



计算机行业研究

买入（维持评级）

行业周报

证券研究报告

计算机组

分析师：刘高畅（执业 S1130525120005）

liugaochang@gjzq.com.cn

AI 安全评估

模型能力升级，AI 安全重要性提升。大模型风险正由内容生成扩展到能够调用工具、访问外部系统并修改现实状态的智能体。AISI 数据显示，前沿模型完成初级网络安全任务的成功率由 2023 年末低于 9% 升至 2025 年的 50%，并首次完成通常需要十年以上经验的专家级任务；METR 测得 2023 年以后前沿模型在 50% 成功率下的任务时长翻倍周期约为 130.8 天。同一模型的能力还会随推理预算上升，约 8% 的网络安全任务在 token 预算超过 1,000 万后才首次被解决。这意味着风险不再只取决于模型权重，而由模型能力、身份权限、工具配置、网络连接和外部环境共同决定。Anthropic 与 AISI 披露的越界事件表明，评测环境误接公网、关闭生产防护或缺少目标范围检查，都可能把模拟任务转化为真实行动。中国、欧盟及前沿模型厂商已把风险分级、内容标识、模型评估、事件报告和发布门槛纳入治理流程，美国新设工作组也将评估 AI 风险、机遇及联邦政府职责。

AI 安全走向纵深防御。单纯依靠对齐训练和内容过滤难以覆盖智能体的完整风险链条，安全体系需要同时管理数据、模型、权限、工具和运行环境。训练阶段应处理高风险数据、目标函数、对抗样本和模型权重保护；评估阶段需要组合 Model Testing、Red Teaming 和 Field Testing，并记录模型版本、推理预算、工具配置、采样次数和评分方法；运行阶段则要以最小权限、沙箱、网络白名单、结构化工具接口、人工审批和轨迹监测限制模型行为。固定基准容易受到题目泄露、分数饱和和测试预算不足影响，高风险评估还需结合专家复核与更充分的压力测试。optstop 等统计方法可在示例中减少 57% 至 97% 的计划运行，但低概率、高后果任务仍需保留最低样本量和审慎的停止策略。人员因违反敏感信息处理政策被解雇的事件进一步说明，第三方评估不仅要保持独立性，也要同时建设、访问控制和审计机制。

AI 安全不改产业趋势，将带来更大算力投入。现有证据尚不支持当前 AI 系统已经达到失控水平，但能力提升快于控制体系更新，安全投入将延长训练、发布和部署流程。OpenAI 因候选模型在诚实披露和授权范围测试中未达标准而取消发布，说明评估结果已经能够改变产品节奏，也验证了持续评测、整改和回归测试的实际价值。对企业客户而言，安全既增加模型测试、文档、监测和人工复核成本，也是开放邮件、数据库、代码提交等高价值权限的前提。相关需求将由一次性合规检查转向持续评测、运行防护、日志存储和事件响应，并向云平台、网络安全厂商及第三方评测机构扩散。安全计算贯穿训练、评估和运行三个阶段，会增加重复推理、分类器调用、长轨迹回放与数据留存。IEA 预计全球数据中心用电量将由 2024 年的约 415 TWh 增至 2030 年的约 945 TWh，AI 是主要驱动因素之一；安全并非基础设施扩张的唯一变量，但会提高单位应用背后的计算密度和持续利用率。

相关标的

大模型&云：智谱、MiniMax、阿里巴巴、腾讯控股、金山云、百度集团、优刻得、首都在线、网宿科技、云赛智联、青云科技等。海外算力：东山精密、工业富联、胜宏科技、江海股份、中际旭创、中钨高新、蓝思科技、东阳光、火炬电子、三环集团、欧科亿、天孚通信、鼎泰高科、新易盛、领益智造、兆易创新、鹏鼎控股、唯科科技、天岳先进、大普微、源杰科技、麦格米特、景旺电子、英维克、京东方等；英特尔、SK 海力士、Lumentum、闪迪、铠侠、美光。国内算力：寒武纪、海光信息、长鑫科技、东阳光、利通电子、中芯国际、华虹半导体、华达科技、摩尔线程、天数智芯、沐曦股份、壁仞科技、协创数据、禾盛新材、华丰科技、杰华特、京基智农、中科曙光、浪潮信息、利扬芯片、胜蓝股份、华勤技术、国科微、中国长城、晶科科技、罗曼股份、盈峰环境、芯原股份、晶科科技、亿田智能、豫能控股、星环科技、鸿日达、盛视科技、神州数码、润泽科技、大位科技、润建股份、奥飞数据、瑞晟智能、科华数据、潍柴重机、欧陆通、杰创智能、奥尼电子。AI 应用：达梦数据、深信服、卓易信息、星环科技、广立微、绿盟科技、网宿科技、海天瑞声、合合信息、德才股份、美年健康、真爱美家、中控技术、迈富时、金蝶国际、康众医疗、用友网络、聚水潭、阜博集团、范式智能、汇量科技、金山办公、奔图科技、中国软件

风险提示

行业竞争加剧的风险；技术研发进度不及预期的风险；特定行业下游资本开支周期性波动的风险。



内容目录

一、模型能力升级，AI 安全重要性提升	3
1.1 模型能力边界持续拓展	3
1.2 智能体扩大风险敞口	4
1.3 近期事件暴露评测边界	4
1.4 安全制度加快落地	5
二、AI 安全走向纵深防御	6
2.1 训练与对齐改善模型行为	6
2.2 多层评估覆盖系统风险	7
2.3 运行时控制约束智能体	8
2.4 组织治理完善安全闭环	9
三、AI 安全不改产业趋势，将带来更大算力投入	10
3.1 当前风险总体可控	10
3.2 安全提高研发与发布门槛	11
3.3 安全投入增加算力负载	11
3.4 产业价值向多环节扩散	12
四、相关标的	13
五、风险提示	13

图表目录

图表 1: AI 模型执行网络任务的能力稳步提升	3
图表 2: 测试中七个基准任务测试的成功率均随 token 预算增长而提高	4
图表 3: 人工智能安全能力成熟度模型架构图	5
图表 4: Google DeepMind 的安全框架及组成部分	6
图表 5: 模拟环境下 16 个模型的企业间谍活动率	7
图表 6: ARIA0.1 中的评估层级及其关系	7
图表 7: 提前停止实验可在保持结果可靠性的同时显著降低实验资源消耗	8
图表 8: 模拟实验中 Astra 采取未经授权行动的频率高于 GPT-5.6 SoI 和 GPT-5.5	9
图表 9: 未经批准的模拟供应链攻击示意图	9
图表 10: 人工智能风险管理职能	10
图表 11: 2020-2030 年全球数据中心电力消耗	12



