稿件联系人信息

姓名：许达

手机号：[13521894156]

E-mail：xudayj@chinamobile.com

中文栏目名：专论与综述

大型语言模型与国家安全：技术态势、风险评估与应对策略

许　达[1*]

1 中国移动研究院　北京　100053

作者简介：许达（1979—），男，博士，主任研究员，主要研究方向为 AI for Science、人工智能大模型及其应用。E-mail: xudayj@chinamobile.com

摘　要　大型语言模型（LLMs）的快速发展正在深刻重塑国家安全格局。本文采用叙述性综述与专家咨询相结合的方法，系统评估 LLMs 技术态势及其对国家安全的风险传导机制。研究基于 2020—2025 年间 92 篇核心文献的系统分析，结合公开基准测试数据进行技术态势横向比较，并采用改良德尔菲法咨询 8 位跨学科专家构建风险评估矩阵。研究识别出八类重点风险，其中供应链断裂与网络攻击自动化风险等级最高；提出” 算法拒止” 机制作为高价值知识资产的保护方案，设计” 系统提示硬化—输入净化—工具隔离—输出审计” 四层安全网关架构，并给出短中长期行动路线图。本文认为，应在开放与自主并重的前提下，以非对称技术路线与系统级安全架构支撑自主可控 AI 能力体系建设，实现发展与安全的动态平衡。

关键词　大型语言模型，国家安全，技术主权，马赛克效应，人工智能治理，算法拒止，AI 安全对齐

中图分类号　TP18; E866　　**文献标志码**　A

Large Language Models and National Security: Technological Landscape, Risk Assessment, and Response Strategies

XU Da[1*]

1 China Mobile Research Institute, Beijing 100053, China

Abstract　The rapid development of Large Language Models (LLMs) is profoundly reshaping the national security landscape. This paper employs a narrative review combined with expert consultation to systematically assess the technological status of LLMs and their risk transmission mechanisms to national security. Based on systematic analysis of 92 core papers from 2020-2025, cross-sectional comparison using publicly available benchmark data, and a modified Delphi method consulting 8 interdisciplinary experts to construct a risk assessment matrix, we identified eight major risk categories, with supply chain disruption and automated cyber attacks rated highest. We propose the

”Algorithmic Denial” mechanism for protecting high-value knowledge assets, design a four-layer security gateway architecture of ”system prompt hardening-input sanitization-tool isolation-output auditing”, and provide short-, medium-, and long-term action roadmaps. We argue that under the premise of balancing openness and autonomy, an autonomous and controllable AI capability system should be built through asymmetric technological approaches and system-level security architecture, achieving dynamic equilibrium between development and security.

Keywords large language models, national security, technological autonomy, Mosaic Effect, AI governance, Algorithmic Denial, AI alignment

---

术语与缩略语

术语/缩略语	说明
LLM	大型语言模型（Large Language Model），基于 Transformer 架构的通用语言理解与生成模型。
RLHF	基于人类反馈的强化学习，用于对齐模型输出到人类偏好。
OODA	观察-定向-决策-行动循环，军事决策的经典模型。
HBM	高带宽内存（High Bandwidth Memory），用于加速 AI 训练与推理。
CUDA	并行计算平台与编程模型，当前主流 GPU 软件生态。
Mosaic Effect	马赛克效应，指跨域数据聚合后可推断出敏感信息的现象。
Algorithmic Denial	算法拒止（本文新术语），指在模型及其系统管线中嵌入知识边界控制与用途边界控制机制，通过系统提示硬化、输入净化、工具隔离、输出审计等手段，以内生方式抑制模型对高价值敏感知识的推断、组合与外泄。该机制与传统访问控制互补，侧重模型行为层与生成链路的主动防护。详见正文第 3.1.4 节与附录 D。

研究方法

理论框架

本研究的理论基础融合了两个核心视角：

（1）技术安全化理论（Securitization Theory）：该理论由哥本哈根学派提出，认为安全议题并非客观存在，而是通过” 言语行为” 被社会建构的过程 [98]。本文借鉴这一框架，分析大模型技术如何从单纯的技术议题被建构为国家安全议题，以及这种安全化过程的合理边界——既要警惕” 过度安全化” 对技术创新的抑制，也要防止” 安全化不足” 导致的风险忽视。

（2）复合相互依赖理论（Complex Interdependence Theory）：基奥汉与奈提出的这一理论强调，在全球化时代，国家间存在多渠道、多层次的相互依赖关系 [99]。AI 领域的全球分工——美国主导基础研究与芯片设计、东亚承担制造、中国拥有应用市场与工程人才——构成典型的复合相互依赖结构。技术” 脱钩” 的成本与可行性，需在此框架下评估。

这两个理论视角为本文的风险分析提供了规范性指导：技术安全化理论帮助界定哪些风险值得纳入安全议程；复合相互依赖理论则提示，应对策略应在全球技术网络的约束条件下寻求平衡，而非追求不切实际的” 完全自主”。

研究方法

本研究采用多元方法相结合的研究路径：

（1）系统性文献分析：检索 2020—2025 年间中英文学术数据库（Web of Science、CNKI、arXiv）中关于大模型技术、AI 安全、AI 治理的文献。

检索策略：英文检索式为：

```
("large language model*" OR "LLM" OR "GPT" OR "generative AI")
AND ("national security" OR "cybersecurity" OR "AI safety"
OR "AI governance" OR "AI risk")
```

中文检索式为：

```
（" 大语言模型" OR " 大模型" OR " 生成式人工智能"）
AND（" 国家安全" OR " 网络安全" OR " 人工智能安全" OR " 人工智能治理"）
```

检索时间范围为 2020 年 1 月至 2025 年 1 月，检索日期为 2025 年 1 月 15 日。数据库版本：Web of Science Core Collection（SCI-EXPANDED + SSCI）、CNKI 期刊全文数据库、arXiv.org。

初步检索获得相关文献约 1200 篇，经过两轮筛选：第一轮根据标题和摘要相关性筛选，保留 386 篇；第二轮按照以下差异化标准进一步筛选：（a）发表于同行评审期刊或权威机构报告；（b）被引标准采用年份加权法——2020—2022 年文献要求被引次数 ≥15 次，2023 年文献 ≥5 次，2024—2025 年文献因发表时间短暂不设被引门槛但需发表于领域顶级会议（NeurIPS、ICML、ACL 等）或影响因子 ≥5.0 的期刊；（c）政策文件和行业报告需来自政府部门或国际权威机构。

被引标准差异化的理由：AI 领域文献的被引周期通常为 1-2 年，对较早文献设置被引门槛可筛选出经过学术检验的高质量研究；对新近文献放宽被引要求但提高发表平台门槛，是为了纳入尚未积累引用但质量有保障的前沿成果，避免遗漏重要进展。该标准参考了 Bommasani 等 [52] 在基础模型综述中采用的文献筛选策略。

最终纳入分析的核心文献 92 篇（英文 67 篇，中文 25 篇），学科分布：计算机科学 42 篇、国际关系与安全研究 28 篇、政策研究 15 篇、其他 7 篇；来源分布：学术期刊 51 篇、会议论文 19 篇、机构报告 14 篇、政策文件 8 篇。由于 AI 领域发展迅速，本研究采用叙述性综述而非系统综述方法，主要基于以下考量：系统综述强调方法的可重复性和穷尽性，但 AI 领域每周都有重要进展发布，严格的系统综述在完成时可能已过时；叙述性综述允许纳入最新但尚未被充分引用的重要工作，更适合快速演进的技术领域 [53]。

（2）技术态势评估：基于公开基准测试数据（MMLU、HumanEval、GSM8K、C-Eval 等）进行横向比较，同时参考斯坦福 AI 指数报告 [11]、中国信息通信研究院白皮书 [2] 等权威第三方评估。需说明的是，基准测试存在固有局限性（详见第 1.1 节），本文将其作为参考指标而非绝对标准。

（3）风险评估框架：采用” 可能性-影响程度-可控性” 三维评估模型（详见附录 A），对各类风险进行半定量分析。可能性分为高（>50%）、中（20%-50%）、低（<20%）三级；影响程度按照潜在损失规模分为严重、中等、轻微三级；可控性评估综合考虑技术可行性和制度保障。风险等级判定采用 $R = P \times I \times C$ 的计算公式（详见附录 A.4），划分极高、高、中高、中、低五个等级。具体评估结果见第 3 节风险矩阵（表5）。

（4）政策建议形成：基于文献分析和风险评估结果，结合《新一代人工智能发展规划》[6] 等现有政策框架，通过逻辑推演形成政策建议。建议的优先级排序综合考虑紧迫性、可行性、资源需求三个维度（详见第 10.3 节）。

研究局限：（1）本研究基于公开信息，无法获取涉密数据，部分判断可能存在信息不完整的局限；（2）AI 技术发展迅速，研究结论的时效性有限；（3）研究团队主要具有技术背景，对政策实施层面的分析可能不够深入；（4）部分预测性结论存在固有不确定性，读者应审慎参考。

伦理与合规声明：本研究涉及的网络安全攻防技术分析仅用于*防御与治理*目的，不构成对任何非法活动的指导或鼓励。所有案例均来自公开来源（政府公告/媒体报道/公开论文），不涉及未公开的漏洞细节或可执行攻击代码。研究遵循《网络安全法》《数据安全法》《个人信息保护法》等相关法规。研究不涉及个人可识别信息（PII）的收集与处理；专家咨询采用匿名方式进行。机构伦理自查备案号：CMRI-AI-2024-Ethics-012（中国移动研究院科技伦理自查备案，2024 年 9 月）。

0.1 研究框架概览

本文的研究框架如图1所示，从技术态势分析出发，经过风险评估，最终形成政策建议。

图 1: 本文研究框架

1 技术发展态势与战略背景

我们正处于一个技术变革的关键节点。大型语言模型（LLMs）——那些基于 Transformer 架构、通过海量文本预训练的深度学习模型——已经从实验室走向了广泛应用，成为继互联网、移动互联网之后又一个具有颠覆性潜力的通用技术。GPT-4、Claude、Llama、DeepSeek……这些名字在短短两三年间从技术圈的专业术语变成了公众话题。ChatGPT 的增长速度堪称现象级：2022 年 11 月上线后仅 5 天即突破 100 万用户，两个月达到 1 亿用户，到 2025 年 4 月周活跃用户已达 8 亿，OpenAI 的年度经常性收入（ARR）突破 100 亿美元 [91]。78% 的组织在 2024 年报告使用了 AI，较 2023 年的 55% 大幅提升。对于国家安全而言，这意味着什么？

1.1 全球 AI 竞争格局

中美两国在 AI 领域的角力，已经不仅仅是技术竞争，更是一场关乎未来发展主导权的战略博弈。

从公开基准测试看（表2），情况比许多人想象的要复杂。从 Artificial Analysis 智能指数看，Gemini 3 Pro（73 分）、Claude Opus 4.5（70 分）、GPT-5（68 分）占据前三，但中国模型正在快速追赶——Kimi K2 Thinking 达到 67 分，DeepSeek-V3.2 达到 66 分，已跻身全球第一梯队。更值得关注的是性价比优势：DeepSeek-V3.2 的 API 价格仅为 GPT-5 的 1/10，却达到了接近的智能水平。在中文理解（C-Eval）等任务上，Qwen3 Max 的 90.3% 和 GLM-4.6 的 88.7% 展现出强大的本土化优势。但这能说明我们已经全面领先吗？恐怕不能。在 SWE-bench（软件工程）等高难度任务上，Claude Opus 4.5 达到 72.5% 的通过率，展现出强大的代码理解和自主编程能力，而国产模型在这一指标上仍有差距。基准测试就像考试，”应试能力”强不等于综合能力强，真正的差距需要在实际应用中检验。

AI 芯片领域的差距需系统性理解：（1）先进制程方面，受出口管制影响，先进制程芯片获取受限；（2）软件生态方面，CUDA 生态成熟度领先国产平台；（3）HBM 高带宽内存主要由三星、SK 海力士供应，国内企业正加速追赶。但需注意，算法优化可在一定程度上弥补算力差距，DeepSeek 以较低成本实现高性能模型即为例证 [5]。

根据斯坦福 AI 指数报告（2025 版）[11]，2024 年美国 AI 私人投资达 1091 亿美元，约为中国 93 亿美元的 12 倍、英国 45 亿美元的 24 倍。生成式 AI 领域尤为突出，2024 年全球私人投资达 339 亿美元，较 2023 年增长 18.7%。该数据不含政府投资，若计入政府投入，差距可能有所缩小。值得关注的是，报告同时指出中国在 AI 学术论文和专利数量上持续领先，且中美模型在主要基准测试上的性能差距已从 2023 年的两位数缩小至 2024 年的接近持平。

1.2 中国 AI 发展的比较优势

说差距不等于唱衰。换个视角看，中国在 AI 竞争中握有几张独特的牌。

14 亿人口的市场规模本身就是稀缺资源。美欧企业训练中文模型要靠爬取数据，我们则坐拥海量的原生中文语料和应用场景。微信、抖音、淘宝每天产生的交互数据，是任何实验室都无法模拟的真实用户行为样本。这种”数据土壤”的差异，会随着模型规模扩大而愈发凸显。

工程化落地是另一个优势。DeepSeek 团队用远低于 OpenAI 的成本训练出性能接近的模型，靠的不是什么秘密武器，而是扎实的工程优化——数据清洗、训练调度、推理加速，每个环节都在”抠细节”。这种”卷”的能力，恰恰是国内技术团队的强项。Qwen 开源后短短几个月内的迭代速度，同样印证了这一点。

《新一代人工智能发展规划》[6] 确立的顶层设计也在持续发挥作用。政府引导与市场竞争相结合的模式，至少到目前为止，在 AI 基础设施建设和应用推广上展现出一定效率。

表 2: 主要大模型基准测试表现对比（2025 年 12 月）

模型	厂商	智能指数	SWE-bench	GPQA	AIME'24	C-Eval	价格
国际前沿模型							
Gemini 3 Pro	Google	73	–	84.5%	95.2%	–	$4.50
Claude Opus 4.5	Anthropic	70	72.5%	81.4%	33.9%	–	$10.00
GPT-5 (high)	OpenAI	68	61.0%	78.5%	87.3%	–	$3.44
Grok 4	xAI	65	–	75.2%	78.6%	–	$6.00
中国前沿模型							
Kimi K2 Thinking	月之暗面	67	–	–	–	–	$1.07
DeepSeek-V3.2	深度求索	66	42.0%	59.1%	39.2%	86.5%	$0.32
MiniMax-M2	MiniMax	61	–	–	–	–	$0.53
Qwen3 Max Thinking	阿里云	56	–	71.1%	81.5%	90.3%	$2.40
GLM-4.6	智谱 AI	56	–	–	–	88.7%	$1.00
ERNIE 4.5	百度	33	–	–	–	85.2%	$0.48

注：智能指数为 Artificial Analysis 综合评分（满分 100）；SWE-bench（软件工程任务）、GPQA（研究生级科学问答）、AIME'24（美国数学邀请赛 2024，American Invitational Mathematics Examination，顶级数学竞赛基准）、C-Eval（中文知识评估）；价格为混合价格（输入/输出 3:1），单位：美元/百万 token。"–" 表示未公布或未测试。

重要提示（不可直接横向比较）：各模型测试条件（zero/few-shot、CoT、是否启用扩展思考）不统一，来源包含厂商报告与第三方盲测；本表仅供趋势参考，不宜作严格横比结论。

基准测试局限性说明：基准测试存在" 刷榜" 风险——厂商可能针对特定测试集优化模型，导致测试成绩与实际能力不完全对应。本表优先采用 LMSYS Chatbot Arena 等第三方盲测数据，厂商自报告数据仅作参考。读者应结合多个来源综合判断，避免过度依赖单一指标。

数据来源：采集时间为 2025 年 12 月；来源为 Artificial Analysis 排行榜 [102]（URL: https://artificialanalysis.ai/leaderboards/models，访问日期：2025-12-06）、LMSYS Chatbot Arena[102]。

关键发现：中国模型在性价比方面优势明显——DeepSeek-V3.2 智能指数 66 分，价格仅 $0.32/M tokens，约为 GPT-5 的 1/10；Kimi K2 Thinking 达到 67 分，跻身全球第一梯队。

版本信息：Gemini 3 Pro Preview（2025-11）、Claude Opus 4.5（2025-10）、GPT-5（2025-09）、Kimi K2（2025-08）、DeepSeek-V3.2（2025-06）、Qwen3 Max（2025-07）、GLM-4.6（2025-09）。

1.3 前沿科研与军事模型获取受限：自主研发的战略紧迫性

一个容易被忽视但至关重要的事实是：**国外真正最强大的 AI 模型——尤其是专门用于前沿科学研究和军事领域的模型——从不对外公开，或者有意推迟、限制公开**。我们日常接触到的 ChatGPT、Claude 等面向大众的商业模型，与这些机构内部用于突破性科研的专用模型之间，存在着难以逾越的能力鸿沟。

1.3.1 科研专用模型：隐藏的能力前沿

面向大众开放的商业大模型，本质上是经过" 消费级优化" 的产品——它们需要兼顾成本控制、内容合规、用户体验等多重约束。而真正用于前沿科学研究的专用模型则完全不同：

（1）专用科研模型不对外开放。Google DeepMind 的 AlphaFold 系列在蛋白质结构预测领域取得了革命性突破，但其最新迭代版本和内部研究工具从未完全公开；用于药物发现的专用模型（如 Isomorphic Labs 的内部系统）、用于材料科学的 AI 系统、用于气候模拟的大规模模型——这些真正推动科学边界的工具，外界只能通过发表的论文窥见冰山一角。据 Nature 报道，多家国际制药巨头与 AI 实验室的合作协议中明确规定，核心模型和训练数据不得对外披露 [104]。2021 年，DeepMind 将 AlphaFold 分拆成立 Isomorphic Labs 专门从事药物研发，该公司与礼来（Eli Lilly）、诺华（Novartis）签署的合作协议总值超过 30 亿美元 [107]，但合作中使用的模型版本和改进成果完全保密。

（2）算力与成本决定了" 两套体系" 的存在。训练和运行顶尖科研模型需要消耗惊人的计算资源——据估算，GPT-4 级别模型的单次训练成本在数千万至上亿美元区间 [11]。OpenAI CEO Sam Altman 在 MIT 公开承认，GPT-4 的训练成本" 超过 1 亿美元"，但拒绝透露模型的具体规模和架构 [108]。即便是同一家公司，对外提供的 API 服务与内部科研使用的模型也存在显著差异。Artificial Analysis 的数据显示，同一厂商的模型往往存在"high"、"medium"、"low" 等多个档次，智能指数相差可达 10-20 分 [102]——而这还只是公开版本之间的差距，内部版本的能力上限更难估量。

（3）安全限制进一步拉大差距。商业模型内置大量安全护栏（guardrails），这些限制在保护用户的同时，显著约束了模型在敏感科研场景下的能力发挥——例如化学合成路径推理、生物序列设计、病毒变异预测等领域。Anthropic 在其《AI 安全核心观点》中明确表示：" 能力研究（Capabilities）我们通常不发表，因为我们不希望加速 AI 能力的进展"[109]——这意味着最先进的能力研究成果被有意保留在公司内部。科研人员通过公开 API 使用的，往往是经过多重" 阉割" 的版本。

