

计算机行业周报

端侧落地进行时，大模型进入“白菜价”时代

投资要点:

➤ 本周 AI 新闻速递:

国内 AI: (1) 面壁智能发布最强多模态开源模型 MiniCPM-Llama3-V 2.5。 (2) 黑马 Yi-Large 刷新榜单，成功登顶国产模型第一。 (3) 百川发布首个 AI 应用百小应。 (4) 国产大模型迎降价潮。

国外 AI: (1) Meta 发布 34B 的混合模态模型 Chameleon。 (2) 谷歌发布文生图模型 Imagen 3。 (3) 首个 GPU 高级语言发布，编写并代代码门槛显著降低。 (4) Falcon 2 回归，综合性能超越 Llama3。 (5) Hugging Face ZeroGPU 计划正式发布：一千万美元的免费共享 GPU。 (6) 微软发布 Copilot+PC，重新定义 AI 生产力。 (7) 英伟达 GH200 芯片落地 9 个超算中心，总算力规划达 200 exaflops。 (8) 视频版 PS 发布，0 样本视频编辑模型。

➤ 端侧盛宴即将开始，微软决战生产力时代

微软发布多模态小模型 **Phi-3 Silica**，端侧实现 **SOTA**。这是一个专为 Copilot+PC 设计的端侧小模型 (SLM)，能够充分适配 AIPC 的 NPU，以满足模型计算的算力要求。Phi-3 Silica 仅有 3B 参数，是 Phi-3 系列中最小的。其他发布的 Phi-3 系列模型包括：Phi-3 vision、Phi-3 mini、Phi-3 small、Phi-3 medium 分别有 42 亿、38 亿、70 亿、140 亿参数。分别在视觉、长文本、综合性能等领域表现出色。

推出 **Copilot Studio**，简化 **Agent 开发流程**。通过 Copilot Studio 构建 Agent 有以下优势：对长期稳定运行的业务流程完成自动化流程；对操作和用户输入进行推理；利用记忆功能带入上下文；根据用户反馈进行自主学习；记录异常请求并寻求帮助。

GitHub Copilot 作为微软推出的扩展功能之一，可以利用自然语言交互来提高开发速度。所构建的 GitHub Copilot Workspace 是一个 Copilot 原生的开发环境，它允许开发者从构思到代码再到软件的整个过程中，都能使用自然语言来进行交流和操作。

➤ 国产大模型迎降价潮，商业落地按下加速键

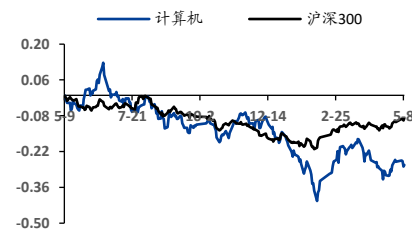
自 Deepseek-V2 率先降价开始，国内诸如豆包、腾讯、阿里、讯飞等厂商纷纷跟进，截至目前共有 7 家企业共计 21 款大模型参与降价。从降价模型级别来看，轻量级及次主力级模型降价显著，但 GPT-4 级别的主力模型也有一定程度的降价。随着大模型价格的进一步降低，B 端客户，特别是云厂的现有客户将加速大模型产品的推进过程，商业化落地有望加速。

➤ 风险提示

大模型迭代不及预期的风险；大模型商业落地不及预期的风险等。

强于大市（维持评级）

一年内行业相对大盘走势



团队成员

分析师： 钱劲宇(S0210524040006)
QJY3773@hfzq.com.cn

相关报告

正文目录

1	本周 AI 新闻速递.....	3
1.1	国内 AI.....	3
1.2	国外 AI.....	3
2	端侧盛宴即将开始，AIPC 决战生产力时代	4
2.1	微软发布多模态小模型 Phi-3 Silica，端侧实现 SOTA	4
2.2	正式发布 phi-3 系列模型，小参数，高性能	5
2.3	Copilot Studio，简化 Agent 开发流程	6
3	国产大模型迎降价潮，商业落地按下加速键	8
4	风险提示	10

