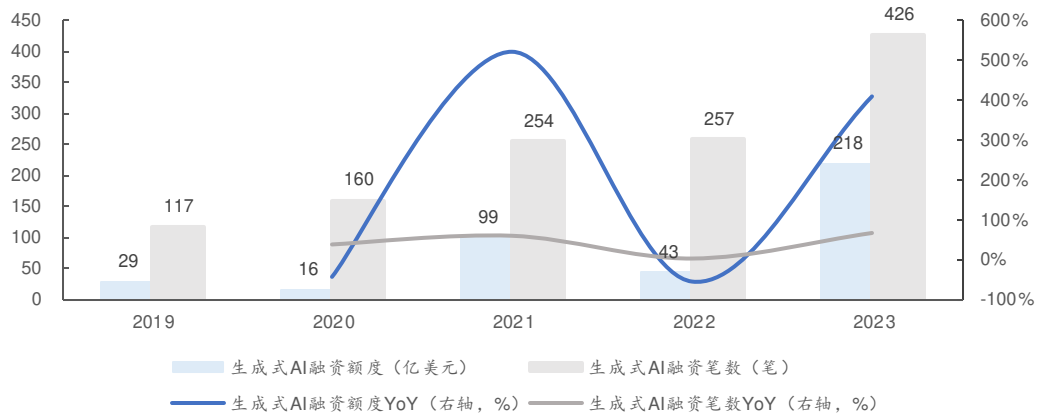
- 推理挑战：MoE 模型虽然可能拥有大量参数，但在推理过程中只使用其中的一部分，这使得它们的推理速度快于具有相同数量参数的稠密模型。然而，这种模型需要将所有参数加载到内存中，因此对内存的需求非常高。

3. 如何商业落地：借力模型开源及 B 端合作，寻找高人工替代率的场景

3.1 开源模型 vs 闭源模型？——Scaling Law 不再 work 之后，找场景或优于做模型

本轮 AI 底层模型创业需求依赖资本密集的人才与算力持续投入。据 BofA GLOBAL RESEARCH，2023 年，全球生成式 AI 公司融资额度高达 218 亿美元，同比 22 年提升 4 倍，超过 19~22 年 4 年融资总额；2023 年全球生成式 AI 公司融资笔数多达 426 笔，同比提升 65.8%。我们认为，融资笔数同比增速大幅低于融资额度说明 2023 年 AI 创业公司平均融资额度较大，可能与 AI 大模型创业公司对人才、算力需求较大所致，变相说明本轮 AI 模型创业相对资本密集，对于持续高额融资的需求较为旺盛。

图表 26：2023 年生成式 AI 融资额度与融资笔数快速提升



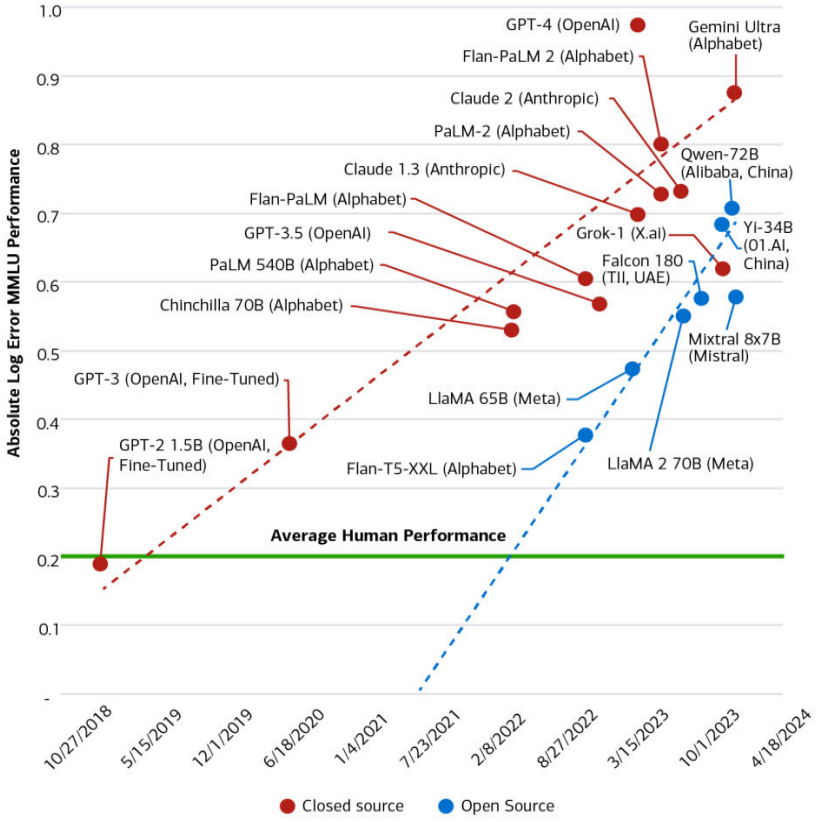
来源：BofA GLOBAL RESEARCH，CB Insights，国金证券研究所

