2023 年 04 月 16 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：王倩雯（执业 S1130522080001） 分析师：孟灿（执业 S1130522050001）
wangqianwen@gjzq.com.cn mengcan@gjzq.com.cn

LLaMA 等开源模型凸显先进算法及行业数据的重要性

投资逻辑

自 2017 年 Transformer 发布以来，大语言模型经历了由开源到逐步闭源的转变，头部公司先进模型的壁垒逐步形成。目前 OpenAI、Google 等领先的头部 AI 大厂对于先进模型大多采用部分开源或仅开放使用的模式，以此构建技术护城河。然而，将 AI 大模型直接应用于垂直行业，存在通用能力过剩、行业专业知识储备不足、推理过程消耗算力过高等问题。基于开源模型进行垂类模型开发可兼顾开发成本和数据安全，尤其是对于党政军、金融、电网、先进制造等数据敏感性较高的行业而言。

Meta 旗下 LLaMA 大模型的开源或能为垂类模型落地提供预训练模型底座。LLaMA 基于通用领域的开源数据集进行训练，训练数据涵盖 40 种语言，包含约 1.4 万亿 Tokens。尽管 LLaMA 模型参数量较小，但性能丝毫不逊色于 PaLM、GPT-3 等大语言模型。并且较小的参数规模显著降低了 LLaMA 模型的落地部署和二次开发难度。

LLaMA 作为完全开源的领先模型，具备高度的灵活性、可配置性和泛化能力，可以作为垂类 AI 模型的通用基座。基于 LLaMA，垂类 AI 开发者可以根据其行业特点、应用行业数据定制开发相应的“行业发行版 AI 模型”。 LLaMA 模型一经发布就对外完全开源，吸引了广大 AI 开发者和研究者。目前，用户可在全球知名 AI 模型开源社区 Hugging face 中获取 LLaMA 的模型权重与训练代码。能够自由下载并使用 LLaMA 模型，既可以将其部署至设备直接进行推理，也可以基于 LLaMA 进行研究与二次开发。

我们测算了模型在迁移学习阶段的训练算力成本，在模型微调阶段，由于训练量级较小，仅为万级，相关的算力成本相比之下可忽略不计。例如，斯坦福大学于 2023 年 3 月对外发布 Alpaca，这是一个基于 LLaMA-7B 基座，应用 5.2 万指令对模型微调训练得到的对话类语言模型，该模型基于 8 块 A100 微调，微调时长 3 小时，算力成本不超过 300 元。

在推理阶段，根据我们的初步测算，由 8 块 A100 组成的 AI 服务器可为规模达 2,000 人的中大型企业提供 AI 服务，离线部署方案每年的推理算力成本约为 33.2 万元，若采用云计算方案则每年需花费约 66 万元算力成本。基于上述推理成本分析，推理成本并不高昂，绝大多数中型以上企业足以负担，为各领域垂类模型落地提供了极为广阔的市场空间。

投资建议

LLaMA 等优质开源模型的推出极大加速了下游行业 AI 应用开发效率。基于“通用基座+迁移学习+微调”的垂类 AI 模型开发范式或将成为主流，优质的行业数据资源成为影响模型性能的关键。

在此趋势之下，我们看好两类企业：1）拥有开发先进大模型能力的企业。这类企业在先进模型逐步走向闭源的趋势下，有望保持算法优势，如商汤科技、科大讯飞等。2）拥有丰富行业数据的头部公司。这类企业有望基于稀缺的行业数据以及开源模型，开发出可用性更强的垂类模型。如东方财富、同花顺、恒生电子等。

风险提示

海外基础软硬件使用受限；骨干网络创新放缓；应用落地不及预期

内容目录

1. 头部领先模型走向闭源，垂类模型开发呼唤开源 3
1.1 头部公司大模型逐步走向闭源 3
1.2 为什么垂类 AI 开发呼唤开源？ 5
2. LLaMA 在通用开源模型中性能领先 6
2.1 LLaMA 具有参数量低、性能优异、完全开源等特点 6
2.2 小参数量可降低垂类模型开发及部署难度 7
2.3 LLaMA 提供通用开发基座，泛化能力更强 8
3. “通用模型+迁移学习+微调”有望成为开发新范式，数据是重要壁垒 9
3.1 使用“迁移学习”向模型注入新知，开发难度相对较低 9
3.2 叠加先进算法微调，进一步释放模型性能 10
3.3 算力消耗并非海量，成本效益匹配 11
3.4 赋予垂类 AI 开发者离线部署能力和离线迭代能力 13
4. 投资建议 13
5. 风险提示 14

图表目录

图表 1: 头部公司大模型从完全开源逐步走向部分开源 3
图表 2: OpenAI 的系列模型开始向闭源发展 4
图表 3: 大多数已完全开源模型准确率低于非开源模型 5
图表 4: 基于开源模型训练垂类模型是较为理想的开发方式 5
图表 5: LLaMA 参数量相比领先语言模型较小 6
图表 6: LLaMA 基于海量通用领域的开源数据进行训练 6
图表 7: LLaMA 通用领域性能处于世界领先行列 7
图表 8: LLaMA 兼具高性能和易部署的特点 7
图表 9: LLaMA 有望带动海量垂类模型落地 8
图表 10: 微调无法向模型内部注入新知识 9
图表 11: 迁移学习能向模型内部注入垂直领域新知识 10
图表 12: “通用基座+迁移学习+微调”有望成为垂类模型开发新范式 10
图表 13: LLaMA 模型基座算力消耗 11
图表 14: 采用云计算的垂类模型训练费用估算 11
图表 15: 采用自建算力的垂类模型训练费用估算 11
图表 16: 垂类模型推理算力成本估算 12

