火山引擎FORCE总结及API收入预期

行业研究 · 行业专题

计算机 · 人工智能

投资评级：优于大市（维持评级）

证券分析师：熊莉
021-61761067
xiongli1@guosen.com.cn
S0980519030002

摘要

国信证券 GUOSEN SECURITIES

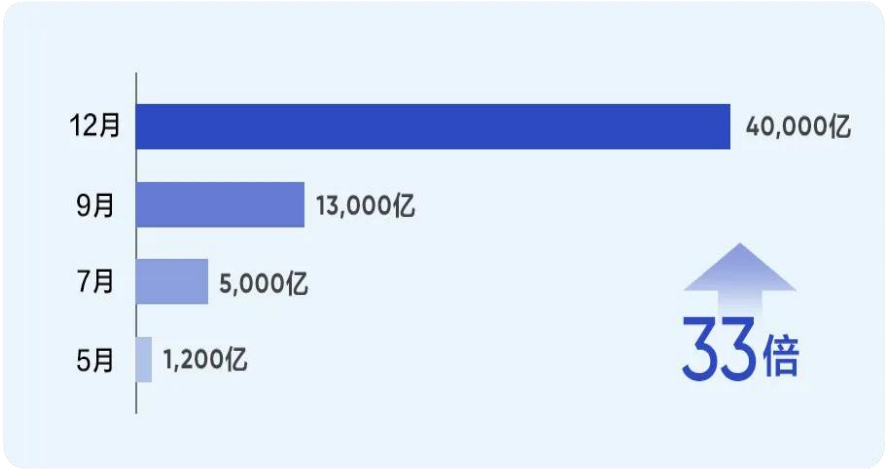
- **豆包大模型更新发布，提升AI能力，推动多行业应用和开发者生态的发展。**字节原动力大会上一系列产品发布更新：豆包视觉理解模型支持文本和图像输入，精准识别物体与场景，具备推理和复杂计算能力；豆包通用模型Pro较5月提升32%，全面对齐GPT-4o；火山引擎基于豆包大模型推出全域AI搜索，支持多模态理解，提升搜索效率；扣子1.5完善开发者生态，支持多种应用形态发布；火山引擎新服务提升AI体验，优化计算、存储和安全，数据飞轮2.0为大模型训练提供优质数据支持。
- **风险提示**：AI应用落地不及预期、市场需求不及预期、行业竞争加剧、宏观经济波动。

目录

豆包大模型推进迅速，各场景迅速渗透

- **模型使用量：**截至12月中旬，豆包通用模型的日均tokens使用量已超过4万亿，较五月首次发布时日均1200亿增长了33倍，大模型应用正在向各行各业加速渗透。
- **智能终端合作情况：**豆包大模型已经与八成主流汽车品牌合作，并接入到多家手机、PC等智能终端，覆盖终端设备约3亿台，来自智能终端的豆包大模型调用量在半年时间内增长100倍。
- **多场景快速渗透：**最近3个月，豆包大模型在信息处理场景的调用量增长了39倍，帮助企业更好的分析和处理内部及外部的数据；客服与销售场景增长16倍，帮助企业更好的服务客户，扩大销售；硬件终端场景增长13倍，AI工具场景增长9倍，学习教育等场景也有大幅增长。