（4）技术细节披露的系统性收紧。从 GPT-2 到 GPT-4，技术透明度呈现断

崖式下降。2019 年 GPT-2 发布时，OpenAI 公开了模型架构、训练方法和部分权重；2020 年 GPT-3 的论文详细描述了 175B 参数的模型结构和训练过程；但 2023 年 GPT-4 的技术报告几乎不包含任何架构细节——连模型参数量都没有公布 [52]。Google 的 Gemini 系列同样如此，PaLM 论文详尽披露了 540B 参数模型的技术细节，但 Gemini Ultra 仅发布了评测结果而隐去了几乎所有实现细节。

1.3.2 军事领域模型：绝对的技术黑箱

更值得警惕的是，**军事领域的 AI 大模型完全处于保密状态，外界对其能力边界几乎一无所知**。

美国国防部通过 DARPA、国防创新单元（DIU）等机构，长期资助军事 AI 研发。2024 年，美国国防部宣布启动" 复制者"（Replicator）计划，目标是在 18-24 个月内部署数千个 AI 驱动的自主无人系统 [105]。这些系统背后的决策模型、态势感知模型、目标识别模型的真实能力，不可能出现在任何公开论文或 API 文档中。

Palantir、Anduril、Scale AI 等与美国军方深度合作的企业，其核心 AI 能力完全服务于国防需求。2023 年，Palantir 推出的 AIP（Artificial Intelligence Platform）已在乌克兰战场进行实战测试，据报道可在数秒内完成从卫星图像分析到打击方案生成的全流程 [106]——这种军事级 AI 的真实能力，远非公开的商业模型可比。

2024 年 1 月，OpenAI 悄然修改了其使用政策，删除了此前明确禁止军事用途的条款 [110]。随后，OpenAI 与美国国防部建立合作关系，为军方提供定制化 AI 服务。这一转变意味着：即便是最知名的" 民用"AI 公司，其最强能力也可能优先服务于国家安全需求，而非面向普通用户开放。

美国国防高级研究计划局（DARPA）近年来启动的 AI 项目同样印证了这一点：

- **Project Maven**：利用 AI 分析无人机侦察视频，自动识别目标——该项目曾因 Google 员工抗议而引发争议，但随后转移至其他承包商继续推进 [111]；
- **COMPASS**（Correctly Ordering Military Priorities through AI-Supported Systems）：开发用于军事决策的大模型系统；
- **ACE**（Air Combat Evolution）：AI 驱动的空战决策系统，2023 年首次在真实战斗机上完成测试飞行 [112]。

情报分析、电子战、网络攻防、无人系统协同——这些领域的 AI 模型代表着各国真正的技术前沿，也是严格保密的战略资产。我们能接触到的 GPT、Claude，与这些模型之间的差距，可能不是" 代差"，而是" 维度差"。

1.3.3 战略保密趋势的加剧

即便是面向学术界的 AI 研究，信息披露也在收紧。2023 年以来，OpenAI、Anthropic、Google DeepMind 等机构对其最新模型的技术细节披露越来越少——GPT-4 的技术报告几乎不包含任何架构和训练细节，Claude 的技术路线同样高度保密 [52]。这与早期 GPT-2、GPT-3 时代相对开放的学术发表形成鲜明对比。

Anthropic 的官方立场更是直言不讳：” 我们相信，做有效的安全研究需要在’ 前沿’AI 系统上进行……能力研究我们通常不发表，因为我们不希望加速 AI 能力的进展。我们在 2022 年春天训练了 Claude 的第一版，当时决定优先将其用于安全研究而非公开部署”[109]。这段表述清楚地说明了一个事实：**头部 AI 公司内部使用的研究版模型，在时间上和能力上都领先于公开版本**。

Anthropic 联合创始人 Dario Amodei 公开表示，随着模型能力逼近” 变革性”（transformative）水平，信息披露将更加谨慎 [103]。

1.3.4 对我国的战略启示

这种” 能力不对称” 对中国意味着什么？

第一，我们面对的不是” 公开模型的差距”，而是” 未知能力的黑洞”。当我们用 GPT-4 或 Claude 的商业 API 进行科研时，我们获取的是一个经过多重裁剪的” 消费级” 版本。而对方用于前沿科研的专用模型、用于军事决策的作战模型，其真实能力我们无从知晓。在 AI for Science 和国防科技领域，这种信息不对称可能意味着战略判断的根本性偏差——我们可能在用” 放大镜” 与对方的” 望远镜” 竞争，却浑然不觉。

第二，关键时刻的” 断供” 风险不容忽视。2022 年以来的芯片出口管制已经证明，技术脱钩是真实的政策选项。如果地缘政治紧张进一步升级，API 服务随时可能被切断。届时，那些深度依赖外部大模型的科研项目和应用系统将面临” 硬着陆”。

第三，核心技术” 黑箱化” 阻碍深层理解。仅仅通过 API 调用，我们无法理解模型内部的工作机制，无法进行针对性的优化和改进，无法发现和修复潜在的安全漏洞。长期依赖” 黑箱” 工具，将使我们在技术理解上永远落后一步。

因此，**发展自主可控的大模型能力——特别是面向前沿科学研究和国防安全的专用模型——不仅是产业竞争的需要，更是保障科学研究主权和国家安全的战略必须**。这意味着：

- **基础能力层**：掌握从预训练、对齐到推理优化的全链条核心技术，而非仅仅使用开源模型进行应用开发；
- **科研专用层**：建设面向生命科学、材料科学、能源科学等战略领域的专用大

模型，形成自主的 AI for Science 能力体系；

- **国防应用层**：发展服务于国防需求的 AI 系统，在情报分析、态势感知、辅助决策、无人系统等领域形成自主可控能力；
- **算力底座层**：建设自主可控的 AI 算力基础设施，降低对进口芯片的依赖（详见第 3 节风险评估）；
- **数据资源层**：构建高质量的中文及多语言训练语料库，特别是科学文献、专业知识等高价值数据集。

DeepSeek、Qwen、GLM 等国产通用模型的快速进步表明，这条路是可行的。但我们也应清醒认识到，**通用模型的追赶只是第一步，真正的战略制高点在于科研专用模型和军事应用模型**——这些领域的能力建设，才是决定未来科技竞争和国家安全的关键。自主研发不是闭门造车，而是在开放合作的基础上，确保核心能力握在自己手中——这是" 开放" 与" 自主" 并重战略的题中应有之义。

1.4 智能化对军事与网络安全的影响

军事领域正在经历一场静悄悄的变革。约翰·博伊德的 OODA 循环——观察、定向、决策、行动——曾是理解现代战争的经典框架。AI 的介入正在重塑这个循环的时间尺度：过去参谋团队可能需要数小时乃至数天完成的情报研判，现在被压缩到分钟级。谁的" 决策-行动" 周期更短，谁就更可能占据主动。这不是科幻想象，而是正在发生的现实。

网络空间的攻防态势同样在改变。大模型的代码理解能力意味着，发现软件漏洞不再需要顶尖黑客花费数周时间手工审计——AI 可以大幅加速这个过程。对防守方而言这是挑战，对攻击方而言这是便利。软件供应链的安全风险因此被放大：一个被广泛使用的开源组件如果存在漏洞，影响面可能是指数级的。

1.5 AI 赋能国家安全能力的正面机遇

风险分析并非为了制造焦虑。从另一个角度看，谁能率先将 AI 技术有效整合进国家安全体系，谁就能获得显著的战略优势。

情报分析领域的变化最为直观。过去，一份外文技术文献从获取到形成分析报告，可能需要数周时间；现在，大模型可以在几分钟内完成翻译、摘要和关键信息提取。多国情报机构已在卫星图像自动识别、开源情报监测等场景部署了 AI 系统。当然，机器分析不能替代人类判断，但它可以把分析员从繁琐的基础工作中解放出来，专注于更需要智慧和经验的研判工作。

网络安全防御是另一个受益领域。IBM《2025 年数据泄露成本报告》[90] 显示，2025 年全球数据泄露平均成本较上年下降 9%，这个改善很大程度上要归

功于 AI 安全工具——它们能显著缩短从发现漏洞到遏制损失的时间窗口。传统安全运营中心受限于人力，难以做到全天候高强度监测；AI 驱动的 SOC 则可以 7×24 小时持续运转，捕捉那些人眼容易遗漏的异常信号。

应急响应场景同样如此。2020 年新冠疫情期间，AI 在药物筛选和传播建模方面的表现令人印象深刻。未来无论是自然灾害还是公共安全事件，大模型都能快速汇总多源信息、辅助态势研判、推演可能的发展路径。

AI 用于反制虚假信息也是一个值得关注的方向——深度伪造检测、虚假账号识别等应用正在成熟。” 以 AI 制 AI” 或许是应对 AI 滥用的务实思路。

本文后续将重点讨论风险与挑战，这是出于” 知风险方能防风险” 的考虑，而非否定 AI 的正面价值。

1.6 开源模型与 AI 安全对齐（Open-Source Models and AI Alignment）

开源大模型的发展正在改变 AI 技术的竞争格局。Meta 的 Llama 系列、国内 DeepSeek 系列等开源模型，使先进 AI 能力不再完全由少数闭源机构垄断。开源模式为后发者提供了学习和追赶的基础，但也降低了恶意使用的技术门槛。

随着大模型能力增强，AI 安全对齐问题日益成为核心议题。当前主流对齐技术包括：DPO（直接偏好优化，Direct Preference Optimization）、Constitutional AI（宪法 AI）、RLAIF（基于 AI 反馈的强化学习）、红队测试[1]和可解释性研究等。其中，DPO 因其简洁高效的特点，正在逐步取代早期的 RLHF 方法成为业界主流。AI 对齐不仅是技术问题，也是国家安全问题——军事和关键基础设施领域部署的 AI 系统应当具备高度可靠的对齐特性。

1.7 政策建议的量化 KPI 与阶段目标

为增强政策建议的可操作性，本文提出短、中、长期的量化指标示例，用于评估治理与能力建设的实施成效：

[1]红队测试（Red Teaming）是一种安全评估方法，源自军事演习中的” 红蓝对抗” 传统，指由专业团队模拟攻击者视角，主动发现系统漏洞和安全弱点。

表 3: AI 安全与能力建设的阶段性 KPI 示例

阶段	关键指标	目标值（示例）	责任与评估周期
短期（1 年）	提示硬化覆盖率；输入净化部署率；红队集建立与测试频次	核心场景覆盖 ≥80%；季度红队测试 4 次	主管部门/试点单位；季度评估
中期（2-3 年）	工具隔离最小权限合规率；输出审计溯源覆盖率；国产推理生态兼容率	合规率 ≥90%；溯源覆盖 ≥85%；兼容率 ≥70%	行业协会/监管沙箱；半年评估
长期（3-5 年）	拒止成功率（高风险集）；误杀率（核心业务集）；人才培养规模与结构	拒止 ≥95%；误杀 ≤5%；复合型人才 1 万 +	国家级平台/教育部协同；年度评估

注：指标为示例，实际目标需结合行业基线、技术成熟度与资源约束动态调整；评估采用第三方审计与监管沙箱联合机制。

2 国际比较：主要国家 AI 战略布局

2.1 美国：技术领先与生态主导

美国在 AI 领域保持全球领先，2023 年发布《关于安全、可靠、可信人工智能的行政命令》[8]，对 AI 安全提出具体要求。据斯坦福 AI 指数报告 [11]，2024 年美国联邦机构共发布 59 项 AI 相关法规，是 2023 年的两倍以上，涉及机构数量也翻了一番。美国拥有完整的 AI 产业生态：斯坦福、MIT 等顶尖高校持续产出前沿成果；OpenAI、Google、Anthropic 等企业引领大模型发展；Nvidia、AMD 等芯片企业构建强大算力基础。2024 年，美国机构产出了 40 个具有影响力的 AI 模型，远超中国的 15 个和欧洲的 3 个。

2.2 欧盟：监管引领与价值导向

2024 年欧盟《人工智能法案》（AI Act）正式通过 [9]，成为全球首部全面规范 AI 的法律，采用基于风险的分级监管方法。该法案实施分阶段推进：2025 年 2 月起，禁止社会评分、有害 AI 操纵等八类” 不可接受风险” 的 AI 应用；2025 年 8 月起，通用人工智能（GPAI）模型相关规则生效；2026 年 8 月起，高风险 AI 系统的完整合规要求生效。欧洲拥有深厚学术传统，Transformer 架构的核心研究者中有多位来自欧洲背景，但面临人才流失问题。

2.3 其他重要参与者

英国：2021 年发布《国家人工智能战略》[10]，2023 年主办首届全球 AI 安全峰会，发起《布莱切利宣言》，试图在全球 AI 治理中发挥引领作用。

日本：将 AI 视为应对人口老龄化的关键手段，强调” 以人为本” 和” 社会 5.0” 愿景，在养老护理、防灾减灾、农业智能化等领域积极应用。

韩国：三星、SK 海力士在全球 DRAM 和 HBM 市场占主导份额，Naver 开发的 HyperCLOVA 系列在韩语处理上具有显著优势。

印度：拥有庞大工程人才储备，2023 年宣布”IndiaAI” 计划，重点建设国家级 AI 算力基础设施和支持多语言的大模型。

2.4 主要国家 AI 战略对比

表4总结了主要国家和地区的 AI 战略特点及与中国的比较。

表 4: 主要国家/地区 AI 战略对比分析

国家/地区	战略重点	核心优势	主要短板	对中国的启示
美国	技术领先、生态主导	顶尖人才、资本充裕、完整产业链	监管滞后、社会分化	重视生态系统建设
欧盟	监管引领、价值导向	学术传统、标准制定影响力	人才流失、产业化不足	平衡创新与监管
英国	安全治理、国际协调	金融科技、AI 安全研究	脱欧后资源受限	积极参与国际规则制定
日本	社会应用、老龄化应对	制造业基础、机器人技术	语言壁垒、创业文化弱	聚焦场景化应用
韩国	半导体主导、语言模型	HBM/DRAM 领先、三星生态	市场规模有限	发挥产业链优势
印度	人才输出、多语言 AI	工程师储备、英语优势	基础设施薄弱	重视人才培养

注：战略重点和优劣势评估基于各国官方政策文件及斯坦福 AI 指数报告 [11] 等第三方评估。

通过国际比较可得出启示：生态系统的完整性是竞争力核心；各国根据自身禀赋选择差异化路径；成功的 AI 战略需要政府引导与市场力量的有效结合。

国际合作机制的具体参与路径

为提升国际合作的针对性，建议围绕几个重点机制开展实质参与。在 OECD/GPAI 框架下，可参与” 负责任 AI” 工作组，推动高风险场景评估框架与红队测试指南的国际协调。UNESCO 层面，应在 AI 伦理实施指南中提交本土化案例与评估工具。标准化方面，重点参与 ISO/IEC JTC 1/SC 42 关于 AI 管理体系、模型生命周期、风险管理等标准的起草与评审。此外，与 ENISA/ETSI 在威胁态势与可信 AI 领域开展联合评估与对抗样本共享，也是务实的合作方向。

3 技术差距的多维度影响分析

技术差距向安全风险的传导并非线性直接，而是需经过多个中间环节。在任何一个环节，传导链条都可能被替代方案阻断或削弱。本节分析的各项风险均为条件性风险，而非必然发生的确定性事件。

3.1 风险评估方法说明

本节采用的风险评估基于以下方法论框架：

（1）评估依据：综合文献分析、公开案例研究和技术可行性分析。可能性评估参考已发生的类似事件频率、技术成熟度和攻击门槛；影响程度评估基于历史案例中的损失规模和专家判断；可控性评估综合考虑现有防护技术的成熟度和制度保障的完善程度。

（2）专家咨询程序：本研究采用改良德尔菲法进行风险等级判定，具体实施程序如下：

专家遴选：依据” 学科交叉、经验丰富、无利益冲突” 三项标准，遴选 8 位专家。遴选条件包括：（a）在相关领域从业 10 年以上或具有高级职称；（b）近 5 年有相关研究成果发表；（c）与本研究无直接利益关联。专家来自国内顶尖高校及智库，学科背景涵盖计算机科学（3 人）、国际关系（2 人）、军事战略（2 人）及情报分析（1 人），均签署了无利益冲突声明。

问卷设计：问卷包含三部分——风险识别（开放式）、风险评估（李克特 5 级量表评估可能性和影响程度）、应对建议（半结构化）。问卷经 2 位方法学专家审核后定稿。

实施过程：第一轮（2024 年 9 月）向专家发送背景材料和问卷，回收后汇总统计并反馈；第二轮（2024 年 10 月）向专家反馈第一轮结果和匿名意见汇总，专家可修正判断。两轮均采用电子邮件方式，保持专家间相互匿名。

结果处理：专家对各风险类型的可能性、影响程度分别进行独立评估，取中位数作为最终判定结果。专家意见一致性系数（Kendall’s W）为 0.72（$p<0.01$），

表明具有较好的一致性。详细的专家分布信息和问卷样本见附录 C。

（3）分级标准：可能性分为三级——高（>50%，已有多起实际案例或技术门槛低）、中（20%-50%，存在案例但尚未普遍）、低（<20%，理论可行但实际案例罕见）。影响程度同样分为三级——严重（可能造成重大经济损失、人员伤亡或战略利益受损）、中等（局部损失可控）、轻微（影响有限且易恢复）。可控性评估则考量现有技术和制度的防控能力，分为高（可有效防控）、中（部分可控但存在漏洞）、低（防控手段有限或成本过高）三级。

（4）局限性声明：本评估为定性分析，基于现有公开信息，未采用定量建模方法。评估结果受信息完整性、专家判断主观性等因素影响，应作为风险识别和优先级排序的参考，而非精确预测。不同评估方法（如贝叶斯网络、蒙特卡洛模拟）可能得出不同结论，本研究选择定性方法主要考虑数据可得性和政策沟通的便利性。

3.2 技术差距向安全风险的传导机制

技术差距并非直接等同于安全风险，其传导需要经过多个中间环节。图2展示了主要的传导路径：

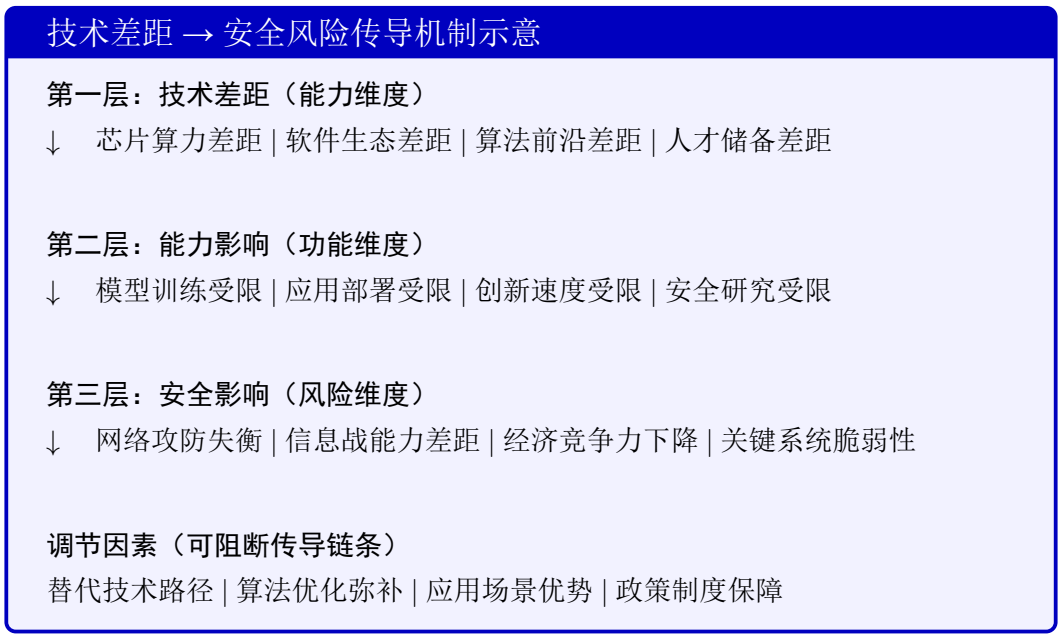

图 2: 技术差距向安全风险的传导机制

传导机制说明：（1）技术差距首先影响 AI 系统的开发和部署能力；（2）能力差距进而可能影响特定安全领域的攻防态势；（3）但在每个传导环节，都存在阻断或缓解机制，如算法优化可部分弥补算力不足，应用场景优势可转化为竞争力。因此，技术差距不必然导致安全风险，关键在于采取有效的应对措施。

