人工智能行业点评

超配

清华和蚂蚁携手AI安全合作，数据安全和AI监管是重要基础

◆ 行业研究·行业快评 ◆ 计算机 ◆ 投资评级：超配（维持评级）

证券分析师： 熊莉 021-61761067 xiongli1@guosen.com.cn 执证编码：S0980519030002
证券分析师： 库宏垚 021-60875168 kuhongyao@guosen.com.cn 执证编码：S0980520010001

事项：

2023年4月7日，清华大学与蚂蚁集团签署合作协议，双方将在“下一代互联网应用安全技术”方向展开合作，聚焦智能风控、反欺诈等核心安全场景。双方将携手攻坚可信AI、安全大模型等关键技术，致力于解决AI时代的互联网安全科技难题。

国信计算机观点：1）清华大学和蚂蚁集团携手聚焦在可信AI和安全通用大模型两个核心领域，安全问题是AI发展的前提条件。2）ChatGPT数据泄露、欧洲各国对ChatGPT的监管加强、三星数据泄露等事件，进一步推动了全球对AI安全的建设，OpenAI也发布了自己关于AI安全的建设方法。3）除了AI发展本身的安全性，AI也会带来攻防场景的改变：攻方AI降低了网络安全攻击门槛；守方AI赋能安全运营，同时提升防御技术。新技术应用，进一步推动安全刚需属性，带来持续安全需求。4）投资建议：国内AI应用逐步扩大，数据安全和AI监管同样是重点。随着国内百度、360、阿里、华为、腾讯等互联网公司发布或者计划发布大模型，AI相关应用也会逐步起量。针对AI滥用风险，数据安全、密码、AI监管均是重要技术发展方向，重点关注安全板块标的，如启明星辰、绿盟科技、天融信、深信服、奇安信、安恒信息、亚信安全、美亚柏科、三未信安等。5）风险提示：国内AI应用发展低于预期；宏观经济下行导致各行业IT支出下降；政策和法律等推进缓慢。

评论：

◆ 清华大学和蚂蚁集团联合突破AI安全技术

人工智能为网络安全带来新挑战。信息化、数字化，以及当前的智能化发展再持续带来科技突破的同时，也带来了大量安全隐患，尤其是当前ChatGPT引发的关于安全的问题，成为关注焦点。未来AI更广泛和深度的应用，对AI自身的安全可信、风控技术提出更高的要求。因此，清华大学与蚂蚁集团联合，将围绕“下一代互联网应用安全技术”长期攻坚，先期聚焦在可信AI和安全通用大模型两个核心领域：

在可信AI领域，双方将联合攻克安全对抗、博弈攻防、噪声学习等核心技术，加强AI模型的可信保障机制，助力提升规模化落地中的AI模型的可解释性、鲁棒性、公平性及隐私保护能力。

在安全通用大模型领域，双方将基于互联网异构数据，构建面向网络安全、数据安全、内容安全、交易安全等多领域多任务的安全通用大模型，以打造强推理能力，解决跨多风险领域的安全任务拓展；覆盖互联网行为全周期，实现端到端的安全防控运营智能化提效。

AI本身的安全性值得关注。在最近的“人工智能大模型技术高峰论坛”上，清华大学教授陶建华表示，AI大模型也有一定的技术风险，比如大模型本身可能存在安全漏洞，可以发生数据投毒攻击，对抗样本攻击等，对国家、企业、个人信息存在风险。因此，清华团队一直关注AI模型本身的安全，清华大学计算机科学与技术系长聘副教授黄民烈团队建立了大模型安全分类体系，并从系统层面和模型层面等打造了大模型安全框架。降低AI对人类社会的负面影响，关键在于大模型底座；该大模型安全分类体系设定了不安全对话场景，如犯罪违法、歧视仇恨等方面，让模型具备正确的问题回复策略，不进行误导。清华团队将打造中文大模型的安全风险评估的Leaderboard，为国内对话大模型的安全评估提供公平公开的测试平台，并提供针对中文对话的安全场景等。