图：豆包大模型API调用量迅速提升



资料来源：火山引擎，国信证券经济研究所整理

图：豆包大模型在不同场景迅速渗透

资料来源：火山引擎，国信证券经济研究所整理

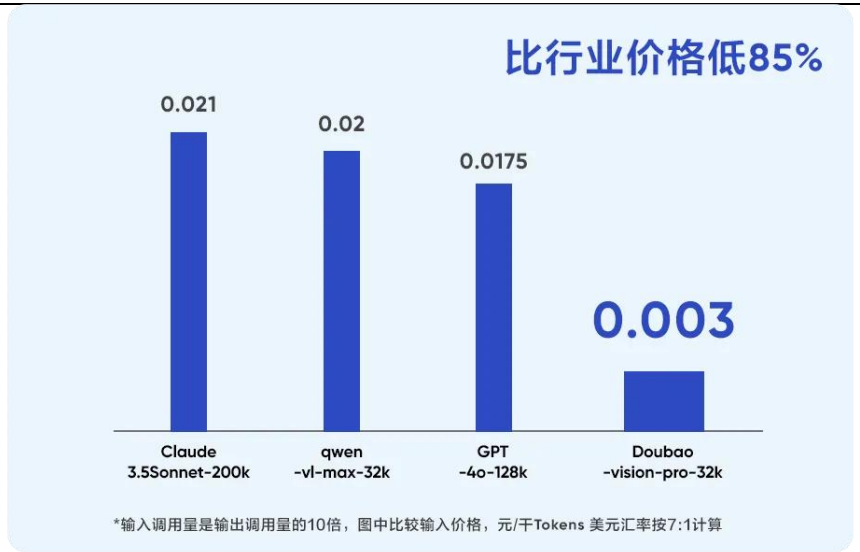
豆包视觉理解模型发布，拓宽大模型场景边界

- **豆包视觉理解模型发布，用户可以同时输入文本和图像相关的问题。模型能够综合理解并给出准确的回答，将极大地简化应用的开发流程，在金融、医疗、教育、旅游等诸多行业有广阔的应用前景。**
- **内容识别能力：**精准识别物体类别、形状及物体间关系，理解场景含义，能够推理出物体信息，如通过影子识别猫，通过光照识别丁达尔效应，识别并科普现实对象。
- **理解和推理能力：**识别文字和图像信息实现复杂逻辑计算，解答微积分、分析论文图表、解决物理题及处理真实代码等。
- **视觉描述与创作能力：**模型能根据图像描述创作内容，如为文创产品写祝福语、根据涂鸦创作故事、讲述物体背后的文化故事，以及创作古风诗歌等。
- **价格：**豆包视觉理解模型输入价格仅为0.003元/千tokens，1块钱可处理284张720P的图片，比行业价格便宜85%，目前已经接入豆包App和PC端产品中。