开源模型快速追赶闭源模型，开源模型性能优化速度快于闭源模型。AI 底层模型创业客观上、依托投资人的持续投入，以支撑模型训练对于尖端人才与大规模算力的需求。

- 一方面，性能卓越的爆款应用（killer app）往往需要基于足够强大的模型能力，甚至于会有 MaaS（Model as a Service，模型即服务）说法的出现——“好模型等于好应用”➔每家 AI 公司都应该自研大模型；
- 另一方面，在国内外众多创业者投身底层模型研发的过程中，AI 大模型第一梯队的领跑者也会阶段性地开源已有的较先进的模型，以塑造围绕自身的开发者生态——在 Scaling Law 不再 work 的世界中，开源模型迟早会追平（或无限接近）闭源模型性能➔不必重新造轮子，中长期看找应用场景优于卷大模型。

据 BofA GLOBAL RESEARCH，目前开源模型性能优化速度快于闭源模型，我们认为，目前第一梯队 AI 大模型纷纷进军万亿参数，且不远的将来大模型将逐步逼近十万亿参数收敛值，对于本轮 AI 科技浪潮而言，找场景或优于做模型。

图表27：开源模型性能改善速度快于闭源模型



来源：BofA GLOBAL RESEARCH，国金证券研究所

3.2 如何定义一个好场景？——“幻觉”尚未消除的世界，高人工替代率或为重点

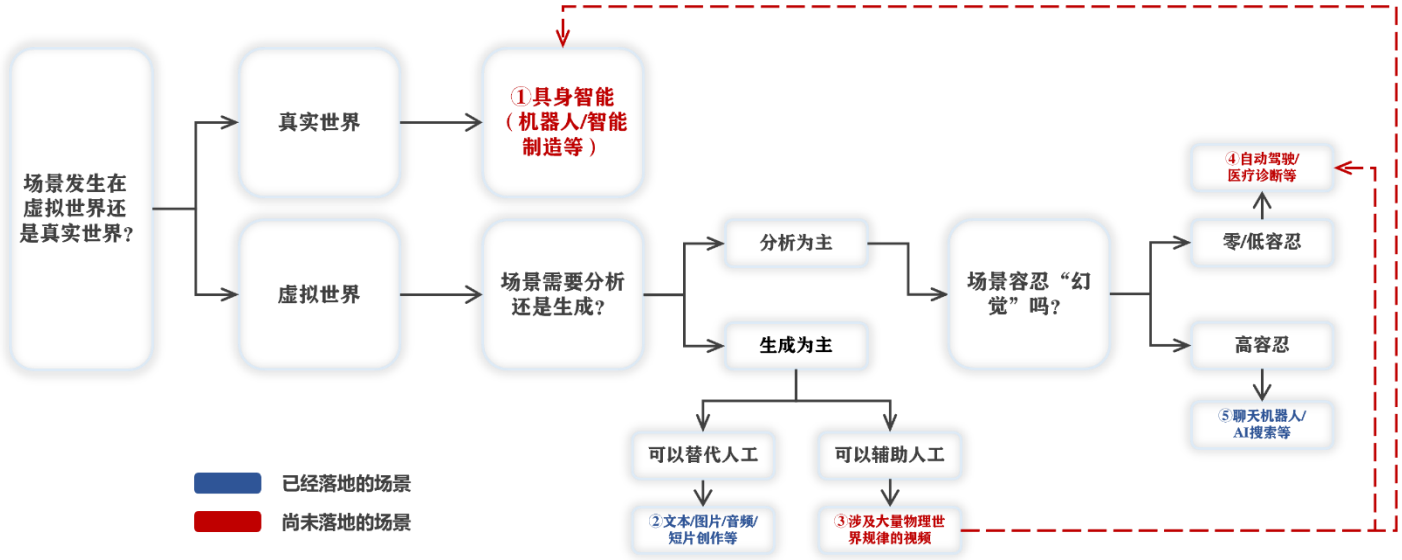
基于未来开源模型性能表现终将追平或接近闭源模型能力这一假设，我们认为以中长期视角来看，找到一个能够将 AI 落地且可以产生商业化收入的场景对于大多数 AI 厂商而言或将成为更优的投入方向。

