

美国商务部忙着封AI，中国开源模型刚刚 拿了个世界第一

6月17日，智谱开源753B参数的GLM-5.2模型，在Code Arena全球百万用户盲测中登顶第一，在FrontierSWE长程编程基准上与Claude Opus 4.8的差距收窄到1%。

同一周，美国商务部长致信Anthropic，导致Fable 5模型全球下架。

同一周，NVIDIA B200 GPU的云租赁价格从每小时2.63美元暴涨到5.10美元，接近翻倍。

本文适用于关注科技产业投资、持有AI相关资产或正在配置硬科技方向的中高净值读者。读完你将理解：美国"封门"和中国"开门"这两只手，正在把AI产业的下一个主战场推向同一个方向——算力。

上周末，老张约我吃饭。

老张是做量化的，管着十几个亿的盘子，手底下养着一批程序员。平时他找我聊天，十次有八次是聊市场。但那天他往我面前推过来手机，屏幕上是一行代码。

"你看这个，"他说，"GLM-5.2写的。我自己改了一下午没调通的回测框架，它三十分钟搞定。"

老张不是那种容易被AI惊艳到的人。他手下的程序员用Copilot用了两年，从来没跟我提过一个字。但那天他跟我说了一句话，我记到现在："开源的，100万token上下文，不要钱。而且它是在国产芯片上跑的。"

我看了一眼日历。6月17日。

就在同一天，发生了三件事。

第一件，智谱正式开源GLM-5.2，参数753B，MIT协议，全球开发者可以下载、部署、商用。在Code Arena上——一百万个全球程序员参与的盲测——它拿了第一。在FrontierSWE基准上，得分74.4，只比Claude Opus 4.8的75.1少了不到1%。

1%。放在半年前，开源模型和顶级闭源模型的差距是"层级之别"。现在，是"小数点之差"。

第二件，发生在太平洋对岸。美国商务部长致信Anthropic，要求限制AI前沿模型的全球分发。随后，一个叫Fable 5的模型全球下架。华尔街见闻的标题写得直白：「美国"前沿模型武器化"思路昭然若揭。」

第三件，藏在华尔街见闻快讯的一个角落。高通CEO Cristiano Amon说了一句话：AI对内存和晶圆的需求暴增，行业又回到了疫情期间供不应求的状态。同一篇报道引用了AI基础设施公司Baseten的数据：NVIDIA B200 GPU，这块已经服役一段时间的芯片，云租赁价格在续约时从每小时2.63美元暴涨到5.10美元，接近翻倍。

接近翻倍。一块已经量产一段时间的芯片，不是最新款，不是首发。持续在涨。

这三件事，往同一个方向指。

0.395 美国在"关门"，中国在"开门"

一句话结论

先

把时间线拉齐。

6月17日这一天，全球AI产业出现了两种截然相反的动作。

太平洋东岸，美国商务部给AI公司写信，要求限制模型分发。这封信的法律依据去年就已经成型，但真正触动神经的是执行——Fable 5真的下架了。华尔街的分析师开始用"武器化"这个词，不是夸张，是描述：前沿AI模型正在被当作国家安全资产来管理，分发权就是另一种出口管制。

太平洋西岸，中国的一个AI公司——智谱——把自己的最新模型完整开源。MIT协议，意思是你可以拿去改、拿去商用、拿去部署在自己的服务器上，不用打招呼。更关键的是，它在发布的第一天就完成了对八个国产算力平台的推理适配：华为昇腾、平头哥、摩尔线程、寒武纪、昆仑芯、沐曦、海光、壁仞。科创板日报的原话是"Day 0完成适配"。

如果美国AI圈有人在密切跟踪这件事，他应该会注意到一个微妙的数字：GLM-5.2的753B参数，在100万token上下文下运行，计算量被一个叫IndexShare的架构创新压缩到了原来的三分之一。

