证券研究报告

2024 年 07 月 21 日

华鑫证券 CHINA FORTUNE SECURITIES

# OpenAI 发布 GPT-4o mini，引领大模型普及时代

—计算机行业周报

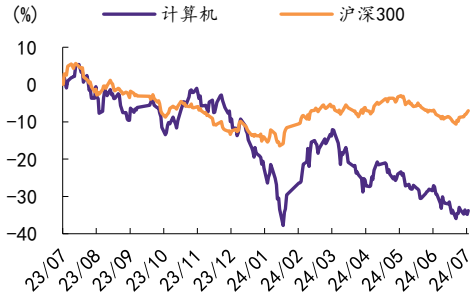
## 推荐(维持)

分析师：宝幼琛 S1050521110002
baoyc@cfsc.com.cn

### 行业相对表现

| 表现 | 1M | 3M | 12M |
|---|---|---|---|
| 计算机(申万) | -6.2 | -9.1 | -33.3 |
| 沪深 300 | 1.2 | -0.1 | -7.4 |

### 市场表现




资料来源：Wind，华鑫证券研究

### 相关研究

1、《计算机行业周报：AI 助理包围 WAIC2024，共建共治智能体生态》2024-07-14

2、《计算机行业周报：商汤发布流式多模态大模型日日新 5.5，国内首次全面对标 GPT-4o》2024-07-07

3、《计算机行业周报：首款 Transformer 专用芯片 Sohu 亮相，10 倍于 B200 速度成为最快 AI 芯片》2024-06-30

## 投资要点

### AI 应用：OpenAI 发布 GPT-4o mini，引领大模型普及时代

7 月 19 日，OpenAI 官宣推出 GPT-3.5 Turbo 替代品—GPT-4o mini，即 GPT-4o 更小参数量的简化版本。ChatGPT 的免费用户、Plus 用户和 Team 用户都能够使用 GPT-4o mini。下周，企业版客户也将获得使用 GPT-4o mini 的权限。 GPT-4o mini 拥有低成本和快速响应能力，适用于多种应用场景。GPT-4o mini 支持需要连续或同时调用多个模型的应用程序，处理大量上下文信息，以及通过快速实时的文本回复与客户进行互动，能够处理多达 128K token 的长上下文，且对非英文内容的支持更友好。GPT-4o mini 与业内其他模型对比优势明显：1）在基准测试中， GPT-4o mini 在推理基准结果 MMLU 上得分为 82%，而 Gemini Flash 为 77.9%，此前主打极高性价比的 Claude Haiku 为 73.8%；同时 GPT-4o mini 在数学推理和编程任务方面表现出色，超越市场上其他小模型；2）GPT-4o mini 多模态推理测评表现结果优秀，得分超过 Gemini Flash 和 Claude Haiku。同时，GPT-4o mini 价格也大幅下降， GPT-4o mini 每 100 万输入 token 价格为 15 美分，每 100 万输出 token 价格为 60 美分，比 GPT-3.5 Turbo 便宜超过 60%。随着 GPT-4o Mini 的普及，GPT-3.5 正逐渐成为历史，预示着人工智能技术的新发展阶段。

### AI 融资动向：World Labs 获天使轮融资 1 亿美元

本期 AI 初创公司的融资中，World Labs/ Vectara 融资额前二，分别为 1 亿/0.25 亿美元。World Labs 主要利用类似人类的视觉数据处理技术，使 AI 具备高级推理能力，将尝试通过开发类似人类的视觉数据处理技术，在 AI 中创造“空间智能”。Vectara 提供一个端到端的生成式人工智能平台，专注于检索增强生成技术，该平台旨在为受监管行业（如健康、法律和制作业）提供一种安全、可靠、可信赖的 AI 解决方案。 这笔资金用于推进检索增强生成技术，加强内部创新，提高市场推广资源，并拓展其他地区的业务。

