仅供内部参考，请勿外传！

万联证券 WANLIAN SECURITIES

大模型能力与成本协同优化，AI 商业化验证持续增强

——计算机行业跟踪报告

同步大市(维持)

2026 年 09 月 01 日

行业核心观点：

上周，沪深300下跌0.21%，创业板指下跌3.42%，申万计算机行业指数上涨1.38%，分别跑赢沪深300和创业板指1.59和4.80个百分点，在申万各一级行业中排名第20位。阿里千问AI平台上线Qwen3.8-Flash，并同步开源Qwen3.8-Flash-Next。据公司披露，其训练成本仅约为Qwen3.7-Plus的九分之一，并在多项编程及办公任务评测中超过Qwen3.7-Plus和Claude Opus 4.6。智谱同期上线并开源GLM-5.3-Flash，该模型是GLM-5系列首个原生多模态模型；模型综合能力超过GLM-5.2，API定价仅为GLM-5.3的十分之一，并通过视觉反馈强化前端开发、文档生成和专业办公任务中的自我检查及修改能力；进行了多项架构更新，在保持精准长上下文能力的同时，大幅降低长上下文服务成本。我们认为，阿里与智谱本次模型升级强化了大模型进入生产流程所需的经济性和任务交付能力：得益于架构更新，两款模型均在保持较强编程、办公及Agent能力的同时显著降低训练或调用成本，为长文本处理、代码开发和高频Agent任务提供更具性价比的模型底座；GLM-5.3-Flash通过引入视觉反馈和输出自检能力，将模型的任务边界由生成代码或文档内容延伸至检查界面效果、验证操作结果并交付完整成品，有望提高复杂任务的完成率及输出可靠性。两者共同降低开发者和企业部署Agent应用的成本与门槛，并推动模型厂商通过API调用、开发工具订阅及金融、办公等行业解决方案扩大商业化收入。MiniMax发布2026年中期业绩公告，上半年实现收入1.17亿美元，同比增长283.1%；其中，开放平台及其他基于AI的企业服务收入同比增长703.1%至7393万美元，占总收入的63.4%，AI原生产品收入同比增长100.9%至4264万美元。公司毛利率由上年同期的12.1%提升至17.9%，主要受益于基础设施效率改善；中国内地以外市场收入占比达到60.8%。同期公司研发开支同比增长138.8%至2.97亿美元，经调整净亏损同比扩大111.2%至2.93亿美元。我们认为，MiniMax开放平台及企业服务收入快速增长并成为主要收入来源，反映开发者API调用、Token套餐及企业客户需求正在加快转化为收入，收入结构由主要依赖AI原生产品逐步拓展至面向开发者和企业的模型服务。公司海外收入占比较高，表明国产模型厂商具备通过全球化产品及开发者平台扩大用户覆盖的潜力；毛利率改善则初步体现推理基础设施效率提升对模型服务盈利能力的正向作用。公司研发投入仍明显高于收入规模，当前仍处于以高研发投入和算力支出换取模型能力、用户规模及市场份额的阶段，后续还需关注收入增长的可持续性以及推理成本优化、商业化进程推进对盈利能力的提升。**中长期视角下，**继续聚焦AI产业、数据产业两大投资主线。

