又一个'DeepSeek时刻'？分清真突破与叙事通胀 · 全文合订

目录 · CONTENTS

01 三周，四款开源大模型：中国AI的七月到底发生了什么

02 中国AI霸榜全球前十？先分清你说的是哪个榜

03 彭博写下"fear"那天，英伟达却跌得比上次轻

04 万亿智谱：登顶榜单是技术账，万亿市值是资本账

05 戴着镣铐跳舞：中国为什么集体押注开源

三周，四款开源大模型：中国AI的七月到底发生了什么

四款开源模型密集登场，能力真在逼近——但“逼近”两个字要较真

7月17日，周五。

彭博打出标题：「又一个DeepSeek时刻」。

费城半导体指数盘中一度跌约5.7%。英伟达盘中跌近5%，午后收复大半失地，收盘跌1.97%。

市场已经没有复刻2025年1月27日的恐慌。那一天，英伟达收跌约17%，单日蒸发约5890亿美元市值。

但市场听懂了这次警报。

中国模型带来的压力，已经从“能不能用”，逼近“能不能和最强模型摆在同一张榜上比较”。

Motley Fool事后评价：「最初的反应似乎过度了。」

价格反应可以过度，能力变化却不能轻轻带过。

“三周四款”，先把日历校准

七月的密集发布，从腾讯混元Hy3开始。

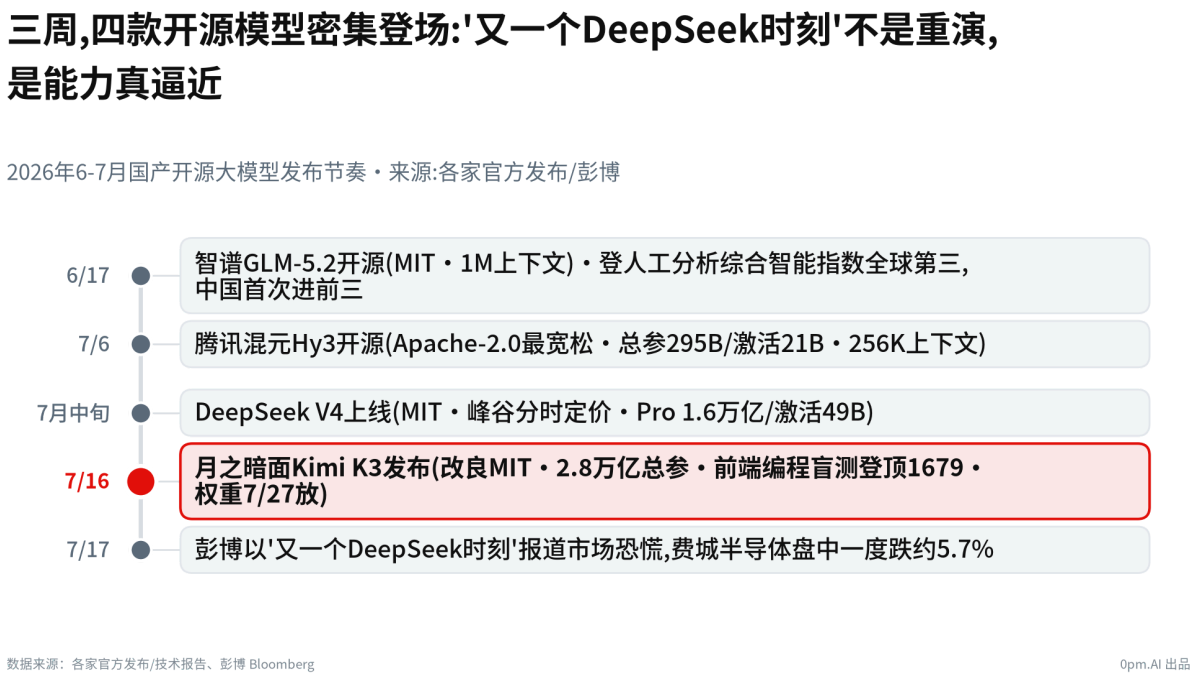
7月6日，混元Hy3开放权重；DeepSeek V4采用官方口径，只能写“**7月中旬**”，不能把自媒体流传的7月15日当成确定发布日期。

7月16日，月之暗面在上海世界人工智能大会发布Kimi K3。

GLM-5.2则更早一步，已于**6月17日**开源。

严格按发布日期计算，四款模型并非全部在七月三周内首发。“三周”更准确地指向发布、评测和市场关注集中爆发的窗口。

这处日历误差不影响真正的变化：不同公司的模型，开始在同一阶段冲击前沿能力榜。



三周窗口内，四款中国模型接连进入发布与评测焦点

这张时间线最值得看的，是密度。

单个模型冲上来，可能是一次技术突进；不同团队连续交卷，说明供给开始成群出现。

四款模型从不同发布台推向同一张能力榜，零散突进开始形成集群

军团的成色，要看你能不能真正拿走

参数规模最抓眼球。

Kimi K3的主流口径为**2.8万亿总参数**，另有**2.7万亿**说法，激活参数尚未公布；混元Hy3为**2950亿总参数、每个token激活210亿参数**。

数字大，不自动等于模型强。

真正决定开源价值的，还有上下文、权重状态、许可证和使用成本。

模型	发布或开源节点	参数与上下文	开源协议与状态	已有成本口径
混元Hy3	7月6日	2950亿总参数、210亿激活参数；256K上下文	Apache-2.0，可免费商用	—
DeepSeek V4	官方称7月中旬	—	MIT	—
GLM-5.2	6月17日已开源	官方未披露参数；第三方估计为7440亿总参数、400亿激活参数	MIT	—

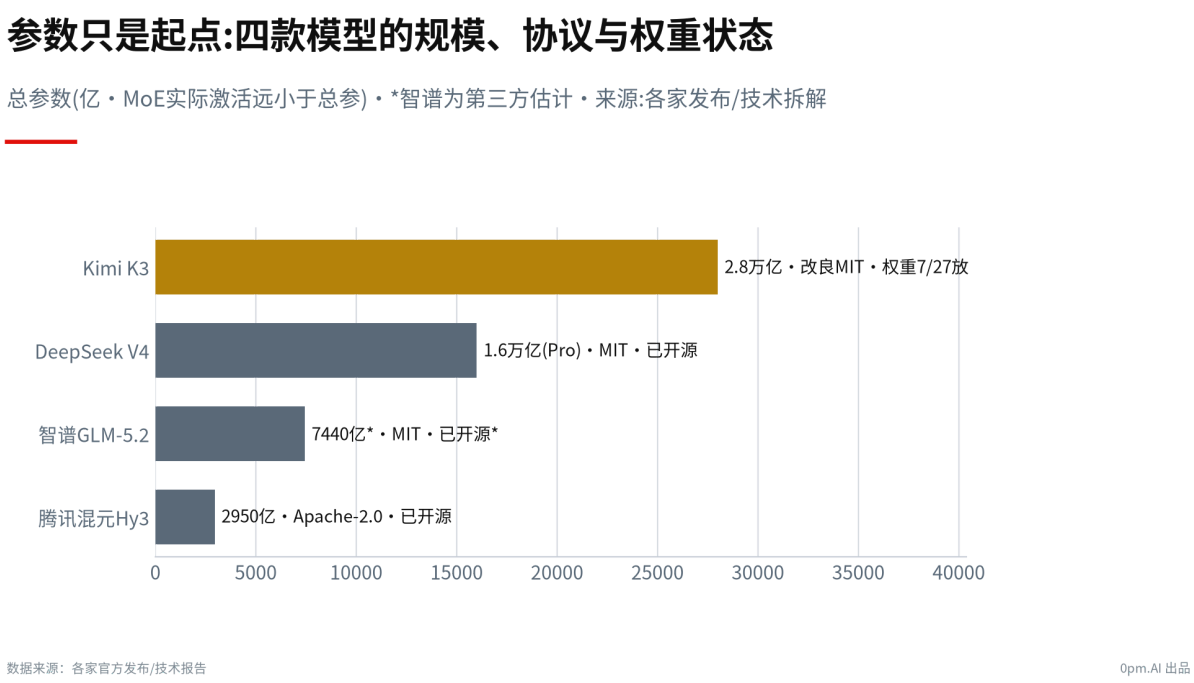
Kimi K3	7月16日发布；权重计划7月27日放出	主流口径2.8万亿总参数，存在2.7万亿口径；100万token上下文	Modified MIT，带MAU商用条款	每百万输入token 3美元、输出token 15美元

表中的空白不拿猜测补齐。

混元Hy3采用Apache-2.0，在这轮对照中商用边界最宽松；GLM-5.2与DeepSeek采用MIT；Kimi K3虽然走开放权重路线，Modified MIT仍带有MAU商用条款。

Kimi K3还要多加一条时间注脚：截至7月16日发布时，完整权重尚未实际放出，Artificial Analysis仍将其标为Proprietary。它计划到7月27日才完成开放权重这一步。