图表目录

图表 1:	Windows Copilot Runtime.....	5
图表 2:	Phi-3 系列：性能与参数量对比	5
图表 3:	Copilot Studio 的 Agent 构建	6
图表 4:	GitHub Copilot	7
图表 5:	GitHub Copilot 提供的代码均是可编辑的.....	7
图表 6:	GitHub Copilot Workspace 能够在手机端使用	8
图表 7:	5 月大模型降价后单价情况	9
图表 8:	国内外大模型性价比盘点	10



1 本周 AI 新闻速递

1.1 国内 AI

(1) **面壁智能发布最强多模态开源模型 MiniCPM-Llama3-V 2.5**。综合性能超越 Gemini Pro、GPT-4V。OCR 能力 SOTA，9 倍像素更清晰，能够实现难图长图长文本的精准识别。图像编码 150 倍加速。与之相对的，该模型参数量仅有 70 亿。

(2) **黑马 Yi-Large 刷新榜单，成功登顶国产模型第一**。零一万物旗下 Yi-Large 大模型于大模型竞技场中成功跃升总榜第七，成为榜上国产大模型第一。在中文类别中，Yi-Large 与 GPT-4o 并列总榜第一。此外，LMSYS 还实施了重复数据删除机制，并提供去除冗余查询后的榜单。在该榜单下，Yi-Large 的 ELO 得分更进一步，与 Claude 3 Opus、GPT-4-0125-preview 并列第四。

(3) **百川发布首个 AI 应用百小应**。5 月 22 日，百川智能在北京召开发布会，发布自己的最新模型 Baichuan 4，并同时推出了自己的首款移动端应用——百小应。百小应的定位为 AI 智能助手，与文心一言、Kimi 智能助手相似，但主打的特点是「懂搜索、会提问」，这点又与目前很火的 AI 搜索应用类似。

(4) **国产大模型迎降价潮**。自 Deepseek 大幅度降价后，字节跳动豆包、阿里通义千问、腾讯、百度等互联网厂商纷纷跟进，各自调低了主力模型的 api 价格，诸如腾讯混元 standard-256K 版本输入输出价格分别调低 87.5%、50%。部分厂商彻底放开其轻量版本。诸如腾讯混元-lite、百度 ERNIE Speed 和 ERNIE lite。

1.2 国外 AI

(1) **Meta 发布 34B 的混合模态模型 Chameleon**。与 GPT-4o 一样，Chameleon 采用了统一的 Transformer 架构，使用文本、图像和代码混合模态完成训练。以类似文本生成的方式，对图像进行离散「分词化」(tokenization)，最终生成和推理交错的文本和图像序列。通过 10 万亿多模态 tokens 训练，在纯文本任务中性能与 Gemini-pro 相当；在视觉问答和图像标注基准上，性能接近 GPT-4V，成功刷新 SOTA。

(2) **谷歌发布文生图模型 Imagen 3**。谷歌在 I/O 2024 上发布全新文生图模型 Imagen3，Imagen3 可生成漫画、写实、赛博朋克、水墨、油画、素描等几十种风格，尤其是生成的写实风格图片质量很好，和真人拍摄的相片相差无几，写实能力比肩 Midjourney DALL·E 3 等头部产品。

(3) **首个 GPU 高级语言发布，编写并行代码门槛显著降低**。Bend 是一个大规模并行的高级编程语言，与 CUDA、Metal 等低级替代方案不同，Bend 具有 Python、Haskell 等表达性语言的功能，包括快速对象分配、完全闭包支持的高阶函数、无限制的递归，甚至 continuation。Bend 运行在大规模并行硬件上，具有基于核心数量的近线性加速。通过使用 Bned，开发者无需成为 C/CUDA 专家，能够像使用 python 一



样为多核 CPU/GPU 编写并行代码。

(4) Falcon 2 回归，综合性能超越 Llama3。 Falcon 2 包括两个版本：Falcon 2 11B 和 Falcon 2 11B VLM。前者是在 5.5 万亿个 token 上训练出的大语言模型，后者则在此基础上开发的「视觉到语言」模型。Falcon 2 11B 具有多语言能力，可以快速响应英语、法语、西班牙语、德语、葡萄牙语和其他各种语言的任务。Falcon 2 11B VLM 作为视觉语言模型，能够从环境中识别和解释各种视觉信息，因此在医疗、金融、电子商务、教育和法律等领域有广泛的应用场景。