投资要点：

产业动态：（1）AI 大模型：8 月 28 日，腾讯混元发布并开源新一代大

行业相对沪深 300 指数表现



数据来源：聚源，万联证券研究所

相关研究

多模态模型加速进入生产工作流，AI 商业化进程持续提速

国产大模型加速走向专业工作流，AI 商业化闭环逐步形成

AI 产业规模加速扩张，模型服务商业化与多模态应用持续推进

分析师： **夏清莹**
执业证书编号：S0270520050001
电话：(0755) 8322 3620
邮箱：xiaqy1@wlzq.com.cn

仅供内部参考，请勿外传！

语言模型 Hy4 preview。**（2）AI 大模型：**千问 AI 平台上线 Qwen3.8-Flash，在编程、Agent 等真实任务中达到极致性价比。**（3）AI 大模型：**当地时间 8 月 27 日，谷歌宣布推出 Gemini Omni 1.1 Flash 视频生成 AI 模型，最高可生成 4K 视频。**（4）AI 大模型：**MiniMax 截至 2026 年 6 月 30 日止六个月，总收入同比增长 283.1%，由 30.4 百万美元增至 116.6 百万美元。**（5）AI 大模型：**8 月 26 日，智谱上线并开源 GLM-5.3-Flash（320B-A18B），这是 GLM-5 系列的首个原生多模态模型。

行业估值：从估值情况来看，申万计算机行业 2026 年 8 月 28 日 PE-TTM 为 119.10 倍，低于 2023-2025 年历史 PE-TTM 的均值 158.59 倍。

风险提示：中美科技摩擦；地缘政治风险；产业生态建设不及预期；应用落地不及预期；数据隐私安全风险；市场竞争加剧。

仅供内部参考，请勿外传！

正文目录

1 行业周观点.....4
2 市场行情回顾.....4
3 产业动态.....6
3.1 AI 大模型：腾讯混元 Hy4 preview 发布，稳居开源模型第一梯队.....6
3.2 AI 大模型：Qwen3.8-Flash 上线千问 AI 平台.....6
3.3 AI 大模型：谷歌推出 Gemini Omni 1.1 Flash 视频生成 AI 模型，最高 4K 分辨率.....7
3.4 AI 大模型：MiniMax 发布 2026 年中期业绩公告.....7
3.5 AI 大模型：GLM-5.3-Flash，前沿智能进入普惠时代.....7
4 风险提示.....8

图表 1： 沪深 300、创业板指、申万一级行业指数涨跌幅（%）（上周）.....5
图表 2： 沪深 300、创业板指、申万一级行业指数涨跌幅（%）（2026 年初至 2026 年 8 月 28 日）.....5
图表 3： 申万计算机行业估值情况（PE-TTM，单位：倍）.....6

仅供内部参考，请勿外传！

1 行业周观点

阿里千问AI平台上线Qwen3.8-Flash，并同步开源Qwen3.8-Flash-Next。据公司披露，其训练成本仅约为Qwen3.7-Plus的九分之一，并在多项编程及办公任务评测中超过Qwen3.7-Plus和Claude Opus 4.6。智谱同期上线并开源GLM-5.3-Flash，该模型是GLM-5系列首个原生多模态模型；模型综合能力超过GLM-5.2，API定价仅为GLM-5.3的十分之一，并通过视觉反馈强化前端开发、文档生成和专业办公任务中的自我检查及修改能力；进行了多项架构更新，在保持精准长上下文能力的同时，大幅降低长上下文服务成本。我们认为，阿里与智谱本次模型升级强化了大模型进入生产流程所需的经济性和任务交付能力：得益于架构更新，两款模型均在保持较强编程、办公及Agent能力的同时显著降低训练或调用成本，为长文本处理、代码开发和高频Agent任务提供更具性价比的模型底座；GLM-5.3-Flash通过引入视觉反馈和输出自检能力，将模型的任务边界由生成代码或文档内容延伸至检查界面效果、验证操作结果并交付完整成品，有望提高复杂任务的完成率及输出可靠性。两者共同降低开发者和企业部署Agent应用的成本与门槛，并推动模型厂商通过API调用、开发工具订阅及金融、办公等行业解决方案扩大商业化收入。MiniMax发布2026年中期业绩公告，上半年实现收入1.17亿美元，同比增长283.1%；其中，开放平台及其他基于AI的企业服务收入同比增长703.1%至7393万美元，占总收入的63.4%，AI原生产品收入同比增长100.9%至4264万美元。公司毛利率由上年同期的12.1%提升至17.9%，主要受益于基础设施效率改善；中国内地以外市场收入占比达到60.8%。同期公司研发开支同比增长138.8%至2.97亿美元，经调整净亏损同比扩大111.2%至2.93亿美元。我们认为，MiniMax开放平台及企业服务收入快速增长并成为主要收入来源，反映开发者API调用、Token套餐及企业客户需求正在加快转化为收入，收入结构由主要依赖AI原生产品逐步拓展至面向开发者和企业的模型服务。公司海外收入占比较高，表明国产模型厂商具备通过全球化产品及开发者平台扩大用户覆盖的潜力；毛利率改善则初步体现推理基础设施效率提升对模型服务盈利能力的正向作用。公司研发投入仍明显高于收入规模，当前仍处于以高研发投入和算力支出换取模型能力、用户规模及市场份额的阶段，后续还需关注收入增长的可持续性以及推理成本优化、商业化进程推进对盈利能力的提升。

