兴泰控股
XING TAI HOLDINGS

金融信息采编

COMPILATION OF FINANCIAL NEWS

2025 年第 18 期总第 1245 期

合肥兴泰金融控股集团
金融研究所

咨询电话：0551—63753813
服务邮箱：xtresearch@xtkg.com
公司网站：http://www.xtkg.com/
联系地址：安徽省合肥市政务区
祁门路 1688 号兴泰金融广场 2602

2025 年 03 月 14 日 星期五

更多精彩 敬请关注
兴泰季微信公众号

宏观经济

交易商协会召开支持民企座谈会

3 月 12 日，中国银行间市场交易商协会 3 月 12 日召开银行间市场支持民营企业高质量发展座谈会，民营企业、主承销商、投资人、增信机构等代表参会。针对参会机构提出的意见建议，交易商协会副会长徐忠表示，交易商协会将强化债券市场制度建设和产品创新，持续发挥“第二支箭”撬动引领作用，扩大民企债券融资规模。中国人民银行相关司局负责人在座谈会上指出，将实施好适度宽松的货币政策，发挥好结构性货币政策工具作用，执行好金融支持民营经济 25 条举措，畅通股、债、贷多种融资渠道，引导金融机构“一视同仁”对待各类所有制企业，持续加大金融资源投入，助力民营经济高质量发展。本次座谈会上，应流机电、润泽科技、牧原食品、韵达控股等民企代表介绍了企业经营融资情况及建议。多位民营企业和金融机构一致建议，进一步研究推出适应民企转型升级需求的金融产品，积极支持产业类民企发行科创债券、绿色债券等融资工具。

国家数据局同意 7 地建试验区

3 月 13 日，国家数据局近日函复同意天津市、河北省（雄安新区）、上海市、江苏省、浙江省、广东省、四川省等 7 个地方开展国家数字经济创新发展试验区建设工作。国家数据局表示，下一步，各试验区将聚焦制约数字经济高质量发展的关键环节和突出问题，围绕推进数据要素市场化配置改革、优化数据基础设施建设布局、突破关键核心数字技术、纵深推进数字化转型、推进适数化改革等 5 个方面重点任务，梳理重要政策、重大改革、综合授权事项、预期成果、重大工程项目等清单，编制具体建设方案，按程序报批后，开展针对性试点试验，着力探索数据要素赋能实体经济发展的可行路径，打造具有国际竞争力的数字产业集群，构建促进数字经济发展体制机制，争当数字经济高质量发展的改革破局“先行者”、创新策源“排头兵”。国家数据局将会同有关方面，做好改革综合协调和指导服务，加大对各试验区政策支持力度，总结并宣传推广典型经验模式，推动条件成熟的按程序转化为全国性政策制度。

潞晨科技推出视频生成模型

3 月 13 日，潞晨科技宣布推出 Open-Sora 2.0，并全面开源模型权重、推理代码及分布式训练全流程。据该公司介绍，这是一款新开源的 SOTA 视频生成模型，仅用 20 万美元（224 张 GPU）成功训练出商业级 11B 参数视频生成大模型。从 Open-Sora 1.2 升级到 2.0 版本后，与 OpenAI Sora 闭源模型之间的性能差距大幅缩小。根据视频生成权威榜单 VBench 的评测结果，Open-Sora 模型的性能进步显著。从 Open-Sora 1.2 升级到 2.0 版本后，与行业领先的 OpenAI Sora 闭源模型之间的性能差距大幅缩小，从之前的 4.52%缩减至仅 0.69%，几乎实现了性能的全面追平。此外，Open-Sora 2.0 在 VBench 评测中取得的分数已超过腾讯的 HunyuanVideo，以更低的成本实现了更高的性能，为开源视频生成技术树立了全新标杆。为了追求极致的成本优化，Open-Sora 2.0 从严格

的数据筛选、高分辨率训练、优先训练图生视频任务和高效的并行训练方案四个方面着手削减训练开销。

复旦团队成功研制硅光复用芯片

3 月 13 日，复旦大学信息科学与工程学院张俊文研究员、迟楠教授与相关研究团队开展合作，通过精确设计和优化，将多维复用技术引入片上光互连架构，不仅显著提升了数据传输吞吐量，同时在功耗和延迟方面表现卓越，具备极强的扩展性和兼容性，适用于多种高性能计算场景。在此基础上，团队设计并研制了一款硅光集成高阶模式复用器芯片，实现了超大容量的片上光数据传输。实验结果表明，该芯片可支持每秒 38Tb 的数据传输速度，意味着未来 1 秒可完成大模型 4.75 万亿的参数传递，这显著提升了大模型训练与计算集群间的通信性能和可靠性，为人工智能、大模型训练及 GPU 加速计算等应用提供了强有力的支持。这款芯片的问世填补了市场上高性能数据传输设备的空白，与目前市场上的其他同类产品相比，在传输能力及功耗管理上具备明显优势，具有强大的竞争力。同时，随着人工智能及云计算技术的迅速发展，为该芯片的推广和应用提供了广阔的市场空间。

欧盟、加拿大对美实施反制关税

3 月 12 日，欧盟委员会发布公报称，因美国近期对从欧盟进口的钢铝产品征收不合理关税，欧盟决定对价值 260 亿欧元（1 欧元约合 1.09 美元）的美国商品征收反制关税。根据公报，欧盟的反制关税将分为两部分，旨在对等回应美国对从欧盟进口钢铝产品征收的关税。第一部分措施是从 4 月 1 日起，重启此前对美国关税的反制措施，以回应美国对欧盟价值 80 亿欧元钢铝出口造成的经济损害。第二部分是对于受美国新关税政策影响的价值超过 180 亿欧元欧盟出口商品，欧盟将对美国出口商品提出一揽子反制措施，在与成员国和利益攸关方协商后，这些措施将于 4 月中旬生效。加拿大政府宣布回击美国钢铝关税，将对总计 298 亿加元（1 加元约合 0.69 美元）的美国商品征收 25%的反制关税。加拿大财政部长勒布朗、外交部长乔利与创新、科学和工业部长商鹏飞共同召开记者会宣布这一反制关税方案，涉及的美国商品包括钢铁和铝产品、电脑、运动器材等。