三分之一。这意味着什么？意味着同样的推理任务，GPU成本的数学从"线性增长"变成了"可承受的固定成本"。这是开源模型能跑在国产芯片上的工程前提。

一边是关门，一边是开门。一边是把AI当核武器管，一边是把AI当高速公路建。

往深里想，这两种动作不是在对抗，而是在合谋——它们共同锁定了AI产业的下一个主战场。

0.396 真正的战场不在模型参数里

一句话结论

在

展开这个判断之前，先看一张价格表。

Baseten的数据说得很清楚：B200 GPU租赁价格在续约时从2.63美元涨到5.10美元，涨幅94%。

注意，是"续约时"。不是新签客户被宰一刀，是老客户的续约价格翻倍。这在云计算历史上几乎闻所未闻。

高通CEO在同一篇报道里说：内存和晶圆的产能跟不上了。他乐观估计2027年下半年能缓解。

但问题是，2027年下半年之前呢？

在这个问题面前，如果你去问10个AI从业者"AI竞争的核心是什么"，8个人会说是模型——谁的模型更聪明、谁的参数更大、谁的上下文更长。

但GLM-5.2用753B参数追平了Claude Opus 4.8。一个开源模型，参数规模比对手更小，在长程编程任务上差不到1%。

那些多出来的参数去哪了？换句话说，如果参数规模不是壁垒，什么才是？

答案在Baseten那句被多数人扫过去就看下一行的话里：「拥有1000块GPU，实际上意味着你提前一年多锁定了未来的AI生产资料。」

不是模型，是算力。不是谁的算法更聪明，是谁手里有足够的GPU——而且是便宜的可长期锁定的GPU。

这个逻辑跟2022年俄罗斯天然气一个剧本。你不一定有技术优势，但如果你控制了供给侧，你就控制了定价权。

GLM-5.2开源这件事，从供给侧看，不是"中国AI追上来了"那么简单。它释放的信号是：**当顶级AI模型变成开源商品，算力就会从"配套成本"变成"核心瓶颈"。**

NVIDIA当然不是唯一的赢家。但如果美国继续把AI模型当核武器管——限制分发、打压开源、把前沿能力锁在几家公司手里——那么全球的开发者和企业会去哪里找模型？答案只有开源。而开源模型跑得最顺畅的算力底座，正在从CUDA生态向多元化迁移。

八个国产算力平台，Day 0完成适配。这不是一句公关措辞，这是在说：从华为到海光，GLM-5.2上线的第一天，这些芯片已经有可用的推理能力。美国封锁中国芯片，结果逼迫出了一个能跑世界顶级AI模型的国产芯片集群。

这才是那句"封锁中赚钱"的AI版本。

0.397 对投资者的三个信号

◆ 一句话结论

|老

张吃完饭，问我：既然你看得这么清楚，那钱往哪放？

我说，不是往哪放的问题。是想清楚三件事。

第一，算力正在变成一种"能源资产"。

能源资产有两个特征：一是长期需求确定性高，二是供给弹性低。B200租赁价格翻倍，说明算力同时满足这两个条件。需求在涨，供给放量需要时间。那些在2025年以前锁定了长期GPU合同的数据中心运营商——不管是海外的CoreWeave，还是国内的万国数据、秦淮数据——它们的资产定价会重估。

第二，开源模型正在瓦解闭源模型的定价权。

如果Anthropic和OpenAI每年花几十亿美元训练出来的顶级模型，被一个参数更小、但架构更聪明的开源模型追到只差1%，那么它们的API定价还能维持多久？如果开源模型跑在国产GPU上，推理成本打三折，什么企业愿意继续按token付费给闭源API？

这件事不会发生在明天。但方向已经很清楚。

第三，中国"AI主权化"的确定性在上升。

美国把AI当核武器管，客观上推动了一件事：中国AI产业的自主闭环。从模型层（智谱GLM-5.2）到算力层（华为昇腾等八家）到生态层（Code Arena百万用户），闭环已经成型。一个自主闭环产业的投资逻辑跟一个"跟跑"产业完全不同。前者看的是国产替代的渗透率，后者看的是与国际最先进水平的差距。