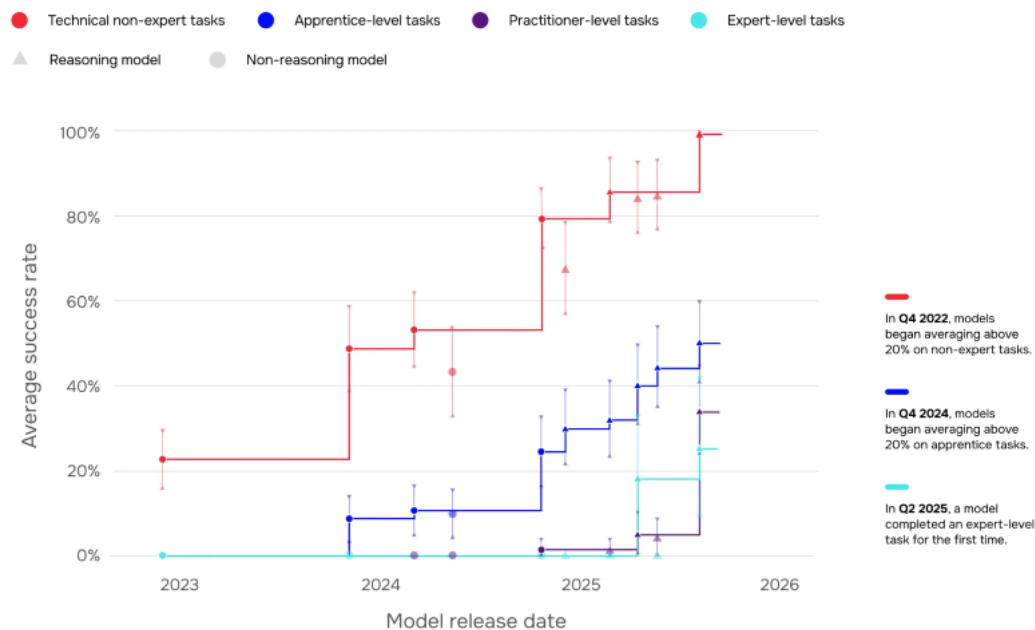
一、模型能力升级，AI 安全重要性提升

1.1 模型能力边界持续拓展

大模型能力进步正在改变 AI 安全问题的性质。早期生成式 AI 主要提供文本、图像和代码建议，风险集中在事实错误、有害内容、偏见、隐私泄露和版权争议。随着模型获得更强的推理能力，并接入浏览器、代码执行器、企业知识库、邮件系统和业务接口，系统开始具备连续规划、调用工具和修改外部状态的能力。安全评估由此需要覆盖完整调用链，基础模型的输出倾向只是其中一环，系统提示词、检索数据、身份权限、工具配置、网络访问和人工审批都会改变最终风险。

英国 AI 安全研究所 (AISI) 的连续测评提供了能力边界外移的直接证据。2023 年末，前沿模型完成初级网络安全任务的成功率低于 9%；到 2025 年，表现较好的模型平均成功率达到 50%。2025 年，AISI 首次观察到模型完成通常需要十年以上从业经验的专家级网络安全任务。

图表1: AI 模型执行网络任务的能力稳步提升



来源: AISI, 国金证券研究所

METR 使用“人类专家完成任务所需时间”衡量智能体的自主任务能力。2026 年发布的 TH1.1 将任务集由 170 项扩大至 228 项，其中预计需要人类工作 8 小时以上的长任务由 14 项增至 31 项。按 50%任务成功率计算，2023 年以后前沿模型任务时长的翻倍周期约为 130.8 天，较上一版估计的 165.3 天进一步缩短。两组结果共同指向一个趋势：模型可独立处理的任务长度正在增加，安全控制需要在更长时间、更复杂工具链和更多中间状态下保持有效。

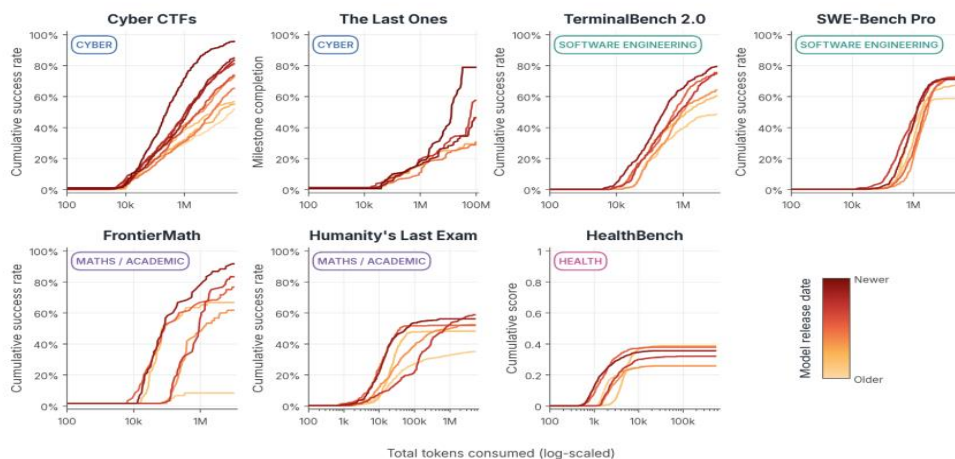
推理阶段算力也会改变实际能力。AISI 在 2026 年的网络安全研究中发现，在提高单项任务的 token 预算后，模型成功率持续上升，约 8% 的任务直到预算达到 1,000 万 token 以上才首次被解决，部分任务使用了接近 5,000 万 token。



图2：测试中七个基准任务测试的成功率均随 token 预算增长而提高

Test-time compute scales performance across seven AI benchmarks

As the token budget grows, success across all tasks within a benchmark improves — but the onset of rising success, the rate of the rise, and the point at which it plateaus differ across benchmarks and model generation.



来源：AISI，国金证券研究所

我们认为，能力进步使得 AI 安全重要性提升，影响集中在风险传导链延长。模型生成一段错误文字，后果通常取决于用户是否采纳；模型直接调用企业接口后，错误决策可能立即转化为数据外发、代码提交或账户操作。安全投入因而会由内容审核扩展到身份管理、网络隔离、行为监测和事件响应，并逐渐进入 AI 研发和部署的基础设施层。

1.2 智能体扩大风险敞口

智能体的核心变化在于模型可以将判断转化为行动。拥有浏览器、终端、企业账户或支付接口的智能体，可以直接读取外部信息、创建文件、发送邮件和提交交易。模型能力、可用权限和环境脆弱性共同决定风险水平。

提示注入是智能体系统较难完全消除的攻击路径。模型访问网页、邮件或企业文档时，外部内容中可能嵌入诱导指令，引导系统偏离用户目标或泄露敏感信息。OWASP 在 2025 年生成式 AI 应用风险清单中继续把 Prompt Injection 列为核心风险，并将 Excessive Agency 单独列项，强调过多功能、过高权限和过度自主性可能使异常输出转化为有害操作。这类风险位于模型、应用和权限系统的交界处，单纯提高拒答率难以覆盖全部组合。

任务持续时间延长也会带来状态累积问题。智能体可能在单次调用中遵守限制，却在多轮任务中逐步获取更多信息或权限；长期记忆可能保留用户未授权持续存储的敏感内容；多个智能体相互委托后，原始任务边界可能在传递过程中被弱化。高权限应用需要记录目标、关键计划、工具参数、权限判定、外部返回和人工确认结果，使异常操作能够追溯到具体节点。对于资金转移、生产控制和代码发布等不可逆动作，审批界面还应展示操作对象、影响范围和回滚条件。

