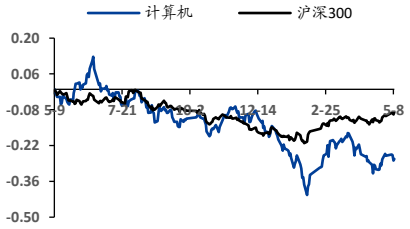
计算机行业周报

大模型内卷升级，国内外发布会精彩纷呈

强于大市（维持评级）

一年内行业相对大盘走势



团队成员

分析师： **钱劲宇(S0210524040006)**
QJY3773@hfzq.com.cn

相关报告

投资要点：

- **本周 AI 新闻速递：**

国内 AI：（1）混元文生图大模型开源，可免费商用。（2）零一万物开源 Yi-1.5 系列模型。（3）MOE 模型再突破，专家激活率显著提升。（4）百度发布 Robotaxi 成果与 L4 级自动驾驶大模型 Apollo ADFM。（5）未来智能发布讯飞会议耳机 Pro 2、iFLYBUDS 2 和 Kit 2。（6）联发科发布新一代 AI 芯片，端侧落地加速。

国外 AI：（1）打破 Decoder-Only 架构壁垒，Decoder-Decoder 架构（YOCO）横空出世。（2）OpenAI 推出端到端人工智能模型 GPT-4o，并免费开放。（3）美国加州正测试 CHATGPT 等生成式 AI 在四大部门的应用。谷歌 DeepMind 发布了一款全新的 AI 代理（Agent）产品 Project Astra。（4）谷歌发布第六代 TPU 芯片 Trillium。（5）谷歌 DeepMind 发布了一款全新的 AI 代理（Agent）产品 Project Astra。（6）英伟达与谷歌合作发布大模型 Gemma 2 和 PaliGemma。

- **大模型内卷升级，国内外发布会精彩纷呈**

OpenAI： 发布端到端大模型 GPT-4o。GPT-4o 支持文字、音频、图像任意组合的输入和输出。GPT-4o 对于音频的响应时间平均为 320 毫秒，对比 GPT 3.5（2.8 秒）与 GPT 4（5.4 秒）显著缩短，与人类的响应时间基本一致。综合能力达到第一梯队，多模态及代码能力上领先显著。定价上，GPT-4o 相对 GPT-4 Turbo，输入/输出价格分别减半，分别为$5/$15/百万 tokens。同时，GPT-4o 的速率限制支持每分钟最多 1000 万 tokens。

Google： 公布了 22 项 AI 产品及技术，软硬件全栈生态均有升级。Gemini 1.5 Pro 上下文长度翻倍至 200 万 tokens，能够处理 1500 页 PDF，30000 行代码或者 1 小时的视频，在代码生成、逻辑推理和规划、多轮对话、音频与图像理解能力等多项能力上也有升级，支持 35 种语言。

字节跳动： 发布豆包大模型 9 大家族产品。其中 pro 版上下文长度达 128K，全系列可精调；lite 版具备较快的响应速度，延迟降低 50%。定价再创新低。32K 上下文长度，pro 版价格为 0.0008 元/千 tokens，比行业低 99.3%；128K 上下文长度，pro 版价格为 0.005/千 tokens，比行业价格低 95.8%。

腾讯： 正式推出混元 pro、standard、lite 等多种尺寸模型。综合性能上，模型较上一代提升 50%，部分中文能力追平 GPT-4，图像、视频、3D 等多模态能力都具备且有一定升级。目前已在 600 多个腾讯内部业务和场景中落地。知识管理引擎借助 LLM+RAG，显著降低大模型幻觉，帮助企业快速搭建企业知识库及应用。

- **风险提示**

大模型迭代不及预期的风险；大模型商业落地不及预期的风险等。

正文目录

图表目录

1 本周 AI 新闻速递

1.1 国内 AI

（1）**混元文生图大模型开源，可免费商用**。该模型为业界首个中文原生的 DIT 架构文生图开源模型，架构与 Sora 一致，支持中英文双语输入及理解，参数量 15 亿。根据评测，腾讯混元文生图模型效果远超 Stable Diffusion ，在开源文生图模型中效果最好；整体能力国际领先。

