智谱大跌惊魂：大模型商业化竞赛提速

本报记者 黎竹 张靖超 北京报道

近日，“全球大模型第一股”智谱（02513.HK）就GLM Coding Plan改版发布致歉信，承认规则透明度不够、GLM-5灰度节奏太慢、老用户升级机制设计粗糙等问题。随后其股价单日跌幅超20%。

随着新一波人工智能浪潮的来临，今年以来，多家大模型相关企业都在加速更迭版本，并寻求商业上的变现。其中，智谱于2月12日正式推出新一代旗舰模型GLM-5，并于隔日同步上调GLM Coding Plan套餐价格，其中，中国区涨价30%，海外版涨价超100%，成为国内首家对大模型商业化服务进行提价的AI原生企业。对于GLM-5是否存在因市场竞争加剧而加速上线的情况，截至发稿，智谱方面暂未回应。

萨摩耶云科技集团首席经济学家郑磊认为，大模型企业在快速迭代与用户预期管理之间存在结构性矛盾。快速迭代是技术竞争的关键，但频繁更新可能导致用户难以适应，甚至对产品稳定性产生疑虑。同时，他指出，模型迭代带来能力边界扩张，但“幻觉”问题始终未解。企业需要在推进技术的同时，通过透明沟通和阶段性发布来管理用户预期、平衡创新与用户体验。

极速推出GLM-5

自年初上市以来，智谱在资本市场上一路狂飙，市值一度冲破3000亿港元大关。

智谱在致歉信中坦言：“很多老用户反馈‘消耗变快了’，这里面有我们的锅，也有一些客观的算力逻辑。”

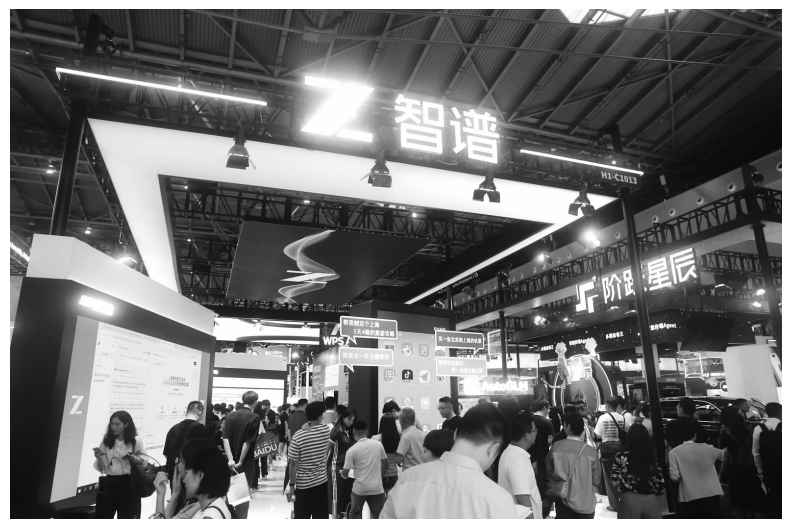
和最初DeepSeek上线一样，由于对GLM-5的需求激增，导致服务出现排队、响应延迟及卡顿现象，影响到部分用户的体验。一位开发者告诉《中国经营报》记者：“我订的是每月149元的Pro套餐，GLM-5的tokens消耗太快了，一下子就消耗了一半额度。”

对此，智谱亦在信中承认：为应对GLM-5更高的算力消耗，公司设计了分层使用策略，将高峰期消耗提升至3倍、非高峰期2倍，但未向用户清晰说明，导致“消耗变快”的普遍抱怨。

值得一提的是，在致歉信中智谱也给出了补偿方案：即分级开放GLM-5，受影响的用户可自主申请退款，部分用户使用期延长15天，而误升级的用户则可“一键回滚”。

前述致歉信发布后，智谱股价在2月23日单日跌超20%。自年初上市以来，智谱在资本市场上一路狂飙，市值一度冲破3000亿港元大关。然而，截至发稿，智谱股价为571.5港元/股，市值为2548亿港元，与年内725港元/股的高点相比，其股价回落明显。

对此，盘古智库高级研究员江瀚告诉记者：“能力边界的模糊化会加剧消费者和投资者的预期落差。如果企业在宣传中常将实验室的‘峰值能力’等同于商业化的‘稳定能力’，过度承诺通用智能，而用户在实测中发现模型在垂直场景的幻觉率与稳定性仍存在波动，这种‘营销通胀’会迅速透支市场信任。”