结合我们上一篇 AI 行业研究报告——《AI 应用落地的商业模式探索》与本篇报告前述章节的讨论，我们认为 AI 落地场景大致可以分为 5 类：

- 第一类-真实世界中基于具身智能的应用场景。
- 第二类-虚拟世界中场景更需要“生成”为主，且现阶段可以替代人工的场景。
- 第三类-虚拟世界中场景更需要“生成”为主，且现阶段不可以替代人工的场景。
- 第四类-虚拟世界中场景更需要“分析”为主，且容错率较低的场景。
- 第五类-虚拟世界中场景更需要“分析”为主，且容错率较高的场景。

已经落地的场景往往对“幻觉”具备一定的容忍度。只有第二类和第五类场景是目前 AI 能够应用落地且可以产生商业化收入的。第二类场景例如生成某种风格类型的小说/插画/音乐以及基于 Sora 等多模态模型得到的短片。第五类场景例如 ChatGPT 或者 Character.ai 等满足效率工具与角色扮演需求的聊天机器人，以及例如 Perplexity 等 AI 辅助搜索。我们认为，以上两类场景之所以能够在现阶段落地的核心原因是它们均能够在一定程度上容忍“幻觉”（Hallucination，指在 AI 生成或反馈结果当中存在的不符合常理的情况），其中，对于第二类场景所对应的文本/图像/音乐/视频创作（错题生成也可以被视作文本创作的一部分）而言，其本身便不存在唯一性的最优解；而对于第五类场景所对应的问答互动与信息总结需求而言，固然存在更优的回答与更有效率的信息归纳方式，但用户对于 AI 偶尔出现不符合常理的反馈仍有一定的宽容度。

图表28：AGI 演进过程中的应用场景分类



来源：国金证券研究所

尚无法落地的场景需要解决“幻觉”所产生的问题。对于尚无法应用落地或至少不能形成商业化收入的第一/三/四类场景而言，我们认为最重要的是要解决“幻觉”所带来的问题。

- 对于第一类场景而言，基于具身智能的机器人置身于真实的物理世界当中，其每一个行为动作都可能会对工厂与居家安全带来风险（比如，一个错误的参数反馈可能导致生产事故，一个错误的指令理解可能伤害到居家住户），因此天然对于“幻觉”的容忍度很低。
- 对于第三类场景而言，尽管多数依赖生成式 AI 的创意工作都已经得到落地，但我们依然可以在 Sora 对外披露的视频中看到不符合物理世界常识的画面出现，涉及大量物理世界规律的长视频制作目前依然无法完全取代人工。
- 对于第四类场景而言，部分直接涉及人类生命安全的领域，如自动驾驶与医疗诊断也天然对“幻觉”具有较低的容忍度。