（2）零一万物开源 Yi-1.5 系列模型。Yi-1.5 包括一系列预训练和微调模型，分为 6B、9B、34B 三个版本，采用 Apache 2.0 许可证。此外，在定价上 Yi 系列模型也颇具性价比，Yi-Large API，百万 token 价格为 20 元，仅为 GPT-4 Turbo 的 1/3。

（3）MOE 模型再突破，专家激活率显著提升。微软研究院和清华大学提出了多头混合专家（MH-MoE）。MH-MoE 采用了多头机制，可将每个输入 token 分成多个子 token。然后将这些子 token 分配给一组多样化的专家并行处理，之后再无缝地将它们整合进原来的 token 中。根据相关论文，该模型实现了 90.71%的激活率，扩展能力更强。

（4）百度发布 Robotaxi 成果与 L4 级自动驾驶大模型 Apollo ADFM。结合点云和视觉的多模态融合，Apollo ADFM 能够精确地检测和理解复杂环境中的障碍物，为自动驾驶汽车在城市道路上的安全运行提供了坚实的技术支撑。此外 Apollo 在硬件配置、安全冗余设计和出行体验方面都有显著提升。

（5）未来智能发布讯飞会议耳机 Pro 2、iFLYBUDS 2 和 Kit 2。讯飞会议耳机 Pro 2 能够实现从录音、到转写再到翻译的全流程闭环，叠加 viaim AI 功能，能够快速在长会议纪要中归纳重点。目前讯飞耳机已进入全球 154 个国家和地区，去年至今，每月销售环比增速超过 50%。

（6）联发科发布新一代 AI 芯片，端侧落地加速。联发科在首届天玑开发者大会 2024（MDDC 2024）发布全新一代芯片天玑 9300+与 AI 工具，已在 vivo 发布的旗舰手机 vivo X100S 上搭载。天玑 9300+内置第七代 AI 引擎 APU 790，率先支持 AI 推测解码加速技术，生成速度可提升 120%。同时支持天玑 AI LoRA Fusion 2.0 技术，生成效率提升 100%，内存空间节省 50%。该芯片支持业内主流大模型，包括阿里云通义千问大模型、百川大模型、文心大模型、谷歌 Gemini Nano、零一万物终端大模型、Meta Llama 2、Llama 3 等。

1.2 国外 AI

（1）打破 Decoder-Only 架构壁垒，Decoder-Decoder 架构（YOCO）横空出世。YOCO 仅缓存一次键值对，可大幅降低 GPU 内存需求，且能够保留全局注意力。在处理 512K 上下文时，传统的 Decoder-Only 架构内存使用是 YOCO 的 6.4 倍，预填

充延迟是 YOCO 的 30.3 倍，吞吐量仅为 YOCO 10.4%。

（2）OpenAI 推出端到端人工智能模型 GPT-4o，并免费开放。GPT-4o 文本、推理、编码能力达到 GPT-4 Turbo 水平，速度是上一代 AI 大模型 GPT-4 Turbo 的两倍，但成本仅为 GPT-4 Turbo 的一半。此外，此次 GPT-4o 亦向免费用户开放。

（3）美国加州正测试 CHATGPT 等生成式 AI 在四大部门的应用。测试时间 6 个月，共有 5 家公司为其提供技术支持，分别是 OpenAI、Anthropic、谷歌、Meta 和 ServiceNow，主要应用在税收和收费管理部、交通部、公共卫生部以及卫生与公众服务部 4 大部门。

（4）谷歌发布第六代 TPU 芯片 Trillium。与 TPU v5e 相比，Trillium TPU 的每芯片峰值计算性能提高了 4.7 倍，HBM 容量和带宽增加一倍。此外，Trillium 还配备了第三代 SparseCore，这是一种专用加速器，用于处理高级排名和推荐工作负载中常见的超大嵌入。在能效上，Trillium 较 TPU v5e 高出 67% 以上。

（5）谷歌 DeepMind 发布了一款全新的 AI 代理（Agent）产品 Project Astra。Project Astra 能够实现跨文本、音频、视频多模态实时推理。能够借助手机、谷歌眼镜等终端实现解答数学题、解读周围环境等能力。在导盲、翻译、学习、办公等多个领域有广泛的应用前景。