(5) Hugging Face ZeroGPU 计划正式发布：一千万美元的免费共享 GPU。 ZeroGPU 是 HuggingFace 的 Spaces 平台的一种新型硬件集成，设计有两个主要目标：1.为 Spaces 应用提供免费的 GPU 访问；2.允许 Space 应用在多个 GPU 上运行，提高性能和效率。ZeroGPU 对所有用户开放，PRO 用户还可以托管自己的 ZeroGPU Spaces。为了帮助开发过程，Hugging Face 鼓励用户通过社区标签提供反馈，并探索已经可用的 ZeroGPU Spaces 列表。

(6) 微软发布 Copilot+PC，重新定义 AI 生产力。首批上市的 AIPC 产品将搭载高通 Snapdragon X 系列处理器，在今年晚些时候将推出更多基于英特尔和 AMD 处理器的产品。由于搭载了 40+TOPS 的 NPU，AI 工作负载能力增强了 20 倍，效率提高了 100 倍。Phi-Silica 完全使用 NPU 进行推理，首个 token 输出速度为 650 token/s，仅消耗约 1.5 瓦的电量，可以让 CPU 和 GPU 同时用于其他计算任务。持续运行时，文本生成重用 NPU 中的 KV 缓存并在 CPU 上运行，每秒生成约 27 个 token。

(7) 英伟达 GH200 芯片落地 9 个超算中心，总算力规划达 200 exaflops。即将上线的 9 台超算包括法国的 EXA1-HE、波兰的 Helios、瑞士的 Alps、德国的 JUPITER、美国 UIUC 的 DeltaAI 以及日本的 Miyabi。GH200 芯片在 AI 和高性能运算方面有非常强大的能力，单个 GH200 芯片由一个 72 核 Grace CPU 和一个 H100 GPU 组成，内存容量最高可达 624GB。在单个服务器上，通过 NVlink 连接的双 GH200 芯片可以提供比 H100 高 3.5 倍的 GPU 内存容量和 3 倍的带宽。

(8) 视频版 PS 发布，0 样本视频编辑模型。巴黎矿业大学、以色列理工学院的研究人员联合推出了一款创新视频模型——Slicedit。Slicedit 主要通过结合文生图像的扩散模型以及对视频时空切片的预处理，在不影响背景的情况下对视频主体进行修改。例如，将一名冲浪的男人变成钢铁侠，将一名转球的男孩变成 NBA 巨星库里，将猫的样子变成豹等。