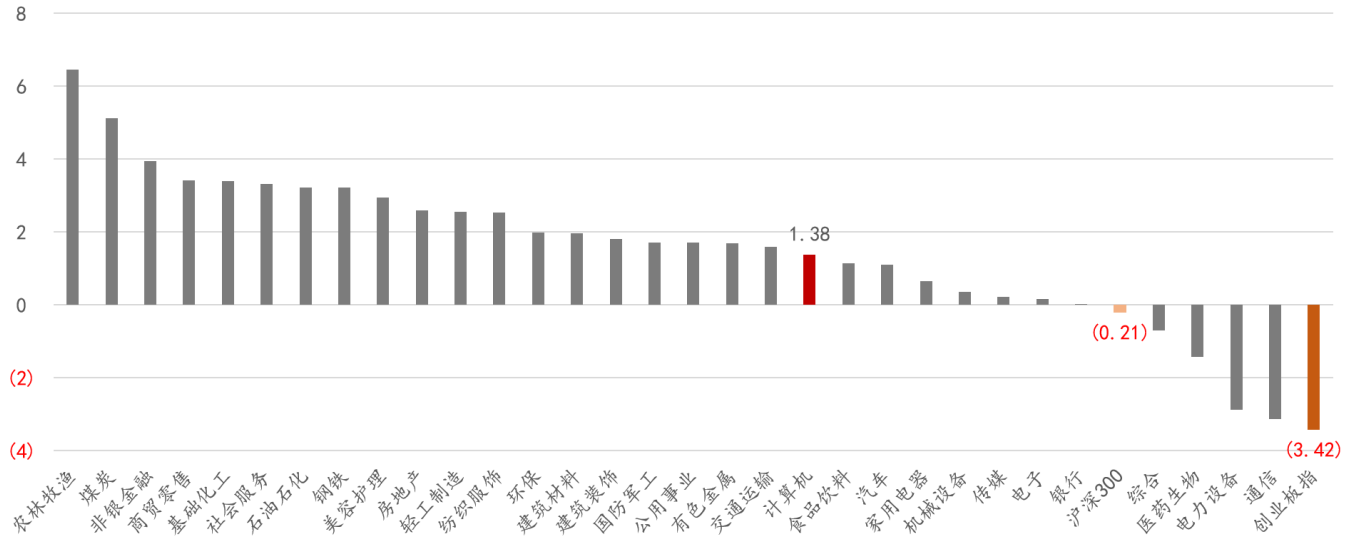
中长期视角下，继续聚焦AI产业、数据产业两大投资主线。1）AI产业：主要关注国内AI大模型的技术升级和产品迭代、头部AI大模型厂商的生态建设、智能驾驶、具身智能等新型智能终端及智能体的应用落地进程，以及人工智能治理的生态建设。2）数据产业：主要关注数据基础设施建设、高质量数据集等数据要素资源的开发利用、数据要素的价值释放及应用落地以及数据隐私安全。

2 市场行情回顾

申万计算机行业跑赢沪深300和创业板指，周涨跌幅位于申万31个一级行业的中下游。上周，沪深300下跌0.21%，创业板指下跌3.42%，申万计算机行业指数上涨1.38%，分别跑赢沪深300和创业板指1.59和4.80个百分点，在申万各一级行业中排名第20位。