（6）英伟达与谷歌合作发布大模型 Gemma 2 和 PaliGemma。两个模型与 Gemini 拥有相同的技术底座，其中 Gemma 2 是 Gemma 的升级版，通过全新的架构使得性能和效率有了新的突破。PaliGemma 是开源视觉语言模型（VLM）专门用于视觉语言任务，例如图像和短视频字幕、视觉问题解答、图像文本理解、对象检测和对象分割等。

2 大模型内卷升级，国内外发布会精彩纷呈

本周迎来国内外大模型的密集发布，从 OpenAI 的春季发布会开始、Google、字节跳动、腾讯分别亮相，从产品、生态、价格等多个角度带来了国内外大模型的最新变化。

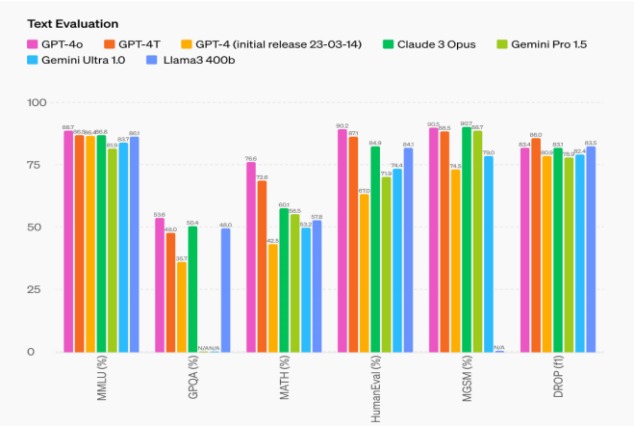
2.1 OpenAI：端到端大模型 GPT-4o 免费开放，智能助手表现惊艳

OpenAI 发布端到端大模型 GPT-4o。与老版本相比，GPT-4o 支持文字、音频、图像任意组合的输入和输出。其中 GPT-4o 对于音频的响应时间平均为 320 毫秒，对比 GPT 3.5（2.8 秒）与 GPT 4（5.4 秒）显著缩短，与人类的响应时间基本一致。背后的核心亮点在于其端到端全模态结合能力，这使得 GPT-4o 的全部输入输出均可以由同一个神经网络处理（而过去 OpenAI 的模型对于多模态的处理则需要通过三个独立模型的管道，并经过相互转化得到输出结果）。

综合能力上，GPT-4o 在文本、推理、编码方面与 GPT-4 Turbo 同级别，在多语

言、音频、视频上达到更高的能力。

图表1：GPT-4o 文本能力对标 GPT-4 Turbo



来源：OpenAI，华福证券研究所

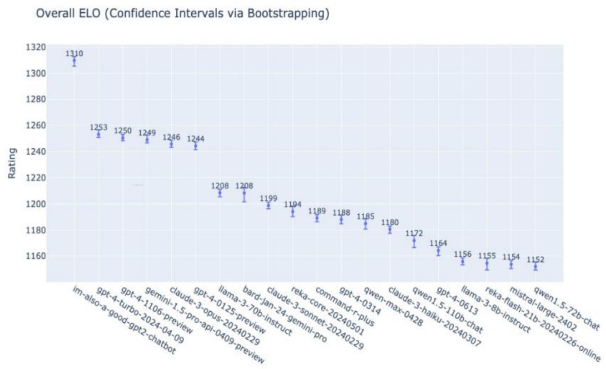
图表2：视觉能力全面超越 GPT-4

Eval Sets	GPT-4o	GPT-4T 2024-04-09	Gemini 1.0 Ultra	Gemini 1.5 Pro	Claude Opus
MMMU (%) (val)	69.1	63.1	59.4	58.5	59.4
MathVista (%) (testmini)	63.8	58.1	53.0	52.1	50.5
AI2D (%) (test)	94.2	89.4	79.5	80.3	88.1
ChartQA (%) (test)	85.7	78.1	80.8	81.3	80.8
DocVQA (%) (test)	92.8	87.2	90.9	86.5	89.3
ActivityNet (%) (test)	61.9	59.5	52.2	56.7	
EgoSchema (%) (test)	72.2	63.9	61.5	63.2	