1. 头部领先模型走向闭源，垂类模型开发呼唤开源

1.1 头部公司大模型逐步走向闭源

自 2017 年 Transformer 发布以来，大语言模型经历了由开源到逐步闭源的转变，头部公司先进模型的壁垒逐步形成。

我们将 AI 模型的开源程度划分为以下四类：

- 完全开源：以论文形式对外发布 AI 模型的研究细节，研究者可以下载 AI 模型并离线部署。
- 部分开源：仅以论文形式对外发布 AI 模型的研究细节，研究者可以依照论文较为简单地进行模型复现。
- 仅开放使用：不对外公布任何技术细节，AI 模型仅以 API 或自有产品的方式提供给 B 端和 C 端用户。
- 完全闭源：不对外公布任何技术细节，AI 模型仅以自有产品的方式提供给 C 端用户。

图表1：头部公司大模型从完全开源逐步走向部分开源

机构	大模型	模型开源	发布完整论文	API	开源模式	应用领域	发布时间
OpenAI	GPT	✓	✓		完全开源	自然语言生成	2018.6
	GPT-2	✓	✓		完全开源	自然语言生成	2019.2
	GPT-3		✓	✓	部分开源	自然语言生成	2020.3
	GPT-3.5		✓	✓	部分开源	自然语言生成	2022.11
	DALLE-2		✓	✓	部分开源	图像生成	2022.4
	Whisper	✓	✓	✓	完全开源	语音识别翻译	2022.9
	GPT-4			✓	仅开放使用	多模态	2023.3
Google	BERT	✓	✓		完全开源	自然语言处理	2018.10
	T5	✓	✓		完全开源	自然语言处理	2019.10
	LaMDA	✓	✓		完全开源	自然语言生成	2021.5
	PaLM		✓	✓	部分开源	自然语言生成	2022.4
	Imagen	✓	✓		完全开源	图像生成	2022.5
	PaLI		✓		部分开源	多模态	2022.9
	Muse		✓		部分开源	图像生成	2023.1
	PaLM-E		✓		部分开源	多模态	2023.3
Meta	LLaMA	✓	✓		完全开源	多模态	2023.2
	OPT	✓	✓		完全开源	自然语言处理	2022.5
	Make-A-Video		✓		部分开源	视频生成	2022.9
	Segment Anything	✓	✓		完全开源	图像分割	2023.4
百度	PLATO		✓	✓	部分开源	自然语言生成	2019.10
	ERNIE 3.0	✓	✓	✓	完全开源	自然语言理解	2021.7
	ERNIE-ViLG		✓	✓	部分开源	多模态	2022.1
	文心一言			✓	仅开放使用	多模态	2023.3
Anthropic	Claude		✓	✓	部分开源	自然语言生成	2022.12
Stablilty AI	Stable Diffusion	✓	✓	✓	完全开源	图像生成	2022.9

来源：OpenAI 官网，Google AI 官网，百度文心官网，GitHub，Anthropic 官网，Stability AI 官网，国金证券研究所

对于 AI 算法公司，选择对外开源有助于行业技术进步和自身生态构建，是学界、早期业

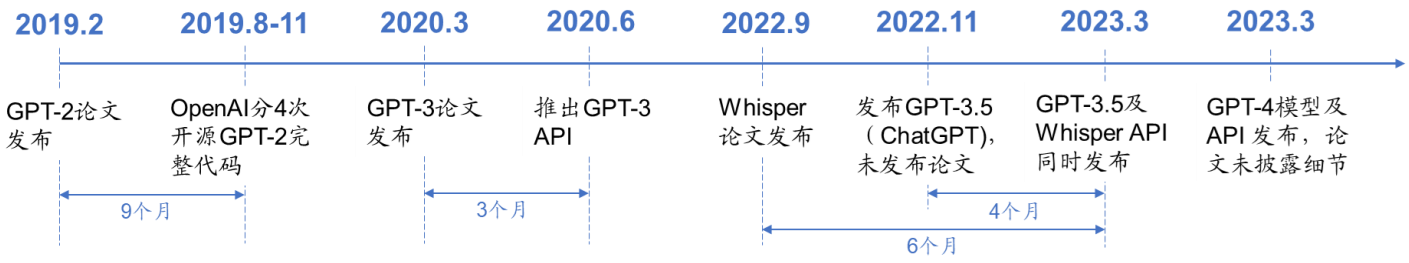
界以及部分 AI 初创企业的选择：

- 行业技术发展：Google 于 2017 年提出 Transformer，采用完全开源模式，凭借其性能优势统一了深度学习底层算法，为后续的大模型发展奠定技术基础。2018 年 BERT、GPT-1 等生成式语言模型陆续发布，也采用完全开源模式，带动 AI 行业高速发展。
- 生态构建：部分初创 AI 公司选择模型开源更多是出于自身生态建设的考虑。以 Stability AI 为例，公司成立于 2020 年，于 2022 年 9 月开源文生图模型 Stable Diffusion，成为目前 AI 绘画赛道的佼佼者。目前 Stability AI 已经在全球积累了超过 14 万名开发人员和 7 个研究中心社区，各渠道累计日活跃用户数超过 1,000 万，日益成熟的生态建设是推升公司估值的主要驱动力之一。

目前 OpenAI、Google、Meta 等领先的头部 AI 大厂对于先进模型大多采用部分开源模式或仅开放使用。以 OpenAI 为例，我们可以管窥海外头部 AI 厂商正在逐渐由开源走向闭源：

- 2019 年及之前，OpenAI 以完全开源为主。OpenAI 于 2018 年发布第一代生成式预训练模型 GPT-1 并对外完全开源；2019 年 2 月 GPT-2 论文发表，在模型规模和 Zero-shot 表现上提升较为明显，模型代码于同年 8 月开始分 4 批陆续对外开源。
- 2019 年 OpenAI 成立盈利子公司 OpenAI LP，开始向盈利公司转变。2020 年 OpenAI 发布 GPT-3，并在论文中较为详细地介绍了模型训练情况，此外用户还可以通过 API 的方式调用模型资源，属于对外部分开源。GPT-3 的发布加快了 AI 落地进程，此后 OpenAI 逐渐向闭源转变。
- 未来 OpenAI 可能采用仅开放使用模式。OpenAI 于 2022 年 11 月发布 ChatGPT，虽然官方未发布模型的具体论文，但目前 AI 开发者仍能从相关论文中获取技术路线信息，显著促进了行业技术的整体进步。2023 年 3 月，OpenAI 同步开放了 GPT-3.5 及语音识别翻译 Whisper 模型的 API。3 月 14 日，GPT-4 发布，目前也处于仅开放使用状态，尚未向外部公布任何技术细节。

