

中国科技行业

本文内容由 Ryan Lee (李昊谦) 提供

中信证券财富管理(香港)

产品及投资方案部

电话: (852) 2237 9250 / 电邮: wminvestmentsolutions@citics.com.hk

Ox Alpha 发布

摘要

中信证券财富管理与中信里昂研究观点一致。根据中信里昂研究在 2026 年 8 月 27 日发布的题为《Ox Alpha unveiled》的报告，智谱（Z.AI）在匿名发布中推出其首个原生多模态模型 GLM 5.3 Flash，广为人知的名称为“Ox Alpha”（“牛来”）。其在具备视觉能力加持下提供卓越的成本调整后性能，并且很可能已成为 Flash 模型的新全球标准。随着中国模型迭代节奏加快，研究指出其全球受欢迎程度和强劲需求，这从 OpenRouter 上不断上升的收入份额和强劲的经常性收入（ARR）可见一斑：继 Z.AI 于 7 月达到 10 亿美元里程碑后，MiniMax 和 MoonShot 于 8 月初分别达到 8 亿美元和 10 亿美元 ARR。

Flash 模型将掀起波澜

- Z.AI 推出 GLM 5.3 Flash，广为人知的名称为 Ox Alpha。在较小参数规模（320B/18B active）下，其以 GLM 5.2 十分之一的成本实现更优性能，并以四分之一的成本在编码能力上匹敌 Opus 4.8。按每百万输入/输出 tokens API 定价 0.8 元/2.8 元计算，其较 DeepSeek V4 Flash 便宜约 40%。
- 这一性价比提升由一种为超低成本推理设计的更高效架构驱动。其结合稀疏和线性注意力，以降低长上下文服务成本，同时利用 mHC 增强模型可扩展性。每层 KV cache 大小和每层注意力计算较 GLM 5.3 分别下降 4 倍和 3 倍。
- 同日，阿里巴巴展示了多模态 Qwen 3.8 Flash-Next（125B/6B active，配备 51B N-gram embedding），按每百万输入/输出 tokens API 定价 1 元/3 元。

从基于文本到多模态

- GLM 5.3 Flash 是 Z.AI 首个原生多模态模型，基于 30T 多模态数据训练。其也针对专业服务工作流程进行了优化，例如金融和法律。
- 视觉编码拓宽了编码任务范围，提升其观察环境的能力，并增强模型在前端设计以及涉及 pdf、docx、xlsx 格式等办公工作中的能力。
- 截至目前，所有中国前沿实验室均至少有一个视觉/多模态模型，包括 DeepSeek V4 Flash Vision（8 月 21 日）、Qwen 3.8 27B（8 月 14 日）、K3（7 月 16 日）、M3（6 月 1 日）。

国产芯片上的推理

- 过去六天内，Ox Alpha 在 OpenRouter 和 OpenCode 上的 token 排行榜居首，全球使用总量达到 65trn，为有史以来最大的发布。在 OpenRouter 上，包含 Ox Alpha 在内的 Z.AI 周 token 份额（23%）超过 DeepSeek（16%）。
- 据 Z.AI 称，该推理服务完全运行于国产硅基芯片之上，并由自研高带宽网络和基于 SGLang 的专用推理引擎支持。
- 根据本地媒体 LatePost，GLM 5.3 Flash 很可能利用了来自海光、华为和摩尔线程的 100,000 枚国产芯片集群提供推理服务。
- Z.AI 在其技术文件中指出，该集群的硬件效率和 token 单位经济性已达到可与主流英伟达 GPU 相比的水平。

OpenRouter 数据说明了什么？

- 随着 OpenRouter 的周 token 使用量接近 99trn（年初至今 14 倍），在全球超过 400 个模型支持下，合计 ARR 达到 10 亿美元（年初至今 4 倍）。
- 截至 8 月 23 日当周，在 Z.AI、DeepSeek 和 MoonShot 强劲增长带动下，中国模型的合计 token/收入份额进一步升至 72%/31%。
- Anthropic/OpenAI/Google 模型的收入份额分别为 38%/18%/10%，相比之下，其各自峰值份额分别为 66%/19%/21%。
- 中国模型总支出年初至今激增 20 倍，由 token 消耗增加 41 倍驱动。尽管如此，中信里昂认为中国实验室盈利仍偏低，原因在于算力供应有限，这使第三方推理服务提供商和云合作伙伴受益。