图：视觉理解模型能力增强，场景拓宽

资料来源：火山引擎，国信证券经济研究所整理

图：视觉理解模型价格远低于同行

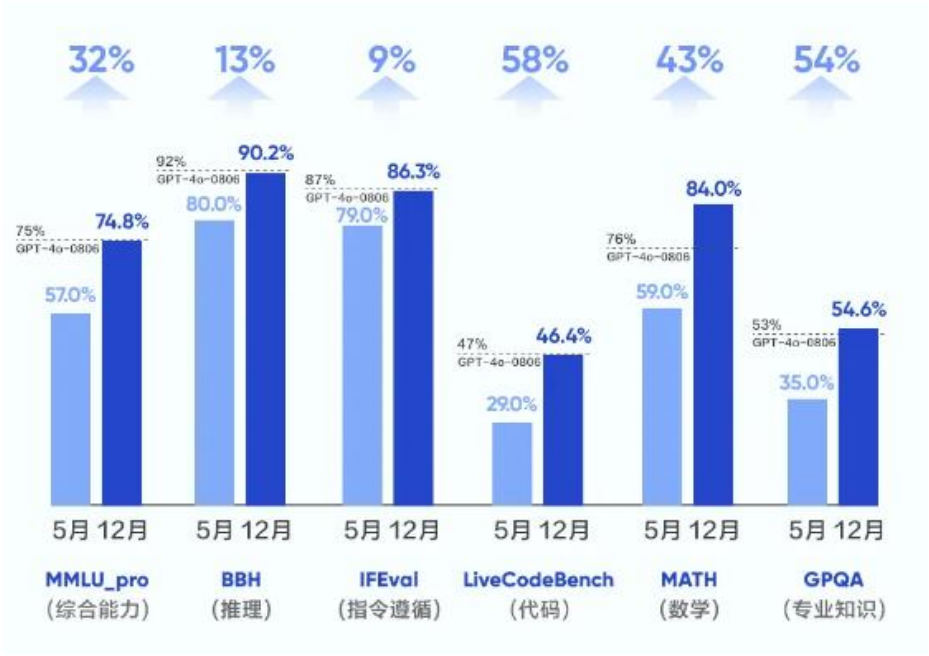


资料来源：火山引擎，国信证券经济研究所整理

大模型家族全面升级：主力模型Doubao-pro升级，对齐 GPT-4o

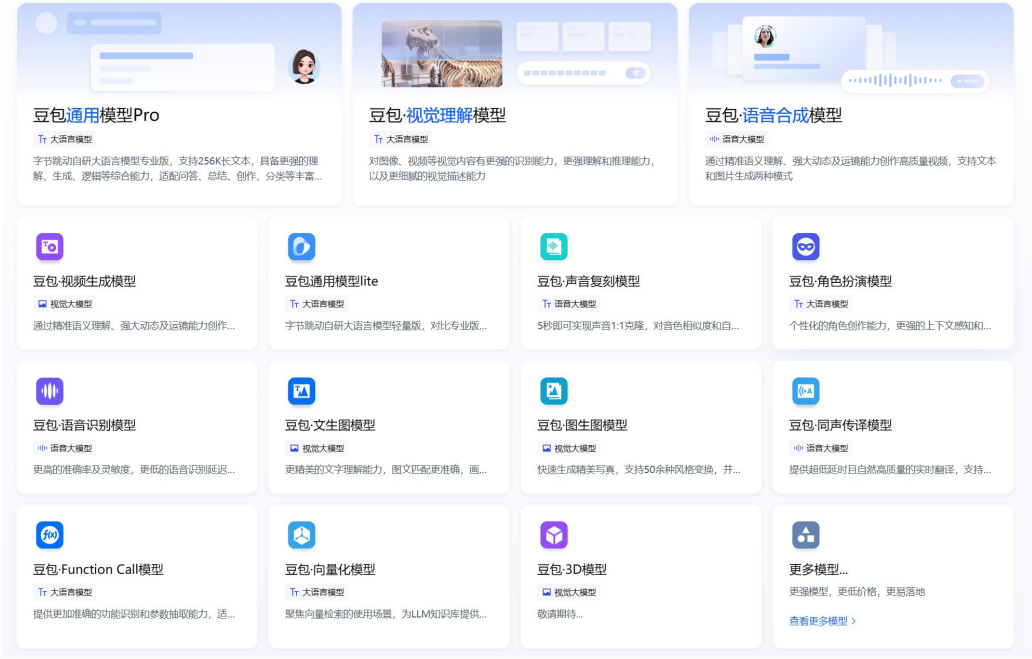
- **豆包通用模型pro完成新版本迭代，综合任务处理能力较5月份提升32%，在推理上提升13%，在指令遵循上提升9%，在代码上提升58%，在数学上提升43%，在专业知识领域能力提升54%。**
- 豆包通用模型Pro已全面对齐GPT-4o的能力，但使用价格远低于后者，推理输入价格为0.0008元/千tokens。豆包大模型具备行业领先的大模型能力，保持成本优势；应用全面，覆盖广泛通用任务场景；支持联网问答、角色扮演和工具调用，满足精细化业务需求。

图：单POD IT耗能测算



资料来源：《NVIDIA DGX SuperPOD Data Center Design》，国信证券经济研究所整理

图：豆包大模型家族



资料来源：火山引擎官网，国信证券经济研究所整理

大模型家族全面升级：音乐、文生图等模型更新发布

- **豆包·音乐模型4.0发布，从 “高光片段” 走向“完整歌曲”：**1）支持包括前奏、主歌、副歌、间奏、过渡段的3分钟全曲创作；2）歌词局部修改，仍能适配原有旋律；3）全曲风格、情感和音乐逻辑保持一致，曲风连贯。
- **豆包·文生图模型2.1发布：**1）一键P图：对中英文、专有名词指令理解精度高；聚焦目标，编辑效果质量高；可实现多元风格，美观自然；2）一键海报：高质量精准生成中文；字体与图片内容巧妙融合；模型最快做到6秒出图，极速生成海报。
- **veOmniverse+豆包·3D生成模型发布：**该模型采用3D-DiT架构，可生成高质量3D模块。该模型与火山引擎数字孪生平台veOmniverse结合使用，可以高效完成智能训练、数据合成和数字资产制作，成为一套支持AIGC创作的物理世界仿真模拟器。通过快速批量生成并上传至云空间，布局师可实时调用并完成场景设计，提升创作效率。