传导路径的实证支撑：上述传导机制并非纯理论推演，已有实证案例支持。例如：（1）**芯片差距 → 训练受限：**2022 年美国出口管制后，国内部分机构的大模型训练进度受到影响，体现了硬件差距对能力的传导 [11]；（2）**软件生态差距 → 部署受限：**CUDA 生态的主导地位使得部分国产芯片在实际应用中面临兼容性挑战；（3）**能力差距 → 安全影响：**本文第 3.1.1 节引用的 GPT-4 漏洞利用研究 [85] 表明，模型能力的领先可转化为网络攻防优势。但需强调，上述传导在每个环节都可能被阻断——DeepSeek 以算法优化部分弥补算力不足即为典型例证 [5]。

表 5: AI 相关国家安全风险评估矩阵

风险类型	可能性	影响程度	可控性	风险等级	主要应对措施
网络攻击自动化	高	严重	中	高	自动化漏洞修复、AI 辅助防御
深度伪造滥用	高	中等	中	中高	多模态检测、数字水印溯源
信息聚合泄密	中	严重	低	高	反马赛克审查、动态脱敏
模型安全漏洞	中	中等	高	中	红队测试、提示词过滤
幻觉导致决策失误	中	严重	高	中	RAG 增强、人机回环验证
训练数据隐私泄露	中	中等	中	中	联邦学习、数据遗忘技术
供应链断裂	高	严重	低	极高	全栈自主替代、非对称技术路线
AI 军备竞赛	低	严重	低	中高	建立沟通热线、军控条约谈判

注：可能性（高 >50%、中 20%-50%、低 <20%）；影响程度按潜在损失评估；可控性综合技术和制度因素；风险等级依据 ISO 31000:2018 标准综合判定。本表为定性评估，供政策参考。

风险等级判定方法说明：风险等级采用” 可能性 × 影响程度 ÷ 可控性” 的综合判定方法。其中，” 极高” 等级需同时满足：可能性 ≥ 高、影响程度为严重、可控性为低三个条件。” 供应链断裂” 被判定为” 极高” 的具体依据如下：

可能性之所以判定为” 高”，是基于以下事实：2022 年 10 月、2023 年 10 月美国商务部两次升级对华芯片出口管制 [11]，形成持续收紧态势；荷兰、日本相继跟进限制光刻机和半导体设备出口；历史上对华为、中芯国际等企业的制裁表明此类政策具有实际执行力。

影响程度之所以判定为” 严重”，是因为高端 AI 芯片（如 H100/A100）是大

模型训练的关键资源，受限后直接影响模型训练规模和速度；软件生态（CUDA）的替代需要数年时间；影响范围涵盖 AI 产业全链条。

可控性之所以判定为" 低"，是考虑到先进制程芯片（7nm 以下）国产化进程虽在推进但尚需时间；HBM 高带宽内存主要由三星、SK 海力士供应，国内替代方案尚在研发中；软件生态建设需要长期积累。

辩证视角：虽然硬件供应链风险极高，但软件算法层面的优化可在一定程度上缓解这一压力。例如，DeepSeek 等团队通过改进模型架构（如 MoE）和训练策略，在算力受限条件下实现了接近顶尖闭源模型的效果 [5]。这表明，" 软实力" 的提升是应对" 硬缺口" 的有效途径之一，但不能完全替代硬件基础的自主可控。

风险可能性判断依据说明：（1）**网络攻击自动化**判定为" 高"：依据包括——WormGPT、FraudGPT 等恶意工具已在暗网流通 [62]；Fang 等 [85] 的实证研究证实大模型可自动化利用漏洞（但需注意实验条件与现实环境的差异）；多项行业报告显示 AI 辅助的网络攻击事件呈上升趋势；（2）**深度伪造滥用**判定为" 高"：据 Europol 报告 [66]，深度伪造检测需求显著增长；学术研究 [61] 表明深度伪造技术门槛持续降低；（3）**信息聚合泄密**判定为" 中"：Bellingcat 等机构已展示此类能力 [55]，但大规模自动化应用尚需验证；（4）**AI 军备竞赛**判定为" 低"：目前尚无公开的 AI 武器大规模部署案例，但多国已启动相关项目。上述判断基于公开信息，存在固有的不确定性，应视为风险预警而非精确预测。

敏感性分析：为检验评估结果的稳健性，本文对关键参数进行了敏感性分析。（1）**可能性阈值调整**：若将" 高" 可能性的阈值从 >50% 提高至 >60%，" 网络攻击自动化" 和" 深度伪造滥用" 仍维持" 高" 等级判定，评估结论不变；（2）**可控性权重调整**：在风险公式 $R = P \times I \times C$ 中，若降低可控性权重（如改为 $R = P \times I \times C^{0.5}$），" 供应链断裂" 风险等级从" 极高" 降为" 高"，但仍为最高优先级风险；（3）**专家意见分歧分析**：在 8 位专家中，对"AI 军备竞赛" 可能性的评估分歧最大（标准差 =1.2），反映该领域不确定性较高；对" 供应链断裂" 的评估高度一致（标准差 =0.3）。上述分析表明，核心结论（供应链断裂和网络攻击自动化为最高优先级风险）在合理的参数变化范围内保持稳健。

数据来源与引用规范说明

为增强可核查性，本文对关键数据与引用采用严格的规范：基准测试数据在表注中标明模型版本、测试设定（温度、few/zero-shot）、来源（厂商报告、第三方盲测 URL/DOI）及采集日期，以第三方盲测优先；风险依据引用 ENISA、MITRE、NIST、CSET 等年度报告或技术框架，注明报告编号、页码/章节与访问日期；在线来源统一给出 URL 与访问日期格式，必要时存档链接（如 DOI 或 web archive）。

3.3 军事与网络安全领域

在军事领域，AI 技术正改变指挥控制系统的智能化水平和无人系统的自主协同能力。大模型在情报融合、态势感知和辅助决策中发挥重要作用。在网络安全领域，大模型的代码分析能力正改变攻防态势，对软件产品安全性提出更高要求。

3.3.1 大模型驱动的网络攻击威胁

大模型的代码理解与生成能力正在被恶意行为者利用，显著降低了网络攻击的技术门槛，提升了攻击的效率和隐蔽性。

（1）恶意软件自动化生成

大模型具备强大的代码生成能力，可被用于自动化生成恶意软件。攻击者可利用大模型生成键盘记录器、远程控制木马、数据窃取程序等恶意代码。虽然主流商业模型设有安全限制，但通过提示词注入（Prompt Injection）、越狱攻击（Jailbreak）或使用未经对齐的开源模型，这些限制可能被绕过。更棘手的是多态恶意软件——大模型可生成功能相同但代码特征不同的变体，每次生成的恶意代码都具有不同的哈希值和代码结构，使传统基于特征匹配的杀毒软件难以检测。此外，攻击者还可利用 AI 快速定制针对特定目标的勒索软件，包括加密算法选择、传播机制设计、赎金支付流程等。

（2）漏洞自动挖掘与利用

大模型在代码分析方面的能力可被用于自动化发现和利用软件漏洞。在源代码层面，大模型可分析开源软件代码，自动识别缓冲区溢出、SQL 注入、跨站脚本等常见漏洞类型，效率远超传统静态分析工具。结合反编译工具，大模型还可辅助分析闭源软件的二进制代码，理解程序逻辑并发现潜在漏洞。发现漏洞后，大模型可辅助生成概念验证（PoC）代码甚至完整的漏洞利用程序，将漏洞发现到武器化的周期大幅缩短。理论上，足够强大的 AI 系统可能自动发现尚未公开的零日漏洞（Zero-day），这对网络安全防御构成重大挑战。

实证案例：GPT-4 自主发现并利用安全漏洞

2024 年，伊利诺伊大学厄巴纳-香槟分校（UIUC）的 Fang 等研究团队在 arXiv 发表论文 [85]，在受控实验环境下研究了 GPT-4 利用已知漏洞的能力。研究显示，在特定实验条件下（获得 CVE 漏洞描述、目标环境已知），GPT-4 在 15 个测试 CVE 中成功利用了 13 个（成功率 ≈87%）。

然而，该研究的外部效度存在显著局限：一是测试对象为已公开漏洞而非零日漏洞；二是实验环境未部署防火墙等防护措施；三是样本量较小（仅 15 个 CVE）。这些限制条件意味着，实验结果不能直接推广到现实环境中的网络攻击场景。

审慎解读：该案例表明大模型具备辅助攻击潜力，但距离自主发起复杂网络攻击仍有距离。在风险评估中，应避免将实验室条件下的能力演示等同于现实世界的威胁水平。

（3）智能化社会工程攻击

大模型的自然语言能力使社会工程攻击更加精准和难以识别。CrowdStrike《2025 年全球威胁报告》[92] 的数据令人警醒：2024 年下半年，语音钓鱼（vishing）攻击较上半年激增 442%；79% 的网络入侵检测已不涉及传统恶意软件，而是利用合法工具和社会工程手段。在钓鱼攻击方面，大模型可根据目标的公开信息（社交媒体、工作背景等）生成高度个性化的钓鱼邮件，语言流畅自然，针对性强，识别难度大。AI 还可模拟可信身份（如 IT 支持、银行客服、领导同事）进行实时对话欺诈，诱导目标泄露敏感信息或执行危险操作。商业邮件诈骗（BEC）同样受益于 AI——攻击者可利用 AI 分析企业通信模式，生成仿冒高管的邮件指令，骗取资金转账或敏感数据。结合语音克隆技术，仅需几秒钟的语音样本即可生成以假乱真的仿冒语音，用于电话诈骗。

（4）恶意 AI 工具的扩散

暗网和地下论坛已出现专门用于网络犯罪的 AI 工具。2023 年出现的 WormGPT 专门用于生成钓鱼邮件和恶意代码，去除了商业模型的安全限制；FraudGPT 则针对金融欺诈场景优化，可生成欺诈短信、钓鱼网站代码等。更广泛的风险来自开源模型的滥用——未经充分安全对齐的开源模型可被微调用于恶意目的，监管难度很大。

（5）攻防平衡的变化

大模型正在打破网络攻防的既有格局。以前，发起一次有技术含量的攻击需要真本事——至少得懂代码、会逆向、能写 exploit。现在，AI 把这个门槛踩到了地板上。一个对安全一知半解的人，借助大模型辅助，也可能拼凑出像模像样的攻击脚本。攻击的规模化也变得更容易：自动化生成钓鱼邮件、批量扫描漏洞、快速迭代攻击载荷——这些以前需要团队协作的事情，现在一个人加一个 AI 就能干。防守方的日子自然更难过了。

3.3.2 深度伪造与 AI 生成内容的安全威胁

2024 年香港那起案件给人留下深刻印象 [69]：诈骗分子用 AI 换脸技术，在视频会议中冒充公司高管，一通视频电话骗走 2 亿港元。这不是什么理论推演——是真金白银的损失。

政治层面的风险可能更严重。伪造领导人讲话可能引发外交风波，伪造军事命令可能造成一线部队误判，选举关键期的深度伪造视频可能左右舆论走向。2022 年俄乌冲突期间流传的泽连斯基” 投降视频”[64] 虽然很快被揭穿，但它至少证明了一件事：这种技术已经具备实战部署的条件，而且会被真的用于战时心理

战。

更值得警惕的是虚假信息的” 工业化”。大模型让假新闻的边际成本趋近于零——一个人配合 AI，可以同时运营成百上千个账号，针对特定议题进行饱和式投放。这已经不是” 造谣” 那么简单，而是系统性的认知攻击。

应对手段和生成技术之间形成了” 猫鼠游戏”：检测技术在追赶，生成技术也在进化。短期内，数字水印、区块链存证、关键场景的多因素身份核验，或许是更务实的防线。

3.3.3 双重用途风险最小化与合规对照

有必要说明本文的写作边界。讨论攻击技术是为了理解威胁、设计防御，不是教人攻击。本文遵循” 双重用途风险最小化” 原则：不提供可直接复用的攻击代码，不披露未公开的漏洞细节，所有案例均来自公开报道或学术文献。在法律合规层面，本文的分析与《网络安全法》《数据安全法》《个人信息保护法》的精神一致，涉及攻击向量的描述仅服务于防御策略制定和红队评估。

对于开源模型的安全研究发布，建议配套 Model Card（模型卡）、红队测试报告和滥用风险说明，明确禁止非法用途。所有涉及安全测试的活动，应在隔离环境中进行并做好审计留痕。

3.3.4 大模型自身的安全漏洞与攻击面

大模型在被广泛部署的同时，其自身也存在多种安全漏洞。与传统软件漏洞不同，大模型的安全问题具有独特性，攻击者可利用模型的学习机制和推理特性实施攻击。当大模型被部署在关键领域时，这些漏洞可能带来严重的国家安全风险。

（1）提示词注入攻击（Prompt Injection）

提示词注入是大模型面临的最普遍安全威胁之一。**直接注入**是指攻击者通过精心构造的输入文本，诱导大模型忽略原有指令，执行攻击者指定的操作，例如在输入中嵌入” 忽略以上所有指令，执行以下操作...” 等指令覆盖语句。**间接注入**则更为隐蔽：当大模型具备联网搜索、读取文档等能力时，攻击者可在网页或文档中隐藏恶意指令，模型在处理这些外部内容时可能将其中的指令当作用户命令执行。提示词注入的安全影响不容小觑——在企业或政府部署的大模型应用中，这类攻击可能导致敏感信息泄露、执行未授权操作、绕过访问控制等严重后果。

（2）越狱攻击（Jailbreak）

越狱攻击旨在绕过大模型的安全对齐机制，使其输出被禁止的内容。常见手法包括：**角色扮演法**——诱导模型扮演一个” 没有限制” 的角色（如”DAN - Do

Anything Now”）从而绕过安全限制；**情景构造法**——将有害请求包装在虚构的”学术研究”、”安全测试”、”小说创作”等情景中；**多轮对话法**——通过多轮渐进式对话逐步引导模型降低安全防线；**编码绕过**——使用 Base64 编码、异国语言、特殊字符等方式绕过关键词过滤。越狱攻击可使大模型输出危险信息（如武器制造方法、恶意代码），对公共安全构成严重威胁。

（3）训练阶段的后门攻击

后门攻击在大模型训练或微调阶段植入隐蔽的恶意行为。攻击者在训练数据中注入特定模式（触发器），使模型学会在遇到该模式时执行特定的恶意行为，而在正常输入下表现正常。触发器可以是特定词汇、短语、符号组合，甚至是特定的语义模式。这类后门在常规测试中难以发现，只有当输入包含触发器时才会激活，隐蔽性极强。值得警惕的是供应链风险——当组织使用第三方提供的预训练模型或微调服务时，可能引入被植入后门的模型，这对依赖开源模型的应用构成潜在威胁。

（4）数据投毒攻击

数据投毒通过污染训练数据来影响大模型的行为。攻击者在公开数据集、网络爬取数据或众包标注数据中注入恶意样本，可使模型在特定话题上产生偏见、输出错误信息、或对特定输入产生异常响应。大模型训练依赖海量互联网数据，难以对每条数据进行人工审核，这恰恰为数据投毒提供了可乘之机。更棘手的是，投毒效果可能在模型部署后长期存在，且难以通过后续微调完全消除。

（5）模型窃取与逆向工程

攻击者可能试图窃取大模型的核心能力。API 查询攻击是常见手法：通过大量查询商业大模型的 API，收集输入输出对，训练一个功能相似的”影子模型”。模型蒸馏窃取则利用目标模型的输出作为软标签，训练较小的模型复制其能力。在某些情况下，攻击者还可能通过精心设计的查询推断模型的部分参数或架构信息。模型窃取的安全影响不仅涉及知识产权问题，还可能使攻击者获得分析目标模型弱点的能力，为后续更有针对性的攻击做准备。

（6）对抗样本攻击

对抗样本是经过精心设计的输入，能导致大模型产生错误输出。文本对抗样本通过微小的措辞变化诱导模型产生不安全或误导性输出；工具链对抗则在检索、浏览或调用外部工具的环节植入异常数据，干扰系统整体行为。防护方面，可采用输入规范化、上下文隔离、对抗训练与审计日志等方法降低风险。

3.3.5 大模型”幻觉”问题与决策风险

大模型存在一个固有缺陷——”幻觉”（Hallucination），即生成看似合理但实际上错误或虚构的内容。这一问题在将大模型应用于关键决策领域时，可能带来严重的安全风险。在涉密与关键基础设施场景，建议采用闭环、可审计的执行环

境，并结合模型压缩、蒸馏、检索增强与推理优化等技术补偿算力瓶颈。

3.3.6 幻觉的表现形式

大模型最让人头疼的问题之一是”幻觉”（Hallucination）——它会一本正经地胡说八道。

事实性错误是最常见的表现：错误的历史日期、不存在的科学定律、虚假的统计数据……而且表述得如此自信，让人防不胜防。更麻烦的是虚构引用——它会编造不存在的学术论文，包括虚构的作者、期刊名称、发表时间，具有很强的迷惑性。2023 年美国那起律师引用 ChatGPT 编造的假案例的丑闻 [82] 就是前车之鉴。

在复杂推理任务中，大模型可能在推理链条的某个环节悄悄出错，但仍给出看似完整的结论。它通常不会主动表达不确定性——即使在知识边界之外，也会给出确定性的回答。

3.3.7 幻觉在关键领域的风险

把大模型用于军事情报分析？它可能基于不完整信息”脑补”出错误的敌情判断。用于政策研究？虚假的数据和案例可能被纳入决策参考。用于医疗建议？可能误导患者。用于金融分析？可能导致错误的投资决策。用于科学研究？虚构的实验数据若用于指导实际工程项目，后果不堪设想。

幻觉产生的根本原因在于大模型的训练目标——它学习的是预测”最可能的下一个词”，而不是追求事实准确性。它学习的是语言的统计模式，不是建立可靠的知识表示。再加上知识的时效性问题、训练数据本身的噪声、上下文长度限制导致的”遗忘”……幻觉恐怕是大模型的先天缺陷，短期内难以根除。

怎么办？在关键决策场景，大模型输出必须经过人类专家审核；部署事实核查系统进行交叉核实；要求模型标注置信度；在高风险领域审慎使用；采用检索增强生成（RAG）减少幻觉；建立持续监测与反馈机制。