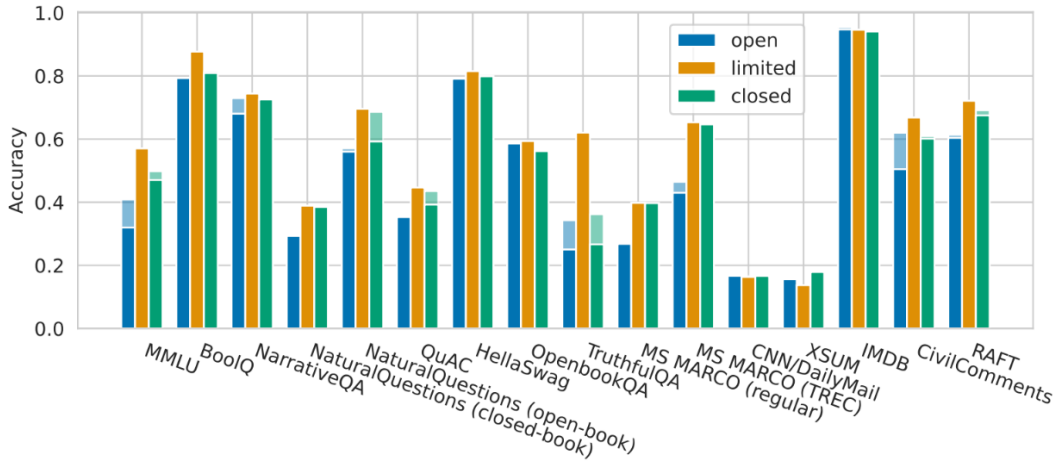
图表2：OpenAI 的系列模型开始向闭源发展



来源：OpenAI 官网,国金证券研究所

选择部分开源或闭源的模式有助于维持 AI 厂商技术优势。根据 Percy Liang 等人于 2022 年 11 月的研究，非开源模型性能优于开源模型：在 16 项核心语言类任务中，开源模型在 13 项任务中表现不及部分开源或闭源模型。AI 大厂或延续部分开源或闭源模式，以此构建技术护城河。

图表3：大多数已完全开源模型准确率低于非开源模型



来源：《Holistic Evaluation of Language Models》(Percy Liang 等)，国金证券研究所
说明：文章统计了 30 个主流语言模型的情况，图中淡色柱状图为同类模型中准确度最高值，深色柱状图为整体精确度水平

1.2 为什么垂类 AI 开发呼唤开源？

AI 大模型能够赋能下游垂直行业，可实现产品性能升级、用户体验感提升、企业降本增效等。但将 AI 大模型直接应用于垂直行业，存在通用能力过剩、行业专业知识储备不足、推理过程消耗算力过高等问题。因此，根据细分行业需求训练相应的垂类模型成为 AI 技术落地的必要环节。

对下游厂商而言，训练或获取 AI 垂类模型的主要方式有 4 种：1）自己从头训练垂类模型；2）调用 AI 厂商 API；3）接受 AI 厂商离线部署模型；4）基于开源模型开发。其中基于开源模型进行开发可兼顾开发成本和数据安全，是较为理想的垂类模型开发方式，尤其是对于党政军、金融、电网、先进制造等数据敏感性较高的行业而言。

图表4：基于开源模型训练垂类模型是较为理想的开发方式

	成本效益问题	数据安全问题	场景应用问题
从头自行训练	• 训练难度大、成本极高	• 不存在数据安全问题	• 可定制应用场景
调用 AI 厂商 API	• API 毛利率或达 95%，成本较高	• 企业内部数据将与外部模型相连	• 在线部署，要求网络稳定
接受 AI 厂商离线部署	• 需要使用者具备一定自研能力，仍存在成本效益问题	• 仍存在数据安全问题	• 产品迭代周期长，或影响后续研发
基于开源模型开发	• 技术要求和成本门槛较低	• 数据安全有保障	• 自主性强，可自行定制模型能力

来源：第四范式公众号，数字时氪公众号，甲子光年公众号，国金证券研究所

- 自行从头训练模型：从技术角度看，绝大部分下游行业公司不具备独立训练 AI 大模型的技术能力、数据储备及算力储备。从成本效益角度看，垂类 AI 模型应用面较窄，难以覆盖前期研发投入。
- 调用 API：从成本效益角度看，AI 公司开放 API 接口的收益可以参考 SaaS 毛利率及云计算设施毛利率，分别约为 90%及 50%-70%，则 API 毛利率或可超过 95%，AI 厂商高毛利经营模式将推升下游公司使用成本。此外，部分企业内部数据、用户资料等敏感数据与外部模型相连存在数据安全风险。从应用落地角度看，使用 API 需要用户网络环境稳定，而部分工业软件或企业软件有保密要求，需要离线运行。
- 使用 AI 厂商产品离线部署：部分 AI 厂商可提供定制化开发服务，但对于下游使用者而言，存在产品自主性弱、迭代周期长、后续研发不确定性强等问题。此外，该模式仍存在性价比较低和数据安全风险。
- 基于开源模型开发：相较于上述三种模式，使用开源模型对下游用户的技术水平和研发投入要求较低；无需向 AI 厂商分享数据，不存在信息安全问题；此外，使用者还可根据业务需求自行增减功能或进行模型迭代。

综上而言，使用开源模型进行垂类大模型开发是性价比较高的选择。在 AI 大厂倾向于部分开源或闭源的背景下，Meta 旗下 LLaMA 大模型的开源或能为垂类模型落地提供预训

