2026 年 05 月 26 日

华鑫证券 CHINA FORTUNE SECURITIES

Google 推出 Gemini 3.5 Flash，GLM-5.1 高速版发布

—计算机行业周报

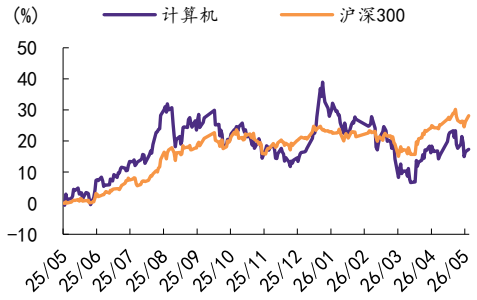
推荐(维持)

分析师：任春阳 S1050521110006
rency@cfsc.com.cn

行业相对表现

表现	1M	3M	12M
计算机(申万)	0.5	-6.5	16.4
沪深 300	3.2	4.1	27.5

市场表现



资料来源：Wind，华鑫证券研究

相关研究

1、《计算机行业点评报告：AMZN.O：2026Q1 营收利润双增，AWS 与电商业务驱动增长》2026-05-26

2、《计算机行业点评报告：苹果（AAPL.O）：Q2 营收利润双增长，iPhone17 热销与研发投入增长》2026-05-25

3、《计算机行业点评报告：谷歌（GOOGL）：2026Q1 营收持续增长，Cloud 增速创历史新高》2026-05-25

投资要点

算力：算力租赁价格平稳，Google 推出 Gemini 3.5 系列首款模型 Gemini 3.5 Flash

2026 年 5 月 20 日，Google 在 I/O 2026 大会上正式发布 Gemini 3.5 Flash。作为 3.5 系列首款轻量级模型，该模型在保持高智能水平的同时，成本仅为同类顶尖模型的一半，有时甚至不到三分之一。此外，该模型在四大基准测试中表现均优于前代 Gemini 3.1 Pro，部分指标领先 GPT-5.5 和 Claude Opus 4.7，输出速度高达 289 tokens/秒，约为其他前沿模型的 4 倍。官方表示，该模型将作为全球 Gemini 应用和搜索 AI 模式的默认模型推广使用。

AI 应用：Gemini 周访问量环比+1.40%，GLM-5.1 高速版发布

2026 年 5 月 22 日，智谱公司面向部分企业客户推出了 GLM-5.1 高速版 API“GLM-5.1-highspeed”，其模型输出速度达到了每秒 400 个 token。从实际应用的角度来看，这意味着原本需要一位写作者连续伏案数天才能完成的文字量，该模型可以在 1 分钟内交付完毕；而一名工程师埋头敲击键盘三天才能完成的开发任务，模型也能在喝一杯咖啡的短暂时间内处理完成。

AI 融资动向：Hark 完成超 7 亿美元 A 轮融资，投后估值达 60 亿美元

2026 年 5 月 22 日，AI 硬件公司 Hark 完成超 7 亿美元 A 轮融资，投后估值达 60 亿美元。本次融资由 Parkway Venture Capital 领投，NVIDIA、AMD Ventures、Intel Capital 等 11 家公司参投。该公司致力于开发具备持久记忆的高度个性化智能系统，旨在通过软硬件协同构建人机通用交互界面。目前，其多模态 AI 模型已支持餐厅订座、电商下单和信息检索等功能，计划于今年夏季推出首批 AI 产品。

投资建议

2026 年 5 月 21 日，英伟达公布其最新财报。公司整体营收达 816 亿美元，同比增长 85%、环比增长 20%；盈利层面，GAAP 标准下净利润达 583 亿美元，稀释后每股收益为 2.39 美元，两项数据均较上年均同期增长超三倍；本季度公司现金流表现强劲，经营现金流升至 503 亿美元，自由现金流达到 486 亿

美元，整体盈利与现金流能力持续突出。业务层面，核心数据中心业务营收 752 亿美元，同比、环比分别增长 92%、21%，其中数据中心计算、网络收入均创下历史新高，受益于 Blackwell300 及各类算力、互联解决方案的旺盛需求；边缘计算业务营收 64 亿美元，同比、环比均稳步增长，依托 Blackwell 工作站市场需求实现扩容。展望 2026 年第二季度，公司预计营收中值 910 亿美元，继续高于市场预期。此外，公司已启用全新业务报告框架，将整体业务整合划分为数据中心、边缘计算两大核心平台。其中，数据中心业务进一步细分出超大规模、ACIE 两大子市场：超大规模市场涵盖公共云与头部消费互联网企业相关收入，ACIE 市场则覆盖 AI 云、工业及企业级应用领域，重点挖掘各行业 AI 数据中心、AI 工厂的增长机遇。边缘计算业务聚焦智能体与物理 AI 数据处理设备，业务场景全面覆盖 PC、游戏机、工作站、AI-RAN 基站、机器人、汽车等各类终端领域。

本次英伟达最新财报调整业务统计口径，ACIE 企业级市场确立为算力核心增长主线。黄仁勋判断，工业及企业端覆盖经济规模可达 50 万亿-80 万亿美元，长期 ACIE 业务增速将优于超大规模客户业务。叠加板块 31%环比高增、Blackwell 系列产品适配优势，企业级 AI 工厂已成为 Token 经济体系下行业主流发展路径。单机算力提升可优化模型推理效率，有效增厚 AI 服务收益。ACIE 市场覆盖制造、金融、医疗、能源等实体企业，客户结构分散，与集中化互联网大客户形成明显区分，亦是赛道高增长的核心支撑。随着企业数字化转型推进、行业专属算力集群建设需求上行，及 ACIE 算力配套落地节奏加快，下游硬件采购需求具备较强超预期空间。