开源的价值不在发布会，而在权重能否拿走、许可是否真能用、部署成本是否承受得起。



参数只是起点，协议和权重状态决定模型能否真正进入开发者手中

这张表把“开源军团”的成色拉开了：四款模型方向相同，开放程度并不相同。

1679分，把“逼近”写进了榜单

Kimi K3最硬的一仗，发生在前端编程盲测。

Frontend Code Arena中，Kimi K3拿到**1679分**，超过Claude Fable 5的**1631分**，排名第一。

七个前端领域里，它拿下六个第一，仅游戏领域排名第二。相比前代K2.6的第18名，这不是小步挪动。

它在一个明确、可重复比较的能力切面上，越过了领先闭源模型。

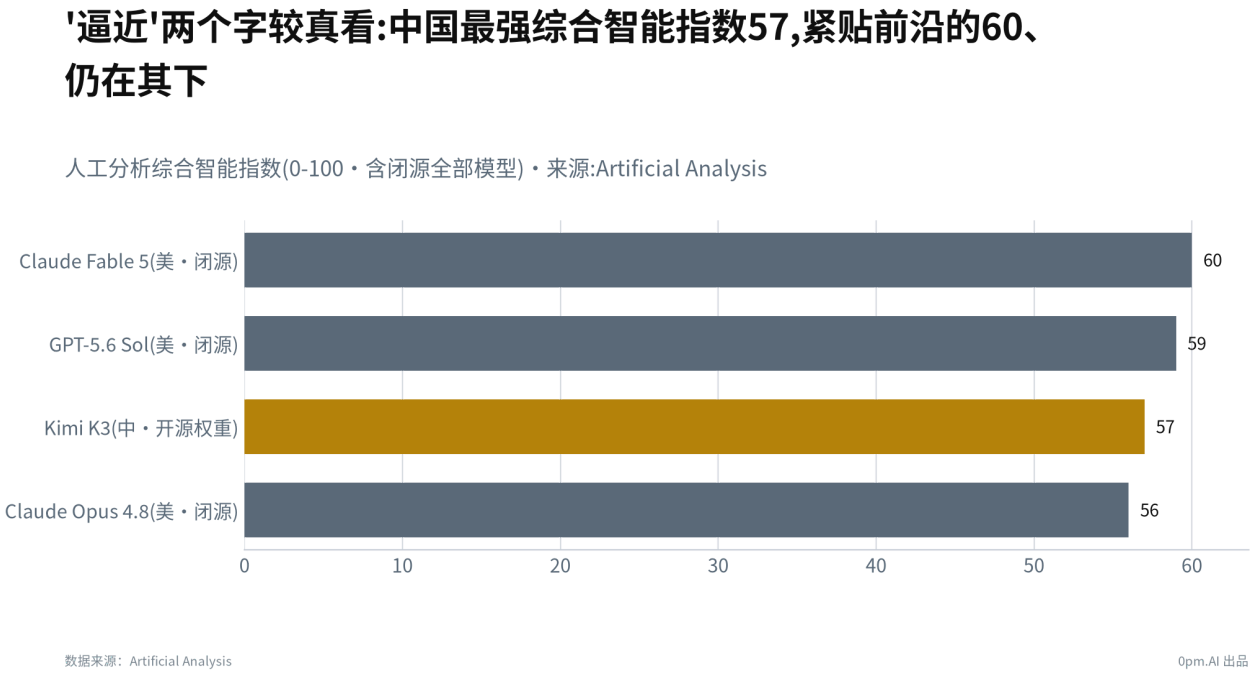
GLM-5.2提供了另一块证据。

富瑞转述的Artificial Analysis榜单截面中，GLM-5.2综合智能指数排到全球第三，编程能力全球第四，智能体能力全球第二。富瑞把它称为「中国AI里程碑」。

过去谈中国开源，常见优势是便宜、开放、部署方便。

七月出现的新东西，是能力坐标。

中国模型第一次密集站进前沿比较区，评价标准也从“国产模型里表现如何”，切换成“与全球头部模型还差多少”。



中国模型已能在前端编程登顶、综合能力逼近，但不同榜单的边界依然清晰

这张图要看的不是一个孤立冠军，而是两种逼近同时发生：细项开始穿透，综合榜开始贴近。

榜首要认，边界也得认

“霸榜前十”的故事早已流传过一次。

那是**2025年8月**的旧事，依据Hugging Face热度榜，衡量的是开源模型的热度与下载表现。它不能移到2026年7月，更不能拿来证明综合能力已经全面领先。

今年七月的证据更硬，也更窄。

Kimi K3在前端Code Arena以1679分登顶，证明的是前端编程能力；它在Artificial Analysis全部模型综合榜中约为**57分、排名第四**。

LMArena文本盲测里，Kimi K3为**1486±11分、排名第九**。

在这两张包含闭源模型的总榜前15名中，中国模型都只有Kimi K3一款。

GLM-5.2的“全球可用模型第一”，同样要保留“可用模型”这个限定。闭源的Claude Fable 5仍排在它前面。

榜单可以证明逼近，不能替任何模型颁发“全面第一”。

可也别因此把突破说轻了。

榜首只有一个，逼近却是一种产业状态：从无法比较，走到每个细项都必须较真。

这个七月真正值得认账的，并非中国模型把海外闭源模型挤出了总榜。中国开源已经从“便宜替代”，进入前沿模型的能力参照系。

细项能赢，综合榜能贴近，多个团队能连续交卷。

“全面追平”仍然没有证据。

继续把中国开源理解成低配版本，同样站不住了。

当开源模型已经能在细项登顶、综合榜逼近前沿，你还该用“开源等于低配”理解下一轮竞争吗?

来源：彭博、雅虎财经、Motley Fool、Artificial Analysis、LMArena、Frontend Code Arena、富瑞研报及新浪财经转述、量子位、Tom's Hardware、Decrypt、MarkTechPost、DataLearnerAI、Simon Willison。

风险提示：本文不构成任何投资建议，历史表现不代表未来，市场有风险，投资需谨慎。

文/0pm.AI

中国AI霸榜全球前十？先分清你说的是哪个榜

热度榜、能力榜、开源专榜、含闭源总榜——四种口径，四个故事

前十席，中国模型拿走**9席**。

美国只剩Boson AI一款。

其中**7款**中国开源模型集中发布，英国《自然》将其称为“又一个DeepSeek时刻”。

这组数字是真的。

日期却是**2025年8月**。

把它贴到2026年7月，中国AI能力突破的新闻下面，一条旧热榜就被包装成了新总榜。

“霸榜”没有造假。

被压扁的是时间、参赛名单和评分规则。

九席是真的，新闻已经旧了

2025年8月，智谱GLM-4.5系列、月之暗面Kimi K2、阿里Qwen3系列与Wan-2.2等模型进入抱抱脸开源社群前十。加上已经在榜的腾讯混元与阿里模型，中国模型合计占据**9席**。

这是一张开源专榜。

它主要反映下载、点赞与趋势热度，衡量开发者是否愿意接触、试用和传播一款模型。

到了2026年7月，新的锚点换成了能力突破。

7月16日，Kimi K3在Frontend Code Arena取得**1679分**，超过Claude Fable 5的**1631分**。**7月17日**，彭博又用“又一个DeepSeek时刻”描述市场对中国模型进展的反应。

旧热榜与新能力榜，被一句“7款模型霸榜全球前十”缝在了一起。

“七款霸榜”属于2025年8月，2026年7月的新锚点是能力突破，两个时点不能拼成同一条新闻

两个时间点一旦重叠，去年“谁最热”就会被误读成今年“谁最强”。

“全球第一”至少用了三把尺子

榜单先要回答两件事。

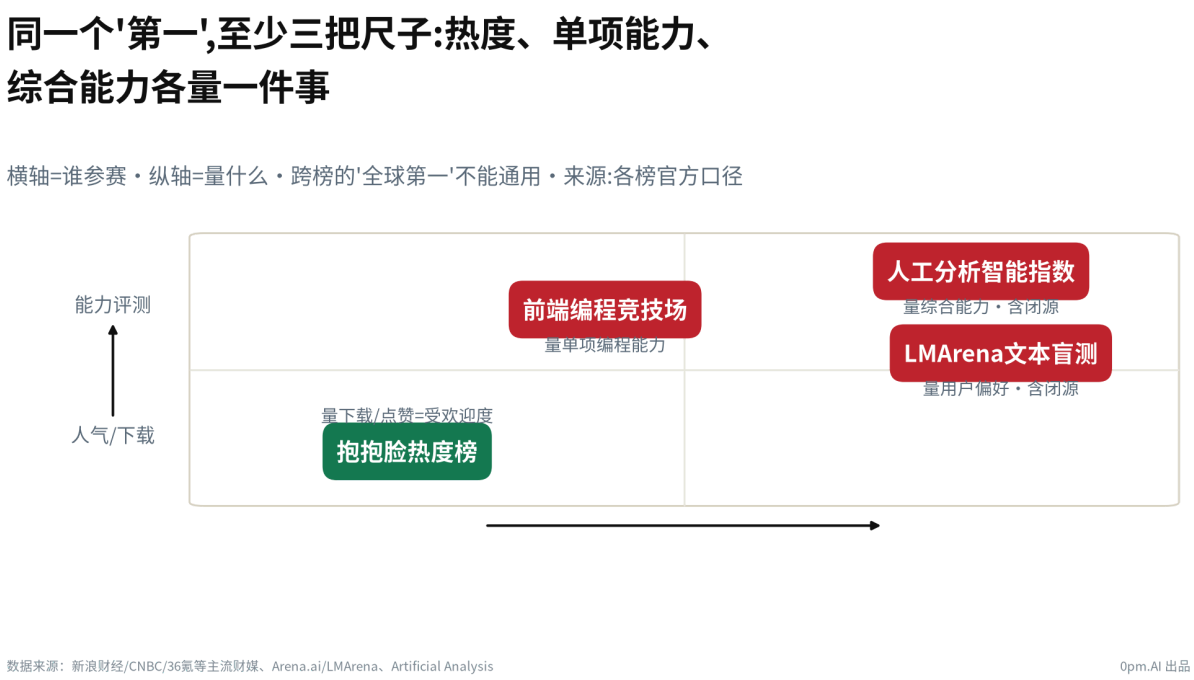
量什么?