仅供内部参考，请勿外传！

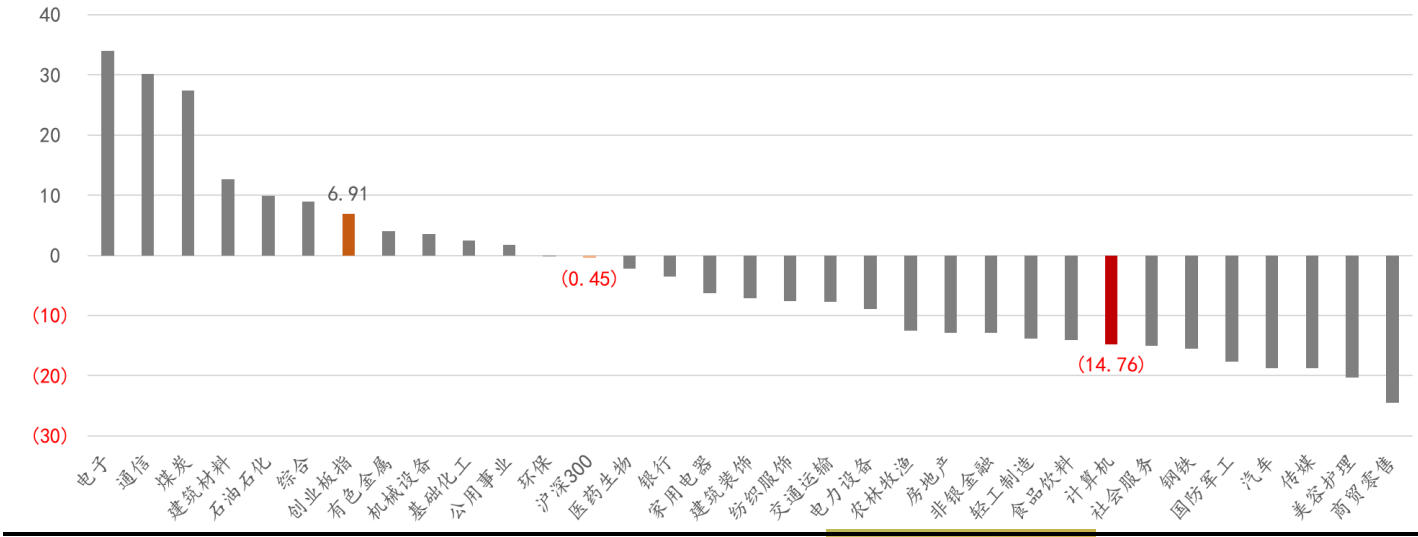
图表1：沪深300、创业板指、申万一级行业指数涨跌幅（%）（上周）



资料来源：iFinD，万联证券研究所

申万计算机行业2026年初至8月28日累计下跌14.76%，位于申万31个一级行业的中下游。2026年初至8月28日，沪深300下跌0.45%，创业板指上涨6.91%，申万计算机行业指数下跌14.76%，分别跑输沪深300和创业板指14.31和21.66个百分点，在申万各一级行业中排名第24位。