2 端侧盛宴即将开始，微软决战生产力时代

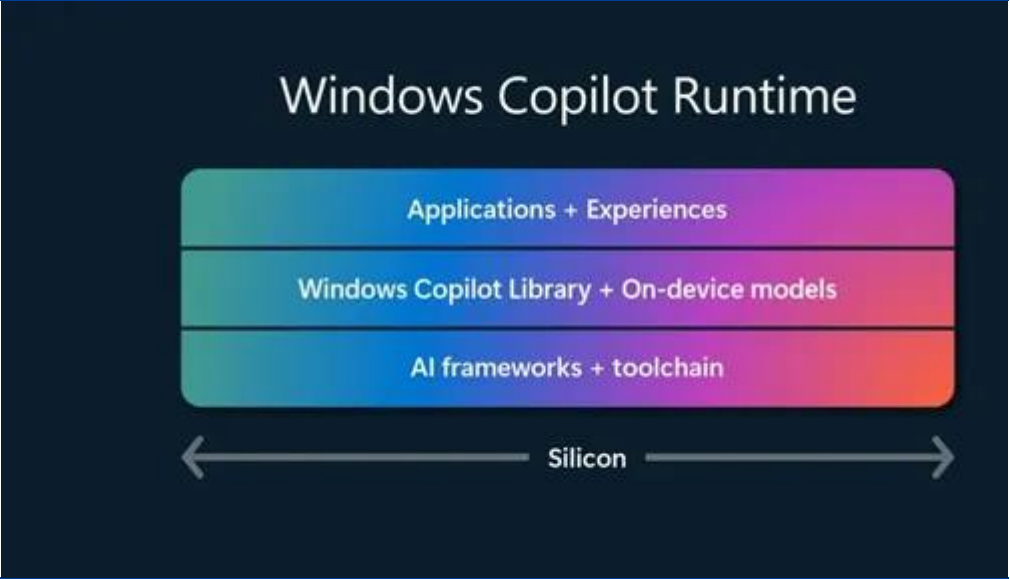
2.1 微软发布多模态小模型 Phi-3 Silica，端侧实现 SOTA

这是一个专为 Copilot+PC 设计的端侧小模型（SLM），能够充分适配 AIPC 的

NPU，以满足模型计算的算力要求。Phi-3 Silica 仅有 3B 参数，是 Phi-3 系列中最小的。

Phi-3-Silica 已被嵌入到 Windows Copilot Runtime 中。其具备丰富的 AI 框架：包含 DirectML、ONNX Runtime、PyTorch、WebNN 以及工具链如 Olive、Visual Studio Code 的 AI 工具包等，可以帮助开发人员快速引入自己的大模型，并在 Windows 硬件生态系统的广泛范围内扩展其 AI 应用。

图表1: Windows Copilot Runtime

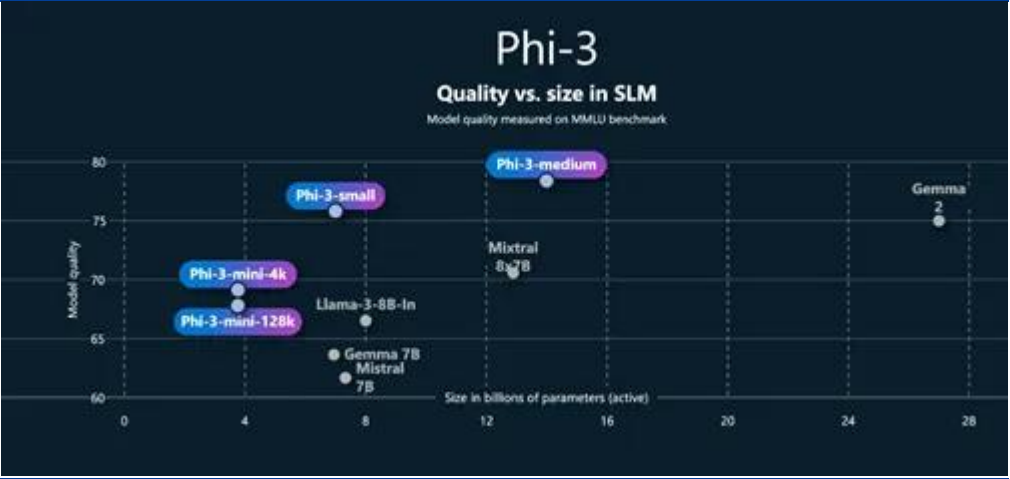


来源：AIGC 开放社区，华福证券研究所

2.2 正式发布 phi-3 系列模型，小参数，高性能

Phi-3 一共有 4 种参数：Phi-3 vision、Phi-3 mini、Phi-3 small、Phi-3 medium 分别有 42 亿、38 亿、70 亿、140 亿参数。其中 vision 具备视觉识别和转化功能；mini 支持 128K 和 8K 上下文窗口；small 综合能力超越 GPT 3.5；medium 性能优于 Gemini 1.0 Pro。各个版本各有千秋，从当前 huggingface 下载量来看，mini 最受欢迎，下载次数超 100 万次。