### 投资建议

行业研究

**GB200 出货量预期持续上调，AI 应用需求乐观。**近期美联储释放转鸽信号，导致美股科技股板块有所调整。但基本面角度看，海外大厂都将导入英伟达 Blackwell 架构 GPU 打造 AI 服务器，量能持续超预期。英伟达扩大 Blackwell 架构 GPU 投片量之际，就终端整机服务器机柜数量来看，包括 GB200 NVL72 及 GB200 NVL36 服务器机柜出货量同步大增，由原预期合并出货 4 万台，大增至 6 万台，增幅高达五成，当中以 GB200 NVL36 总量达 5 万台为数最多。

建议关注关注以 AI 为核心的龙头厂商科大讯飞（002230.SZ）、有望迎来需求爆发的 AI 应用金桥信息（603918.SH）、高速通信连接器业务或显著受益于 GB200 放量的鼎通科技（688668.SH）。

## 风险提示

1）AI 底层技术迭代速度不及预期。2）政策监管及版权风险。3）AI 应用落地效果不及预期。4）推荐公司业绩不及预期风险。

| 公司代码 | 名称 | 2024-07-21 股价 | EPS | | | PE | | | 投资评级 |
|---|---|---|---|---|---|---|---|---|---|
| | | | 2023 | 2024E | 2025E | 2023 | 2024E | 2025E | |
| 002230.SZ | 科大讯飞 | 39.30 | 0.28 | 0.40 | 0.56 | 140.36 | 98.25 | 70.18 | 买入 |
| 002368.SZ | 太极股份 | 17.14 | 0.79 | 1.01 | 1.28 | 21.70 | 16.97 | 13.39 | 买入 |
| 603918.SH | 金桥信息 | 9.37 | 0.33 | 0.49 | 0.80 | 28.39 | 19.12 | 11.71 | 买入 |
| 688668.SH | 鼎通科技 | 36.55 | 0.67 | 1.04 | 1.41 | 54.55 | 35.14 | 25.92 | 买入 |

资料来源：Wind，华鑫证券研究

# 正文目录



# 图表目录

# 1、 算力动态：算力租赁价格平稳

## 1.1、 数据跟踪：算力租赁价格平稳

**本周算力租赁价格环比持平。**具体来看，显卡配置为 A100-40G 中，腾讯云 16 核+96G 价格为 28.64 元/时，阿里云 12 核+94GiB 价格为 31.58 元/时；显卡配置为 A100-80G 中，恒源云 13 核+128G 价格为 8.50 元/时;阿里云 16 核+125GiB 价格为 34.74 元/时；显卡配置为 A800-80G 中，恒源云 16+256G 的租赁较为紧张。

**图表 1：本周算力租赁情况**

| 显卡配置 | CPU | 内存 | 磁盘大小（G） | 平台名称 | 价格（每小时） | 价格环比上周 |
|---|---|---|---|---|---|---|
| A100-40G | 16 | 96 | 可自定，额外收费 | 腾讯云 | 28.64/元 | 0.00% |
| | 12 核 | 94G | 可自定，额外收费 | 阿里云 | 31.58/元 | 0.00% |
| A100-80G | 13 | 128 | 系统盘：20G<br>数据盘：50GB | 恒源云 | 8.50/元 | 0.00% |
| | 16 核 | 125G | 可自定，额外收费 | 阿里云 | 34.74/元 | 0.00% |
| A800-80G | 16 | 256 | 系统盘：20G<br>数据盘：50GB | 恒源云 | - | - |

资料来源：腾讯云，阿里云，恒源云，华鑫证券研究

# 2、 AI 应用动态：OpenAI 发布 GPT-4o mini，引领大模型普及时代

## 2.1、 流量跟踪：文心一言访问量环比增长+13.15%

本期（2024.7.10-2024.7.16）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（638.9M）、Bing（310.4M）和 Discord（265M）；访问量环比增速前三位分别为文心一言（13.15%）、Perplexity（4.05%）和 Deepl（2.33%）；平均停留时长前三位分别为 Character.AI（0:15:54）、Canva（0:09:43）和 Notiona AI（0:08:21）;平均停留时长环比增速前两位分别为文心一言(15.91%)、Perplexity（2.48%）和 Kimi（0.63%）。