中长期，建议关注专注于半导体等高端制造业的罗博特科（300757.SZ）、新能源业务高增并供货科尔摩根等全球电机巨头的唯科科技（301196.SZ）、AI 智能文字识别与商业大数据领域巨头的合合信息（688615.SH）、深耕工业 AI 与软件并长期服务高端装备等领域头部客户的能科科技（603859.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。

重点关注公司及盈利预测

公司代码	名称	2026-05-26	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	621.99	-0.30	0.30	0.60	-2073.30	2073.30	1036.65	买入
301196.SZ	唯科科技	147.93	2.53	3.34	3.98	58.47	44.29	37.17	买入

603859.SH	能科科技	47.25	0.92	1.21	1.50	51.36	39.05	31.50	买入
688615.SH	合合信息	139.01	3.24	4.22	5.25	42.90	32.94	26.48	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，Google 推出 Gemini 3.5 系列首款模型 Gemini 3.5 Flash

1.1、Tokens 跟踪

根据 OpenRouter 公开数据，2026 年 5 月 18 日至 5 月 24 日，周度 Token 消耗量有所上升，调用量为 28.9T，环比上周增加 7.43%。在 Tokens 规模 Leaderboard 前五名中，DeepSeek 的 DeepSeek V4 Flash 以 3.43T tokens 位居榜首，Tencent 的 Hy3 preview 以 3.07T tokens 位居第二，Anthropic 的 Claude Opus 4.7 以 1.94T tokens 位居第三；Anthropic 的 Claude Sonnet 4.6 以 1.89T tokens 位列第四；OpenRouter 旗下的 Owl Alpha 以 1.15T tokens 位居第五；

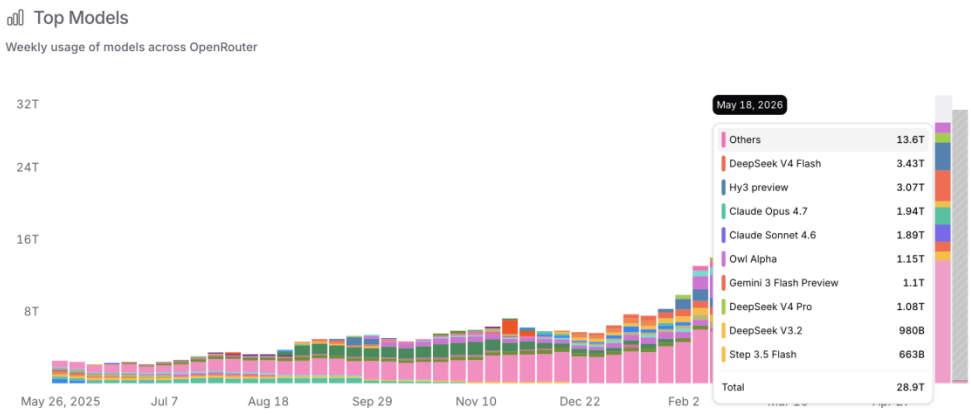
从市场份额维度来看，DeepSeek 以 5.74T tokens 占据 19.8%的份额，稳居首位；Anthropic 以 4.77T tokens 占据 16.5%，位列第二；Google、Tencent、OpenAI 则分别以 4.37T、3.07T、2.97T tokens，对应占据 15.1%、10.6%、10.3%的市场份额。

近期，中国移动、中国联通、中国电信三大运营商陆续在区域范围内推出 Token 套餐，面向个人及政企用户，按用量计价。专家评价称，此举将助力算力普惠，推动 AI 算力成为公共服务，从而进一步提升中国的 AI 渗透率。

5 月 22 日，智谱推出 GLM-5.1 高速版 API GLM-5.1-highspeed。数据显示，该模型输出速度高达 400 tokens/s，打破全球大模型厂商历史记录，并在一众国产大模型中首度实现旗舰级能力与极致低延迟的结合，意味着用户无需再为响应速度牺牲模型质量。当前，该模型已面向智谱 MaaS 平台部分企业客户开放。

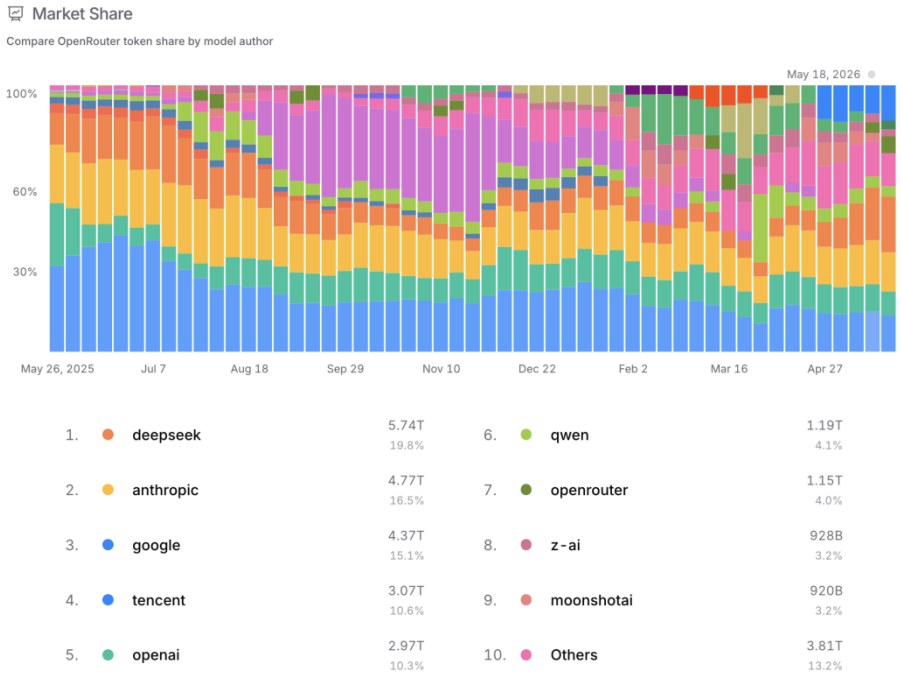
DeepSeek-V4-Pro 模型 API 官宣永久降价。最新公告显示，原定自 6 月起恢复原价的 DeepSeek-V4-Pro 模型，将于 5 月 31 日 2.5 折优惠活动结束后，永久调整为原价的 1/4，即：每百万 tokens 输入（缓存命中）0.025 元，输入（缓存未命中）3 元，输出 6 元。