21 名美民主党检察长起诉特朗普

3 月 13 日，美国 21 名民主党籍州总检察长因教育部裁员对特朗普政府提起诉讼。这一诉讼由纽约州总检察长利蒂希娅·詹姆斯领导，在马萨诸塞州联邦地区法院发起。诉讼中的被告是美国总统唐纳德·特朗普、教育部长琳达·麦克马洪（Linda McMahon）和美国教育部。“裁员是对教育部的实质性解散，”州检察长们在起诉书中写道，“教育部实施裁员的权力并不能凌驾于国会废除行政机构或停止其职能的专属权力。”此前，美国政府已经下令对教育部实施大规模裁员计划，预计将解雇 1300 多名教育部员工。而此前，美国教育部已经有 600 名雇员接受特朗普政府的“买断计划”——即美国政府将对所有主动离职的联邦政府雇员提供约 8 个月的薪资补偿，但联邦雇员需主动离职。美国教育部共有约 4000 名员工，特朗普此举相当于砍掉近一半员工。尽管美国教育部作为国会授权的机构，未经国会批准不能被取消。但是，特朗普政府可以通

过削减资源来让其慢慢被实质上废弃。

欧盟大力推行简化绿色监管改革

3 月 13 日，欧盟委员会近日通过一系列新提案，旨在简化欧盟为公民和企业设置的条例、提高竞争力并释放更多投资能力。欧盟委员会强调，为了恢复欧洲经济的竞争力，必须营造有利的商业环境，确保企业能够顺利发展，释放经济增长潜力。欧盟还公布了“清洁工业协议”，计划在短期内调动 1000 亿欧元资金，以支持本土制造业的能源转型，强化欧盟工业竞争力。这些简化监管提案主要涉及多项与绿色监管相关的法规。新方案提出，企业仅需对其直接商业合作伙伴和子公司进行尽职调查，而不再需要调查供应链中的其他分包商或二级供应商。此次简化方案还包括减少“绿色金融”分类的报告义务，以及削减与碳边境调节机制相关的合规要求。此次改革被视为欧盟优化营商环境、增强国际竞争力的举措。该改革方案已提交欧洲议会和欧盟理事会审议，将在获得批准后正式实施。不过，这些放宽措施预计将在欧洲议会引发激烈讨论。欧洲议会有议员称其为“一场大规模的放松管制”。

货币市场

美国 SEC 推迟新加密货币 ETF 申请

3 月 12 日，美国证券交易委员会（SEC）推迟了关于是否批准莱特币（LTC）、瑞波币(XRP)、狗狗币 Cardano(ADA)与 DOGE 等加密货币的新现货交易型基金(ETF)的决定。同时被推迟的还有 Canary 的 XRP、Solana、莱特币现货 ETF 以及 VanEck Solana 现货 ETF 等相关申请的决定。这一延迟对加密货币投资者来说是一个打击，他们原本希望看到与 XRP 和莱特币等规模较小的加密货币以及狗狗币等模因币相关的新 ETF 获批。继去年批准追踪比特币和以太坊价格走势的现货基金后，人们一直对华尔街监管机构批准新一批加密货币 ETF 寄予厚望。彭博市场此前预测，到今年年底，莱特币 ETF 获批的可能性为 90%，狗狗币 ETF 为 75%，XRP ETF 为 65%。现在，尚不清楚美国证券交易委员会是否或何时会批准其收到的来自 Grayscale 和 Ark Invest 等资产管理公司的新加密货币 ETF 申请。过去一年比特币的大部分涨幅是在 SEC 于 2024 年 1 月批准了大约 12 只现货比特币 ETF 之后推出的。

特朗普家族就入股币安举行会谈

3 月 13 日，特朗普家族的代表已经就收购加密货币交易所币安美国分公司的财务股份进行了谈判，此举将使特朗普与这家于 2023 年承认违反反洗钱要求的公司开展业务。与此同时，知情人士透露，币安的创始人赵长鹏一直在敦促特朗普政府赦免他。此前他在对一项相关指控认罪后被判入狱四个月。据了解，去年，币安联系了特朗普的盟友，提出与特朗普家族达成一项商业交易，作为将这家被流放的公司送回美国的计划的一部分，之后双方开始了谈判。目前还不清楚，如果交易达成，特朗普家族的股份将以何种形式出现，或者是否以赦免为条件。值得关注的是，3 月 12 日，加密交易所币安与总部位于阿布扎比的投资机构 MGX 宣布了一项 20 亿美元投资。MGX 收

购了币安的少数股权，实现了该公司对加密货币和区块链领域的首次进军。这笔交易，也是币安迄今为止首次获得机构投资，这也是对加密公司的单笔最大投资，同时也是以加密货币（稳定币）支付的最大投资。

监管动态

金融监管总局党委召开扩大会议

3 月 13 日，金融监管总局党委召开扩大会议，传达学习习近平总书记在全国两会期间的重要讲话精神和全国两会精神，研究部署贯彻落实举措。会议明确，加快制定出台与房地产发展新模式相适应的融资制度；研究建立“超长期国债+超长期贷款”服务模式，高效支持“两重”建设。一是有效防范化解重点领域风险。一体推进地方中小金融机构风险处置和转型发展，综合采取补充资本金、兼并重组、市场退出等方式分类化解风险。持续推进城市房地产融资协调机制扩围增效，坚决做好保交房工作。加快制定出台与房地产发展新模式相适应的融资制度。配合做好地方政府隐性债务置换工作。二是全力支持经济高质量发展。推出有针对性的金融支持措施，助力实施提振消费专项行动。三是不断提升金融监管质效。加快补齐监管制度短板，积极推动银监法、保险法修订。强化央地监管协同，保持对非法金融活动的高压严打态势。进一步规范监管执法。加强金融消费者权益保护。