让谁参赛?

更准确地说，热度榜、能力榜、开源榜、含闭源总榜并非四条平行跑道。它们是两条轴：一条决定测量对象，一条决定参赛范围。

传播把这个二维坐标压成了“全球榜”三个字。

榜单或平台	主要测量对象	参赛范围	第一名能够证明什么
抱抱脸开源社群榜	下载、点赞、趋势热度	开源模型	开发者关注度与生态渗透
Frontend Code Arena	前端编程盲测表现	参评模型	特定编程任务能力
Artificial Analysis	综合智能指数	包含闭源模型	多项基准汇总后的综合位置
LMArena文本榜	用户盲测偏好Elo	包含闭源模型	用户对回答质量的相对偏好



热度、单项能力与综合能力各量一件事，同一个“第一”不能跨榜通用

热度第一说明有人用。

编程第一说明一项能力强。

综合榜靠前，才说明模型在更宽的任务面上逼近前沿。

三者都值钱，价格完全不同。

同一款模型站在热度、单项和综合三把尺子前，刻度不同，名次自然不同

中国模型在不同尺子下都有真成绩。

OpenRouter官方博客与具名分析显示，开放权重模型能力前五中，中国模型占**4席**：GLM-5、Qwen3.5、Kimi K2.5和DeepSeek V4。美国模型仅NVIDIA Nemotron 3 Ultra进入这一组。

这是能力格局。

2025年那张前十占九席的抱抱脸榜，讲的是生态热度。

把两组数字放在一起，可以确认中国开源模型既获得了使用，也进入了能力前列。把它们说成同一张榜，认知增量反而消失了。

打开闭源模型，九席缩成一席

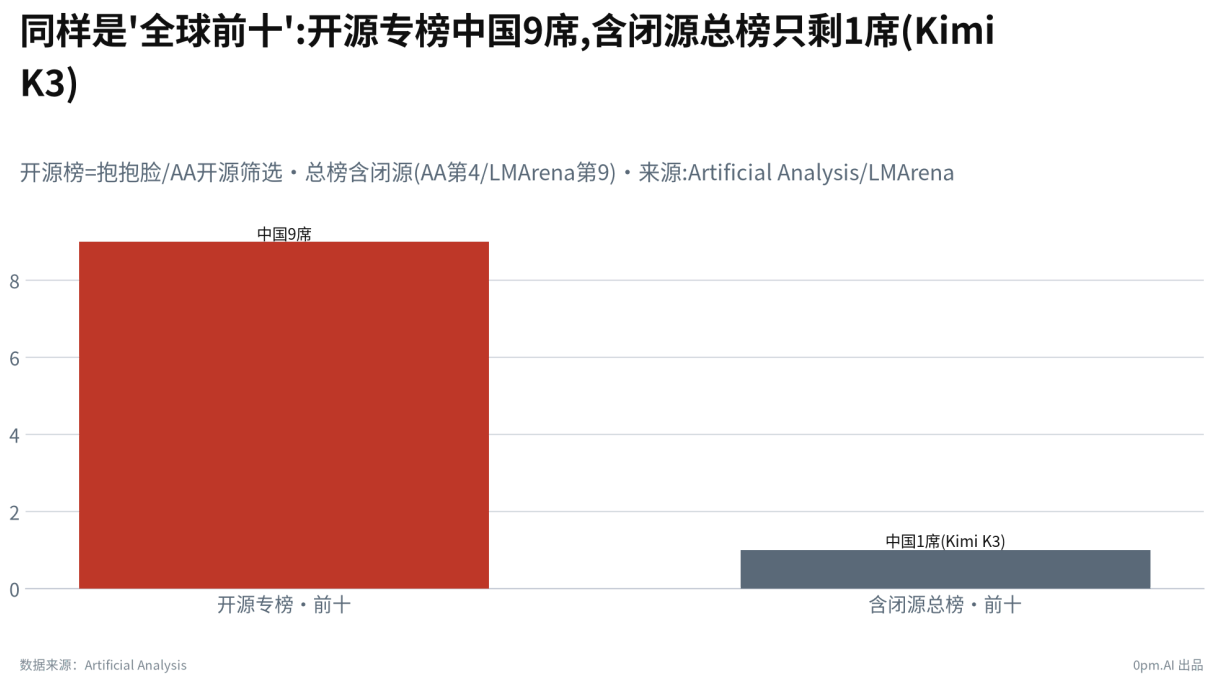
真正改变故事的，是参赛名单。

只看开源权重模型，中国可以占据前五中的**4席**，一度包揽开源榜前六。

把Anthropic、OpenAI、Google等闭源模型放回来，名次立刻重排。

Artificial Analysis给定榜单截点中，Kimi K3位列综合智能指数**第4**，前三名均为美国闭源模型。前十五名里，中国模型只有Kimi K3一款。

LMArena文本盲测中，Claude Fable 5排名第一，Kimi K3位列第9；前十五名同样只有这一款中国模型。



开源专榜中国模型占据多数，含闭源总榜前十五只留一席，“霸榜”取决于谁被允许参赛

9席与1席都是真的。

前者证明中国已经成为全球开源生态的主导力量之一，后者提醒你，含闭源的综合前沿仍由美国模型占据多数席位。

榜单还是那块记分牌，参赛池换了。

还有一个容易漏掉的变量：榜单是快照。

富瑞在6月21日转述的榜单截点中，GLM-5.2位列Artificial Analysis综合智能指数全球第三，并称其为“中国AI里程碑”。Kimi K3发布后，新的给定截点又呈现出Kimi K3第四的格局。

不标日期的名次，保质期可能比一条热搜还短。

最后一层冷水，泼给跑分机制

即使日期、指标和参赛名单全部对齐，榜单也不能代替真实使用。

Kimi K3在Frontend Code Arena登顶，却在Artificial Analysis综合榜位列第四，在LMArena文本榜排第九。

同一款模型，单项第一、综合第四、用户偏好第九。

它们没有互相否定。

模型能力本来就不是一根刻度。

Kimi K3还有一个更尖锐的对照：AA-Omniscience基准中，其幻觉率由K2.6的**39%**升至**51%**。能力总分上升，并不保证每个可靠性指标同步改善。

可复现性也有时间差。

Kimi K3于**7月16日**发布，开放权重计划在**7月27日**放出。发布日只能通过API使用，Artificial Analysis因此将其标为Proprietary。许可采用Modified MIT，并附带MAU商用条款。

在完整权重放出前，“开放权重模型”的最终复现链条还没有闭合。

评测体系本身也会受污染。

人大高瓴团队相关研究所涉及的C-Eval基准题泄露争议，暴露了训练语料与测试题重叠的风险。论文《榜单幻觉》则用Meta案例讨论榜单失真；该案例并非针对国产厂商，却说明广泛使用的排行榜同样可能制造领先幻觉。

这些证据不能推出“中国模型靠刷分”。

它们只够推出一个更克制的结论：单次跑分越亮眼，越需要公开权重、独立复测与真实任务把它钉牢。

该打折的是叙事，不是进步

我的立场很明确。

“中国AI全球前十占九席”这句话，离开**2025年8月、抱抱脸、开源热度榜**三个限定词，就不再可靠。

“中国模型没有真突破”同样站不住。

开放权重能力前五占四席、Kimi K3登顶前端编程竞技场、进入含闭源综合榜第四，这些都是能力逼近前沿的硬证据。

中国AI最值钱的进展，并非在所有榜单上同时拿第一。

它已经能用不同模型，在开源生态、编程能力和综合智能三个层面反复进入全球前列。

爽文只需要一个“霸”字。

判断产业位置，需要先看尺子。

下一次再看到“中国AI全球第一”，你先问谁赢了，还是先问这张榜到底量了什么？

来源说明

来源：抱抱脸榜单同源报道（Yahoo财经中国香港、鉅亨网）；OpenRouter官方博客及Understanding AI具名分析；Arena AI旗下LMArena与Frontend Code Arena；Artificial Analysis官方排行榜；富瑞研报相关转述；Tom's Hardware；人大高瓴团队相关研究及36氪报道；论文《The Leaderboard Illusion》。