◆ **AI 应用安全关注与日俱增，涉及内容监管、数据安全等多个领域**

AI 安全事件近期频发，监管和安全持续被重视。近期关于 AI 安全主要有三类事情引发关注：

1）**ChatGPT 数据泄露**：ChatGPT 在 3 月 20 日临时中断服务，主要是开源库中存在一个漏洞，致使一些用户可以看到另一用户的聊天记录标题，导致数据泄露。ChatGPT 现在已经是超过 1 亿用户的新超级应用，和其他的互联网应用一样，同样存在的数据泄露等安全风险。目前数据泄露是最常见的网络安全问题，在各个应用和行业均存在，ChatGPT 作为现象级应用，将推动产业对数据泄露和安全的建设。

2）**ChatGPT 引发各国加强监管**：3 月 31 日，意大利个人数据保护局宣布从即日起禁止使用 ChatGPT，限制 OpenAI 处理意大利用户信息，并开始立案调查。意大利认为平台没有就收集处理用户信息进行告知，缺乏大量收集和存储个人信息的法律依据。相应的，德国联邦数据保护专员发言人也表示，出于对数据安全保护的考虑，德国可能会效仿意大利，暂时禁用 ChatGPT。对于新兴的 AI 应用，在监管和法律上其实必然存在一定的滞后，但是用户数据安全和隐私保护是必然要重点建设的。欧盟目前正在抓紧制定关于人工智能的法律，美国政府也在讨论人工智能的“风险和机遇”。

3）**三星半导体使用 ChatGPT 数据泄露**：三星员工使用 ChatGPT，涉及两起“设备信息泄露”和一起“会议内容泄露”事件。该事件主要是三星员工“主动投喂”了相关数据和信息，例如将录制的会议内容输入 ChatGPT，让其帮助生成会议纪要等。因此相关信息可能已存入 ChatGPT 学习资料库中。这类事件存在的安全风险主要关于公司内部管理，因此很多企业逐步禁止将公司资料输入 ChatGPT。

OpenAI 持续加强自身安全建设。近日 OpenAI 在官网也发布了自己对安全建设的方法《Our approach to AI safety》。第一、构建安全的 AI 系统，使用诸如基于人类反馈的强化学习等技术改进模型行为，并构建广泛的安全监控系统，积极与政府合作，制定最佳监管形式。第二、从现实世界中学习以改善保障措施，通过各类应用持续反馈迭代和改进。第三、保护儿童，要求使用 OpenAI 工具的人必须满 18 岁，或者父母同意下并满 13 岁；OpenAI 不允许技术用于生成仇恨、骚扰、暴力等内容，最新模型 GPT-4 在响应不允许内容的请求方面减少了 82%的可能性。第四、尊重隐私，OpenAI 努力在可行的情况下从训练数据集中删除个人信息，对模型进行微调以拒绝请求私人个体的个人信息。第五、提高事实准确性，通过利用用户对被标记为错误的 ChatGPT 输出的反馈作为主要数据来源，OpenAI 已经提高了 GPT-4 的事实准确性；与 GPT-3.5 相比，GPT-4 生成事实内容的可能性提高了 40%。第六、持续的研究与参与，政策制定者和 AI 提供商需要确保 AI 的开发和部署在全球范围内得到有效治理。

◆ **AI 快速发展，为攻防双方带来新变化**

除了 AI 发展本身的安全性，AI 也会带来攻防场景的改变：

攻方：AI 降低了网络安全攻击门槛。由于 ChatGPT 极强的内容生成能力，根据网络安全公司 Darktrace 的最新报告，攻击者能够通过 ChatGPT，通过增加文本描述、标点符号和句子长度，高效生成更加难以分辨的钓鱼邮件，攻击量增加了 135%。除此之外，国外已经出现利用 ChatGPT 编写恶意软件，黑客甚至不需要编码就可以生成一个恶意软件的数十万次迭代。网络安全也是一种社会工程学，利用 ChatGPT 类 AI 工具的生成内容，几乎可以让非专业人士也能轻松产生简单攻击行为，因此 AI 可能会导致网络攻击的成本和门槛大幅降低，网络安全风险进一步加剧和泛滥。