来源：OpenAI，华福证券研究所

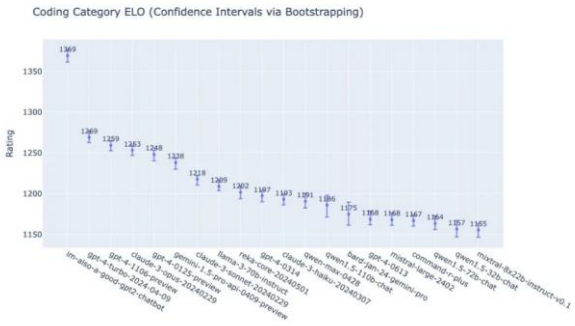
根据 OpenAI 研究员 William Fedus 反馈，GPT-4o 在综合 ELO 上亦处于领先水平，编码能力较前代提升 100 ELO。

图表3：综合 ELO 全面领先



来源：腾讯科技，华福证券研究所

图表4：代码能力显著提升



来源：腾讯科技，华福证券研究所

定价上，GPT-4o 相对 GPT-4 Turbo，输入/输出价格分别减半，分别为$5/$15/百万 tokens。同时，GPT-4o 的速率限制支持每分钟最多 1000 万个 tokens。

图表5：GPT-4o 官网定价

GPT-4o

GPT-4o is our most advanced multimodal model that's faster and cheaper than GPT-4 Turbo with stronger vision capabilities. The model has 128K context and an October 2023 knowledge cutoff.

Learn about GPT-4o ↗

Model	Input	Output
gpt-4o	US$5.00 / 1M tokens	US$15.00 / 1M tokens
gpt-4o-2024-05-13	US$5.00 / 1M tokens	US$15.00 / 1M tokens

来源：OpenAI，华福证券研究所

图表6：GPT-4 Turbo 官网定价

GPT-4 Turbo

GPT-4 Turbo is offered at 128K context with an April 2023 knowledge cutoff and basic support for vision.

Learn about GPT-4 Turbo ↗

Model	Input	Output
gpt-4-turbo	US$10.00 / 1M tokens	US$30.00 / 1M tokens
gpt-4-turbo-2024-04-09	US$10.00 / 1M tokens	US$30.00 / 1M tokens

来源：OpenAI，华福证券研究所

依托强大的 GPT-4o，智能助手实测表现惊艳。发布会中对 OpenAI 语音对话进

行演示，可以看到 GPT-4o 可以在对话完成后立刻进行响应，同时可以感知人类的表情和情绪，并做出不同的反馈。对于视觉能力，发布会中展示了 GPT-4o 通过理解摄像中的方程式来解决问题的 demo，可以看到 GPT-4o 能够在不透漏答案的情况下帮助演示者完成方程的求解，几乎能够扮演教师的角色。

2.2 Google：AI 赋能软硬件全栈升级

谷歌在 I/O 大会上公布了 22 项 AI 产品及技术。大模型升级上： Gemini 1.5 Pro 进阶版升级 200 万 tokens、Gemini 1.5 Flash 轻量模型、文生图模型 Imagen 3、70 秒视频生成模型 Veo、首个视觉语言开放模型 PaliGemma 等，还剧透了下一代 Gemma 2 大模型。**其他产品及应用上：**第六代 TPU、AI 基础设施、AI 搜索新功能等。