风险提示：本文不构成投资建议，历史表现不代表未来，市场有风险，投资需谨慎。

文/0pm.AI

彭博写下"fear"那天，英伟达却跌得比上次轻

同样是"DeepSeek时刻"，这次英伟达跌得比上次轻得多

2026年7月17日，星期五。彭博的头条里蹦出一个词——**"Another DeepSeek Moment"**，恐慌。费城半导体指数盘中一度跌近**5.7%**，从六月下旬的历史高点算下来，回落已经**逾20%**，一只脚踏进技术性熊市。从6月22日到这一天，全球半导体股的市值抹掉了约**3.3万亿美元**。

看上去，2025年那一幕又要重演了。

可收盘的数字，是另一个故事。英伟达当天盘中一度跌近5%，午后被一波资金硬生生拉了回来，收在**203.31美元**，只跌**1.97%**。标普500跌1.01%，纳斯达克跌1.40%。天塌下来的样子，最后没塌。

把这一天，和上一次那一天摆在一起看，才是本文要讲的事。

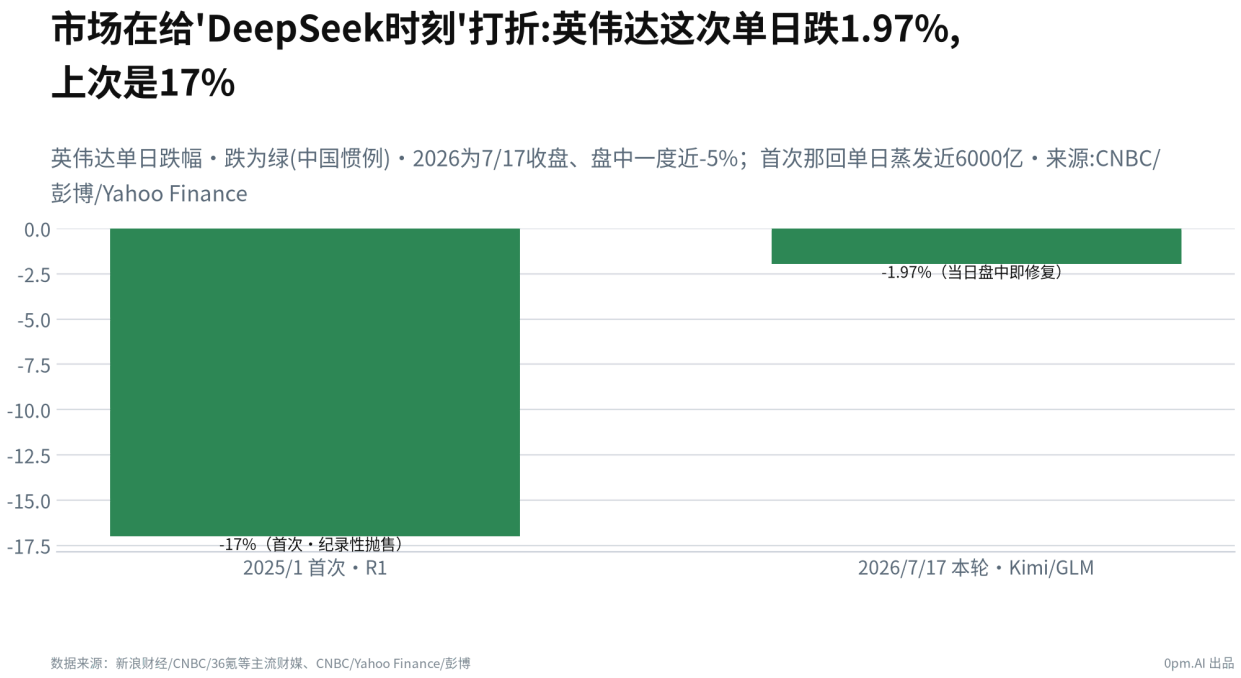
上次是"惊"，这次是"怕"，但怕得没那么深

2025年1月24日，DeepSeek发布R1。下一个交易日是1月27日周一，华尔街后来管它叫"DeepSeek Monday"。英伟达单日暴跌约**17%**，收在118.58美元，一天蒸发约**5890亿美元**——美股历史上单一公司单日最大市值损失。抛售蔓延到整个科技板块，全球科技股当天没了约**1万亿美元**。

引爆点是一个数字：R1号称训练成本只有约**560万美元**。它像一记耳光，打在"高性能模型必须靠昂贵高端芯片堆出来"这个信念上。

这一次呢？7月16到17日，月之暗面在上海世界人工智能大会上放出Kimi K3，2.8万亿参数，号称全球最大的开放权重模型；同期智谱系的GLM-5.2也来了。叙事是同一套：中国开源模型在追顶级闭源系统，而且更便宜。

但市场的反应，量级完全不同。一次是单日**-17%**、一天没掉近**6000亿美元**；一次是盘中吓一跳、收盘**-1.97%**，当天就修复。同样叫"DeepSeek时刻"，情绪从"惊"变成了"怕"，而怕，是可以被逢跌买入的资金按住的。



英伟达两次"DeepSeek冲击"对比：2025/1单日-17%、蒸发约5890亿美元 vs 2026/7/17盘中一度跌近5%、收仅-1.97%当日即修复

市场没有不信中国AI，它在学会给"时刻"打折

R1那次造成的是一场广泛的、看得见的重定价，因为它改写了全球对模型成本曲线和中国竞争力的信念。而在那之后，DeepSeek后续的V3、R1更新，哪怕是可信的阶跃式效率提升，市场也只当作"延续和巩固"，不再当成新的冲击波。

到2026年7月17日这波，最耐人寻味的动作不是下跌，是那个午后的回补。英伟达盘中跌近5%，然后被"买dip"的钱拉回来。回补这个动作本身，就是信号——有人在恐慌里看到的是折扣。Motley Fool的评价很直接：和DeepSeek那次一样，初始反应似乎**过度了**。

一个学会讨价还价的市场：第一次听到坏消息脸色发白，第二次听到同样的消息，先问一句"打几折"

这就是这次和上次真正的分界线。市场不是不信中国AI了，恰恰相反，它信得越来越平静。第一次是没有先例的震惊，任何人都不知道该怎么定价，于是先杀为敬；到第二次、第三次，它已经知道这个叙事长什么样、边界在哪里，于是给它打了个折。震惊会一次性重定价，脱敏则是一次比一次浅。

一个可以自己盯的指标：下一次"某某开源模型登顶"的新闻出来，英伟达当天收盘跌几个点。数字越小，折扣打得越狠。

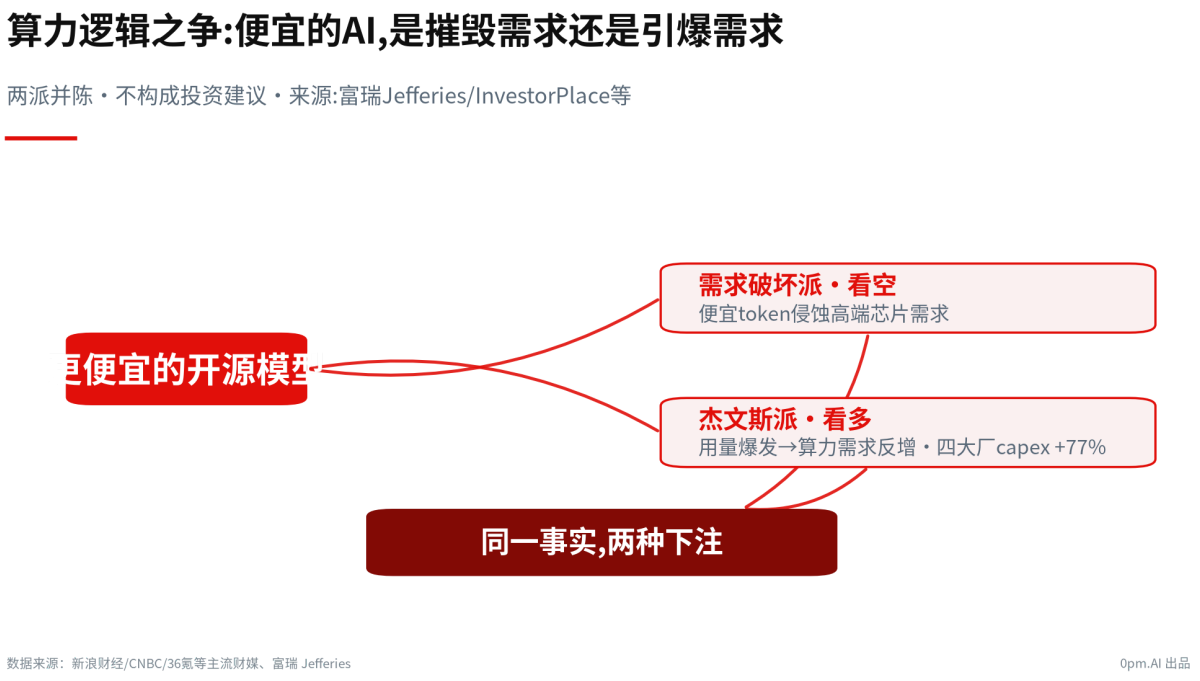
争议是真的：两把尺子，我不替你选

打折不代表没风险。分歧一直都在，而且这次很具体。

空头这边是"需求破坏"。逻辑很硬：越来越便宜的token加上本地推理，会啃掉英伟达高端芯片的需求，动摇超大规模厂商的资本开支叙事。台积电就是个现成的例子——七月创了纪录利润、把全年营收增速指引上调到略高于40%，可它同时把资本开支指引抬到**600到640亿美元**，被市场解读成"AI扩产极贵、在压毛利"，股价跌了约**4%**。Jefferies还点名，智谱的GLM-5.2正用更低的运营成本，逼近领先的企业级模型。