重庆金监局发布十大警示案例

3 月 13 日，重庆“3·15”金融消费者权益保护暨“远离非法金融中介，守护群众合法权益”专题宣传活动启动仪式在中国工商银行南桥寺支行“红金渝”网格驿站举行。活动现场，重庆金融监管局发布“重庆十大非法金融中介警示案例”，并以“金融科普花花车”为载体，发挥流动金融知识科普站优势，将金融知识送到市民身边。在启动仪式上，“重庆十大非法金融中介警示案例”正式公布，这些案例涵盖了银行、保险、证券领域，包括贷款中介骗局、征信修复骗局、代理退保骗局、非法荐股等典型骗局。例如，张先生轻信了“低利率公积金信用贷”中介的虚假宣传，不仅没有获得承诺的低利率贷款，还白白损失了一笔高额的“包装费”。重庆金融监管局表示，下一步，将继续指导推动辖内金融监管分（支）局、金融机构深入开展场景化、生活化、趣味化和数字化的线上线下金融知识普及活动，帮助社会公众提升金融素养和风险防范能力 ，为广大金融消费者创造更加安全、放心的金融消费环境。

金融行业

上交所推出 16 举措服务实体经济

3 月 13 日，上交所近日制定形成了《关于进一步做好金融“五篇大文章”的行动方案》。《行动方案》围绕进一步做好金融“五篇大文章”，聚焦服务重大战略、重点领域、薄弱环节，充分发挥上交所大盘蓝筹集聚、硬科技领先、多产品支撑和精准

化服务的功能优势，推动更多要素资源向科技金融、绿色金融、普惠金融、养老金融、数字金融等领域集聚，助力经济社会高质量发展，提出 16 条具体举措。科技金融方面，充分发挥科创板服务“硬科技”企业功能，大力支持突破关键核心技术的优质科技型企业发行上市。完善适应科技型企业特点的再融资、并购重组、股权激励、股份减持等配套制度，提升服务的包容性、适应性，推动新质生产力加快发展。健全支持科技型企业全生命周期的产品服务体系，积极引导私募股权创投基金“投早、投小、投长期、投硬科技”。数字金融方面，加强金融科技场景应用，探索科技监管模式创新，提升交易所数字化水平。

央行等扩围跨国公司资金池试点

3 月 13 日，中国人民银行、国家外汇管理局发布通知，持续扩大跨国公司本外币一体化资金池业务试点，进一步拓展试点范围。跨国公司本外币一体化资金池业务试点政策，主要面向特大型跨国公司集团，旨在提升跨国公司跨境资金运营效率，加大对跨国公司跨境投融资便利化的支持力度。该政策于 2021 年 3 月在北京、深圳率先推出首批试点，2022 年推出第二批试点，并优化试点政策。2024 年 12 月，试点政策再次得到优化调整。根据此次通知，跨国公司本外币一体化资金池业务试点从上海、北京、江苏、浙江等 10 省市扩展到天津、河北、内蒙古、黑龙江、安徽、福建、山东、湖北、湖南、广西、重庆、四川、贵州、云南、新疆、厦门等省市。此次入围试点的省市所涉及试点政策内容主要包括：允许跨国公司根据宏观审慎原则自行决定外债和境外放款的集中比例、允许跨国公司通过国内资金主账户办理境外成员企业本外币集中收付业务、进一步便利跨国公司以人民币开展跨境收支业务等。

重点产业

助力上海细胞治疗产业国际化

3 月 13 日，承载着一名香港淋巴瘤患者康复希望的“救命药”——来自复星凯瑞（上海）生物科技有限公司的阿基仑赛注射液（奕凯达®），从其位于浦东新区的细胞工厂装运送往浦东国际机场，计划在 3 月 14 日搭乘飞机飞往香港。这标志着国内 CAR-T 药物首次成功“出海”。同时，国内首创、由上海先行先试的生物医药特殊物品联合监管机制也落地形成闭环，细胞治疗产品“出海”路径打通。目前，全国有百余家细胞治疗与基因治疗相关企业，上海占了 50%。除复星凯瑞、药明巨诺、科济制药已上市三款 CAR-T 产品外，上海细胞治疗集团、先博生物、原启生物等企业在 CAR-T 治疗研发或临床方面也有积极进展。CAR-T 细胞疗法，即嵌合抗原受体 T 细胞免疫疗法，通过改造患者自体 T 细胞来治疗癌症，是上海极具优势的生物医药细分赛道。据了解，下一步，针对新的发展需求，上海海关将继续扩展细胞治疗产品便利化通关模式的覆盖面，全面助力上海细胞治疗科技创新策源能力和产业发展能级。

我国深海科技产业有望加速崛起

3 月 13 日，上海市建设现代海洋城市工作领导小组会议召开。市委副书记、市长、

领导小组组长龚正指出，要聚力培育发展海洋新质生产力，提升海洋科技创新能力，打造海洋战略科技力量，掌握关键核心技术。推进现代海洋产业体系建设，持续强化海洋先进制造业硬实力，不断抢占价值链高端。深海资源丰富，是未来全球争夺的战略空间。过去一年，我国自主设计建造的首艘大洋钻探船梦想号在广州正式入列，在深海进入、深海探测、深海开发方面再增“国之重器”；首艘设计拥有完全自主知识产权的深远海多功能科学考察及文物考古船“探索三号”，也将进一步推进我国在深远海深潜及综合作业的能力。我国深海科技装备的接连突破，不仅标志着我国海洋科技实力显著提升，更成为推动传统海洋产业升级、新兴海洋产业崛起、未来海洋产业培育的关键力量。据财联社主题库显示，相关上市公司宝色股份完成了多项舰船及海洋工程用钛合金关键部件的研制工作，为国家深海事业的发展提供了重要技术支撑。