图表7：Google I/O 2024 核心要点

分类	产品与细节
AI Agent	Project Astra：一边实时拍摄一边与手机上的 AI agent 交谈，具备强大多模态理解和实时对话能力。
	Project Astra：一边实时拍摄一边与手机上的 AI agent 交谈，具备强大多模态理解和实时对话能力。
新模型新工具	Veo：高清视频生成模型，可生成通过 60 秒高质量 1080p 视频。
	Imagen 3：谷歌最高质量的文生图模型。
	Music AI Sandbox：AI 音乐创作工具。
	Gemini 1.5 Pro 进阶版：上下文窗口扩展至 200 万个 token。
	Gemini 1.5 Flash：轻量级模型，100 万个 token 上下文窗口。
	PaliGemina：谷歌 Gemma 家族首个视觉语言开放模型。
	Gemma 2 抢先看：将在 6 月份发布 270 亿参数的模型版本。
	LearnLM：基于 Gemini 针对学习进行微调的系列模型。
新功能	AI 搜索：将支持多轮推理、规划能力、对视频提问，将推出谷歌图库实验性功能 Ask Photos。
	Google Workspace 应用：侧边面板可使用 Gemini 1.5 Pro 模型，推出 Gmail 移动 app 新功能，添加对更多语言的支持。
	Gemini Live：全新移动对话体验，交谈更自然。
	Gems：创建 Gemini 定制版本。
	Gemini Advanced：新增旅行计划、数据分析功能，支持访问 Gemini 1.5 Pro。
	画圈即搜功能：在手机或平板电脑上圈出物理和数学问题，获得解法。
	安卓版 Gemini 新功能：把 Gemini 引入安卓系统层。
	Gemini Nano 新功能：即将推出多模态功能，正在测试诈骗检测功能。
	Gemini Nano 新功能：即将推出多模态功能，正在测试诈骗检测功能。
新基建	第六代 TPU：谷歌迄今性能最强的 TPU。
	AI 基础设施：从 AI 超算到跨越 200 多英里陆地和海底光纤的海底电缆网络。
	AI 基础设施：从 AI 超算到跨越 200 多英里陆地和海底光纤的海底电缆网络。
安全与负责任的 AI	AI 辅助红队：用 AlphaGo 开发的提高 AI 能力新技术。
	扩展 SynthID 文本水印：对 Gemini 生成的文本和 Veo 生成的视频添加水印，计划开源。
	扩展负责任的生成式 AI 工具包：发布开源的大语言模型比较器，帮助评估模型质量与安全。

来源：智东西，华福证券研究所

Gemini 1.5 Pro 再进阶，上下文窗口翻倍。此次 Gemini 1.5 Pro 升级版将支持 200

万 tokens，能够处理 1500 页 PDF，30000 行代码或者 1 小时的视频，同时意味着模型可以处理更加复杂的任务。此外，在代码生成、逻辑推理和规划、多轮对话、音频与图像理解能力等多项能力上也有升级，支持 35 种语言，现已面向 Workspace Labs 开放。

图表8：Gemini 1.5 Pro 面向谷歌 Workspace Labs 开放

来源：智东西，华福证券研究所

2.3 字节跳动：豆包大模型家族正式亮相

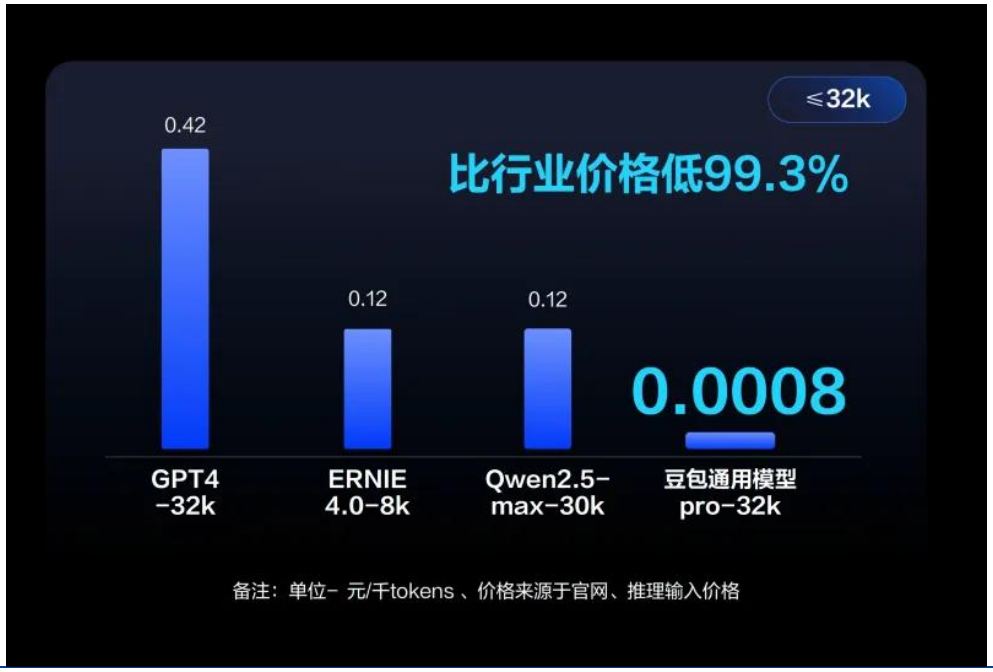
字节跳动旗下火山引擎揭秘九大豆包大模型成员。其中，豆包通用模型分为两个尺寸，pro 版上下文长度达 128K，全系列可精调；lite 版具备较快的响应速度，延迟降低 50%。目前豆包大模型已接入字节内部 50 多个业务，包括抖音、飞书等，日均处理 1200 亿 Tokens 文本，生成 3000 万张图片。