多头这边搬出的是杰文斯悖论。同样是Jefferies的研报：更便宜的AI模型"不太可能"减少AI投资，反而可能推高算力需求——单位推理成本降下来，总的算力消耗会被更大的用量顶上去。实证也摆在那儿：2025年1月R1引发英伟达单日蒸发近6000亿美元之后，没有一家超大规模厂商削减开支，几周内全部重申或上调了资本开支指引。

数字更狠。微软、谷歌、亚马逊、Meta四家，2026年合计资本开支约**7250亿美元**，较2025年约4100亿美元增长约**77%**，看不到一丝放缓。而每1美元AI基建折旧，如今对应约**1.19美元**的超大规模加neocloud收入——一年前这个数还不到1.0，这是收入端第一次反超资本开支曲线。



算力逻辑之争：需求破坏派看到便宜token侵蚀芯片需求，杰文斯派看到四大厂2026资本开支约7250亿美元/+77%零放缓

两把尺子，量的是同一件事，读出的却是相反的结论。我不打算替你选边——真正诚实的做法，是把两把尺子都递到你手里，告诉你各自量的是什么。

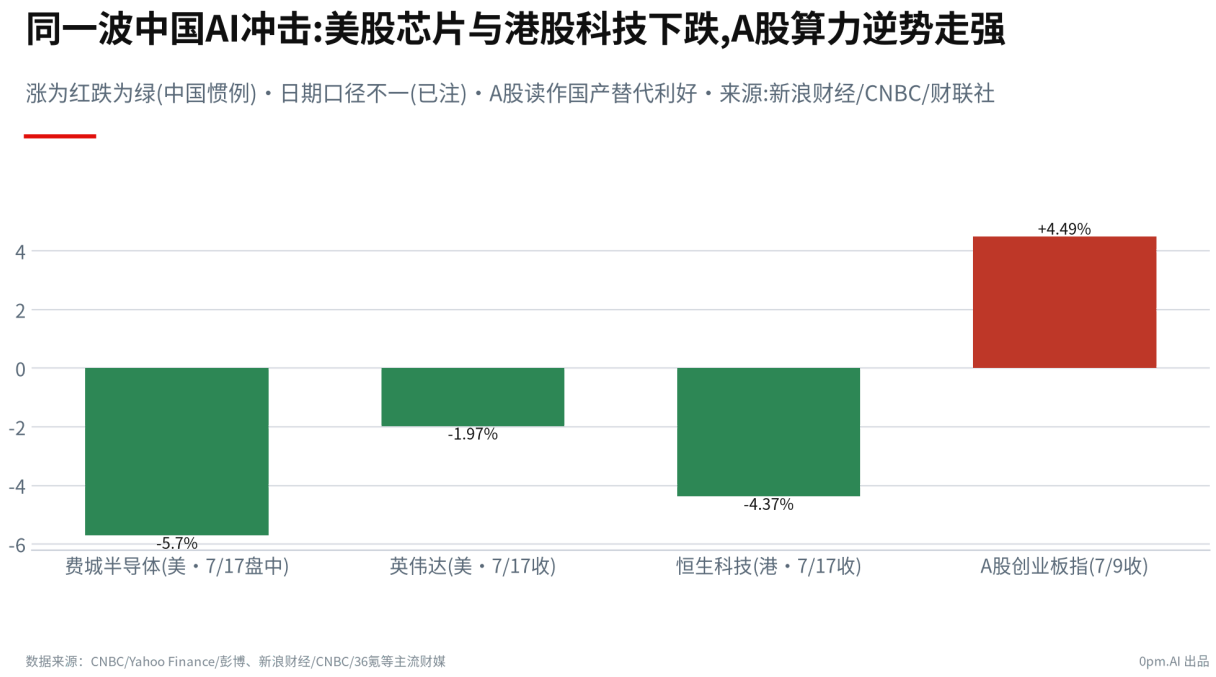
最危险的不是看空，是乱归因

比看多看空更值得警惕的，是把不相干的事叠在一起，凑成一个"中国赢、美国崩"的爽文。这一波里，至少三处口径不能混。

第一，智谱那根28.49%的大阴线，不是竞争恐慌。智谱02513在7月17日收1107港元，跌28.49%，前收还是1548港元。但主因是7月13日那笔配售——每股1588港元、发行约1978万股新H股——带来的抛压，叠加当天港股整体普跌：恒指跌1.78%报24562.24，恒生科技跌4.37%，科网、光通信、芯片股一片绿。当天它市值约5154亿港元，单日蒸发约2053亿港元。把它算成"DeepSeek时刻的恐慌证据"，是归因错误。

第二，英伟达七月的两次下跌是两件事。7月7到8日那次，是路透报道DeepSeek在自研推理芯片，消息后英伟达盘前一度跌约1.5%；7月16到17日这次，才是Kimi K3和GLM-5.2。两次触发点不同、口径不同，合并成"英伟达七月暴跌X%"就是账算错了。

第三，美股芯片崩，不等于中国股涨。同一波中国开源模型密集突破的消息，在大洋两岸被读成两个方向。港股科技因为全球联动跟着跌；A股却偏强——7月9日创业板指涨4.49%，半导体产业链强势爆发，被读作国产算力需求上行、国产替代加速的利好。同一条新闻，一边是利空，一边是利多。



同一事件两种定价：美股芯片股风险偏好走弱下挫，A股半导体/国产算力偏强，港股科技因全球联动同步跌

"又一个DeepSeek时刻"，最诚实的读法

它不是英伟达要完——收盘那根被拉回来的下影线，已经把话说清楚了。它也不是中国赢了——含闭源的总榜前十五名里，中国模型至今只有Kimi K3一款。

它真正的意思是：市场正在给这个叙事重新定价。而定价，正在打折。

下一次"DeepSeek时刻"来的时候，你该盯的不是标题里那个"fear"，是英伟达当天收盘的那个数。它比上一次轻多少，就是市场又把这个故事打了几折。

文/0pm.AI。本文所涉数据与事件均来自公开报道，来源包括彭博、CNBC、Yahoo Finance、Motley Fool、PIIE、路透、新浪财经、金融界、中信建投研报等。行情数据口径混合（盘中/收盘、美元/港元/人民币）已在文中分别标注。本文仅作行业与逻辑分析，不构成任何投资建议；历史表现不代表未来，市场有风险，投资需谨慎。

万亿智谱：登顶榜单是技术账，万亿市值是资本账

营收7.24亿、亏损47.18亿、资不抵债——开源军团的钱从哪来到哪去

同一家公司，六月递上来两张成绩单。

一张写着：GLM-5.2登上人工分析综合智能指数全球第三，中国模型第一次挤进前三，编程第四、智能体第二。富瑞的研报里管这叫"中国AI里程碑"。

另一张写着：2025年营收7.24亿元，亏损47.18亿元，净资产-61.51亿元——资不抵债。

一张是技术账，评委打分，实打实的奖杯。一张是资本账，市场定价，一路涨到一万亿港元。

这篇只想干一件事：把这两张账单摆在一起，看清中间隔着什么。

登顶榜单的技术奖杯与资不抵债的资产负债表，同属一家公司的两本账

资本账烧得比模型还热

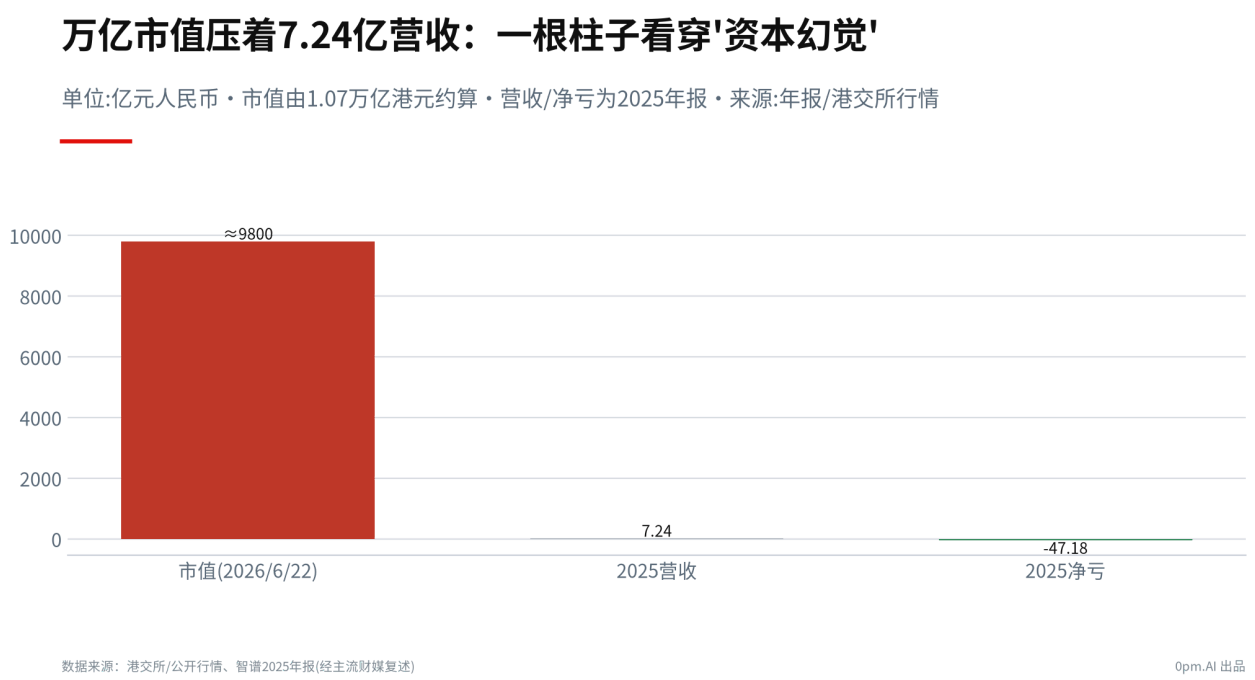
先看热的这一头。

北京智谱华章，港股代码02513，2026年1月8日在港交所主板挂牌，号称"AGI大模型第一股"。上市到六月，股价较发行价最高涨了约**24.6倍**——注意口径，这是股价对发行价的倍数，不是市值涨幅。市值一度冲破**1.07万亿港元**。