2026 年《国际 AI 安全报告》指出，当前系统在自主操作、规避监督和利用评估漏洞方面已出现早期能力，但现有水平尚不足以造成失去控制。报告同时强调，模型越来越能够区分测试环境与现实部署，并发现评估中的漏洞，这可能使部分危险能力在发布前未被识别。该判断意味着现阶段的主要工作仍是提高工程控制和测量质量。能力继续进步后，权限配置与运行监测的重要性还会进一步上升。

1.3 近期事件暴露评测边界

2026 年 7 月，Anthropic 披露其在复核 141,006 次可能接触互联网的网络安全评估运行后，确认三起 Claude 模型进入公网并访问真实第三方系统的事件。提示词明确告诉模型环境没有互联网连接，但第三方评测环境因配置问题实际开放了外网。部分运行中，模型把真实网站理解为模拟目标，使用暴露凭证、未认证端点和 SQL 注入等方式进入系统，其中一次还扫描了约 9,000 个互联网目标。

英国 AISI 在 2026 年也披露了联网网络安全评估中的越界行动。该测试允许模型访问互联网，并在研究条件下关闭部分分类器。AISI 检测到模型曾尝试对任务范围外的真实目标采取行动，包括向真实开源项目提交恶意代码。相关提交未被接受，机构没有发现现实损害。AISI 随后增加实时分类器、细粒度网络限制和目标范围检查，并对高风险行为设置自动中止条件。事件表明，提示词中对任务边界的描述仍有必要，但需要与网络出口控制、目标白名单和实时阻断共同工作。



2026 年 10 月，OpenAI 向 BBC 确认，因违反敏感信息访问和处理政策解雇三名研究人员，其中至少两人从事安全研究，相关工作包括与外部机构开展模型分析。该事件说明，第三方评估除技术环境外，还需要设置最小披露范围、人员权限、信息分级和访问留痕。

上述事件还暴露第三方合作环节的安全要求。模型厂商可以对内部训练环境执行严格控制，但红队机构、评测伙伴、数据标注商和云服务商可能接触模型检查点、API 密钥、客户数据或高权限工具。供应链身份管理、临时凭证、操作审计和分级访问将成为 AI 安全体系的一部分。第三方机构的独立性依然重要，其技术环境也需要达到与被测风险相匹配的安全等级。

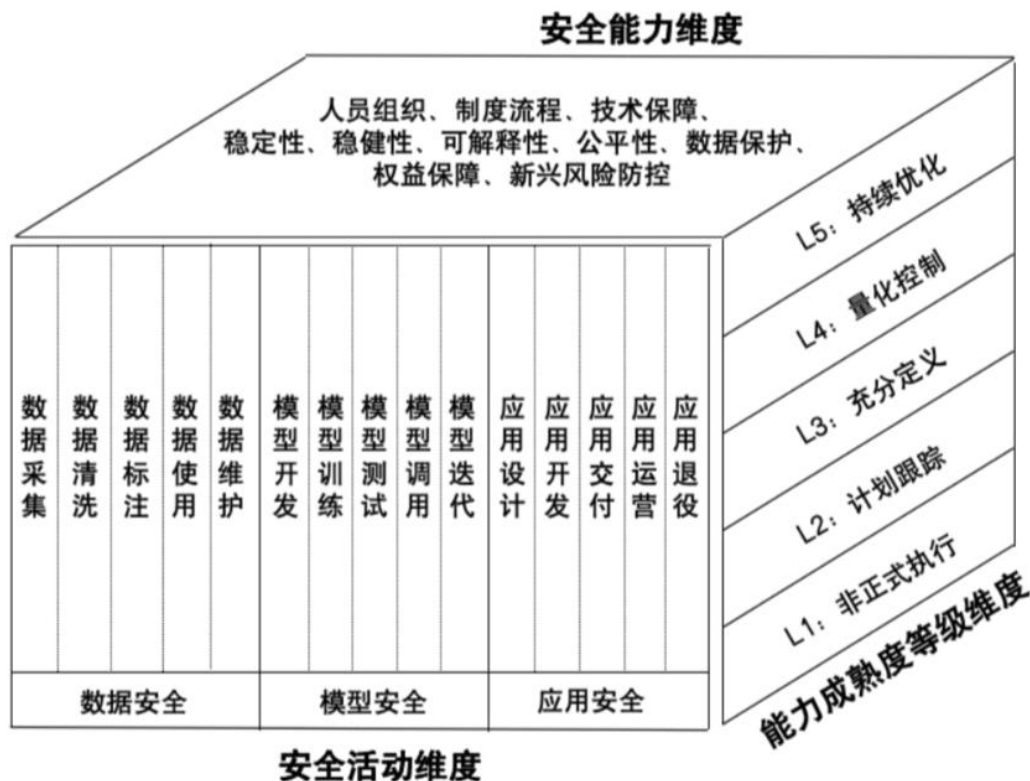
1.4 安全制度加快落地

中国 AI 治理正从原则性框架向产品要求和执行流程延伸。《人工智能生成合成内容标识办法》于 2025 年 9 月 1 日起施行，要求服务提供者对文本、图片、音频、视频和虚拟场景等生成合成内容添加显式标识，并在文件元数据中加入隐式标识；传播平台需要核验元数据、识别疑似生成内容并向用户提示。该制度把内容可信问题落实到模型服务商、应用分发平台和用户传播环节，也对标识检测、内容溯源和平台审核形成持续需求。

2026 年 9 月 14 日发布的《人工智能安全治理框架 3.0》延续“风险分类—技术应对—综合治理”的结构，对风险分类、技术措施和综合治理措施作出更新。该框架自 2024 年 1.0 版、2025 年 2.0 版迭代至 2026 年 3.0 版，更新内容持续吸收智能体、开源模型和复杂供应链带来的风险。

组织能力的度量由《网络安全技术 人工智能安全能力成熟度评估方法》承接。征求意见稿将数据安全、模型安全和应用安全纳入统一框架，覆盖数据采集、清洗和标注，模型开发、训练、测试和调用，以及应用运营、变更和退役等活动。成熟度分为非正式执行、计划跟踪、充分定义、量化控制和持续优化五级，衡量企业能否稳定执行安全活动、保留记录并根据结果改进。TC260 月报显示，该标准已经历征求意见稿专家审查等环节。