图表2：沪深300、创业板指、申万一级行业指数涨跌幅（%）（2026年初至2026年8月28日）



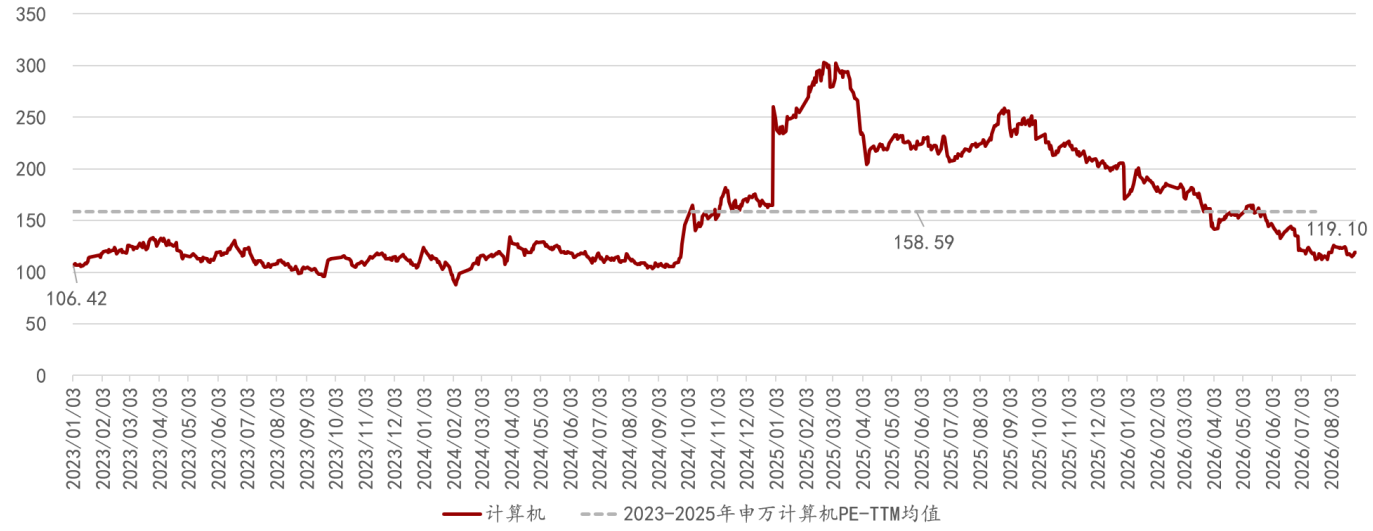
资料来源：iFinD，万联证券研究所

行业估值水平低于历史中枢水平。从估值情况来看，申万计算机行业2026年8月28日PE-TTM为119.10倍，低于2023-2025年历史PE-TTM的均值158.59倍。

人在草木间

仅供内部参考，请勿外传！

图表3：申万计算机行业估值情况（PE-TTM，单位：倍）



资料来源：iFinD，万联证券研究所

3 产业动态

3.1 AI 大模型：腾讯混元 Hy4 preview 发布，稳居开源模型第一梯队

8月28日，腾讯混元发布并开源新一代大语言模型Hy4 preview，总参数770B、激活参数49B，上下文长度突破1M，在代码、办公、科学等真实生产力任务上展现出卓越能力。Hy4 preview在模型尺寸、上下文长度、数据规模上都进行了显著的扩展，预训练和后训练的共同进步带来了智能水平的又一次巨大提升，稳居开源模型第一梯队。混元持续与CodeBuddy/WorkBuddy等产品深度协同，优化生产力场景的真实用户体验。在腾讯内部组织的163名内部专家、203个工程任务的盲测中，Hy4 preview（均分2.99/4）略优于GLM 5.3（均分2.92/4.00）和Kimi K3（均分2.94/4.00）。Hy4 preview定位“为生产力而生”的模型，通过与软件工程、游戏、金融、安全等顶尖腾讯专家的高质量数据共建，以及与WorkBuddy等产品的深度协同（Co-Design），模型在各类真实生产力任务上取得显著进步。（资料来源：“腾讯技术工程”公众号）

3.2 AI 大模型：Qwen3.8-Flash 上线千问 AI 平台

千问AI平台上线Qwen3.8-Flash，在编程、Agent等真实任务中达到极致性价比。

最新价格如下：

输入每百万Tokens0.8元，输出每百万Tokens2.7元，缓存命中每百万Tokens0.1元。Qwen3.8-Flash是一个多模态混合专家（MoE）模型，采用了下一代（Next）模型架构，千亿总参数仅激活60亿（6B），创下模型效率的全球新基准。