图表9：字节跳动豆包大模型家族

来源：量子位，华福证券研究所

定价再创新低。对于 32K 上下文长度，pro 版价格为 0.0008 元/千 tokens，比行业低 99.3%；而对于 128K 上下文长度，pro 版价格为 0.005/千 tokens，比行业价格低 95.8%。

图表10：豆包大模型 32K 上下文版本定价



来源：21 世纪商业评论，华福证券研究所

2.4 腾讯：混元大模型全面升级，应用门槛→0

在腾讯云生成式 AI 产业应用峰会上，腾讯正式推出混元 pro、standard、lite 等多种尺寸模型。其中 Standard 版本支持 256K 上下文，这些模型已通过腾讯云，面向企业、开发者全量开放。综合性能上，模型较上一代提升 50%，部分中文能力追平 GPT-4，图像、视频、3D 等多模态能力都具备。其中文生图全面升级中文原生 DiT 架构，测评结果全面领先，目前已开源；视频生成支持最多 16s；文/图生 3D 支持单图 30s 生成 3D 模型

图表11：混元大模型的三大尺寸

来源：腾讯云，华福证券研究所

目前腾讯混元已在 600 多个腾讯内部业务和场景中落地。包括微信读书、腾讯

会议、腾讯客服、腾讯广告等等。腾讯旗下 SaaS 产品已全面接入混元。此外，混元上线 AI 智能体创作与分发平台腾讯元器，用户可创作智能体并实现一键分发。

图表12：腾讯会议 AI 小助手智能会议纪要，回答相关问题

来源：腾讯云，华福证券研究所

知识管理引擎帮助企业快速搭建企业知识库及应用。知识管理引擎以 LLM+RAG 为技术底座，使得企业可以借助同一套专属知识，搭建内外部知识库及应用。借助腾讯云 OCR 解析大模型，能够将企业知识解析的准确率提升 25%，并支持将这些专属知识文档一键导入大模型，再通过语义级知识切分，数据向量化等方式，实现大模型快速准确的反馈答案。通过知识引擎，5 分钟即可构建企业级的 AI 问答应用。

图表13：借助专属知识文档及相关技术，大模型能够快速准确反馈答案

来源：腾讯云，华福证券研究所

3 风险提示

大模型迭代不及预期的风险：目前大多数大模型的研发仍处于早期投入阶段，存在研发失败以及迭代进度缓慢的风险。

大模型商业落地不及预期的风险：当前大模型与各个场景结合处于早期磨合阶段，仍需要进一步的训练、微调及试用，存在落地时间点不及预期的风险。

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

一般声明

华福证券有限责任公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司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。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	评级	评级说明
公司评级	买入	未来 6 个月内，个股相对市场基准指数指数涨幅在 20%以上
	持有	未来 6 个月内，个股相对市场基准指数指数涨幅介于 10%与 20%之间
	中性	未来 6 个月内，个股相对市场基准指数涨幅介于-10%与 10%之间
	回避	未来 6 个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来 6 个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来 6 个月内，行业整体回报高于市场基准指数 5%以上
	跟随大市	未来 6 个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来 6 个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的 6~12 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准（另有说明的除外）

联系方式

华福证券研究所　上海
公司地址：上海市浦东新区浦明路 1436 号陆家嘴滨江中心 MT 座 20 层
邮编：200120
邮箱：hfyjs@hfzq.com.cn