守方：AI 赋能安全运营，同时提升防御技术。以微软推出的 Security Copilot 为例，该助手可以帮助防御者识别网络入侵，更好地理解每天收集到的海量信号和数据。AI 助手可以帮助甲方安全运维团队更快速和高效处理内部安全事件，对于海量日志、告警、事件的分析也会更有价值，减轻团队运维工作。除此之外，安全从业者也可以利用 ChatGPT 关联网络连接、恶意 IP 地址等，增强威胁情报能力。同时，ChatGPT 德国等工具也能加强逆向工程能力，大大提高反恶意软件工作的效率。因此 AI 的融入，也为了网络安全建设方带来新的便利。

新技术应用，进一步推动安全刚需属性，带来持续安全需求。AI 当前已经成为产业界一致方向，以 ChatGPT 为代表，AI 应用已经进入“iPhone 时刻”。更广泛和更深度的 AI 应用，一方面是自身本身的安全需求，必然需要法律、系统等全方位的建设和监管；另一方面，AI 对社会将造成广泛的外部性，对攻防双方安全技术的应用均会有大量变化。尤其是 AI 未来的普及，必然带动安全需求的持续提升和演化。整个安全行业将持续受益于 AI 技术的发展。

◆ **投资建议：**

国内 AI 应用逐步扩大，数据安全和 AI 监管同样是重点。随着国内百度、360、阿里、华为、腾讯等互联网公司发布或者计划发布大模型，AI 相关应用也会逐步起量。针对 AI 滥用风险，数据安全、密码、AI 监管均是重要技术发展方向。数据安全是 AI 发展的前提保障，建议关注数据安全企业：启明星辰、绿盟科技、天融信、深信服、奇安信、安恒信息、亚信安全、山石网科、安博通等。密码方面，数据加密，包括隐私计算也可能会引入相关模型训练或者 AI 应用中，建议关注三未信安、吉大正元等。在 AI 监管层面，重点关注美亚柏科，公司牵头编制国家通信行业标准《深度图像视频检测算法引擎安全性评测方法》，且研发了“AI-3300 慧眼视频图像鉴真工作站”，支持当前绝大部分深伪视频图像篡改方法的检测。未来公司进一步针对对各类 AIGC 内容的检测、AI 生成文本的检测技术及产品进行布局。

◆ **风险提示：**

国内 AI 应用发展低于预期；宏观经济下行导致各行业 IT 支出下降；政策和法律等推进缓慢。

相关研究报告：

《计算机行业 2023 年 4 月投资策略-BloombergGPT 发布，建议关注具有数据优势的细分龙头》 ——2023-04-02
《人工智能专题报告——生成式人工智能产业全梳理》 ——2023-03-28
《人工智能行业点评-英伟达 GPU、DGX 云、AI 工厂三驾马车发布，AI 算力和应用再迎跃迁》——2023-03-26
《人工智能行业点评-OpenAI 访问限流，GPT-4 算力测算》 ——2023-03-20
《人工智能行业点评-Microsoft 365 Copilot 发布，国内外 AI 应用有望加速落地》 ——2023-03-19

免责声明

分析师声明

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。

国信证券投资评级

类别	级别	说明
股票 投资评级	买入	股价表现优于市场指数 20%以上
	增持	股价表现优于市场指数 10%-20%之间
	中性	股价表现介于市场指数 ±10%之间
	卖出	股价表现弱于市场指数 10%以上
行业 投资评级	超配	行业指数表现优于市场指数 10%以上
	中性	行业指数表现介于市场指数 ±10%之间
	低配	行业指数表现弱于市场指数 10%以上

重要声明

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 层
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