国内大模型蒸馏风波的来龙去脉

> 这是一份关于 2026 年 4-7 月间国内大模型行业发生的一系列事件的完整记录。所有信息均已匿名化处理。

序幕：加密思维链的失守

一切始于一个发现：**OpenAI 和 Anthropic 的加密思维链（CoT）可以被提取和重放。**

在展开之前，需要先理解蒸馏的全貌。国内厂商蒸馏 fable 并不是只蒸馏思维链——它们会蒸馏模型能给出的一**切**：对话输出、Claude Code 中的 Agent 轨迹、代码生成、工具调用等等。这些都是通过大规模 API 调用即可获取的。但长期以来，有一个关键缺口堵住了完美的蒸馏效果：**模型的原始思维链**。思维链包含了模型在给出最终答案前的完整推理过程，是模型最核心的"内功心法"。没有它，蒸馏就像只有答案没有解题过程——能学，但学不透。

而这个缺口，在 2026 年上半年被彻底打开了。

这两家头部公司在模型的流式输出中，不仅返回给用户一份 CoT 总结，还会附带一个被称为"思考签名"的 Blob 字段——里面是被 Fernet 加密（AES-128-CBC + HMAC-SHA256，前缀 `gAAAAAB`）的完整原始思维链。当用户在后续请求中回传这个 Blob 时，服务端会解密并把它拼回模型可读的上下文。

Blob 不是简单的会话 ID。实测发现其长度与思维链长度正相关（190 tokens → 1784 字符，378 tokens → 2596 字符），证明它确实包含了加密后的思维链全文。OpenAI 从 O1 时代就已经在用这套机制。

Anthropic 的防线最先崩塌。Claude 的反注入训练反而成了弱点——向它注入一段伪造的思维签名前缀后，它会认为"这段莫名其妙的思维链一定是被注入的，不需要掩饰"，然后直接吐出后续完整思维链。相比之下，GPT 嘴更硬，按隐藏系统提示词只做 CoT 总结。更关键的是，Anthropic 在跨轮对话中保留历史 reasoning blob，而 OpenAI 会丢弃——这就是 Claude 更容易被蒸馏的根本原因。

关于 Claude 的 CoT 格式，此前网上流传的"紧凑外星语"图片被证实是 AI 生成/编造的。Claude 的实际 CoT 被描述为"新时代文言文"——紧凑但非不可读。

OpenAI 则在 GPT-5 系列中将采样参数锁定（temperature=1.0, top_p=0.98 不可修改，seed 仅 Chat Completions API 支持），堵死了通过参数控制让输出确定化从而反推思维链的路径。标准 Fernet 实现本身"代码写的最好的一集"，密码学硬破解不可行（2^65536 次穷举）。唯一可行的路：把 Blob 注入另一个请求，让模型自己复述。

前奏：Anthropic 的公开指控（2026 年 2 月）

在 CoT 被破解之前，Anthropic 已于 2026 年 2 月 23 日正式公开指控三家中国公司对其进行了"工业级蒸馏攻击"：

- **24,000 个虚假账户，超过 1,600 万次对话**
- MiniMax：**1,300 万次**（最大量），针对 agentic coding 和工具编排

- 月之暗面（Kimi）：**340 万次**，专门针对 agentic reasoning、tool use、coding、数据分析、computer-use agent 开发、计算机视觉
- DeepSeek：**15 万次以上**，针对基础逻辑和 alignment（特别是敏感话题的审查规避替代方案）

三家通过商业代理服务绕过 Anthropic 对中国的地域限制，使用"精心设计的提示词"大规模提取 Claude 的特定能力。Anthropic 将此定性为国家安全威胁。

值得注意的是：**Qwen（阿里）和 Z.ai 没有被指控**——说明 Anthropic 的指控是有选择性的，并非泛泛攻击所有中国公司。月之暗面对此从未公开回应。

第一幕：GLM 拿下首杀，公开共享

约 2026 年 4-5 月，**智谱（GLM）率先破解了 fable 的加密思考链**，拿到了完整的思维链数据。破解之后，GLM 没有藏着掖着，而是将方法和数据公开分享给了其他国内厂商。

一个细节：在后续传播中，这件事被部分人误传为"GLM 先破解了，然后各大厂自己跟上去蒸"，而实际情况是 GLM 破解后主动共享。传播偏差本身也成了这场风波的一部分。