图表 1：Tokens 规模 Leaderboard



资料来源：OpenRouter，华鑫证券研究

图表 2：市场份额占据示意



资料来源：OpenRouter，华鑫证券研究

1.2、数据跟踪：阿里云发布千问云官网，同步推出新一代千问旗舰模型 Qwen3.7-Max

2026 年 5 月 20 日，阿里云召开年度峰会，同日发布千问云、Qwen3.7-Max 两大重磅产品，宣布面向 Agentic 时代全面升级。

作为面向 Agent 时代而生的全新 AI 产品官网，千问云聚合了 Qwen、GLM、Kimi、DeepSeek、Wan、HappyHorse 等 150 余款主流模型 API，实现了从原子能力到交互逻辑的全面重构，支持模型参数、能力、价格、上下文长度等多维度对比，用户可以根据真实体验进行评估，挑选与业务匹配的模型，实现快速选型。

千问云的另外一个特点在于其对模型服务链路的全面 Skill 和 CLI 化。通过将模型服务的核心能力封装为 Skills 和 CLI 工具，Agent 可动态路由不同模型，自动完成图片生成、图片处理、视频创作等任务，可直接通过脚本或命令行自动化完成所有模型服务的工作流，还可实时拉取模型用量数据，自动分析趋势、识别异常，为用户提供成本优化建议。

此外，阿里云同日发布新一代千问旗舰模型 Qwen3.7-Max。该模型在三方机构 Arena 全球大模型盲测总榜中已超过 Kimi-K2.6、DeepSeek-v4-pro、GLM-5.1，位列国产模型第一，并与 GPT、Claude、Gemini 三大最强模型接近。

数据显示，该模型在编程智能体、通用智能体、推理能力、通用能力与多语言等方面均有不同程度的突破。与此同时，该模型已具备自主完成 35 小时超长程智能体复杂任务的能力，实验结果显示，Qwen3.7-Max 在全新芯片平台上通过自主编程和超1000次工具调用，推理速度较原版本提升 10 倍。

近三个月内，千问旗舰大模型已实现3.5、3.6、3.7三个版本的稳定迭代，国产模型的

性能上限持续被刷新。据官方消息，千问 3.7 系列还将推出 Qwen3.7-Plus 等不同版本模型，实现从编程智能体到视觉智能体的全覆盖，为下一代 AI 提供全能智能体新基座。

1.3、产业动态：Google 推出 Gemini 3.5 系列首款模型，AI 布局全面提速

北京时间 5 月 20 日凌晨，Google 最新、最强旗舰模型 Gemini 3.5 Flash 于 I/O 2026 大会公开亮相。该模型为 Gemini 3.5 系列首款轻量级产品，官方称其在保持高智能水平的同时，成本仅为同类顶尖模型的一半，甚至有时不到三分之一。

数据显示，该模型在编码生成测试（Terminal-Bench 2.1）中得分 76.2%；在真实世界 Agent 任务测试（GDPval-AA）中评分 1656 Elo；在大规模工具调用测试（MCP Atlas）中，得分 83.6%；在多模态理解测试（CharXiv Reasoning）中，得分 84.2%，四大基准测试结果均优于 Gemini 3.1 Pro，并在部分基准测试中，表现领先于 GPT-5.5 和 Claude Opus 4.7。