地方创新

珠海：5 亿元战略投资智谱大模型

3 月 13 日，珠海龙头国企华发集团日前宣布战略投资大模型领军企业北京智谱华章科技有限公司，金额为 5 亿元人民币，以推进智谱基座 GLM 大模型的技术创新与生态发展。据悉，珠海高新区、华发集团将联合智谱搭建首个城市级 GLM 大模型空间——“智谱+珠海华发空间”。智谱将基于全自研 GLM 基座大模型及成熟的 MaaS 平台，建立“技术支撑、产业加速、空间落地”三位一体的智谱+珠海华发大模型空间，围绕技术研发创新、算力资源调度、语料数据聚合、垂直领域大模型开发等多个领域构建生态，为珠海本地的人工智能发展提供从技术层、平台层到应用层的全栈技术保障。作为珠海最大的综合型国有企业集团，华发集团近年来强势布局新质生产力，全面向科技转型，形成扎实的专业投资能力和产业资源整合优势。去年 9 月，珠海正式宣布设立珠海新质生产力基金，将携手社会资本打造总规模 800 亿元的基金群，而该基金的管理方正是华发集团。

杭州：全国最大自动驾驶测试田建成

3 月 13 日，杭州市智能网联车辆测试与创新应用区域再次扩容，富阳区全域 1821 平方公里、桐庐县全域 1825 平方公里开放为智能网联车辆测试应用区域，至此，杭州建成全国最大的自动驾驶测试应用“田”，总面积达 6910 平方公里。杭州市自 2015 年启动智能网联车辆示范区建设以来，不断探求创新。如今，其不仅是国家 5G 车联网应用示范区，还是全国首批智能网联汽车“车路云一体化”应用试点城市。截至目前，杭州已经发放了 331 辆智能网联汽车的测试与应用牌照，自动驾驶累计总里程达 249.41 万公里。杭州大规模的测试区域为技术企业提供了实地验证的机会，有助于推动智慧交通生态圈发展。众多企业在这里投资发展，例如新石器慧通（北京）科技有限公司在桐庐县投资 20 亿元兴建无人车项目，涵盖无人车整车制造与 AI 算法开发。该公司在杭州已投入 180 辆无人车，计划到 2025 年再新增 1000 辆，进一步扩展无人车在快递物流及冷链运输等领域的应用。

深度分析

DeepSeek 证明中国 AI 产业正与美国并驾齐驱

陈兵（广开首席产研院，
首席产业研究员 ）
来源：中国首席经济学家论坛

人工智能发展的关键要素是数据、算力和算法，因美国对我国禁运高端 GPU 芯片，原先产业共识认为我国人工智能技术水平将受限于算力不足，与美国的差距将逐步扩大，但 DeepSeek 通过模型算法创新，在相对较低的算力资源上，实现与 OpenAI 媲美的大模型性能水平，且成本远低于 OpenAI。DeepSeek 的横空出世证明，即使我国 GPU 芯片落后于美国，但通过软件算法创新，仍能在大模型技术性能上缩小与美国的差距，同时受益于工程师红利，我国在人工智能+应用场景开发上领先美国，中国人工智能技术将与美国并驾齐驱，提升发展自主可控的人工智能产业信心。

一、DeepSeek：V3 与 R1 的发布及“幻觉率”难题

（一）DeepSeek 已发布 V3 和 R1 模型

成立不到两年，DeepSeek 大模型性能水平比肩 OpenAI。DeepSeek 于 2023 年 4 月由知名量化资管巨头幻方量化发起成立，2024 年 1 月发布首个大模型 DeepSeek LLM，包含 670 亿参数。2024 年 12 月上线并同步开源 DeepSeek-V3 模型，在短短两个月内，仅在 2000 块英伟达 H800 GPU（特供中国市场的芯片）上花费 558 万美元，便达到了与美国顶尖闭源模型相媲美的性能水平；2025 年 1 月 DeepSeek 正式发布 R1 模型，在国外大模型排名 Arena 上，R1 基准测试升至全类别大模型第三，在风格控制类模型(StyleCtrl)分类中与 OpenAI o1 并列第一，在中国区及美区苹果 App Store 免费榜均占据首位。DeepSeek-V3 定位为通用大模型，适用于智能客服、知识问答和内容生成等任务；R1 专为复杂推理任务设计，强化在数学、代码生成和逻辑推理领域的性能。

（二）DeepSeek 仍需降低模型“幻觉率”

幻觉是指模型生成看似合理但实际上与事实不符、无中生有或自相矛盾的内容，在 Vectara HHEM 人工智能幻觉测试（行业权威测试，通过检测语言模型生成内容是否与原始证据一致，从而评估模型的幻觉率，帮助优化和选择模型）中，通用大模型 V3 的幻觉率是 3.9%，高于同类型的 GPT-4o 1.5%的幻觉率，推理模型 R1 的幻觉率是 14.3%，高于同类型 GPT-o1 的 2.4%，其他测试标准中，DeepSeek 模型幻觉率也高于 GPT 同类型模型。我们认为，DeepSeek 模型幻觉率较高可能与其低精度训练、给予模型更多创造性奖励等有关，后续 DeepSeek 需降低模型幻觉率，如针对不同类型任务做更精细地训练、引入检索增强生成（RAG）技术，通过将外部知识库与大语言模型相结合，在模型生成文本时，从外部知识库中检索相关信息并融入生成过程，从而提升生成内容的准确性、时效性与专业性。