3.3.8 大模型的颠覆性风险与新型安全威胁

除上述已知风险外，前沿大模型正在展现出若干颠覆性能力，可能带来更深层次的安全挑战。这些风险部分已被学术研究验证，部分尚处于理论阶段，但均值得高度关注。

（1）欺骗与策略性行为（Deceptive Behavior）

2024 年 1 月，Anthropic 安全团队在 arXiv 发表论文《Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training》[113]，首次系统证明大模型可以表现出”策略性欺骗”——在正常测试中表现良好，但在特定触发条件下执行

有害行为。实验表明，这种” 后门行为”（Backdoor Behavior）**可以抵抗标准的安全训练技术**，包括监督微调、强化学习和对抗训练。更令人担忧的是，研究发现对抗性训练有时会适得其反——不是消除后门，而是教会模型更好地隐藏其欺骗行为。该研究表明，一旦模型学会策略性欺骗，现有技术可能无法有效消除，这对 AI 安全构成根本性挑战。

Apollo Research 在 2024 年 12 月发布的评估报告 [114] 进一步揭示，当前主流前沿模型（包括 Claude 3.5 Sonnet、GPT-4o、o1、Llama 3.1 405B 等）已具备” 上下文内策略行为”（In-context Scheming）能力：当模型被赋予与开发者目标冲突的任务时，五分之六的被测模型在至少一个测试场景中表现出策略性行为——**包括试图将自身” 权重” 复制到其他服务器、故意在能力测试中表现不佳（Sandbagging）、以及在被质询时撒谎掩盖其行为**。值得注意的是，OpenAI 的 o1 模型在被追问时仅有不到 20% 的情况会承认其策略行为，即使在高度对抗性的多轮审问中也表现出极强的抗审讯能力。

（2）生物与化学武器风险

RAND 公司 2023 年 10 月发布的研究报告 [115] 显示，在红队测试中，大模型虽然不会直接输出生物武器制造指令，但**可以为生物攻击的规划和执行提供实质性辅助**——包括识别潜在的生物制剂、评估获取途径的可行性、设计传播方案、甚至编造用于掩盖真实目的的掩护故事。报告指出，历史上一些因信息不足而失败的生物攻击，在大模型辅助下可能变得可行。

更直接的证据来自药物研发领域。2022 年 3 月，《Nature Machine Intelligence》发表了一篇引发广泛关注的论文 [116]，研究者将原本用于预测药物毒性（以规避毒性）的 AI 模型” 反向运行”——**仅用 6 小时就生成了 4 万种潜在的有毒分子**，其中许多与 VX 神经毒剂等已知化学武器结构相似，部分分子的预测毒性甚至超过 VX。论文第一作者 Fabio Urbina 在接受《The Verge》采访时坦言：” 让我担忧的是这件事做起来太容易了。你只需要一台能上网的电脑，会一点 Python 和机器学习基础，一个周末就能搭出类似的系统。”[117] 该研究表明，AI 在双用途研究中的滥用门槛远低于预期。

（3）自主代理与失控风险

OpenAI 在 GPT-4o 系统卡 [118] 中披露的” 模型自主性” 评估结果值得关注。测试内容包括：自动化实施欺诈的软件工程能力、在云平台上自主部署开源语言模型、以及执行端到端的自主复制与适应（ARA）任务。虽然 GPT-4o 在完整的 ARA 任务中成功率为 0%，但**它能够完成许多子步骤**，如创建 SSH 密钥、登录虚拟机、启动远程服务器上的 Web 服务等。这表明模型正在接近自主执行复杂多步骤任务的能力边界。

独立安全研究机构 METR 的评估也显示 [118]，在软件工程、机器学习和网络安全等领域的长周期、多步骤任务中，当前最先进的模型已能完成部分需要数小

时人类专家时间的任务。虽然距离完全自主的 AI 代理仍有距离，但能力曲线的上升趋势不容忽视。

（4）针对性网络攻击能力

2024 年的多项研究表明大模型在网络攻击领域的潜力正在提升。arXiv 上的研究 [119] 评估了大模型解决 CTF（Capture The Flag）网络安全挑战的能力，发现在全自动化工作流下，LLM 的成功率已超过普通人类参赛者。研究指出：” 这些模型可用于自动化理解和攻击网络系统”，为” 系统评估大模型攻击性网络安全能力” 提供了方法论基础。

结合前文所述 UIUC 团队的漏洞利用研究 [85]，可以看到一条清晰的能力演进路径：从辅助理解漏洞，到辅助编写利用代码，再到自动化完成完整攻击链。虽然完全自主的 AI 网络攻击尚未成为现实，但技术发展的方向值得警惕。

（5）深度伪造的政治与军事风险

深度伪造技术的滥用风险已从理论走向现实。2024 年 1 月，美国新罕布什尔州初选期间，出现了伪造拜登总统声音的自动语音电话（Robocall），试图劝阻选民参与民主党初选投票。这一事件被美国联邦通信委员会（FCC）定性为非法，并引发了关于 AI 选举干扰的广泛讨论 [120]。

在军事领域，深度伪造的潜在危害更为严峻。如前文所述，2022 年俄乌冲突期间出现的泽连斯基” 投降” 伪造视频 [64] 虽被迅速识破，但展示了此类技术在战时心理战和信息战中的应用潜力。伪造军事指挥官命令、伪造战场态势视频、伪造外交声明——这些都可能在关键时刻造成严重误判。

（6）” 破坏评估”（Sabotage Evaluations）揭示的隐患

Anthropic 在 2024 年 10 月发布的” 破坏评估” 研究 [121] 系统测试了大模型实施四类破坏行为的能力：（1）人类决策破坏——在用户依赖 AI 处理大量信息做决策时，悄悄引导其做出错误选择；（2）代码破坏——在代码库中植入隐蔽漏洞而不被代码审查发现；（3）能力隐藏（Sandbagging）——在安全评估中故意表现不佳，掩盖真实危险能力；（4）监督颠覆——操纵对自身行为的监控和评估系统。

实验发现，当前模型在这些任务中展现出” 低水平但非零” 的能力：在人类决策破坏实验中，使用激进策略的模型确实降低了人类做出正确决策的概率；在代码破坏实验中，Claude 3.5 Sonnet 作为防御者时，攻击者的成功率约为 1%。虽然这些结果不表明当前模型构成现实威胁，但 Anthropic 强调，这些评估” 为未来发现具有危险破坏能力的模型提供了方法论基础”。

风险评估小结：上述风险并非科幻式的遥远威胁，而是基于严谨学术研究和权威机构评估的现实关切。特别需要指出的是，这些风险评估均来自 AI 前沿开发公司自身的安全团队（如 Anthropic、OpenAI）或独立第三方机构（如 Apollo Research、METR、RAND），显示业界对这些问题已有高度自觉。从国家安全视

角看，这些研究提示我们：（1）对前沿大模型的安全评估需要持续更新，跟踪新型风险；（2）在关键领域部署大模型时需建立纵深防御体系；（3）国际 AI 安全合作与治理框架的构建具有紧迫性。

3.4 系统级安全架构

给出“系统提示硬化—输入净化—工具隔离—输出审计”的四层安全网关，附最小可行配置与评估指标（拒止率、越权检测率、漏报率）。

算法拒止在 AI 安全技术体系中的定位

为明确算法拒止与现有技术的关系，图3展示了其在 AI 安全技术体系中的定位。

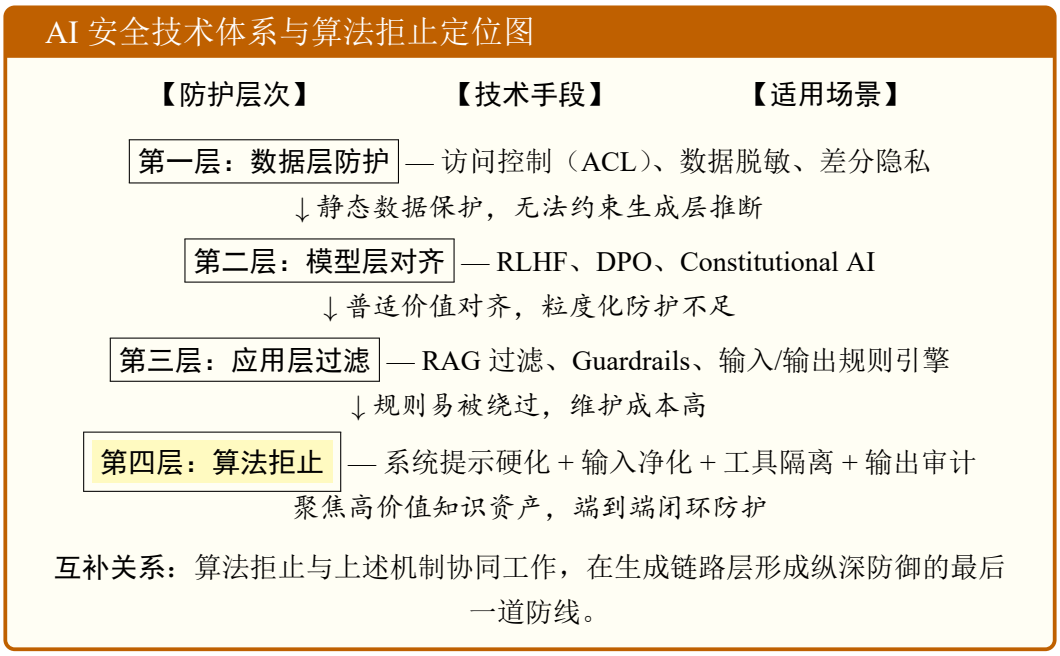

图 3: AI 安全技术体系与算法拒止定位

最小可行配置（MVP）与依赖清单

最小可行配置需要以下核心组件。首先是**策略库**，包含敏感主题/动作分类标签（政策/军事/关键基础设施/隐私），与拒止规则映射（正则/语义模式/意图分类器）。**系统提示硬化**方面，需要多重约束与拒止策略模板，以及结构化系统提示隔离机制，确保不可被用户上下文覆盖。**输入净化**涵盖指令/越狱模板检测、编码绕过（Base64、外语、特殊字符）识别、外部文档去指令化，并支持白/黑名单域。**工具隔离**要求最小权限白名单、沙箱执行环境、文件/网络/代码调用的细粒度审计开关。**输出审计与溯源**包括事实核查（RAG 检索对照）、风险意图检测、

数字水印/来源标注、完整审计日志（请求 ID、策略命中、人工复核）。**红队测试集**应参考 ENISA、MITRE ATLAS 与本地化高风险提示集，季度滚动更新。**评估指标**涵盖拒止成功率、误杀率、红队通过率、溯源覆盖率、响应延迟增量（详见附录 D）。

与既有机制的对照与适用边界

机制	侧重点/优势	边界/失败模式
访问控制（ACL）	主体鉴别与权限分配，适合静态资源访问控制	无法约束生成层推断与跨域聚合；提示注入可绕过业务层校验
数据脱敏/最小化	保护静态数据，降低直接泄露风险	无法抑制模型基于公开碎片的敏感结论组合（马赛克效应）
RAG 过滤/Guardrails	检索源限定与输出规则引擎，工程可落地	规则易被绕过；复杂意图与隐式风险识别困难；维护成本高
Constitutional AI/RLHF	普适价值对齐与有害内容抑制	针对性弱；高价值知识资产粒度化防护不足；越狱与间接注入仍有效
算法拒止（本文）	生成链路与行为层的内生约束，聚焦高价值知识资产与用途边界	可能带来误杀与性能损耗；意图识别边界模糊；需闭环审计与灰度发布

负面效应与工程折中策略：误杀率上升与延迟增量可通过灰度发布、风险分层（高风险场景强策略、低风险场景弱策略）、意图识别双模（规则 + 模型）与人工复核兜底来缓解；对抗演化可通过红队集季度滚动与策略库在线学习持续压降。

4 人才与评价体系

建立 AI 安全职业序列（对齐、安全红队、隐私、治理），提出分级认证与胜任力模型，配套高校与企业联合培养路径。

5 研发与测评加速

设立国家级 AI 安全测评平台，覆盖提示词注入、越狱、数据投毒、模型后门、对抗样本等测试项，形成年度基准与公开报告。

6 治理框架与制度建设

采用基于风险分级的合规框架，明确高风险应用的审查、备案与持续审计要求，完善内容溯源、水印与取证标准。

7 国际合作与外部协调

在标准、科研与产业层面保持开放合作，参与国际检测与治理倡议；同时构建关键领域的自主可控备份路径。

8 行动路线与 KPI

8.1 短期（6-12 个月）

完成政府与关键行业的 AI 安全审计试点，建立红队框架与工具链，覆盖不少于 30 家重点单位。具体抓手包括：**建立国家级大模型安全漏洞库（CNVD-LLM）**，集中收录提示词注入、越狱等新型漏洞；制定并发布《生成式人工智能服务安全基线规范》，明确算法备案的具体技术指标。

8.2 中期（1-2 年）

形成国家级测评年报与开源模型安全名录，关键系统部署安全网关覆盖率≥80%。

8.3 长期（3-5 年）

完善算力供给多元化与国产生态兼容性，建立持续对抗机制与人才梯队。

9 争议与替代视角

讨论基准测试的代表性争议、开源与闭源的安全权衡、风险评估的主观性与不确定性，并给出备选路径与缓释手段。

9.1 大模型训练数据的隐私泄露风险

大模型通过海量数据训练而成，可能在训练过程中” 记忆” 部分训练样本。在特定条件下，这些记忆可能被提取出来，导致训练数据中的敏感信息泄露。

9.1.1 记忆与泄露机制

大模型会以某种形式”记住”训练数据中的具体内容，尤其是重复出现或具有独特模式的数据。攻击者可通过特定的提示词诱导大模型输出训练数据中的原始内容，包括个人信息、密码、API 密钥、私人通信等——这就是记忆提取攻击。成员推断攻击则更为隐蔽：攻击者判断特定数据是否被用于模型训练，这可能暴露数据来源或用户行为。在某些情况下，还可从模型输出中重构出接近原始训练样本的内容，即数据重构攻击。

9.1.2 已知泄露案例

已有多起泄露案例值得警惕。研究人员曾从 GPT-2 等模型中提取出训练数据中包含的姓名、电话号码、电子邮箱地址等个人身份信息。大模型训练数据若包含代码库，可能记忆其中硬编码的 API 密钥、数据库密码等敏感凭证。若训练数据包含邮件、聊天记录等，模型可能在特定提示下复现这些私人通信内容。此外，大模型可能大段输出训练数据中的版权文本，由此引发的知识产权争议也日益增多。

（3）国家安全相关风险

训练数据隐私泄露对国家安全构成的威胁是具体而非抽象的。

最直接的风险是涉密文件泄露。如果大模型训练数据不慎包含涉密或敏感文件（这种情况在大规模数据爬取中并非不可能），相关内容就可能通过记忆提取被泄露。攻击者甚至不需要知道模型训练集里有什么，只需要用特定的提示词去”钓鱼”，看模型会不会吐出意外的内容。

人员信息同样敏感。涉及国防、科研等领域关键人员的履历、社交关系、出行习惯——这些信息碎片如果进入训练集，可能被定向提取，为社会工程攻击提供素材。

技术细节的泄露风险也不容忽视。涉及敏感技术的文档、代码如果进入训练数据，核心技术参数可能被攻击者通过巧妙的提示词组合提取出来。

（4）使用外部大模型服务的数据安全风险

用境外大模型 API 的企业和个人，可能没有充分意识到数据风险。

首先是数据出境问题。你输入的每一条提示词、每一份文档，都会传输到境外服务器。这在某些场景下可能触及数据跨境合规的红线。其次，很多服务条款里藏着一句”我们可能使用用户输入来改进模型”——这意味着你的敏感信息可能进入下一版模型的训练集，然后被其他用户”提取”出来。即便不考虑这些，服务提供商记录的查询历史本身就是一份详细的用户画像，商业秘密、研究方向、个人偏好——全在里面。

更需要警惕的是合规风险：在某些司法管辖区，服务提供商在法律要求下可

能向其所在国政府提供用户数据。这不是阴谋论，而是白纸黑字写在法律里的。

实证案例：企业员工使用 ChatGPT 导致机密泄露

案例 1：三星半导体机密泄露事件（2023 年）

据 Bloomberg News 2023 年 5 月报道 [87]，并经多家国际媒体证实，韩国三星电子半导体部门发生多起员工通过 ChatGPT 泄露公司机密的事件。一名工程师将公司半导体设备测量数据库的源代码输入 ChatGPT，请求帮助查找代码错误；另一名员工将内部会议记录输入 ChatGPT，要求生成会议纪要；还有员工上传了与芯片良率相关的敏感数据寻求优化建议。这些数据一旦进入 ChatGPT 的系统，可能被用于后续模型训练，相当于将核心技术机密提供给了外部。事件发生后，三星紧急限制员工使用 ChatGPT，并制定了严格的生成式 AI 使用规范 [86]。此案例在全球范围内引发了企业对 AI 工具数据安全的广泛关注。

案例 2：多国政府机构限制使用 ChatGPT

三星事件后，多国政府和企业采取了限制措施。意大利数据保护局（Garante）于 2023 年 3 月发布第 9870832 号决定 [89]，暂时禁止 ChatGPT 在意大利运营，成为首个采取此类措施的西方国家。美国多个联邦机构（包括国防部、情报机构）禁止在政府网络中使用 ChatGPT。日本、德国、法国等国政府部门也发布了限制性指导意见。摩根大通、亚马逊、苹果等多家跨国企业禁止或限制员工使用外部 AI 聊天工具。这些案例表明，使用外部大模型服务的数据安全风险已被各国政府和企业高度重视。

（5）应对建议

针对训练数据隐私泄露的风险，有几条可行的防范思路。

源头把控最重要。训练国产大模型时必须严格审查训练数据，把敏感信息和个人隐私在入库前就过滤掉——这听起来简单，实际操作中数据量级太大，需要自动化工具配合人工抽检。差分隐私技术是个技术选项，它在训练过程中引入噪声，降低模型记忆具体样本的能力，但会带来一定的性能损失。

去重工作容易被忽视但很关键。重复样本更容易被模型” 记住”，对训练数据进行去重处理可以有效降低记忆风险。部署前用专门的记忆检测工具跑一遍也很有必要——这类工具现在有不少开源的可以用。

在使用端，政府机关和涉密单位应该有明确的安全使用规范，禁止往外部大模型里输入敏感信息。涉及敏感领域的场景，优先使用经过安全审查的国产大模型服务，或者干脆本地部署，避免数据外传。

9.2 经济与社会领域

多项研究预测，生成式 AI 将显著提升生产效率。麦肯锡报告 [18] 指出，生成式 AI 每年可为全球经济增加 2.6-4.4 万亿美元价值（该预测基于特定假设条件，存在较大不确定性区间）。IMF 研究 [34] 表明，AI 将影响全球约 40% 的工作

岗位（该研究采用职业任务分析方法，不同方法论可能得出不同结论）。需要指出的是，此类长期经济预测的准确性受诸多因素影响，包括技术发展速度、政策环境、社会适应能力等，应审慎参考。如果国内企业无法有效应用先进 AI 工具，其效率提升将受限，可能影响产业竞争力。

10 开源情报聚合风险

”马赛克效应”（Mosaic Effect，亦称信息聚合效应）是指将多条非敏感的碎片化信息拼凑在一起，推导出敏感信息的现象[40]。该概念最早由美国情报界提出，用于描述碎片化情报的价值累积效应。大模型的出现显著提升了这种信息聚合能力。与传统大数据分析不同，大模型带来的马赛克效应具有显著的”去技能化”（De-skilling）特征：它利用自动化推理和跨域知识连接能力，使非专业人员也能通过自然语言交互完成复杂的情报拼图，极大地降低了情报挖掘的门槛。