综合前述，我们认为，假如第三类场景中的“幻觉”得以解决或至少控制在足够低的范围内，将有助于反哺第一与第四类场景进行数据训练，从而加速 AI 的落地进展。

3.3 如何处理“幻觉”？——Scaling Law 信仰派 vs 引入知识图谱改良派

在处理模型幻觉、进而实现 AGI 的路径方面，学界主要存在着两派声音——基于连接主义的“Scaling Law 信仰派”与基于符号主义的“引入知识图谱改良派”：

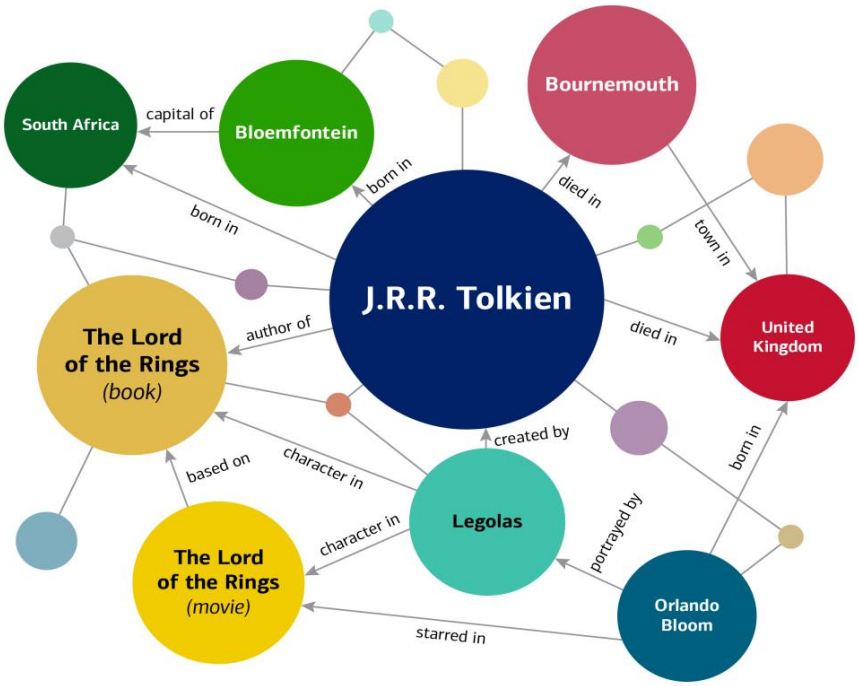
- 连接主义（Connectionism）：又称为神经网络或并行分布处理，是一种模仿人脑神经网络结构和功能的人工智能方法。它的核心思想是通过大量简单的、相互连接的处理单元（类似于神经元）来实现复杂的智能行为。这些处理单元之间的连接强度代表了信息的权重，而智能则体现在这些单元如何通过学习和调整连接强度来处理信息。用一个简单的比喻，连接主义就像是一张由许多节点（神经元）组成的大网。每个节点都可以接收和发送信号，而节点之间的连接则决定了信号如何传递。当这张网接收到输入信号时，它会通过调整节点之间的连接强度来学习新的模式和任务，就像人脑学习新知识一样。连接主义认为，现阶段的“幻觉”只是模型参数与训练数据集的规模未达到理想情况导致的，Scaling Law 将会持续改善模型效果直至“消除”“幻觉”。
- 符号主义（Symbolism），也称为逻辑主义或规则主义，是一种基于符号处理的人工智能方法。它的核心思想是认为智能行为可以通过对符号的操作和处理来实现。这些符号代表了现实世界中的对象、概念或事件，而智能则体现在如何通过逻辑规则对这些符号进行有效的组合、推理和转换。举个例子，符号主义就像是我们使用的语言和数学公式。我们通过文字和公式来表达思想和解决问题，而符号主义 AI 则通过预设的规则和逻辑来操作这些符号，从而实现智能行为。比如以 Yann LeCun 为代表的 Meta、Google、Stanford 等科学家认为目前的生成式 AI 模型没有真正理解内容。