免责声明

本文由中信证券财富管理（香港）（“**中信证券财富管理（香港）**”）撰写。中信证券财富管理（香港）是中信证券经纪（香港）有限公司（“**中信证券经纪（香港）**”）的财富管理品牌。本文为一般性资讯，仅供参考之用。本文具有教学性质，但不被视为对任何司法管辖区的任何特定投资产品、策略、计划特征或其他目的的建议或推荐，中信证券财富管理（香港）或其任何子公司也不承诺会参与任何此处提及的交易。任何示例都是通用的、假设的并且仅用于说明目的。本文尚未经香港证券及期货事务监察委员会审核。

本文是由投资／产品专员而非分析师撰写的文章汇编。它不构成研究报告，也不应被解释为研究报告，也不旨在提供专业、投资或任何其他类型的建议或推荐。本文件所载观点为投资／产品专员个人观点，或投资／产品专员对公司观点的理解，可能与中信证券财富管理（香港）的官方观点和利益（包括投资研究部门的观点）存在差异，也可能不存在差异。

适合性、风险及适当性

本文不含足够资讯以支持投资决策，并且不得依赖本文来评估投资任何证券或产品的优劣。本文仅提供一般资讯，未针对任何接收者量身定制，亦未考虑任何接收者的投资目标、投资组合持仓、策略、财务状况或需求。本文具有教育性质，并非中信证券经纪（香港）或其附属公司参与任何交易的建议、推荐或承诺。任何范例均为通用、假设及仅供参考之用。

读者应自行独立评估法律、监管、税务、信用及会计方面的影响，并与其财务专业人员共同决定本文提及的任何投资是否适合其个人目标。投资者应确保在进行任何投资前获取所有相关资讯。本文中所涉及的任何预测、数据、观点、技术或策略仅供参考，基于某些假设及当前市场状况，可能会在未通知的情况下发生变更。

信息来源被认为是可靠的，但不保证其准确性，也不对任何错误或遗漏承担责任。请注意，投资涉及风险，投资的价值和收益可能会根据市场情况和税收协议而波动，投资者可能无法收回投资的全部金额。过去的表现和收益率都不是当前和未来结果的可靠指标。

本文不应被解释为在任何司法管辖区向任何不合法接受此类邀请或招揽的人发出购买或出售任何证券的邀请或招揽。此处包含的信息均未在任何司法管辖区获得批准。

责任限制

中信证券财富管理（香港）对于您或任何第三方在使用此材料或其内容时遭受的任何损失或损害（包括但不限于直接损失、间接损失、或由此导致的间接损失、利润损失或机会损失）不承担任何责任。此材料是根据您的请求提供给您的，并适用于熟悉相关市场且愿意并能够承担投资相关风险的个人。在您充分了解所有涉及风险并独立确定该投资适合您之前，请勿进行任何交易。请注意，我们并非您指定的投资顾问。

法律主体声明

香港：本资料由中信证券经纪（香港）有限公司（中央编号 AAE879）在香港发布及分发，该公司受香港证券及期货事务监察委员会监管。

附属披露

本文可能包含先前发布的信息，包括由 CLSA Limited（中央编号 AAB893）发布的信息，该公司受香港证券及期货事务监察委员会监管。中信证券经纪（香港）与 CLSA Limited 是关联实体，均属于中信证券国际有限公司（“**CSI**”）的共同所有或控制之下。

中信证券经纪（香港）与 CSI 及其子公司（“**CSI 集团**”）内的其他实体之间可能存在潜在的利益冲突，这些冲突可能涉及多个领域，包括投资产品、企业咨询服务、证券交易以及提供给本文提及公司（们）的其他金融服务。此外，CSI 集团可能在本文讨论的公司（们）中持有利益或头寸。

中信证券经纪（香港）遵循一项整个集团的政策（可能会不时更新），该政策阐明了 CSI 集团实体如何管理或消除任何实际或潜在的利益冲突。此政策旨在确保 CSI 集团发布的任何研究报告或评论保持公正。

针对不同司法管辖区的声明

香港：根据任何时间下所适用的法律和法规，中信证券财富管理（香港）、其各自的关联公司或与中信证券财富管理（香港）有关联的公司或个人可能在发布本资料前使用了此处包含的信息，并且可能不时购买、出售或在所提及的任何证券或相关证券中拥有重大利益，或在目前或将来可能产生或曾经有业务或财务关系，或可能或已经向本文提及的实体、其顾问和／或任何其他关联方提供投资银行、资本市场和／或其他服务。因此，投资者应注意中信证券财富管理（香港）和／或其各自的关联公司或公司、或此等个人可能存在一项或多项利益冲突。

未经中信证券经纪（香港）有限公司事先书面授权，任何人不得以任何目的复制、分发或销售本文。

版权所有 2026。中信证券经纪（香港）有限公司。保留所有权利。