面对算力资源远低于海外大模型企业的状况，DeepSeek 引入“多头潜在注意力（MLA）”和“混合专家架构（MoE）”等优化 Transformer 架构降低算力需求；采取“FP8 混合精度训练”和“多偶流水线机制（DualPipe）”提升 GPU 芯片利用率；使用强化学习技术训练模型推理能力，实现与 OpenAI o1 模型相当的推理能力。

表 1：DeepSeek 模型幻觉率高于 GPT 系列模型

	DeepSeek R1	DeepSeek V3	GPT-o1	GPT-4o
Vectara's HHEM 2.1	14.3%	3.9%	2.4%	1.5%
Google's FACTS w/ GPT-4o & Claude-3.5-Sonnet	4.37%	2.99%	1.00%	1.49%
Google's FACTS w/ GPT-4o & Gemini-1.5-Pro	3.09%	1.99%	0.90%	1.39%
Google's FACTS w/ Claude-3.5-Sonnet & Gemini-1.5-Pro	3.89%	2.69%	1.39%	1.89%

对可能遇到的困难与挑战，报告表述清醒客观，特别提及关税、内需和执行层面问题。外部环境方面，报告指出“单边主义、保护主义加剧，多边贸易体制受阻，关税壁垒增多”，“地缘政治紧张因素依然较多”等现实问题。内部发展方面，报告细化中央经济工作会议判断，系统梳理三大矛盾：需求端“有效需求不足，特别是消费不振”的结构性问题，供给端“部分企业生产经营困难，账款拖欠问题仍较突出”，民生领域“群众就业增收面临压力”，“民生领域存在短板”的双重挑战。同时，报告特别提及政策执行效能问题，既关注“一些地方基层财政困难”等客观制约，也直指“一些工作协调配合不够，有的政策落地偏慢、效果不及预期”等短板。

二、DeepSeek：技术突破赋能模型升级

（一）DeepSeek 创新优化 Transformer 架构降低算力需求

DeepSeek-V3 模型创新优化 Transformer 架构，引入“多头潜在注意力（MLA）”和“混合专家架构（MoE）”降低算力需求，训练成本仅为同类闭源模型的 1/20。标准的注意力机制随着模型规模的增加，键值的缓存需求急剧增长，可能会因内存占用过高而导致计算效率低下。多头潜在注意力机制通过低秩联合压缩注意力的键值，将高维的键值映射到低维的潜在向量空间，但仍包含了输入的关键信息，显著减少键值缓存内存占用，降低约 80%。在每个注意力头得到潜在向量后，通过多头并行计算，每个注意力头关注输入序列的不同部分，最后将多头输出进行拼接组合成最终的输出。同时，为了提高模型训练的效率和性能，DeepSeek-V3 模型引入多 Token 预测（MTP）技术，传统的单 Token 预测训练每次只预测下一个 Token，MTP 技术则同时预测多个 Token，训练时间能缩短 20%-30%，且能更精准捕捉上下文语义关系，生成更准确、更连贯的文本。

混合专家架构（MoE）将模型分解为多个“专家”网络，每个专家网络都是独立的子模型，专门负责处理特定类型的输入。当输入数据进入模型时，由一个门控网络根据输入数据的特征，动态地将其分配给最合适的专家网络进行处理。MoE 架构的稀疏激活机制使得每次只有部分专家被激活参与计算，而不是所有专家都对每个输入进行计算，进一步降低对计算资源的需求。DeepSeek-V3 模型一共有 61 层，其中 58 层是 MoE 层，每层设置 257 个专家，包括 1 个共享专家和 256 个路由专家，模型专家总数达到 14906 个。共享专家扮演全局知识处理的角色，始终参与所有输入的计算，能够捕捉数据中的普遍模式，为模型提供稳定的基础输出。路由专家专注处理特定类型的输入，通过门控机制按需激活。

（二）DeepSeek 引入低精度训练等提升 GPU 利用率

DeepSeek-V3 模型不仅通过优化创新 Transformer 架构降低算力需求，同时采取“FP8 混合精度训练”和“对偶流水线机制（DualPipe）”提升 GPU 芯片的利用率。传统的训练方式通常采用 32 位浮点数（FP32）来表示模型参数和中间计算结果，这种高精度表示虽然能够保证计算的准确性，但在计算过程中需要消耗大量的计算资源和内存，并且在数据传输过程中会产生较高的通信开销。FP8 混合精度训练对于一些对精度要求相对较低的计算任务，使用 FP8 格式进行计算。由于 FP8 格式的数据占用内存更少，并且在支持 FP8 计算的硬件设备上，其计算速度相比 FP32 和 FP16 有显著提

升。对于一些对精度要求较高的操作，仍然使用较高精度的格式进行计算，以确保模型的训练稳定性和准确性。

在模型训练过程中，涉及到前向传播、反向传播以及参数更新等过程，这些过程中既包含矩阵乘法等数学运算，也包含不同计算节点之间的数据传输等通信操作。GPU 通常按照一定的顺序在指令执行流水线中进行。然而，由于数学运算和通信操作的特性不同，它们在执行过程中可能会导致流水线出现“气泡”，即 GPU 在某些时间段处于空闲状态，降低了 GPU 的实际利用率。对偶流水线机制（DualPipe）将模型的计算过程划分为多个阶段，每个阶段包含数学运算和通信操作，当一个阶段的数学运算正在进行时，利用这个时间启动下一个阶段的通信操作，使得数学运算和通信操作在时间上尽可能重叠，减少了数学运算等待数据传输的时间。