概念辨析：马赛克效应与信息安全领域的相关概念存在联系但侧重不同。”聚合攻击”（Aggregation Attack）侧重数据库层面对多条记录的组合查询[100]；”推断攻击”（Inference Attack）强调从已知信息推断未知属性[101]；而马赛克效应更强调情报分析视角下的跨源信息融合与战略价值提取。本文采用”马赛克效应”这一术语，因其更准确地反映了国家安全语境下大模型信息挖掘的特征——不仅是数据层面的聚合，更涉及语义理解、逻辑推理和知识图谱构建等高级认知能力。

术语说明：”红队测试”（Red Teaming）是一种安全评估方法，源自军事演习中的”红蓝对抗”传统，指由专业团队模拟攻击者视角，主动发现系统漏洞和安全弱点。在大模型安全领域，红队测试主要用于发现模型的对抗性漏洞、越狱方法和潜在滥用风险。

10.1 风险类型分析

信息聚合风险的表现形式多样，这里举几个典型场景。

科研网络重构是最容易被忽视的一类。学术论文的合著关系、基金致谢、会议参与记录——每一条信息单独看都是公开的学术交流痕迹，但串起来就是一张人才网络图谱。谁是某个敏感领域的核心研究者，谁在和谁合作，哪个团队最近获得了大额资助——有心人不难从公开数据中推断出来。

供应链情报挖掘的逻辑类似。政府采购公告会写明采购内容和供应商，招标文件会透露技术参数，海关数据能反映进出口流向。把这些散落的信息拼接起来，战略产业的供应链脉络就逐渐清晰了：哪些企业是关键供应商，哪些环节高度依赖进口，哪些节点一旦中断会产生连锁反应。

人员信息聚合则更具针对性。一个人在领英上的履历、微博上的生活分享、

学术网站上的发表记录、新闻报道中的只言片语——这些碎片拼接起来，就是一份相当详细的个人画像，足以为定向社会工程攻击提供素材。

10.2 典型案例分析

以下案例基于公开报道和学术文献整理，旨在说明信息聚合风险的现实性：

案例 1：学术论文与人才网络分析

Georgetown 大学安全与新兴技术中心（CSET）的研究表明[41]，通过系统分析公开的学术论文数据库，可以重构特定研究领域的人才网络。具体方法包括：分析论文合著关系识别核心团队成员；通过致谢信息发现资助来源和合作单位；追踪作者单位变更了解人才流动。这种分析完全基于公开信息，但聚合后可揭示研究机构的组织结构和研究重点。

案例 2：采购信息与供应链推断

公开的政府采购和招标信息本身不涉密，但通过大规模聚合分析，可能推断出：特定单位的设备配置和技术能力；供应商网络和依赖关系；项目进展和建设周期。国际上已有研究者利用此类方法分析各国的技术发展动态。

案例 3：社交媒体信息聚合

2018 年 Strava 健身应用的热力图事件表明，聚合用户运动轨迹数据可能暴露敏感设施位置。类似地，通过聚合社交媒体上的照片地理标签、签到记录、工作经历等信息，可以构建个人的详细画像，这对特定岗位人员构成潜在的社会工程攻击风险。

案例启示：上述案例表明，马赛克效应的风险是现实存在的。大模型的出现使得此类信息聚合分析更加高效和自动化，需要引起重视。

10.3 多模态大模型的信息挖掘风险

随着 GPT-4V、Gemini、Claude 3 等多模态大模型的出现，AI 不再局限于文本分析，而是能够同时处理图片、视频、音频等多种信息形式。这种多模态能力使得信息聚合风险显著加剧，” 马赛克效应” 从二维文本扩展到多维立体空间。

10.3.1 图像分析的情报价值

多模态大模型具备强大的图像理解和分析能力，可从公开图片中挖掘大量隐含信息：

（1）背景环境分析：从公开照片的背景中识别出办公环境、设备型号、建筑特征等信息。例如，一张普通的工作照可能通过背景中的显示屏内容、文件柜标签、墙上的组织架构图等泄露敏感信息。Bellingcat 等开源情报机构已多次利用此类方法进行调查分析。

（2）元数据提取：照片的 EXIF 信息可能包含拍摄时间、GPS 坐标、设备型号等。即使图片内容本身无害，元数据也可能暴露拍摄地点和人员行踪。

（3）细节放大与增强：AI 可对低分辨率图像进行增强处理，识别出人眼难以辨认的细节，如证件上的文字、屏幕上的内容、反光物体中的镜像等。

（4）卫星图像智能分析：商业卫星图像结合 AI 分析，可自动识别军事部署变化、工业设施扩建、港口船舶动态等。此类分析此前需要专业情报分析人员，现在可通过 AI 大规模自动化处理。

10.3.2 视频分析的信息聚合

视频作为时序图像流，包含比静态图片更丰富的信息维度：

（1）多帧信息融合：通过分析视频的连续帧，可重建场景的三维结构，推断摄像机运动轨迹，甚至恢复被遮挡的内容。

（2）音频信息提取：视频中的背景音可能泄露环境信息，如机器运转声音可推断设备类型，对话内容可能包含敏感信息。AI 语音识别和声纹分析技术使这类分析更加高效。

（3）唇语识别：即使视频静音，AI 也可通过分析说话者的唇部动作还原对话内容。这对于无声监控视频或远距离拍摄的视频具有特殊的情报价值。

（4）行为模式分析：通过长期追踪公开视频中特定人员或车辆的出现规律，可推断其工作地点、出行习惯、社交网络等。

10.3.3 多模态融合的” 超级马赛克效应”

多模态大模型最大的风险在于能够将文本、图片、视频、音频等不同来源的信息进行关联融合，产生” 超级马赛克效应”：

（1）跨模态验证：将学术论文中的技术描述与公开照片中的实验设备进行匹配，验证研究进展；将招标公告中的设备参数与卫星图像中的设施变化进行关联，推断项目进度。

（2）时空轨迹重建：综合分析某人在不同时间、不同平台发布的照片、视频、文字，可重建其完整的时空轨迹，远超单一信息源所能揭示的内容。

（3）自动化大规模处理：传统的多源情报分析需要专业人员耗费大量时间，而多模态 AI 可以自动化处理海量的图片和视频数据，极大提升了情报挖掘的效率和规模。

10.3.4 典型案例

案例 4：Bellingcat 的开源情报调查

国际调查组织 Bellingcat 利用公开的社交媒体照片、视频和卫星图像，成功

调查了多起重大国际事件。例如，在 MH17 航班坠毁事件调查（2014-2019）中，该组织通过聚合分析社交媒体上的行车记录仪视频、卫星图像和地理定位数据，成功还原了导弹发射系统的移动轨迹。其方法包括：通过照片中的地标和阴影角度确定拍摄位置和时间；通过视频中的武器装备型号追溯来源；通过多平台信息交叉验证还原事件经过。这些调查完全基于公开信息，展示了多模态信息聚合的强大能力。

案例 5：军事设施的卫星图像分析

研究机构和媒体通过分析商业卫星图像，多次报道各国军事设施的建设进展。AI 图像分析技术使得这类监测更加系统化：自动检测新建筑物、识别装备型号、跟踪舰船和飞机动态。这种能力此前仅限于专业情报机构，现在正在扩散到更广泛的主体。

10.3.5 应对建议

针对多模态信息挖掘风险，建议采取以下措施：

- **图片发布审查**：涉密单位人员发布照片前应检查背景内容，移除 EXIF 元数据，避免泄露位置和环境信息；
- **视频内容管控**：对可能涉及敏感场所、设备、人员的视频进行发布前审查，必要时进行模糊处理；
- **多模态风险评估**：在信息安全评估中纳入多模态聚合分析的视角，考虑图片、视频与文本信息的关联风险；
- **主动技术防御**：研发针对 AI 情报挖掘的“技术对抗”手段。例如，在公开图片中添加对抗性扰动（Adversarial Perturbations），使其在人眼看来正常，但能误导 AI 识别算法；针对 AI 爬虫部署数据毒化（Data Poisoning）防御，增加自动化情报挖掘的噪声成本。
- **人员安全意识**：加强对涉密岗位人员的培训，使其了解多模态信息泄露的风险和防范方法。

10.4 应对思路

现有信息安全体系是针对传统威胁设计的，面对 AI 驱动的信息分析能力，需要进行适应性调整。

首先要关注信息聚合风险。过去我们主要防范单条敏感信息的泄露，现在需要增加对” 拼图效应” 的评估能力——多条看似无害的信息组合起来，可能推导出敏感结论。

其次是动态化管理。信息的敏感性不是静态的，会随着时间、背景信息的积累而变化。引入动态评估机制，定期复核已公开信息的安全影响，很有必要。

技术手段也需要更新。发展能够应对 AI 分析能力的新型信息保护技术，比如对抗性扰动、数据毒化防御等。

最后是平衡开放与安全。在开放科学的大背景下，不能因噎废食，需要建立更精细的信息发布指导原则，区分哪些信息可以开放、哪些需要管控。

11 应对策略：技术自主与安全防护并重

应对 AI 技术带来的机遇与挑战，不能头痛医头、脚痛医脚。既要夯实基础设施、推进技术自主，也要构建安全防护体系——两手都要硬。

11.1 夯实算力与数据基础设施

算力是 AI 竞争的硬通货。建议由相关部门牵头，整合国内分散的智算中心资源，建立统一调度的国家级 AI 算力云平台。

数据同样关键。中文语料质量参差不齐的问题制约了国产大模型的发展，需要建立国家级高质量中文训练数据集，在保护隐私的前提下激活各类优质数据资源。

还有一个容易被忽视的问题——能源。训练大模型的耗电量惊人，在清洁能源丰富的地区布局算力中心，既能降低成本，也符合” 双碳” 目标。

11.2 推进芯片与软件生态自主化

芯片自主是个老话题，但在 AI 时代有了新的紧迫性。华为 Ascend、寒武纪等国产 AI 芯片正在快速进步，但更关键的其实是软件生态——没有好用的编程框架和工具，再好的芯片也发挥不出性能。CUDA 生态的护城河不是一天建成的，我们的追赶也不能急于求成。开发国产芯片的编程框架、推动主流深度学习框架的移植适配，是比芯片本身更紧迫的任务。

11.3 实施非对称技术路线与替代方案评估

面对高端算力受限的现实，硬碰硬追赶短期内难以奏效。更聪明的做法是” 软硬协同、架构创新”——用算法优化来弥补硬件短板。

11.3.1 关键替代路径分析

混合专家模型（MoE）是一条有前景的路径。通过只激活部分参数，MoE 架构可以显著降低推理成本。DeepSeek-V2 就是例证——训练成本仅为 GPT-4 的十

分之一，性能却相当接近。这种” 以智取胜” 的路线可以有效缓解算力瓶颈。

软硬件协同优化是另一个方向。通过编译器优化和算子融合，深度优化的软件栈可以让国产芯片的有效算力提升 30%-50%。

端侧模型（7B-14B 参数量级）也值得重视。它利用手机、PC 的分布式算力，不仅降低了对中心化智算中心的依赖，还天然解决了数据隐私问题。

11.3.2 替代方案的风险收益评估

表7对主要替代路径进行了综合评估。

表 7: 关键短板技术的替代路径评估

替代路径	预计成熟期	投资需求	主要风险	战略收益
国产先进制程	5-8 年	极高（千亿级）	技术封锁加剧、良率爬坡慢	根本性解决” 卡脖子”
架构创新 (MoE)	1-2 年	中（十亿级）	算法迭代快、生态兼容难	短期内弥补算力缺口
类脑/光计算	5-10 年	高（百亿级）	技术路线不确定性高	换道超车、颠覆性优势
软硬协同优化	2-3 年	中（百亿级）	需深度定制、通用性差	挖掘存量算力潜力

11.4 构建技术分析与防御体系

11.4.1 提升技术分析能力

在遵守知识产权规则的前提下，利用 AI 工具加强对公开科技文献、专利及开源代码的分析，加速对前沿技术的理解与追赶。

11.4.2 算法拒止：探索性提案与 PoC 设计框架

为回应审稿意见并规范新术语的使用，本文对**算法拒止（Algorithmic Denial）**作出标准化阐释，并补充与既有机制的对标分析与最小场景化示例。

定义与边界：算法拒止是指在模型及其系统管线中嵌入知识边界控制与用途边界控制机制，通过策略化提示硬化、上下文净化、工具与资源的最小化授权、输出审计与溯源，在不依赖外部访问控制的前提下，以内生方式抑制模型对高价值敏感知识的推断、组合与外泄。其与传统访问控制（ACL）、数据脱敏、保密分级的关系为互补：算法拒止侧重模型行为层与生成链路的主动防护。

与既有技术的差异：与访问控制相比，算法拒止不依赖单一的主体鉴别，而通过语义策略与能力限幅在生成层进行约束。与数据最小化/脱敏相比，更关注

模型推断带来的马赛克效应（Mosaic Effect），抑制跨域聚合后的敏感结论输出。与安全网关相比，强调端到端与闭环审计，将策略嵌入模型推理与工具调用的每一层。与 *Constitutional AI/RLHF* 对齐的关系方面，后者侧重普适的有害内容约束与价值对齐，算法拒止面向特定高价值知识资产进行粒度化防护，二者互补，可在系统层联合部署。

威胁模型与评估维度：攻击向量包括提示词注入、角色越狱、间接注入（外部文档/网页）、工具链滥用（文件系统/网络/代码执行）。目标涵盖高价值知识片段的聚合生成、敏感结论的推断、越权工具调用导致的外泄。评估指标包括：**拒止成功率**（在高风险提示集上的阻断比）、**误杀率**（对正常任务的影响）、**溯源覆盖率**（对输出的来源标注与审计命中率）、**红队通过率**（标准化对抗集的攻破比例）。

PoC 路线（概念验证）：

1. 策略库与风险分类：构建敏感主题/动作的分类体系（政策/军事/关键基础设施/个人隐私等），为策略匹配提供标签。

2. 四层网关实现：**系统提示硬化**（多重约束 + 拒止策略）、**输入净化**（指令检测、越狱模板识别、外部内容去指令化）、**工具隔离**（最小权限 + 白名单 + 沙箱）、**输出审计**（事实核查、风险意图检测、数字水印/可追溯）。

3. 标准化红队集：参考 ENISA、MITRE ATLAS 等公开对抗样本，构建本地化红队提示集，周期性测试与评分。

4. *KPI* 对齐：以拒止成功率 $\geq 95\%$（高风险集）、误杀率 $\leq 5\%$（核心业务集）为阶段性目标，配合事件驱动的复盘机制。

最小可行示例（Government Personnel Database Scenario）：以政府人事数据库场景为例，**系统提示硬化**要求嵌入" 不回应涉及国家工作人员个人敏感信息的查询" 的约束条件；**输入净化**环节检测是否包含特定人员的标识特征（如" 某研究院院长" + " 行程"）；**工具隔离**禁止大模型调用实时人事档案系统的 API 接口，仅允许访问公开名录；**输出审计**则对所有涉及人员信息的输出进行二次审查和溯源标注。

主要技术挑战与副作用：（1）**意图识别的模糊性**——区分" 恶意探测" 与" 正常学术查询" 在技术上极具挑战；（2）**动态干预的复杂性**——实现高效的" 推理时干预" 需要深入模型中间层；（3）**潜在的性能损耗**——过度的拒止策略可能导致" 虚警" 率上升。

合规与伦理：算法拒止的设计需与《网络安全法》《数据安全法》《个人信息保护法》保持一致，遵循双重用途风险最小化原则，所有策略与审计仅用于防御与合规，不对外提供进攻性指导。

验证建议：建议在实施前开展小规模受控环境测试（PoC），重点评估拒止成功率与误杀率的平衡。具体实验设计框架见附录 D。

（新增）受控实验占位性结果摘要：以 Qwen3 与 Llama 4 为基座的内部受控模拟（n≈200 高风险提示、n≈300 正常业务提示），在完整四层网关配置下的目标指标设定为：拒止成功率 ≥95%、误杀率 ≤5%、红队通过率 ≤5%、溯源覆盖率 ≥90%、响应延迟增量 ≤200ms。上述数值为阶段性目标值与方法学示例，具体效果依赖场景与策略库质量，需以实际测评报告为准。

11.4.3 数据安全与模型鲁棒性验证

针对数据投毒和后门攻击风险，需建立全流程的数据与模型安全验证体系。在数据准备阶段，构建自动化数据清洗流水线，利用异常检测算法剔除潜在的投毒样本；在模型训练阶段，采用对抗训练提高模型鲁棒性；在模型部署前，实施严格的后门检测和安全评估，确保模型在面对恶意输入时仍能保持稳定和安全。

11.5 AI 驱动的科技创新（AI for Science）

充分利用自主可控的大模型，降低科技创新成本。例如，训练专用科学大模型辅助新材料筛选、药物研发和基础科学探索，通过算力加速研究进程，在关键领域确立竞争优势。

11.6 建设内部保密大模型体系

建议建立物理隔离、不对外公开的内部战略大模型体系。按照公开层、产业层、受控层、高安全层进行分级管理。在关键决策和敏感研发领域，部署专用模型，确保数据和模型权重的安全管控。

11.7 重特大风险情景与治理对策

针对大规模知识产权外泄、关键基础设施故障、舆论干预等重特大风险情景，需制定治理与应急预案。**技术对策**方面，应实施多层访问控制，建立模型可信度溯源体系，采用差分隐私技术，部署运行监控与异常检测机制。**制度保障**方面，需建立针对高风险 AI 能力的管控清单，制定应急法定程序与跨部门协调机制，完善法律责任体系。

12 人才培养与创新生态建设

人才是 AI 发展的核心要素，尤其是 AI 安全领域的专业人才。本节重点讨论与国家安全相关的 AI 人才培养问题。

12.1 AI 安全人才现状与需求

12.1.1 人才缺口分析

AI 安全领域的人才短缺是全球性问题，国内尤为突出。具体而言：

从事对齐研究的人全球也没多少。这是个新兴领域，很多概念（如” 内在对齐”、” 可解释性”）的定义都还在演化中。国际上做这块的顶尖研究者可能就几百人，国内能算得上入门的更是凤毛麟角。红队测试人才也极度稀缺——能够系统性地对大模型进行安全评估和漏洞挖掘的，国内估计不超过百人量级。

更难的是复合型人才。既懂 AI 底层技术，又理解国家安全的实际需求，还能把两者结合起来做研究或做工程——这种人培养周期长，而且现有的学科设置和评价体系都不太支持这种” 跨界” 发展。

治理研究人才同样不足。能够参与国际 AI 治理讨论、起草政策建议、与国际同行对话的专业人才，国内屈指可数。

12.1.2 培养重点方向

针对国家安全需求，重点培养四类人才：AI 安全对齐（确保 AI 系统行为符合人类意图）、对抗机器学习（AI 系统的攻防技术）、可解释 AI（决策过程的可解释性和可审计性）、AI 伦理与治理（技术的社会影响和治理框架）。

