付费名额难抢 免费版有卡顿 专业版不接智能体

AI大模型之困

民生调查

重庆晨报民生在线 扫码关注

难事、烦事、委屈事、不平事、新鲜事告诉我们，记者帮你办

今年，算力却悄悄成了年轻人消费的新时尚。只不过，有钱不一定买得到——智谱AI的Coding Plan套餐开售即秒空，阿里云百炼Pro版每天早上9:30补货，几分钟就没了，抢购难度堪比春运抢票。

一边是需求火爆，年轻人、中老年人都在往AI上靠，恨不得人手一个智能助手；另一边是名额紧俏、一“票”难求。这种“排队花钱”的奇观，到底咋回事？记者连续调查几天发现，AI大模型的快速普及，加上国产大模型在国际市场上竞争力不断增强，让算力变成了紧俏货。厂商这边推会员，用户那边抢名额——抢不到的干瞪眼，抢到的也不省心：免费版卡得想摔手机，花了钱的专业版接个智能体还动不动被封号。

A 付费名额难抢
有钱都花不出去

6月17日，智谱AI官宣上线并开源新一代旗舰大模型GLM-5.2。Coding Plan分三档：Lite版49元/月、Pro版149元/月、Max版469元/月；通用对话智谱清言VIP 79元/月、SVIP 229元/月。结果呢，名额根本抢不到。通义千问也一样，每天定时放货秒光。阿里云百炼Coding Plan更夸张——每天早上9:30补货，Pro版200元一个月，几分钟就没了。通道开着，钱就是花不出去——买个AI会员比抢春运票还难，真是活久见。

这事儿在技术圈里炸了锅。有人说，买会员买出了彩票开奖的感觉。

6月15日，两江新区一家互联网公司的算法工程师冯宇，为了给团队搞到智谱Coding Plan的Max档，连着守了五个早上。每天9:28准时打开阿里云百炼页面，手指头就搁在“立即购买”按钮上，眼睛都不敢眨。结果9:30一到，库存直接从“可购买”跳成“已售罄”，他连付款页面长啥样都没瞧见。冯宇自嘲道：“守着重庆的机房，抢不到北京发来的API名额，比抢洪崖洞的火锅位还难。”——这可不是个例，2026年的AI大模型订阅市场，活生生整出了一场“抢购大战”：国产大模型的付费套餐，开售即秒空。

智谱GLM Coding Plan从上线那天起就僧多粥少。GLM-5.1口碑爆了之后，“国产最强编程模型”的名号一传开，大批开发者蜂拥而至，套餐一放出来就没了。Lite、Pro、Max三档，哪一档都得靠抢，而且一票难求已经是常态。

通义千问那边也好不到哪里去。阿里云百炼Coding Plan，最早Lite版20来块钱一个月，现在Pro版200块一个月，每天早上9:30补货，一秒就没了。技术员田晓西说：“每天就放那么点量，手慢的根本抢不到。”有网友开玩笑：“抢AI会员比抢春运火车票还离谱。”

B 免费版有卡顿
不花钱就得“慢慢熬”

6月16日，重庆大学城某高校研究生林霖（化名），最近正赶毕业论文。他用通义千问免费版上传了一份120页的PDF文献，想让AI帮他总结一下核心观点。结果呢，等了快6分钟才收到回复，而且回答还缺胳膊少腿——免费版的上下文窗口被截断了，后半部分的实验数据全丢了。林霖哭笑不得：“我在图书馆从早上坐到下午，AI的回答还没出来，我自己都把论文啃完了。这哪是AI助手啊，简直就是个慢速复读机。”——要是说付费是“抢不到”，那免费就是“用不爽”。

通义千问官方虽然嘴上说C端基础功能“继续永久免费，没有每日对话次数限制”，但后面还跟了一句“高峰时段可能限流”。可实际用起来，这个“可能”基本等于“一定”。有评测数据显示，通义千问免费版在工作日白天限流特别明显：响应慢得要命，延迟4.7秒以上；处理20页PDF比付费版多花两倍多时间；上下文窗口被砍到32K，“连一份完整的招股说明书都读不完，关键条款动不动就丢了”。金融类问答准确率只有64.3%，比千问Max模型低了快30个百分点。

