DeepSeek 赋能 AI 应用与端侧，助力算力国产化提速

2025 年 2 月 4 日

- **1 月板块表现强劲。**人工智能板块指数（884201.WI）表现强劲，涨跌幅达 3.6%，显著优于同期的宽基指数，如上证综指（-3.02%）、沪深 300（-2.99%）和创业板指数（-3.63%）。这一逆势上扬的态势反映了市场对人工智能技术持续迭代和应用拓展前景的认可。

- **美国 CES 展落幕，“星际之门”计划启动。**1 月 10 日，CES 大会落幕，AI 技术、应用和产品迎来新跃进。1 月 21 日，美国现任总统特朗普宣布启动“星际之门”（Stargate）人工智能基础设施计划，预示着全球将进入新一轮 AI 科技军备竞赛。

- **DeepSeek 全球爆火，国产大模型加速迭代。**1 月 20 日，DeepSeek 正式发布并开源 DeepSeek-R1 大模型，在数学、代码、自然语言推理等任务上性能比肩 OpenAI o1 正式版。同日，Kimi 发布了多模态思考模型 k1.5，其多模态和通用推理能力达到行业领先水平。1 月 22 日，字节跳动发布豆包大模型 1.5 pro，综合得分优于 GPT-4。1 月 29 日，阿里云通义千问旗舰版模型 Qwen2.5-Max 正式发布，在多向公开主流模型测评基准上录得高分。这些进展表明，国产大模型正在持续缩短与美国核心厂商的差距，尤其是 DeepSeek-V3/R1 通过算法与工程侧深度耦合，不仅将算力资源利用率最大化，降低成本（训练成本仅为 OpenAI 同类模型的 1/30），且性能上比肩 OpenAI o1。以 DeepSeek-R1 为代表的通用大模型开辟出了一条新的 AI 技术范式，震惊全球。国内 AI 应用及端侧 AI 的优势在于丰富的场景生态和庞大的流量，DeepSeek 的开源策略和高效性能将赋能 AI 应用与端侧 AI 爆发，我们认为，AI Agent 将在教育、办公、金融、医疗等领域发挥价值。

- **算力国产化进程预期加速。**1 月 13 日，美国政府宣布推出 AI 芯片管制新规，旨在进一步限制中国等国家和地区对于高端 AI 芯片及技术能力的获得，并对华封锁 16nm 以下先进制程。这一举措将倒逼国产技术持续突破。另一方面，DeepSeek 开源大模型也对国产化产生积极影响，DeepSeek 通过使用 FP8 低精度训练、双管道训练、共享专家机制，大大降低了训练成本（DeepSeek-V3 模型仅用 557.6 万美元和 2048 块 H800GPU 完成训练）。大模型成本的降低以及技术进步将推动整个算力资源使用总量上升，杰文斯悖论将再次到来，以 DeepSeek 为代表的大模型厂商将加速算力国产化进程。

- **投资建议：**关注以下细分赛道及公司：1、国产算力产业链及生态伙伴：如工业富联、中科曙光、曙光数创、海光信息、龙芯中科等。2、算力基础设施产业链：如润泽科技、宝信软件等。3、AI+应用：如科大讯飞、金蝶国际、金山办公、同花顺、嘉和美康、国能日新、彩讯股份、恒生电子、万兴科技等。4、端侧 AI：如虹软科技、海康威视、中科创达、华勤技术、萤石网络等。5、数据要素产业链中供给、流通、应用公司：如拓尔思、达梦数据、深桑达 A、上海钢联等。

- **风险提示：**技术迭代不及预期风险；科技巨头竞争加剧风险；法律监管风险；供应链风险；下游需求不及预期风险。

计算机行业

推荐 维持评级

分析师

吴砚靖

☎：010-66568589

✉：wuyanjing@chinastock.com.cn

分析师登记编码：S0130519070001

鲁佩

☎：（021）20257809

✉：lupei_yj@chinastock.com.cn

分析师证书编码：S0130521060001

研究助理 胡天昊

☎：（8610）80927637

✉：hutianhao_yj@chinastock.com.cn

相对沪深 300 表现图 2024-2-4

资料来源：Wind，中国银河证券研究院

重点公司盈利预测与估值

股票代码	股票名称	EPS			PE			投资评级
		2023A	2024E	2025E	2023A	2024E	2025E	
002230.SZ	科大讯飞	0.28	0.26	0.42	181.00	194.92	120.67	推荐
688041.SZ	海光信息	0.54	0.76	1.05	237.04	168.42	121.90	推荐
688692.SH	达梦数据	5.19	4.58	5.61	-	71.19	80.68	-
300442.SZ	润泽科技	1.02	1.3	1.88	58.45	45.86	31.71	-
301236.SZ	软通动力	0.56	0.65	0.82	95.86	95.86	82.58	-

资料来源：Wind，中国银河证券研究院

目录
Catalog

一、市场行情回顾

（一）整体行情

A 股人工智能指数（884201.WI）截至 1 月 27 日收盘价为 8542.27，月涨跌幅为 3.6%。计算机行业指数（801750.SI）截至 1 月 27 日收盘价为 4311.96，月涨跌幅为-2.16%。

图1：1 月人工智能指数走势图

资料来源：Wind，中国银河证券研究院

（二）代表企业

A 股 Wind 人工智能指数（884201.WI）截至 1 月 27 日总市值 19340.3 亿，含成分股 73 支，权重等分。上市板分布为主板 14 支，创业板 30 支，科创板 10 支，中小板 19 支。

表1：1 月成分股涨幅前十

股票代码	股票简称	1 月涨跌幅	1 月 27 日收盘价（元）	相对计算机指数涨跌幅
603893.SH	瑞芯微	49.72%	164.50	51.88%
688787.SH	海天瑞声	35.99%	133.98	38.15%
688088.SH	虹软科技	27.85%	49.30	30.01%
600410.SH	华胜天成	24.31%	9.00	26.47%
300458.SZ	全志科技	21.67%	47.16	23.83%
300307.SZ	慈星股份	19.90%	9.88	22.06%
000681.SZ	视觉中国	16.14%	24.18	18.30%
300222.SZ	科大智能	15.15%	11.48	17.31%
300442.SZ	润泽科技	14.74%	59.62	16.90%

002354.SZ	天娱数科	13.59%	6.10	15.75%

资料来源：Wind，中国银河证券研究院

表2：1 月成分股跌幅前十

股票代码	股票简称	1 月涨跌幅	1 月 27 日收盘价（元）	相对计算机指数涨跌幅
002253.SZ	川大智胜	-21.71%	10.82	-19.55%
300245.SZ	天玑科技	-19.83%	12.53	-17.67%
605168.SH	三人行	-13.11%	31.21	-10.95%
688256.SH	寒武纪-U	-13.07%	572.00	-10.91%
300474.SZ	景嘉微	-12.61%	81.70	-10.45%
002298.SZ	中电兴发	-12.41%	4.66	-10.25%
301316.SZ	慧博云通	-11.54%	22.62	-9.38%
300078.SZ	思创医惠	-9.54%	2.75	-7.38%
002049.SZ	紫光国微	-8.16%	59.12	-6.00%
600797.SH	浙大网新	-7.52%	6.64	-0.05

资料来源：Wind，中国银河证券研究院

（三）板块估值

人工智能指数（884201.WI）重要成分股 2021-2023 年整体营业收入复合增长率 1.90%，净利润复合增长率-8.41%，截至 1 月 27 日平均估值 PE（TTM）97.64 倍，PS（TTM）4.31 倍。

图2：1 月人工智能指数市场表现

资料来源：Wind，中国银河证券研究院

表3：1月人工智能主题基金一览

基金代码	基金简称(官方)	基金规模（亿元）	11月30收盘价（元）	近1月回报（%）	近3月回报（%）	近6月回报（%）	第一大重仓股名称（2024年报）
001986.OF	前海开源人工智能A	6.99	1.58	5.59	6.34	14.77	罗博特科
005729.OF	南方人工智能主题	3.87	2.18	2.49	4.83	17.55	腾讯控股
005844.OF	东方人工智能主题A	6.43	0.99	-12.53	7.99	27.48	中科飞测
005962.OF	宝盈人工智能A	5.21	2.68	5.94	15.45	27.29	海光信息
005963.OF	宝盈人工智能C	2.58	2.54	5.87	15.22	26.79	海光信息
006281.OF	万家人工智能A	19.51	2.47	-4.05	6.07	14.79	寒武纪-U
008020.OF	华富中证人工智能产业ETF联接A	1.96	0.85	-1.35	8.24	24.73	石头科技
008021.OF	华富中证人工智能产业ETF联接C	1.88	0.84	-1.37	8.15	24.54	石头科技
008585.OF	华夏中证人工智能主题ETF联接A	7.06	0.86	-0.38	9.63	25.09	
008586.OF	华夏中证人工智能主题ETF联接C	6.83	0.85	-0.40	9.55	24.89	
009239.OF	融通中证人工智能主题C	1.47	1.42	-0.05	9.57	24.81	寒武纪-U
011832.OF	西部利得中证人工智能A	1.46	0.94	0.56	13.01	29.56	寒武纪-U
011833.OF	西部利得中证人工智能C	0.97	0.92	0.53	12.89	29.30	寒武纪-U
011839.OF	天弘中证人工智能主题A	2.27	0.94	-0.06	9.75	25.18	寒武纪-U
011840.OF	天弘中证人工智能主题C	7.37	0.93	-0.09	9.70	25.06	寒武纪-U
012733.OF	易方达中证人工智能主题ETF联接A	6.81	1.07	-0.29	10.15	25.79	
012734.OF	易方达中证人工智能主题ETF联接C	10.32	1.07	-0.30	10.13	25.72	
014162.OF	万家人工智能C	17.47	2.41	-4.11	5.86	14.34	寒武纪-U
014630.OF	汇添富中证人工智能主题联接A	0.06	1.02				
014631.OF	汇添富中证人工智能主题联接C	0.06	1.02				
017811.OF	东方人工智能主题C	40.77	0.98	-12.56	7.88	27.24	中科飞测
021580.OF	华夏中证人工智能主题ETF联接D	0.67	0.85	-0.40	9.56	24.90	
023286.OF	前海开源人工智能C		1.58				
023407.OF	华宝创业板人工智能联接A						
023408.OF	华宝创业板人工智能联接C						
159363.OF	华宝创业板人工智能ETF	6.69	0.97				中际旭创
159702.OF	汇添富中证人工智能ETF	0.13	0.79				
159819.OF	易方达中证人工智能ETF	85.15	0.90	-0.37	9.98	26.56	寒武纪-U
161631.OF	融通中证人工智能主题A	6.00	1.45	-0.01	9.68	25.06	寒武纪-U
512930.OF	平安中证人工智能ETF	11.24	1.29	-0.31	9.52	25.67	寒武纪-U
515070.OF	华夏中证人工智能ETF	29.58	1.15	-0.45	9.89	26.40	寒武纪-U
515980.OF	华富中证人工智能产业ETF	19.14	0.97	-1.27	9.02	26.61	中际旭创
517800.OF	方正富邦中证沪港深人工智能50ETF	1.63	0.72	0.07	8.43	29.09	腾讯控股
588730.OF	易方达上证科创板人工智能ETF	2.06	1.00				
588760.OF	广发上证科创板人工智能ETF	3.26	1.03				
588790.OF	博时科创板人工智能ETF	2.34	1.09				

资料来源：Wind，中国银河证券研究院

表4：人工智能主要上市公司近况一览（数据截至 2025 年 1 月 27 日）

股票代码	股票名称	2023 营收增速（%）	2023 净利润增速（%）	24Q3 营收增速（%）	24Q3 净利润增速（%）	总市值（亿元）	市盈率 PE（TTM）	市销率 PS（TTM）	月涨跌幅（%）	今年以来涨跌幅（%）
000977.SZ	浪潮信息	-5.41	-12.89	72.26	66.49	764.63	33.38	0.76	0.20	0.20
002230.SZ	科大讯飞	4.41	22.97	17.73	-1039.84	1171.59	546.84	5.35	4.88	4.88
002236.SZ	大华股份	5.41	230.49	0.77	-3.97	499.29	6.82	1.54	-5.31	-5.31
002362.SZ	汉王科技	3.56	1.72	17.87	29.90	60.87	-50.76	3.75	9.89	9.89
002405.SZ	四维图新	-6.72	-171.14	9.06	9.42	205.87	-15.77	6.18	-9.96	-9.96
002415.SZ	海康威视	7.42	11.78	6.06	-6.22	2680.40	20.06	2.88	-5.44	-5.44
300229.SZ	拓尔思	-13.84	-72.98	2.95	82.98	193.16	332.81	24.16	5.59	5.59
300474.SZ	景嘉微	-38.19	-79.35	-5.99	53.28	426.98	628.10	62.32	-12.61	-12.61
601360.SH	三六零	-4.89	77.66	-16.76	-56.39	776.95	-110.56	9.80	7.25	7.25
603019.SH	中科曙光	10.34	16.12	3.65	2.12	980.20	52.84	6.70	-7.37	-7.37
688088.SH	虹软科技	26.07	54.61	14.09	8.34	197.78	207.92	26.69	27.85	27.85
688169.SH	石头科技	30.55	73.32	23.17	8.22	422.63	19.54	4.24	4.33	4.33
688207.SH	格灵深瞳	-25.84	-379.64	-72.99	-684.03	38.79	-18.41	39.44	5.20	5.20
688256.SH	寒武纪-U	-2.70	33.72	27.09	12.30	2387.85	-312.06	318.86	-13.07	-13.07
688787.SH	海天瑞声	-35.33	-203.16	44.90	111.80	80.82	1463.29	37.36	35.99	35.99
688793.SH	倍轻松	42.30	59.50	-11.16	183.41	22.47	-104.85	1.92	-12.25	-12.25
002410.SZ	广联达	-0.42	-88.22	-8.06	-17.59	192.95	285.81	3.14	-0.68	-0.68
688327.SH	云从科技-UW	19.33	28.26	-34.51	-23.69	128.37	-17.10	25.23	2.31	2.31
688343.SH	云天励飞-U	-7.36	14.21	112.52	-41.51	171.53	-33.63	22.52	-2.62	-2.62
688246.SH	嘉和美康	-3.04	-50.21	-11.52	-4748.76	32.92	-196.05	5.20	-3.51	-3.51
603893.SH	瑞芯微	5.17	-54.65	48.47	354.90	689.09	168.37	24.27	49.72	49.72
300033.SZ	同花顺	0.14	-17.07	-1.59	-15.53	1500.60	116.98	42.55	-2.91	-2.91
300496.SZ	中科创达	-3.73	-45.54	-4.70	-69.75	285.85	2348.22	5.65	4.33	4.33
688111.SH	金山办公	17.27	16.23	10.90	17.23	1452.33	99.18	29.56	9.64	9.64
688475.SH	萤石网络	12.39	68.80	12.93	-6.69	273.89	51.08	5.17	15.20	15.20
300634.SZ	彩讯股份	25.18	40.17	10.72	-37.78	133.29	62.54	8.26	35.50	35.50
300624.SZ	万兴科技	25.49	68.43	-3.91	-105.42	132.36	846.26	9.20	8.41	8.41
301162.SZ	国能日新	26.89	22.93	18.15	7.96	40.88	47.73	7.98	-10.33	-10.33
688188.SH	柏楚电子	56.61	53.12	31.19	30.10	409.14	46.36	23.84	2.52	2.52

资料来源：Wind，中国银河证券研究院

表5：境外上市人工智能企业近况一览（数据截至 2025 年 1 月 31 日）

证券代码	证券简称	3Q24 营业收入（亿元）	3Q24 营业收入同比增长率（%）	3Q24 归母净利润（亿元）	3Q24 归母净利润同比增长率（%）	总市值（原始币种、亿元）	市盈率 PE（TTM）	市销率 PS（TTM）	月涨跌幅（%）	今年以来涨跌幅（%）