**图表 2：2024.7.10-2024.7.16 AI 相关网站流量**

| 应用 | 应用类型 | 归属公司 | 周平均访问量（M） | 访问量环比 | 平均停留时长 | 时长环比 |
|---|---|---|---|---|---|---|
| ChatGPT | 聊天机器人 | OpenAI | 638.9 | 2.01% | 6:28 | -0.77% |
| Bing | 搜索 | 微软 | 310.4 | -0.06% | 6:46 | -0.73% |

| | | | | | | |
|---|---|---|---|---|---|---|
| Discord | 游戏社区 | 微软 | 265 | -0.90% | 6:24 | 0.00% |
| Canva | 在线设计 | Canva | 121.3 | 1.25% | 9:43 | -9.75% |
| Github | 代码托管 | 微软 | 103.5 | 0.19% | 7:18 | -0.45% |
| Gemini | 聊天机器人 | 谷歌 | 65.33 | 0.03% | 5:26 | 0.31% |
| Character.AI | 聊天机器人 | Character.AI | 70.01 | -3.98% | 15:54 | 0.32% |
| NotionAI | 文本/笔记 | Notion | 36.91 | -0.43% | 8:21 | 0.20% |
| QuillBot | 释义工具 | QuillBot | 10.43 | -2.16% | 5:21 | 0.63% |
| Kimi | 聊天机器人 | Moonshot AI | 5.504 | 1.93% | 2:55 | -0.57% |
| DeepI | 翻译工具 | DeepL | 58.39 | 2.33% | 8:16 | -0.80% |
| 文心一言 | 聊天机器人 | 百度 | 3.57 | 13.15% | 1:42 | 15.91% |
| Perplexity | AI 搜索 | Perplexity | 19.25 | 4.05% | 6:53 | 2.48% |

资料来源：similarweb,华鑫证券研究

## 2.2、产业动态：OpenAI 发布 GPT-4o mini，引领大模型普及时代

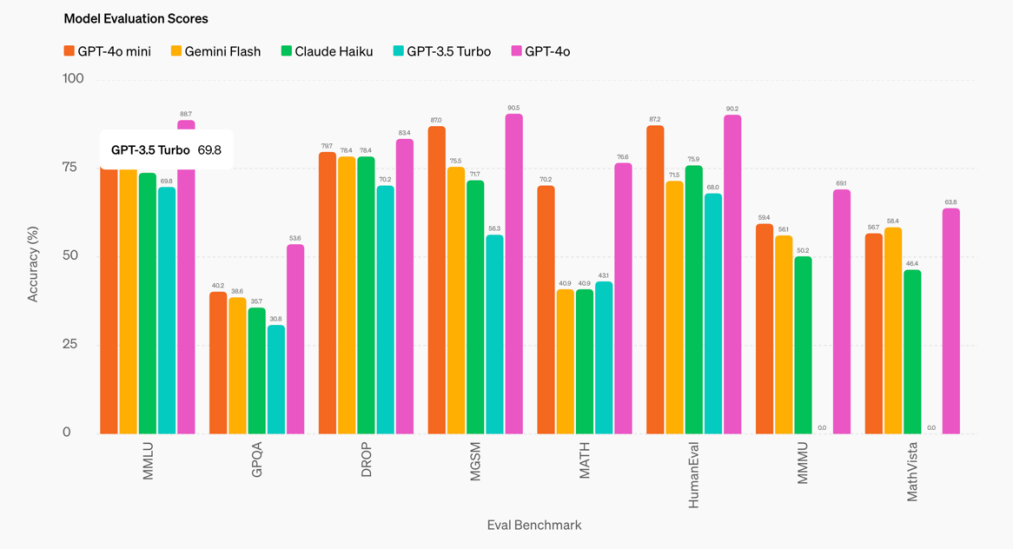
7 月 19 日，OpenAI 官宣推出 GPT-3.5 Turbo 替代品—GPT-4o mini，即 GPT-4o 更小参数量的简化版本。ChatGPT 的免费用户、Plus 用户和 Team 用户都能够使用 GPT-4o mini。下周，企业版客户也将获得使用 GPT-4o mini 的权限。