Kimi那边呢？免费用户每个月只能深度研究1次、做3次OK Computer、3次PPT。有用户吐槽：“免费版就是个试吃装——尝个味道还行，想吃饱？做梦。”

C 专业版不接智能体
花了钱也白花

渝中区化龙桥数字文创产业园的技术负责人吴波，多方蹲守抢购才拿下售价469元/月的智谱GLM Coding Plan Max顶配套餐，自行免费部署安装开源智能体OpenClaw（业内俗称“龙虾”），希望依托这套组合实现开发、测试流程全自动运行。

Max作为该产品线最高档位，官方配置更大调用额度与更高并发权限，本就适配自动化智能体场景。但投入使用仅两天，便频繁收到429调用超限提醒，本质是智能体自动批量发起请求，短时间并发量突破该账号速率阈值；使用至第四天账号进一步被平台限制，封禁理由为检测到异常高并发访问请求。吴波颇为无奈，花费顶配会员费用，却无法顺畅运行智能体任务，付费性价比偏低。

需要说明的是，OpenClaw本体开源免费，使用成本仅来自对接大模型产生的API调用扣费；且此类第三方智能体并非智谱官方适配应用，自行高频调用容易触碰API使用协议，触发平台风控策略，并非付费Max会员本身存在功能缺失。

这样的案例还有很多。开发者刘军最初选购GLM Coding Plan Lite套餐，因高频开发调用很快触及平台滚动5小时调用额度上限，随后续费升级Pro包年套餐（总价约1430元）。按照官方规则，Pro版每5小时可调用400次Prompt、每周总额度2000次，额度远高于Lite档位。但他接入自动化任务持续高频调用，升级当日便再度触达5小时滚动限额。刘军由此产生直观感受，付费升级并未带来可用额度实质性提升，性价比大打折扣。

从最早的Manus，到爆火的OpenClaw（龙虾），再到Hermes（爱马仕），AI智能体一个接一个刷屏。名字换来换去，核心就一个——让AI直接接管电脑，自己跑通复杂任务。可等用户真花了钱、接入智能体，等来的却是429限流和封号警告。有开发者一句话点破：“专业版要是不能顺畅接入智能体，那跟免费版有啥区别？”

延伸

算力需求爆发式增长 一票难求有望破解

回到开头那个问题：一边是全民AI需求井喷，一边是算力一票难求——这种供不应求的尴尬，到底是短期现象还是长期困局？

答案显然是前者。中国互联网协会专家咨询委员会常务副主任邵广禄给出了一个让人提气的预测：算力市场在未来几年将保持40%以上的高增长。

不只是规模在扩张，国产AI大模型的性价比优势也在持续放大。国产头部AI模型在Token消耗量上已稳居全球第一梯队，凭借高性价比使得增长速率更加陡峭。根据OpenRouter数据，在2026年6月第一周全球大模型调用量排名前五的模型中，中国占据四席，合计贡献了前五名总调用量的86.47%。另有数据显示，主流国产模型API调用价格较国际同类产品低60%至80%。以DeepSeek为例，其V4-Pro输入价格降至0.025元/百万Token，输出价格不足GPT-5.5的3%。

产业质感也在升级。AI算力行业人士判断，未来2~3年，我国有望在大规模AI算力集群商用落地领域实现快速追赶并完成局部反超。

说到底，今天这些“抢不到、花钱也憋屈”的现象，是一个新兴行业青春期必经的成长烦恼。等十万卡集群亮起来、算力网连起来、国产芯片硬起来，AI普及才算是真正落地了。到那时候，算力就像水和电一样，拧开龙头就有——这个未来，不遥远。

上游财经-重庆晨报记者 郑三波