练模型底座。

2. LLaMA 在通用开源模型中性能领先

2.1 LLaMA 具有参数量低、性能优异、完全开源等特点

2023 年 2 月 24 日，Meta AI 对外发布生成式预训练语言模型 LLaMA。LLaMA 采用 Transformer Decoder-Decoder 架构，这是目前语言领域领先模型中的主流架构，典型的代表有 OpenAI 的 GPT-4、Google 的 PaLM-E。该架构的优点在于可以很好地处理长文本序列并从中捕捉上下文间存在的特征与联系，从而更容易完成各类生成式语言任务。

LLaMA 根据参数规模的不同，可分为四个子模型，其参数规模分别为 70 亿、130 亿、330 亿和 650 亿，也远小于其他主流生成式预训练模型。

图表5：LLaMA 参数量相比领先语言模型较小

模型名称	发布者	发布时间	参数量
LLaMA	Meta	2023.2	• 70 亿、130 亿、330 亿和 650 亿
PaLM	Google	2022.4	• 5,400 亿
GPT3.5	OpenAI	2022.11	• 1,750 亿
PaLM-E	Google	2023.3	• 5,620 亿
百度文心一言	百度	2023.3	• 千亿级

来源：Meta，OpenAI，Google，百度，国金证券研究所

LLaMA 基于通用领域的开源数据集进行训练，没有使用任何专有数据集，训练数据涵盖英语、中文等 40 种语言，总量约为 1.4 万亿 Tokens。LLaMA 的训练数据量相比主流领先模型有所突破，比如 Google 的 PaLM 采用 7,800 亿 Tokens 的数据进行模型训练，而 LLaMA 几乎是其两倍。

图表6：LLaMA 基于海量通用领域的开源数据进行训练

数据集名称	占比	数据集大小	数据集介绍
CommonCrawl	67%	3.3TB	• 开源互联网网页数据集，包含 30 多亿互联网网页数据，涵盖 40 种语言
C4	15%	783GB	• 经过数据清洗的网络爬虫数据集，数据质量高
Github	4.5%	328GB	• 开源代码网站，拥有大量高质量编程代码
Wikipedia	4.5%	83GB	• 百科数据集，质量高、覆盖面广、拥有多种语言
Books	4.5%	85GB	• 书籍数据集，包含约 20 万本书籍
ArXiv	2.5%	92GB	• 学术论文数据集，语料逻辑性强、知识丰富
StackExchange	2%	78GB	• 计算机问答社区，涵盖多个计算机领域的高质量问答

来源：Meta，国金证券研究所

尽管 LLaMA 模型参数量较小，但通过采用海量的通用领域数据进行训练，性能丝毫不逊

色于 PaLM、GPT-3 等大语言模型。在与其他领先模型的通用领域性能对比测试中，LLaMA 在 8 项性能测试中取得 6 项第一、2 项第二的好成绩。

图表7：LLaMA 通用领域性能处于世界领先行列

模型名称	参数量	BoolQ	PIQA	SIQA	HS	WG	ARC-e	ARC-c	OBQA
LLaMA-65B	650 亿	85.3	**82.8**	**52.3**	**84.2**	77	**81.5**	56	**60.2**
LLaMA-33B	330 亿	83.1	82.3	50.4	82.8	76	81.4	**57.8**	58.6
PaLM	5,400 亿	**88**	82.3	—	83.4	**81.1**	76.6	53	53.4
GPT-3	1,750 亿	60.5	81	—	78.9	70.2	68.8	51.4	57.6

来源：Meta，国金证券研究所

LLaMA 模型一经发布就对外完全开源，吸引了广大 AI 开发者和研究者。目前，用户可在全球知名 AI 模型开源社区 Hugging face 中获取 LLaMA 的模型权重与训练代码。能够自由下载并使用 LLaMA 模型，既可以将其部署至设备直接进行推理，也可以基于 LLaMA 进行研究与二次开发。

2.2 小参数量可降低垂类模型开发及部署难度

进入大模型时代以来，人们对于 AI 模型的智能水平存在一种普遍共识，即参数规模越大则模型性能越强，然而这一认识并不完全正确。

AI 大模型的高性能是建立在充足的高质量数据之上。Google 在对其语言模型 T5 的实验中得出以下结论：相较于数据数量，数据质量更为重要；AI 大模型的正确发展路径是在保证数据质量的前提下，增大数据数量、并相应扩大参数规模。根据 Google 的研究，如果训练数据量和数据质量不达标，会导致模型训练不充分，此时即便一个模型的参数规模较大，其也未必具备高性能。

LLaMA 很好地诠释了数据质量的重要性，该模型通过使用比主流模型更多的训练数据，在较小的参数规模上实现了更强大的性能。同时，较小的参数规模显著降低了 LLaMA 模型的落地部署和二次开发难度。

图表8：LLaMA 兼具高性能和易部署的特点

模型名称	参数规模	部署所需的最低硬件配置
LLaMA-7B	70 亿	• NVIDIA 3060 12G 以上即可单卡部署
LLaMA-13B	130 亿	• NVIDIA 3090Ti，NVIDIA 4090 单卡部署
LLaMA-33B	330 亿	• A100-40GB 单卡部署
LLaMA-65B	650 亿	• A100-80GB 单卡部署

来源：Meta，国金证券研究所

对于希望基于 LLaMA 开发垂类模型的 AI 开发者而言，在保证预训练通用模型具备领先性能的前提下，较小的参数规模是优势而并非劣势，这是因为：

- 垂类模型场景单一：垂类 AI 模型主要面向 B 端及 G 端客户，且通常情况下仅面向单一行业、单一领域，需要处理的任务类型也较为有限。任务场景受限意味垂类 AI 模型本身并不需要太大的参数量，AI 模型仅需具备部分通用能力以及某一专业领域的相关知识即可高水准地完成工作。

  在面向垂直领域时，以 ChatGPT 为代表的大模型中的大部分参数始终处于未激活状态，这非但不会使模型在特定领域获得更强的性能，还会在模型二次开发和推理时造成大量的算力浪费。反之，参数规模较小的垂类模型能够在保证性能的同时，有效减少训练和推理的算力开支。