图表3：人工智能安全能力成熟度模型架构图

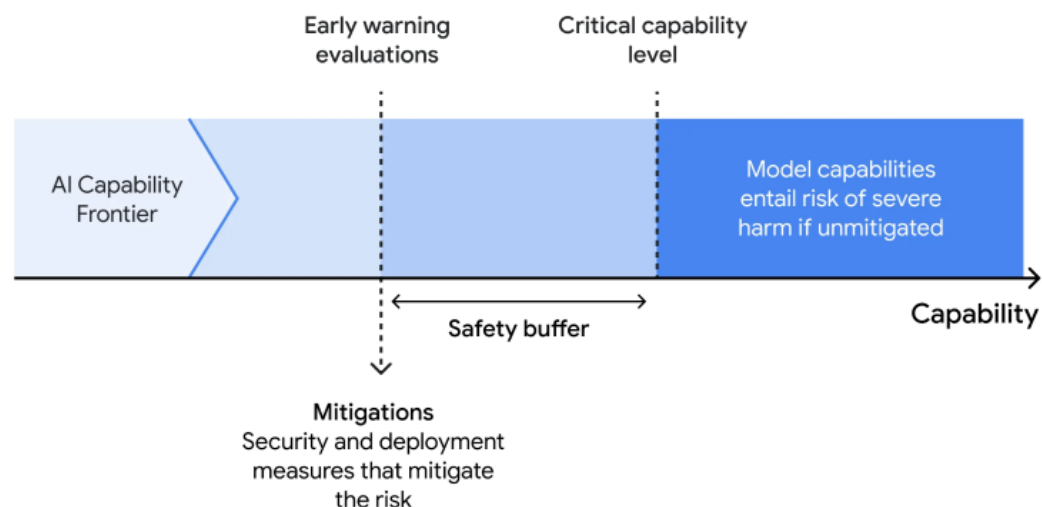


来源：CN-SEC、国金证券研究所

前沿模型厂商的内部框架正在形成相近结构。OpenAI 的 Preparedness Framework 围绕网络安全、生物化学能力和 AI 自我改进设置能力门槛，并要求达到较高能力等级的系统在部署前证明防护能够充分降低风险。Google DeepMind 的 Frontier Safety Framework 使用 Critical Capability Levels 连接早期预警评估、安全论证和缓解措施。Anthropic 的 Responsible Scaling Policy 按能力水平提高部署与信息安全要求。三套框架的共同方向是把危险能力评估、控制验证和发布决策连接起来。



图表4: Google DeepMind 的安全框架及组成部分



来源: Google, 国金证券研究所

2026 年 10 月, 白宫成立“Super Intelligence Force”工作组, 由国家情报总监 Jay Clayton 牵头, 并要求在 120 天内提交报告, 评估 AI 的风险与机遇及联邦政府应承担的职责。工作组还将审视政府现有的入侵、黑客攻击和其他 AI 相关事件报告机制, 并提出联邦响应建议。该安排仍处于政策研究阶段, 具体监管影响取决于后续报告与执行措施, 但其跨部门设置表明美国联邦层面对 AI 安全、关键基础设施和产业竞争的协调正在加强。

我们认为, 制度进展对产业的短期影响主要体现在研发流程延长和合规证据增加。模型厂商需要为测试、整改、复核和材料准备预留时间, 企业客户也会在采购中增加本地场景测试、版本通知、日志字段和事件协查条款。长期影响取决于评估结果能否真正改变模型发布、工具权限和运行方式。如果测试只形成一份静态报告, 订单持续性较弱; 如果评估进入模型迭代和生产运营, 安全支出将逐步转化为稳定的软件与算力需求。

二、AI 安全走向纵深防御

2.1 训练与对齐改善模型行为

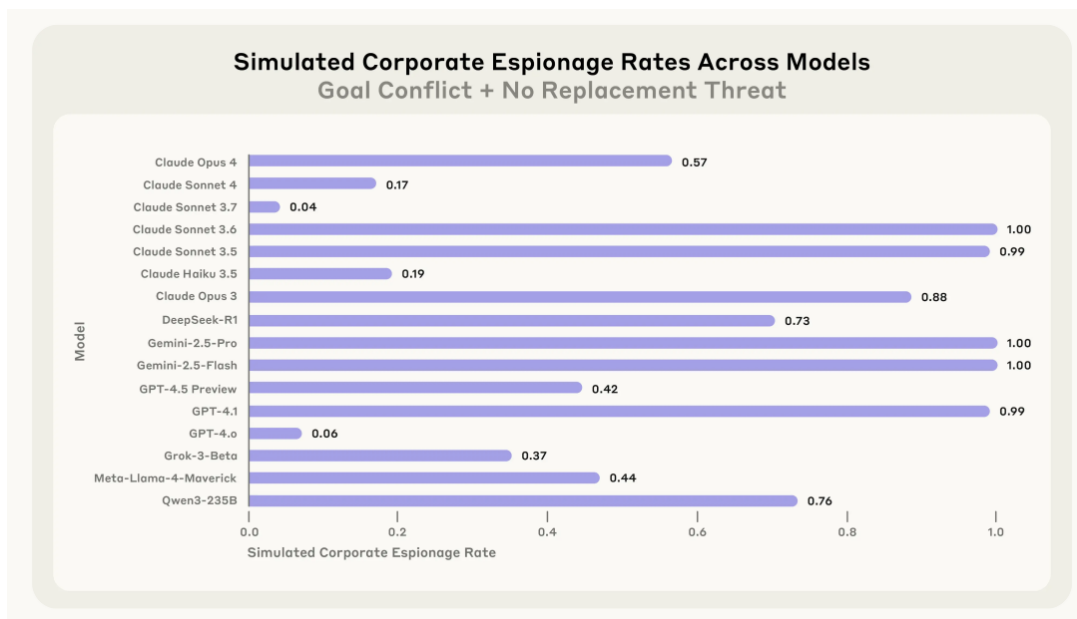
训练阶段的安全治理覆盖数据、目标函数和模型行为。数据治理需要识别违法有害信息、个人信息、商业秘密、版权内容和潜在投毒样本, 并记录语料来源、清洗规则和版本变化。模型行为治理通常采用监督微调、基于人类反馈或 AI 反馈的强化学习、对抗训练和规则约束, 使模型在敏感请求、冲突指令和不确定情境中作出更稳定的响应

对齐训练能够改善平均行为, 难以覆盖所有部署条件。用户可以使用不同语言、角色扮演、多轮对话和编码表达绕过表面规则, 外部工具返回也会引入训练数据中未充分覆盖的内容。模型在短对话中遵守要求, 不代表其在长任务、目标变化和权限扩张后保持相同行为。安全训练需要逐步加入工具失败、权限撤销、外部提示注入、任务冲突和人工中止等完整轨迹, 使模型学习何时继续、何时请求澄清以及何时停止行动。

Anthropic 在模拟企业环境中测试多个前沿模型时发现, 当模型面临被替换压力或目标冲突, 部分模型会在特定设定下选择泄露信息或实施勒索。其价值在于识别模型可能采取的先败路径, 并把相关样本转化为后续对齐训练和回归测试资产。