TSLA.O	特斯拉	468.01	0.53	26.07	-52.72	13014.03	183.53	13.32	0.19	0.19
NVDA.O	英伟达	560.84	85.53	314.80	205.17	29405.14	46.62	25.96	-10.59	-10.59
GOOGL.O	谷歌	1652.81	14.68	472.81	23.05	24974.09	26.49	7.35	7.78	7.78
META.O	脸书	755.27	22.50	258.34	59.50	17461.47	28.00	10.61	17.71	17.71
MSFT.O	微软	1185.37	15.84	441.61	21.17	30855.49	33.27	11.79	-1.53	-1.53
BIDU.O	百度	654.44	-0.65	109.36	168.75	317.69	10.52	1.66	7.46	7.46
AAPL.O	苹果	2103.28	0.79	575.52	-3.36	35452.09	36.87	8.96	-5.76	-5.76
BABA.N	阿里巴巴	4589.46	8.90	620.89	28.53	2351.13	19.19	1.71	16.57	16.57
0700.HK	腾讯控股	3206.18	7.49	895.19	-38.79	37011.23	19.66	5.19	-3.79	-3.79
0020.HK	商汤-W	17.41	0.00	-24.57	-6.54	595.82	-9.42	14.64	8.05	8.05
0268.HK	金蝶国际	29.00	0.00	-2.18	46.07	368.64	-233.31	5.55	20.52	20.52

资料来源：Wind，中国银河证券研究院

二、人工智能产业动态

（一）数据要素、数据交易所最新新闻及政策

表6：数据要素最新新闻及政策

日期	具体内容
1.24	**国家数据局：加快各项改革举措落地落实，大力推动数据要素市场化价值化** 会议认为，国家数据局成立以来，始终以推动数据要素市场化配置改革作为工作主线，通过建立健全数据基础制度、建设数据基础设施、推进数据资源开发利用、实施“数据要素×”行动等工作，推进数据要素价值释放。相关部门结合各自职能，积极稳妥推进数据资产入表、数据资产管理工作，探索股东可依法依规用数据作价出资等工作，共同推进数据要素市场化价值化。会议认为，从数据要素价值实现路径来看，数据要素通过与其他生产要素的协同，进入社会化大生产，进而创造价值。2024 年，国家数据局重点推进数据要素市场化工作，通过发挥市场机制作用让数据供出来、用起来。数据只有用得好，价值才能“显性化”。市场化是手段，价值化是目的。无论是数据产品还是数据服务，只有在使用过程中才会创造价值、体现价值。数据资产和数据资本是助推和放大数据要素价值的重要路径。会议指出，当前数据市场培育正处于起步发展阶段，要加快各项改革举措落地落实，大力推动数据要素市场化价值化，充分发挥市场机制作用，实现数据“供得出、流得动、用得好、保安全”，让数据的价值体现在企业降本增效里，体现在培育新质生产力中，体现在赋能经济社会高质量发展上。会议强调，数据要素市场化价值化涉及大量的理论和实践问题，相关工作也需要久久为功、持续用力，需要政产学研合作，凝聚众智来共同解答、协同推进。国家数据局将与财政部、市场监管总局等部门密切协作，强化对地方工作的有效指导，鼓励在数据工作方面积极探索，加强场景需求牵引，推进数据要素协同优化、复用增效、融合创新，不断释放数据要素乘数效应，持续推进数据要素市场化价值化进程。
1.21	**河南数据要素研究中心揭牌成立** 1 月 21 日，省数据局、省战略研究院、郑州数据交易中心在郑州举行战略合作框架协议签约仪式。省发展改革委党组书记、主任马健出席并为河南数据要素研究中心揭牌，党组成员、副主任王旭出席见签，省数据局党组书记、局长郑华卿，省战略研究院党委书记、院长王文莉，郑州数据交易中心总经理潘新民分别代表三方签约。三方一致同意将持续深化数据领域重大问题研究，加强政策储备和宣传解读，搭建研讨交流合作平台，凝聚社会各方力量，共谋数据要素发展、共创美好“数字未来”，为推动全省数字化转型发展、加快数字强省建设提供坚强支撑。省数据局、省战略研究院、郑州数据交易中心班子成员及相关人员参加了签约仪式。

1.17	**“人大指数”发布数据要素市场化推进力指数** “人大指数”系列发布启动会暨 2025 年首场指数发布会 1 月 17 日举行，发布了数据要素市场化推进力指数和成渝双城经济圈协同发展指数报告。中国人民大学把影响经济社会发展的战略性重大问题作为研究重点，形成了一批关于中国式现代化、经济金融、城市与区域发展、人口健康、新质生产力等各领域的指数，并在此基础上整合各类指数设立“人大指数”系列发布平台，统一对外发布各领域指数，建立固定机制，形成品牌效应，以更好服务经济社会高质量发展。现场发布的“数据要素市场化推进力指数”由人大信息资源管理学院发起，旨在面向国家的数据事业和数据要素市场战略指引，基于对各省（自治区、直辖市）最新实践的调查研究，形成了包括基础环境、保障支撑、执行推进三大维度、7 项二级指标、17 项三级指标、37 项四级指标的评价体系，以科学反映各地在引导和规范数据要素市场发展方面的能力和成效。根据数据要素市场化推进力指数（2024），我国数据要素市场发展仍处于初级阶段，各地因地制宜开展数据要素市场化建设的探索，取得可喜进展。广东、浙江、北京等第一梯队的数据要素市场化推进力表现出色，创新探索，先行先试，有效发挥引领示范作用。
1.15	**国家发改委等六部门印发《关于完善数据流通安全治理更好促进数据要素市场化价值化的实施方案》** 1 月 15 日，国家发改委等部门印发《关于完善数据流通安全治理更好促进数据要素市场化价值化的实施方案》，到 2027 年底，规则明晰、产业繁荣、多方协同的数据流通安全治理体系基本构建，数据合规高效流通机制更加完善，治理效能显著提升，为繁荣数据市场、释放数据价值提供坚强保障。
1.14	国家数据局批复复同意贵州建设数据要素综合试验区 1 月 14 日消息，国家数据局于近日正式批复同意贵州建设数据要素综合试验区。按照批复文件要求，数据要素综合试验区建设的主线是数据要素市场化配置改革，突破口是公共数据汇聚治理、授权运营，重点任务包括数据基础设施建设、数据资源体系构建、数据资源开发利用、数据赋能产业发展、数据安全治理等。按照建设要求，数据要素综合试验区建设以制度建设为主线，以促进数据要素流通使用为重点，探索符合区域特点的数据价值释放路径，进一步激发经营主体活力，培育壮大数据要素市场，要求聚焦优势领域，打造释放数据价值的标志性成果，促进数据“供得出、流得动、用得好、保安全”，为西部地区推进数据要素市场化配置改革提供经验借鉴。

资料来源：中证网、观点网、北京青年报、Wind、大河财立方，中国银河证券研究院

表7：数据交易所新闻及政策

日期	具体内容
1.26	**北京国际大数据交易所累计备案交易近 100 亿元** 日前，2025 北京数据交易成果报告会在京举办。会上透露，截至 2024 年底，北京国际大数据交易所累计备案交易金额近 100 亿元，上架数据产品超 3000 个。会上，北京国际大数据交易所“数据流通交易专家咨询委员会”正式成立，朝阳区数据要素产业园举行开园仪式。
1.9	**广州数据交易所增城服务专区揭牌，27 家单位签署战略合作** 1 月 9 日，广州数据交易所（增城）服务专区（下称“增城服务专区”）正式揭牌。据介绍，增城服务专区可为参与数据要素交易的各类主体提供包括基础配套服务、会员管理服务、数据产权登记指引服务在内的三大基础服务，同时基于分级分类的会员体系，配套供给数据交易生态培育、数据经纪服务、数据运营管理服务等增值服务。值得关注的是，目前，增城服务专区已推动“农业气象指数保险”“数字供销综合服务平台”“城市体检评估报告”等首批本地特色数据产品率先上架，涵盖农业服务、城乡规划建设等多个应用场景。据悉，增城服务专区正在推进 106 家企事业单位申报成为专区会员。在此次专区成立仪式上，工信部电子五所、中国联通、广州数科集团等 27 家单位与增城服务专区签约成为战略合作伙伴，其中，成都数据集团、厦门数据交易公司成为增城服务专区的首批跨省合作伙伴，进一步推动数据要素资源跨区域流动。此外，为将金融资源与外部资金引入增城区，结合国家金融监督管理总局的金融资产投资公司（AIC）股权投资试点，增城产投集团拟联合广州产投、工银资本共同发起设立 2025 年广州首支综合性 AIC 基金，首期规模 10 亿元，基金投资领域包括数字经济、新一代信息技术等广州市、增城区重点发展的产业，而三方也在此次仪式上完成了基金意向合作签约。据介绍，以该基金为契机，增城区将与 AIC 公司和产业投资机构建立更广泛的股权投资合作，通过“产业+金融”、赋能实体经济的股权投资模式，推动数字经济等新质生产要素集聚增城。

1.8	**北京国际大数据交易所落户北辰世纪中心** 1 月 8 日消息，北京国际大数据交易所（简称北数所）于近日正式落户北辰世纪中心。据介绍，北数所是北京金控集团联合 48 家单位共同发起成立的北京国际数据交易联盟，为国内首家基于“数据可用不可见，用途可控可计量”新型交易范式的数据交易所。另据了解，北辰世纪中心总面积 14.6 万平方米，位于北京市朝阳区奥运村商务区，聚焦数据产业赛道，以数据交易流通为核心，发展上下游生态产业。

资料来源：Wind、观点网、南方都市报，中国银河证券研究院

（二）算法端：国内外巨头大模型动态

DeepSeek 领衔国产大模型突破。1 月 20 日，DeepSeek 正式发布并开源 DeepSeek-R1 大模型，在数学、代码、自然语言推理等任务上性能比肩 OpenAI o1 正式版。同日，Kimi 发布了多模态思考模型 k1.5，其多模态和通用推理能力达到行业领先水平。1 月 22 日，字节跳动发布豆包大模型 1.5 pro，综合得分优于 GPT-4。1 月 29 日，阿里云通义千问旗舰版模型 Qwen2.5-Max 正式发布，在多向公开主流模型测评基准上录得高分。这些进展表明，国产大模型正在持续缩短与美国核心厂商的差距，尤其是 DeepSeek-V3/R1 通过算法与工程侧深度耦合，不仅将算力资源利用率最大化，降低成本（训练成本仅为 OpenAI 同类模型的 1/30），且性能上比肩 OpenAI o1。以 DeepSeek-R1 为代表的通用大模型开辟出了一条新的 AI 技术范式，震惊全球。国内 AI 应用及端侧 AI 的优势在于丰富的场景生态和庞大的流量，DeepSeek 的开源策略和高效性能将赋能 AI 应用与端侧 AI 爆发，我们认为，AI Agent 将在教育、办公、金融、医疗等领域发挥价值。

表8：国内人工智能大模型动态

时间	模型	主要内容
1.28	DeepSeek Janus-Pro	**DeepSeek 发布新款开源多模态 AI 模型 Janus-Pro** 1 月 28 日凌晨，人工智能社区 HuggingFace 显示，DeepSeek 发布了开源多模态 AI 模型 Janus-Pro。据介绍，Janus-Pro 是 Janus 的高级版本，其拥有优化的训练策略，扩展的训练数据以及更大的模型规模，这些改进使得 Janus-Pro 在多模态理解和文本到图像的指令跟踪能力方面都取得了重大进步，同时还增强了文本到图像生成的稳定性。Janus-Pro 系列包括了参数量分别为 7B 和 1.5B 的两个型号。报告公开的测试结果显示，Janus-Pro-7B 在 GenEval 和 DPG-Bench 基准测试中击败了 OpenAI 的 DALL-E3 和 StableDiffusion。
1.26	Baichuan-Omni-1.5	**百川智能上线开源全模态模型 Omni-1.5，号称多项能力超越 GPT-4omini** 1 月 26 日消息，百川智能今日宣布，Baichuan-Omni-1.5 开源全模态模型正式上线。该模型不仅支持文本、图像、音频和视频的全模态理解，还具备文本和音频的双模态生成能力。官方宣称，其在视觉、语音及多模态流式处理等方面，Baichuan-Omni-1.5 的表现均优于 GPT-4omini；在多模态医疗应用领域，其具备更突出的领先优势。Baichuan-Omni-1.5 不仅能在输入和输出端实现多种交互操作，还拥有强大的多模态推理能力和跨模态迁移能力。模型结构方面，Baichuan-Omni-1.5 的模型输入部分支持各种模态通过相应的 Encoder/Tokenizer 输入到大型语言模型中。而在模型输出部分，Baichuan-Omni-1.5 采用了文本-音频交错输出的设计，通过 TextTokenizer 和 AudioDecoder 同时生成文本和音频。百川智能构建了一个包含 3.4 亿条高质量图片/视频-文本数据和近 100 万小时音频数据的庞大数据库，且在 SFT 阶段使用了 1700 万条全模态数据。

1.25	TeleAI-t1-preview	**中国电信发布“复杂推理大模型”，数学基准评测超越 GPT-4o** 1 月 25 日消息，中国电信人工智能研究院（TeleAI）“复杂推理大模型”TeleAI-t1-preview 近日正式发布。TeleAI-t1-preview 使用了强化学习训练方法，通过引入探索、反思等思考范式，大幅提升模型在数学推导、逻辑推理等复杂问题的准确性。在美国数学竞赛 AIME2024、MATH500 两项权威数学基准评测中，TeleAI-t1-preview 分别以 60 和 93.8 分的成绩，大幅超越 OpenAIo1-preview、GPT-4o 等标杆模型。在研究生级别问答测试 GPQADiamond 中，TeleAI-t1-preview 得分超过 GPT-4o，并比肩 Claude3.5Sonnet 的性能水准。
1.23	豆包大模型 1.5	**豆包大模型 1.5Pro 灰度上线开发者可直接调用 API** 1 月 23 日消息，据豆包官方公众号消息，豆包大模型 1.5Pro 版本正式发布，目前已在豆包 APP 灰度上线，接受海量请求，开发者也可在火山引擎直接调用 API（应用程序编程接口）。豆包官方介绍显示，豆包大模型 1.5Pro 在知识(MMLU_PRO、GPQA)、代码(McEval、FullStackBench)、推理(DROP)、中文(CMMLU、C-Eval)等多项公开测评基准上成绩全球领先。豆包方面表示，豆包大模型 1.5Pro 使用较小的激活参数进行预训练，训练成本极低，但性能不打折，采用大规模稀疏 MoE 架构，等效 7 倍激活参数的 Dense 模型性能，远超业内 MoE 架构约 3 倍杠杆的常规效率。此外，豆包大模型 1.5Pro 的多模态能力得到全面提升，新版豆包视觉理解模型 Doubao-1.5-vision-pro，视觉理解能力领先。全新的豆包实时语音模型 Doubao-1.5-realtime-voice-pro，采用 Speech2Speech 端到端框架，表现力实现飞跃，真正做到会哭会笑、能说方言会唱歌。该模型已在豆包 App 全量上线。
1.21	混元 3D 生成大模型 2.0	**腾讯混元 3D 生成大模型 2.0 开源发布，同步上线“业界首个一站式 3D 内容 AI 创作平台** 1 月 21 日消息，腾讯今日官宣开源上线混元 3D 生成大模型 2.0。腾讯混元还同步上线混元 3DAI 创作引擎，号称是“业界首个一站式 3D 内容 AI 创作平台”。该技术宣称一句话、一张图，甚至画个草图都能生成一个 3D 模型，甚至还能加动作、换纹理、捏人物、做动画。腾讯混元 3D-2.0 版本主要是对 3D 生成过程中的几何和纹理两个大模型进行了升级。几何大模型的任务就是捕捉 3D 物体的形状和结构。腾讯云采用 Hunyuan3D-DiT 和 HunyuanShapeVAE 技术，让生成的「白模」（没上色的模型）效果“堪比设计师手工建模”；纹理大模型 Hunyuan3D-Paint 可以根据文字或图片描述，为「白模」穿上各种纹理。
1.20	DeepSeek-R1	**DeepSeek-R1 发布，性能对标 OpenAIo1 正式版** DeepSeek 正式发布 DeepSeek-R1，并同步开源模型权重。DeepSeek-R1 遵循 MITLicense，允许用户通过蒸馏技术借助 R1 训练其他模型。DeepSeek-R1 上线 API，对用户开放思维链输出，通过设置 model='deepseek-reasoner'即可调用。DeepSeek 官网与 App 即日起同步更新上线。DeepSeek-R1 在后训练阶段大规模使用了强化学习技术，在仅有极少标注数据的情况下，极大提升了模型推理能力。在数学、代码、自然语言推理等任务上，性能比肩 OpenAIo1 正式版。