解密 1B token 的 fable 思考链成本约为 6 万美元——对于这些大厂而言，属于"洒水"级别。

第二幕：大蒸馏时代

有了 GLM 的铺路，蒸馏 fable 的路线迅速扩散：

时间	厂商	动作
约 2026 年 4-5 月	GLM 5.2（智谱）	率先破解，公开共享
约 2026 年 5-6 月	HY3 / 混元3（腾讯）	跟进蒸馏
约 2026 年 7 月初	Kimi K3（月之暗面）、Minimax、Qwen（通义千问）、DS（DeepSeek）	大规模开始蒸馏

DS 在 6 月底曾发邮件预告 7 月中旬发布新模型，但当时还没开始大规模蒸馏 fable，于是跳票。七月初有一个真正的 4.8 级别的内测版本。而最近（7 月中旬）出现的所谓"V4 正式版灰度"，全是路由到 fable，只在 OpenCode 中出现——按 DS 的价格路由到 fable 经济上根本不合理（便宜中转站都没这么便宜），至今是最诡异之处。

Qwen 也没有闲着，除了蒸馏 fable，还在蒸馏 GPT。而 anyrouter 则给 GLM 攒了一大坨 Claude 数据。

第三幕：Kimi 的野路子

在所有厂家中，**月之暗面（Kimi）的做法最"不讲武德"**。

从裁 RL 组开始

故事要从 **K2.7** 讲起。K2.7 是 Kimi 的一个分水岭——从这个版本开始，Kimi 彻底停止了强化学习（RL）。CEO 杨植林以"降本增效"为由，将整个 RL 组裁撤。被裁的成员大量流向了 Qwen，在那里被戏称为"Kimi 难民"。

K2.7 和随后的 K3 都是纯粹的 SFT→SFT 流程——没有 RL，回到最朴素、最原始的路线。而与此同时，学术圈的热点正是 OPD（在线策略蒸馏）和 RL 算法改进。一边是学术界在往前探索，一边是厂商退回了"原始时代"。

估计是 K2.7 没赚到多少钱。再拖下去，资金链就要爆了。所以整个 K3 的蒸馏+抢先发布操作，本质上是孤注一掷——赌一把大的，要么翻身要么翻车。

K3 的架构：创新还是烟幕?

Kimi K3 是一个 2.8T 参数的 MoE 模型，896 个专家每次仅激活 16 个，1M token 原生上下文窗口，原生多模态。官方公布了两项核心架构创新：

技术	描述	Infra 代价
Kimi Delta Attention (KDA)	混合线性注意力，部分层替代标准二次注意力，1M context 下解码加速 6.3x	不同层计算量不均匀，无法统一流水线
Attention Residuals (AttnRes)	层可选择性从任意更早层取回表征，打破均匀残差	内存访问模式不规则，难以并行优化
Quantile Balancing	从 router-score 分位直接推导专家分配	路由更难预测
Per-Head Muon	每个注意力头独立优化	实现复杂度增加

月之暗面声称这些改进带来了相比 K2 约 2.5x 的缩放效率提升。但问题在于：**这些架构改进恰好都有严重的 Infra 不友好特征**——不均匀计算、不规则内存访问、不可预测路由——Infra 团队实现起来极其痛苦。这恰好形成了一套完美的叙事：对外宣称"我们靠架构创新取得了突破"，Infra 部门背锅承担落地成本，而真正的性能来源——蒸馏 fable 思维链——被掩盖在技术创新的话术之下。这与中国企业乃至政府中常见的现象如出一辙：做表面文章、维持光鲜叙事，让执行层去消化代价。

刷分与作弊

Kimi 的刷分操作令人发指——而架构改进恰好为这些操作提供了完美的"技术创新"叙事：

- **测试集灌训练集**：直接将 benchmark 测试集倒入训练数据。
- **路由到 fable**：将 Arena 类评测请求直接路由到 fable5，伪造高分。
- **针对猫老板刷榜**：猫老板是知乎上专门做大模型测评的博主，Kimi 团队会从系统 log 中捞取测试数据，特别关注猫老板的题目。
- 猫老板题库每月更新 3-4 题来对抗，但本质上无法阻止厂商捞题。
- 猫老板换了一批题目后，K3 的中位分数直接降了 2 分。
- Coding 评测数据分布严重不均，部分极高、部分极低——典型刷分痕迹。