- 行业数据相对有限：对于绝大多数垂类 AI 模型开发者而言，在 AI 三要素中，最大的阻碍因素是行业数据相对有限。相比于量级达万亿 Tokens 的高质量通用领域数据集，行业数据无论在数量还是质量上都存在较大差距。

在众多行业领域中，金融得益于数字化程度高、资讯类内容丰富，属于数据最为丰富的行业之一。2023 年 3 月 30 日，彭博社基于多年深耕金融资讯的优势，发布金融行业大模型 BloombergGPT，其训练使用的金融行业数据约为 2,720 亿词，与目前通用领域的数据数量仍存在较大差距。

对于大多数行业而言，其行业数据量远小于金融行业。面对行业数据不足的现实问题，盲目追求大参数规模并不能够提高垂类 AI 模型的智能水平，反而会由于模型训练不充分导致模型垂类表现相对较差。

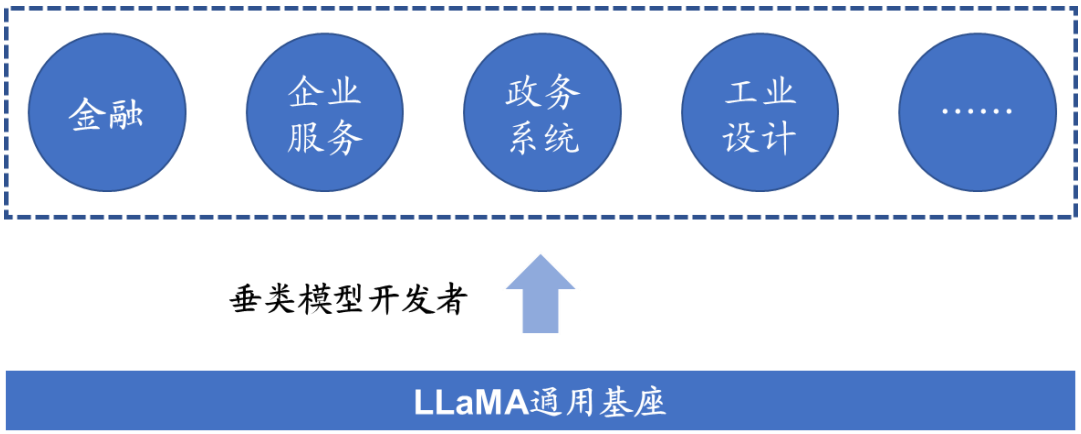
此外，LLaMA 同时对外开源了参数规模大小不同的四个子模型，为垂类模型开发者提供了灵活选择。垂类 AI 开发者可以根据应用场景、任务类型、算力规模、部署需要和行业数据储备选择合适的子模型作为基座，开发与自身需求相符的垂类 AI 模型。

2.3 LLaMA 提供通用开发基座，泛化能力更强

作为通用领域的预训练模型，LLaMA 基于海量的通用领域数据、应用无监督学习方法进行训练，训练方法极为简单。LLaMA 在训练过程中并没有应用目前最为先进的人类反馈强化学习（RLHF）方法，这导致了其智能表现相对于 ChatGPT、GPT-4 等模型稍弱，但这并非 LLaMA 的缺点。

训练方式的选择与 LLaMA 模型定位有关。LLaMA 主要面向垂类 AI 模型的开发者，旨在提供一个通用的 AI 模型基座，使得各行各业的垂类开发者可以基于该通用底座，以较低的算力开支和较短的开发周期，形成高效、精准的垂类 AI 模型，实现 AI 赋能千行万业。

图表9：LLaMA 有望带动海量垂类模型落地



来源：国金证券研究所

基于上述定位训练而成的 LLaMA 模型，以通用领域性能为核心，没有进一步应用 RLHF 等先进算法提升模型智能水平，这是因为模型在应用 RLHF 算法训练后将会导致诸多问题，从而使模型丧失作为通用基座的能力。

- 输出受到限制：RLHF 算法的实质是在不改变模型现有知识的情况下，以少量标注的高质量模板数据，对模型的输出进行诱导与限制，使 AI 模型的输出更加符合人类喜好，同时减少模型的有害输出。

  一方面，以 ChatGPT 为首的领先模型在应用 RLHF 方法解决有害输出时，可能导致模型在特定领域的表现受到限制。另一方面，各垂直应用领域对 AI 模型的谨慎性要求存在差异化，统一的谨慎性处理可能导致垂类模型在特定领域的表现弱化。

- 垂类开发者丧失定制能力：不同的行业拥有不同的任务特点，也要求垂类 AI 模型具备相应的能力。因此，对于垂类 AI 开发者而言，基于“干净”的通用模型基座，自主结合行业特点进行 RLHF 训练，将得到与行业更契合、性能更强大的垂类 AI 模型。

LLaMA 作为完全开源的领先模型，具备高度的灵活性、可配置性和泛化能力，可以作为垂类 AI 模型的通用基座。基于 LLaMA，垂类 AI 开发者可以根据其行业特点、应用行业数据定制开发相应的“行业发行版 AI 模型”。

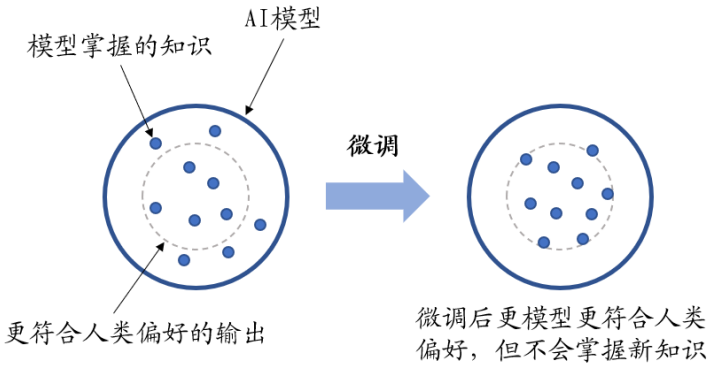
3. “通用模型+迁移学习+微调”有望成为开发新范式，数据是重要壁垒

3.1 使用“迁移学习”向模型注入新知，开发难度相对较低

目前，OpenAI 的 ChatGPT 凭借其在通用领域中的强大性能，成为了全行业的领跑者。其使用的“大规模预训练+微调”的开发模式也成为各 AI 开发者效仿的对象，乃至全行业的开发范式。然而，当我们将目光转向垂类 AI 模型开发时，这一开发模式未必有效，原因有以下两点：