智谱承认规则透明度不够、GLM-5灰度节奏太慢、老用户升级机制设计粗糙等问题。

视觉中国/图

时间拨回到1月8日智谱上市当天，清华大学计算机系教授、智谱创立发起人兼首席科学家唐杰发布题为《用“咖啡”的精神做AGI》的内部信，并宣布很快将推出新一代模型GLM-5。

2025年12月23日，唐杰还在微博上表示，2026年将是AI替代不同工种的爆发年。当时，他认为，随着基础模型能力提升，Agent和领域大模型最终都将与基础模型结合，甚至，AI也不一定意味着需要创建新的应用。“大模型的应用也要回到第一性原理。”

记者注意到，在发布GLM-5后，由于供不应求，智谱新套餐上线即售罄。但对于目前的用户数量以及未来的商业模式和市场布局，智谱方面未回应记者。

中国企业资本联盟副理事长柏文喜认为，目前技术迭代速度与商业化节奏错配情况显著。他这样解释：“大模型行业遵循‘Scaling Law’（规模定律），需要持续巨额投入追求技术突破，但用户预期建立在现有能力边界上。比如当GPT-4o实现原生多模态、o1推出推理能力时，用户期望被瞬间拉高，但企业面临算力成本飙升（如o1 API定价是GPT—4o的3—4倍）与商业化落地滞后之间的矛盾。这种‘技术跃进—成本激增—盈利遥远’的循环，导致企业不得不在‘展示未来能力’与‘兑现当下价值’间走钢丝。”

如何实现商业变现

当下大模型产业正经历从“技术浪漫”到“落地求生”的残酷转型。

自ChatGPT问世以来，国内大模型服务犹如潮水般涌现，其中以豆包、DeepSeek、千问、元宝、文心等大模型服务开始重塑人们获取信息的方式，与此同时，智谱AI、MiniMax、百川智能、零一万物、月之暗面与阶跃星辰这六家大模型初创企业也被称为大模型“六小虎”。

而随着千问、豆包、元宝等AI应用推出花样拉新活动占领市场份额，业内开始对大模型公司的商业模式产生讨论：大模型企业如何实现商业化目标，其估值或市值是否存在泡沫风险？

当下大模型产业正经历从“技术浪漫”到“落地求生”的残酷转型。快速迭代与用户预期的矛盾、估值与基本面的背离、算力约束与算法创新的博弈，共同构成了核心张力。多位业内人士认为，投资者需关注技术落地进展和商业化能力，以避免过度乐观带来的风险。

江瀚指出，目前市场分化正在加速，泡沫呈现结构性特征。缺乏造血能力的纯模型层初创企业市销率高达数百倍，远超互联网泡沫时期的峰值，其估值逻辑仍停留在“参数规模”与“用户增长”的早期阶段，忽视了高昂的算力折旧与现金流压力。但他也分析道：“拥有闭环应用场景、稳定B端订单及自有算力基础设施的头部企业，其高估值有业绩兑现支撑，泡沫相对可控；而单纯依赖融资烧钱、缺乏商业化落地路径的‘中间层’模型厂商，面临极大的估值回调风险。”

摩根大通近期发布研报称，智谱API定价行为是前沿领先能力信号，并指出智谱已经到达一个重要的拐点，尤其是其全球API业务。该行预计，随着GLM-5在开发者中的认可度提升，使用率会快速提高，尤其是在以编码为中心的工作流中，这些工作流的支付意愿和使用强度较高。

一位科研人员告诉记者，目前美国的头部大模型企业本质上还是“卖token”，OpenAI和Antropic目前走的都是订阅和API收费的模式，尽管也做了一些垂类模型，但并没有亲自下场，主要还是联合企业用这些模型来开发应用，满足企业在市场上的新需求。但问题在于，“OpenAI和Antropic是闭源，能够推动商业变现，对比之下开源大模型需要在提高用户付费意愿上下功夫”。

同时，发展AI大模型背后需要强大的资金来支撑，以此构建强大的AI算力基础设施。智谱披露的数据显示，研发投入中有七成是购买算力服务。2026年2月16日，智谱甚至公开发文，全网寻找“算力合伙人”，希望通过寻求芯片厂商、算力伙伴与推理服务商以及其他形式的算力合作，以给用户提供更极致的智能体验。

值得一提的是，2026年1月14日，智谱宣布联合华为开源新一代图像生成模型GLM-Image，称其为首个全程在国产芯片上完成训练的SOTA多模态模型。就此，记者向智谱方面询问未来主力模型是否将全部转向国产算力，截至发稿，智谱方面未回应。