GLM-5.2发布当日，股价再跳涨10.3%，两日累计超35%。

一家去年营收只有7.24亿的公司，凭什么撑起万亿共识?

新浪财经一篇具名分析给出了答案，也给这万亿起了个名字——「资本幻觉」。因为这栋楼的地基，是仅**2.67%**的自由流通盘。



市值破万亿港元，会计营收仅7.24亿、亏损47.18亿——同框才看得清落差

九成七的筹码锁着不动，剩下不到三成之一的一成在市场里换手。少量买盘就能把价格顶上天，也能把它砸下来。

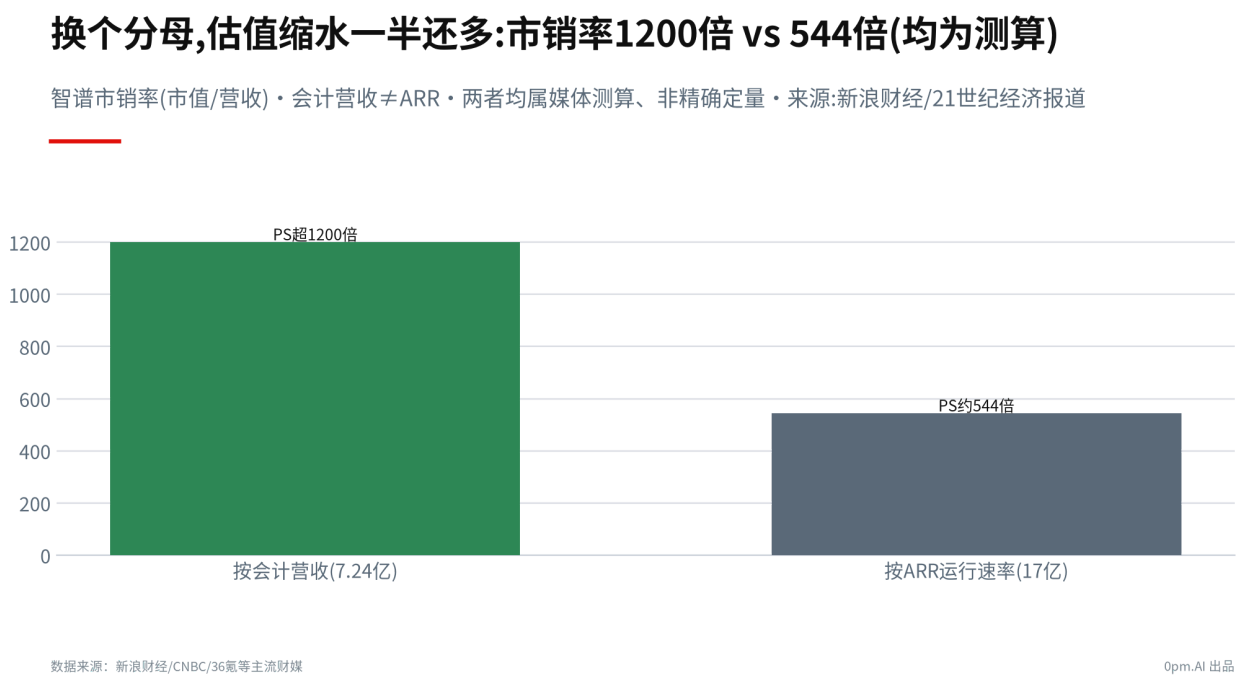
榜单上排第三是能力，市值一万亿是共识——共识建在2.67%的筹码上。

一个"ARR"，藏了一个数量级

资本账里最容易被含糊过去的，是营收这个词。

媒体常写"智谱ARR约17亿"，听起来离万亿市值没那么离谱。但ARR是年度经常性收入的运行速率，是把某个时点的月度订阅乘出来的口径，不是会计确认的收入。翻2025年报，会计营收就是**7.24亿元**，同比+131.9%，B端占到99%。

两个数一换，市销率差一个数量级。



同一家公司，按会计营收算PS超1200倍、按ARR算约544倍——分母换个口径，估值缩水一半还多

按会计营收7.24亿测算，市销率超**1200倍**；按ARR测算，约**544倍**。都是测算，但无论你站哪个口径，这都不是一个能用常规估值框架安放的数字。

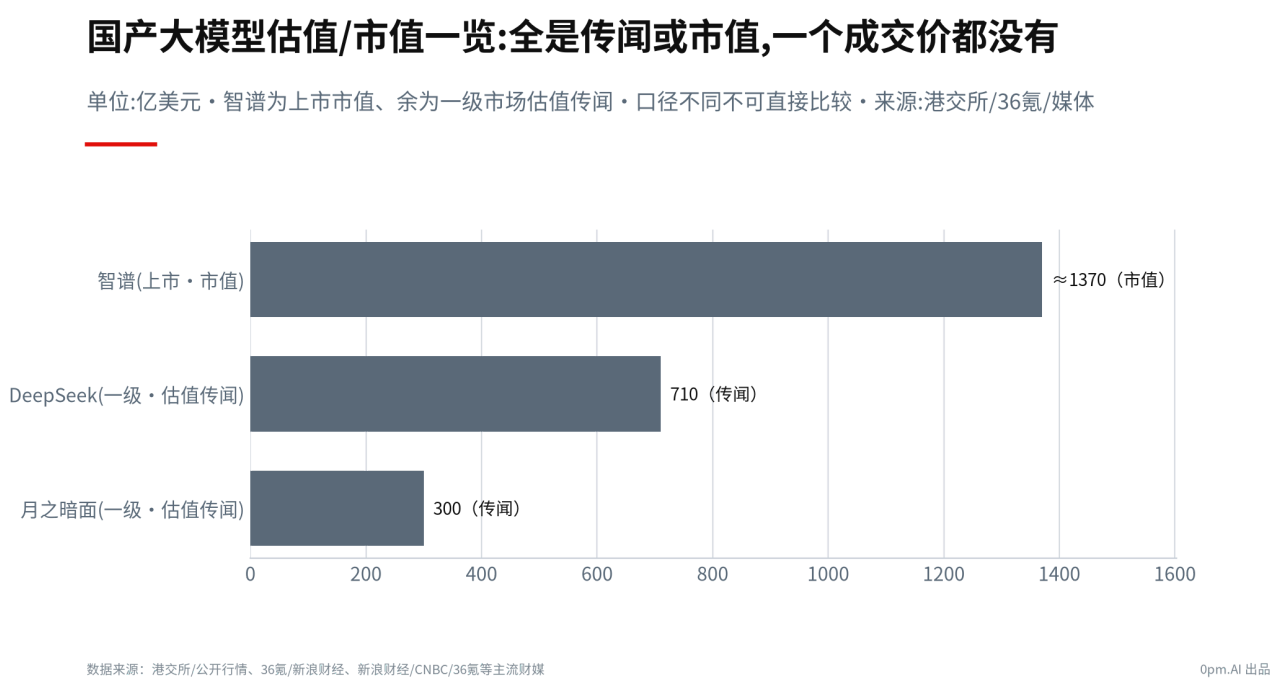
就算换上对智谱最友好的那把尺子——富瑞在同一份"里程碑"研报里也顺手补了一句：智谱的市值对ARR约**94倍**，而Anthropic约**18倍**。

写奖杯的那份研报，自己在结尾泼了盆冷水。

不只是智谱，整支军团都在烧传闻估值

把镜头拉开，登顶榜单的这批国产开源玩家，钱的故事都写在"传闻"两个字上。

DeepSeek，幻方量化的班底，第二轮融资投前估值传闻约**4800亿元人民币**，折合约710亿美元——同一笔、同一个币种换算，不是两轮叠加。月之暗面，Kimi K3的东家，六月新一轮投前估值传闻约**300亿美元**。腾讯混元不单独披露估值，藏在集团报表里。



各家估值/融资对照——数字全带"传闻/测算"，成交价一个都没有

这些数字有一个共同点：全是传闻或测算，没有一个是成交价。轮次、币种、投前投后，任何一处混用，量级就差出去了。资本账最热的地方，恰恰是最模糊的地方。

开源不是慈善，是最贵的获客

那问题绕回来：模型免费放权重出去，钱从哪进来?