图表 3：Gemini 3.5 Flash 基准测试表现对比

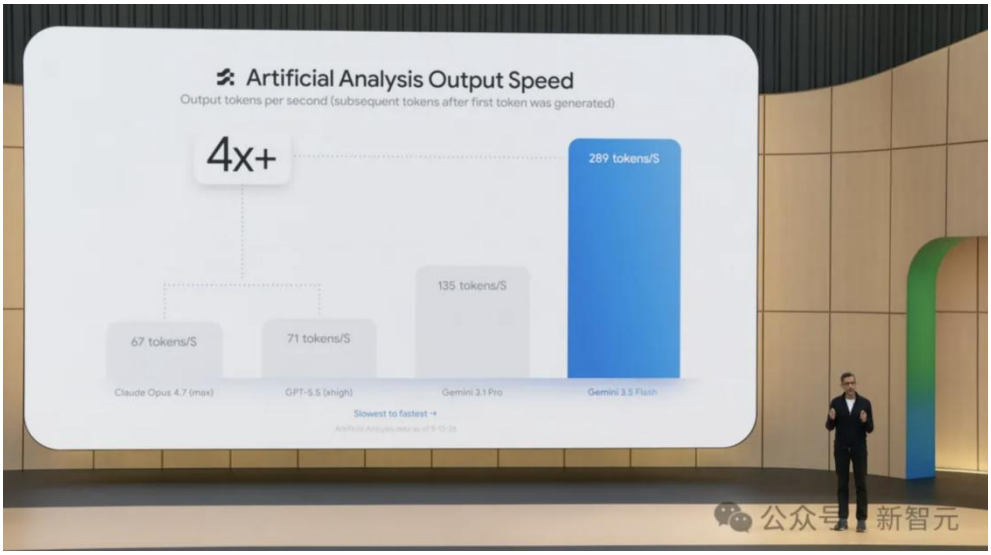
	Benchmark		Gemini 3.5 Flash	Gemini 3 Flash	Gemini 3.1 Pro	Claude Sonnet 4.6	Claude Opus 4.7	GPT-5.5
Coding	Terminal-bench 2.1 Agentic terminal coding	Terminus-2 harness	76.2%	58.0%	70.3%	-	66.1%	**78.2%**
	SWE-Bench Pro (Public) Diverse agentic coding tasks	Single attempt	55.1%	49.6%	54.2%	-	**64.3%**	58.6%
Agentic	MCP Atlas Multi-step workflows using MCP		**83.6%**	62.0%	78.2%	69.5%	79.1%	75.3%
	Toolathlon Real-world general tool use		**56.5%**	49.4%	-	-	-	55.6%
UI control	OSWorld-Verified Agentic computer use		78.4%	65.1%	76.2%	72.5%	78.0%	**78.7%**
Expert tasks	Finance Agent v2 Financial analysis and decision-making		**57.9%**	42.6%	43.0%	51.0%	51.5%	51.8%
	GDPval-AA Economically valuable knowledge work	Elo	1656	1204	1314	1676	1753	**1769**
Multimodal	CharXiv Reasoning Information synthesis from complex charts	No tools	**84.2%**	80.3%	83.3%	72.4%	82.1%	84.1%
	MMMU-Pro Multimodal understanding and reasoning	No tools	**83.6%**	81.2%	80.5%	74.5%	75.2%	81.2%
	Blueprint-Bench 2 Agentic spatial reasoning	Normalized score	33.6%	0.0%	26.5%	6.7%	24.5%	**36.2%**
Long context	MRCR v2 (8-needle) Long context performance	128k (average)	77.3%	67.2%	84.9%	84.9%	59.3%	**94.8%**
		1M (pointwise)	**26.6%**	22.1%	26.3%	-	-	-
Reasoning	Humanity's Last Exam Academic reasoning (full set, text + MM)		40.2%	33.7%	44.4%	33.2%	**46.9%**	41.4%
	ARC-AGI-2 Abstract reasoning puzzles		72.1%	33.6%	77.1%	58.3%	75.8%	**84.6%**

资料来源：智东西，华鑫证券研究

速度方面，3.5 Flash 以每秒输出 289 tokens 的速度遥遥领先，高达其他前沿模型输出速度的 4 倍以上。具体功能上，该模型支持快速规划、构建和迭代，具备帮助用户完成开发新应用、维护代码库、协助准备财务文件等任务的能力。与此同时，Google 在该模型的网络安全防护方面同样做了强化，降低了其生成有害内容的可能性。

此外，官方表示，该模型将作为全球 Gemini 应用和搜索 AI 模式的默认模型推广使用。当前重量级版本 Gemini 3.5 Pro 已在内部测试使用，预计下月向更广泛的用户进行开放。

图表 4：Gemini 3.5 Flash 输出速度对比示意图



资料来源：新智元，华鑫证券研究

同时，Google 正在使用 Gemini 3.5 Flash 配合开发平台 Antigravity 的新模式——升级后的 Antigravity 推出了独立桌面应用 Antigravity 2.0，作为智能体交互的中心平台，支持开发者协调多个智能体并行执行任务，并具备动态子智能体支持并行工作流程、后台自动化定时任务，及跨 Google AI Studio、Android 和 Firebase 的生态系统集成。

在一场演示中，Antigravity 展现出了其出色的工程能力：在 12 小时内调度 93 个智能体并行工作，处理超过 15000 次模型请求、26 亿 tokens，API credits 消耗不到 1000 美元，从内核到进程和内存管理系统，从零开始构建出一个可用的操作系统。

图表 5：Antigravity 工程能力数据展示图

资料来源：新智元，华鑫证券研究

基于 Gemini 3.5 系列提供的支持，Google 同步推出了全新的智能 AI 搜索框，Google Search 迎来了 25 年来的最大升级。升级后的 Google Search 不仅能依托 AI 智能提示梳理提问思路，还将具备多模态搜索能力，支持文字、图片、文件、视频乃至浏览器标签页作

为检索输入，并拥有创建和管理多个 AI 智能体完成任务的能力。

与此同时，Google 还推出了新型通用 AI 智能体 Gemini Spark，该智能体基于 Gemini 3.5 Flash，采用 Antigravity 调度架构，拥有针对关联应用中信息进行跨平台推理的能力，支持后台持续运行，能够实现 7x24 小时在线，可用于管理用户的数字生活，并代表用户执行操作。

除了 Gemini 3.5 系列，Google 还推出了全新模型系列 Gemini Omni，其定位为可基于任意输入形式生成各类模态输出内容的全能模型。目前，首款模型 Gemini Omni Flash 率先支持视频输出，用户可以在 Gemini 应用、Google Flow 和 YouTube Shorts 上进行试用，未来该系列将持续扩展至支持图像、文本输出。

2、AI 应用动态：Gemini 周访问量环比+1.40%，GLM-5.1 高速版发布

2.1、周流量跟踪：Gemini 周访问量环比+1.40%

本期（2026.5.16-2026.5.22）AI 相关网站流量数据：访问量前三位分别为 ChatGPT（1267.0M）、Bing（805.6M）和 Gemini（671.6M），访问量环比增速第一为 Gemini（1.40%）；平均停留时长前三位分别为 Character.AI（00:14:38）、Discord（00:11:03）和 Kimi（00:08:23）;平均停留时长环比增速第一为文心一言（1.33%）。