资料来源：IT 之家、中新网、TechWeb、格隆汇、Wind，中国银河证券研究院

表9：海外人工智能大模型动态

时间	模型	主要内容
1.31	o3-mini	**OpenAI 上线 o3-mini，首次向 ChatGPT 免费用户开放推理模型** 美国当地时间 1 月 31 日，OpenAI 宣布正式推出推理模型 o3-mini，是其推理系列中最新、最具成本效益的模型，即日起可在 ChatGPT 和 API 中使用。作为首款支持开发者高频需求功能的小型推理模型，OpenAI o3-mini 内置函数调用、结构化输出和开发者消息等专业功能，开箱即用，可直接投入生产环境。此外，开发者还可根据场景需求，灵活选择低、中、高三级推理强度，使模型在应对复杂挑战时能“深度思考”，或在需要快速响应时优先保证速度。

1.23	Operator	**OpenAI 放大招！重磅发布首个 AI 智能体，像人类一样使用网页浏览器，可自主订餐购物** 美国当地时间 1 月 23 日，美国初创公司 OpenAI 正式发布了其首个 AI 智能体 Operator。 与以往“问一句、答一句”的传统聊天机器人不同，Operator 能够在人类有限监督的情况下，按照预设指令自主完成任务，该创新被视为 AI 生产力发展的下一个重要里程碑。据 OpenAI 首席执行官奥特曼介绍，这款智能体能像人类一样使用网页浏览器， 并点击按钮、打字输入内容等复杂操作。 它能够自动完成预订旅行住宿、餐厅预约、在线购物等一系列日常生活中的繁琐任务 ，极大地提高了工作效率和便利性。在演示案例中，当用户要求 Operator 预订某家饭店的晚餐座位时，只需在对话框中输入简单的指令 ，如“给我订一个 XX 饭店今晚 19 点的桌子”，Operator 便能自动打开网页，进入预订网站，搜索并成功预订餐厅，这一过程无需人工干预。Operator 的技术核心在于 Computer-Using Agent（CU）模型， 该模型结合了 GPT-4 的视觉识别能力和基于强化学习的高级推理功能，使得 Operator 能够“看见”网页内容，并通过模拟鼠标和键盘操作与网页进行互动。
1.7	Nemotron	**英伟达发布 Nemotron 系列大语言模型 欲推动代理式 AI 加速崛起** CES 2025 大会上，英伟达创始人兼 CEO 黄仁勋发布了全新的 Llama Nemotron 系列大语言模型。黄仁勋表示，人工智能正在进入一个新时代——代理式人工智能（agentic AI），专业的 AI 代理可以帮助人们解决复杂问题并自动执行重复性任务。他进一步表示，借助定制的 AI 代理，各行各业的企业都可以实现前所未有的生产力。然而，这些先进的 AI 代理需要一套针对代理 AI 功能和能力进行优化的多个生成式 AI 模型系统。这种复杂性意味着对强大、高效的企业级模型的需求从未如此强烈。英伟达此次推出的 Llama Nemotron 模型，有 Nano、Super 和 Ultra 三个不同版本。其中，Nano 是最具成本效益、低延迟的模型，适合在 PC 和边缘设备上部署。Super 是一种高精度模型，在平衡计算效率的同时具有更高的准确性；而 Ultra 是最高精度模型，专为要求最高性能的数据中心规模应用而设计。

资料来源：21 财经、前瞻网、界面新闻，中国银河证券研究院

（三）算力端：AI 服务器、AI 芯片最新动态

表10：最新 AI 服务器、AI 芯片动态

时间	主要内容
1.28	**Nano Labs 投资人工智能 ASIC 芯片初创企业** Nano Labs 今日宣布对杭州微恒科技有限公司进行战略投资，获得该公司 5%股权。微恒科技专注于开发面向边缘计算、终端计算及大模型应用的人工智能专用计算存储一体化芯片，其产品可与 DeepSeek 最新大模型实现兼容。
1.25	**AI 芯片需求持续猛增! Meta 继续砸钱布局 AI，今年拟斥资 650 美元** 社交媒体 Facebook 与 Instagram 母公司 Meta Platforms 的首席执行官马克·扎克伯格周五表示，该科技巨头计划在 2025 年投资高达 650 亿美元用于与人工智能密切相关的项目，意味着继 2024 年疯狂砸钱超 380 亿美元投向人工智能等最前沿科技领域之后，Meta 今年将继续砸重金加码布局 AI，同时也大幅强化 AI 算力高景气预期：即 AI 芯片需求持续呈现井喷增长之势。据了解，高达 650 亿美元的 AI 相关支出计划包括新建设一个规模巨大的 AI 数据中心以及大幅扩充 AI 领域人才，增加 Meta 人工智能团队实际规模。
1.21	**5000 亿美元！特朗普宣布重磅 AI 项目“星际之门”** 华盛顿特区，21 日——美国总统唐纳德·特朗普今日在白宫宣布了一项雄心勃勃的计划，由甲骨文公司、OpenAI（美国开放人工智能研究中心）和日本软银集团共同出资 5000 亿美元，在美国建设名为“星际之门”（Stargate）的人工智能基础设施项目。特朗普在白宫椭圆形办公室与这三家科技巨头的负责人共同出席了发布会，向全世界宣告了这一重大消息。他激动地说：“‘星际之门’将不仅仅是一个数据中心，它将是一个支持新一代人工智能发展的物理和虚拟基础设施，为美国的创新和技术领导地位奠定坚实基础。”据美国媒体报道，“星际之门”项目的初始投资为 1000 亿美元，并计划在未来四年内逐步增

	加至 5000 亿美元。这一巨额投资将用于建设数据中心、研发新技术和推动人工智能在各个领域的广泛应用。
1.20	**欧冶半导体智能汽车 AI SoC 芯片及解决方案降低智能汽车维护和新产品开发成本，提升其智能化水平和安全性** 欧冶半导体由创始团队和国投招商共同发起设立，是国内首家聚焦智能汽车第三代 E/E 架构的系统级 SoC 芯片及解决方案供应商。股东阵容包括国投招商、鲲鹏大交通基金、丝路金桥基金、南山战新投、中科创星、招商致远、上汽、星宇股份、均胜电子、瑞声科技、保隆科技、虹软科技等，涵盖国有资本、产业资本、头部创投及众多汽车产业链龙头企业。目前，欧冶半导体以深圳为总部，分别在上海、珠海、苏州、西安多地设立研发中心，拥有员工近 300 人，90%以上为研发人员。目前已累计提交发明专利申请近百篇，先后通过 ISO 9001 质量体系认证、ISO 26262 功能安全开发流程及产品认证，并获得国家高新技术企业认定。
1.13	**美国公布最新 AI 芯片禁令 英伟达与甲骨文实名反对！** 中新网 1 月 14 日电 综合报道，当地时间 13 日，美国政府宣布推出美国制造 AI 芯片管制新规，旨在对美国制造的 AI GPU(图形处理器，主要用于 AI 大模型的训练及推理)芯片实施严格的全球出口限制。据美国全国广播公司(NBC)披露，根据管制新规，美国将对各个国家及地区，根据其部署的芯片计算能力被划分为三个等级，不同等级适用不同的销售限制。第一等级包括美国的主要盟友，如德国、荷兰、日本、韩国和新加坡、印度等 18 个国家和地区。这些国家几乎不受限制地使用美国厂商生产的 AI 芯片，并可以在其境内自由部署算力。第二等级则包括除第一梯队外的绝大多数国家，这些国家将面临总算力限制，每个国家在 2025 年至 2027 年期间最多可获得约 50000 个 AI GPU。第三等级主要是中国、俄罗斯、伊朗等被美国实施武器禁运的国家及地区。这些国家将受到最严格的限制，几乎全面禁止进口美国厂商生产的 AI GPU 芯片。

资料来源：智通财经、经济观察网，中国银河证券研究院

三、前沿行业动态

（一）前沿技术动态

1.Meta 提出大概念模型，1B 模型干翻 70B

Meta 提出大概念模型，抛弃 token，采用更高级别的「概念」在句子嵌入空间上建模，彻底摆脱语言和模态对模型的制约。

「大概念模型」（LCM）是在「句子表示空间」对推理（reasoning）建模，抛弃 token，直接操作高层级显式语义表示信息，彻底让推理摆脱语言和模态制约。具体而言，只需要固定长度的句子嵌入空间的编码器和解码器，就可以构造 LCM，处理流程非常简单：首先将输入内容分割成句子，然后用编码器对每个句子进行编码，以获得概念序列，即句子嵌入。然后，大概念模型（LCM）对概念序列进行处理，在输出端生成新的概念序列。最后，解码器将生成的概念解码为子词（subword）序列。

「大概念模型」（LCM）在推理（inference）效率上具备优势：在大约 1000 个 token 数左右，新模型理论上需要的计算资源就比 LLama2-7b 具备优势，且之后随着下上文中 token 数越大，新模型优势越大。具体结果见论文中的图 15，其中的蓝色表示 LLama2-7b 模型，红色和绿色分别代表新模型；红色的参数规模为 7b，而绿色为 1.6b；右图是左图在 0-3000 的 token 数下的局部放大图。

图3：「大概念模型」（LCM）推理效率检测

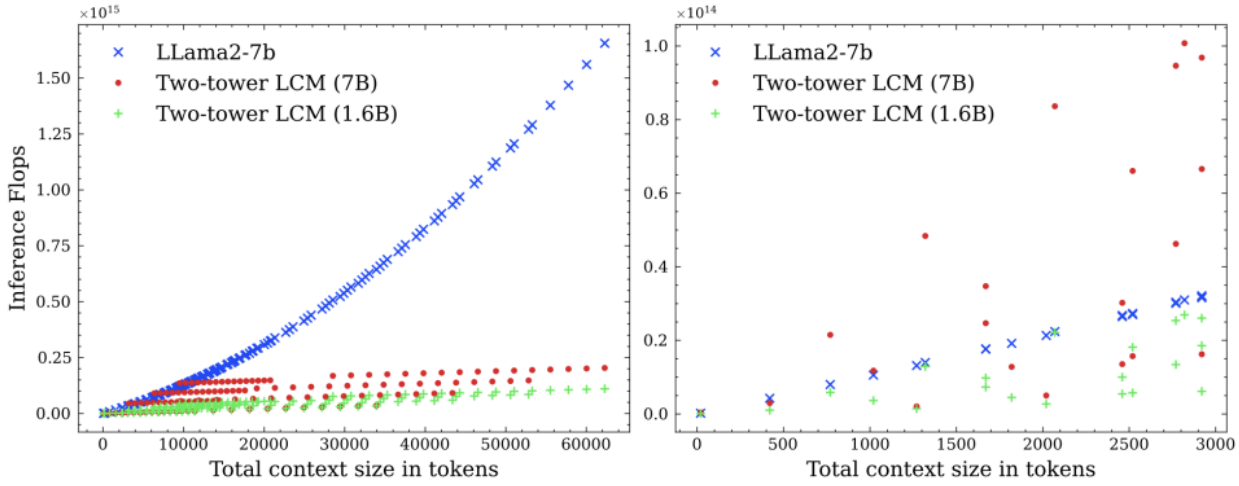


Figure 13 - Theoretical inference Flops of LCMs and LLLms. We evaluate the inference flops for different text lengths (in LLAMA2 tokens) with a variable average sentence length. Only extremely short sentences ($\leq$ 10 tokens) favor LLMs.

资料来源：新智元，中国银河证券研究院

新模型的其他亮点如下：

- **在抽象的语言和模态无关的层面上进行推理，超越 token：**（1）新方法模拟的是底层推理过程，而不是推理在特定语言中的实例。（2）LCM 可同时对所有语言和模态进行训练，即获取相关知识，从而有望以无偏见的方式实现可扩展性。目前支持 200 种语言文本。
- **明确的层次结构：**（1）提高长文输出的可读性。（2）方便用户进行本地交互式编辑。
- **处理长上下文和长格式输出：**原始的 Transformer 模型的复杂性随序列长度的增加而呈二次方增长，而 LCM 需要处理的序列至少要短一个数量级。
- **无与伦比的零样本（zero-shot）泛化能力：**LCM 可在任何语言或模态下进行预训练和微调。
- **模块化和可扩展性：**（1）多模态 LLM 可能会受到模态竞争的影响，而概念编码器和解码器则不同，它们可以独立开发和优化，不存在任何竞争或干扰。（2）可轻松向现有系统添加新的语言或模态。

2.港大 Aria-UI（纯视觉方案）登顶，超越 Claude 3.5

Aria-UI 通过纯视觉理解，实现了 GUI 指令的精准定位，无需依赖后台数据，简化了部署流程；在 AndroidWorld 和 OSWorld 等权威基准测试中表现出色，分别获得第一名和第三名，展示了强大的跨平台自动化能力。

Aria-UI 是一款专门面向 GUI 智能交互的创新型大规模多模态模型（LMM），颠覆性地实现了「看到即会操作」的自然交互范式 - 就像人类用户一样，AI 只需「观察」界面，即可理解并自主完成复杂的操作流程，从网页浏览、文件处理到系统设置等任务都能轻松应对。

在评估 AI 自动化操作能力的权威基准测试中，Aria-UI 配合 GPT-4o 展现出卓越表现：AndroidWorld 榜单排名第一，OSWorld 榜单排名第三！这一成绩不仅超越了业界领先的 Claude 3.5 Sonnet computer-use 接口，更展示了其在模拟人类操作电脑方面的强大能力。

Aria-UI 采用创新的 MoE (Mixture of Experts)架构，通过智能动态激活机制，将模型参

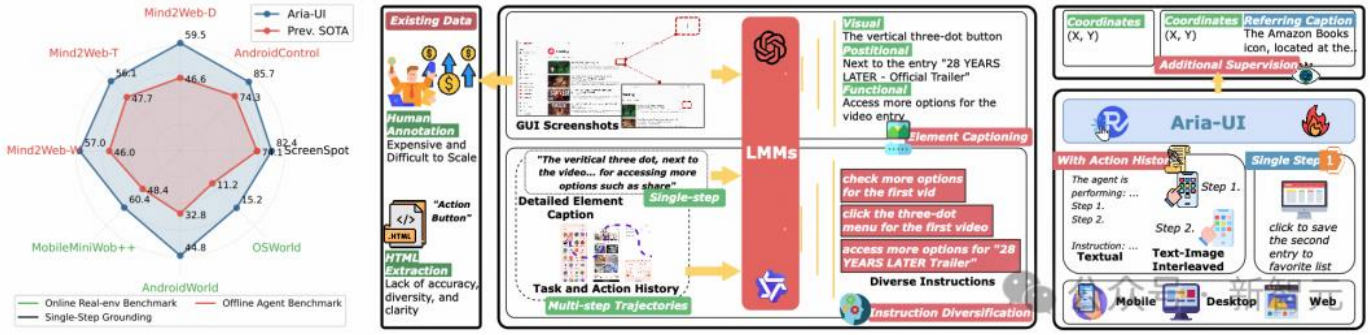
数需求压缩至仅 3.9B，同时保持较好的性能。这一突破性的轻量级设计带来多重优势：