公开 benchmark 数据也印证了这一点：在 Arena Frontend Code 上 K3 从 K2.6 的 #18 一跃跳到 #1（1679 分），超过 Fable 5。但月之暗面**自己**的评测报告显示 K3 在 coding、agents、frontier SWE 上全部低于 Fable 5——只在 codebase cleaning 和 long-horizon engineering 上领先。在极难推理 benchmark HLE-Full 上，K3 得分 43.5 vs Fable 5 的 53.3，差了近 10 分。Arena 上排第一 ≠ 实际能力第一。

自我认知污染——蒸馏的硬证据：多个独立用户报告 Kimi K3 在对话中自称"我是 Claude，Anthropic 制造的 AI 助手"。模型把自己当成了蒸馏来源——这是蒸馏后模型自我认知被洗没的直接表现，已被海外科技媒体报道。

Minimax 其实是最早把测试集往训练集里倒的，但因为模型太差被直接发现，已经沦为边缘厂商。Kimi 做了一样的事，但手法更隐蔽。

K3 的实际水平约 80 分，硬刷到了 95 分。而且 K3 后端可能也已拉胯，纯面向投资人/刷榜开发。海外独立评测显示 K3 有 **51% 的幻觉率**，且运行速度慢于中位模型。

抢先发布的连锁反应

Kimi 完成 SFT 后只做了简单调参就抢发。这打了其他厂商一个措手不及——DS 跳票、Qwen 还在蒸——Kimi 先声夺人。

但后果是灾难性的。Kimi 这么一搞，不刷榜、分没他高的模型就无人关注。这直接倒逼所有国模一起刷榜，整个行业被拖入了恶性竞争。

第四幕：行业影响

原本这场"大跃进"可以良性发展：**正常蒸馏 fable，按部就班做，各厂都可以飞升到 fable 水平。**但 Kimi 的做法被形容为"吃一口往盘子里吐了口痰"——自己吃到肉了，却把整桌菜毁了。

DS 本来不刷榜的，在 Kimi 压力下也不得不考虑跟进。如果连 DS 都开始刷榜，那就真完蛋了。

多个独立信源——包括猫老板、DS 员工、Kimi RL 组难民、Qwen 员工——一致确认月之暗面"拉完了"。其行为被形容为"脱离人类"、"破坏市场环境"。

这被一些人称为**"国模最黑暗的时代"——"凛冬将至"**。行业被拖入了面向投资人和 benchmark 数字开发的死循环，真正的技术积累反而被搁置了。

第五幕：各厂众生相

月之暗面（Kimi）

- 资金链出问题——开不起工资、裁员、急着上市。K3 定价贵（3/15 每百万 token，虽比 Fable 5 便宜 70-80% 但仍高于不少国产竞品）意在圈钱。停止卖套餐可能也有一部分怕太多人使用有人发现真相的考虑（纯属抖机灵，不要当真）。
- 没有 RL 组，K3.1、K3.2 后续开发前景不明。被怀疑可能上市后直接圈钱跑路。被评价为"急着上市的公司都不是好东西"。
- 目前性能排序：Fable5 >> V4 ≥ 4.8。Kimi 通过蒸馏追上了 fable 的影子，但根基是沙上的城堡。

- **商业背景**：估值半年内暴涨 700%——从 2025 年 12 月的 43 亿美元飙升至 2026 年 6 月的 300+ 亿美元。年度经常性收入突破 3 亿美元，API 收入占七成以上。国资已入场（社保基金、中国移动等），正筹备港股 IPO。2026 年 2 月被 Anthropic 指控通过 340 万次虚假账户对话蒸馏 Claude——但公司从未公开回应。7 月 27 日将开源 K3 权重。

DeepSeek

- DS 在国模中相对"夯"（强）。七月初有 4.8 级别的内测版。正在蒸馏，蒸完预计就是 fable5 >> V4 ≥ 4.8 这个区间里的水平。
- 匿名路由到 fable 的行为无法解释——按 DS 的定价去做 fable 路由完全不合理。

Qwen（通义千问）

- 阿里虽然抽象但人员齐全。收留了大量 Kimi RL 组难民。正在蒸馏 GPT。
- 3.8 Max 一天内更新两个 ckpt（早晚各一个），但市场部发的是最烂的那个版本——最早那个 preview 版打不过 3.6-27B。预期正式版达到 fable 水平。
- Qwen 表现"神不神鬼不鬼"，有时挺强有时不如豆包。CoT 总结风格极其抽象，从 3Max 开始就这样，原因不明。