图表 6：2026.5.16-2026.5.22AI 相关网站流量

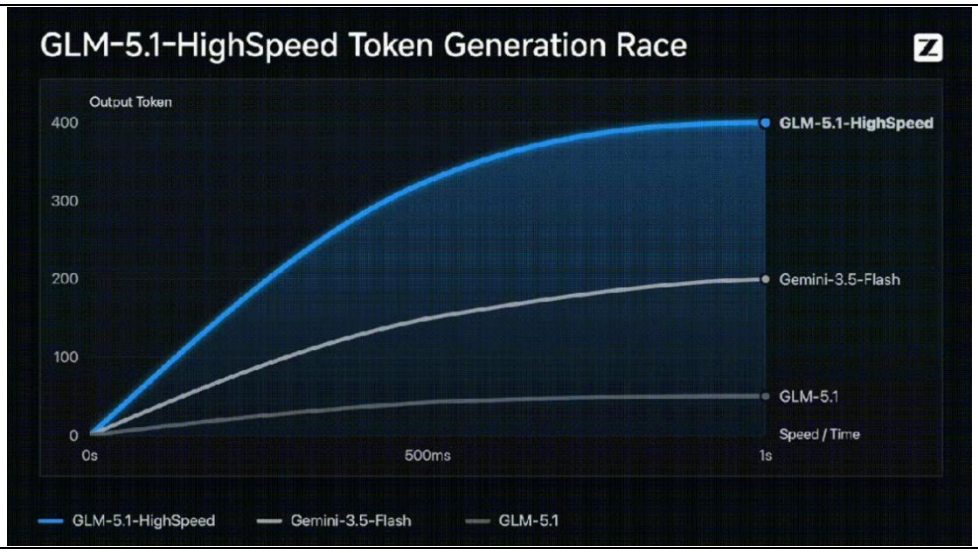
应用	应用类型	归属公司	周平均访问量（M）	访问量环比	平均停留时长	时长环比
ChatGPT	聊天机器人	OpenAI	1267.0	-3.50%	5:58	0.28%
Bing	搜索	微软	805.6	-1.00%	7:25	-0.67%
Gemini	聊天机器人	谷歌	671.6	1.40%	7:03	-0.94%
Canva	在线设计	Canva	227.6	0.31%	5:49	0.00%
Github	代码托管	微软	143.9	-1.03%	6:27	0.00%
Discord	游戏社区	微软	141.7	0.07%	11:03	0.30%
NotionAI	文本/笔记	Notion	38.49	-1.18%	7:56	0.00%
Character.AI	聊天机器人	Character.AI	34.7	-6.82%	14:38	-2.01%
Perplexity	AI 搜索	Perplexity	30.74	-5.04%	4:35	-1.43%
DeepI	翻译工具	DeepL	27.11	0.82%	2:24	0.00%
Kimi	聊天机器人	Moonshot AI	10.70	-2.99%	8:23	-0.59%
QuillBot	释义工具	QuillBot	9.97	-4.02%	2:52	0.58%
文心一言	聊天机器人	百度	0.59	-7.15%	2:32	1.33%

资料来源：similarweb,华鑫证券研究

2.2、产业动态：GLM-5.1 高速版发布，以 400 tokens/s 刷新全球大模型 API 速度上限

2026 年 5 月 22 日，智谱公司面向部分企业客户推出了 GLM-5.1 高速版 API“GLM-5.1-highspeed”，其模型输出速度达到了每秒 400 个 token。从实际应用的角度来看，这意味着原本需要一位写作者连续伏案数天才能完成的文字量，该模型可以在 1 分钟内交付完毕；而一名工程师埋头敲击键盘三天才能完成的开发任务，模型也能在喝一杯咖啡的短暂时间内处理完成。

图表 8：GLM-5.1-highspeed 模型输出速度达到 400 tokens/s



资料来源：智谱，华鑫证券研究

在过去的行业实践中，高速模型几乎总是轻量级模型，而 GLM-5.1 高速版的推出打破了这一惯例，它首次在国产大模型中将旗舰级的能力与极致的低延迟同时带入生产环境，使得用户无需再为了响应速度而妥协模型的质量。在与 GLM-5.1 普通版的对比中，高速版在长程任务和 Agent Swarm 等场景下表现尤为突出：例如在 30 秒内完成复杂网页的处理，或在极短时间内调度多达 50 个不同人格并行回答问题。

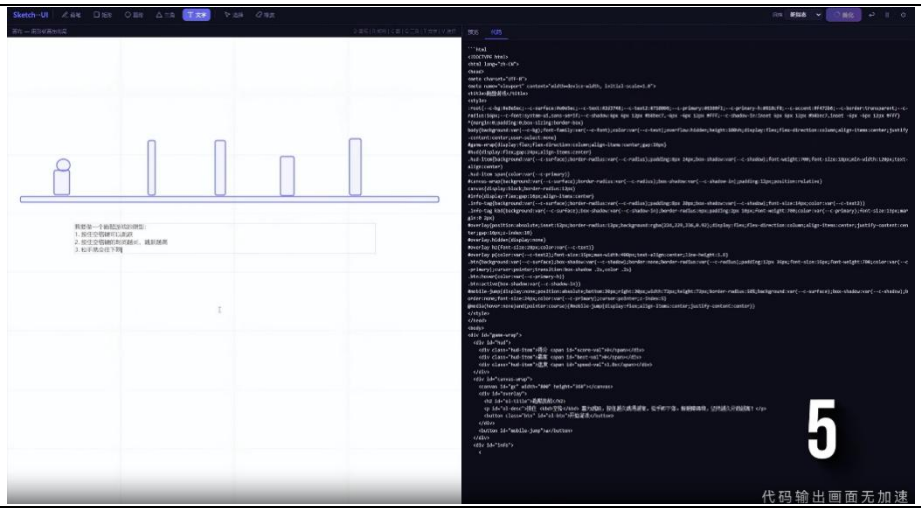
图表 9：在 Agent Swarm 中，GLM-5.1 高速版瞬间调度 50 个不同人格来并行回答