GPT-4o mini 拥有低成本和快速响应能力，适用于多种应用场景。GPT-4o mini 支持需要连续或同时调用多个模型的应用程序，处理大量上下文信息，以及通过快速实时的文本回复与客户进行互动。目前，GPT-4o mini 已在 API 中提供了文本和图像处理功能，后续还将逐步增加对视频和音频的支持。该模型能够处理多达 128K token 的长上下文，知识库截止日期为 2023 年 10 月，且对非英文内容的支持更友好。

其中 GPT-4o mini 部分性能测试结果如下：

**1）基准测试结果：** GPT-4o mini 在推理基准结果 MMLU 上得分为 82%，而 Gemini Flash 为 77.9%，此前主打极高性价比的 Claude Haiku 为 73.8%。GPT-4o mini 在数学推理和编程任务方面也同样表现出色，超越市场上的其他小型模型。在 MGSM 数学推理能力基准测试中，GPT-4o mini 得分达到了 87.0%，而 Gemini Flash 的得分为 75.5%，Claude Haiku 的得分为 71.7%。GPT-4o mini 在 HumanEval 基准测试中同样再次展现优势，得分达到 87.2%，而 Gemini Flash 的得分为 71.5%，Claude Haiku 的得分为 75.9%。

**图表 3：GPT-4o mini 模型测评对比**

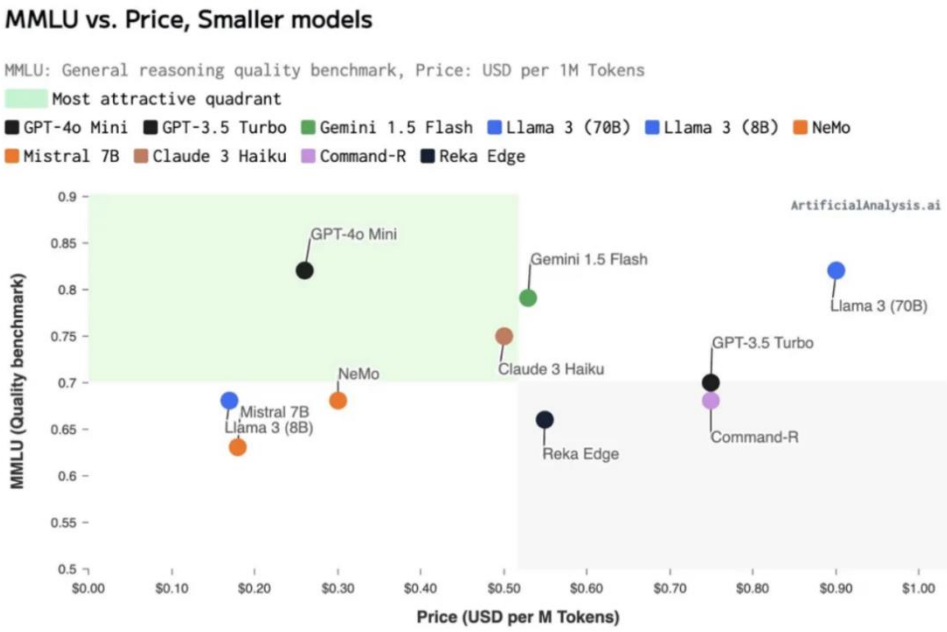



资料来源：OpenAI，华鑫证券研究

**2）多模态推理测评：**在多模态推理 MMMU 中，Gemini Flash 得分为 56.1%，Claude Haiku 得分为 50.2%，而 GPT-4o mini 得分为 59.4%，性能优越。此外，GPT-4o mini 在大模型盲测竞技场 LMSYS 中的表现也优于 GPT-4T 01-25。

**3）GPT-4o mini 价格降低：**除性能增强外，GPT-4o mini 的价格也大幅下降。GPT-4o mini 每 100 万输入 token 价格为 15 美分，每 100 万输出 token 价格为 60 美分，比 GPT-3.5 Turbo 便宜超过 60%。GPT-4o mini 生成一本 2500 页的书，价格只需 60 美分。