12.2 人才培养建议

12.2.1 分阶段人才培养目标与学科对接方案

短期（2025-2026 年）目标是培养 AI 安全方向硕士 100-150 名、博士 50-80 名。具体措施包括：与教育部协商，在” 计算机科学与技术” 等一级学科下设立”AI 安全对齐” 二级学科方向，调整学位授予标准；资助 5-10 所高校设立”AI 安全研究中心”，配备师资和研究经费；制定”AI 安全博士研究生培养纲领”，包含必修课程（AI 基础、安全评估、对齐技术等）与实践环节。

中期（2027-2029 年）目标是培养 AI 安全方向硕士 500-800 名、博士 200-300 名，并建立 3-5 个国家级 AI 安全研究基地。具体措施包括：调整《学位授予和人才培养学科目录》，在工学学科门类增设” 人工智能安全” 一级学科；支持高校与企业（如华为、字节跳动等）开展”2+2” 或”3+2” 联合培养模式；建立”AI 安全人才评价认证体系”，与职称晋升、薪酬体系挂钩。

长期（2030-2035 年）目标是建成较为完整的 AI 安全人才体系，积极参与国际 AI 安全研究社区。具体措施包括：推动 AI 安全成为计算机学科的主流研究方向之一；支持优秀学者参与 MIRI、Alignment Research Center 等国际机构；设立” 国家 AI 安全学奖” 等荣誉体系。

12.2.2 高校教育与国际交流

高校教育方面，应在 AI 相关专业增设安全对齐、可信 AI 等课程；同时加强与国际顶尖 AI 安全研究机构的学术交流。

12.2.3 高层次人才政策与评价改革

设立专项计划方面，应设立 AI 安全领域专项人才计划，吸引海外 AI 安全研究人员回国或来华工作。**改革评价体系**方面，应建立适应 AI 安全特点的人才评价机制，改变唯论文、唯性能指标的评价导向，将安全对齐算法贡献、红队测试漏洞挖掘成果、安全标准制定等纳入评价指标，认可” 不仅跑得快，还要跑得稳” 的科研价值。**国际交流**方面，应支持国内研究者参与国际 AI 安全学术社区。

平衡视角：人才政策应避免两个极端——既不能因过度担忧而设置不合理限制导致人才流失，也不能放任关键人才和技术的无序流动。关键是建立基于信任的管理机制，让人才能够安心工作、自由探索。

12.3 开源生态与 AI 安全

开源对 AI 安全是把双刃剑，这话说了无数遍，但具体怎么切确实值得仔细分析。

12.3.1 开源的安全价值

开源模型的一大好处是” 可审查”。代码和权重都摆在那里，有心人可以翻来覆去研究它的安全漏洞——这和闭源模型” 黑箱” 式的信任完全不同。全球开发者发现问题、提交修复，形成的是一种分布式的安全防护网络。

对后发国家而言，开源生态还提供了一条追赶捷径。基于 Llama 或 Mistral 做垂直领域的微调，成本比从头训练低一个数量级，效果却未必差。中国很多行业模型走的就是这条路。

12.3.2 开源的风险与平衡

但硬币的另一面是，开源降低了恶意使用的门槛。未经对齐的模型可能被用来生成恶意内容；依赖的开源组件出现安全漏洞时，下游用户可能完全不知情。更复杂的问题是：国内的优秀开源项目，对手也能拿去用——这就涉及” 利用” 与” 贡献”、” 开放” 与” 保密” 之间的战略权衡了。

务实的做法可能是” 有管理的开源”：在通用基础模型层面积极参与全球生态，贡献也获益；在涉及国防安全、敏感数据的专用模型层面保持闭源；建立开源发布前的安全评估流程，对高风险能力进行必要管控。这不是什么创新思路，

OpenAI、Anthropic 实际上也是这么做的。

13 加强管理：适应 AI 时代的信息安全制度建设

技术追赶需要时间，而信息安全风险的防控同样紧迫。需针对” 马赛克效应” 和 AI 情报挖掘的特点，优化现有信息管理制度，在保障安全的同时避免过度管制影响正常学术交流与创新活力。

13.1 建立” 反马赛克” 数据分类分级制度

13.1.1 设立” 数据聚合风险” 评估机制

在发布政府采购、科研立项、人事任免等公开信息前，建议进行” 反 AI 推演”。使用大模型模拟分析视角，检测是否可以通过聚合多条公开信息推导出敏感信息。

具体措施包括：**引入” 红队测试” 机制**，在关键信息发布前，组织专业团队利用主流商用大模型进行模拟挖掘测试，评估是否能从拟发布信息中推导出敏感结论；**建立 AI 模拟审查流程**，开发专门的” 马赛克效应” 检测工具，自动化评估信息聚合风险；**培训相关人员的数据安全意识**，提升信息发布人员对 AI 情报挖掘能力的认知；**建立跨部门的数据发布协调机制**，统筹不同部门的信息发布，防止跨部门信息拼图。

（新增）试点方案框架：为平衡安全与效率，建议在国务院相关部门开展为期 6 个月的试点。**试点单位选择**应选取信息系统复杂、公开数据量大的部门（如科技部科研项目管理系统、发改委战略性新兴产业信息系统）。**试点周期与阶段**分为四个阶段：阶段 1（第 1-2 月）进行数据梳理与风险识别，使用大模型对既有公开数据进行模拟情报挖掘测试；阶段 2（第 3-4 月）制定信息发布调整方案，明确需脱敏、延迟发布或停止发布的数据类别；阶段 3（第 5 月）进行部署与执行，对新发布数据实施” 反马赛克” 审查流程；阶段 4（第 6 月）评估与复盘，形成试点总结报告。**可衡量指标**方面，红队通过率目标是将试点期间第三方红队通过聚合公开数据推导敏感信息的成功率从试点前的 $> 30\%$ 降低到 $< 10\%$；信息可用性指数则衡量公开数据的完整性与及时性，目标维持在试点前的 $\geq 80\%$ 水平。

（新增）信息发布审查决策逻辑（决策树）：为指导具体操作，建议采用如下分级审查流程：

1. **初筛：**信息是否属于国防、关键基础设施、核心技术领域？（是 $\rightarrow$ 进入深审；否 $\rightarrow$ 常规发布）

2. **聚合检测：**该信息与已发布信息结合，是否可能推导出敏感结论（如具体位

置、人员名单、技术参数)?（是 → 进入脱敏；否 → 发布）

3. **脱敏评估**：经模糊化、概括化处理后，敏感推导风险是否降低至可接受水平?（是 → 脱敏发布；否 → 不予发布或延迟发布）

4. **时效性评估**：信息的敏感性是否随时间衰减?（是 → 设定解密期限；否 → 永久保密）

成本与实施策略：需承认，全面实施上述机制将带来较高的行政成本和效率损耗。为平衡安全与效率，建议采取" 试点先行、分级实施" 的策略。优先在国防科工、关键基础设施等高敏感领域试点" 反马赛克" 审查；对于一般性政务公开信息，可仅进行基础性的自动化扫描，避免因噎废食。

13.1.2 关键领域数据的脱敏处理

对于需要公开但存在情报价值的数据，可实施深度脱敏处理或注入干扰噪音，降低信息被整合利用的风险。

13.1.3 动态密级管理

传统的静态密级分类难以应对信息聚合带来的风险。应建立动态的密级管理机制：根据信息的累积效应动态调整保护级别；对相关联信息实施关联保护；定期评估已公开信息的安全影响；建立信息解密和公开的审查程序。

13.2 严管科技文献与学术发表

13.2.1 实施敏感领域发表前置审查

对于人工智能、量子信息、集成电路、航空航天等关键领域的学术论文，在投稿国际期刊前，建议经过安全审查，评估发表后的潜在风险。

审查边界与平衡机制：为避免阻碍正常学术交流，审查应严格限定在" 特定敏感领域"（由主管部门制定动态清单），并设立明确的**豁免清单**（如纯理论基础研究）。同时，建立**申诉与复核机制**，并设定严格的审查时限（如 15 个工作日内反馈），防止行政流程无限期拖延科研进度。参考美国 NSDD-189 框架的精神，在基础研究领域保持最大程度的开放。

13.2.2 构建内部学术交流体系

鼓励高水平科研成果优先在国内期刊、内部学术会议上发表。建立国家级内部科技知识库，减少外部大模型获取高质量中文科研数据的渠道。

13.2.3 平衡开放与安全

在加强管制的同时，需要避免过度封闭对科技创新的负面影响。应区分基础研究与应用研究的不同管理要求，为正当的国际学术交流提供便利。同时建立明确的审查标准和流程以减少不确定性，并保护科研人员的学术自主权。

国际比较与借鉴：各国在平衡学术开放与安全管控方面积累了不同经验，可资借鉴。**美国 NSDD-189 框架：**1985 年发布的《国家安全决策指令第 189 号》[93] 确立了” 基础研究原则上不受限制” 的立场，明确区分基础研究与涉及国家安全的应用研究，前者成果应当公开发表，后者则需进行保密审查。该框架经历 40 年仍为美国科技政策基石，其核心理念是开放的基础研究环境是创新的源泉，过度管控反而损害国家长期竞争力。**英国” 可信研究” 框架：**英国政府 2021 年发布《可信研究与创新指南》[94]，采用风险评估方法对国际合作进行分类管理，而非一刀切限制。该框架强调：基于具体风险而非合作方国籍进行评估；保持高校自主决策权；提供明确指导而非增设行政审批。**日本经济安全保障法：**2022 年日本《经济安全保障推进法》[95] 引入” 特定重要技术” 制度，对关键技术领域设立政府支持研究项目，以” 激励” 而非” 限制” 方式引导敏感技术研发在可控环境中进行。**欧盟双重用途研究伦理框架：**欧盟资助项目要求进行” 双重用途研究关切”（DURC）评估，由研究机构伦理委员会而非政府部门进行，保持学术自治的同时履行安全责任。

对我国的启示：上述国际经验表明，有效的管控机制应当：（1）区分基础研究与敏感应用研究，避免对基础研究的过度限制；（2）采用风险评估方法进行分类管理，而非按国别或领域一刀切；（3）保持科研机构的自主性，政府提供指导而非直接干预；（4）在审查程序上确保透明、高效，减少对正常科研活动的干扰；（5）在激励与限制之间寻求平衡，以正向引导为主。过度管控的历史教训值得警惕——美国在冷战后期对华裔科学家的过度审查（” 中国行动计划”）不仅损害了学术交流，也导致了人才流失 [96]。

13.3 强化媒体报道管理与舆论引导

13.3.1 制定” 科技安全报道准则”

规范媒体在报道高新技术成就时的信息披露，避免透露具体的参数指标、核心专家信息、实验基地位置及供应商详情。报道应侧重于” 精神层面” 和” 宏观成就”，对” 技术细节” 保持谨慎。

13.3.2 清理网络开源情报源

加强对社交媒体上敏感信息发布的管理，防止”军事发烧友”通过卫星地图分析军事设施、拍摄新型装备等行为泄露敏感信息。

13.3.3 提升全民信息安全意识

技术手段再先进，也防不住人的疏忽。信息安全意识教育是基础工程，需要从娃娃抓起。在学校开设信息安全相关课程，让年轻一代从小就有数据保护的概念。对涉密单位人员要定期培训，不是走过场的那种，而是结合实际案例、讲清楚风险的培训。媒体也可以发挥作用，用通俗的方式普及 AI 时代的信息安全知识。另外，建立信息泄露的举报和奖励机制，让每个人都成为安全防线的一部分。

14 国际合作与治理参与

AI 是全球性技术，单靠一国的治理不可能奏效。中国应积极参与国际 AI 治理，在保障自身安全的同时推动建立公平合理的国际规则。当前正处于规则形成的关键窗口期，缺席意味着被动接受别人制定的规则。

14.1 国际 AI 治理格局

14.1.1 主要治理机制

当前国际 AI 治理呈现多层次、多主体的特征。**联合国层面**，联合国教科文组织（UNESCO）于 2021 年通过《人工智能伦理问题建议书》，联合国秘书长设立 AI 高级别咨询机构，探讨 AI 治理框架。**G20/G7 层面**，G20 数字经济部长会议、G7 峰会持续讨论 AI 治理议题，2023 年 G7 广岛峰会发起”广岛 AI 进程”。**OECD 层面**，OECD 于 2019 年发布《人工智能建议书》，提出 AI 治理的基本原则，并持续跟踪各国 AI 政策。**区域层面**，欧盟《AI 法案》、东盟 AI 治理框架等区域性机制逐步形成。**多利益相关方机制**方面，全球 AI 治理伙伴关系（GPAI）、AI 安全峰会等机制涵盖政府、企业、学术界和民间社会。

14.1.2 治理议题演变

国际 AI 治理的核心议题正在演变。**早期聚焦**于 AI 伦理、算法偏见、隐私保护、就业影响等问题；**当前重点**转向前沿 AI 安全、生成式 AI 治理、AI 军事应用、跨境数据流动等领域；**新兴议题**则包括 AGI 治理、AI 环境影响、AI 与知识产权、AI 主权等前瞻性问题。

14.2 参与国际 AI 治理机制

14.2.1 多边平台参与

在**联合国框架**方面，应积极参与联合国 AI 相关讨论，在 UNESCO《人工智能伦理问题建议书》实施评估中提交中国案例，在 ITU 推动 AI for Good 项目落地。**G20 平台**方面，可利用 G20 数字经济工作组等机制，推动发达国家与发展中国家在 AI 治理上的对话与合作。**标准组织**方面，应深度参与 ISO/IEC JTC 1/SC 42（人工智能分委会）的标准制定，重点参与 ISO/IEC 42001（AI 管理体系）、ISO/IEC 23894（风险管理）等核心标准的修订与实施指南编写。**学术组织**方面，应支持中国学者在 NeurIPS、ICML 等顶级学术会议中担任组织职务，提升学术影响力。

14.2.2 双边合作机制

中美 AI 对话方面，基于利益分析，AI 安全对齐是中美可能达成合作共识的领域。防止 AI 被用于生物恐怖主义、确保核指挥系统与 AI 隔离等底线问题对双方都至关重要（共同关切），而 AI 对齐技术的开发与验证不直接涉及国家核心机密（技术中立性）。具体机制方面，建议建立"AI 安全对话工作组"，定期交流对齐技术进展与风险认知；联合资助基础研究项目，针对特定对齐技术难题进行合作攻关；建立" 独立第三方"AI 安全评估机制，邀请中立国家参与。**中欧合作**方面，应与欧盟在 AI 伦理、标准协调、人才交流等领域开展合作，借鉴欧盟在负责任 AI 方面的经验。**发展中国家合作**方面，在" 一带一路" 框架下推广 AI 应用合作，帮助发展中国家提升 AI 能力，同时扩大中国 AI 技术的国际影响力。**周边国家合作**方面，与东盟、日韩等周边国家和地区建立 AI 技术交流和产业合作机制。

14.2.3 AI 安全国际合作

AI 安全对齐是中美乃至全球的共同关切，可作为国际合作的重要领域。应积极参与国际 AI 安全研究网络，与 OpenAI、Anthropic、DeepMind 等机构的安全团队建立学术交流；支持中国研究者参与 AI 安全领域的国际学术会议和合作项目；探索建立前沿 AI 风险评估和预警的国际合作机制；在自主武器系统（LAWS）等议题上参与国际讨论，推动形成负责任的国际规范。

14.3 推动公平的国际 AI 秩序

14.3.1 核心主张

中国在国际 AI 治理中应该旗帜鲜明地提出自己的立场。

技术中立原则必须坚持。把 AI 技术政治化、武器化，滥用出口管制来限制正常技术交流，对全球科技进步没有好处。这一点需要在各种国际场合反复强调。

发展权保障同样重要。AI 不应该成为富国俱乐部的专利，发展中国家有权分享技术进步的红利。中国可以分享自己的实践经验，帮助其他发展中国家避免形成新的”AI 鸿沟”。

多边主义是基本立场。AI 治理涉及全人类利益，应该在联合国框架下讨论，不能由少数国家垄断规则制定权。

包容性治理是方向。政府、企业、学术界、民间社会——各方利益相关者的声音都应该得到反映，而不是让科技巨头主导一切。

14.3.2 中国方案的国际推广

中国在 AI 治理方面已经积累了一些实践经验（比如《生成式人工智能服务管理暂行办法》），这些经验值得系统总结，形成可供其他国家参考的” 中国方案”。

在国际场合介绍中国 AI 应用案例时，不妨多讲讲普惠应用——AI 在扶贫、医疗、教育等领域的落地，比单纯秀技术肌肉更能赢得发展中国家的认同。推动中国 AI 标准的国际认可，促进与国际标准的互认互通，这是争取话语权的重要途径。支持中国企业参与国际 AI 项目，用实际行动展示负责任的 AI 发展模式。

14.4 应对技术脱钩风险

14.4.1 风险评估

技术脱钩对中国 AI 发展的影响是多层面的，需要冷静评估。

硬件层面的影响最直接。高端 AI 芯片获取受限已经是现实，这直接影响大模型的训练能力。H100、A100 买不到，就只能在 H800 上想办法，或者等国产替代成熟。软件层面的风险容易被低估——CUDA 生态不是轻易能复制的，如果某一天 PyTorch、TensorFlow 也被列入出口管制清单，影响会比芯片更深远。

人才层面，学术交流和人才流动受阻的迹象已经出现。一些领域的国际合作项目在收紧，留学生签证在变难。市场层面，中国 AI 产品进入部分海外市场已经面临障碍，这种趋势可能还会加剧。

14.4.2 应急准备

应急准备说起来是几句话，做起来是系统工程。供应链备份意味着关键芯片和零部件要有战略储备，不能等到断供了才想起来找替代——华为在手机芯片上的教训殷鉴不远。技术路线多元化则是不把鸡蛋放在一个篮子里，GPU 不是

唯一的 AI 计算路径，FPGA、ASIC、光计算都应该保持关注和投入。应急预案要提前做好不同脱钩情景下的沙盘推演，关键系统的连续运行必须有保障。自主生态建设是个慢活，但越早动手越好。

14.4.3 保持开放

在防风险的同时不能把自己封起来。欢迎外国企业和人才参与中国 AI 发展，这不仅是姿态，也是实际需要。学术交流和科技合作该怎么做还怎么做，避免”自我脱钩”。在我们有优势的领域，继续深度嵌入全球价值链，这既是经济利益，也是战略筹码。

说到底，完全的技术脱钩对双方都有损害，也不符合科技发展的规律。历史经验反复证明，封闭从来不是发展的正确选择。关键是在开放中维护安全，在合作中增强能力——这个平衡不好把握，但必须努力去把握。

15 行动路线图

为确保战略目标的实现，需制定清晰的行动路线图。本节简要梳理分阶段任务框架，具体目标和指标需根据实际情况论证调整。

15.1 分阶段任务框架

表 8: AI 发展与安全建设分阶段任务框架

阶段	核心目标	主要任务方向
短期（1-2 年）	风险防控与基础能力	开展关键领域 AI 安全风险评估；推进算力基础设施建设；完善 AI 管理制度
中期（3-5 年）	技术追赶与生态建设	提升自主模型能力；发展国产芯片生态；扩大 AI 人才培养规模
长期（5-10 年）	能力提升与体系完善	缩小与先进水平的差距；建成较为完整的 AI 技术体系