图：音乐模型4.0，支持全曲创作

资料来源：火山引擎，国信证券经济研究所整理

图：豆包·3D生成模型，快速生成3D场景

资料来源：豆包大模型团队，国信证券经济研究所整理

基于豆包·视频生成模型，即梦成为“想象力的相机”

- **视频生成模型开放：**9月24日，豆包·视频生成模型发布，此后，该模型通过即梦APP和网页端已对C端用户开放使用。面向企业客户和开发者，豆包·视频生成模型将于2024年1月依托火山引擎正式对外开放服务。
- **基于豆包·视频生成模型，为用户带来创新体验：**即梦AI是剪映于今年5月上线的AI内容平台，支持通过自然语言及图片输入，生成高质量的图像及视频。即梦希望成为“想象力世界”的相机，帮助有想法的人轻松表达、自由创作。
- **即梦支持动态海报生成，文字、画面、排版全面兼顾**
  - 全新的海报生成功能：用户只需通过一句话就可在几分钟内轻松生成设计师水平的海报。即梦支持长提示词理解，用户可通过以引号输入想要生成的具体文字、增加更多描述等方式，满足层次感更强、更有创意的海报需求。
  - 一键变成为动态海报：即梦还拥有将静态海报一键变成为动态海报的能力，让图片具备更强表现力。