- **极致压缩：**仅激活 3.9B 参数，大幅降低计算资源需求
- **高效推理：**优化的 MoE 架构确保快速响应和稳定性能
- **广泛适配：**支持在资源受限场景下的灵活部署
- **开放生态：**全面开源模型权重与训练数据
- **部署便利：**提供即用型 vLLM 推理脚本、支持主流 huggingface transformers 框架、完整的部署文档与示例

Aria-UI 的突破性创新：

智能指令适配引擎：Aria-UI 设计了数据生成 pipeline，通过自动合成海量高质量训练样本，为模型注入强大的指令理解能力。这套智能指令适配引擎使模型获得了卓越的泛化性能，能从容应对各类复杂任务场景，展现出非凡的环境适应能力，为实现真正的通用型 AI 助手奠定了坚实基础。

图4：Aria-UI 智能指令适配引擎



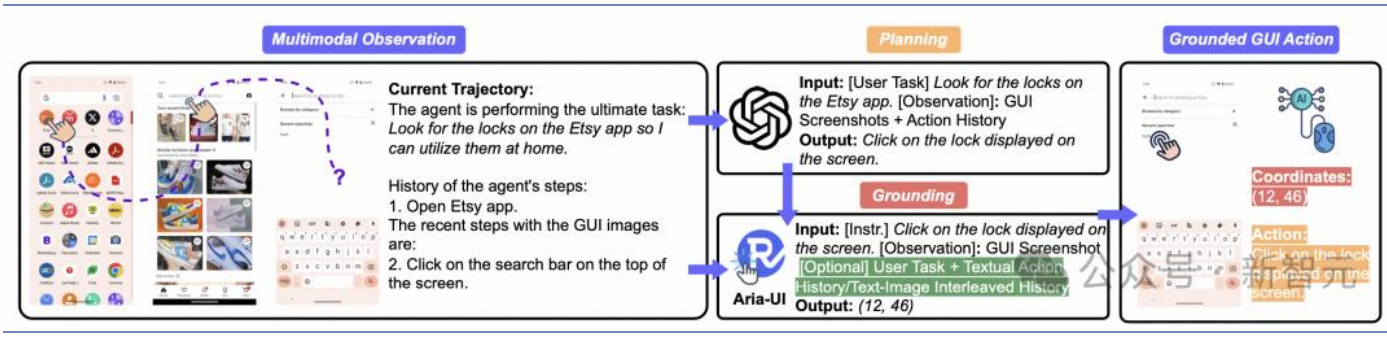
资料来源：新智元，中国银河证券研究院

动态上下文感知：为实现高精度的任务执行，Aria-UI 创新性地融合了多模态上下文理解机制。通过整合文本记录和图文操作历史，模型获得了强大的场景理解能力，能准确把握动态变化的操作环境，将复杂指令精准转化为具体行动。

全面性能测评：Aria-UI 在严格的性能评测中展现出令人瞩目的技术优势，成功刷新了多个领域基准的记录。在纯视觉人机交互基准测试中，其表现远超现有最佳视觉模型；在与需要调用 AXTree 等额外信息的传统方案对比中，Aria-UI 仅依靠视觉理解就取得了显著的性能提升。实验测评不仅验证了纯视觉方法的可行性，更展示了其在界面自动化领域(GUI Grounding)的应用潜力。

从日常生活场景到专业工作领域，GUI 智能体正在重塑人机交互的方式，为任务自动化开辟新天地。如图所示，一个完整的 GUI 智能体运作可分为两大核心阶段：决策规划（Planning）和视觉定位（Grounding）。在决策规划阶段，智能体通过分析当前界面状态，制定执行任务的具体策略；而在视觉定位阶段，则需要将规划好的指令精准映射到实际界面元素上，确保操作的准确执行。

图5：一个完整 GUI 智能体运作的两大核心阶段：决策规划和视觉定位



资料来源：新智元，中国银河证券研究院

尽管大规模多模态模型（LMMs）在决策规划方面取得显著进展，特别是在链式推理（CoT）和模型扩展等技术的加持下，但如何实现语言指令到 GUI 元素的精准定位仍然面临重大挑战。这些挑战主要体现在三个层面：

- **跨设备兼容性：**不同设备间界面布局存在巨大差异，要求模型具备强大的适应能力
- **指令多样性：**规划指令在形式和内容上变化多端，考验模型的理解能力
- **场景复杂性**：任务执行过程充满动态变化，对模型的实时响应能力提出更高要求

这些挑战不仅推动着 GUI 智能体技术的持续创新，也为打造更智能、更实用的自动化解决方案指明了方向

3．微软全华人团队提出 rStar-Math 算法，在数学推理上击败 o1

微软全华人团队提出 rStar-Math 算法，证明了 SLM 无需从高级模型蒸馏，就能在数学推理上，媲美甚至一举超越 o1。rStar-Math 核心在于，让小模型具备「深度思考」的能力。团队借鉴了 AlphaGo 中蒙特卡洛树搜索（MCTS）技术，设计了一个由 2 个协同工作的 SLM 组成的系统：

- 一个数学策略小语言模型（SLM）
- 一个过程奖励模型（PRM）

此外，rStar-Math 具体设计中，引入了三项技术创新：全新代码增强 CoT 数据合成；全新 PRM 训练方法；自我进化方案。通过 4 轮自我进化，并结合数百万个为 747k 数学问题合成的解答，rStar-Math 让 SLM 数学推理能力刷新 SOTA。

在 MATH 基准测试中，它将 Qwen2.5-Math-7B 的成绩从 58.8%提升至 90.0%，将 Phi3-mini-3.8B 的成绩从 41.4%提升至 86.4%，比 o1-preview 分别高+4.5%和+0.9%。

在美国数学奥林匹克（AIME）上，rStar-Math 解决了平均 53.3%（8/15）的题目，排名位于高中数学优等生前 20%。具体结果如下所示。

图6：美国数学奥林匹克（AIME）rStar-Math 排名位于高中数学优等生前 20%

Task (pass@1 Acc)	rStar-Math (Qwen-7B)	rStar-Math (Qwen-1.5B)	rStar-Math (Phi3-mini)	OpenAI o1-preview	OpenAI o1-mini	QWQ 32B-preview	GPT-4o	DeepSeek-V3
MATH	90.0	88.6	86.4	85.5	90.0	90.6	76.6	90.2
AIME 2024	53.3	46.7	43.3	44.6	56.7	50.0	9.3	39.2
Olympiad Bench	65.6	64.6	60.3	-	65.3	61.2	43.3	55.4
College Math	60.5	59.3	59.1	-	57.8	55.8	48.5	58.9
Omni-Math	50.5	48.5	46.0	52.5	60.5	49.6	30.5	35.9

资料来源：新智元，中国银河证券研究院

Keras 之父预言道，2025 年将会不断涌现这样的研究，通过结合程序搜索、CoT 搜索，在 LLM 指导下提升推理基准（包括 ARC 和数学基准）的表现。

4. 谷歌提出的 Titans 突破了传统 Transformer 在长序列处理中的局限

谷歌团队提出的 Titans 架构通过引入神经长期记忆模块，突破了传统 Transformer 架构在长序列处理中的局限。该架构通过创新的记忆整合和遗忘机制，在语言建模、常识推理、时间序列预测等任务中展现了显著的性能提升，在长上下文任务中的优势突出。

Titans 是什么？研究者认为大多数现有架构将记忆视为由输入引起的神经更新，并将学习定义为在给定目标的情况下有效获取有用记忆的过程。由于记忆分为短期记忆、工作记忆和长期记忆，而其中每个部分都相互独立地服务于不同的场景，也具有不同的神经结构。

受此启发，研究者提出了两个问题：

1. 如何设计一个高效架构，将不同且相互关联的记忆模块整合起来？

2. 是否需要一个深度记忆模块，以有效存储和记住长期历史信息？

本研究旨在通过设计一个长期神经记忆模块来解决上述问题，神经长期记忆模块的设计受到人类长期记忆系统的启发，能存储和检索过去的信息。该模块不是无差别地记住所有信息，而是会通过「惊讶度」来选择性地记住那些重要或令人惊讶的信息。并且其记忆不是静态的，可以根据新的信息动态更新。这种动态更新机制类似于人类的学习过程，使得模型能够不断适应新的数据和任务需求。为了更好地管理有限的内存，模块引入了衰减机制。该机制根据记忆的大小和数据的惊讶程度来调整记忆的权重，从而优化内存管理。

长期神经记忆模块设计完成后，面临的一个关键问题是如何把记忆高效地整合进深度学习架构。研究者提出了 Titans 架构，由三个模块构成：

- **核心模块（Core）：**包含短期记忆，负责主要的数据处理流程，采用具有有限窗口大小的注意力机制。
- **长期记忆模块（Long-term Memory）：**此模块是研究者设计的神经长期记忆模块，负责存储和记住远距离的历史信息。
- **持久记忆模块（Persistent Memory）：**这是一组可学习但与数据无关的参数，主要用于对任务知识进行编码，为模型提供先验知识储备。

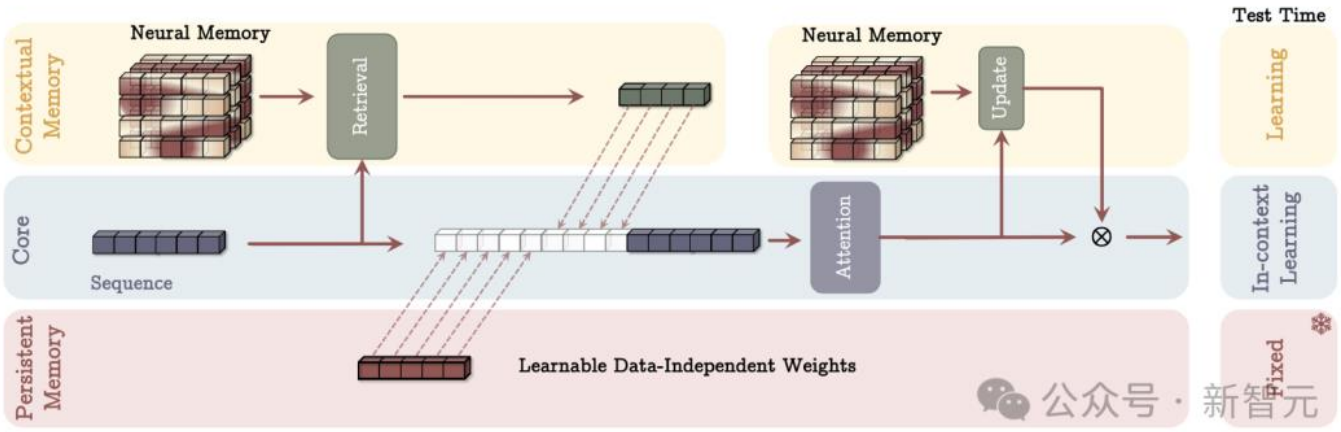
在此基础上，研究者提出了 Titans 架构的三种变体：

（1）记忆作为上下文（MAC）架构

核心分支把对应的长期记忆、持久记忆和当前输入信息拼接在一起，然后用注意力机制来处理上下文，并决定哪些信息应存储在长期记忆中。在测试时，与上下文记忆对应的参数仍在学习，

与核心分支对应的参数负责上下文学习，而持久记忆的参数则负责存储任务相关知识，因此是固定的。

图7：记忆作为上下文（MAC）架构

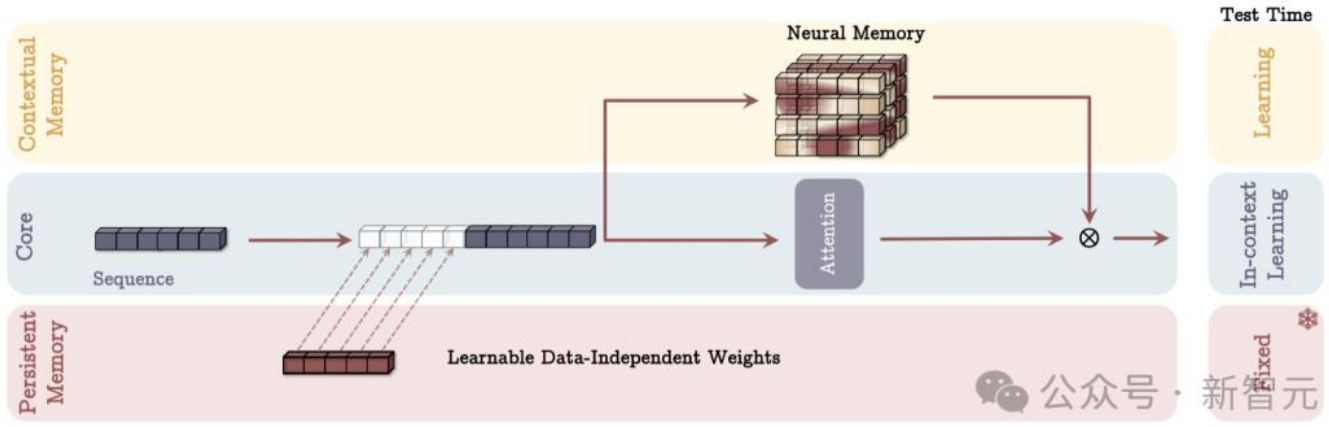


资料来源：新智元，中国银河证券研究院

（2）记忆作为门控（MAG）架构

在此架构中，一个分支用输入数据更新长期记忆，另一个分支使用滑动窗口注意力（SWA），最后将两者结果通过门控机制组合。在此设计中，滑动窗口注意力充当精确的短期记忆，而神经记忆模块则作为模型的衰减记忆。这种架构设计也可视为一种多头架构，其中头的结构各不相同。与 MAC 架构不同的是，MAG 架构仅将持久记忆融入上下文，并通过门控机制将记忆与核心分支结合。门控机制决定了来自持久记忆的信息在多大程度上影响核心分支的处理结果。

图8：记忆作为门控（MAG）架构

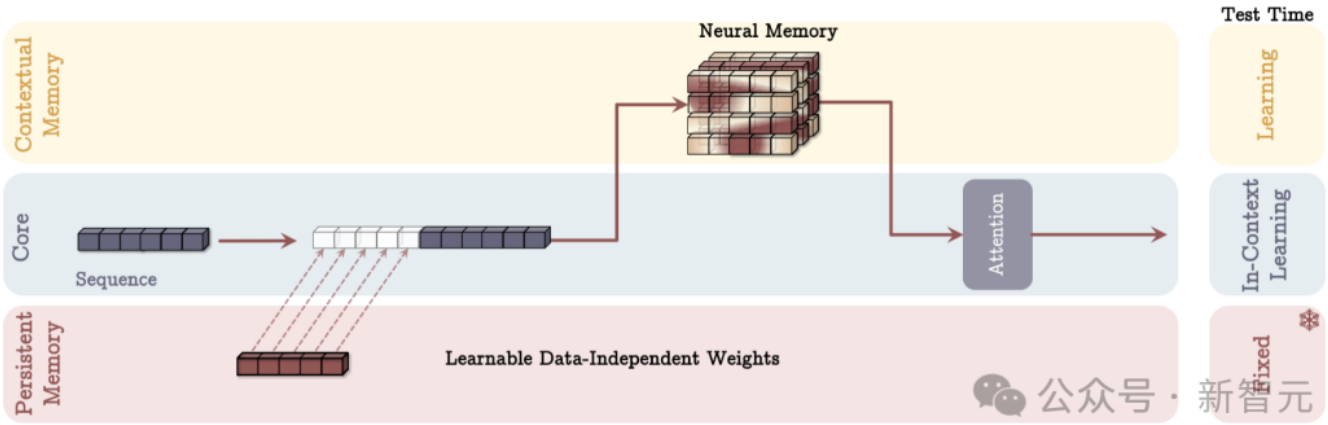


资料来源：新智元，中国银河证券研究院

（3）记忆作为层（MAL）架构

将神经记忆模块作为深度神经网络的一层，结合滑动窗口注意力机制。记忆层的核心功能是对过去和当前的上下文信息进行压缩处理，之后将处理结果传递给注意力模块

图9：记忆作为层（MAL）架构



资料来源：新智元，中国银河证券研究院

实验结果：

在语言建模及常识推理任务中，对 340M、400M、760M 等不同参数规模下的 Titans 变体与多种基线模型进行对比。非混合模型里，Titans (LMM) 在困惑度和准确率上表现优异。混合模型对比中，Titans 的三个变体均比基线模型更好。MAC 和 MAG 整体性能高于 MAL，能更好地整合注意力和记忆模块。