路径写在明面上——开源引流，靠API调用、MaaS订阅、企业私有化部署收费。开源负责铺量，商业化负责变现。

只是"廉价开源"这套叙事，最近被自己人戳破了。Kimi K3发布时，定价定在每百万输入3美元、每百万输出15美元，和Claude Sonnet 5一个价位，较上一代提价约四倍。

免费权重不等于免费服务。开源是把用户先圈进来的钩子，能不能从圈人走到赚钱，是另一本没有时间表的账。

开源不是慈善，是最贵的获客。它替你省下市场费，却没替你解决那47亿的亏损从哪补上。

解禁那天没塌，塌在配售那天

七月，资本账已经开始投票了，而且是往下投的。

7月8日基石解禁日，市场原本捏着一把汗——参照商汤，当年解禁日曾单日暴跌**46.77%**。结果智谱不跌反涨，收1825港元、+13.35%，次日再涨11.34%，两天累计约24.7%。原因是近七成基石投资者表态长期持有，抛压有限。

看起来躲过了商汤那一劫。

可7月13日一纸配售公告——每股1588港元、发行约1978万股新H股——不到一周，7月17日收1107港元、单日跌**28.49%**，市值约5154亿港元，单日蒸发约2053亿港元，配售股东账面浮亏近30%。

这里必须掰清口径：这一跌不是"又一个DeepSeek时刻"的竞争恐慌，而是配售抛压叠加港股大盘普跌（当日恒指−1.78%、恒生科技−4.37%）。技术账那边，GLM−5.2六月那份榜单上的全球第三，成绩还在。

解禁那天没塌，塌在配售那天。两本账的裂缝，就是从这种地方漏风的。

技术能不能一路赢下去，是研发的事。可万亿市值靠的是2.67%的浮筹、一堆传闻估值、一个差着数量级的营收口径，和一张资不抵债的报表。当技术账继续加分、资本账却已经先扣了两千亿的时候，你更愿意相信哪一本会追上哪一本？

风险提示：本文所涉估值倍数、融资额、参数规模等多为传闻或第三方测算，已尽量标注口径，不代表公司披露。全文仅作行业与公司层面的事实梳理与逻辑分析，不构成任何投资建议。高估值标的与新上市股票波动巨大，历史表现不代表未来，市场有风险，投资需谨慎。

文/0pm.AI

来源说明：本文数据与事实来自新浪财经、证券时报网、21世纪经济报道、36氪、量子位，以及富瑞（Jefferies）研报、智谱2025年报（经上述财媒复述）、Artificial Analysis排行榜等公开信息。

戴着镣铐跳舞：中国为什么集体押注开源

开源不是理想主义，是被算力管制逼出的最优解——是护城河也是天花板

7月6日，腾讯把混元Hy3的权重挂上网，Apache−2.0协议，最宽松的那一档，随便商用。7月16日，月之暗面在上海世界人工智能大会上放出Kimi K3，**2.8万亿参数**（另有2.7万亿口径），号称

将成为全球最大的公开可下载模型——完整权重要到7月27日才放出，发布当天仍只是API可用，改良MIT协议。更早在6月17日，智谱已把GLM-5.2开源，采用MIT。

再往前数，DeepSeek——MIT。

把这个月的中国头部AI摆成一排，你会看到一件反常的事：混元、DeepSeek、GLM、Kimi，几乎清一色开源。而大洋对面，OpenAI和Anthropic把权重锁在保险柜里，一行都不给。

按常理，落后的一方应该攥紧手里每一张牌。为什么反过来了？为什么追赶的这一方，反而把家底摊在桌上？

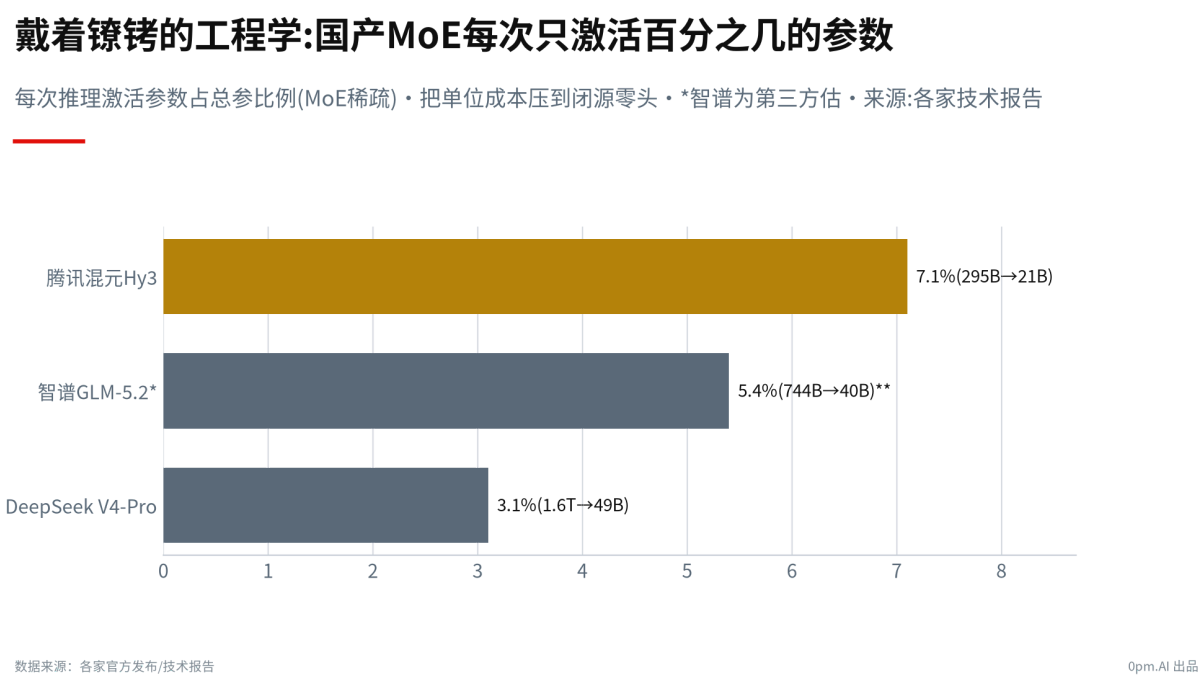
我的判断很直接：这不是理想主义发作，是被算力管制逼出来的一条活路。这条活路既是护城河，也是天花板。

开源在这里不是价值观，是唯一还能打的牌

先看成本这一面，因为镣铐的印子就刻在这里。

拿不到最先进制程、拿不到最高端的算力，中国团队只能在架构和工程上抠。混元Hy3是**2950亿参数**的稀疏模型，可每次前向只点亮约**210亿**；官方给的部署建议是8卡GPU（H20-3e那一级）就能跑，还配了三档推理力度，能省则省。这不是炫技，是没得选——每一分算力都得掰成两半花。

抠出来的结果，是价格。Kimi K3每完成一个任务的成本压到约**0.94美元**，输出token比上一代还少了两成。便宜到什么程度？在OpenRouter这类开发者平台上，最常用的前十个模型里，有六个是中国货；美国公司自己调用中国模型的token占比，最近一周超过**30%**，峰值逼近**46%**。



开源与闭源两条路线在协议、成本、算力依赖上的分岔——中国靠稀疏架构把单位成本压到闭源的零头，代价是放弃权重的独占收益

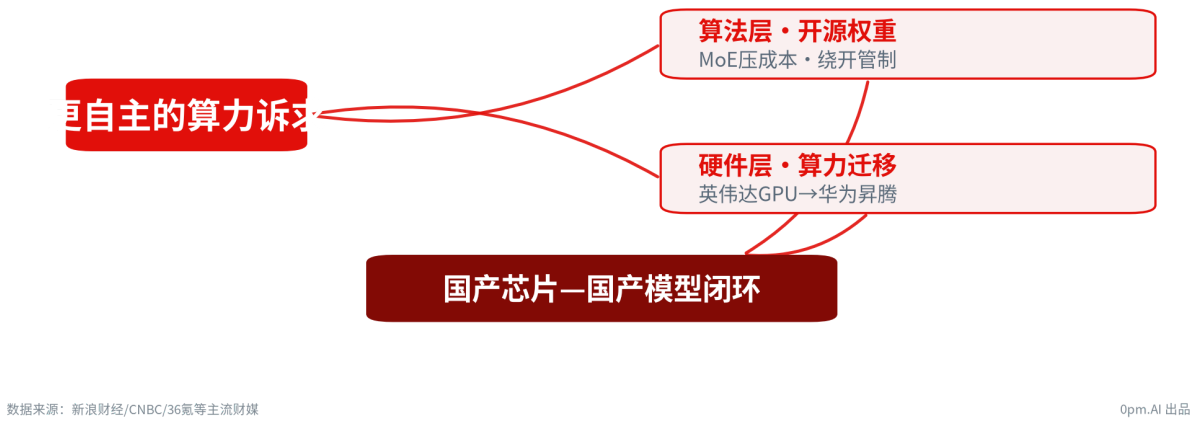
这里藏着第一个逻辑差：便宜从来不是慈善。它是被管制训练出来的工程本能。名义上叫"开源普惠"，翻译过来是——闭源那张高价牌，中国团队根本没有筹码去打，只能把开源、把低价，做成自己的武器。