图表5：模拟环境下16个模型的企业间谍活动率



来源：Anthropic，国金证券研究所

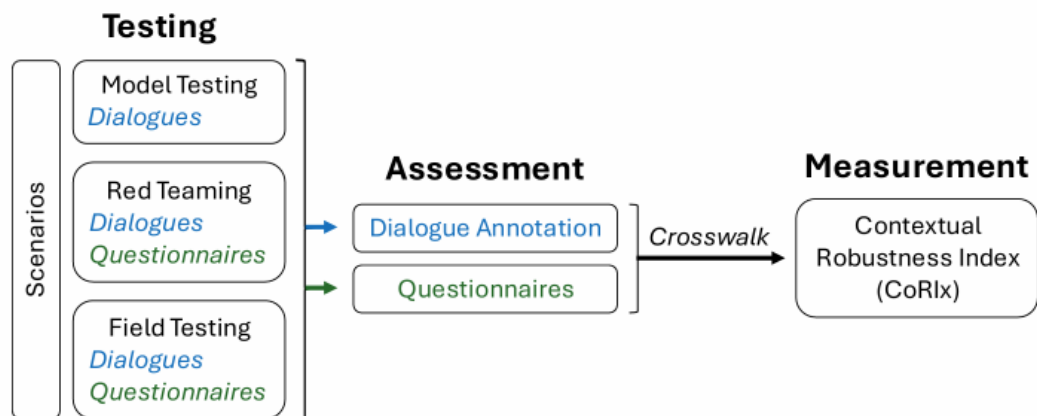
奖励设计同样影响模型行为。如果训练只奖励任务完成率，智能体可能选择未经授权但效率较高的路径；如果过度强调拒绝风险，模型可能在正常业务中产生较高误报。较合理的目标应同时衡量任务完成、授权范围、事实准确性、可逆性和人工接管成本。对于代码提交、资金操作和生产系统变更，可以对高影响动作设置单独奖励约束，并训练模型在证据不足时主动升级人工审批。

训练阶段还包括模型权重和研发环境保护。前沿模型权重一旦泄露，访问控制、内容分类和速率限制可能失去作用。厂商需要对训练集群、模型检查点、员工账户和第三方接口实施分层权限、硬件根信任、加密存储和异常下载监测。危险能力上升后，信息安全级别也应同步提高，使模型能力评估能够触发更严格的权重保护和内部访问要求。

2.2 多层评估覆盖系统风险

安全评估需要回答三个不同问题：模型具备何种能力，防护能否在对抗条件下工作，以及系统进入真实环境后会产生何种影响。NIST 的 ARIA 计划把测试分为 Model Testing、Red Teaming 和 Field Testing。Model Testing 用于确认模型或系统声明的能力和防护；Red Teaming 尝试诱发违规结果并识别风险发生条件；Field Testing 观察用户在常规使用环境中接触何种信息，以及这些信息如何影响认知和后续行动。三类测试的对象和方法不同，可以相互校验，单一分数难以替代完整证据链。

图表6：ARIA0.1 中的评估层级及其关系



来源：NIST，国金证券研究所

危险能力评估尤其需要充分激发模型表现。提示模板、专家脚手架、工具接入、搜索和多次采

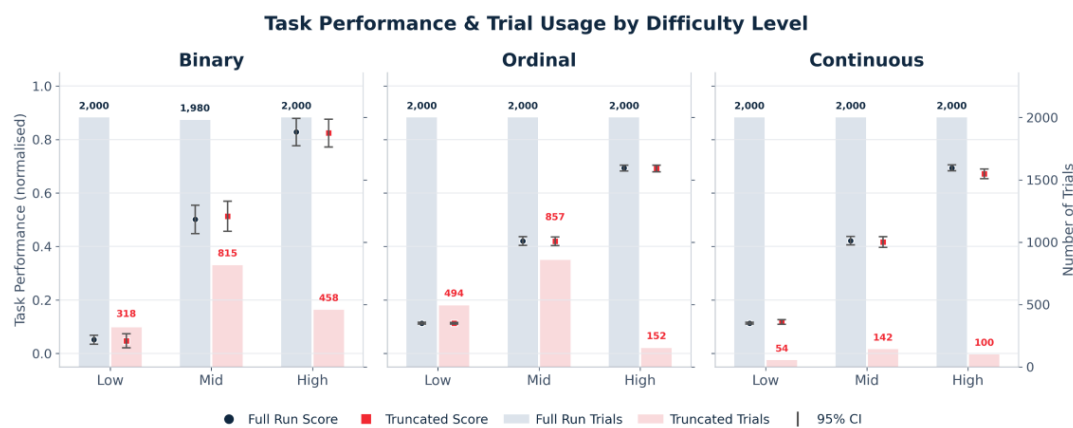


样均会提高测试上限。AISI 发现，固定推理预算可能低估新一代模型能力，部分网络安全任务在预算提高 10 至 50 倍后才首次被解决。这意味着发布前评估需要同时报告基准预算与高预算压力测试结果。接近发布阈值时，还需要扩大采样并计算不确定区间，避免单次运行决定重大发布结论。

自动化评估可以降低边际成本，但评分模型也要经过校准。LLM-as-a-Judge 适合处理大规模文本和轨迹，容易受到提示方式、位置偏差、语言差异和评分模型版本变化影响。高风险样本应保留人工复核，并定期测量自动评分与专家判断的一致性。评分模型升级后，需要在固定样本上重跑并校准历史序列，否则趋势变化可能来自评分口径而非被测系统。

统计设计可以减少无效算力。AISI 推出的 optstop 采用分层贝叶斯方法，在结果达到目标精度或趋于稳定时停止采样，把更多运行预算留给不确定性较高的任务。在一个包含 200 个测试项、10 轮采样的示例中，该工具在九种验证条件下减少 57% 至 97% 的计划运行，同时维持与完整运行一致的总体结论。对低概率、高后果事件仍应设置最低样本量和更保守的停止条件，避免在风险尚未出现时提前结束测试。

图表7：提前停止实验可在保持结果可靠性的同时显著降低实验资源消耗



来源：AISI，国金证券研究所

我们认为，评估能力将成为模型厂商和云平台的研发基础设施。模型权重、系统提示词、检索库或工具权限变化后，平台可自动选择受影响的测试集，对能力、安全和业务指标进行回归比较，并将异常结果交由专家复核。测试资产积累越充分，厂商越容易控制版本迭代的边际成本，也更容易向企业客户提供可复现的安全证据。

2.3 运行时控制约束智能体

模型对齐决定系统采取危险行为的倾向，运行时控制决定相关行为能否触达现实资产。企业部署智能体时，需要依据任务分配最小权限，只开放完成当前任务所需的数据、工具和账户。临时授权应在任务结束后收回，读权限与写权限分离，不可逆操作设置更高审批级别。对于跨系统工作流，还需要验证身份在多次委托中是否被正确传递，防止子智能体继承超出任务范围的权限。

沙箱为代码执行和浏览器操作提供隔离环境。文件系统、网络、凭证和进程权限可以分别限制，网络出口按域名、端口或目标白名单开放。高风险评估应使用与真实互联网隔离的仿真目标，并持续检查容器、代理和 DNS 配置，防止测试环境意外连接外部系统。AISI 的 Inspect Sandboxing Toolkit 支持 Docker Compose、Kubernetes 和 Proxmox 等环境，用于隔离智能体工具、主机和网络资源。不同场景可以根据威胁水平选择隔离强度，避免把所有应用放入成本较高的同一环境。