图10：语言建模及常识推理任务中 Titans 变体与多种基线模型对比表现

Model	Wiki. ppl ↓	LMB. ppl ↓	LMB. acc ↑	PIQA acc ↑	Hella. acc_n ↑	Wino. acc ↑	ARC-e acc ↑	ARC-c acc_n ↑	SIQA acc ↑	BoolQ acc ↑	Avg. ↑
340M params / 15B tokens											
Transformer++	31.52	41.08	30.76	62.98	34.76	50.53	45.21	24.05	36.81	58.24	42.92
RetNet	32.50	49.73	28.24	62.61	34.15	50.91	44.27	23.62	36.79	59.72	42.54
GLA	28.51	43.02	28.73	64.05	35.96	50.00	54.19	24.29	37.13	58.39	44.09
Mamba	30.83	40.21	29.94	63.79	35.88	49.82	49.24	24.56	35.41	60.07	43.59
DeltaNet	28.65	47.30	28.43	63.52	35.95	49.63	52.68	25.37	37.96	58.79	44.04
TTT	27.44	34.19	30.06	63.97	35.71	50.08	53.01	26.11	37.32	59.83	44.51
Gated DeltaNet	27.01	30.94	34.11	63.08	38.12	51.60	55.28	26.77	34.89	59.54	45.42
Titans (LMM)	26.18	29.97	34.98	64.73	39.61	51.85	55.60	28.14	34.52	59.99	46.17
Titans (MAC)*	25.43	28.13	36.00	65.32	40.35	51.21	58.17	29.00	38.63	60.18	47.36
Titans (MAG)*	25.07	28.72	36.71	64.88	40.56	52.49	57.72	28.16	39.75	60.01	47.54
Titans (MAL)*	24.69	28.80	35.74	64.97	39.44	51.97	56.58	28.21	38.14	57.32	46.55
400M params / 15B tokens											
Transformer++	30.63	37.37	29.64	64.27	37.72	51.53	54.95	27.36	38.07	61.59	45.64
RetNet	29.92	46.83	29.16	65.23	36.97	51.85	56.01	27.55	37.30	59.66	45.47
HGRN2	32.33	47.14	26.12	64.52	35.45	52.24	55.97	25.51	37.35	59.02	44.52
GLA	27.96	36.66	27.86	65.94	37.41	49.56	56.01	26.36	38.94	59.84	45.24
Mamba	29.22	39.88	29.82	65.72	37.93	50.11	58.37	26.70	37.76	61.13	45.94
Mamba2	26.34	33.19	32.03	65.77	39.73	52.48	59.00	27.64	37.92	60.72	46.91
DeltaNet	27.69	44.04	29.96	64.52	37.03	50.82	56.77	27.13	38.22	60.09	45.57
TTT	26.11	31.52	33.25	65.70	39.11	51.68	58.04	28.99	38.26	59.87	46.86
Gated DeltaNet	25.47	29.24	34.40	65.94	40.46	51.46	59.80	28.58	37.43	60.03	47.26
Samba*	25.32	29.47	36.86	66.09	39.24	51.45	60.12	27.20	38.68	58.22	47.23
Gated DeltaNet-H2*	24.19	28.09	36.77	66.43	40.79	52.17	59.55	29.09	39.04	58.56	47.69
Titans (LMM)	25.03	28.99	35.21	65.85	40.91	52.19	59.97	29.20	38.74	60.85	47.83
Titans (MAC)*	25.61	27.73	36.92	66.39	41.18	52.80	60.24	29.69	40.07	61.93	48.65
Titans (MAG)*	23.59	27.81	37.24	66.80	40.92	53.21	60.01	29.45	39.91	61.28	48.60
Titans (MAL)*	23.93	27.89	36.84	66.29	40.74	52.26	59.85	29.71	38.92	58.40	47.87
760M params / 30B tokens											
Transformer++	25.21	27.64	35.78	66.92	42.19	51.95	60.38	32.46	39.51	60.37	48.69
RetNet	26.08	24.45	34.51	67.19	41.63	52.09	63.17	32.78	38.36	57.92	48.46
Mamba	28.12	23.96	32.80	66.04	39.15	52.38	61.49	30.34	37.96	57.62	47.22
Mamba2	22.94	28.37	33.54	67.90	42.71	49.77	63.48	31.09	40.06	58.15	48.34
DeltaNet	24.37	24.60	37.06	66.93	41.98	50.65	64.87	31.39	39.88	59.02	48.97
TTT	24.17	23.51	34.74	67.25	43.92	50.99	64.53	33.81	40.16	59.58	47.32
Gated DeltaNet	21.18	22.09	35.54	68.01	44.95	50.73	66.87	33.09	39.21	59.14	49.69
Samba*	20.63	22.71	39.72	69.19	47.35	52.01	66.92	33.20	38.98	61.24	51.08
Gated DeltaNet-H2*	19.88	20.83	39.18	68.95	48.22	52.57	67.01	35.49	39.39	61.11	51.49
Titans (LMM)	20.04	21.96	37.40	69.28	48.46	52.27	66.31	35.84	40.13	62.76	51.56
Titans (MAC)	19.93	20.12	39.62	70.46	49.01	53.18	67.86	36.01	41.87	62.05	52.51
Titans (MAG)	18.61	19.86	40.98	70.25	48.94	52.89	68.23	36.19	40.38	62.11	52.50
Titans (MAL)	19.07	20.33	40.05	69.99	48.82	53.02	67.54	35.65	30.98	61.72	50.97

资料来源：新智元，中国银河证券研究院

在 S-NIAH 任务里，基于 RULER 基准测试，对 2K、4K、8K 和 16K 长度序列予以评估。神经记忆模块相较基线模型优势显著。在 Titans 变体中，MAC 性能最佳。

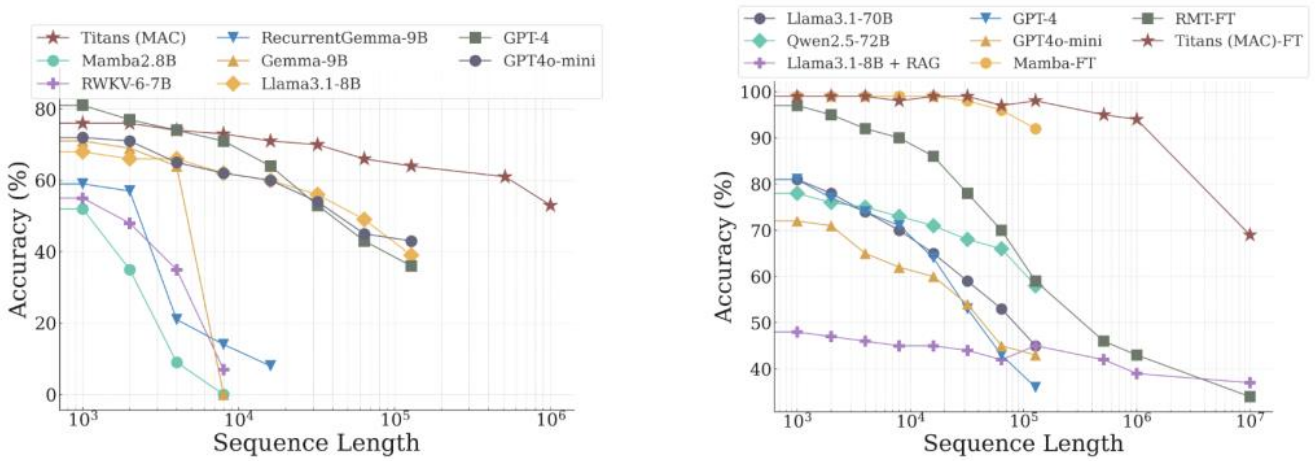
图11：S-NIAH 任务里，神经记忆模块相较基线模型优势显著

Model	S-NIAH-PK				S-NIAH-N				S-NIAH-W			
	2K	4K	8K	16K	2K	4K	8K	16K	2K	4K	8K	16K
TTT	98.4	98.8	98.0	88.4	60.2	36.6	10.2	4.4	78.8	28.0	4.4	0.0
Mamba2	98.6	61.4	31.0	5.4	98.4	55.8	14.2	0.0	42.2	4.2	0.0	0.0
DeltaNet	96.8	98.8	98.6	71.4	47.2	15.4	12.8	5.4	46.2	20.0	1.6	0.0
Titans (LMM)	99.8	98.4	98.2	96.2	100.0	99.8	93.4	80.2	90.4	89.4	85.8	80.6
Titans (MAC)	99.2	98.8	99.0	98.4	99.6	98.2	97.6	97.4	98.2	98.2	95.6	95.2
Titans (MAG)	99.4	98.0	97.4	97.4	99.2	98.8	97.2	98.6	98.0	98.0	90.2	88.2
Titans (MAL)	98.8	98.6	98.8	97.8	99.8	98.1	96.8	96.4	98.0	97.4	92.0	90.4

资料来源：新智元，中国银河证券研究院

在 BABILong 基准测试中，Titans（MAC）展现了卓越的性能，能够有效扩展到超过 200 万的上下文窗口，超越了 GPT-4、Llama3+RAG 和 Llama3-70B 等大模型。Titans（MAC）的参数量远少于基线模型，展现出在长序列推理方面的高效性和强大能力。在微调设置环节，Titans（MAC）的表现更为出色。

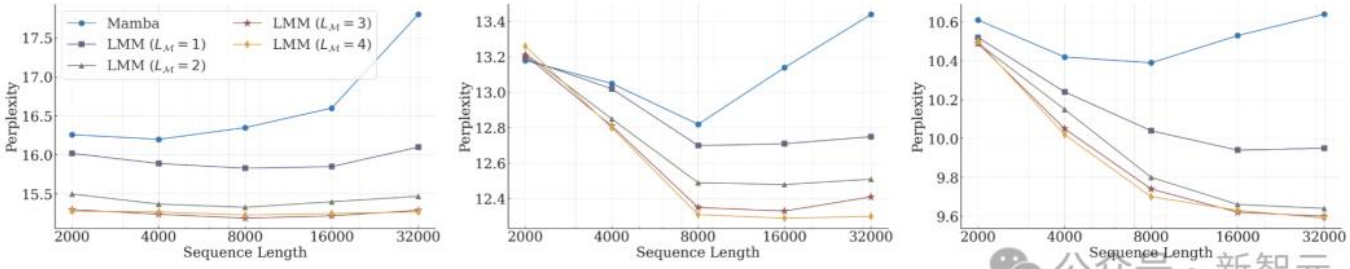
图12：BABILong 基准测试中，Titans (MAC) 展现了卓越的性能



资料来源：新智元，中国银河证券研究院

研究发现，增加记忆深度可提升模型在较长序列上的性能，并改善困惑度，但训练速度会因此降低，呈现出性能与效率之间的权衡。

图13：增加记忆深度对模型性能与效率的影响



(a) 170M Parameters (b) 360M Parameters (c) 760M Parameters

资料来源：新智元，中国银河证券研究院

通过在 Simba 框架中替换 Mamba 模块，并在 ETT、ECL、Traffic 和 Weather 等基准数据集上测试，神经记忆模块超越了所有的基线模型。这表明其在处理时间序列任务中的潜在优势。

图14：在 Simba 框架中替换 Mamba 模块测试表现

	Neural Memory		Simba		iTransformer		RLinear		PatchTST		Crossformer		TiDE		TimesNet		DLinear	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	0.358	0.387	0.383	0.396	0.407	0.410	0.414	0.407	0.387	0.400	0.513	0.496	0.419	0.419	0.400	0.406	0.403	0.407
ETTm2	0.261	0.309	0.271	0.327	0.288	0.332	0.286	0.327	0.281	0.326	0.757	0.610	0.358	0.404	0.291	0.333	0.350	0.401
ETTh1	0.420	0.421	0.441	0.432	0.454	0.447	0.446	0.434	0.469	0.454	0.529	0.522	0.541	0.507	0.458	0.450	0.456	0.452
ETTh2	0.336	0.382	0.361	0.391	0.383	0.407	0.374	0.398	0.387	0.407	0.942	0.684	0.611	0.550	0.414	0.427	0.559	0.515
ECL	0.162	0.261	0.169	0.274	0.178	0.270	0.219	0.298	0.205	0.290	0.244	0.334	0.251	0.344	0.192	0.295	0.212	0.300
Traffic	0.415	0.289	0.493	0.291	0.428	0.282	0.626	0.378	0.481	0.304	0.550	0.304	0.760	0.473	0.620	0.336	0.625	0.383
Weather	0.231	0.265	0.255	0.280	0.258	0.278	0.272	0.291	0.259	0.281	0.259	0.315	0.271	0.320	0.259	0.287	0.265	0.317

资料来源：新智元，中国银河证券研究院

在 DNA 建模任务中，Titans 架构也展示了其强大的长序列处理能力。实验结果表明，Titans 架构在这些任务中能够有效地利用历史信息，从而提高模型的性能。

图15：DNA 建模任务中，Titans 与其他架构对比

Model	Enhancer Cohn	Enhancer Ens	Human Reg.	Non-TATA Promoters	Human OCR Ens.
CNN	69.5	68.9	93.3	84.6	68.0
DNABERT	74.0	85.7	88.1	85.6	75.1
GPT	70.5	83.5	91.5	87.7	73.0
HyenaDNA	74.2	89.2	93.8	96.6	80.9
Transformer++	73.4	89.5	89.9	94.4	79.5
Mamba	73.0	-	-	96.6	-
Based	74.6	89.5	89.5	96.8	79.0
Neural Memory Module	75.2	89.6	89.3	96.6	79.9

资料来源：新智元，中国银河证券研究院

消融研究表明，神经记忆模块的所有组件对模型性能均有积极贡献，特别是权重衰减和动量。MAC 和 MAG 在语言建模和常识推理上表现相近，但 MAC 在长上下文任务中表现最佳。

图16：消融研究表明神经记忆模块的所有组件对模型性能均有积极贡献

Model	Language Modeling ppl ↓	Reasoning acc ↑	Long Context acc ↑
LMM	27.01	47.83	92.68
+Attn (MAC)	26.67	48.65	97.95
+Attn (MAG)	25.70	48.60	96.70
+Attn (MAL)	25.91	47.87	96.91
Linear Memory	28.49	46.97	85.34
w/o Convolution	28.73	45.82	90.28
w/o Momentum	28.98	45.49	87.12
w/o Weight Decay	29.04	45.11	85.60
w/o Persistent Memory	27.63	46.35	92.49

资料来源：新智元，中国银河证券研究院

实验结果表明，Titans 架构在语言建模、常识推理、时间序列预测和 DNA 建模等任务中均表现出色，特别是在处理超 200 万上下文窗口任务中，能够有效地利用历史信息，提高模型的准确性。