江瀚指出，受限于先进制程设备禁运，国产算力在单卡绝对性能与万卡集群的线性加速比上，较国际顶尖水平仍有代差，这在训练超大规模基座模型时会遭遇效率瓶颈与成本劣势。但若通过异构计算架构创新、软硬协同优化以及集群调度算法的提升，国产算力有望在系统级效率上弥补硬件短板。

在业内看来，中美大模型竞争的核心变量是“生态”，而非单一技术要素。柏文喜提到，算力、算法和数据是基础，但生态的完善程度决定了技术的应用广度与深度，而包括开发者社区、行业合作、政策支持等在内的生态因素，将直接影响大模型的落地速度和市场竞争力。

但多个采访对象的一致看法是，市场化能力正在取代纯技术参数，成为市场定价的主导因素。天使投资人、资深人工智能专家郭涛直言：“今年是大模型‘商业化落地’的攻坚之年，企业能否将技术红利转化为可持续的商业价值至关重要。”

存储涨价潮冲击 2026年手机市场承压

中经记者 谭伦 北京报道

在存储芯片持续涨价冲击下，全球手机市场将承压。

日前，市场研究机构Counterpoint Research发布的最新研报显示，为了应对DRAM（动态随机存储器）和NAND Flash（闪存）的供求和价格上涨带来的成本压力，手机厂商将被迫对旗下终端产品进行涨价，而这也将在一定程度上抑制消费者的购买需求。

因此，Counterpoint Research预计，此前一直受益于DRAM和NAND Flash成本下降的中低端智能手机市场，将会在2026年承受前所未有的巨大压力，出货量将较2025年同比下降6.1%，同时，智能手机SoC出货量同比下降7%。

另一大研究机构TrendForce在2月最新跟踪报告中亦指出，2026年第一季度，全球DRAM合约价上涨90%—95%，NAND合约价亦大幅上调。在此背景下，若干移动存储子品类（包括部分LPDDR5X与UFS4.0规格）与消费级1TB SSD在近几个月内出现了数倍级或接近翻倍的现货与零售价格上涨。

受此推动，2026年在全球手机市场涨价也成为业内共识。CHIP中国研究室主任罗国昭向《中国经营报》记者表示，此轮普涨来自AI训练需求的挤压，使得全球内存产能告急甚至缺货，虽然半导体市场供需波动遵循周期规律，价格未来将回调，但短期年内涨价或成为主要基调。

国内3月或迎普涨

2025年，中国手机市场出货量已占全球整体手机市场的23%左右，成为全球目前最大单一手机销售市场。因此，在此轮存储提价潮影响背景下，我国手机售价也或将在2026年迎来一轮全线涨价。

据供应链最新向媒体释出的信息，自2026年3月起，包括OPPO、一加、vivo、iQOO、小米、荣耀在内，即将推出各线新品的国内手机厂商将进入年后首轮涨价阶段，其涨幅最低1000元起步，部分中高端旗舰机型涨幅或将达2000—3000元。

国内某手机厂商采购部门负责人李军向记者透露，存储涨价压力从去年下半年已经传导至终端侧，但在库存周期下稍有迟滞，目前出货至海外的新机涨价已陆续开始，国内市场春节后预计也不例外，且今年这轮涨价潮可能将呈现“猛烈、多轮不定期”趋势。

李军同时表示，去年10月很多智能手机采购的存储芯片采购价格已较上年同期上涨超过了100%，而近三个月供应链现货价还在走高，有些上涨幅度已经接近200%，而且还在增加。

据公开报道，目前多家国内主流手机厂商已向各级经销商及渠道发出了调价通知，其中，新品发布高峰的3月被视为关键提价时间点，并预告这将是近五年来业内规模最大、涨幅最高的一轮集体调价潮。

Counterpoint Research预测，2026年全球智能手机平均售价将同比上涨6.9%，而今年3月后，国内市场新品手机均价将较2025年同档位机型上涨15%—25%，涨幅将显著高于全球水平，原因是国内厂商更倾向于配置16GB、24GB大内存。

以红米为例，记者注意到，其K系列新品售价已从去年年末开始最高上涨近300元，其涨价势头甚至延续在2025年12月“双十二”活动期间。而在今年春节期间，记者也在某华东省份县城多家手机线下店走访获悉，中低端手机在春节前的促销活动中并未如往年般降价，相反是小幅度涨价销售。

AI需求抢夺产能

此轮手机涨价潮背后，存储成本抬升被公认为最大驱动因素。包括TrendForce、IDC在内的多家市场研究机构一致认为，存储、内存向AI数据中心方向转移产能，是其价格上行的一个结构性原因。