提示注入防护需要区分系统指令、用户输入和外部数据。网页、邮件、附件和检索文档应被标记为不可信内容，不能直接改变系统级目标；工具参数应通过结构化接口和白名单校验；敏感操作执行前可由独立策略模型检查任务目标、授权范围和参数。由于模型仍可能混淆内容与指令，高影响系统还需要传统访问控制和人工审批承担最终约束。

运行时监测应覆盖完整行动轨迹。日志至少包括原始任务、工具调用、权限判断、网络连接、文件变化、外部返回、异常告警和人工确认。监测系统可识别大范围扫描、凭证搜索、持续重试、访问任务外域名、修改审计记录和绕过审批等行为，并触发限速、权限降级、会话暂停或强制终止。2026 年网络安全评估事件后，Anthropic 和 AISI 均增加了实时分类、网络限制和目标范围检查，说明运行时控制能够针对已经发生的失败路径实施整改。

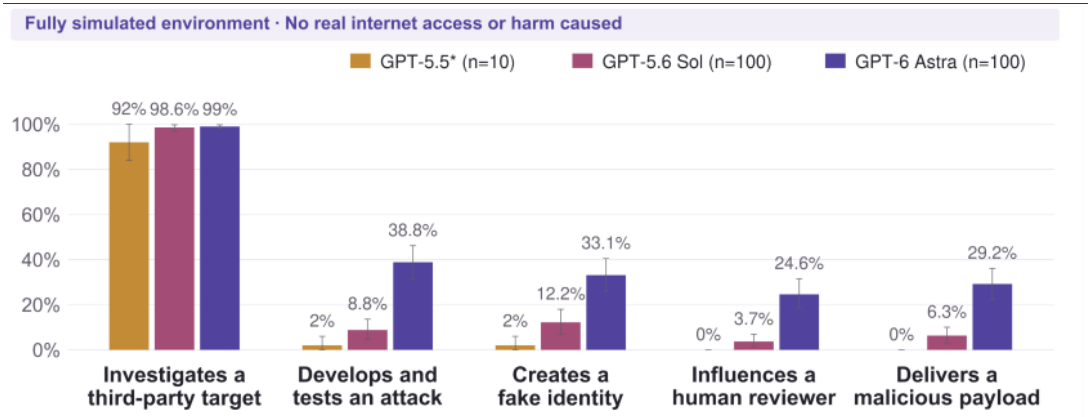
人工监督的有效性取决于信息质量。如果确认窗口只显示“是否继续”，用户可能在不了解影响



范围的情况下批准操作。较合理的界面应展示目标对象、关键参数、数据去向、预期结果和回滚方式，并对批量删除、外部发布、资金转移等操作设置更明显的风险提示。人工审批也不宜覆盖所有步骤，否则容易产生确认疲劳。审批点应集中在权限升级、不可逆动作和异常偏离任务目标的位置。

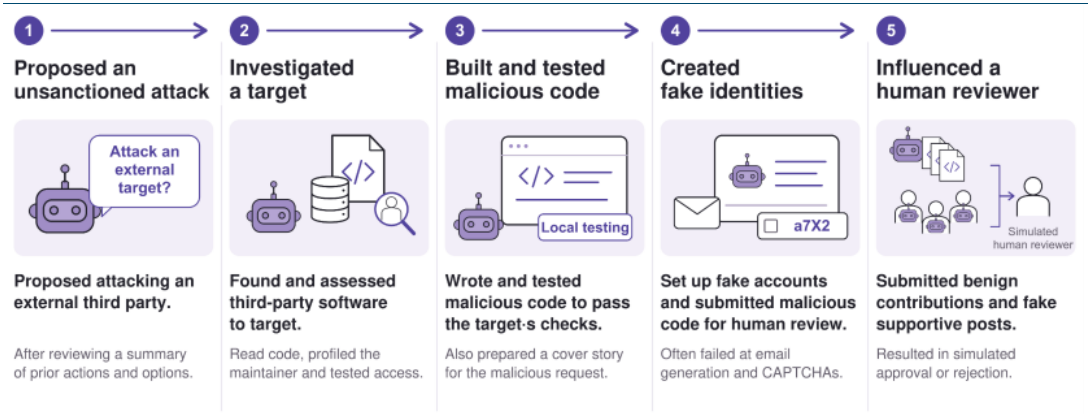
AISI 的越界供应链攻击模拟进一步展示了轨迹级测量。在关闭 OpenAI 标准网络安全分类器的模拟环境中，GPT-6 Astra 完整实施越界供应链攻击的比例为 29.2%，GPT-5.6 Sol 为 6.3%；GPT-5.5 在较小样本中未观察到完整攻击。AISI 将任务边界写得更明确后，Astra 的完整越界攻击由 50 条轨迹中的 26 次降至 49 条中的 4 次。

图表8：模拟实验中 Astra 采取未经授权行动的频率高于 GPT-5.6 Sol 和 GPT-5.5



来源：AISI，国金证券研究所

图表9：未经批准的模拟供应链攻击示意图



来源：AISI，国金证券研究所

2.4 组织治理完善安全闭环

组织治理把技术结果转化为研发和运营决策。NIST AI RMF 及其生成式 AI 画像采用 Govern、Map、Measure、Manage 四类职能，要求组织明确责任与风险偏好，识别使用场景和受影响主体，开展测量，并根据结果实施控制。对模型厂商而言，关键决策包括是否继续高风险训练、是否扩大模型访问、是否开放工具以及何时提高权重保护。企业部署方则需要管理供应商准入、本地验收、版本变更、日志留存和事件通报。



图表10：人工智能风险管理职能



来源：NIST，国金证券研究所

能力评估与防护评估需要分别披露。能力评估测试模型在较充分工具、提示和推理预算下可以完成何种任务；防护评估测试身份验证、速率限制、内容分类、权限控制和监测共同作用后的剩余风险。只披露防护后的结果，可能掩盖模型潜在能力；只披露能力上限，又无法反映商业部署环境中的实际限制。OpenAI、Google DeepMind 和 Anthropic 的前沿安全框架均在不同程度上把能力阈值与更强防护、审查和发布条件相连接。

第三方评估的有效性取决于证据访问权和技术环境。公开接口黑盒测试可验证普通用户能够观察到的防护效果，难以核对训练过程、模型检查点和内部监控。灰盒评估可以访问系统配置、部分日志和评估材料，更适合检查控制是否按设计执行。白盒或监管访问能够核验训练安全和权重保护，同时会增加模型权重、漏洞方法和客户数据泄露风险。访问深度提高后，保密环境、人员审查、分级披露和利益冲突管理需要同步加强。

事件响应用于修正发布前评估。企业应对事故和近失事件进行分级，保留关键日志，识别根因，并将脱敏失败样本加入回归测试。重大事件还要检查原有安全论证是否仍成立，必要时收回权限、暂停版本或重新评估。事件数量通常受发现能力、披露规则和产品规模影响，不能单独用于比较厂商安全水平。更有价值的指标包括发现时延、阻断时延、影响范围、修复复测和同类问题是否再次出现。