图表29：连接主义 VS 符号主义

来源：国金证券研究所

“改良派”认为，知识图谱可能用以减轻“幻觉”问题。知识图谱是一种存储信息并展示相关信息源之间关系的方法。知识图谱具有一个集中准确的知识来源，并且能够将不同格式的信息进行结构化的组织。AI 大模型有时会“很有信心”提供一些不准确的信息。知识图谱从多个来源摄取大量事实信息，并在它们之间建立联系，将知识图谱与大模型整合，将促使大模型内部的概念之间形成逻辑连接。理想状况下，AI 大模型可以利用包括结构化和非结构化数据在内的各种信息来源，生成更准确的输出。知识图谱不像 AI 大模型那样的概率引擎，其基于一个准确的知识中心进行推理和解释，进而也可以减少 AI 大模型训练对大量标记数据集的需求。

图表30：知识图谱通过机器学习和自然语言处理来构建节点、边和标签的全面视图

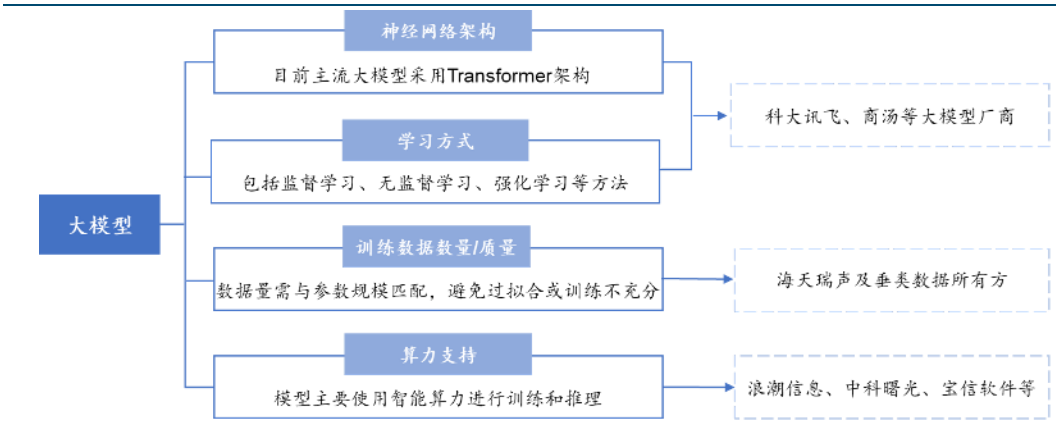


来源：BofA GLOBAL RESEARCH，国金证券研究所

4. 投资建议

目前大模型正处在 Emerging AGI 阶段，多模态融合是现阶段发展的重点方向。在模型性能提升方面，无论是继续沿 Scaling Law 推进，还是探索神经网络骨干架构和细分算法的创新，均需要大模型厂商与数据工程类、算力支持类公司合作推进。