**图表 4：MMLU 测试结果 vs 价格**




资料来源：APPSO，华鑫证券研究

**GPT-4o Mini 引领大模型普及时代，GPT-3.5 渐成过往。**GPT-4o Mini 的推出预示着大模型技术的大面积普及，以更低的成本门槛，使得更广泛的用户能够享受到先进的 AI 服务。在安全性上，OpenAI 通过与社会心理学和错误信息研究等领域的专家的合作，对模型进行了严格的风险评估和优化，同时引入了创新技术以强化模型的防御能力。随着 GPT-4o Mini 的普及，GPT-3.5 正逐渐成为历史，预示着人工智能技术的新发展阶段。

# 3、 AI 融资动向：World Labs 获天使轮融资 1 亿美元

本期 AI 初创公司的融资中，World Labs/ Vectara 融资额前二，分别为 1 亿/0.25 亿美元。World Labs 主要利用类似人类的视觉数据处理技术，使 AI 具备高级推理能力，将尝试通过开发类似人类的视觉数据处理技术，在 AI 中创造“空间智能”。Vectara 提供一个端到端的生成式人工智能平台，专注于检索增强生成技术，该平台旨在为受监管行业（如健康、法律和制作业）提供一种安全、可靠、可信赖的 AI 解决方案。 这笔资金用于推进检索增强生成技术，加强内部创新，提高市场推广资源，并拓展其他地区的业务。

**图表 5：本周 AI 初创公司的融资动态**

| 应用 | 应用类型 | 领投方 | 融资轮 | 融资额 | 目前累计融资额 | 目前估值 |
|---|---|---|---|---|---|---|
| 上海感图网络科技有限公司 | AI 高端制造 | 地方政府产业基金 | C2 轮 | 数亿元 | —— | —— |
| 耀速科技 | AI 生物科技 | 鼎泰集团 | 天使+轮 | 亿元级人民币元 | —— | —— |
| Vectara | 企业生成式 AI 平台 | FPV Ventures Race Capital | A 轮 | 2500 万美元 | —— | —— |
| Mira | 去中心化 AI 基础平台 | BITKRAFT Ventures Framework Ventures | 种子轮 | 900 万美元 | 900 万 | —— |
| World Labs | AI 空间智能 | Andreessen Horowitz | 天使轮 | 1 亿美元 | —— | 10 亿美元 |
| Adaptive | AI 财务自动化平台 | Emergence Capital | A 轮 | 1900 万美元 | —— | —— |