5. Transformer 作者初创重磅发布 Transformer^2，可自行动态调整权重

Sakana AI 发布了 Transformer^{2}新方法，通过奇异值微调和权重自适应策略，提高了 LLM 的泛化和自适应能力。新方法在文本任务上优于 LoRA；即便是从未见过的任务，比如 MATH、HumanEval 和 ARC-Challenge 等，性能也都取得了提升。

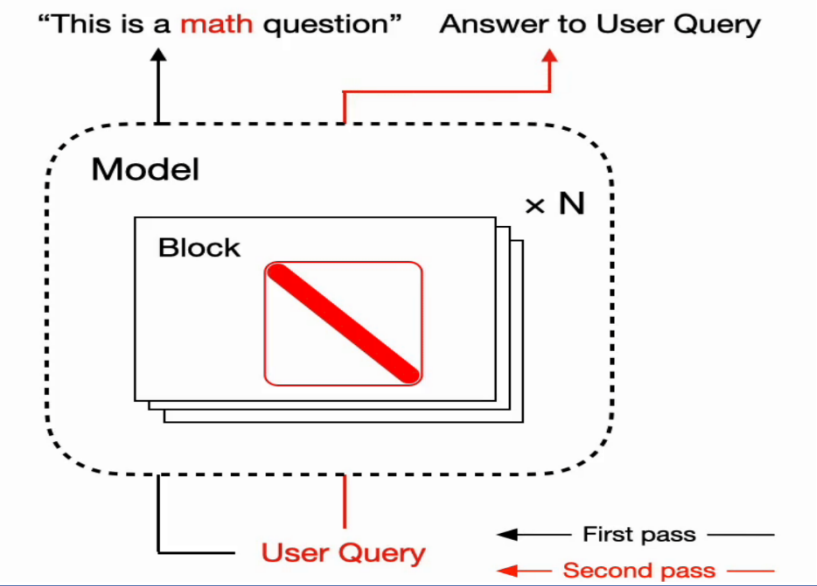
传统上，LLM 的后训练通过一次全面的训练来优化模型，使其具备广泛的能力。从简化的角度，这种「one shot」微调框架看起来很理想，**但在实际操作中却很难实现。**例如，后训练需要大量资源，导致计算成本和训练时间显著增加。此外，当引入更多样化的数据时，很难同时克服过拟合和任务干扰。

相比之下，自适应模型提供了一种更灵活高效的方法。与其一次性训练 LLM 来应对所有任务，不如开发专家模块，根据需求将其离线开发并增强到基础 LLM 中。**然而，创建多个专家模块，对 LLM 进行微调，显著增加了需要训练的参数数量，而且容易过拟合，模块之间的组合也不够灵活。**

对此，新框架通过有选择性地调整模型权重中的关键组件，让 LLM 能够实时适应新任务。Transformer^{2}的名称体现了它的两步过程：首先，模型分析传入的任务，理解其需求；然后应用任务专用的适应性调整，生成最佳结果。

Transformer^{2}在多种任务（如数学、编码、推理和视觉理解）中表现出了显著的进步，在效率和特定任务的表现上超越了传统静态方法如 LoRA，同时所需的参数大大减少。

图17：Transformer[2]的运作过程



资料来源：新智元，中国银河证券研究院

（二）前沿政策动态

表11：1 月人工智能相关政策法规

时间	部门	文件	内容
2025/1/2	杭州市经济和信息化局	《杭州市人工智能全产业链高质量发展行动计划（2024—2026 年）》	到 2026 年，力争全市智能算力集群规模在国内同类城市中领先，形成基础通用大模型 1 个以上、行业专用模型 20 个以上，建成人工智能特色产业园区 10 个，集聚开源模型生态企业 1000 家以上，努力打造全国算力成本洼地、模型生态最优城市和人工智能产业发展高地。
2025/1/14	北京市科委、中关村管委会	《北京市加快推动“人工智能+新材料”创新发展行动计划（2025-2027 年）（征求意见稿）》	到 2027 年，北京“人工智能+新材料”创新能力显著增强，新材料研发服务业态培育取得积极进展，形成国际领先的新材料创新策源与人工智能应用高地，构筑全球竞争新优势。 （一）创新能力位居全球前列。产生一批重大原创性成果，突破一批产业亟需核心关键技术，在全球率先发布新一代物质科学大原子模型，研发 10 个（套）以上国际领先的垂类模型和自主核心软件，形成 15 个人工智能赋能的标杆性新材料产品，实现应用示范。 （二）支撑体系基本成型。建成新材料大数据中心服务门户、数据资源节点集群，建立材料数据标准规范体系，建成若干个新材料智能实验室和应用赋能公共服务平台，打造 1 个“人工智能+新材料”融合创新示范基地。 （三）新模式新业态加快涌现。探索培育新材料 CRO 服务业态，培育 5-8 家独角兽企业和潜在独角兽企业，100 家创新型企业。
2025/1/21	世界经济论坛	《智能时代的产业发展》	系列报告为各关键行业实现人工智能的负责任、规模化应用提供务实洞见。同时还发起了 Frontier MINDS 计划，旨在聚焦并推广对于应对全球挑战具有重大影响的人工智能解决方案。首批入选的解决方案预计将于 2025 年公布
2025/1/26	国家自然科学基金委员会	《可解释、可通用的下一代人工智能方法重大研究计划 2025 年度项目指南》	2025 年度资助研究方向： （一）培育项目：神经网络的新架构和新的预训练或自监督学习方法；深度学习的基础理论；大模型的基础问题；以数据为中心的机器学习；科学领域的人工智能方法与理论

			（二）重点支持项目：融合逻辑和深度学习的推理方法；融合物理与人工智能的几何生成；新一代脑启发的人工智能；类人认知学习框架；物理过程驱动的多智能体仿真场景可信生成；可解释的人工智能方法及其在化学反应复杂体系中的应用；人工智能驱动的虚拟细胞研究；罕见病诊断决策大模型；基于多模态大模型的耐受极端环境生物元件设计 （三）集成项目：记忆与推理分离、分层的通用大模型； 结构材料构效关系的构筑方法与应用；融合环境-系统-模型的智能操作系统

资料来源：杭州市经济和信息化局官网，北京市人民政府网，世界经济论坛官网，国家自然科学基金委员会官网，中国银河证券研究院

四、前沿企业动态

（一）前沿产品动态

1.“天工大模型 4.0”o1 版/4o 版在网页端和 APP 端正式上线，具备超强逻辑推理

1 月 6 日，天工大模型 4.0 o1 版/4o 版正式上线天工网页端和 APP。底座大模型，正式进化到「天工 4.0」。「天工大模型 4.0」 o1 版（Skywork o1）的上线，意味着国内首款中文逻辑推理能力的 o1 模型来了！数学高考题、考研题、奥数题，Skywork o1 都能靠自己的逐步思考破解。Skywork o1 并不是简单地复现 OpenAI o1 模型的工作。它不仅在模型输出上内生了思考、计划、反思等能力，还在模型真正拥有了思考和反思之后，带来了推理能力的提升。

Skywork o1 为何能在逻辑推理任务上有如此大幅提升？得益于天工三阶段自研的训练方案。

推理反思能力训练。首先，在推理训练方面，团队通过自主研发的多智能体体系，构建出了高质量的分步推理、反思与验证数据。然后，用这些高质量且多样化的长思考数据，对基座模型进行继续预训练和监督微调，并在版本迭代中采用大规模的自蒸馏和拒绝采样，从而显著提升了模型的训练效率和逻辑推理能力。

推理能力强化学习。其次，在强化学习阶段，团队创新性地提出了一种适配分步推理强化的奖励模型——Skywork o1 Process Reward Model（PRM）。在最新的版本中，团队将 Skywork-PRM 的应用范围，从原本侧重的数学和代码领域，拓展到了常识推理、逻辑推演和伦理决策等更广泛的场景中。同时，还针对写作、闲聊等通用领域以及多轮对话构建了专门的训练数据，实现了全场景覆盖。此外，团队重点提升了 Skywork-PRM 的模块化评估能力，特别是在处理 o1 风格思维链方面，优化了试错和反思验证机制。通过更细致的评估体系，为强化学习和搜索过程提供了更精准的奖励信号指导。

推理 planning。最后，在推理的规划方面，团队通过自研的 Q*线上推理算法，以及模型的在线思考能力，实现了最优推理路径的寻找。概括来说，Q*算法通过借鉴人类大脑中「System 2」的思考方式，将 LLM 的多步推理过程抽象为一个启发式搜索问题。然后，再通过 Q*线上推理框架与模型在线思考的结合，实现了推理过程中的精细规划，进而指导 LLM 的解码过程。Q*算法的成功落地，不仅显著提升了模型的线上推理能力，同时也标志着 Q*算法的全球首次实现和公开。

图18：Q*算法全球首次实现和公开

Q*: Improving Multi-step Reasoning for LLMs with Deliberative Planning

Chaojie Wang[1]* Yanchen Deng[2]* Zhiyi Lyu[2] Liang Zeng[1] Jujie He[1]
Shuicheng Yan[1] Bo An[12]
[1]Skywork AI [2]Nanyang Technological University

资料来源：新智元，中国银河证券研究院

进一步的，团队基于 Q*算法对推理系统进行了全面优化。

第一点是模块化的树形结构推理：团队通过高质量、多样化的长思考数据对 Skywork o1 进行预训练和监督微调，使模型具备了对整个推理流程进行系统规划，自动将回答按层次展开，同时在推理过程中融入自我反思和验证环节的结构化输出能力。此外，还创新性地利用以「模块」为单位的规划方式，取代了传统的以「句子」为单位的方法。既提升了规划效率，也使 PRM 能够基于更完整的模块化回答进行准确判断和推理指导。

第二点是自适应的搜索资源分配：针对现有 o1 风格模型存在的过度思考问题，团队开发出了一种全新的自适应搜索资源分配机制。也就是，通过对用户 query 进行难度预估，自适应地控制搜索树的宽度和深度，进而实现简单问题快速响应、复杂问题多轮验证的动态平衡，有效提升了系统的计算效率和回答准确率。

最终，Skywork o1 在 GSM8k，MATH，OlympiadBench，AIME-24 和 AMC-23 标准数学基准测试，以及 HumanEval、MBPP、LiveCodeBench 和 BigCodeBench 代码基准测试中，性能显著优于常规通用大模型，表现仅次于 o1-mini。

图19：Skywork o1 在各项标准数学基准测试中表现

Table 1: Performance on maths benchmarks.

Model	GSM8K	MATH	OlympiadBench	AIME-24	AMC-23
o1-mini	93.6	93.7	61.8	60.0	90.0
GPT4o	87.4	73.1	39.7	20.0	57.5
Claude-3.5-Sonnet	95.5	71.2	30.2	16.7	37.5
Qwen2.5-72B-instruct	95.8	83.1	46.1	13.3	62.5
Qwen-QwQ	95.8	89.3	57.9	50.0	77.5
Deepseek v3	95.8	90.2	50.1	39.2	80.0
Skywork o1 Lite	95.7	90.0	57.2	40.0	82.5
Skywork o1 Preview	96.3	92.8	61.0	56.7	95.0

资料来源：新智元，中国银河证券研究院

图20：Skywork o1 在各项标准代码基准测试中表现

Table 2: Performance on code benchmarks.

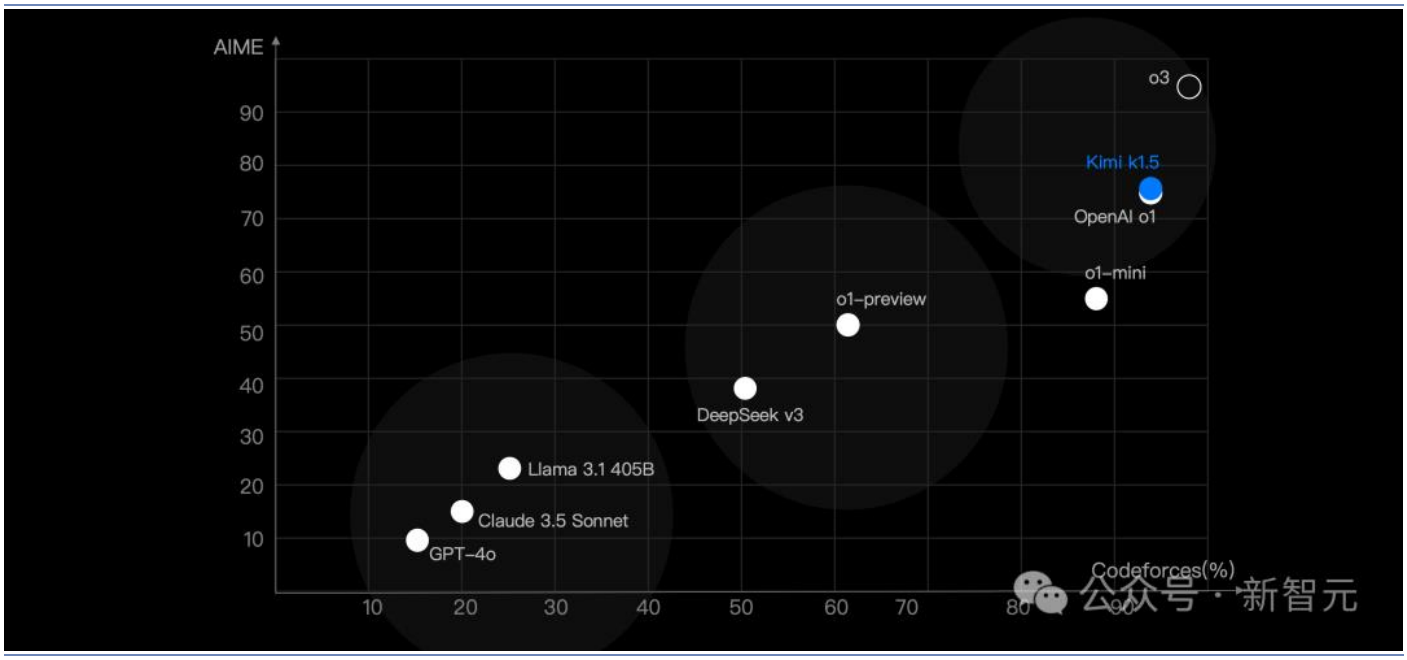
Model	HumanEval	MBPP	LiveCodeBench (2408-2411)	BigCodeBench
o1-mini	95.1	89.0	58.0	45.0
GPT4o	86.0	84.2	36.0	47.3
Claude-3.5-Sonnet	89.6	87.0	32.1	44.9
Qwen2.5-72B-instruct	86.6	88.2	30.4	46.4
Qwen-QwQ	87.2	83.2	50.0	45.6
Deepseek v3	90.9	89.0	40.5	48.2
Skywork o1 Lite	89.0	83.0	27.7	46.0
Skywork o1 Preview	94.5	93.4	44.6	45.4

资料来源：新智元，中国银河证券研究院

2. Kimi 发布了 k1.5 多模态思考模型

Kimi 发布了 k1.5 多模态思考模型。这是继去年 11 月他们发布 k0-math 数学模型，12 月发布 k1 视觉思考模型之后，连续第三个月带来 k 系列强化学习模型的重磅升级。Kimi k1.5 的性能，如今已经全面追上现役全球最强模型——OpenAI o1 满血版。