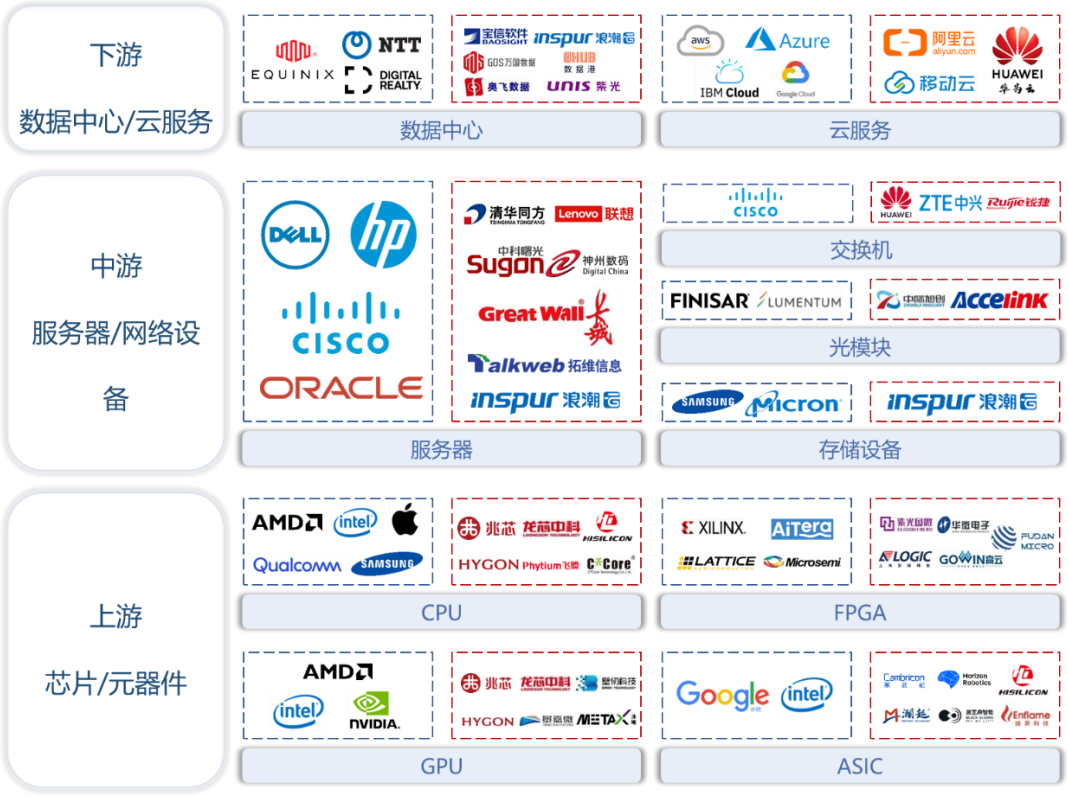
图表31：大模型向 AGI 演进，模型训练产业链有望持续收益



来源：国金证券研究所

- **大模型厂商：**国内头部 AI 大模型上市公司包括科大讯飞、商汤等。其中科大讯飞于 2023 年 5 月发布讯飞星火大模型，至 24 年 1 月模型已升级至 V3.5 版本，根据公司测评，在七大核心能力上都获得了全面提升，在数学、语言理解和语音交互能力上还超越了 GPT-4 Turbo。商汤于 23 年 4 月发布日日新系列大模型，24 年 2 月模型更新至 V4.0 版本，其中日日新商量大语言模型支持 128K 语境窗口长度，综合整体评测成绩水平比肩 GPT-4。
- **数据工程厂商及拥有丰富垂类数据的公司：**海天瑞声是我国语音类基础数据服务领域头部企业，目前公司已推出 DOTS-LLM 大模型服务平台，平台包括数据采标、数据管理、模型训练和模型评测四大功能，旨在为千行百业数字化转型赋能。此外，在微调环节需要使用垂类行业数据，以提升模型在具体应用中的性能，拥有丰富行业数据积累的公司也有望收益。
- **算力产业链：**上游包括芯片和元器件，中游包括服务器和网络设备，下游包括 IDC 以及云服务等。其中，浪潮信息是算力系统供应商，支持多元异构算力、可适配多种架构的 AI 加速芯片；中科曙光积极建设“全国一体化算力服务平台”，致力于链接遍布各地各类算力中心；宝信软件多年专注于自主研发工业互联网平台宝联登 xln³ Plat 和 AI 中台。我们在《算力深度报告一：算力研究框架-产业链全梳理》进行了详细投资标的梳理。

图表32：算力产业图谱



来源：中国信通院，国金证券研究所
说明：途中蓝色虚线方框内为境外公司，红色虚线方框内为中国公司

对于 AI 下游应用厂商而言，选择基于开源模型开发，或者与海内外顶级模型厂商进行合作，即能够实现大模型技术赋能已有业务。因为，选择合适的落地场景更为重要。目前虽然大模型在实际应用中仍存在“幻觉”问题，但已经在教育、企业服务、办公、金融等众多领域落地应用，建议持续关注。