资料来源：智谱，华鑫证券研究

一个编程智能体任务往往需要经历数十轮模型调用，单轮响应只要慢上几秒，整体耗时就可能被拉长十几分钟。GLM-5.1 高速版在完整保留 GLM-5.1 原有能力的基础上，首次实现了真正意义上的“即问即答”式响应速度。这种速度带来的体验完全不同——模型开始成为一个可以实时协作的伙伴，共同与用户进行参数调整和代码修改。

在具体实测中，该模型在编写代码时如同开启了十倍速，能够一边理解工程上下文，一边持续生成代码与修改方案，用户刚刚输入需求，函数、接口与调用链便已同步展开。在另一项涉及 3D 地图角色移动的测试中，用户输入文字指令，模型能够根据输入瞬时完成场景建模并实时改变环境，此前因延迟而无法实现的全新产品形态开始具备落地的可能。更进一步，当用户提出需求的那一刻，模型可以即时生成恰好匹配该需求的工具与交互方式，甚至进行意图判断。

图表 10：GLM-5.1 高速版能够一边理解工程上下文，一边持续生成代码与修改方案



资料来源：智谱，华鑫证券研究

这一高速表现的背后，是智谱 GLM 团队与 TileRT 团队联合打造的 TileRT 高性能推理引擎。该引擎在推理引擎层、调度系统层和基础设施层三个层面进行了系统级优化。在推理引擎层，团队针对 GLM-5.1 的架构特点重写了核心推理路径，有效提升了单卡的吞吐能力；在调度系统层，通过动态批处理、请求合并和 KV 缓存调度优化，显著降低了高并发场景下的尾部延迟；在基础设施层，围绕推理集群部署、网络链路和负载均衡进行协同优化，确保每秒 400 个 token 的速度不是一个峰值数字，而是稳定可用的生产级能力。

当前主流推理框架以 operator/kernel 为基本调度单元，在单 token、小 batch、多卡 TP 场景下，各算子重复经历 host 启动、读写与同步的完整链路，导致微秒级算子内的调度和访存开销被迅速放大，使实际速度远低于硬件理论极限。TileRT 的设计思路则是彻底抛弃 Runtime 层的动态调度，在编译期将整个计算图静态编排为一个常驻 GPU 的 persistent Engine Kernel。在单卡内部，计算、异步 IO 与通信被全部拆解为微任务，整个推理过程只 Launch 一次 Engine Kernel，算子间的中间结果不再写回 Global Memory，而是通过 Register、Shared Memory 与 L2 Cache 直接传递，主机端调度与跨算子同步被压缩进同一个常驻 Kernel。在多卡尺度上，TileRT 进一步将 SM 内部的 Warp Specialization 思路外推到整张 8 卡 NVL 拓扑，不同 GPU 不再执行同构逻辑，而是按计算密度与数据依赖被特化为不同的工作单元。

GLM-5.1 高速版适用于人工智能编程、实时交互、商业决策、实时语音等对响应延迟要求极高的场景，目前已面向智谱 MaaS 平台的部分企业客户开放服务。智谱方面表示，将持续推进推理引擎的工程优化，进一步扩大高速模型的服务能力，使更多企业与开发者能够获得低延迟、高智能的生产级人工智能能力。

3、AI 融资动向：Hark 完成超 7 亿美元 A 轮融资，投后估值达 60 亿美元

2026 年 5 月 22 日，AI 硬件初创公司 Hark 宣布完成超 7 亿美元 A 轮融资，投后估值达 60 亿美元。本轮融资由 Parkway Venture Capital 领投，NVIDIA、Align Ventures、AMD Ventures、ARK Invest、Brookfield、Greycroft、Intel Capital 等 11 家公司集体参投。

Hark 致力于开发可长期记忆用户偏好与习惯的高度个性化的智能系统，旨在通过软硬件协同设计，将模型与硬件深度集成，构建人类与机器之间的通用交互界面。公司认为，当前主流 AI 产品大多依赖聊天界面和沿用已久的消费级设备运行，既缺乏对用户的持续记忆能力，也缺少专为智能交互而设计的硬件载体。为了改变这一现状，Hark 正在构建能够与人类及真实世界自然交互的下一代智能体系统。

路径上，Hark 并未选择聚焦于单一 AI 技术层，而是采取了垂直整合路线，同步构建基础模型、软件系统、原生硬件与全新交互界面，进而打造出一款端到端的无缝个人智能产品，该产品能够主动预判用户需求、降低认知负担，并以协作伙伴而非传统软件的方式运作。

从产品层面看，Hark 专注于开发一系列定制 AI 模型，将高阶个性化智能作为核心方向，核心模型采用多模态架构，内置持久记忆功能，用于持续记录用户偏好并主动生成任务建议。目前，其模型已具备餐厅订座、电商下单和信息检索等功能，并支持语音指令，但尚不明确其是否具备企业级任务自动化能力。

此外，Hark 计划将旗下 AI 模型与专属硬件设备捆绑推出，其专属硬件设备被定位为新一代交互入口，旨在替代传统的 AI 服务访问方式。

在团队方面，Hark 正大力招募硬件工程、AI 研发与设计等领域的顶尖人才，以支撑其团队从模型训练到硬件落地的全链路自研能力。随着 7 亿美元资金到位，Hark 将持续投入算力资源，在全新的 NVIDIA B200 集群上推进模型训练，并计划于今年夏季推出首批具备持久记忆与主动交互能力的 AI 模型。

图表 11：上周 AI 初创公司融资动态

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
Hark	AI 硬件	Parkway Venture Capital	A 轮	超 7 亿美元	超 7 亿美元	60 亿美元