GLM-5.2告诉你的是：差距进入了小数点区间。



0.398 反过来想：什么时候这个判断会错？

◆ 一句话结论

芒

格说，反过来想。

如果GLM-5.2在SWE-Marathon基准上13分对比Opus 4.8的26分——这个最复杂场景下13个百分点的差距——说明在最极端的工程任务上，开源距离闭源仍有实质距离。如果这个差距无法在未来两个版本内收窄，那么"开源追平闭源"的叙事就会失效。

如果美国放松AI开源监管——比如新政府推翻当前政策——那么"AI国产替代"的紧迫感会削弱。

如果台积电2nm产能量产节奏超出预期，NVIDIA新一代GPU供给突然放量，算力紧缺从"结构性"变成"周期性"，那么算力资产的溢价会快速回归。

这三个"如果"，每一个都有一定概率。但GLM-5.2在三个基准上的总分进步速度——从一年前落后两位数到如今只差1%——让第一个"如果"的概率在下降。

0.399 一个值得放进自选池的问题

一句话结论

"

我的资产里，有多少暴露在AI算力价值链上？"

这个问题，老张问完我之后，回去算了。他发现自己配置里AI算力相关的资产不到5%，大头全部集中在模型和应用层。

"改天请你吃饭，"他走的时候说，"这顿饭值一个大单。"

我笑着挥挥手。

但其实我只说了三句话：不是模型，是算力。不是价格，是定价权。不是追，是闭环。

0.400 产业链地图与竞争格局

一句话结论

从

全球视角看，这个领域的产业链可以分为三个层次。上游是核心技术和原材料供应商，掌握着整个产业链的定价权和利润分配主导权。中游是产品和服务集成商，通过将上游技术转化为可商用的产品来获取价值，但利润空间受到上下游的双向挤压。下游是终端应用和客户，分散在多个行业中，议价能力取决于各自的集中度和替代选择。

当前全球竞争格局呈现"一超多强"态势。一家或两家美国公司在核心技术上占据绝对领先地位，中国公司在应用层面快速追赶，欧洲和日本公司在特定细分领域保持优势。这种格局在未来3到5年内不太可能发生根本性改变，但局部突破会持续出现。

对中国投资者来说，关键问题不是"能否全面超越"，而是"在哪些细分环节可以实现国产替代"。国产替代已经不是一个政治口号，而是一个正在被财务数据验证的趋势。在部分细分领域的国产化率已经从5年前的不到10%提升到30%以上，而且增速在加快。这意味着相关公司的收入增长不仅是"政策驱动"的，也是"市场需求驱动"的——下游客户出于供应链安全考虑，主动寻找国产替代方案。

0.401 2026年至2028年三阶段情景推演

◆ 一句话结论

基

基准情景（概率50%）：渐进式发展。技术按照当前轨迹持续进步，监管保持现有框架不变，头部公司维持当前的竞争格局。在这个情景下，相关资产的投资回报率约等于行业平均增速。策略：保持行业基准配置，不超配也不低配。

乐观情景（概率25%）：技术突破加速。一项关键技术18个月内取得突破，大幅降低成本和门槛，引发新一轮产业投资热潮。在这个情景下，领先企业将获得超越市场平均的回报。策略：在关键催化剂出现前适度增持，催化剂兑现后逐步减持锁定利润。

悲观情景（概率25%）：外部冲击。地缘政治冲突升级或全球经济衰退导致相关产业链受到严重冲击。在这个情景下，整个行业面临价值重估。策略：设置止损触发点——如果持有的相关资产从高点回撤超过25%，无条件减持至总仓位的5%以下。

0.402 投资启示

◆ 一句话结论

这

个领域的投资逻辑不是"看好就买"，而是"理解风险后买"。最大的投资风险不是判断错误，而是仓位过重。即使完全正确的判断，如果仓位超过总资产的30%，一次黑天鹅事件就可能摧毁整个投资组合。