图表33：建议关注 AI 赋能细分场景的龙头企业

赋能行业	受益公司	基本情况
AI+教育	科大讯飞	· 公司于 23 年 5 月 6 日发布星火认知大模型，并同步推出搭载大模型能力的 AI 学习机 T20 Pro，支持中英作文类人批改、数学类人互动辅学、英语类人口语对话等功能。讯飞 AI 硬件在 2023“双 11”全周期销售额同比增长 126%。
	竞业达	· 公司发布星空教育大模型，与学校私有化个性化教学数据深度融合，通过课堂教学、实验教学、考试评价、质量保障、学工就业等关键环节的全方位产品布局为高校赋能，打造 AI 助教、AI 导师、AIGC 数字老师等新型教育教学模式，助力教育教学全流程数字化转型。
	世纪天鸿	· 23 年 4 月上线教师端助教产品小鸿助教，至 24 年 2 月已更新至 4.0 版本，产品通过对话的方式，在包括教案生成、作文批改、文章写作、教学活动策划、读书笔记、文本润色、PPT 大纲、思维导图设计、教师评语编写以及进行中英互译等多种应用场景帮助老师提升工作效率。
AI+企业服务	泛微网络	· 公司推出了基于大语言模型的小 e 智能办公助手，在逐步完成智能问答、智能办公、智能工具等应用场景的基础上，通过“RPA+Agent+LLM”的系统架构，为用户提供业务自动化智能工具。
	致远互联	· 公司于 23 年 9 月正式发布 AI-COP 大模型框架，推出一站式 AI 协同管理与运营服务平台，并联合方寸智能发布国内首个公文领域大模型。同时，致远互联结合大模型+知识图谱能力推出一系列应用，包含工作智能助手、流程智能助手、智能领域应用、低代码搭建智能助手、企业级 AI 工作站五大协同智能应用场景，助力企业提质降本增效。
	用友网络	· 公司发布企业服务大模型 YonGPT，覆盖企业财务、人力和业务，能够为企业带来智能化的业务运营、自然化的人机交互、智慧化的知识生成、语义化的应用生成。
	金蝶国际	· 公司发布财务大模型苍穹 GPT，为企业利用大模型能力提供了完整的工程技术方案，广泛接入百度文心一言、微软 OpenAI 等通用大模型能力。综合平衡企业算力成本、训练成本、模型能力、应用价值等要素，设置百亿级参数，经过专业训练和精调。
AI+办公	金山办公	· 23 年公司发布了基于大语言模型的智能办公应用 WPS AI，锚定 AIGC（内容创作）、Copilot（智慧助理）、Insight（知识洞察）三个战略方向发展，将 AI 在国内办公软件领域率先落地的成果带给用户。