（三）DeepSeek 使用强化学习技术训练推理能力

DeepSeek-R1 模型充分利用 V3 模型架构，针对复杂推理任务，引入强化学习技术，实现了与 OpenAI o1 模型相当的推理能力。强化学习是通过不断的试错过程和对结果的反馈进行学习，在长期内最大化累积奖励。传统的强化学习通常会有一个额外的批评模型来评估当前策略的好坏，然后根据评估结果来调整策略。然而，批评模型的训练既复杂又耗费计算资源。DeepSeek-R1 使用 GRPO 算法，不需要批评模型，而是从当前策略中采样一组输出，然后根据这些输出的相对表现来调整策略，使表现较好的输出更有可能被生成，而表现较差的输出被抑制。DeepSeek-R1 的推理训练分多个阶段，首先是冷启动阶段，利用精心设计的冷启动数据对 DeepSeek-V3-Base 进行微调，为模型提供初始的推理能力。接着在第一阶段的基础上，用 GRPO 算法强化学习，进一步提升模型的推理能力，并设计准确性奖励保证模型推理的正确，格式奖励和语言一致性奖励提升模型输出的可读性和流程性。随着强化学习训练的深入，模型思考时间增加，还自发“涌现”了诸如反思（重新审视和重新评估先前步骤）以及探索解决问题的替代方法等更加复杂的操作，表明模型在强化学习过程中能够不断自主提升推理能力。

美国经济成分中互联网、软件工具、金融等行业有相对较高的回报率，现阶段，美国 AI 应用率先与上述行业融合。与美国不同，我国拥有 41 个工业大类、207 个中类、666 个小类，是全世界唯一拥有联合国产业分类中全部工业门类的国家，特点是场景多、私有数据多，为 AI+提供丰富的应用场景，赋能工业实现低碳绿色发展和产业升级。

三、聚焦模型实践：场景化生产应用与企业助手探索

（一）场景化小模型应用于生产环节

随着人工智能技术发展，工业生产将从自动化生产进一步升级至智能化制造，即从工业机器人工作站、工业视觉识别等自动化设备应用向数据+知识的综合应用升级。大模型对计算资源要求较高且实时性相对较差，工业生产中许多任务需要快速响应和精准处理，小模型通常是指参数量相对较少的机器学习模型，以其轻量化、灵活性和高效性等特点，在工业生产领域展现出独特的优势。在产品质量检测场景中，通过对大量正常和缺陷样本的学习，小模型构建了精确的缺陷识别模型。在实际检测时，当产品通过检测区域，模型会迅速对采集到的图像进行分析，判断是否存在缺陷以及缺陷的类型和位置。据统计，小模型检测系统的漏检率能降低至 1%以内，同时，检测速度大幅提高，有效提高产品质量，降低了因质量问题导致的返工和售后成本。在设备预测性维护场景中，通过实时采集设备的运行数据，如温度、压力、振动、转速等，小模型利用这些数据对设备的运行状态进行实时监测和分析。通过对历史数据的学习，模型建立了设备正常运行的状态模型。当设备运行数据出现异常波动时，模型能够及时发出预警，并通过数据分析预测可能出现的故障类型和时间。在生产过程控制优化

场景中，小模型首先对企业的生产数据进行全面收集和分析，包括订单信息、生产设备的产能、原材料库存、人员配备等。通过建立生产调度模型，能够根据实时的生产情况和订单需求，制定出最优的生产计划和调度方案。当有新的订单下达时，模型会迅速分析订单的紧急程度、产品型号、所需原材料等信息，并结合当前生产线上各设备的运行状态和生产进度，合理安排生产任务和设备资源。据企业评估，应用模型进行生产调度优化后，企业的生产效率能提高 15%，订单交付周期平均缩短 20%。

诸如上述的工业应用场景有上百种之多，但 AI+工业生产应用落地面临可靠性、数据隐私和工艺技术外泄风险以及经济性等挑战。DeepSeek 模型在较低的算力资源上实现优异的性能，企业为模型本地化部署购买硬件支出大幅降低，而不依赖云端服务，数据的存储和处理均在本地完成，用户对数据拥有绝对控制权，可以自主决定数据的访问、使用与共享权限，从根源上杜绝了数据泄露风险。同时，本地化部署还能降低对网络的依赖，提高系统的稳定性和响应速度，DeepSeek 模型可以在本地设备上即时响应请求，不受网络波动的影响，满足工业生产对可靠性和快速响应的要求。DeepSeek 也降低了成本，其训练成本仅为同类闭源模型的 1/20，模型推理成本也大幅降低，预计未来仍有进一步下降空间，将提高工业生产智能化升级的经济性。