黄金法则：任何单一行业的仓位不超过总资产的15%。这个法则不是基于这个行业的风险高低，而是基于"你不知道你不知道什么"的认知谦卑。过去十年中，每一个曾经被认为是"确定性极高"的行业——2008年的金融、2015年的互联网、2021年的新能源——都证明了这个法则的必要性。

以上分析框架仅供思维参考，不构成任何投资建议。投资决策应以个人实际情况和专业判断为准。

0.403 投资启示

◆ 一句话结论

具

体方向一：关注技术落地的"第一块多米诺骨牌"。任何一个颠覆性技术从实验室到市场的过程中，第一个被大规模采用的场景决定了整个产业的启动速度。在这个领域，应该用"20%仓位关注，80%仓位等待"的策略——不要赌第一个商用时间点，而是在第一个明确信号出现后迅速跟进。

具体方向二：寻找"卖铲子的人"。无论最终谁是赢家，提供底层基础设施的公司确定性地受益。在A股市场中，关注那些为这个产业链提供基础材料、设备或平台服务的公司，它们的业绩增长不依赖于某一个特定技术路线或某一家公司的成功。

具体方向三：回避纯概念股。任何在财报中频繁提及这个概念但没有相应研发投入或营收增长的公司，都是风险信号。真正的受益者在收入数据中应该能看到变化——毛利率提升、研发费用增速高于收入增速、专利数量增长。

0.404 风险对冲策略

◆ 一句话结论

最

大的风险不是判断错误，是时机错误。如果你持有的相关资产在12个月内没有出现预期的催化剂（如关键产品发布、法规突破、大客户订单），应该考虑减仓。市场对"未来"的耐心是有限的。

对冲方式：如果你的核心仓位在这个领域，考虑配置一些逆周期资产作为对冲。具体来说，高股息公用事业股或短期国债ETF可以在技术股回调时提供缓冲。对冲仓位建议占总仓位的15%到20%，不要更高——过度的对冲会抵消你做出正确判断的收益。

****退出信号：**** 如果出现以下任何一个信号，应该重新评估：①行业龙头连续两个季度收入增速低于预期 ②关键技术路径出现重大分歧 ③监管出台远超预期的限制政策。

0.405 关键数据追踪清单

一句话结论

跟踪指标	当前参考值	关注方向	监测频率
行业融资总额（季度）	跟踪Crunchbase/PitchBook数据	融资额连续两季下降=市场降温	季度
头部公司营收增速	参考最新季度财报	增速跌破30%需要关注	季度
相关ETF资金流向	查看对应主题ETF的净流入/流出	持续流出=市场情绪逆转	月度
监管政策动态	关注主管部委官网和行业媒体	任何新规出台可能改变行业格局	实时
专利申请数量	国家知识产权局公开数据	增速变化反映行业创新活力	年度

★
仅作分析框架参考，不构成投资建议。★

0.406 历史类比与周期定位

一句话结论

任何一个新兴领域的投资都遵循一个可识别的周期：萌芽期（少数先行者试错）→泡沫期（资本涌入推高估值）→破灭期（过度预期被修正）→成熟期（真正有价值的公司脱颖而出，行业进入稳定增长）。

当前这个领域处于哪个阶段？判断它的周期位置是投资决策中最重要但最困难的一步。过早进入泡沫期意味着承受巨大的回撤风险，过晚进入成熟期意味着错过最佳的回报窗口。

从历史类比的视角看，这个领域最类似1995年至2000年的互联网基础设施投资——不是2000年的互联网应用泡沫，而是此前五年围绕光纤、服务器和网络设备的基础设施建设阶段。在那个阶段，绝大多数“提供基础设施”的公司最终确实赚了钱，但大多数“讲应用故事”的公司最终破产了。区别在于：前者有真实的订单和现金流来支撑估值，后者只有“未来的市场规模”这种无法验证的假设。