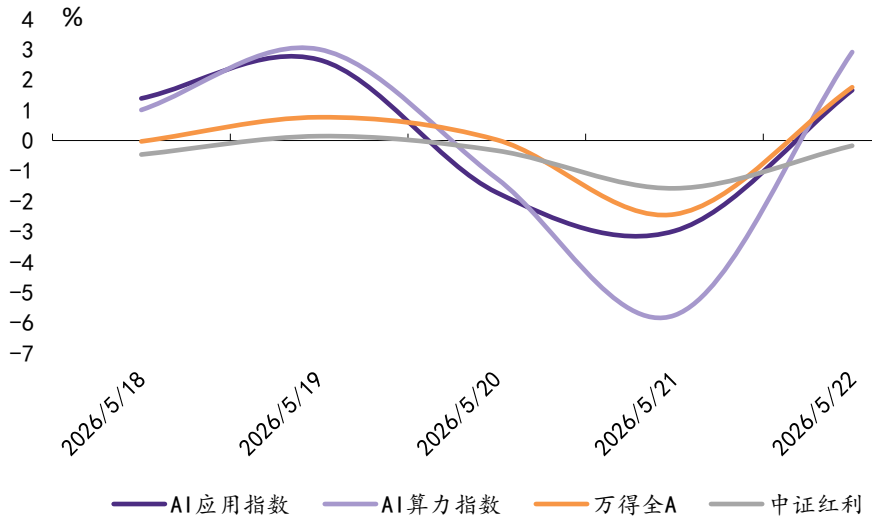
Exa Labs	AI 搜索引擎	a16z	C 轮	2.5 亿美元	约 4.5 亿美元	22 亿美元
Sigma Computing	AI 数据分析	Princeville Capital	E 轮	8000 万美元	超 3.72 亿美元	30 亿美元

资料来源：wind，Saasverse，华鑫证券研究

4、 行情复盘

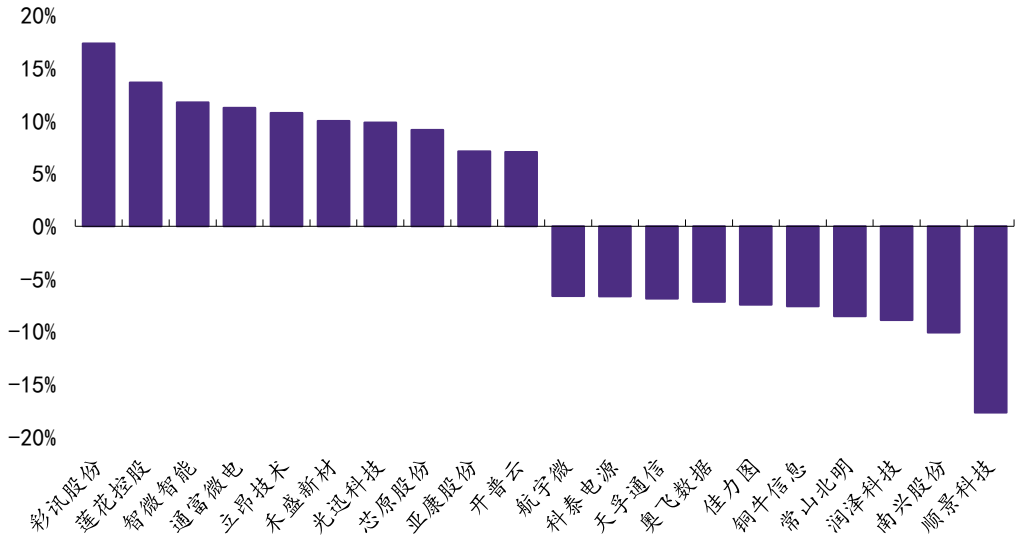
上周（2026.5.18-2026.5.22 日），AI 应用指数/AI 算力指数/万得全 A/中证红利日涨幅最大值分别为 2.66%/3.01%/1.76%/0.14%，AI 应用指数/AI 算力指数/万得全 A/中证红利日跌幅最大值分别为-2.98%/-5.76%/-2.43%/-1.58%。AI 算力指数内部，彩讯股份以 17.35%录得上周最大涨幅，顺景科技以-17.68%录得上周最大跌幅。AI 应用指数内部，美迪凯以 31.77%录得上周最大涨幅，鸿博股份以-15.41%录得上周最大跌幅。

图表 12：上周（2026.5.18-2026.5.22 日）指数日涨跌幅



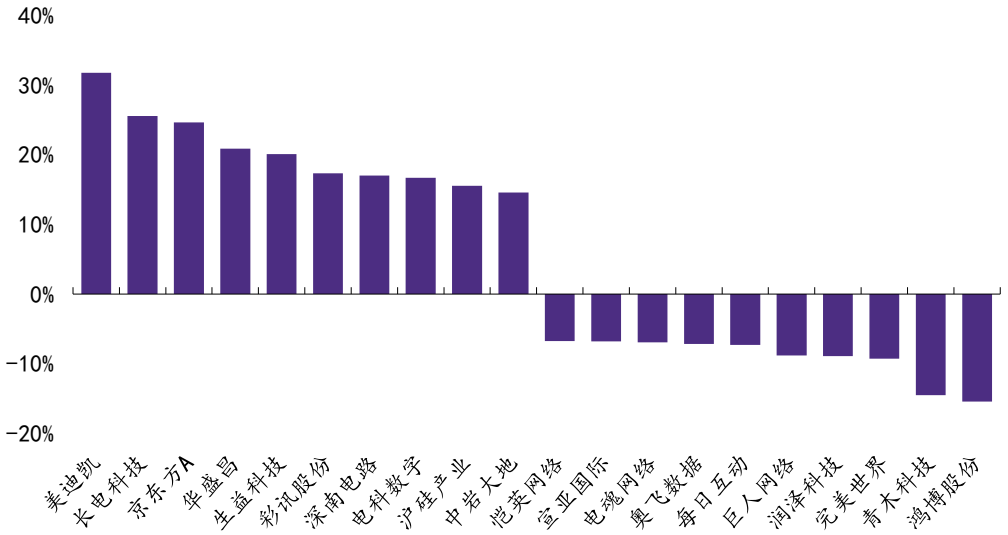
资料来源：wind，华鑫证券研究

图表 13：上周（2026.5.18-2026.5.22 日）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 14：上周（2026.5.18-2026.5.22 日）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