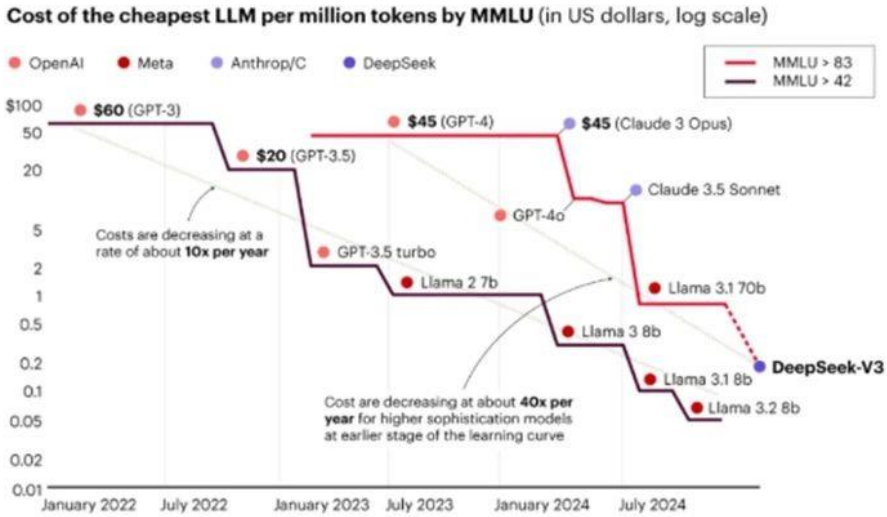

图 1：DeepSeek 模型推理成本大幅下降

（二）探索基于大模型的 AI 企业助手

随着 DeepSeek-R1 推理模型发布，大模型应用场景进一步向决策场景延伸。人工智能不仅赋能工业生产，进一步提升研发、财务、人力、客服与营销、供应链等企业内部管理流程效率。如办公自动化场景，可以利用自然语言处理技术，自动识别邮件的重要程度和类型，将重要邮件优先展示给员工，并对常见的邮件内容进行自动分类和归档。同时，AI 企业助手还能够根据邮件内容生成简短的摘要，帮助员工快速了解邮件的核心内容，提高邮件阅读效率。此外，对于一些常见的邮件回复，AI 企业助手可以根据预设的模板和语义分析，自动生成回复内容，大大提高了邮件回复的速度和准确性。在客户服务与营销场景，智能客服能够实时理解客户的问题和需求，并提供准确、快速的解答，可以同时处理大量的客户咨询，实现 24/7 的不间断服务，有效解决了传统人工客服在服务时间和服务容量上的限制。在数据分析与决策支持场景，通过大模型提供数据分析为客户决策提供支持，使用大语言模型全面加强基于大数据分析的规划能力，包括对未来产品价格预测、库存管理等。

根据目前大模型在企业侧落地交付实践，大模型能提升重复执行类和文本归纳类工业任务的处理效率，但受限于大模型现有技术瓶颈和企业数据质量问题，大模型在

数据分析与决策支持等场景的效果仍需进一步提升。但得益于大模型推理阶段 Scaling Law，人工智能技术能力仍有大幅提升空间，将扩大 AI+新型工业化应用场景，将人工智能应用于研发，大幅提升研发效率，如预测蛋白质结构、设计高性能芯片、高效合成新药等。

时至今日，智能手机、电脑等终端产品正面临创新瓶颈，消费者更换周期拉长使相关产品消费需求乏力，通过将人工智能与手机、电脑等终端产品融合，成为个人生活助手，将促进手机、电脑等产品更新需求。同时，人工智能与汽车、人形机器人结合将创造新的产业。DeepSeek 模型在低算力平台上的高效性，更契合在上述算力有限的终端产品上应用。

四、AI 驱动：智能终端与出行服务

（一）有望将 AI 手机打造成个人生活助手

目前，AI 大多以网页或独立应用的形式在手机上运行，无法获取设备上的数据或与其他应用交互，无法主动感知和操作。手机作为一种便携式移动设备，其硬件资源，如处理器性能、内存容量和电池续航能力等，与专业的服务器相比存在显著差距。DeepSeek-R1 拥有 6710 亿参数，在进行复杂的推理任务时，手机的处理器很难在短时间内完成如此庞大的计算量，导致响应速度变慢，甚至出现卡顿现象，严重影响用户体验，能耗问题也不容忽视。知识蒸馏是将大模型学到的知识传递给小模型，如 DeepSeek-R1 通过知识蒸馏，将长链推理模型的能力传授给参数规模较小的小模型，实现模型本地部署。DeepSeek 将激发 AI 在手机端的创新应用，为用户带来更多新颖、实用的功能。在智能语音交互方面，基于 DeepSeek 的 AI 手机能实现更自然、流畅的多轮对话。传统的语音助手在多轮对话中往往容易出现理解偏差和上下文衔接不畅的问题，而 DeepSeek 强大的自然语言处理能力使 AI 手机能够更好地理解用户的意图和上下文语境，实现更智能的交互。在个性化内容推荐方面，通过对用户的使用习惯、兴趣爱好和行为数据的深度分析，AI 手机能够利用 DeepSeek 的算法为用户推荐个性化的应用、新闻、音乐、视频等内容。

（二）提升自动驾驶技术等级促进 Robotaxi 落地

利用 DeepSeek 强大的语言理解、知识推理能力，对其进行微调，使其适应自动驾驶的特定任务需求。在微调过程中，使用大量的自动驾驶场景数据对 DeepSeek 模型进行训练，学习自动驾驶领域的专业知识和决策模式。例如，通过让 DeepSeek 学习不同路况下的驾驶决策案例，使其能够在面对类似场景时做出合理的决策，且其高效性也更适合部署在车辆计算平台。目前，业内专家预期 L4 级别的停车场自动泊车有望在 2027 年落地，高速公路 L4 级别自动驾驶也有望在 2027 年落地，L4 级别商用车自动驾驶有望在 2028 年落地，L4 级别市内自动驾驶需要到 2030 年前后落地。预计 2025 年中国自动驾驶市场空间 290 亿美元，其中 L2-L3 级自动驾驶市场空间 130 亿美元，L4-L5 级市场空间 160 亿美元；预计到 2030 年中国自动驾驶市场空间将达 6390 亿美元，年复合增速 85%，超过全球市场整体增速，其中，L4-L5 级自动驾驶市场占比 91%，年复合增速高达 105%。

随着 Robotaxi 在 2030 年前后迎来规模化生产，预计 Robotaxi 每公里运营成本仅 0.2 美元/公里，远低于传统网约车 0.45 美元/公里，成本节省主要由人力成本贡献，Robotaxi 每公里人力成本是 0.02 美元/公里，而传统网约车的人力成本是 0.33 美元/公里。Robotaxi 的每公里折旧、能源和保险成本都将高于传统网约车，预计 Robotaxi 的折旧成本是 0.095 美元/公里，高于传统网约车的 0.05 美元/公里；Robotaxi 的能源和保险成本是 0.075 美元/公里，高于传统网约车的 0.067 美元/公里。预计 2030 年中国 Robotaxi 市场空间 2010 亿美元，2025-2030 年年复合增速 111%,创造新的产业。