		WPS AI 已于 23 年第四季度完成备案并正式开启公测。
	福昕软件	· 福昕已于 2023 年 4 月将旗下海外版云产品集成 ChatGPT，可实现文本摘要（局部）/文本润色/文本翻译/文档摘要（全局）/分析文档以问答等 AIGC 功能。集成 ChatGPT 有助于持续推进福昕的订阅制转型与产品增值提价。
	万兴科技	· 万兴于 2022 年底至 2023 全年对旗下各条产品线的多款拳头产品进行 AI 功能更新，发布全新 AI-Native 产品 Kwicut/万兴播爆/万兴智演等，且至少已有 3 款产品明确接入 OpenAI GPT 系列模型。加入 AI 功能之后，多款产品月活/付费率/收入数据有所增长。
	彩讯股份	· 23 年 5 月发布 AI 邮箱 Demo，23 年四季度进行邀请测试，预计相关产品会在 24 年三季度有商业化业绩体现。此外，公司还在探索 AI 云盘产品研发。
AI+金融	同花顺	· 目前公司自研的问财 HithinkGPT 预训练金融语料达万亿 tokens，预计应用于投顾、投研、客服、代码生成、法律咨询等领域。目前同花顺 AI 开放平台目前可面向客户提供数字虚拟人、短视频生成、文章生成、智能金融问答、智能语音、智能客服机器人、智能质检机器人、会议转写系统、智慧政务平台、智能医疗辅助系统等多项 AI 产品及服务。
	恒生电子	· 2023 年公司发布金融行业大模型 LightGPT、金融智能助手光子及智能投研平台 WarrenQ。公司以 Light 技术平台为底座，进一步发展分布式低延时平台、敏捷业务交付平台、高性能数据编织平台。
	宇信科技	· 公司已推开发助手 CodePal、金融数据安全分级分类助手 DataSherpa、AI+信贷助手（客户尽调）、AI+营销助手、大模型应用开发平台。其中 AI+信贷助手使用大语言模型辅助了解对公信贷客户，可以提供客户资料收集、图文识别、财报数据抽取、智能核验、风险分析、尽调报告生成等各项功能。
	凌志软件	· 公司使用 ChatGPT 构建投行智能解决方案，已完成 GPT-3.5 对接，可进行基本文档和招股书财务分析的初稿生成，有望进一步带动国内业务加速增长。
AI+工业软件	中控技术	· 24 年 3 月发布人形机器人整机“领航者 1 号”，完全为国内自主研发，身高 1.5 米，体重 50 千克，硬件端包括新型行星减速器、轻量化仿人机械臂和多自由度灵巧手，其中多自由度灵巧手有 15 个手指关节，6 个主动自由度，指尖力 10N，单手重量 600g，关节速度 150 度/秒。
	索辰科技	· 24 年 2 月成立机器人事业部，开发针对机器人行业的专业软件和解决方案，设计软件涵盖并联机器人本体的完整研发过程，还能提供专门为机器人设计的仿真解决方案。
	中望软件	· 公司计划在下一阶段产品版本中实现对于机器学习优化或命令提效等功能。
AI+安防	海康威视	· 公司发布“观澜”行业大模型，包括 X 光大模型、雷视大模型、音频大模型，以及用震动管线感知手段训练出的光纤大模型，可应用于交通、电力、钢铁、煤炭、安检等场景。
	大华股份	· 公司发布“星汉”大模型，通过融合图像、点云、文本、语音等多模态数据，大幅提升了视觉解析能力，可赋能城市高效治理、运行自治、安全体系升级、生态协同治理等领域。
	萤石网络	· 24 年 3 月公司举办春季新品发布会，推出使用云端大模型的三摄全自动人脸视频锁 DL60FVX Pro，包裹检测、跌倒检测、儿童检测等算法精准度得到显著提升，为用户提供更可靠和精准的安全保障。
AI+网络安全	奇安信	· 公司于 23 年 8 月发布 Q-GPT 安全机器人，重点解决长期困扰政企安全客户的三大问题，即告警疲劳、专家稀缺和效率瓶颈，后续还会持续升级迭代。
	安恒信息	· 公司的恒脑 • 安全垂域大模型已顺利通过华为 AI 框架昇思 MindSpore 相互兼容性测试认证，基于昇腾联合开发的大模型一体机已完成适配。目前很多客户已经开始部署试用。
	永信至诚	· 公司于 24 年 2 月发布“春秋”靶场构建大模型、“春秋”安全竞赛大模型和“春秋”人才测评大模型三款产品，为政企用户提供更智能、更高效、更便捷的网络安全解决方案。
AI+医疗	卫宁健康	· 发布 WinGPT 行业大模型支持 7 大类基础任务与 20 多项子任务，应用场景包括互联网问诊、医疗报告生成、单病种癌症智能辅助诊断等。
	创业慧康	· 与浙大共建慧康-启真大模型，赋能临床决策支持系统 CDSS、智能就医助手、专科电子病历等场景。
	嘉和美康	· 参与 AI 医疗影像公司安德医智破产重组，并推出 AI 电子病历质控产品

来源：ifind，各公司公众号，各公司公告，国金证券研究所

5. 风险提示

- 底层大模型迭代发展不及预期

若底层大模型迭代发展不及预期，可能会影响 AI 应用落地的深度，使其难以在金融、教育、游戏等领域进行更加深入的应用。若底层大模型的蒸馏剪枝发展不及预期，可能会使其难以在边缘硬件上充分发挥性能。

- 国际关系风险

若出于国际关系原因，OpenAI 等海外大模型的调用或其他软硬件的进口受到影响，有可能使得国内 AI 应用的发展不及预期。

- 应用落地不及预期

若相关应用公司不能找到人工智能算法较好的商业应用落地场景，或相关场景客户没有较强的付费意愿，可能算法应用落地会不及预期。

- 行业竞争加剧风险

若相关企业加快技术迭代和应用布局，整体行业竞争程度加剧，将会对行业内已有企业的业绩增长产生威胁。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的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，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究