模型采用了与此前所有千问大模型完全不同的全新架构：

在注意力方面，引入独特的QSA（Qwen Sparse Attention）机制，形成与GDN高效融合的混合注意力架构，可显著降低计算开销，在高缓存命中的1M Tokens长上下文实际场景中可提速8倍以上，又准又快；在模型内部分层通信方面，引入自研的Gated Residual方法，将传统Transformer的1条信息主通道拆分并拓展为4条，让模型可根据需求动态决定读写信息，信息传递更高效，训练也更稳定；在模型参数扩展方面，引入51B的N-gram Embedding参数，为模型Transformer参数提供更高效的查表寻址，在每次token计算时无需参与，以极少的资源消耗“外挂”了一个大规模的“记忆指南手册”，模型参数扩展效率更高；在模型调参方面，模型结构和后期优化协同进行，

仅供内部参考，请勿外传！

收敛效率和训练吞吐大幅提升。（资料来源：“阿里云”公众号）

3.3 AI 大模型：谷歌推出 Gemini Omni 1.1 Flash 视频生成 AI 模型，最高 4K 分辨率

当地时间8月27日，谷歌宣布推出Gemini Omni 1.1 Flash视频生成AI模型，最高可生成4K视频。

新模型亮点如下：

场景扩展：用户可以直接基于已有视频，从结尾处无缝衔接并继续生成后续画面。用户可按每次10秒的步长逐步扩展视频，单条视频的累计最长时长可达40秒；

指定首尾帧：只需指定镜头的起始帧与结束帧，即可实现平滑自然的转场效果与镜头运镜；

生成360p预览：相较于Omni 1.1的标准720p分辨率，生成速度最高可提升60%，且成本仅需原有的三分之一，适用于快速原型验证、分镜脚本迭代以及高速渲染；

最高4K分辨率：可直接输出画质细腻、细节完善的高分辨率1080p或4K最终成片；视频参考：构建视频场景时，支持引入最长3秒的参考视频片段，从而依据参考内容精准维持画面的上下文背景与角色一致性。

价格方面，Gemini Omni 1.1 Flash生成360p视频每秒0.03美元（现汇率约合0.2元人民币），生成720p视频每秒0.1美元（现汇率约合0.67元人民币），生成1080p视频每秒0.15美元（现汇率约合1元人民币），生成4K视频每秒0.3美元（现汇率约合2元人民币）。（资料来源：IT之家）

3.4 AI 大模型：MiniMax 发布 2026 年中期业绩公告

于报告期间，我们发布MiniMax M3，对核心模型产品进行了升级，进一步加强我们在编程、Agentic工作流程及专业工作等方面的能力。之后不久，我们亦发布具有开源权重的MiniMax H3，推进用于商业创作的视频生成技术，并拓宽企业部署与开发者创新的路径。随着推理及Agentic工作负载需求持续增长，开放平台服务的企业客户与开发者群体不断扩大，其对业务的推动作用亦逐渐增强。我们持续深化全球布局，以日益强大且具成本效益的智能产品，为230多个国家与地区的企业客户、开发者及个人用户提供服务。

截至2026年6月30日止六个月，我们的总收入同比增长283.1%，由30.4百万美元增至116.6百万美元，超过2025年全年79.0百万美元的总收入。这一增长反映出我们全球客户及用户群不断扩大、对模型推理的需求快速增长，以及我们能够将模型能力优势转化为全球企业、开发者及个人用户所使用的产品与服务。

开放平台及其他基于AI的企业服务收入同比增长703.1%，由9.2百万美元增至73.9百万美元，占总收入的63.4%，而2025年同期占比为30.3%。该增长乃归因于付费用户及企业客户数量增长、API调用量增加以及我们的Token计划迅速获采用。这一表现证明了在实际业务场景中对我们模型的需求日益增长，以及企业及开发者

工作负载对我们业务的贡献不断增加。AI原生产品收入同比增长100.9%，由21.2百万美元增至42.6百万美元，主要得益于用户参与度提升、付费意愿增强以及海螺AI及其他AI原生产品持续商业化。我们持续推出新产品组合及工具产品，使用户能更直接地将前沿模型能力应用于生产力。（资料来源：MiniMax财报）