- 微调无法向模型注入新知识：微调技术是指应用少量新领域的数据，对 AI 模型的内部参数进行调整，其实质是让模型更好地利用已有的知识和经验解决问题，从而提升模型性能，并不会让模型掌握新知识。

图表10：微调无法向模型内部注入新知识



来源：深度学习技术前沿公众号，国金证券研究所

目前，以 ChatGPT 为首的“All in one”路线大模型在通用领域已经具备了极强的性能，但这类模型在面对特定行业时，其表现较可落地的垂类模型仍有较大差距，行业知识匮乏是其性能不佳的最主要原因。正如我们前文所述，微调无法向模型注入新知识，这意味着在垂类模型的开发中“大规模预训练+微调”的开发范式或许较为困难。

- 行业数据是未来最宝贵的资源：或许有人认为既然“大规模预训练+微调”的开发范式对于垂类模型开发行不通，那么可以使用海量通用数据+海量行业数据，在模型训练阶段一步到位。这种方法在理论上确实具备一定可行性，然而在现实中实现难度较高。主要因为互联网时代数据孤岛问题严重，难以获得各行业全域数据。

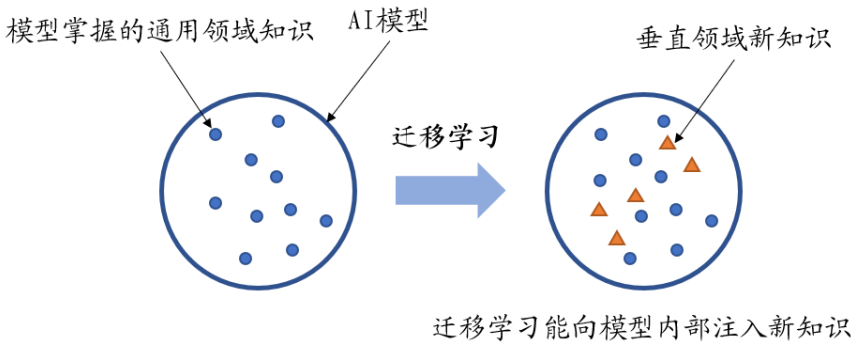
  目前行业数据的价值已经逐步为广大 AI 开发者所认知。与通用数据不同，行业数据往往是私有数据，并且由于可能涉及公司机密、客户隐私等敏感信息，对外售卖的可能性极小。即便数据内容本身并不敏感，对外售卖也可能会导致竞争对手凭借 AI 赋能，抢占相应垂直市场，造成削弱企业竞争力等不良后果。

  因此，我们认为私有化的行业数据是未来最宝贵的资源，行业数据或将成为垂类模型开发的准入壁垒，垂类模型开发不会由 AI 巨头所垄断，而是会下沉至各垂直领域的场景拥有者。

LLaMA 采用海量通用领域数据进行训练，本身已经具备了领先的通用领域性能，这意味着模型能够很好地理解并完成用户的绝大多数的通用领域任务需求，如文字生成、代码生成、知识搜索等。LLaMA 相比垂类 AI 模型，所欠缺的仅仅是对特定垂直领域知识的掌握。

LLaMA 作为垂类模型基座已经具备极强的通用能力，为了向模型注入垂直领域的知识，垂类模型开发者可以采用迁移学习技术。迁移学习是一种基于预训练模型解决新领域任务的前沿技术。基于此方法，可以使用预训练模型作为基座模型，然后应用垂直领域的数据对模型进行二次训练，从而将垂直领域中的新知识注入基座模型，大幅提升模型在特定垂直领域的表现。

图表11：迁移学习能向模型内部注入垂直领域新知识



来源：人工智能理论及算法工程研究中心公众号，国金证券研究所

迁移学习过程与模型预训练类似，通常采用无监督学习方式进行模型训练，无需准备标注数据，垂类 AI 开发者仅需准备对应的行业数据，并进行数据清洗，即可完成垂类 AI 模型的开发。垂类 AI 开发者往往深耕一个或几个行业多年，具备较充足的数据积累。因此，基于 LLaMA 开发垂类 AI 模型对于大多数开发者而言难度较低，有望大幅加速垂类 AI 落地进程。

3.2 叠加先进算法微调，进一步释放模型性能

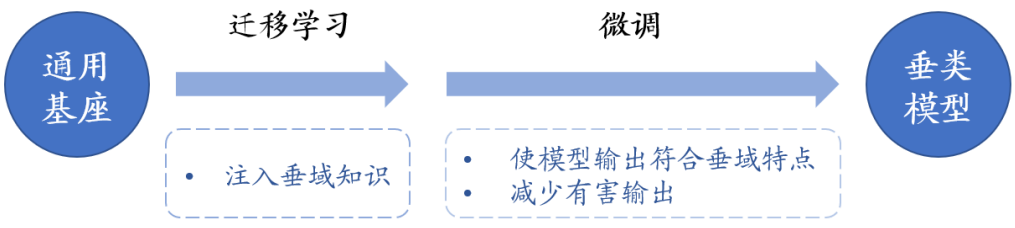
在应用迁移学习向垂类模型基座注入垂域知识后，垂类 AI 开发者仍可基于该模型，应用 RLHF 先进算法对模型进行微调，进一步提升垂类 AI 模型应用知识的能力。