图：即梦全新海报生成功能

资料来源：火山引擎，国信证券经济研究所整理

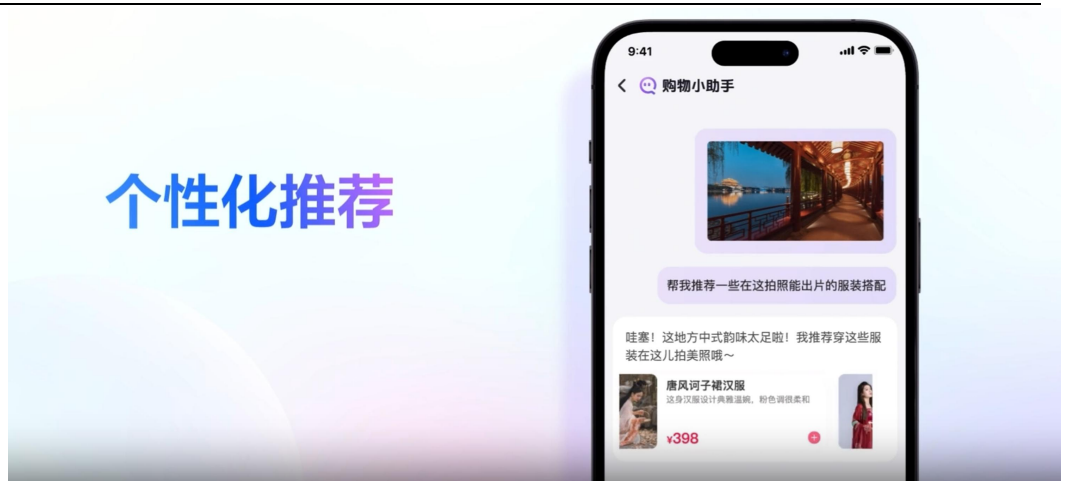
图：即梦海报生成兼顾文字、画面、排版

资料来源：火山引擎，国信证券经济研究所整理

火山引擎AI应用开发平台持续升级，加速大模型的落地运用

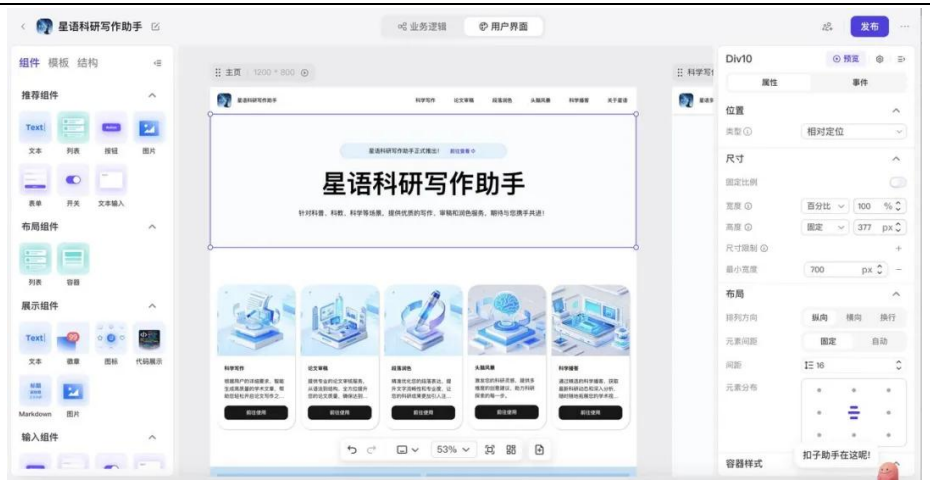
- **火山引擎AI搜推引擎“发现更多，推荐更准，搜索无限可能”**：基于豆包大模型，火山引推出全域AI搜索，提供精准、个性化的搜索推荐，支持文本、图像、音频和视频多模态理解。通过场景化推荐、企业私域信息整合和联网问答服务，搜推引擎帮助企业提升信息获取和搜索效率，具备超大规模吞吐和秒级检索能力。
- **扣子1.5发布，让AI离应用再近一步：**开发者生态不断完善，已吸引超过100万活跃开发者，发布超200万个智能体。全新的AI应用开发环境支持GUI界面搭建，可一键发布为小程序、H5、API等形式。扣子具备强大的多模态能力，提供音视频对话功能，低至1秒的延迟和低成本SDK支持各类硬件接入。此外，扣子提供海量精品模板，覆盖多业务场景。
- **HiAgent1.5发布，敏捷构建企业级 A 原生应用的能力中心：**提供AI转型支持，包括评测体系、100+行业模板和AI咨询。支持企业级插件、灵活集成，利用GraphRAG构建知识图谱，融合CUI和GUI打造智能交互引擎，并支持全栈私有化部署保障安全。