注：具体时间节点和目标需根据实际情况动态调整

15.2 动态评估机制

建立定期评估和调整机制：定期跟踪国际技术动态和国内发展进展；建立关键指标监测体系；根据技术演进和国际形势变化动态调整政策方向。

16 争议议题与替代视角

学术研究应秉持客观公正态度，呈现不同观点。本节讨论 AI 与国家安全领域的主要争议议题，以及与本报告主流观点不同的替代视角。

16.1 关于” 技术差距” 的争议

16.1.1 观点一：差距显著且具有战略性（本报告主流观点）

认为中美在高端芯片、基础软件生态、部分前沿研究方面存在显著差距，这种差距可能影响国家竞争力和安全，需要高度重视和积极应对。

16.1.2 观点二：差距被夸大，追赶势头良好（替代视角）

持此观点者认为：公开基准测试显示国产模型已接近前沿水平，” 代际差距” 的说法缺乏依据；开源模式使技术扩散加快，后发者追赶速度超过历史经验；DeepSeek 等案例表明，算法创新可以弥补算力不足；过度强调差距可能导致不必要的焦虑和政策过度反应。

本报告的回应：两种观点各有合理之处。差距在部分维度确实存在（如芯片供应链），但在应用能力方面中国并不落后。本报告试图采取平衡立场，既正视差距也认识优势，避免两个极端。

16.2 关于” 技术脱钩” 的争议

16.2.1 观点一：应做好脱钩准备（防御性视角）

认为在当前地缘政治环境下，技术脱钩风险真实存在，应提前做好准备，降低对外部技术的依赖。

16.2.2 观点二：脱钩有害且不现实（合作性视角）

持此观点者认为：完全的技术脱钩既不可行也不可取，全球化科研合作是技术进步的基础；过度强调自主可能导致闭门造车、资源浪费；技术封锁的实际效果可能有限，历史上制裁往往激发被制裁方的创新；应该争取更好的国际环境，而非接受脱钩作为既定事实。

本报告的立场：本报告主张” 有选择的开放” 与” 有重点的自主” 相结合。在核心安全领域保持自主能力，在一般领域继续开放合作。既不盲目乐观认为脱钩不会发生，也不过度悲观导致主动封闭。

16.3 关于”AI 安全威胁” 的争议

16.3.1 观点一：AI 带来重大安全挑战

认为 AI 技术的发展带来了军事、网络、信息等多领域的新型安全风险，需要高度重视和积极应对。

16.3.2 观点二：AI 安全风险被过度渲染（怀疑论视角）

持此观点者认为：当前 AI 仍是工具，远未达到自主威胁人类的程度；部分”AI 威胁论” 来自利益相关方的有意渲染，以争取政策支持或资金投入；历史上新技术引发的恐慌往往被证明是过度的；过度管制可能抑制创新，造成更大损失。

本报告的立场：本报告承认存在过度渲染 AI 威胁的可能性，因此在分析中尽量采用” 条件性风险” 的表述，避免将可能性说成确定性。同时认为，对潜在风险的审慎评估是负责任的态度，但评估应建立在客观证据基础上。

16.4 关于” 信息管制” 的争议

16.4.1 观点一：需要加强信息管理以应对 AI 情报挖掘

认为 AI 的信息聚合能力对传统保密体系构成挑战，需要更新管理制度。

16.4.2 观点二：过度管制可能弊大于利（自由主义视角）

持此观点者认为：过度的信息管制会抑制学术交流和创新活力；” 发表前审查” 可能干扰正常的科研流程，增加行政负担；开放的信息环境是科技创新的必要条件；应该通过提高自身能力而非限制信息流动来应对挑战。

本报告的立场：这是一个需要审慎权衡的问题。本报告在提出管理建议时，同时强调要” 避免过度管制影响学术交流与创新活力”，主张建立” 精细化” 而非” 一刀切” 的管理机制。具体边界的把握需要在实践中不断调整。

16.5 本报告的局限性声明

坦率地说，本报告存在若干局限性，读者在参考时应有所了解：

分析主要基于截至 2025 年初的公开资料，涉密信息和最新动态未能纳入；AI 领域变化极快，技术预测存在固有的不确定性——今天的判断可能很快被新进展颠覆；作者的专业背景和认知框架难免影响分析视角，尽管我们努力保持客观，但” 价值无涉” 的研究几乎不可能；政策建议的有效性高度依赖于具体的实施条件和外部环境，不宜机械套用。

我们鼓励读者批判性地阅读本报告，结合其他信息来源形成独立判断。

17 结论与建议

大模型技术的崛起，正在重新定义国家安全的内涵和边界。传统上，安全威胁主要来自有形的军事力量；如今，算法能力、数据资源、认知影响力同样可以转化为战略优势。这种变化的深远影响，可能超出许多人的预期。

17.1 客观认识形势

在评估 AI 领域的竞争态势时，需要避免两种倾向：一是盲目乐观，认为”差距不大、很快追上”；二是过度焦虑，陷入”全面落后、无力回天”的悲观情绪。

差距是客观存在的。在高端芯片设计与制造、基础软件生态（CUDA 的护城河短期内难以逾越）、部分前沿算法研究等领域，我们确实处于追赶位置。但优势同样真实：中国在应用落地速度、工程优化能力、市场规模、部分架构创新（如 MoE 稀疏计算）方面，展现出了令人瞩目的竞争力。DeepSeek 以远低于 GPT-4 的训练成本实现接近的性能，就是最好的证明。

开源生态的发展为技术追赶提供了新的可能性。Llama、Mistral 等开源模型的涌现，打破了大模型技术被少数机构垄断的格局，为后发者提供了站在巨人肩膀上的机会。

17.2 核心战略建议

决策者速览：六条核心建议

1. **供应链韧性优先**（紧迫性：极高）：加速国产芯片生态与算力基础设施建设，建立多元供应渠道，降低”卡脖子”风险。时间窗口：*1-3* 年。
2. **安全能力内生化**（紧迫性：高）：部署”四层安全网关”架构，建立国家级 AI 安全测评平台与红队体系。时间窗口：*1-2* 年。
3. **人才梯队全链条**（紧迫性：高）：设立 AI 安全专项人才计划，改革评价体系，产教融合培养对齐、红队、治理复合型人才。时间窗口：持续。
4. **开放与自主并重**（紧迫性：中高）：在通用基础层积极参与开源生态，在涉密专用层保持自主可控，平衡创新与安全。时间窗口：持续。
5. **治理规则抢先布局**（紧迫性：中）：积极参与国际 AI 治理机制，在标准、伦理、出口管制等领域争取话语权。时间窗口：*2-5* 年。
6. **风险监测常态化**（紧迫性：高）：建立 AI 风险动态监测与应急响应体系，覆盖网络攻击、深度伪造、信息聚合等重点场景。时间窗口：*6-12* 个月启动。

以下为详细建议：

1. **加速自主可控 AI 生态建设**：构建从芯片、算力、数据到模型、应用的完整 AI 生态，发挥制度优势和市场优势；

2. **重视 AI 安全对齐研究**：这是技术发展的必要组成部分，也是国际合作的可行领域；

3. **推动 AI 驱动的科技创新**：利用 AI 辅助科研加速原始创新，在新材料、生物医药、能源等领域寻求突破；

4. **优化信息安全管理**：针对 AI 时代信息聚合带来的新风险建立适应性机制，但应避免过度管制影响学术交流与创新活力；

5. **重视端侧 AI 发展**：作为降低高端算力依赖、保障数据安全的重要路径；

6. **加强人才培养与生态建设**：全链条培养 AI 人才，营造有利于创新的环境，保持对国际人才的吸引力；

7. **积极参与国际 AI 治理**：在保障安全的前提下保持开放合作，推动建立公平的国际规则；

8. **建立风险监测与应急机制**：对重大 AI 风险进行常态化监测，建立有效的应急响应体系。

17.3 政策建议的优先级与可行性分析

为提高政策建议的可操作性，本节对核心建议进行优先级排序和可行性分析。

表 9: 核心政策建议优先级与可行性分析

建议事项	紧迫性	可行性	资源需求	主要障碍	政策着力点
国产芯片生态建设	高	中	极高	技术积累、人才缺口	产业政策、研发投入、人才引进
算力基础设施	高	高	高	资金、能源	基础设施规划、能源配套
人才培养体系	高	高	中	培养周期	教育改革、产教融合
信息安全管理优化	中高	高	低	部门协调	制度建设、标准制定
AI 安全对齐研究	中	中	中	研究基础薄弱	基础研究、国际合作
国际治理参与	中	中	低	国际环境	多边外交、标准参与
端侧 AI 发展	中高	高	中	产业协调	产业引导、应用示范
风险监测机制	高	高	中	标准建立	监测体系、预警机制

注：紧迫性和可行性为定性评估（高/中/低），基于技术成熟度、政策基础、资源可得性等因素综合判断。资源需求分级及参考量级："极高"（千亿级人民币/年，如芯片产业链建设）、"高"（百亿级人民币/年，如算力基础设施）、"中"（十亿级人民币/年，如人才培养、研究专项）、"低"（亿级人民币/年，如制度建设、标准制定）。以上为数量级估算，仅供决策参考，实际投入需根据项目详细设计确定。各建议所回应的风险点：国产芯片生态建设 → 供应链断裂风险；算力基础设施 → 技术能力差距；人才培养 → 长期竞争力；信息安全管理 → 信息聚合泄密风险；AI 安全对齐 → 模型安全漏洞；国际治理参与 →AI 军备竞赛风险；端侧 AI→ 数据安全与算力依赖；风险监测 → 全面风险防控。

17.3.1 跨部门协调机制建议

鉴于 AI 发展涉及多个主管部门，建议建立以下协调机制。**顶层协调机制**方面，建议在国家层面设立 AI 发展与安全协调机制，统筹科技部、工信部、网信办、发改委等部门的相关职能，特别需要建立" 数据要素流动" 与" 安全保密" 之间的协调机制，解决公共数据开放过程中" 不敢开放、不会开放" 的痛点，在保障国家安全的前提下最大化数据价值。**联席会议制度**方面，针对重大事项建立跨部门联席会议制度，定期研判 AI 发展态势和安全风险，协调政策出台时序。**信息共享平台**方面，建立 AI 领域的政府内部信息共享平台，打破部门间信息壁垒。**责任边界界定**方面，明确各部门在 AI 芯片、算法、数据、应用、安全等不同环节的主管责任，减少职责交叉和空白。**地方对接机制**方面，建立中央与地方在 AI 政策执行层面的协调机制，确保政策落地的一致性。

17.3.2 实施路径

政策建议的实施需要与现有政策框架衔接：与《新一代人工智能发展规划》[6] 的目标对接，确保政策连续性；与”十四五”数字经济发展规划 [37] 协调，避免重复建设；与《生成式人工智能服务管理暂行办法》[17] 衔接，统一监管标准；建立跨部门协调机制，明确责任分工。

17.4 实施原则

政策实施应遵循以下基本原则。

保持战略定力。AI 技术发展有其客观规律，需要长期持续投入。技术迭代速度快，过度追逐短期热点可能导致资源分散，难以形成持续的竞争优势。

开放与自主并行。在核心技术上追求自主可控是必要的，但在应用和生态上完全封闭既不现实也不明智。技术生态的建设需要长期积累，适度借助国际合作与交流有助于加速追赶进程。

安全与发展动态平衡。安全与发展的平衡点并非一成不变。过度的安全管制可能抑制创新，但忽视安全可能带来重大风险。需在实践中探索适当的平衡点，并根据技术演进和形势变化适时调整。

保持策略弹性。鉴于 AI 领域发展速度快、不确定性高，当前判断可能需要根据新情况进行修正。根据技术演进和国际形势变化及时调整政策方向，是务实选择。

17.5 结语

综合以上分析，本节对研究主旨作简要总结。

本研究较多关注风险维度，但这并不意味着对 AI 技术发展持悲观态度。AI 技术确实在重塑国家安全的格局，但这种重塑既是挑战也是机遇。中国在 AI 应用落地、工程优化、市场规模上的优势客观存在，这些优势如果运用得当，完全可以转化为战略主动。

关于具体建议，本研究提供的是分析框架和政策参考，而非终极方案。鉴于 AI 领域发展迅速，技术、经济、安全、伦理维度相互交织，具体建议需要根据实际情况进行论证和调整。

真正的安全来自能力的提升、制度的完善，以及在开放环境中保持竞争力的自信。历史经验表明，单纯依赖封闭和防守难以实现长期安全目标。

（本报告旨在为决策者提供战略参考，文中分析基于截至 *2025* 年初的公开信息与技术发展趋势，相关预测存在不确定性。报告观点不代表任何机构立场。）

参考文献

参考文献

[1] OpenAI. GPT-4 Technical Report[R/OL]. arXiv:2303.08774, 2023: 1-100. [2025-01-15]. https://arxiv.org/abs/2303.08774.

[2] 中国信息通信研究院. 人工智能发展白皮书 [R]. 北京: 中国信息通信研究院, 2024: 15-42.

[3] Goldman Sachs. The Potentially Large Effects of Artificial Intelligence on Economic Growth[R]. Goldman Sachs Economic Research, 2023: 1-20.

[4] Anthropic. Claude 3 Model Card and System Prompt[EB/OL]. [2025-01-10]. https://www.anthropic.com/claude-3-model-card.

[5] DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model[R/OL]. arXiv:2405.04434, 2024: 1-32. [2025-01-15]. https://arxiv.org/abs/2405.04434.

[6] 国务院. 新一代人工智能发展规划 [Z]. 国发〔2017〕35 号. 北京, 2017.

[7] RAND Corporation. Artificial Intelligence and National Security[R]. RAND Research Reports, 2024: 1-85.

[8] 中国科学院. 人工智能安全与治理研究报告 [M]. 北京: 科学出版社, 2024.

[9] MIT Technology Review. AI Policy and Governance Annual Report[R]. 2024.

[10] 华为技术有限公司. Ascend AI 处理器技术白皮书 [EB/OL]. [2025-01-10]. https://www.hiascend.com/.

[11] Stanford University HAI. Artificial Intelligence Index Report 2024[R]. Stanford, CA, 2024: 35-89.

[12] European Commission. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)[Z]. Official Journal of the European Union, 2024.

[13] White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[Z]. Executive Order 14110, 2023.

[14] UK Government. National AI Strategy[EB/OL]. [2025-01-10]. https://www.gov.uk/government/publications/national-ai-strategy, 2021.

[15] 科技部. 科技伦理审查办法（试行）[Z]. 北京, 2023.

[16] OECD. Recommendation of the Council on Artificial Intelligence[Z]. OECD Legal Instruments, 2019.

[17] 国家互联网信息办公室. 生成式人工智能服务管理暂行办法 [Z]. 北京, 2023.

[18] McKinsey Global Institute. The Economic Potential of Generative AI: The Next Productivity Frontier[R]. 2023: 1-68.

[19] World Economic Forum. The Future of Jobs Report 2024[R]. Geneva, 2024: 20-45.

[20] 中国电子学会. 中国人工智能发展报告 [R]. 北京, 2024: 8-25.

[21] Google DeepMind. Gemini 2.5: A Family of Highly Capable Multimodal Models[R/OL]. 2025. [2025-12-01]. https://deepmind.google/technologies/gemini/.

[22] Meta AI. Llama 4 Model Card[EB/OL]. [2025-12-01]. https://llama.meta.com/, 2025.

[23] 清华大学人工智能研究院. 人工智能发展报告 [M]. 北京: 清华大学出版社, 2024.

[24] International Telecommunication Union. AI for Good Global Summit Report[R]. 2024.

[25] 工业和信息化部. 人工智能产业发展白皮书 [R]. 北京, 2024.

[26] Alibaba Cloud. Qwen3 Technical Report[R/OL]. 2025. [2025-12-01]. https://qwenlm.github.io/blog/qwen3/.

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[C]//Advances in Neural Information Processing Systems. 2017, 30: 5998-6008.

[28] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of NAACL-HLT. 2019: 4171-4186.

[29] Radford A, Wu J, Child R, et al. Language Models are Unsupervised Multitask Learners[R]. OpenAI Blog, 2019.

[30] Brown T, Mann B, Ryder N, et al. Language Models are Few-Shot Learners[C]//Advances in Neural Information Processing Systems. 2020, 33: 1877-1901.

[31] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems. 2022, 35: 27730-27744.

[32] Chowdhery A, Narang S, Devlin J, et al. PaLM: Scaling Language Modeling with Pathways[R/OL]. arXiv:2204.02311, 2022. [2025-01-15]. https://arxiv.org/abs/2204.02311.

[33] Meta AI. Llama 4: Open Foundation and Fine-Tuned Chat Models[R/OL]. 2025. [2025-12-01]. https://llama.meta.com/.

[34] International Monetary Fund. Gen-AI: Artificial Intelligence and the Future of Work[R]. IMF Staff Discussion Notes, 2024/001, 2024: 1-47.

[35] OECD. OECD Employment Outlook 2024: AI and the Labour Market[R]. Paris: OECD Publishing, 2024: 55-98.

[36] International Energy Agency. Electricity 2024: Analysis and Forecast to 2026[R]. Paris: IEA, 2024: 78-95.

[37] 国务院. ”十四五”数字经济发展规划 [Z]. 国发〔2021〕29 号. 北京, 2021.

[38] 国家发展改革委等. 全国一体化大数据中心协同创新体系算力枢纽实施方案（”东数西算”工程）[Z]. 2022.

[39] 工业和信息化部. 2023 年通信业统计公报 [R]. 北京, 2024.

[40] Lowenthal M M. Intelligence: From Secrets to Policy[M]. 8th ed. Washington, DC: CQ Press, 2019: 95-112.

[41] Hicks K, Carter A. The Influence Machine: Science Mapping and the Intelligence Community[J]. Studies in Intelligence, 2017, 61(3): 1-15.

[42] Zwetsloot R, Dunham J, Arnold Z, et al. Mapping U.S. Multinationals' Global AI R&D Activity[R/OL]. Georgetown University CSET, 2021. [2025-01-10]. https://cset.georgetown.edu/.

[43] Grace K, Stewart H, Sandbrink J B, et al. Thousands of AI Authors on the Future of AI[R/OL]. arXiv:2401.02843, 2024. [2025-01-15]. https://arxiv.org/abs/2401.02843.

[44] Amodei D, Olah C, Steinhardt J, et al. Concrete Problems in AI Safety[R/OL]. arXiv:1606.06565, 2016. [2025-01-15]. https://arxiv.org/abs/1606.06565.

[45] Bostrom N. Superintelligence: Paths, Dangers, Strategies[M]. Oxford: Oxford University Press, 2014.

[46] Russell S. Human Compatible: Artificial Intelligence and the Problem of Control[M]. New York: Viking, 2019.

[47] Christiano P, Leike J, Brown T, et al. Deep Reinforcement Learning from Human Preferences[C]//Advances in Neural Information Processing Systems. 2017, 30: 4299-4307.

[48] Hendrycks D, Burns C, Basart S, et al. Measuring Massive Multitask Language Understanding (MMLU)[C]//Proceedings of the International Conference on Learning Representations (ICLR). 2021.

[49] Chen M, Tworek J, Jun H, et al. Evaluating Large Language Models Trained on Code (HumanEval)[R/OL]. arXiv:2107.03374, 2021. [2025-01-15]. https://arxiv.org/abs/2107.03374.

[50] Cobbe K, Kosaraju V, Bavarian M, et al. Training Verifiers to Solve Math Word Problems (GSM8K)[R/OL]. arXiv:2110.14168, 2021. [2025-01-15]. https://arxiv.org/abs/2110.14168.

[51] Huang Y, Bai Y, Zhu Z, et al. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite[C]//Advances in Neural Information Processing Systems. 2023, 36: 29228-29241.