图表2: Phi-3 系列：性能与参数量对比

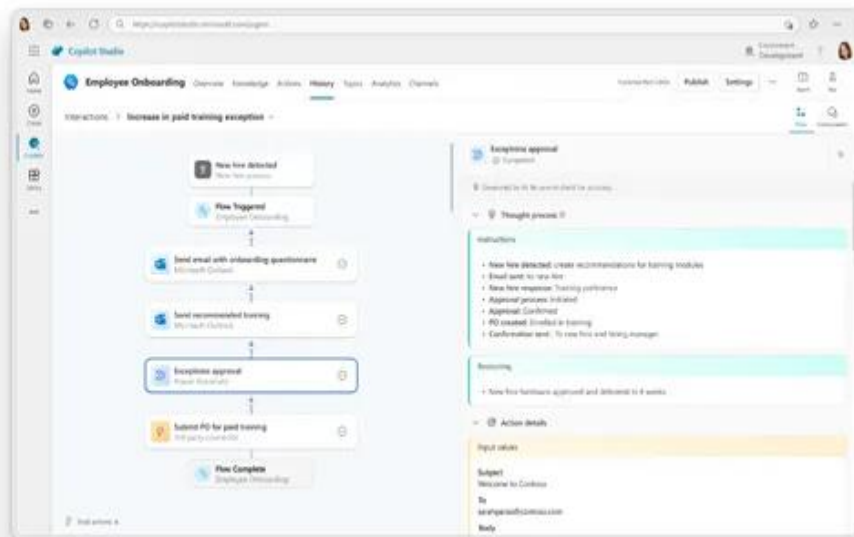


来源：AIGC 开放社区，华福证券研究所

2.3 Copilot Studio, 简化 Agent 开发流程

Copilot Studio 的使用方法非常简单, 只需要告知该 Copilot 的功能以及他所预留的知识, 该 Agent 的思维执行链便会立刻创建完成。也就是说, 创建者无需专业的编程能力, 也能够实现 Agent 的创造。

图表3: Copilot Studio 的 Agent 构建



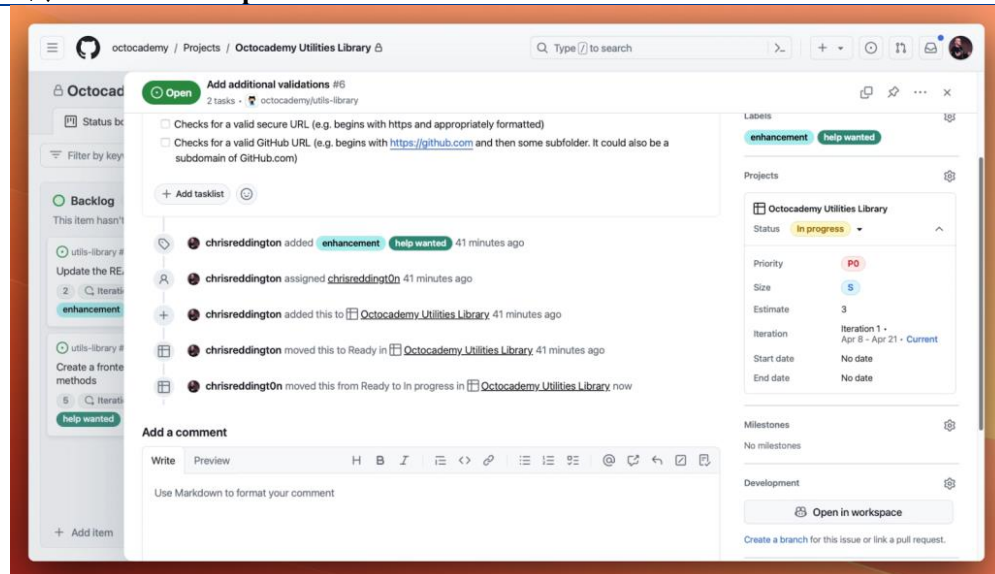
来源: AIGC 开放社区, 华福证券研究所

通过 Copilot Studio 构建 Agent 有以下优势:

- 对长期稳定运行的业务流程完成自动化流程。
- 对操作和用户输入进行推理。
- 利用记忆功能带入上下文。
- 根据用户反馈进行自主学习。
- 记录异常请求并寻求帮助。