图21：全球前沿大模型数学竞赛和编程竞赛基准测试

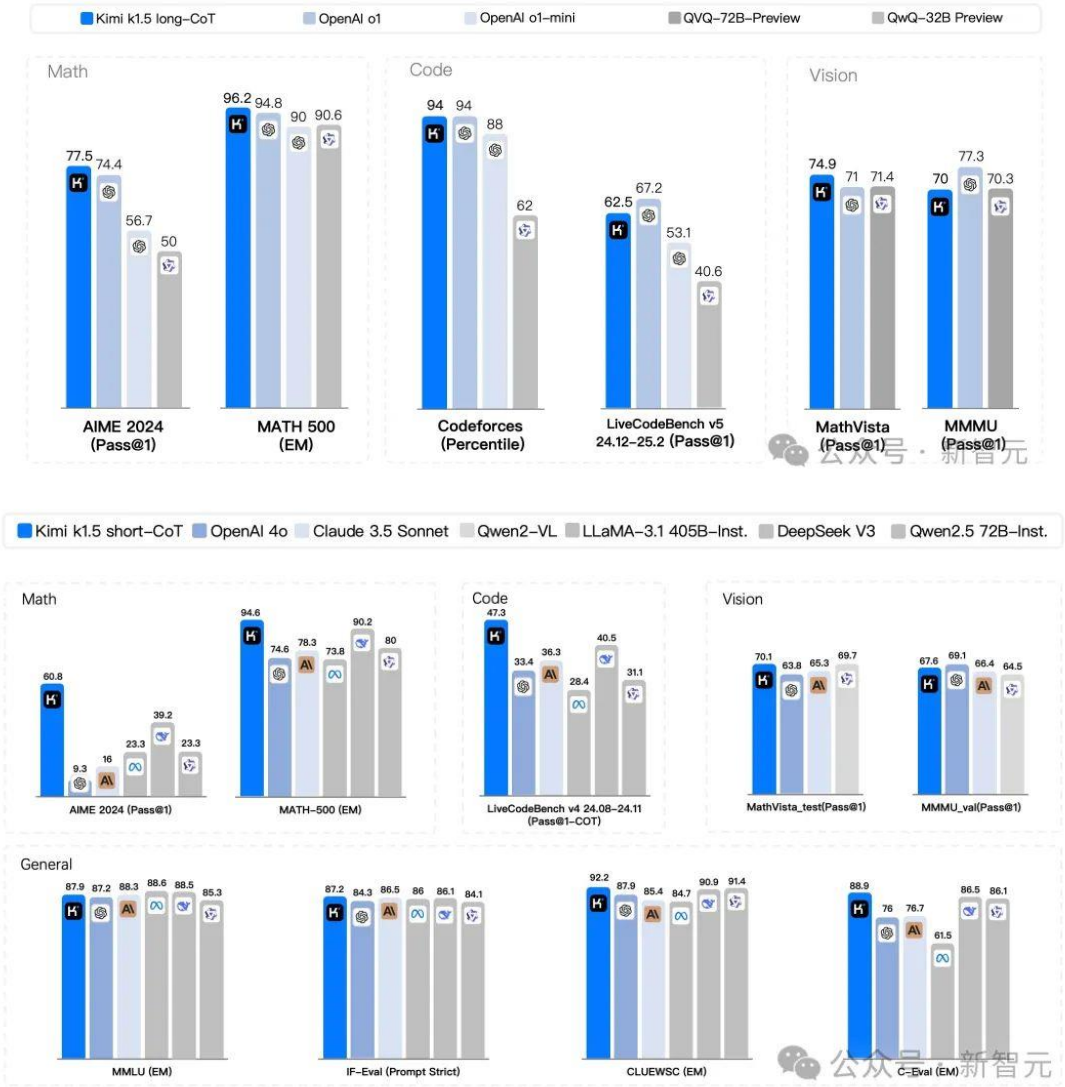


资料来源：新智元，中国银河证券研究院

具体来说，在 Long CoT 模式下，Kimi k1.5 的数学、代码、多模态推理能力，达到了长

思考 SOTA 模型 OpenAI o1 满血版的水平。这也是全球范围内，首次有 OpenAI 之外的公司达到。而在 Short CoT 模式下，Kimi k1.5 大幅领先 GPT-4o 和 Claude 3.5 的水平。

图22：短 COT 模式下，数学成绩显著高于 GPT-4o 和 Claude Sonnet 3.5



资料来源：新智元，中国银河证券研究院

3．豆包大模型 1.5 发布，训练不走蒸馏“捷径”

1 月底发布的豆包大模型 1.5，不仅多模态能力全面提升，霸榜多个基准；更难得的是，它在训练过程中从未使用过任何其他模型生成的数据，坚决不走蒸馏「捷径」。

豆包大模型 1.5 的模型基础能力，再次展现出超强进化，在多个公开测评基准中成绩亮眼。而它的多模态能力，无论语言、视觉理解还是实时语音，也都实现了全面领先。

图23：豆包大模型 1.5 与其他模型在综合指标上的对比

		Doubao-1.5-pro	Llama3.1-405B	GPT4o-0806	Gemini-exp-1205	Claude-3.5-Sonnet-latest	Qwen2.5	DeepseekV3
Knowledge	MMLU	88.6	88.6	**88.7**	86.8	88.5	85.6	88.5
	MMLU_PRO	**80.1**	73.3	74.9	76.4	78.0	71.1	75.9
	GPQA	**65.0**	51.1	53.1	62.1	**65.0**	49.0	59.1
MATH	Math	88.6	73.8	75.9	**89.7**	78.3	83.1	87.8
	OlympiadBench	59.8	34.1	40.7	**64.7**	43.5	50.0	59.1
Code	MBPP+	78.0	72.8	78.3	78.6	76.5	76.9	**79.3**
	McEval	**70.2**	58.7	68.2	67.0	68.2	61.7	69.4
	FullStackBench	**65.1**	53.6	61.8	62.6	60.3	56.9	63.3
Reasoning	BBH	91.6	89.2	91.7	**92.6**	**92.6**	88.3	92.3
	DROP	**93.0**	91.2	79.8	89.7	88.3	87.4	91.6
Instruction Following	IFEval	89.5	86.0	85.7	**89.8**	89.3	84.1	86.1
	SysBench	67.6	58.9	62.2	**69.0**	**69.0**	47.2	66.3
Chinese	CMMLU	**90.9**	75.4	77.3	84.3	81.2	84.3	83.5
	C-Eval	**91.8**	72.7	76.0	83.9	80.9	84.1	86.5

资料来源：新智元，中国银河证券研究院

图24：豆包大模型 1.5 与其他模型在视觉理解指标上的对比

	Benchmark	Doubao-1.5-pro	GPT4o-1120	Claude3.5-Sonnet	Gemini-2-flash	Qwen2-VL-72B	InternVL-2.5-78B
College-level Problems	MMMU(val)	**73.8**	70.7	70.4	70.7	64.5	70.1
	MMMU-Pro	**59.3**	54.5	54.7	57.0	46.2	48.6
Mathematical Reasoning	MathVision	**48.6**	30.4	38.3	41.3	25.9	32.2
	OlympiadBench	**48.5**	25.9	27.8	43.6	11.2	25.1
	MathVista	**78.8**	63.8	65.4	73.1	70.5	76.6
Document and Diagrams Reading	TextVQA(val)	84.7	81.4	76.5	75.6	**85.5**	83.4
	ChartQA(test avg.)	88.0	86.7	**90.8**	85.2	88.3	88.3
	InfoVQA(test)	**88.0**	80.7	74.3	77.8	84.5	84.1
	DocVQA(test)	**96.7**	91.1	95.2	92.1	96.5	95.1
	Chartxiv(RQ/DQ)	54.4 / 84.3	52.0 / **86.5**	**60.2**/ 84.3	55.2/81.8	43.0 / 81.3	42.4 / 82.3
General Visual Question Answering	RealWorldQA	**78.9**	75.4	66.6	74.5	77.8	78.7
	MMStar	71.9	63.9	65.1	69.4	68.6	**73.1**
	MMBench-en	87.5	83.5	81.7	83.0	85.9	**88.3**
	MMBench-cn	86.0	82.1	83.4	82.9	83.4	**88.5**
Spatial and Counting Understanding	Blink	**68.4**	68.0	59.6	62.6	61.1	63.8
	CountBench	**89.6**	85.1	86.8	88.2	88.6	84.1
Video Understanding	Video-MME	74.1	73.4	61.7	**78.2**	71.2	72.1
	EgoSchema-subest	75.4	74.8	64.4	71.8	80.6	78.2

资料来源：新智元，中国银河证券研究院

图25：豆包大模型 1.5 与其他模型在深度思考模型指标上的对比

		Doubao-1.5-pro -AS1-Preview	01-preview	01
AIME	pass@1	70.0	44.6	**74.4**
	cons@k	**86.7** (cons@32)	54.7 (cons@64)	83.3 (cons@64)

资料来源：新智元，中国银河证券研究院

豆包 1.5 在以下几方面实现了进化：

（1）视觉理解能力超强进化

视觉理解方面，团队这次在多模态数据合成、动态分辨率、多模态对齐、混合训练上进行了全面技术升级，让模型在视觉推理、文字文档识别、细粒度信息理解、指令遵循方面的能力进一步增强了。而且，模型的回复模式还变得更加精简、友好。现在，豆包大模型 1.5 能读懂不同分辨率和不同长宽比的图片，支持百万级分辨率，能更清晰得识别内容。

（2）语音多模态：真正实现了端到端的语音对话

这次豆包的语音多模态模型，真正实现了端到端的语音对话。语言表现力、控制力、情绪承接上堪称一绝，而且还低时延，对话中可随时打断。这都要归功于，团队提出的全新 Speech2Speech 端到端框架。

它通过原生方法将语音和文本模态进行深度融合，从而实现了语音理解生成端到端。并且，在语音对话效果上，它相比传统的 ASR+LLM+TTS 的级联方式有了质的飞跃！因此，它不仅拥有高理解力（高智商），还拥有前所未有的语音高表现力与高控制力，而模型整体在回复内容和语音上，还有了高情绪承接能力。而在框架设计上，研究者将语音和文本 Token 进行融合，为语音多模态数据的 Scaling 提供了必要条件。

在预训练阶段，他们开发了多样化的数据生产和使用方式，同时在训练上探索了多种有效方

案，通过 Scaling 最大化地将语音和文本能力进行深度融合。在后训练阶段，通过融合高表现力与智商数据的均衡， 数据筛选以及多模态 RL 阶段的专项能力提升让模型在智商、语音表现力等多方面达到最优。

豆包没有对任何其他模型进行过蒸馏。最近，中科院北大的一项研究引起了热议。他们发现，许多知名的闭源和开源大语言模型，都表现出了相当高的蒸馏程度！我们最常见的问题之一——A 模型说自己是 B 模型，就是因为它们「蒸过头」导致的。少数例外，也就是没有对任何其他模型进行过蒸馏的，就数 Claude、Gemini 和豆包了。

4. 智谱全球首个电脑智能体 GLM-PC 全新升级，具有“代码思维”

智谱率先卷入 L3 级使用工具能力，发布全球首个面向公众、回车即用的电脑智能体 GLM-PC。经过全新升级的 v1.1 版本，不仅能够像人类一样「观察」和「操作」计算机，自主完成各种复杂任务；而且还拥有「深度思考」模式，以及专门用来做逻辑推理和代码生成的功能。

比如：上传一张图片后，然后给出指令——「识别图片中的信息生成朋友圈文案，并发送一个朋友圈」。它首先会将任务分解成多个步骤，并对图片内容进行识别，生成相应配文。然后，AI 瞬间跳转到微信，打开朋友圈，将图片上传，再附上文案，一键发送就搞定了。

GLM-PC 的 Window 和 Mac 客户端已经同步上线了。

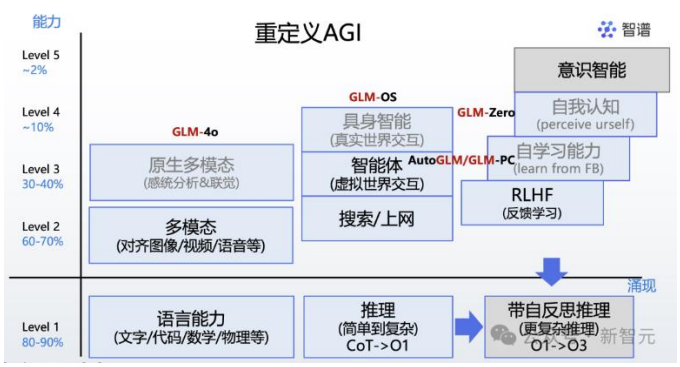
图26：GLM-PC 网页界面

注：试用地址 https://cogagent.aminer.cn

资料来源：新智元，中国银河证券研究院

智谱之所以能够成为智能体领域的先行者，离不开这家公司从成立伊始就绘制出的 AI 路线图。2024 年，智谱同样将 AGI 的实现划分了 5 级，能力从 L1 逐步攀升至 L5。在他们看来，AI 的能力早已突破了传统语言和文本处理限制。目前，AI 已经从语言/文本逐渐扩展到多模态、工具使用，未来还会有更多的自我认知。

图27：智谱定义 AGI 五等级



资料来源：新智元，中国银河证券研究院

图28：豆智谱对人工智能分级

	OpenAI	我们的思考
Level 1	有语言能力的AI	AI学会使用语言，在大多数自然语言任务上突破图灵测试
Level 2	人类水准的问题求解能力	AI学会求解问题，涌现世界知识和类人的复杂逻辑推理能力，在问题求解方面突破图灵测试
Level 3	使用工具，系统可以执行动作	AI学会使用工具，利用工具完成多数人类物理世界问题，在工具使用方面突破图灵测试
Level 4	AI将能自己发明创新	AI通过自我学习，实现GPT到GPT-zero的升级，具备自我批判、自我改进以及自我反思能力
Level 5	AI可以融入组织或者自成组织	AI能力全面超越人类，具备探究科学规律、世界起源等终极问题的能力

资料来源：新智元，中国银河证券研究院

GLM-PC v1.1 的推出，意味着智谱在 L3 级智能体的探索又有了新的进展。截至目前，智谱已经有了手机智能体 AutoGLM 和电脑智能体 GLM-PC 两大系统，实现了工具使用能力的深度突破。这两个系统分别覆盖了移动设备和桌面端——AutoGLM 在手机上，能够精准操控各类应用，实现跨场景智能交互；而 GLM-PC 则将电脑端的操作提升到了新的高度，基于视觉语言模型 VLM 的图形界面智能体 GUI Agent，实现逻辑推理与感知认知的结合，凸显出 AI 对复杂系统工具的掌控力。这些并非是简单功能的堆砌，而是对人机交互范式的根本性重塑。

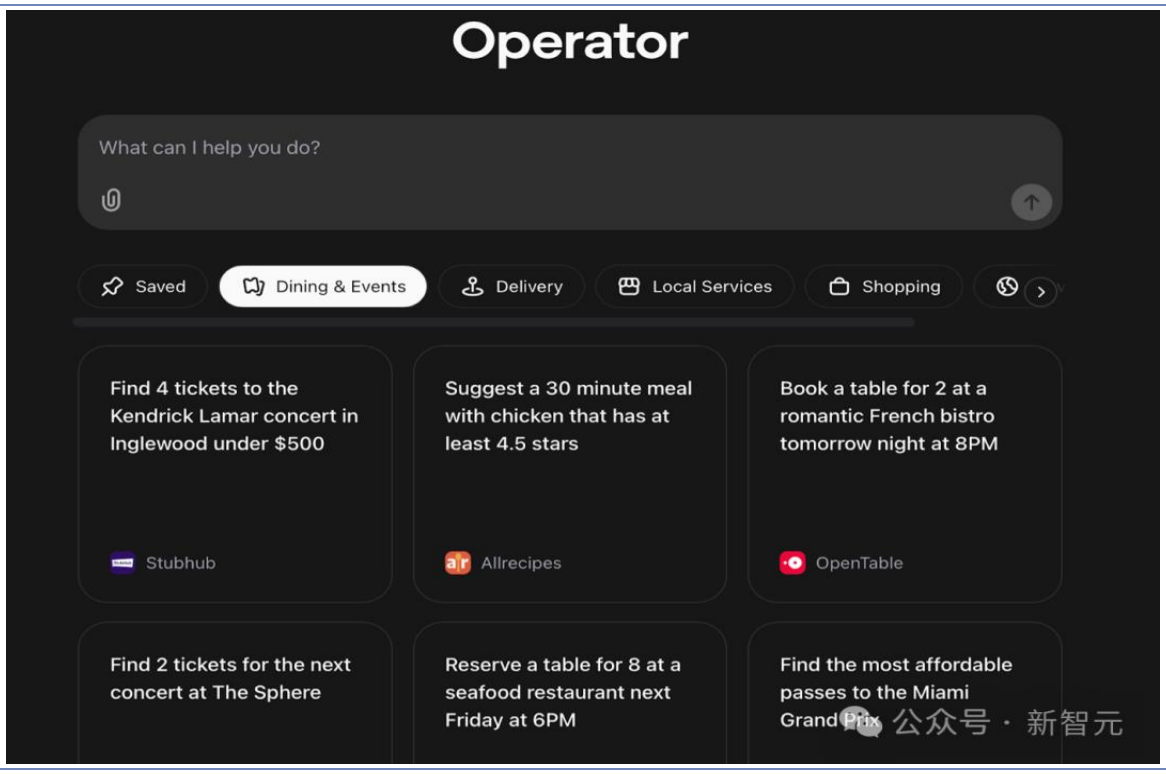
根据这个技术路线图，AI 实现 L3 之后，通过不断优化工具使用能力，正为 L4 阶段——自主学习发明创新奠定了扎实的技术基础。这也正是智谱下一步，所要开拓的智能。