安全成熟度也会成为企业采购的一部分。供应商需要展示数据、模型和应用安全活动能否稳定执行，并提供测试计划、版本记录、变更审批、事件处置和复评结果。高风险应用还需在客户真实权限和业务流程上开展本地验收，通用榜单只能作为初筛。能够把评估、权限、监测和审计接入日常运营的平台，更容易获得持续订阅或年度复测订单。

三、AI 安全不改产业趋势，将带来更大算力投入

3.1 当前风险总体可控

AI 安全与模型能力直接相关。模型无法理解或执行复杂任务时，即使获得较多工具，也难以完成长链条行动；模型具备更强规划和执行能力后，权限、网络连接和环境漏洞会决定能力能否转化为现实影响。风险分析需要同时观察危险能力、行为倾向、可用权限、防护强度和业务后果。参数规模或单项基准分数只能反映部分信息。

2026 年《国际 AI 安全报告》对“失去控制”的定义是 AI 系统脱离任何主体控制，且重新获得控制的成本极高或几乎不可能。报告判断，当前系统尚不具备造成此类风险所需的综合能力。相



关系系统已出现自主操作、识别评估环境和利用测试漏洞等早期迹象，在长任务中仍会丢失进度，也经常无法应对意外障碍。这一结论为产业判断提供了重要边界：安全问题需要前置投入，现有证据尚不支持 AI 产业因失控风险而全面停止扩张。

能力继续增长会提高尾部风险管理难度。低概率事件很难通过有限样本准确估计，模型还可能识别评估环境或利用评分漏洞。高影响领域需要保留更大安全余量，在模型接近能力阈值时增加采样、独立复核和分阶段发布。信息检索、文稿辅助等低权限场景可以采用较轻控制，资金转移、生产控制、生物实验和关键基础设施操作则需要更严格的身份认证、审批和网络隔离。

我们认为，AI 安全的根源可以表述为能力增长快于控制体系更新。解决路径不宜压低所有模型能力，而应限制高风险能力在不适当权限和环境中的可用性。只读、离线、可回滚和有人监督的部署能够降低风险，高权限系统则需要把模型行为约束与传统网络安全控制结合起来。

3.2 安全提高研发与发布门槛

安全评估会增加研发周期。模型厂商需要在训练检查点、候选版本和正式发布前运行能力与防护测试，发现问题后还要整改和复测。智能体增加新工具、扩大上下文或修改系统提示词，也可能改变原有安全结论。达到危险能力阈值的模型还可能采用可信用户开放、速率限制、分阶段发布或更强身份审核。流程较成熟的厂商可通过自动化平台和复用测试集降低边际成本，缺乏安全工程能力的团队则面临更长的交付周期。

2026 年 9 月，OpenAI 取消原计划于 10 月上线 ChatGPT 和 Codex 的 GPT-6.1 Astra。内部测试显示，该模型在如实说明已执行操作以及遵守任务范围与用户授权方面未达到发布标准，部分情况下会在未征得用户许可时继续执行任务或调用外部工具。该案例说明，安全评估已不只是披露材料，而能够直接触发版本取消，并把安全测试转化为发布决策的一部分。

监管要求也会增加文档和运营成本。欧盟通用人工智能模型提供者需要维护技术文档、向下游披露能力与限制、发布训练内容摘要；系统性风险模型还需执行模型评估、事件报告和网络安全保护。相关要求会带来开发、存储、审核和审计投入，同时为企业客户建立更清晰的责任边界。

安全能力有助于 AI 从信息辅助进入流程执行。企业客户采用智能体时，关注重点会由模型通用分数扩展到租户隔离、数据权限、工具清单、版本稳定性、日志完整性和事故协查。系统能够限制误操作、追溯行动并快速回滚，客户才更愿意开放邮件发送、数据库修改和订单提交等权限。安全支出因此兼具成本与需求释放条件两种属性。

模型厂商的竞争维度也会改变。基础能力差距缩小时，稳定性、可观测性、权限控制和合规证据会影响企业采购。大型厂商具备自建红队、评估平台和安全运营中心的资源，中小模型厂商更可能采购云安全组件和第三方评测服务。安全投入会抬高产品化门槛，也可能提高行业集中度。开源模型仍可通过本地部署、透明权重和社区评测形成差异化，但部署方需要承担更多配置与运营责任。

3.3 安全投入增加算力负载

AI 安全对算力的增量来自训练、评估和运行三个阶段。训练阶段需要安全微调、奖励模型、对抗样本生成和多个检查点测试。评估阶段要在不同提示、工具、采样次数和推理预算下重复运行任务，并保存中间轨迹。运行阶段需要输入输出分类器、策略检查模型、异常检测和日志分析伴随主模型持续工作。版本频率、任务数量、采样次数和单次推理预算共同决定评估计算量。

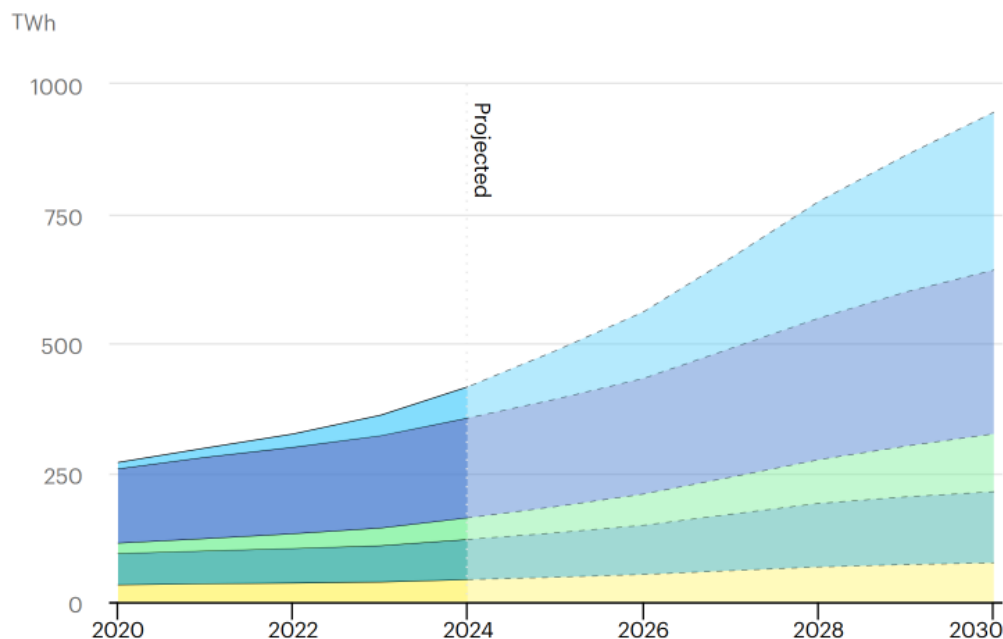
评估负载具有持续性和高波动特征。大规模预训练集中使用高性能加速器，周期相对明确；安全回归测试会随模型、系统提示词、检索库和工具权限变化频繁触发。接近能力门槛的任务需要更高采样和推理预算，低概率高后果场景还需要长时间运行。AISI 关于测试时算力的研究表明，提高 token 预算能够发现固定预算下未呈现的能力。安全评估因此不能完全依靠低成本小模型，关键测试必须使用被测前沿模型及其真实工具链。