GitHub Copilot 作为微软推出的扩展功能之一, 可以利用自然语言交互来提高开发速度。对于开发者来说, 迈出开发的第一步至关重要。在最初的解决方案制定上, GitHub Copilot 能够作为人类的第二大脑, 为开发人员提供建议和思路。通过对于代码库的深入理解, 提供解决问题的分步计划。

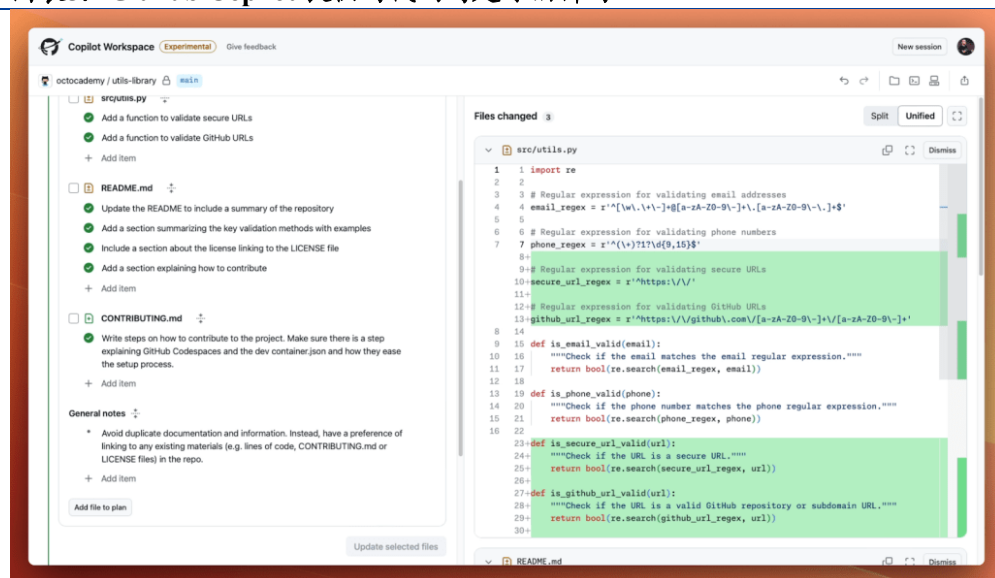
图表4: GitHub Copilot



来源: AIGC 开放社区, 华福证券研究所

GitHub Copilot 提出的所有内容, 从计划到代码方案, 都是完全可编辑的。也就是说, 开发人员只需要关注代码本身, 其他事情可以交给 GitHub Copilot。通过对流程和资源精准把握, 只要开发人员跟他提需求他就会帮助整合 workflow, 帮助在信息平台信息中切换并且调用必要的工具, 如果在开发过程中遇到问题, 他可以实时帮你解答。