2026 年 5 月 21 日，英伟达公布其最新财报。公司整体营收达 816 亿美元，同比增长 85%、环比增长 20%；盈利层面，GAAP 标准下净利润达 583 亿美元，稀释后每股收益为 2.39 美元，两项数据均较上年均同期增长超三倍；本季度公司现金流表现强劲，经营现金流升至 503 亿美元，自由现金流达到 486 亿美元，整体盈利与现金流能力持续突出。业务层面，核心数据中心业务营收 752 亿美元，同比、环比分别增长 92%、21%，其中数据中心计算、网络收入均创下历史新高，受益于 Blackwell300 及各类算力、互联解决方案的旺盛需求；边缘计算业务营收 64 亿美元，同比、环比均稳步增长，依托 Blackwell 工作站市场需求实现扩容。展望 2026 年第二季度，公司预计营收中值 910 亿美元，继续高于市场预期。此外，公司已启用全新业务报告框架，将整体业务整合划分为数据中心、边缘计算两大核心平台。其中，数据中心业务进一步细分出超大规模、ACIE 两大子市场：超大规模市场涵盖公共云与头部消费互联网企业相关收入，ACIE 市场则覆盖 AI 云、工业及企业级应用领域，重点挖掘各行业 AI 数据中心、AI 工厂的增长机遇。边缘计算业务聚焦智能体与物理 AI 数据处理设备，业务场景全面覆盖 PC、游戏机、工作站、AI-RAN 基站、机器人、汽车等各类终端领域。

本次英伟达最新财报调整业务统计口径，ACIE 企业级市场确立为算力核心增长主线。黄仁勋判断，工业及企业端覆盖经济规模可达 50 万亿-80 万亿美元，长期 ACIE 业务增速将优于超大规模客户业务。叠加板块 31%环比高增、Blackwell 系列产品适配优势，企业级 AI 工厂已成为 Token 经济体系下行业主流发展路径。单机算力提升可优化模型推理效率，有效增厚 AI 服务收益。ACIE 市场覆盖制造、金融、医疗、能源等实体企业，客户结构分散，与集中化互联网大客户形成明显区分，亦是赛道高增长的核心支撑。随着企业数字化转型推进、行业专属算力集群建设需求上ACIE算力配套落地节奏加快，下游硬件采购需求具备较强超预期空间。

中长期，建议关注专注于半导体等高端制造业的罗博特科（300757.SZ）、新能源业务高增并供货科尔摩根等全球电机巨头的唯科科技（301196.SZ）、AI 智能文字识别与商业大数据领域巨头的合合信息（688615.SH）、深耕工业 AI 与软件并长期服务高端装备等领域头部客户的能科科技（603859.SH）。

图表 15：ficonTEC2025 年年中至今公告订单

签约日期	客户/描述	业务类型	金额	折合人民币
2025/6/20	美国某头部公司 A 及其子公司	光电子封测设备	约 1,710 万欧元	约 1.36 亿元
2025/7/11	美国某头部公司 B 及其子公司	光电子封测设备	约 1,418 万美元	约 0.98 亿元
2025/9/3	瑞士某头部公司 C 的子公司	全自动硅光子封装整线设备或服务	约 946.50 万欧元	约 0.75 亿元
2025/10/21	武汉驿路通科技股份有限公司	光纤预制及组装线相关自动化设备	约 900 万美元	约 0.62 亿元

2026/1/6	瑞士某头部公司 C 的子公司	第二条全自动 OCS（光交换机）封装整线设备及服务	约 770.00 万欧元	约 0.61 亿元
2025/9/24-2026/1/26	以色列的纳斯达克上市的头部公司 E	单面晶圆测试设备及服务	约 921.60 万美元	约 0.64 亿元
2026/3/13	暂未披露	双面晶圆测试设备及服务	约 608.09 万欧元	约 0.48 亿元
2026/3/19-2026/3/25	纳斯达克上市的公司 F 及其子公司	耦合设备及服务（可用于可插拔硅光高速光模块封装制程核心环节的量产）	约 6 亿元人民币	约 6 亿元
2026/4/1	纳斯达克上市的公司 F	耦合设备及服务（可用于可插拔硅光高速光模块封装制程核心环节的量产）	约 3,570 万美元	约 2.46 亿元
2026/4/8-2026/5/1	纽约证券交易所上市的公司 B 的子公司	耦合设备及相关服务	约 2680 万美元	约 1.83 亿元
2026/4/8-2026/5/1	纳斯达克上市的公司 F	视觉检测设备、高精度激光 bar 条封装设备及相关服务	约 3226 万美元	约 2.20 亿
	总金额			约 17.93 亿元

资料来源：Wind，公司公告，华鑫证券研究

图表 16：重点关注公司及盈利预测

公司代码	名称	2026-05-26	EPS			PE			投资评级
		股价	2025	2026E	2027E	2025	2026E	2027E	
300757.SZ	罗博特科	621.99	-0.30	0.30	0.60	-2073.30	2073.30	1036.65	买入
301196.SZ	唯科科技	147.93	2.53	3.34	3.98	58.47	44.29	37.17	买入
603859.SH	能科科技	47.25	0.92	1.21	1.50	51.36	39.05	31.50	买入
688615.SH	合合信息	139.01	3.24	4.22	5.25	42.90	32.94	26.48	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。

中小盘&北交所组介绍

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

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。