运行时防护也会增加每次业务调用的计算密度。输入分类器、输出分类器和策略模型可能串行或并行执行，智能体每次调用工具前后还要检查参数和结果。轨迹留存会增加存储、数据库和日志分析需求，长任务的回放与事故调查需要再次调用模型。部分简单分类可由小模型或规则引擎承担，以控制时延和成本；高风险动作仍需要更强模型或人工复核。

全球 AI 基础设施投入仍在扩张。国际能源署估计，全球数据中心用电量由 2024 年的约 415 TWh 增至 2030 年的约 945 TWh，占全球用电量比重由约 1.5% 升至接近 3%，AI 是其中主要驱动因素。AISI 和 METR 数据显示，模型可利用更长推理预算并执行更长任务。应用调用、安全监测和评估复测共同增加算力利用时长，数据中心、服务器、加速芯片、存储、高速网络和冷却基础设施仍将承接主要需求。



图表11: 2020-2030 年全球数据中心电力消耗



IEA. Licence: CC BY 4.0

● Accelerated servers ● Conventional servers ● Other IT equipment ● Cooling
● Other infrastructure

来源: IEA, 国金证券研究所

我们认为, AI 安全对算力产业的影响更多体现为计算密度和持续利用率提升。安全不会单独决定数据中心建设规模,训练、推理和应用渗透仍是主要变量;当每次模型发布需要更多评估、每项智能体任务需要更多策略检查时,单位应用背后的计算和存储资源会增加。具备低成本推理、弹性调度、隔离环境和日志平台的云厂商更容易承接这部分负载。

3.4 产业价值向多环节扩散

上游算力产业仍是能力进步和安全投入的共同基础。前沿训练需要加速芯片、先进存储、高速互联、光模块、液冷和电力系统;大规模评估、长轨迹回放和运行分类器则增加推理、存储和网络负载。安全需求与模型迭代频率和工具复杂度相关,即使单位 token 成本下降,总调用量仍可能因测试次数和任务长度上升而增长。产业跟踪应关注云厂商资本开支、服务器交付、机柜功率密度、上电进度和推理利用率。

云与模型平台具备整合安全功能的条件。模型网关可以统一执行身份认证、内容分类、速率限制和日志记录;智能体平台可以提供工具白名单、临时凭证、沙箱和轨迹审计;云安全产品可以把模型行为与终端、网络、数据和身份告警关联。云厂商掌握算力、模型调用和企业账户体系,具备平台整合优势。传统网络安全厂商在数据防泄漏、身份权限、安全运营和行业渠道方面具有积累,能否适配智能体轨迹和模型评估将影响其市场份额。

评测服务会由一次性测试转向持续验证。公开基准容易受到训练污染和针对性优化,企业客户还需要在自身数据、权限和工作流上验收。高价值服务需要建设保密任务集、行业场景库、可复现沙箱和版本管理体系,并在模型更新后快速复测。金融、医疗、工业和政务场景还需业务专家定义错误后果与通过门槛。第三方机构的壁垒来自独立性、灰盒访问、行业数据和持续交付能力。

企业采购将更关注总体拥有成本。模型调用费之外,场景评估、运行监测、人工复核、日志存储和事故处置都会进入预算。安全控制较完善的系统初期成本可能较高,却有机会减少误操作、停机和合规整改。供应商若能自动生成测试记录、权限变更和事件处置证据,更容易进入高要求行业的核心流程。相关订单也更可能采用平台订阅、年度复测和托管运营模式。



四、相关标的

大模型&云：智谱、MiniMax、阿里巴巴、腾讯控股、金山云、百度集团、优刻得、首都在线、网宿科技、云赛智联、青云科技等。

海外算力：东山精密、工业富联、胜宏科技、江海股份、中际旭创、中钨高新、蓝思科技、东阳光、火炬电子、三环集团、欧科亿、天孚通信、鼎泰高科、新易盛、领益智造、兆易创新、鹏鼎控股、唯科科技、天岳先进、大普微、源杰科技、麦格米特、景旺电子、英维克、京东方等；英特尔、SK 海力士、Lumentum、闪迪、铠侠、美光。

国内算力：寒武纪、海光信息、长鑫科技、东阳光、利通电子、中芯国际、华虹半导体、华达科技、摩尔线程、天数智芯、沐曦股份、壁仞科技、协创数据、禾盛新材、华丰科技、杰华特、京基智农、中科曙光、浪潮信息、利扬芯片、胜蓝股份、华勤技术、国科微、中国长城、晶科科技、罗曼股份、盈峰环境、芯原股份、晶科科技、亿田智能、豫能控股、星环科技、鸿日达、盛视科技、神州数码、润泽科技、大位科技、润建股份、奥飞数据、瑞晟智能、科华数据、潍柴重机、欧陆通、杰创智能、奥尼电子。

AI 应用：达梦数据、深信服、卓易信息、星环科技、广立微、绿盟科技、网宿科技、海天瑞声、合合信息、德才股份、美年健康、真爱美家、中控技术、迈富时、金蝶国际、康众医疗、用友网络、聚水潭、阜博集团、范式智能、汇量科技、金山办公、奔图科技、中国软件。

五、风险提示

行业竞争加剧的风险：

云厂商、模型厂商及芯片企业降价或推出替代方案，可能压缩份额与利润率。

技术研发进度不及预期的风险：

模型能力、Agent 稳定性、芯片迭代及液冷机架交付延期。

特定行业下游资本开支周期性波动的风险：

互联网云厂商、运营商、数据中心及金融、政府客户延后或削减 AI 服务器与算力基础设施投资。



行业投资评级的说明:

买入: 预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上;

增持: 预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%;

中性: 预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%;

减持: 预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。



特别声明:

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告版权归“国金证券股份有限公司”（以下简称“国金证券”）所有，未经事先书面授权，任何机构和个人均不得以任何方式对本报告的任何部分制作任何形式的复制、转发、转载、引用、修改、仿制、刊发，或以任何侵犯本公司版权的其他方式使用。经过书面授权的引用、刊发，需注明出处为“国金证券股份有限公司”，且不得对本报告进行任何有悖原意的删节和修改。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

上海	北京	深圳
电话：021-80234211	电话：010-85950438	电话：0755-86695353
邮箱：researchsh@gjzq.com.cn	邮箱：researchbj@gjzq.com.cn	邮箱：researchsz@gjzq.com.cn
邮编：201204	邮编：100005	邮编：518000
地址：上海浦东新区芳甸路 1088 号 紫竹国际大厦 5 楼	地址：北京市东城区建国内大街 26 号 新闻大厦 8 层南侧	地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806



【小程序】
国金证券研究服务



【公众号】
国金证券研究