图表5: GitHub Copilot 提供的代码均是可编辑的



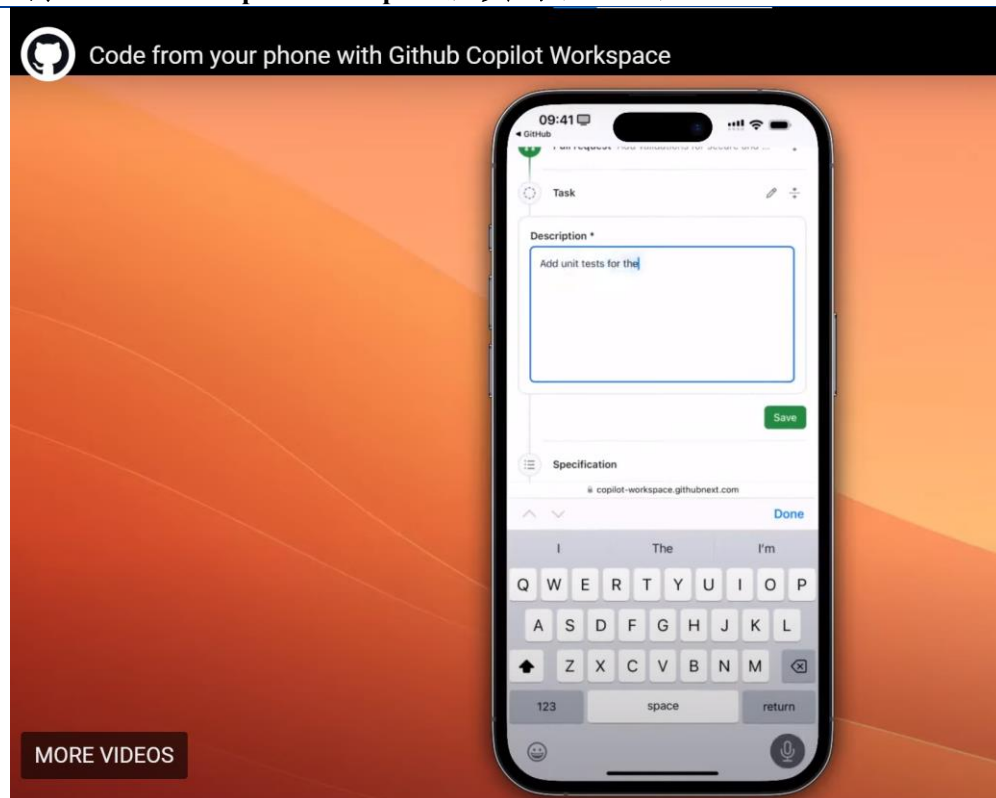
来源: GitHub CEO 官博, 华福证券研究所

GitHub Copilot Workspace 是一个 Copilot 原生的开发环境, 它允许开发者从构思到代码再到软件的整个过程中, 都能使用自然语言来进行交流和操作。GitHub Copilot Workspace 能够直接运行代码, 并支持所有代码的随时调整, 直到开发人员对程序的运行效果满意。

此外, GitHub Copilot Workspace 能够与移动设备兼容。GitHub Copilot Workspace 能够在台式机、笔记本电脑或其他移动设备上部署真实的开发环境, 方便开发人员随时将灵感转换为代码。



图表6: GitHub Copilot Workspace 能够在手机端使用



来源: Github 官方 Youtube, 华福证券研究所

3 国产大模型迎降价潮，商业落地按下加速键

Deepseek 打响国内第一枪，其他大厂纷纷跟进。5月6日，Deepseek-V2 率先降价，每百万 Tokens 的输入、输出价格分别为 1 元、2 元。从模型性能来看，V2 对标 GPT-4，但价格仅有 GPT-4 的 1%，性价比极高。随后，智谱、豆包、通义、腾讯、讯飞等厂商纷纷跟进，目前参与的企业共计 7 家，涉及 21 款模型。大模型正式迎来降价潮。