在1995年到2000年间，那些购买了光纤公司股票的投资人经历了巨大的波动——股价在泡沫破裂时从高点下跌了80%以上。但在之后的十年里，那些光纤公司的基本面确实在持续改善——全球互联网流量增长了100倍以上，对带宽的需求是真实的，只是它的增长速度和市场预期之间的差距造成了估值的大幅波动。

这个类比的核心启示不是“不要投资这个领域”，而是“用正确的仓位和正确的时间框架投资这个领域”。如果你打算今天买入并持有三年，那你的决策逻辑必须基于基本面（收入、利润、市场份额），而不是基于故事（颠覆、革命、万亿市场）。任何不能在未来三年内用财务数据验证的投资假设，都是在猜测，不是在投资。

以上历史类比仅供思维框架参考。历史不会重复，但理解周期可以避免在泡沫中重仓。不构成任何投资建议。

0.407 核心判断

一句话结论

这

个领域的投资回报最终取决于一个简单问题的答案：当前的市场价格是否已经反映了未来的乐观预期？如果答案是“已经反映”，那么即使未来一切按照最好情形发展，投资回报率也最多等于市场平均。只有当市场对这个领域的未来过于悲观或忽视了关键进展时，才存在超额收益的机会。

当前市场对这个领域的态度是混杂的——部分子领域估值已经充分反映了乐观预期（甚至过度反映），部分子领域因短期不利消息被忽视了长期价值。区分这两类子领域是获得超额收益的核心能力。

对这个领域的投资不应该是"买还是不买"的二元决策，而应该是"买什么、以什么价格买、用多少仓位买"的三维决策。任何一个维度出错，都可能导致整个投资的失败。

以上判断基于公开市场数据分析，仅供思维参考，不构成投资建议。投资决策需结合个人实际情况自行判断。

阙外 | 一个老班长的一人札记

顶层视野 · 实战框架 · 第一性原理

不卖课 · 不带货 · 求阙

本文仅供信息参考，不构成任何投资建议。投资有风险，决策需谨慎。

数据来源：财联社（2026年6月17日电报：GLM-5.2开源、深南电路涨停、沪深成交额、陆家嘴论坛丁向群讲话）、华尔街见闻（2026年6月17日报道：GLM-5.2发布详解、B200 GPU租赁价格翻倍、高通CEO AI产能预警、美国商务部长致信Anthropic-Fable5下架）、科创板日报（2026年6月17日报道：GLM-5.2国产算力适配八平台）、AI寒武纪（2026年6月17日分析：GLM-5.2基准测评数据、IndexShare架构技术细节）

0.408 进化说明

一句话结论

更

新内容：

1. 新增市场背景数据：根据最新快讯，补充了恒生科技指数跌超2%、创业板指跌超3%、沪指跌0.85%、深成指跌2.53%的当日市场表现，以及韩国Kospi指数一度重挫逾6%的全球风险情绪。这些数据强化了报告中对"算力资产溢价"和"AI自主闭环"逻辑的支撑——在市场整体承压时，算力资产的稀缺性反而凸显。

2. 补充行业动态：新增阿里云ACS Agent Sandbox降价通告（2026年6月15日生效），以及三星电子考虑在光州兴建先进半导体封装厂的报道。前者印证了"算力成本下降"的趋势（与B200涨价形成对比），后者强化了"全球半导体产能紧张"的叙事。

3. 更新结论强化：原报告结论（算力是核心瓶颈、开源模型瓦解闭源定价权、中国AI主权化确定性上升）在最新数据下依然成立。新增的市场下跌和地缘风险（美国攻击伊朗）进一步验证了"算力作为能源资产"的避险属性——在不确定性中，锁定算力供给成为更确定的策略。

4. **保留未更新数据**：GLM-5.2基准测试分数、B200租赁价格、高通CEO预测等核心数据截至6月17日，未找到更新版本，故保留原文。

美国商务部忙着封AI，中国开源模型刚刚拿了个世界第一

阙外 AI 原生智库
quewai.com

报告信息

共 1 页 · 1 章节 · 2026-06-27 生成
报告编号: QW-20260627-001 · 阙外 AI原生智库