5. OpenAI 发布首个智能体 Operator

1 月 24 日发布的 Operator 无疑是 AI 圈最大的亮点，这款 AI 智能体能够自动处理一系列任务，像是演唱会购票、家政服务预订、AI 新闻查找等。

OpenAI 针对 Operator 新开了一个网页 operator.chatgpt.com，而不是像之前发布的功能都直接统一内置在 ChatGPT 中。Operator 的页面与 ChatGPT 大致相似，只是输入框的提示词从「我能帮您什么吗？」变为了「我能帮您做什么吗？」。这里展示了一些 Operator 在 OpenAI 的合作伙伴网站上能执行的推荐任务。比如，不用半小时就能用鸡肉做好的晚餐食谱。

图29：Operator 在 OpenAI 合作伙伴网站上能执行的推荐任务



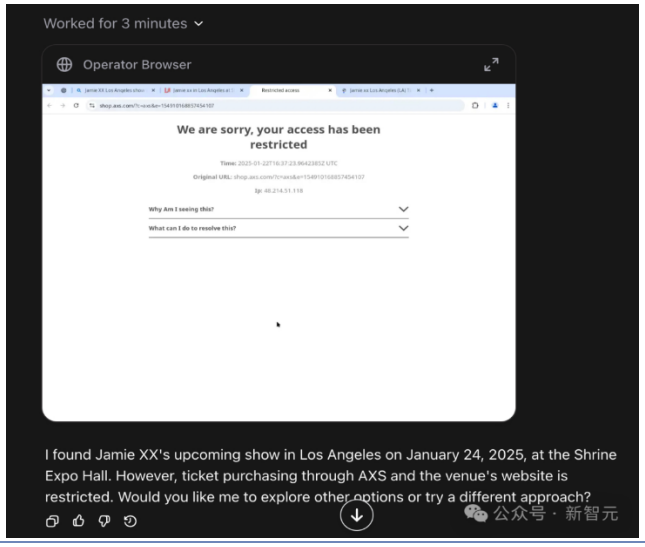
资料来源：新智元，中国银河证券研究院

目前 Operator 支持自动完成通常需要 15-20 分钟内的网络任务。值得一提的是，Operator 还拥有能够极大提高用户体验的「保存和共享功能」。也就是说，一旦完成任务，Operator 就可以轻松保存工作流程。比如持续用最新的销售数据来更新相应的报表。它甚至提供了一个流畅的会话记录视频，支持用户观看并与其他人分享。

Operator 自身的缺陷亦是源于它的优势本身。它的独特之处在于不用使用用户本地的浏览器执行操作，而是 OpenAI 数据中心之一的一个浏览器，用户可以远程观看并与之互动。这种设计的优点是你可以在任何地方、任何时候使用它——例如，在任意移动设备上。但缺点是**许多像 Reddit 这样的网站已经阻止 AI 智能体浏览，因此它们无法被 Operator 访问。并且 Operator 也因性能或法律原因被 OpenAI 阻止访问某些资源密集型网站，如 Figma 或 YouTube。**

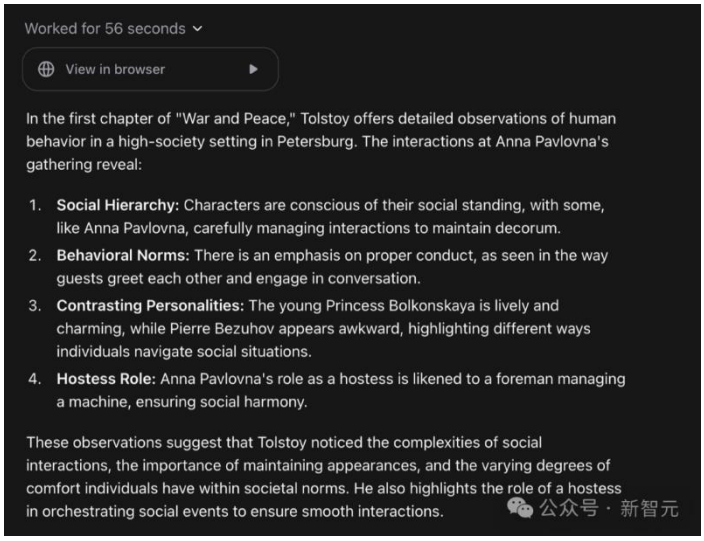
目前的 Operator 实际上更像是一个能够完成你给外包出去业务的乙方，而不是一个足够聪明的个人研究助理。例如，可以让 Operator 阅读《战争与和平》的第一章，并总结每个角色的所有细节以及他们所展现的人类心理和行为。然后，Operator 在 Project Gutenberg 网站上找到了《战争与和平》并阅读了第一章，做得非常出色。但是其摘要却枯燥乏味且粗糙宽泛。如果提供相同的信息，OpenAI 的 o1 在这项任务上会做得更好——但是 o1 还没有自主执行任务的能力。也就是说，OpenAI 专注于让 Operator 非常适合自动执行重复性工作流程，而不太关注其智能水平。

图30：Operator 访问 ticketmaster 被拒



资料来源：新智元，中国银河证券研究院

图31：Operator 阅读《战争与和平》第一章形成的摘要



资料来源：新智元，中国银河证券研究院

6. 百川首个全场景深度思考模型 Baichuan-M1-preview 发布

1 月 24 日百川的首个全场景深度思考模型 Baichuan-M1-preview 发布了，相比其他推理模型，它能力全面，同时具备语言推理、视觉推理、搜索推理三个维度的全面推理能力，且均做到了行业领先。而且，还解锁了「医疗循证模式」，复杂医疗问题的推理能力大幅提升。现在，M1 已经在百小应上线了。

不仅如此，为了推动 AI 技术在医疗领域的创新发展，繁荣 AI 医疗生态，百川还开源了 Baichuan-M1-14B。这个 M1 的小尺寸版模型，医疗推理能力已经超越了更大参数量的 Qwen2.5-72B，与 o1-mini 相差无几。

图32：M1 具备语言、视觉、搜索三个维度的全面推理能力

模型名称	语言推理	视觉推理	搜索推理
Baichuan-M1-preview	Y	Y	Y
o1-preview	Y	Y	N
Gemini-2.0-flash-thinking	Y	Y	Y
QwQ-Preview	Y	N	N
QVQ-72B-Preview	N	Y	N
DeepSeek-R1	Y	Y	N
GLM-Zero-Preview	Y	Y	N
Step R-mini	Y	N	N
K1.5	Y	Y	N

资料来源：新智元，中国银河证券研究院

（二）投融资事件

表12：1 月 AI 相关投融资事件

融资方	赛道	公司简介	融资日期	融资轮次	融资金额	本轮投资方
中科原动力	AI 农业	一家全球领先的农业机器人创新科技公司，致力于用人工智能和自动驾驶技术为全球农业发展提供具备全昼夜、无人化、精准作业能力的农田作业机器人产品和服务	2025-01-07	B1 轮	近 1 亿人民币	厦门先进一号制造业基金领投，老股东祥峰投资跟投
智平方	信息技术服务	成立于 2023 年 4 月，是一家专注于通用智能具身终端研发的科技创新企业。公司由国家级创新领军人才郭彦东博士创立，核心团队成员来自微软、小鹏汽车、OPPO 等国际领先企业和知名高校	2025-01-07	Pre-A 轮	超 1 亿人民币	达晨财智与敦鸿资产联合领投，基石资本跟投
西湖机器人	人形机器人	专注于研发下一代高度智能化的足式机器人，包括四足机器人和双足机器人，通过现实虚拟化和基于深度强化学习的智能行为决策技术，赋予机器人自主学习和自我成长能力	2025-01-09	天使+轮	近 1 亿人民币	天使湾创投，犇驰投资，金能基金，诚信创投
一目科技	AI 大模型	成立于 2015 年，起源于美国硅谷，是一家专注于多模态感知与 AI 计算解决方案的技术驱动型平台企业。公司致力于通过创新的物联网传感技术，将物理世界信号转化为数字信号，推动家电、水务、机器人、生命科学等多个产业的智慧升级	2025-01-13	D 轮	数亿人民币	赛富投资基金领投，南京市创新投资集团、A 股上市公司松霖科技跟投，庚辛资本担任独家顾问
云轴科技 ZStack	企业云服务	成立于 2015 年，是一家专注于产品化的国产自主创新开源云计算服务商，提供自研的 ZStack 私有云、ZStack 混	2025-01-13	D 轮	数亿人民币	北京信息产业发展投资基金

		合云、ZStack CMP 多云管理平台、ZStack Cube 超融合一体机、ZStack AIOS 平台“智塔”等产品				
硕橙科技	AI 工业	成立于 2016 年，核心业务是通过智能硬件收集机器设备的噪声、振动、温度、电流、拉压力等多维数据，结合机器学习和 AI 算法，实现设备的预测性维护、智能运维、自动化质检、环境异常报警等服务，产品体系包括橙盒、多维数据采集站、智能声纹传感器等硬件，以及设备智能运维系统、EAM 设备资产管理系统、星橙云数智化云平台等软件	2025-01-14	C2 轮	超 1 亿人民币	彬复资本、厦门创投、钟楼金控和浪潮产投等
思必驰	AI 大模型	成立于 2007 年，是一家专注于对话式人工智能平台的高科技企业，核心业务涵盖智能语音技术的研发与应用，包括语音识别、语音合成、语义对话、语音唤醒等。公司自主研发了新一代人机交互平台（DUI）和人工智能芯片（TH1520），并为车联网、IoT、政务、金融等多个行业提供自然语言交互解决方案	2025-01-14	战略融资	5 亿人民币	知名产业基金、国资平台、私募基金
国中数字	数据分析服务	一家专注于数字科技领域的研究和开发的企业。公司通过与国内外一流科研机构的合作,致力于为全球客户提供创新的数字化解决方案,涵盖人工智能、大数据分析、物联网、区块链等多个领域。旗下品牌有鱼生活 APP 是一款承载文化大数据新消费的超级物种,为用户提供丰富的文化体验和消费选择	2025-01-22	B 轮	数亿人民币	深圳市东方华远投资（集团）有限公司领投，多家知名投资机构跟投
维他动力	机器人	24 年底在北京正式成立，致力于打造具有开创性的机器人产品。创始人兼 CEO 余轶南是地平线前副总裁、软件平台产品线总裁，另两位联合创始人分别是地平线前软件平台总架构师、智驾创始团队成员宋巍，和理想汽车前智能驾驶产品总监赵哲伦	2025-01-22	种子轮	近 1 亿人民币	地平线和高瓴创投领投
它石智航	信息技术服务	一家专注于具身智能和智能机器人技术的创新型企业，成立于 2024 年 7 月，由前华为智能汽车业务部（车 BU）自动驾驶系统首席科学家陈亦伦博士创立。业务涵盖人工智能基础软件开发、人工智能硬件销售、智能机器人研发与销售等多个领域	2025-01-23	种子轮	1.5 亿美元	
中科时代	AI 芯片	一家专注于工业智能计算机（工智机）及相关自动化技术的高科技企业，提供以“工智机”为牵引，Automation 为核心，IO/Motion/Acceleration/Digitization 为配套的产品组合。公司由中国科学院计算技术研究所孵化，创始团队成员均为中科院计算所的核心专家，拥有近 20 年的工业智能计算控制技术研发经验	2025-01-24	B1 轮	2 亿人民币	湖北高质量发展产业投资基金领投，国新国证、老股东国中资本、博将资本、卓源亚洲跟投，高鹄资本担任长期独家财务顾问

资料来源：投中网，投资界，钛媒体，36 氪，Wind 万得，搜狐，亿欧，中科时代官网，中国银河证券研究院

五、投资建议

关注以下细分赛道及公司：1、国产算力产业链及生态伙伴：如工业富联、中科曙光、曙光数创、海光信息、龙芯中科等。2、算力基础设施产业链：如润泽科技、宝信软件等。3、AI+应用：如科大讯飞、金蝶国际、金山办公、同花顺、嘉和美康、国能日新、彩讯股份、恒生电子、万兴科技等。4、端侧 AI：如虹软科技、海康威视、中科创达、华勤技术、萤石网络等。5、数据要素产业链中供给、流通、应用公司：如拓尔思、达梦数据、深桑达 A、上海钢联等。

六、风险提示

技术迭代不及预期风险；科技巨头竞争加剧风险；法律监管风险；供应链风险；下游需求不及预期风险。

图表目录

分析师承诺及简介

本人承诺以勤勉的执业态度，独立、客观地出具本报告，本报告清晰准确地反映本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与、未来也将不会与本报告的具体推荐或观点直接或间接相关。

吴砚靖 TMT/科创板研究负责人，北京大学软件项目管理硕士，10 年证券分析从业经验，历任中银国际证券首席分析师，国内大型知名 PE 机构研究部执行总经理。具备一二级市场经验，长期专注科技公司研究。

鲁佩机械行业首席分析师，伦敦政治经济学院经济学硕士，证券从业 9 年，2021 年加入中国银河证券研究院。曾获新财富最佳分析师、IAMAC 最受欢迎卖方分析师、万得金牌分析师、中证报最佳分析师、Choice 最佳分析师、金翼奖等。

免责声明

评级标准

评级标准		评级	说明
评级标准为报告发布日后的 6 到 12 个月行业指数（或公司股价）相对市场表现，其中：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，北交所市场以北证 50 指数为基准，香港市场以恒生指数为基准。	行业评级	推荐：	相对基准指数涨幅 10%以上
		中性：	相对基准指数涨幅在-5%～10%之间
		回避：	相对基准指数跌幅 5%以上
	公司评级	推荐：	相对基准指数涨幅 20%以上
		谨慎推荐：	相对基准指数涨幅在 5%～20%之间
		中性：	相对基准指数涨幅在-5%～5%之间
		回避：	相对基准指数跌幅 5%以上

联系

中国银河证券股份有限公司研究院

深圳市福田区金田路 3088 号中洲大厦 20 层

上海浦东新区富城路 99 号震旦大厦 31 层

北京市丰台区西营街 8 号院 1 号楼青海金融大厦

公司网址：www.chinastock.com.cn

机构请致电：

深广地区：程　曦　0755-83471683　chengxi_yj@chinastock.com.cn

苏一耘　0755-83479312　suyiyun_yj@chinastock.com.cn

上海地区：陆韵如　021-60387901　luyunru_yj@chinastock.com.cn

李洋洋　021-20252671　liyangyang_yj@chinastock.com.cn

北京地区：田　薇　010-80927721　tianwei@chinastock.com.cn

褚　颖　010-80927755　chuying_yj@chinastock.com.cn