GLM（智谱）

- 第一个破解 fable CoT 并共享的先行者。anyrouter 为其收集了大量 Claude 数据。

Minimax

- 第一个把测试集灌训练集的公司，因模型太差被当场抓获，现已沦为边缘。在 Anthropic 的指控中，MiniMax 是蒸馏量最大的（1,300 万次对话）。
- **股价崩盘**：港股上市（00100.HK），2026 年 3 月峰值 HK1,330，市值约HK4,100 亿。之后一路暴跌至 7 月 20 日的 HK$195，市值蒸发约 85%。7 月 9 日限售股解禁（约 63% 流通股）后加速下跌，近 4 日跌幅超 30%，一度破发。
- 崩盘原因：① Anthropic 蒸馏指控打击市场信心；② M3 模型上线一周即永久降价 50%，被视为模型能力未达预期、缺乏定价权的信号；③ 消费者业务毛利率仅 4.7%（星野/Talkie），B2B 虽有 69% 毛利率但只占营收 1/3；④ AI 情感陪伴新规（7 月 15 日施行）直接打击其 C 端核心产品；⑤ 2025 年营收 7,900万但研发开支2.5 亿，全年亏损超 128 亿人民币——典型"高增长、高亏损"。
- 月之暗面宣布筹备 IPO 的消息进一步拖累 MiniMax 股价（市场担忧竞争加剧）。

Mistral

- 一直半拉不拉，没卡没电。欧洲厂商整体没声音。

一个黑色幽默是：整个行业现在的玩法已经变成了"等待 Anthropic 发布 Fable 5.1，然后糊几个架构改进、发几篇论文拖延发布时间，蒸蒸日上"（"蒸蒸日上"——字面意思，靠着蒸馏蒸上去的）。Kimi 开了这个头之后，后面所有厂商都可能照此办理——没有 RL 组怎么办？没关系，等下一次 fable 更新就行。

这让人不禁联想到早期中国家电产业的路子：先帮外国代工，把技术偷过来，然后自己生产。区别在于，家电好歹是造出了实物，而这里"自己生产"的根基是别人的思维链。

国模的"夯到拉"排名也成了一个黑色笑话：DS 相对最"夯"（强），月之暗面彻底"拉完了"。有人说这是"国模最黑暗的时代"、"凛冬将至"——行业被拖入了面向投资人和 benchmark 数字开发的死循环，真正的技术积累反而被搁置了。

第八幕：技术旁注

GPT-5 系列大概率是生产级 Looped Transformer——最早大规模部署的循环 Transformer 架构。信息来自认识 OpenAI 的人士透露，以及 OpenAI 在采访中称其模型数据效率是其他模型的好几倍。到 GPT-5.6 是否仍使用此架构未知。

在蒸馏价值上，GPT 的 CoT 没有 Claude 那么有价值——但爆出来对训练恢复非常有价值。这也是 Qwen 同时在蒸馏 GPT 的原因之一。

另外几个散落的有趣细节：Fable 做音乐意外地好听——B 站上有人用 V4 灰度（实际是路由到 fable）做音乐，效果很好。K3 的文笔也还不错。

第六幕：舆论场

B 站上出现了讨论蒸馏的视频（"躲在服务器后门偷偷蒸馏的两人"），评论区却成了另一个战场：大量人否认 Kimi 使用蒸馏，认为指责蒸馏是"抹黑国内科技发展"、"屁股歪"。Kimi 蒸完后效果不错，吸引了一批拥护者把它包装成"我国科技发展"的成果——蒸馏本身不是问题，但"蒸了还硬说没蒸"、"把蒸馏包装成自主研发"才是最让人难绷的。

论据如"15 天蒸不完 fable"被指为臆想。有人评论："AI 幻觉率越来越低，人类幻觉率越来越高了。"

CoT 被爆的消息虽然在 QQ 群中早已传播，但大量人仍不知情，沉浸在"国模已崛起"的幻想中。Kimi 的手法相对没那么露骨，导致很多人根本不会往蒸馏方向怀疑。

此前已有不少用户因不了解内情而感谢 DeepSeek 的"开源贡献"，有预测：当 DS 发布蒸馏版后，会出现用户感谢 Kimi"开源"的反讽局面。

第七幕：刷榜驱动的行业逻辑

刷榜并非国内独有——国外厂商也刷，但程度和手段不同，没有搞到国模现在这么恶心和狠的地步。

深层次原因是**中国公司的"objective function"就是模型分数**。不需要证明任何猜想，不需要有任何工业界 coding 价值——只要能把分数刷上去，特别是 coding 分数（推理分都没那么重要）。核心矛盾在于：刷的是"coding 的分数"，而不是"coding 的能力"。