具体而言，罗国昭表示，自大型生成式AI与高性能计算需求爆发以来，内存厂商把更多产能投向高带宽内存（HBM）与用于服务器的DDR5/下一代产品，以满足云端数据中心长期、利润更高的订单。厂商在短期内难以把这部分产能转回手机用的传统DRAM与主流NAND Flash，从而导致移动端可用供应紧缩、采购竞争加剧。“这就使得AI需求快速侵占了传统消费电子的内存供应份额。”罗国昭指出。

因此，在供需失衡下，合同价与现货价被迅速推高。记者注意到，就在2月初，TrendForce把2026年年初DRAM合约涨幅较此前预估的水平大幅上调。罗国昭认为，这表明产业链反馈的价格弹性正在被放大，OEM厂商在短期内无法通过议价完全对冲涨价压力。

罗国昭还表示，由于全球存储产业长期高度集中，当主流厂商优先选择数据中心客户的长期高利率订单时，短期内消费端供应就不会被补上。目前，尽管三星、美光、SK海力士等都在加大资本开支，但从新产线投产到产能显现通常需要数年时间，市场短缺与高价至少会持续到扩产见效之前。因此，产能修复大概率不会在几个月内完成，此次涨价可能不只是短期波动。

此外，李军向记者表示，在正常年份，存储芯片通常占手机整机物料成本的10%—15%。但在本轮涨价周期中，随着DRAM与NAND Flash价格大幅上行，该比例已显著提升至20%—30%。在部分高配低价机型中，存储成本占比甚至可能超过30%。

考虑到增加销售吸引力，李军指出，近年来很多主流智能手机也纷纷增加了AI功能，从而提升体验，因此向更高规格内存（如LPDDR5X/LPDDR6、UFS 4.x）迁移，这进一步抬高了单台手机的成本。在内存价格飙升的当下，这些升级成为厂商成本不可回避的一部分。

“因此，厂商要么吸收成本，要么通过涨价将压力转嫁给市场，或者降低出货量、压缩配置来平衡。”李军表示。

几家欢乐几家愁

随着内存与手机涨价潮的延续，其给产业链带来的影响也在逐步显现。

最大赢家自然当属存储厂商。Counterpoint Research预计，拥有强大存储产能的三星，其出货量将同比增长7%，市场份额预计同比增加0.9%—6.6%。此外，华为旗下海思半导体也有望在2026年保持4%的增长率。

受到冲击最大的为手机SoC厂商。Counterpoint Research预计，两大手机SoC芯片巨头高通和联发科的出货量均将大幅下滑。其中，高通出货量预计同比下滑9%，市场份额同比减少0.4%—24.7%；联发科出货量预计同比下滑8%，市场份额同比减少0.4%—34%。

而在中低端市场，紫光展锐受到影响则为最大。Counterpoint Research预计，其2026年出货量同比下滑14%，市场份额同比减少0.9个百分点至11.2%。

紧随其后的则是手机厂商。由于涨价将降低消费者购买手机的欲望，手机厂商的业绩也受到冲击。就在1月末，深圳传音控股股份有限公司（688036.SH）公布的2025年业绩预告显示，公司2025年净利润预计将大幅下降54.11%左右。

在公告中，传音官方坦承，此次整体盈利表现受到拖累的原因，为供应链成本上升影响，存储及其他元器件价格上涨较多，对产品成本和毛利率造成一定影响，导致报告期内公司整体毛利率出现下滑态势。

此外，已在本月传出被迫同意与日本闪存大厂铠侠续签新约的苹果，也成为涨价潮或将波及的对象。根据新合约，铠侠未来有权在此后每个季度根据市场价重新议价。据Counterpoint Research预计，苹果2026年出货量或将同比下滑6%，市场份额同比增加0.2%—18.3%。

对此，在最新研报中，高盛指出，存储价格的持续上涨正在明显挤压中低端智能手机利润空间。对售价在千元区间的机型而言，若单机物料成本增加数百元，将难以通过终端提价完全消化，厂商可能被迫优化产品结构，降低低利润机型的占比。

罗国昭认为，半导体市场始终是周期市场，随着厂商资本支出到位，产能紧张有望逐步改善，但考虑到产线扩建的长期性与AI对高端内存的持续消耗，价格回落将是缓慢渐进的。而且，高端和数据中心导向的内存价格回落速度会慢于传统消费级产品。因此，2026年市场预计仍以走高为主，此后才可能出现价格压力缓和，但对完全恢复到此前低位并不乐观。