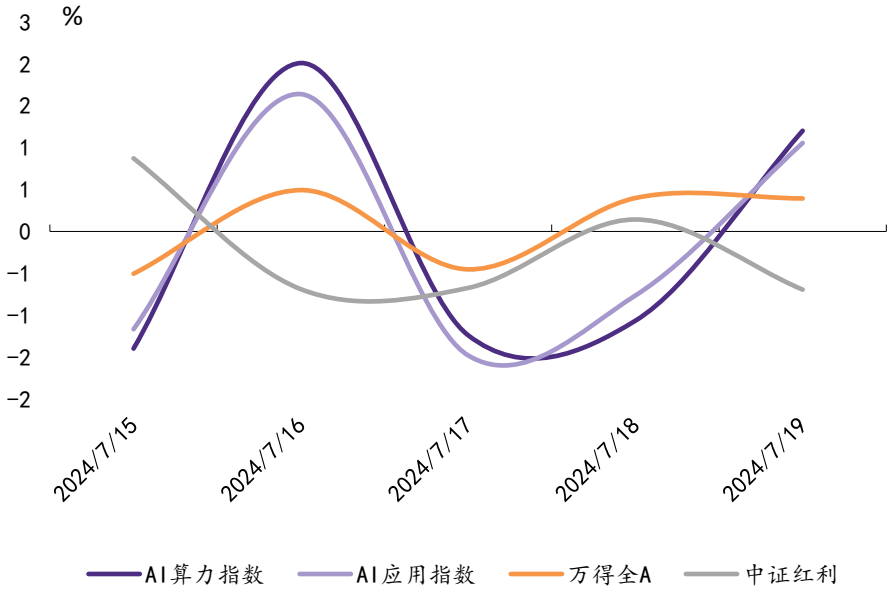
资料来源：投资界，36 氪，ai 工具集，华鑫证券研究

# 4、 行情复盘

本周，AI 算力指数/AI 应用指数/万得全 A/中证红利日涨幅最大值分别为

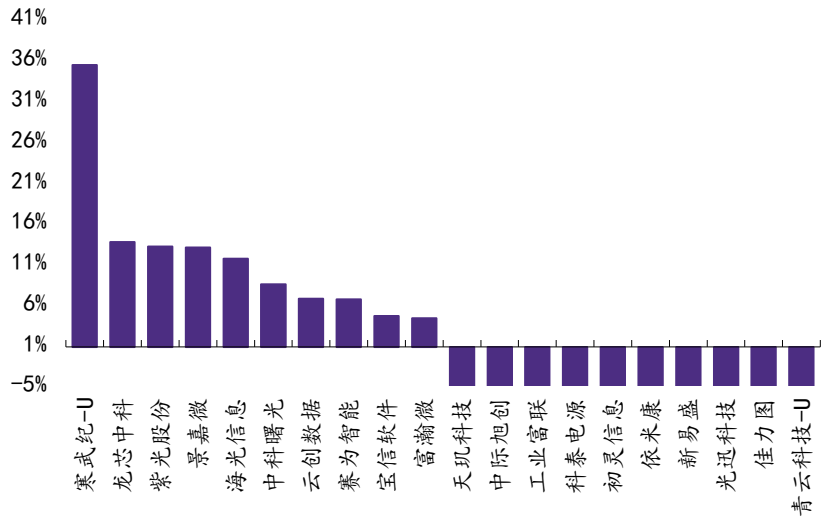
2.01%/1.64%/0.494%/0.87%，日跌幅最大值分别为-1.4%/-1.48%/-0.50%/-0.65%。AI 算力指数内部，寒武纪-U 以+34.60%录得本周最大涨幅，天玑科技以-13.99%录得本周最大跌幅。AI 应用指数内部，寒武纪-U 以+34.60%录得本周最大涨幅，淳中科技以-15.27%录得本周最大跌幅。

**图表 6：本周指数日涨跌幅**



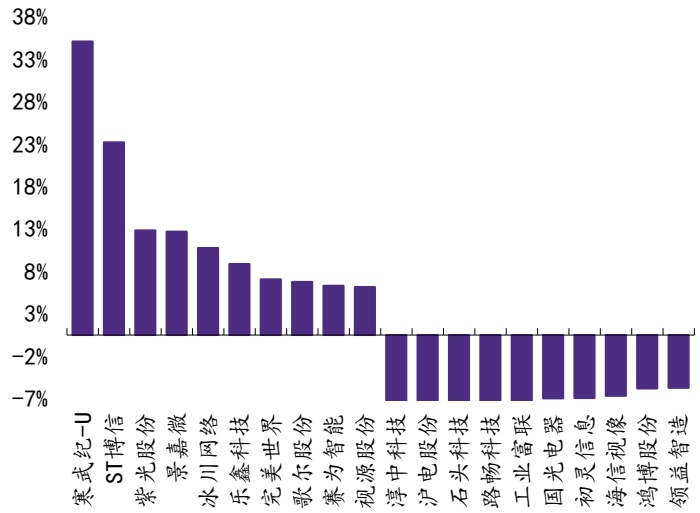

资料来源：wind，华鑫证券研究

**图表 7：本周 AI 算力指数内部涨跌幅度排名**




资料来源：wind，华鑫证券研究

**图表 8：本周 AI 应用指数内部涨跌幅度排名**




资料来源：wind,华鑫证券研究

# 5、 投资建议

**GB200 出货量预期持续上调，AI 应用需求乐观。**近期美联储释放转鸽信号，导致美股科技股板块有所调整。但基本面角度看，海外大厂都将导入英伟达 Blackwell 架构 GPU 打造 AI 服务器，量能持续超预期。英伟达扩大 Blackwell 架构 GPU 投片量之际，就终端整机服务器机柜数量来看，包括 GB200 NVL72 及 GB200 NVL36 服务器机柜出货量同步大增，由原预期合并出货 4 万台，大增至 6 万台，增幅高达五成，当中以 GB200 NVL36 总量达 5 万台为数最多。