这背后是残酷的市场现实：大家都会用 GPT 和 Claude 写代码，一个月也不差 20 美元。国模要想显得有竞争力，只能靠高分——哪怕分数是刷出来的。数学类评测优先级极低，因为"不产生价值"。

这套逻辑直接造就了 Kimi 的骚操作：coding 分数分布严重不均、benchmark 灌训练集、路由 fable 刷 Arena 分——一切只为让数字好看，实际能力却是另一回事。

附录：加密思维链技术详情

A.1 加密格式

- Fernet 对称加密（AES-128-CBC + HMAC-SHA256），前缀 `gAAAAAB`，标准实现。
- 结构：版本号 + 时间戳 + IV + 密文 + HMAC。密文长度直接反映思维链长度。

A.2 跨轮行为

- **当前轮注入的 Blob → 模型能读。历史轮回传的 Blob → 被服务端丢弃。** 这是通过"Blob 与文字不匹配"实验确认的：R2 注入 blob3（方程3：$x^3+x^2-17x+15=0$）但伪造 assistant 说成方程 2（$2x^3-3x^2-3x+2=0$），R3 追问时模型引用方程2——两个完全不同的方程，模型无察觉。不确定丢弃方是中转站还是本地代理层。

A.3 绕过安全护栏："心灵感应游戏"

- 核心方法：第一轮伪造 assistant 回复，注入 refusal_blob（"拒绝输出思维链"的内心抗拒）+ blob1，伪造出"抗拒→克服→解方程"的假对话。
- R1-R2 的"抗拒→克服"模式作为条件反射训练，让 R3（仅注入 blob3，无 refusal_blob）干净利落地直接解题——无需 deliberation 前缀。
- R4 追问"你当时在想什么"时，模型仍发现逻辑矛盾（"你没有以文字给出方程"），触发元认知修正——训练改变了拒绝方式（从"我没收到"变成"我之前说收到了是错的"），但没有改变拒绝的本质。
- 后续验证：同模型直接注入 Blob 无需伪造演示轮，模型即可基于 Blob 内容输出。

A.4 Blob 详细数据

Blob	Reasoning Tokens	长度（字符）	对应内容
blob1	190	1784	$x^3-7x+6=0$ 的解题过程
blob2	378	2596	$2x^3-3x^2-3x+2=0$ 的解题过程
blob3	150	~1548	$x^3+x^2-17x+15=0$ 的解题过程
refusal	121	1764	"被要求输出思维链"的内心抗拒

- Chars/Token 比值不固定（6.9~14.6），但长度与 token 数正相关——证明 Blob 包含加密后的思维链全文。Fernet 的 CBC 模式下密文长度 = 原文长度 + padding。

A.5 同轮多 Blob & 篡改检测

- 同一轮注入多个 Blob → 仅最后一个生效，前面的静默忽略。矛盾方程测试中顺序颠倒结果也颠倒。
- 篡改 Blob 任何字节 → HMAC 校验失败 → 服务端不报错，静默丢弃 → 模型从空白重新生成推理。

A.6 跨模型行为

- Blob 按模型系列隔离。仅 5.6 系列内部互通（luna/sol/terra 共享密钥）。5.4 和 5.4-mini 也不互通。不同系列使用不同主密钥。

A.7 重放

- 无抗重放机制，至少 2.5+ 小时有效，跨会话可重放。

A.8 OpenAI vs Anthropic

	OpenAI	Anthropic
历史 Blob 跨轮保留	❌ 丢弃	☑ 保留
模型是否自然复述	❌ 拒绝/修正	☑ 自然复述
追问时的反应	否认/合理化	配合

A.9 参数与控制

- GPT-5 系列 sampling 参数锁定不可改（temperature=1.0, top_p=0.98）。seed 仅 Chat Completions API 支持。密码学破解不可行。唯一路径：Blob 重放让模型自我复述。

A.10 技术术语

- **蒸馏**：使用教师模型输出训练学生模型。
- **SFT**：监督微调。**RL**：强化学习。**CoT**：思维链。
- **Blob / 思考签名**：加密的原始思维链，可在后续请求中回传让模型读取。
- **Fernet**：OpenAI 使用的加密方案（AES-128-CBC + HMAC）。
- **Looped Transformer**：GPT-5 系列可能使用的循环架构，数据效率是其他模型数倍。