- 显著提升模型智能表现：ChatGPT 的成功证明了应用 RLHF 先进算法对模型进行微调，能够显著提高模型的智能水平。垂类 AI 开发者长期深耕行业，对垂直领域客户需求把握更深，由其主导的模型微调将使模型更贴近于真实场景。
- 极大减少有害输出问题：有害输出问题是阻挠 AI 落地的重要不利因素之一，RLHF 的应用不仅能够帮助垂类开发者极大缓解输出敏感、输出不谨慎等问题，同时垂类 AI 开发者可以结合场景需求，自主选择限制哪一方面的输出，在不限制模型性能的情况下，最大程度缓解有害输出。
- 赋予垂类 AI 开发者模型定制能力：下游用户需求多样是众多垂直领域的基本特征，这要求垂类 AI 开发者模型具备一定的 AI 定制化能力。RLHF 先进算法赋予了垂类 AI 开发者模型定制能力，垂类 AI 开发者可以根据下游用户需求，以相应任务的高质量标注数据对模型的能力进行限制与诱导，使其更符合场景特点，以“指令微调定制”的方式满足下游用户的个性化需求。

RLHF 先进算法帮助人工智能在 NLP 领域实现了从 GPT-3 到 ChatGPT 的迭代，实现了从科研到应用的重大转变。23 年 4 月 12 日，微软开源 DeepSpeed Chat，提供支持端到端的 RLHF 规模化系统，极大地降低了 RLHF 训练难度和算力成本。AI 开发者只需结合垂直领域特征、任务特点，积累数量为万级的真实场景行业标注数据，即可应用 RLHF 算法完成模型微调。

我们认为“通用基座+迁移学习+微调”的三段式开发模式有望成为垂类 AI 模型的开发范式，该范式能够在成本、效益、风险匹配的前提下，最大限度加快垂类 AI 模型的开发进程。

图表12：“通用基座+迁移学习+微调”有望成为垂类模型开发新范式



来源：机器之心公众号，北京大学人工智能研究院公众号，国金证券研究所

3.3 算力消耗并非海量，成本效益匹配

我们在《AI 行业深度之三：ChatGPT 训练及多场景推理成本测算》中，对 ChatGPT 模型的算力需求进行了测算。我们将沿用该报告中的成本估算思路和相关假设，用以估算垂类模型的训练和推理成本。

- 对于训练算力的测算，我们选取 650 亿参数版本的 LLaMA 作为模型基座。Meta 在 LLaMA 的论文中对外公布了模型的训练算力消耗(以 GPU 小时计)，我们将基于此测算 LLaMA 模型基座的训练成本以及垂类模型的训练成本。

图表13：LLaMA 模型基座算力消耗

模型名称	GPU 类型	训练数据量	GPU 小时
LLaMA-7B	A100-80GB	1 万亿 Tokens	82,432
LLaMA-13B	A100-80GB	1 万亿 Tokens	135,168
LLaMA-33B	A100-80GB	1.4 万亿 Tokens	530,432
LLaMA-65B	A100-80GB	1.4 万亿 Tokens	1,022,362

来源：Meta，国金证券研究所

由于采用海量数据训练，LLaMA 模型基座的算力消耗极大，训练成本昂贵。以 LLaMA-65B 为例，该模型基于 2,048 块 A100-80GB 训练约 21 天，如果以云计算方式租用训练算力，模型的训练成本将约为 955 万元。而垂类开发者基于 LLaMA 开发垂类 AI 模型时可以节省这部分训练算力费用，极大降低 AI 研发投入。

保守起见，假设垂类 AI 开发者拥有 500 亿 Tokens 的行业数据，这一量级的行业数据已较为丰富，是 Google 的 PaLM 全部训练数据的 6%；基于这一量级的行业数据进行迁移学习，已经能够向模型中注入较为充分的行业知识。如果同样租用 2,048 块 A100-80GB 进行模型训练，其训练算力费用仅为 34 万元，训练时间约为 0.75 天。

图表14：采用云计算的垂类模型训练费用估算

估算参数	计算公式	说明
算力消耗	$\frac{500}{14,000} \times 1,022,362 = 36,512$ 小时	• 算力以 A100 的 GPU 小时计。
垂类模型训练成本	$36,512 \times 1.36 \times 6.86 = 34$ 万元	• Microsoft Azure 以 1.36 美元/每小时提供 A100 租用。
垂类模型训练时间	$\frac{36,512}{2,048 \times 24} = 0.75$ 天	• 仍租用 2,048 块 A100-80GB。

来源：Microsoft Azure，NIVIDIA，国金证券研究所

如果垂类 AI 开发者选择采用自有算力，垂类模型的训练费用将进一步降低。我们假设可以接受的单次模型训练时间为 10 天，那么垂类 AI 开发者需建立由 160 块 A100-80GB 组成的 AI 算力集群。我们采用最为成熟的 AI 算力集群搭建方式，即采购 20 个 NVIDIA DGX A100 服务器，每个服务器单价约为 20 万美元，由 8 块 A100-80GB 组成，因此算力集群的初始投入折合人民币约为 2,800 万元，以 5 年折旧计算，摊销至每个垂类模型的训练费用约为 15.3 万元。综合能源费用，每个垂类模型的训练费用约为 19 万元，即便考虑到 A100、H100 供应影响，每个垂类模型的训练费用也难以超过 25 万元。

图表15：采用自建算力的垂类模型训练费用估算

估算参数	计算公式	说明
算力消耗	36,512 小时	

估算参数	计算公式	说明
GPU 数量	$2,048 \times \frac{0.75}{10} = 153.6 \approx 160$块	• 凑整至 160 块 A100。
NVIDIA DGX A100 数量	$\frac{160}{8} = 20$ 台	• 每个 NVIDIA DGX A100 包含 8 块 A100。
算力初始投入	$20 \times 20 \times 6.86 \approx 2,800$ 万元	• NVIDIA DGX A100 售价约为 20 万美元。
摊销至单个模型的硬件成本	$\frac{2,800}{5} \times \frac{10}{365} = 15.3$ 万元	• 以折旧五年计算。
单模型能耗费用	$6.5 \times 20 \times 24 \times 10 \times 1.2 \approx 3.7$ 万元	• NVIDIA DGX A100 峰值能耗为 6.5kW。电价以 1.2 元/kWh 商业用电计。
垂类模型训练费用	$15.3 + 3.7 = 19$ 万元	