3.5 AI 大模型：GLM-5.3-Flash，前沿智能进入普惠时代

仅供内部参考，请勿外传！

8月26日，智谱上线并开源GLM-5.3-Flash（320B-A18B），这是GLM-5系列的首个原生多模态模型。大家好像已经习惯将前沿智能与前沿价格绑定，并认为这是技术进步必须支付的成本。今天，GLM-5.3-Flash来了：320B总参数，能力超过GLM-5.2，在全球权威的Artificial Analysis Intelligence Index（AA综合智能指数）中取得57分，进入全球前沿模型能力区间，与Anthropic最受欢迎的模型Claude Opus 4.8得分持平。在自研Z.ai Code Bench体感评估中，其编程表现与Claude Opus 4.8相当。与此同时，GLM-5.3-Flash定价为GLM-5.3的1/10，限时折扣内为GLM-5.3的1/20，为Opus 4.8的1/40。同样智力，1/40价格，前沿智能，第一次不需要省着用。这得益于其全新模型架构。GLM-5.3-Flash的架构专为极低成本而设计，总参数量与GLM-4.5相当（355B vs. 320B），但激活参数量（32B→18B）与层数（92→45）几乎减半。结合智谱最新的30T Token多模态预训练语料，GLM-5.3-Flash能够以更少的计算资源实现更强的性能。GLM-5.3-Flash同时进行了多项架构升级，是首个采用稀疏注意力与线性注意力混合架构的开源前沿模型，在保持精准长上下文能力的同时，大幅降低长上下文服务成本；并采用流形约束超连接（Manifold-Constrained Hyper-Connections，mHC）提升模型Scaling能力。（资料来源：“智谱”公众号）

4 风险提示

中美科技摩擦；地缘政治风险；产业生态建设不及预期；应用落地不及预期；数据隐私安全风险；市场竞争加剧。

仅供内部参考，请勿外传！

行业投资评级

强于大市：未来6个月内行业指数相对大盘涨幅10％以上；
同步大市：未来6个月内行业指数相对大盘涨幅10％至-10％之间；
弱于大市：未来6个月内行业指数相对大盘跌幅10％以上。

公司投资评级

买入：未来6个月内公司相对大盘涨幅15％以上；
增持：未来6个月内公司相对大盘涨幅5％至15％；
观望：未来6个月内公司相对大盘涨幅-5％至5％；
卖出：未来6个月内公司相对大盘跌幅5％以上。
基准指数 ： 沪深300指数

风险提示

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。

证券分析师承诺

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登记为证券分析师，以勤勉的执业态度，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。

免责条款

万联证券股份有限公司（以下简称“本公司”）是一家覆盖证券经纪、投资银行、投资管理和证券咨询等多项业务的全国性综合类证券公司。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。本报告中的信息或所表述的意见并未考虑到个别投资者的具体投资目的、财务状况以及特定需求。客户应自主作出投资决策并自行承担投资风险。本公司不对任何人因使用本报告中的内容所导致的损失负任何责任。在法律许可情况下，本公司或其关联机构可能会持有报告中提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或类似的金融服务。
市场有风险，投资需谨慎。本报告是基于本公司认为可靠且已公开的信息撰写，本公司力求但不保证这些信息的准确性及完整性，也不保证文中的观点或陈述不会发生任何变更。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。分析师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。
本报告的版权仅为本公司所有，未经书面许可任何机构和个人不得以任何形式翻版、复制、刊登、发表和引用。未经我方许可而引用、刊发或转载的引起法律后果和造成我公司经济损失的概由对方承担，我公司保留追究的权利。

万联证券股份有限公司 研究所
上海浦东新区世纪大道 1528 号陆家嘴基金大厦
北京西城区平安里西大街 28 号中海国际中心
深圳福田区深南大道 2007 号金地中心
广州天河区珠江东路 11 号高德置地广场