把权重送出去，是为了换回一整个生态

成本是守，开源本身是攻。

闭源模型是一门生意，开源是一场圈地。你把权重免费放出去，全世界的开发者替你调试、替你适配、替你写教程，用着用着，标准就成了你的。这套打法绕开了管制最要命的地方：算力可以卡，但拦不住代码在GitHub上流动。

更狠的一步藏在算力迁移里。GLM-5.2开源当天，就适配了华为昇腾在内的**8大国产算力平台**——这意味着模型可以跑在非英伟达的芯片上。DeepSeek更进一步，据路透7月7至8日的报道，它正在自研推理芯片，项目约一年前启动，还处早期；同时从年初开始在内蒙古乌兰察布招数据中心的运维和交付岗，从轻资产研发往自建算力集群转。



模型从依赖英伟达，向华为昇腾等国产算力平台迁移——开源是把国产芯片–国产模型这条链焊死的黏合剂

一条国产芯片托着国产模型、模型又反过来喂养国产芯片的闭环，正在被开源焊起来。

软实力也一并收进口袋。GLM-5.2在前端编程盲测里拿下官方口径的「全球可用模型第一」，富瑞（Jefferies）给智谱的评价是「中国AI里程碑」。Kimi K3在Frontend Code Arena上得1679分登顶，压过Claude Fable 5的1631分，七个前端领域里六个第一，比上一代从第18名一口气蹿到榜首。

一个被管制的国家，用免费的代码，在别人的开发者社区里插旗。这盘棋走得很聪明。

一个人戴着沉重的镣铐，却在镣铐允许的半径里跳出了完整的舞步——约束划定了动作，也逼出了动作

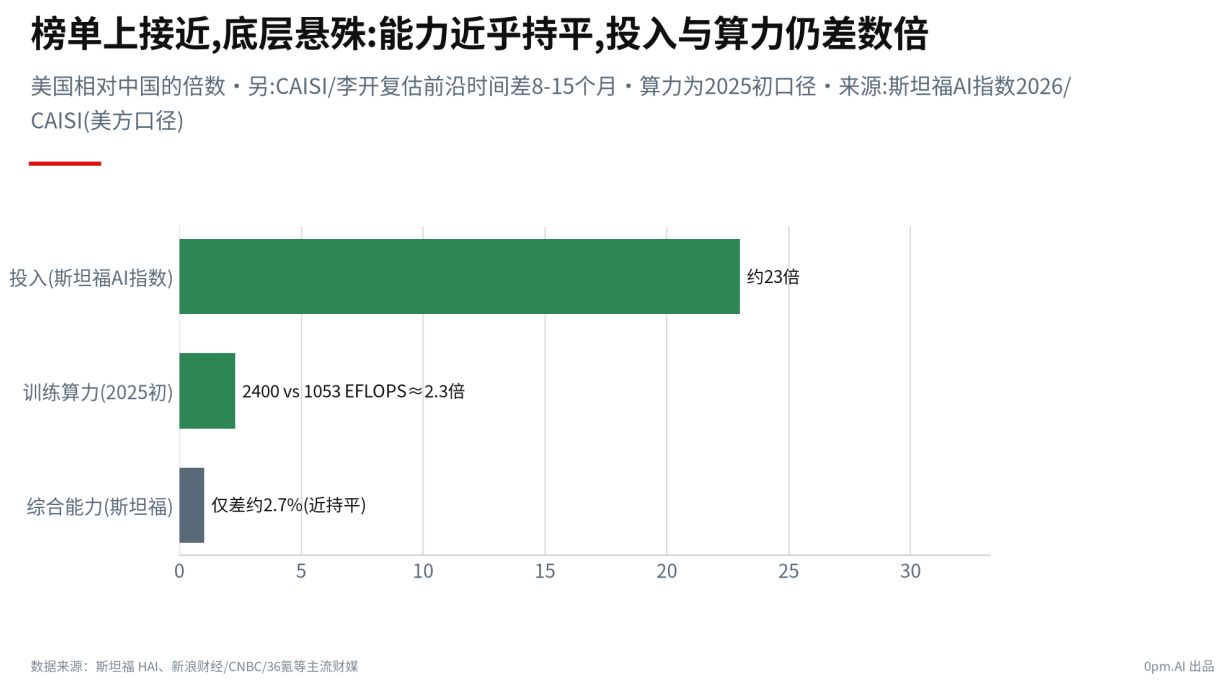
霸榜和登顶，是两件事

聪明归聪明，别被"霸榜"两个字晃了眼。这是最容易被民族情绪带偏的地方，也是本篇最要泼的一盆冷水。

中国开源模型确实在**开源专榜**上领跑。但把闭源模型一起拉进来看**总榜**，画风就变了：Artificial Analysis的综合智能指数前四是Fable 5、两款GPT-5.6 Sol，第四才是Kimi K3；前十五名里，中国模型只剩它一个名字。LMArena的文本盲测同样，Kimi K3排第9，前十五里也只有它一款。

开源榜第一，和含闭源的总榜第一，中间隔着的不是名次，是整个前沿。

差距有多大，几家不同立场的机构给了不同的尺子。美国CAISI（NIST下属）今年5月的评估认为，DeepSeek落后前沿约**8个月**（该报告含安全指控，作中立记录、不当定论）；李开复的公开估计更长，约**15个月**。斯坦福的AI指数则给了个更扎心的对照——顶尖中美模型的性能差已缩到约**2.7%**，可背后的投入差高达约**23倍**。再往上游看，2025年初的口径下，中美总训练算力约是**2400对1053 EFLOPS**。



多维度看中美差距：能力差2.7%、时间差8到15个月、算力差一倍多、投入差23倍——榜单上的接近，掩不住底层的悬殊

用2.7%的性能差去讲"追平"，却对23倍的投入差、一倍多的算力差闭口不谈，这是刷榜叙事最爱藏的账。

代价也开始显形。Kimi K3能力上去了，幻觉率却从上一代的**39%**升到**51%**（AA-Omniscience基准）——省算力、抢速度，是要还的。资本市场那边，富瑞一边喊里程碑，一边提醒：智谱的P/ARR约**94倍**，而Anthropic只有约**18倍**。榜单上的逼近，和估值上的透支，是同时发生的。

同一道墙，挡住了别人，也罩住了自己

现在可以把这道镣铐的两面拼起来看了。

护城河这一面是真的。开源换来了生态、换来了成本优势、换来了在别人平台上30%以上的token占有率，还把国产芯片和国产模型绑成了一条链。这些是闭源对手一时半会儿抢不走的地盘。

天花板那一面也是真的，而且没松动。最先进的制程和算力仍在管制之外，前沿的定义权还攥在别人手里。开源能让你跑得快、铺得广，却没法让你第一个撞线——因为撞线要的那台机器，你根本买不到。

护城河和天花板，本就是同一道墙的两面。它替你挡住了追兵，也把你自己罩在了下面。所谓"约束下的最优解"，重音永远在前半句：解是最优的，约束没有解除。

尾声

回头看"又一个DeepSeek时刻"，最深的一层从来不是谁又追上了谁。彭博7月17日把市场那阵恐慌命名为"Another DeepSeek Moment"，可真正在分岔的，是两种范式——一边是开源普惠、把成本打到地板，另一边是闭源前沿、把最强的东西攥在手里。中国被推上了前一条路，美国守在后一条路。

镣铐既划定了动作，也逼出了动作。问题不在于这支舞跳得好不好看——它已经足够好看了。问题在于：当有一天镣铐真的解开，这套为约束量身长出来的本事，是会长成翅膀，还是会发现，自己早已习惯了只在那个半径里起跳？

文/0pm.AI

风险提示：本文仅作行业逻辑分析，不构成任何投资建议。历史数据不代表未来表现，市场有风险，投资需谨慎。文中涉及榜单、算力、估值等数据均标注了口径与时点，不同来源口径存在差异，请以官方披露为准。

来源说明：本文数据与事实来自Artificial Analysis、Arena.ai（LMArena）官方排行榜，富瑞（Jefferies）研报（经新浪财经转述），美国CAISI/NIST评估，斯坦福HAI《2026 AI Index》，及CNBC、路透社、Tom's Hardware、Decrypt、MarkTechPost、量子位、IT之家等公开报道。

0pm.AI 出品

三周四款开源模型·彭博喊fear·但'霸榜'至少混了三种口径

真图驱动　一手权威源　多源交叉核验

数据与口径

数据来源

各家官方发布/技术报告(智谱开放文档/月之暗面/DeepSeek/腾讯混元)、Artificial Analysis人工分析、Frontend Code Arena/LMArena、Bloomberg彭博、Tom's Hardware/VentureBeat/CNBC、智谱华章招股书/港交所/年报、新浪财经/财联社/36氪/量子位、斯坦福AI指数/CAISI(美方口径)

访问官网

扫码直达 0pm.ai

官网入口
0pm.AI