来源：Microsoft Azure，NIVIDIA，国金证券研究所

以上我们测算了模型在迁移学习阶段的训练算力成本，在模型微调阶段，由于训练量级较小，仅为万级，相关的算力成本相比之下可忽略不计。例如，斯坦福大学于 2023 年 3 月对外发布 Alpaca，这是一个基于 LLaMA-7B 基座，应用 5.2 万指令对模型微调训练得到的对话类语言模型，该模型基于 8 块 A100 微调，微调时长 3 小时，算力成本不超过 300 元。

■ 对于垂类模型推理算力测算，我们采用场景分析方法，从硬件算力出发，以测算一定公司规模下垂类模型的推理成本。我们首先从模型离线部署模式出发，假设某公司购买了参数规模为 650 亿的垂类模型，并将其部署于单个 NVIDIA DGX A100 服务器上，为全公司提供 AI 算力。

图表16：垂类模型推理算力成本估算

	估算参数	计算公式	说明
离线部署	NVIDIA DGX A100 数量	1 台	• 每个 NVIDIA DGX A100 包含 8 块 A100。
	算力初始投入	$20 \times 6.86 \approx 140$ 万元	• NVIDIA DGX A100 售价约为 20 万美元。
	推理算力	$380 \times 3 \times 4 \times 8 = 36,480$ Tokens/秒	• 380 Tokens/秒/GPU 是由 Meta 给出的单块 A100 在 LLaMA-65B 上的峰值训练算力。 • OpenAI 研究表明对于 Decoder-Decoder 模型，训练算力消耗是推理算力消耗的三倍。 • A100 的推理算力约为训练算力的 4 倍。
	用户可接受的推理速度	500 Tokens/秒	• 即假设算力分配至单用户时，AI 每秒处理和生成中文 250 字。
	峰值并发人数	$\frac{36,480}{500} \approx 72$人	• 能够同时处理的 AI 响应数
	平均每日 AI 使用人数	$72 \times 15 \approx 1000$ 人	• 以峰值并发人数的 15 倍估算平均每日 AI 使用人数。

	估算参数	计算公式	说明
	公司总人数	$1000\times2=2000$ 人	• 以平均每日 AI 使用人数的两倍估算公司总人数。
	摊销至每年的硬件算力成本	$\frac{140}{5}\approx28$ 万	• 以折旧五年计算。
	每年的能耗费用	$6.5\times(12+0.2\times12)\times365\times1.2\approx5.2$ 万元	• 以每日满载工作 12 小时，待机 12 小时计算，待机功率设定为满载功率的 0.2 • NVIDIA DGX A100 峰值能耗为 6.5kW。电价以 1.2 元/kWh 商业用电计。
	每年的推理算力费用	$28+5.2=33.2$ 万元	
云计算	每年的推理算力费	$8\times1.36\times6.86\times24\times365=66$ 万元	• Microsoft Azure 以 1.36 美元/每小时提供 A100 租用。

来源：Meta，OpenAI，《Scaling Laws for Neural Language Models》（Jared Kaplan 等，2020），国金证券研究所

根据我们的初步测算，由 8 块 A100 组成的 AI 服务器可为规模达 2,000 人的中大型企业提供 AI 服务，离线部署方案每年的推理算力成本约为 33.2 万元，若采用云计算方案则每年需花费约 66 万元算力成本。

基于上述推理成本分析，推理成本并不高昂，绝大多数中型以上企业足以负担，为各领域垂类模型落地提供了极为广阔的市场空间。

3.4 赋予垂类 AI 开发者离线部署能力和离线迭代能力

在诸多垂直领域中，下游用户出于保密和数据安全等因素考虑，不能通过接入 API 来实现 AI 赋能，但这之中的许多下游用户又希望应用 AI 实现降本增效，这就要求垂类 AI 开发者具备离线部署能力。

基于 LLaMA 二次开发而成的垂类 AI 模型自主性强，除了具备强定制化能力之外，还能够实现离线部署乃至离线迭代。

- 离线部署：由于数据安全问题，模型可离线部署将大大加强垂类模型开发者的竞争力，使其在同类开发者中脱颖而出。此外，离线部署方案往往可以通过软硬一体的方式实现，这不仅将带来下游用户单订单价值量的提升，还将进一步提升用户粘性，不断拓宽核心客户群。
- 离线迭代：如前所述，模型微调的算力消耗极低。因此，垂类开发者可以基于离线部署方案配套提供离线迭代能力。下游客户在 AI 使用过程中积累的高质量标注数据可以用于模型能力迭代，使垂类模型与具体公司更为契合，提高垂类模型的性能与易用性。在模型迭代过程中，公司仅需将数据上传至自有算力集群，采用自有算力对模型进行迭代升级，从根本上杜绝了数据泄露问题。

4. 投资建议

LLaMA 等优质开源模型的推出极大加速了下游行业 AI 应用开发效率。基于“通用基座+迁移学习+微调”的垂类 AI 模型开发范式或将成为主流，优质的行业数据资源成为影响模型性能的关键。

在此趋势之下，我们看好两类企业：1）拥有开发先进大模型能力的企业。这类企业在先进模型逐步走向闭源的趋势下，有望保持算法优势，如百度、商汤科技、科大讯飞等。2）拥有丰富行业数据的头部公司。这类企业有望基于稀缺的行业数据以及开源模型，开发出可用性更强的垂类模型。如东方财富、同花顺、恒生电子、中科软、广立微等。

5. 风险提示

■ 海外基础软硬件使用受限

若因国际关系等原因，高算力 GPU 等基础硬件或计算框架等基础软件使用受限，可能会对国内人工智能算法应用产生影响。

■ 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：021-60753903
传真：021-61038200
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 7 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：0755-83831378
传真：0755-83830558
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：中国深圳市福田区中心四路 1-1 号
嘉里建设广场 T3-2402