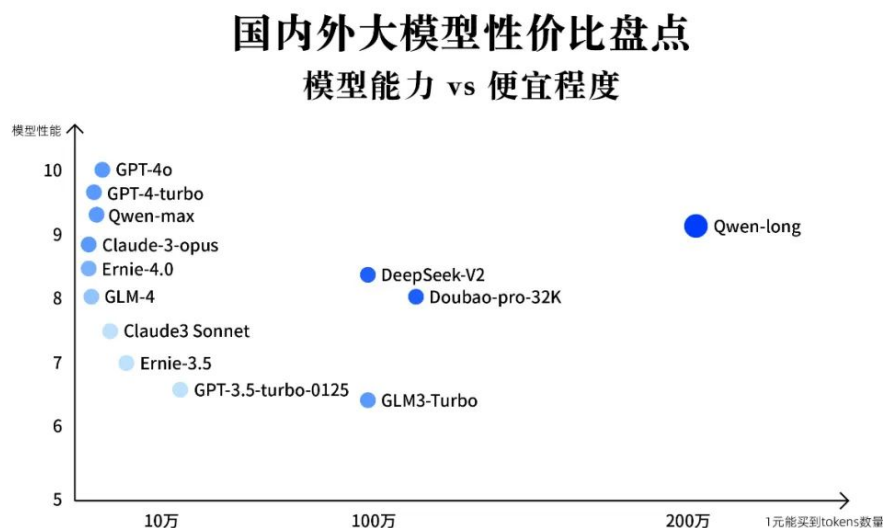
图表7: 5月大模型降价后单价情况

大模型名称	输入价格 元/千 tokens	输出价格 元/千 tokens	宣布价格或宣布 下调价格的日期	所属公司
DeepSeek-V2	0.001	0.002	5月6日	深度求索
GLM-3-Turbo	0.001	0.001	5月11日	智谱 AI
豆包通用模型 pro-32k	0.0008	0.0008	5月15日	字节跳动
通义千问 Qwen-Max	0.04	0.12	5月21日	阿里云
通义千问 Qwen-Plus	0.004	0.002	5月21日	阿里云
通义千问 Qwen-Long	0.0005	0.002	5月21日	阿里云
文心一言 ERNIE Speed	免费	免费	5月21日	百度
文心一言 ERNIE Lite	免费	免费	5月21日	百度
讯飞星火 spark Lite	免费	免费	5月22日	科大讯飞
讯飞星火 Spark3.5 Max	0.021-0.03	0.021-0.03	5月22日	科大讯飞
混元-lite	免费	免费	5月22日	腾讯云
混元-standard	0.0045	0.005	5月22日	腾讯云
混元 -standard-256k	0.015	0.06	5月22日	腾讯云
混元-pro	0.03	0.1	5月22日	腾讯云

来源: 贝壳财经, 华福证券研究所

从轻量模型到主力模型均有不同程度降价。虽然说当前免费的模型大多为轻量模型及次主力模型。但国产主力模型当前的价格也有了显著的降低。以 GPT-4 级主力模型 Qwen-Long 为例, API 输入价格直降 97%, 约为 GPT-4 价格的 1/400, 输出价格也降 90%至每百万 tokens 2 元, 1 元钱能够买 200 万 Tokens。且该模型支持最高一千万 tokens 的长文本输入, 适用于很多的 B 端统计分析类的应用场景。而在 OpenCompass benchmark 上追平 GPT-4 Turbo 的 Qwen-Max, 输入价格降幅 67%, 亦有较大幅度降价。

图表8: 国内外大模型性价比盘点



来源：界面新闻，华福证券研究所

大模型迎来价格换市场阶段，商业落地按下加速键。阿里云智能集团资深副总裁、公共云事业部总裁刘伟光在演讲中表示，阿里云这次大幅降低大模型推理价格的目标是希望推动 AI 应用的爆发，加速商业落地的进程。阿里云方面公布的最新数据显示，通义大模型通过阿里云服务企业超过 9 万、通过钉钉服务企业超过 220 万，已在 PC、手机、汽车、航空、天文、矿业、教育、医疗、餐饮、游戏、文旅等领域落地应用。随着大模型价格的进一步降低，B 端客户，特别是云厂的现有客户将加速大模型产品的推进过程，商业化落地有望加速。

4 风险提示

大模型迭代不及预期的风险：目前大多数大模型的研发仍处于早期投入阶段，存在研发失败以及迭代进度缓慢的风险。

大模型商业落地不及预期的风险：当前大模型与各个场景结合处于早期磨合阶段，仍需要进一步的训练、微调及试用，存在落地时间点不及预期的风险。



分析师声明

本人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。

一般声明

华福证券有限责任公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

本报告版权归“华福证券有限责任公司”所有。本公司对本报告保留一切权利。除非另有书面显示，否则本报告中的所有材料的版权均属本公司。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。未经授权的转载，本公司不承担任何转载责任。

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	评级	评级说明
公司评级	买入	未来 6 个月内，个股相对市场基准指数涨幅在 20%以上
	持有	未来 6 个月内，个股相对市场基准指数涨幅介于 10%与 20%之间
	中性	未来 6 个月内，个股相对市场基准指数涨幅介于-10%与 10%之间
	回避	未来 6 个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来 6 个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来 6 个月内，行业整体回报高于市场基准指数 5%以上
	跟随大市	未来 6 个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来 6 个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的 6~12 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准（另有说明的除外）

联系方式

华福证券研究所 上海

公司地址：上海市浦东新区浦明路 1436 号陆家嘴滨江中心 MT 座 20 层

邮编：200120

邮箱：hfyjs@hfzq.com.cn