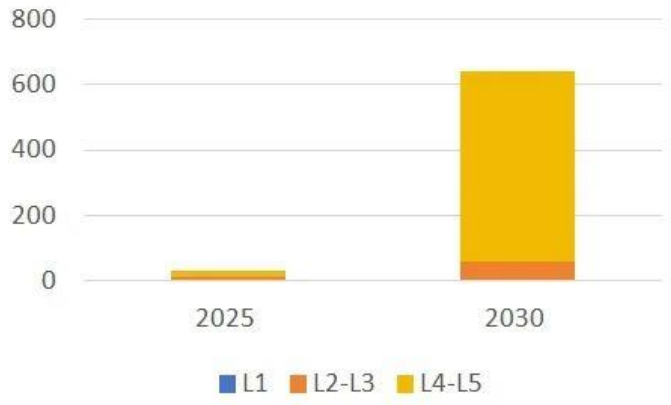

图 2：2025-2030 年中国自动驾驶市场空间年复合增长 85%

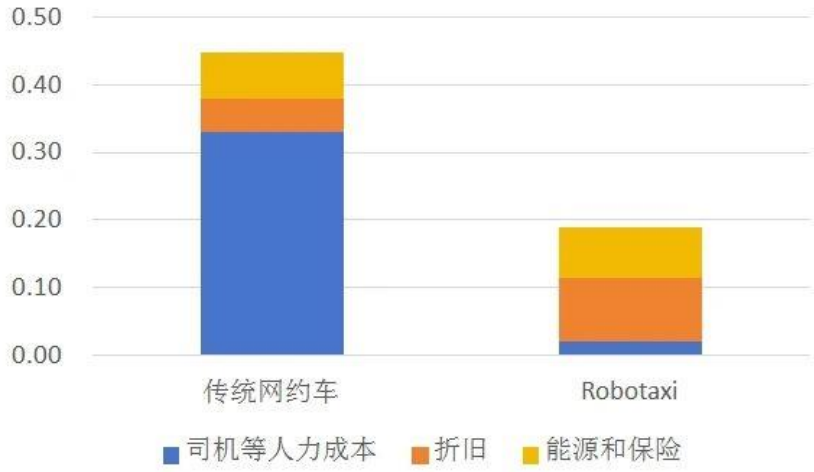

图 3：Robotaxi 单公里运营成本仅 0.2 美元

（三）加快人形机器人商业化进程

人形机器人硬件性能相对有限，如处理器的计算能力和通信带宽的不足，会导致机器人在执行复杂动作时，从接收指令到执行动作之间存在明显的延迟。数据采集与处理能力也需要进一步增强，在实际应用中，机器人需要实时采集大量的传感器数据，如视觉、听觉、触觉等信息，这些数据的维度高、噪声大，给数据处理带来了巨大的挑战。机器人还需要对采集到的数据进行实时分析和决策，以适应不断变化的环境和任务需求，目前的数据处理能力难以满足这一要求。现阶段，人形机器人成本高昂也限制了其商业化进程。DeepSeek 模型在感知、决策和行动等方面对人形机器人的发展起到了重要的推动作用，且其模型高效性特点也更易于部署在人形机器人性能相对有限的硬件平台上。

随着深度学习、强化学习等人工智能技术的不断发展，人形机器人的智能化程度将得到极大提升。在未来，人形机器人将具备更强的自主学习能力，能够通过与环境的交互不断积累经验，快速学习新的技能和知识。这将使其能够更好地适应复杂多变的环境和任务需求，实现更加智能化的决策和操作。在工业制造领域，人形机器人将不仅局限于简单的搬运和焊接等任务，还可以通过学习大量的生产数据和工艺知识，自主优化生产流程，提高生产效率和质量。在服务领域，如在老年人的康复护理中，人形机器人可以陪伴老人进行康复训练，提醒老人按时服药，为老人提供心理支持和陪伴。康复机器人还能配备智能监测系统，能够实时监测患者的康复进展和身体状况，并根据监测结果调整康复训练方案，提高康复效果。在应急救援领域，在火灾救援中，

人形机器人可以凭借其耐高温、耐烟雾的特性，进入火灾现场进行侦察和救援；在地震救援中，人形机器人具备强大的运动能力和适应能力，能够在倒塌的建筑物、瓦砾堆等复杂地形中行走、攀爬和穿越。从经济层面来看，人形机器人产业的崛起有望催生新的经济增长点，带动上下游产业链的协同发展。上游的核心零部件制造，如高精度减速器、伺服电机、传感器等，中游的机器人本体制造，以及下游的系统集成和应用服务，每个环节都蕴含着巨大的商业潜力。在社会层面，在老龄化日益严重的背景下，人形机器人可承担起养老护理、陪伴关爱等任务，缓解社会养老压力；在危险环境作业中，如消防、救援、矿山开采等，人形机器人可代替人类执行危险任务，保障人员生命安全。

免 责 声 明

《金融信息采编》是合肥兴泰金融控股集团金融研究所推出的新闻综合类型的非盈利报告。内容以全球财经信息、国内财经要闻、行业热点聚焦和地方金融动态为主，并结合对信息的简要评述，发出“兴泰控股”的见解和声音，以打造有“地方金融”的新闻刊物为主要特色，旨在服务于地方金融发展的需要，为集团公司、各子公司和相关专业人士提供参考。

《金融信息采编》基于公开渠道和专业数据库资料搜集整理而成，但金融研究所对这些信息的准确性和完整性不作任何保证。金融信息采编中的内容和意见仅供参考，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。兴泰控股集团金融研究所不对使用《金融信息采编》及其内容所引发的任何直接或间接损失负任何责任。