建议关注关注以 AI 为核心的龙头厂商科大讯飞（002230.SZ）、有望迎来需求爆发的 AI 应用金桥信息（603918.SH）、高速通信连接器业务或显著受益于 GB200 放量的鼎通科技（688668.SH）。

**图表 9：重点关注公司及盈利预测**

| 公司代码 | 名称 | 2024-07-21 | EPS | | | PE | | | 投资评级 |
|---|---|---|---|---|---|---|---|---|---|
| | | 股价 | 2023 | 2024E | 2025E | 2023 | 2024E | 2025E | |
| 002230.SZ | 科大讯飞 | 39.30 | 0.28 | 0.40 | 0.56 | 140.36 | 98.25 | 70.18 | 买入 |
| 002368.SZ | 太极股份 | 17.14 | 0.79 | 1.01 | 1.28 | 21.70 | 16.97 | 13.39 | 买入 |
| 603918.SH | 金桥信息 | 9.37 | 0.33 | 0.49 | 0.80 | 28.39 | 19.12 | 11.71 | 买入 |
| 688668.SH | 鼎通科技 | 36.55 | 0.67 | 1.04 | 1.41 | 54.55 | 35.14 | 25.92 | 买入 |

资料来源：wind,华鑫证券研究

# 6、 风险提示

1）AI 底层技术迭代速度不及预期。2）政策监管及版权风险。3）AI 应用落地效果不及预期。4）推荐公司业绩不及预期风险。

## 计算机&中小盘组介绍

**宝幼琛：**本硕毕业于上海交通大学，多次新财富、水晶球最佳分析师团队成员，7 年证券从业经验，2021 年 11 月加盟华鑫证券研究所，目前主要负责计算机与中小盘行业上市公司研究。擅长领域包括：云计算、网络安全、人工智能、区块链等。

**任春阳：**华东师范大学经济学硕士，6 年证券行业经验，2021 年 11 月加盟华鑫证券研究所，从事计算机与中小盘行业上市公司研究

**周文龙：**澳大利亚莫纳什大学金融硕士

**陶欣怡：**毕业于上海交通大学，于 2023 年 10 月加入团队。

## 证券分析师承诺

本报告署名分析师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。

## 证券投资评级说明

股票投资评级说明：

| | 投资建议 | 预测个股相对同期证券市场代表性指数涨幅 |
|---|---|---|
| 1 | 买入 | ＞ 20% |
| 2 | 增持 | 10% — 20% |
| 3 | 中性 | -10% — 10% |
| 4 | 卖出 | ＜ -10% |

行业投资评级说明：

| | 投资建议 | 行业指数相对同期证券市场代表性指数涨幅 |
|---|---|---|
| 1 | 推荐 | ＞ 10% |
| 2 | 中性 | -10% — 10% |
| 3 | 回避 | ＜ -10% |

以报告日后的 12 个月内，预测个股或行业指数相对于相关证券市场主要指数的涨跌幅为标准。

**相关证券市场代表性指数说明：**A 股市场以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以道琼斯指数为基准。

## 免责条款

华鑫证券有限责任公司（以下简称“华鑫证券”）具有中国证监会核准的证券投资咨询业务资格。本报告由华鑫证券制作，仅供华鑫证券的客户使用。本公司不会因接收人收到本报告而视其为客户。

本报告中的信息均来源于公开资料，华鑫证券研究部门及相关研究人员力求准确可靠，但对这些信息的准确性及完整性不作任何保证。我们已力求报告内容客观、公正，但报告中的信息与所表达的观点不构成所述证券买卖的出价或询价的依据，该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。投资者应当对本报告中的信息和意见进行独立评估，并应同时结合各自的投资目的、财务状况和特定需求，必要时就财务、法律、商业、税收等方面咨询专业顾问的意见。对依据或者使用本报告所造成的一切后果，华鑫证券及/或其关联人员均不承担任何法律责任。本公司或关联机构可能会持有报告中所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、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露。

本报告中的资料、意见、预测均只反映报告初次发布时的判断，可能会随时调整。该等意见、评估及预测无需通知即可随时更改。在不同时期，华鑫证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。华鑫证券没有将此意见及建议向报告所有接收者进行更新的义务。