图：AI搜推引擎个性化的搜索推荐



资料来源：火山引擎，国信证券经济研究所整理

图：扣子1.5支持GUI搭建界面，可为小程序、H5、API等多种应用形态



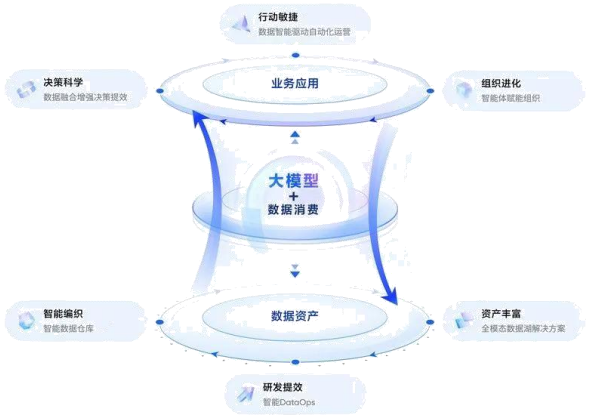
资料来源：火山引擎开发者社区，国信证券经济研究所整理

技术架构面向AI全面转型，AI云与基础设施持续创新

国信证券 GUOSEN SECURITIES

- **火山引擎在基础架构、数据分析等层面带来新服务，为企业打造更便捷、更高效、更安全的AI体验。**
- **AI云原生：打造以A负载为中心的基础架构新范式**
  - 在计算层面，新一代的火山引擎GPU实例，通过vRDMA网络，支持大规模并行计算和P/D分离推理架构，显著提升训练推理效率。
  - 存储上，新推出的EIC弹性极速缓存，能够实现GPU直连，使大模型推理时延降低至1/50；成本降低20%。
  - 在安全层面，火山将推出PCC私密云服务，构建大模型的可信应用体系。
- **数据飞轮2.0：全链路AI开启数智生产力新时代**
  - Data Fabric 驱动下的ChatBI智能体，可以让数据消费变得更简单直接。该方案通过构建完整的智能数据服务体系，打破数据“专业”壁垒，帮助企业内每个业务都能定制专属智能体，持续降低数据使用门槛，提升大模型能力下的数据反馈效率和准确率。
  - 多模态数据湖，拓宽了数据资产的边界，可以实现海量结构化、半结构化及非结构化数据的统一精细化管理，全方位兼容各类数据格式，为LLM 预训练、持续训练和微调全程各个环节提供更好的数据支持。

图：火山引擎数据飞轮2.0模式图



资料来源：字节跳动数据平台，国信证券经济研究所整理

图：火山引擎多模态数据湖解决方案

资料来源：字节跳动数据平台，国信证券经济研究所整理

豆包API调用收入计算

- 我们的计算基于以下假设：1）目前豆包Pro大模型能力已全方位对齐ChatGPT-4o，我们假设豆包通用大模型的参数量已达到万亿以上，与ChatGPT参数量齐平，目前万亿模型普遍采用MoE架构，我们假设每次调用API时使用的模型参数为1100亿；2）数据中心算力调用情况在一天内可出现多次波峰及波谷，我们假设算力设备投建时，所需的算力容量为平均需求的3倍；3）假设数据中心中，单卡平均利用率为45%；4）假设豆包的所有模型API调用平均价格为1.5元/百万Tokens；5）假设未来API调用将有40%为付费调用，剩余60%为免费用户的调用；6）目前国外领先云服务商如微软、Meta等普遍预期计算设备折旧年限为6年，我们假设字节跳动数据中心折旧年限同为6年。**基于上述假设，我们预期在日均100万亿Tokens调用量时，公司年收入可达219亿元。**

图：豆包API调用算力需求及收入

假设日均API调用量为100万亿Tokens	
通用模型	**数据**
豆包通用模型参数量（亿）	1100
预期日均Tokens调用量（亿）	1000000
每日推理总时长（秒）	86400
平均每秒调用Tokens（亿）	11.57
假设峰值算力需求/平均需求（倍）	3
推理所需理论算力=2*模型参数量*每秒调用Tokens/100（EFLOPS）	763.89
H800单卡算力（TFLOPS）	1979
假设单卡利用率（%）	45%
实际需要卡数（万张）	**85.78**
API调用平均单价（元/百万Tokens）	1.5
API调用付费率（%）	40%
日均API调用收入（亿元）	0.6
年收入（亿元）	**219**
单张H800芯片价格（万元）	15
显卡折旧年限（年）	6
年均算力投入（亿元）	**214.44**

假设日均API调用量为1000万亿Tokens	
通用模型	**数据**
豆包通用模型参数量（亿）	1100
预期日均Tokens调用量（亿）	10000000
每日推理总时长（秒）	86400
平均每秒调用Tokens（亿）	115.74
假设峰值算力需求/平均需求（倍）	3
推理所需理论算力=2*模型参数量*每秒调用Tokens/100（EFLOPS）	7638.89
H800单卡算力（TFLOPS）	1979
假设单卡利用率（%）	45%
实际需要卡数（万张）	**857.77**
API调用平均单价（元/百万Tokens）	1.5
API调用付费率（%）	40%
日均API调用收入（亿元）	6
年收入（亿元）	**2190**
单张H800芯片价格（万元）	15
显卡折旧年限（年）	6
年均算力投入（亿元）	**2144.43**

资料来源：火山引擎，英伟达，国信证券经济研究所整理

目录

风险提示

- AI应用落地不及预期、市场需求不及预期、行业竞争加剧、宏观经济波动。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032