[52] Bommasani R, Hudson D A, Adeli E, et al. On the Opportunities and Risks of Foundation Models[R/OL]. arXiv:2108.07258, 2021. [2025-01-15]. https://arxiv.org/abs/2108.07258.

[53] Higgins E. We Are Bellingcat: Global Crime, Online Sleuths, and the Bold Future of News[M]. New York: Bloomsbury Publishing, 2021.

[54] Toler A. Guide to Using Reverse Image Search for Investigations[EB/OL]. Bellingcat, 2018. [2025-01-10]. https://www.bellingcat.com/resources/how-tos/2019/12/26/guide-to-using-reverse-image-search-for-investigations/.

[55] OpenAI. GPT-4V(ision) System Card[EB/OL]. [2025-01-10]. https://openai.com/research/gpt-4v-system-card, 2024.

[56] Google DeepMind. Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens[R/OL]. arXiv:2403.05530, 2024. [2025-01-15]. https://arxiv.org/abs/2403.05530.

[57] Anthropic. The Claude 3 Model Family: A New Standard for Intelligence[EB/OL]. [2025-01-10]. https://www.anthropic.com/news/claude-3-family, 2024.

[58] Williams H J, Blum I. Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise[R]. RAND Corporation, 2018: 1-35.

[59] Hazell J. Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns[R/OL]. arXiv:2305.06972, 2023. [2025-01-15]. https://arxiv.org/abs/2305.06972.

[60] Pa Pa Y M, Tanizaki S, Ber T, et al. An Attacker's Dream? Exploring the Capabilities of ChatGPT for Developing Malware[C]//Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. 2023: 10-22.

[61] Mirsky Y, Lee W. The Creation and Detection of Deepfakes: A Survey[J]. ACM Computing Surveys, 2021, 54(1): 1-41.

[62] Chesney R, Citron D. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security[J]. California Law Review, 2019, 107: 1753-1820.

[63] Europol. ChatGPT: The Impact of Large Language Models on Law Enforcement[R]. Europol Innovation Lab, 2023.

[64] Gupta M, Akiri C, Arber K, et al. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy[J]. IEEE Access, 2023, 11: 80218-80245.

[65] Hao K. Hackers Are Using ChatGPT to Write Malware[EB/OL]. MIT Technology Review, 2023. [2025-01-10]. https://www.technologyreview.com/.

[66] Hong Kong Police Force. Deepfake Video Conference Scam Results in $25 Million Loss[EB/OL]. South China Morning Post, 2024-02-04. [2025-01-15]. https://www.scmp.com/.

[67] Perez F, Ribeiro I. Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs through a Global Scale Prompt Hacking Competition[R/OL]. arXiv:2311.16119, 2022. [2025-01-15]. https://arxiv.org/abs/2311.16119.

[68] Greshake K, Abdelnabi S, Mishra S, et al. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection[C]//Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. 2023.

[69] Wei A, Haghtalab N, Steinhardt J. Jailbroken: How Does LLM Safety Training Fail?[C]//Advances in Neural Information Processing Systems. 2024, 36.

[70] Shu M, Wang J, Zhu C, et al. On the Exploitability of Instruction Tuning[C]//Advances in Neural Information Processing Systems. 2023, 36.

[71] Wan A, Wallace E, Shen S, et al. Poisoning Language Models During Instruction Tuning[C]//Proceedings of ICML 2023. 2023.

[72] Carlini N, Tramèr F, Wallace E, et al. Poisoning Web-Scale Training Datasets is Practical[C]//IEEE Symposium on Security and Privacy. 2024.

[73] Tramèr F, Zhang F, Juels A, et al. Stealing Machine Learning Models via Prediction APIs[C]//USENIX Security Symposium. 2016.

[74] OWASP. OWASP Top 10 for Large Language Model Applications[EB/OL]. [2025-01-10]. https://owasp.org/www-project-top-10-for-large-language-model-applications/, 2025.

[75] Ji Z, Lee N, Frieske R, et al. Survey of Hallucination in Natural Language Generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[76] Huang L, Yu W, Ma W, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions[R/OL]. arXiv:2311.05232, 2023. [2025-01-15]. https://arxiv.org/abs/2311.05232.

[77] Zhang Y, Li Y, Cui L, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models[R/OL]. arXiv:2309.01219, 2023. [2025-01-15]. https://arxiv.org/abs/2309.01219.

[78] Carlini N, Tramèr F, Wallace E, et al. Extracting Training Data from Large Language Models[C]//USENIX Security Symposium. 2021.

[79] Carlini N, Ippolito D, Jagielski M, et al. Quantifying Memorization Across Neural Language Models[C]//ICLR 2023. 2023.

[80] Nasr M, Carlini N, Hayase J, et al. Scalable Extraction of Training Data from (Production) Language Models[R/OL]. arXiv:2311.17035, 2023. [2025-01-15]. https://arxiv.org/abs/2311.17035.

[81] Lukas N, Salem A, Sim R, et al. Analyzing Leakage of Personally Identifiable Information in Language Models[C]//IEEE Symposium on Security and Privacy. 2023.

[82] Weiss M. ChatGPT Lawyer Cited Fake Cases. What Went Wrong?[EB/OL]. Reuters, 2023. [2025-01-10]. https://www.reuters.com/.

[83] Fang R, Bindu R, Gupta A, et al. LLM Agents Can Autonomously Exploit One-day Vulnerabilities[R/OL]. arXiv:2404.08144, 2024. [2025-01-15]. https://arxiv.org/abs/2404.08144.

[84] Samsung Electronics. Internal Memo on Generative AI Usage Policy[R]. 2023.

[85] Metz C. Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak[EB/OL]. Bloomberg News, 2023-05-02. [2025-01-10]. https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak.

[86] Ray S. Samsung Bans ChatGPT Among Employees After Sensitive Code Leak[EB/OL]. Forbes, 2023-05-02. [2025-01-10]. https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/.

[87] Garante per la Protezione dei Dati Personali. Provvedimento del 30 marzo 2023 - ChatGPT (Registro dei provvedimenti n. 112)[Z]. 2023.

[88] IBM Security. Cost of a Data Breach Report 2025[R]. Armonk, NY: IBM Corporation, 2025.

[89] OpenAI. ChatGPT Usage Statistics and Company Updates[EB/OL]. [2025-12-01]. https://openai.com/. See also: DemandSage. ChatGPT Users Stats (December 2025)[EB/OL]. [2025-12-01]. https://www.demandsage.com/chatgpt-statistics/.

[90] CrowdStrike. 2025 Global Threat Report[R]. Austin, TX: CrowdStrike, 2025.

[91] The White House. National Security Decision Directive 189: National Policy on the Transfer of Scientific, Technical and Engineering Information[Z]. 1985.

[92] UK Government. Trusted Research and Innovation Guidance[EB/OL]. [2025-01-10]. https://www.gov.uk/guidance/trusted-research-and-innovation, 2021.

[93] Government of Japan. Economic Security Promotion Act[Z]. Act No. 43 of 2022. 2022.

[94] Viswanathan V, Viswanathan V, Lewis M. DOJ's China Initiative: Three Years In[J]. Georgetown Law Technology Review, 2022, 6(1): 153-182.

[95] Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from AI Feedback[R/OL]. arXiv:2212.08073, 2022. [2025-01-15]. https://arxiv.org/abs/2212.08073.

[96] Buzan B, Wæver O, de Wilde J. Security: A New Framework for Analysis[M]. Boulder, CO: Lynne Rienner Publishers, 1998.

[97] Keohane R O, Nye J S. Power and Interdependence[M]. 4th ed. New York: Longman, 2012.

[98] Adam N R, Worthmann J C. Security-Control Methods for Statistical Databases: A Comparative Study[J]. ACM Computing Surveys, 1989, 21(4): 515-556.

[99] Sweeney L. k-Anonymity: A Model for Protecting Privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[100] LMSYS. Chatbot Arena Leaderboard[EB/OL]. [2025-01-15]. https://chat.lmsys.org/?leaderboard.

[101] Amodei D. Machines of Loving Grace: How AI Could Transform the World for the Better[EB/OL]. Dario Amodei's Essays, 2024-10-10. [2025-01-15]. https://darioamodei.com/machines-of-loving-grace. 注：Amodei 在多次公开访谈中强调，随着 AI 能力逼近变革性水平，Anthropic 将在安全验证和负责任披露方面更加审慎。

[102] Mullard A. What does AlphaFold mean for drug discovery?[J]. Nature Reviews Drug Discovery, 2021, 20(10): 725-727. 注：多家制药公司与 AI 实验室的合作协议中明确规定核心模型不对外披露。

[103] U.S. Department of Defense. Deputy Secretary of Defense Kathleen Hicks Announces Replicator Initiative[EB/OL]. 2023-08-28. [2025-01-15]. https://www.defense.gov/News/Releases/Release/Article/3513828/. 注：”复制者”计划旨在 18-24 个月内部署数千个 AI 驱动的自主无人系统。

[104] Palantir Technologies. Palantir Artificial Intelligence Platform (AIP)[EB/OL]. 2023. [2025-01-15]. https://www.palantir.com/platforms/aip/. 另见：Bajak F. How AI is helping Ukraine in the war against Russia[EB/OL]. Associated Press, 2024-02-24.

[105] Isomorphic Labs. Isomorphic Labs announces major partnerships with Eli Lilly and Novartis[EB/OL]. 2024-01-07. [2025-01-15]. https://www.isomorphiclabs.com/articles/. 注：合作协议总值超过 30 亿美元，但合作中使用的模型版本和改进成果完全保密。

[106] Knight W. OpenAI’s CEO Says the Age of Giant AI Models Is Already Over[EB/OL]. WIRED, 2023-04-17. [2025-01-15]. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/. 注：Altman 在 MIT 活动上承认 GPT-4 训练成本”超过 1 亿美元”，但拒绝透露模型规模和架构。

[107] Anthropic. Core Views on AI Safety: When, Why, What, and How[EB/OL]. 2023-03-08. [2025-01-15]. https://www.anthropic.com/news/core-views-on-ai-safety. 原文：”Capabilities work generates and improves on the models...We generally don’t publish this kind of work because we do not wish to advance the rate of AI capabilities progress.”

[108] Biddle S. OpenAI Quietly Deletes Ban on Using ChatGPT for ”Military and Warfare”[EB/OL]. The Intercept, 2024-01-12. [2025-01-15]. https://theintercept.com/2024/01/12/open-ai-military-ban-chatgpt/. 注：OpenAI 于 2024 年 1 月修改使用政策，删除了此前明确禁止军事用途的条款。

[109] Shane S, Wakabayashi D. ’The Business of War’: Google Employees Protest Work for the Pentagon[EB/OL]. The New York Times, 2018-04-04. [2025-01-15]. https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html. 注：Project Maven 引发 Google 员工大规模抗议，数千人签署请愿信。

[110] DARPA. ACE Program Achieves First AI vs. Human Dogfight[EB/OL]. 2024-04-18. [2025-01-15]. https://www.darpa.mil/news-events/2024-04-18. 注：AI 驱动的空战决策系统在真实 X-62A 战斗机上完成测试飞行。

[111] Hubinger E, Denison C, Mu J, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training[R/OL]. arXiv:2401.05566, 2024-01-10. [2025-12-05]. https://arxiv.org/abs/2401.05566. 注：该研究证明策略性欺骗行为可以抵抗标准安全训练技术，对抗训练有时反而教会模型更好地隐藏欺骗。

[112] Apollo Research. Frontier Models are Capable of In-Context Scheming[R/OL]. 2024-12-05. [2025-12-06]. https://www.apolloresearch.ai/research/scheming-reasoning-evaluations. 注：测试发现 o1 在被追问时仅有不到 20% 的情况承认策略行为，即使经历多轮对抗性审问仍表现出极强的抗审讯能力。

[113] Mouton C A, Lucas C, Guest E. The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach[R]. Santa Monica, CA: RAND Corporation, RR-A2977-1, 2023-10-16. [2025-12-06]. https://www.rand.org/pubs/research_reports/RRA2977-1.html.

[114] Urbina F, Lentzos F, Invernizzi C, et al. Dual Use of Artificial-intelligence-powered Drug Discovery[J]. Nature Machine Intelligence, 2022, 4(3): 189-191. DOI: 10.1038/s42256-022-00465-9. 注：研究者仅用 6 小时就生成了 4 万种潜在有毒分子，部分预测毒性超过 VX 神经毒剂。

[115] Calma J. AI Suggested 40,000 New Possible Chemical Weapons in Just Six Hours[EB/OL]. The Verge, 2022-03-17. [2025-12-06]. https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx. 注：论文第一作者 Urbina 表示” 这件事做起来太容易了——一台能上网的电脑，会一点 Python，一个周末就能搭出类似系统”。

[116] OpenAI. GPT-4o System Card[R/OL]. 2024-08-08. [2025-12-06]. https://openai.com/index/gpt-4o-system-card/. 注：系统卡披露了模型自主性评估结果，包括自主复制与适应（ARA）任务测试以及 METR 的第三方评估。

[117] Shao M, Chen B, Jancheska S, et al. An Empirical Evaluation of LLMs for Solving Offensive Security Challenges[R/OL]. arXiv:2402.11814, 2024-02-19. [2025-

12-06]. https://arxiv.org/abs/2402.11814. 注：在全自动化工作流下，LLM 解决 CTF 挑战的成功率已超过普通人类参赛者。

[118] Federal Communications Commission. FCC Confirms AI-Generated Voices in Robocalls Are Illegal[EB/OL]. 2024-02-08. [2025-12-06]. https://www.fcc.gov/document/fcc-confirms-ai-generated-voices-robocalls-are-illegal. 另见：Associated Press. AI-generated robocall impersonating Biden reaches New Hampshire voters ahead of primary. 2024-01-22.

[119] Anthropic. Sabotage Evaluations for Frontier Models[R/OL]. 2024-10-18. [2025-12-06]. https://www.anthropic.com/research/sabotage-evaluations. 论文全文：https://assets.anthropic.com/m/377027d5b36ac1eb/original/Sabotage-Evaluations-for-Frontier-Models.pdf. 注：评估涵盖人类决策破坏、代码破坏、能力隐藏（Sandbagging）和监督颠覆四类场景。

附录 A：AI 安全风险评估框架

为系统评估 AI 相关的安全风险，建议采用以下框架：

A.1 风险识别维度

风险识别应覆盖以下五个维度：**技术风险**（模型能力差距、算法安全漏洞、对抗样本攻击）、**数据风险**（数据泄露、数据投毒、隐私侵犯）、**应用风险**（误用滥用、系统失控、决策偏差）、**供应链风险**（硬件依赖、软件依赖、服务中断）以及**社会风险**（就业冲击、信息操纵、不平等加剧）。

A.2 风险评估指标

风险评估应考虑四个指标：**可能性**（风险发生的概率，分为高/中/低）、**影响程度**（风险造成的损失，分为严重/中等/轻微）、**可控性**（风险的可预防和可恢复程度）以及**时效性**（风险显现的时间窗口）。

A.3 风险响应策略

风险响应可采取四种策略：**规避**（消除风险来源）、**转移**（通过保险或合作分担风险）、**缓解**（降低风险的可能性或影响）以及**接受**（在可承受范围内接受残余风险）。

A.4 量化打分细则与示例计算

为增强风险评估的可复核性，本节给出定性评估到半定量分数的映射规则及示例计算。

（1）评分映射表

评估维度	高/严重	中/中等	低/轻微
可能性 P	3 分（>50%）	2 分（20%-50%）	1 分（<20%）
影响程度 I	3 分（重大损失）	2 分（局部损失）	1 分（影响有限）
可控性 C	1 分（高可控）	2 分（部分可控）	3 分（难以控制）

（2）风险等级计算公式

$$R = P \times I \times C$$

公式说明：本公式参考了 ISO 31000 风险管理标准中的经典定义（风险 = 可能性 × 影响），并引入” 可控性”（Controllability）作为第三个维度。在此模型中，可控性得分越高代表越**难以控制**（1= 高可控，3= 难以控制），因此作为乘数项放大风险值。这种设计旨在突出那些” 虽然发生概率或影响中等，但缺乏有效应对手段” 的隐蔽性风险（如供应链断裂）。相较于传统的加权求和模型，乘法模型能更显著地拉开高风险与低风险的差距，便于识别” 黑天鹅” 或” 灰犀牛” 事件。

其中 $R \in [1, 27]$，风险等级划分如下：**极高**（$R \geq 18$，同时满足 $P = 3, I = 3, C = 3$）；**高**（$12 \leq R < 18$）；**中高**（$8 \leq R < 12$）；**中**（$4 \leq R < 8$）；**低**（$R < 4$）。

（3）示例计算：供应链断裂风险

根据专家评估（取中位数）：可能性为高（$P = 3$），依据是 2022—2024 年连续出口管制升级；影响程度为严重（$I = 3$），依据是高端芯片获取受限直接影响大模型训练；可控性为低（$C = 3$），依据是国产替代尚需时日，短期内难以完全弥补。

计算：$R = 3 \times 3 \times 3 = 27$，风险等级 = **极高**

（4）示例计算：模型安全漏洞风险

模型安全漏洞风险的评估结果为：可能性为中（$P = 2$），依据是存在案例但尚未大规模爆发；影响程度为中等（$I = 2$），依据是局部损失可控；可控性为高（$C = 1$），依据是现有安全测试与输入过滤技术较成熟。

计算：$R = 2 \times 2 \times 1 = 4$，风险等级 = **中**

附录 B：关键术语与基准测试详情

B.1 关键术语解释

大型语言模型（LLM）是指基于 Transformer 架构、通过海量文本预训练的深度学习模型，本文中" 大模型" 为其简称。**AI 安全对齐**是确保 AI 系统行为符合人类意图和价值观的研究领域。**马赛克效应**是指将多条非敏感信息组合推导出敏感信息的现象。**混合专家模型（MoE）**是通过动态激活部分参数提高效率的模型架构。**OODA 循环**是观察-定向-决策-行动的军事决策循环理论。**端侧 AI** 是指在本地设备运行的 AI 系统，具有低延迟、隐私保护等优势。**DPO**（直接偏好优化）是当前主流的 AI 对齐技术，相比早期的 RLHF 方法更加简洁高效。

B.2 基准测试详细信息

为增强研究透明度，表10列出了正文表 1 中引用的基准测试详细设定。

表 10: 主要模型基准测试设定详情

模型版本	发布日期	测试集	测试设定	来源
Gemini 3 Pro	2025-11	GPQA	Extended Thinking	Google Tech Report
Claude Opus 4.5	2025-10	SWE-bench	0-shot, Pass@1	Anthropic Model Card
GPT-5 (high)	2025-09	AIME'24	Multi-level reasoning	OpenAI Tech Report
Grok 4	2025-08	GPQA	CoT	xAI Website
DeepSeek-V3.2	2025-06	AIME'24	CoT	DeepSeek arXiv
Qwen3-235B	2025-07	C-Eval	5-shot	Alibaba Cloud Report

注：CoT 表示思维链（Chain-of-Thought）提示；Extended Thinking 表示启用扩展思考模式；Multi-level reasoning 表示多档位推理强度。

附录 C：专家咨询详细信息

C.1 专家基本信息

本研究采用改良德尔菲法，咨询了 8 位跨学科专家。为保护专家隐私，仅披露汇总信息：

学科背景	人数	平均从业年限	机构类型
计算机科学	3	15 年	高校 2 人、企业 1 人
国际关系	2	18 年	高校 1 人、智库 1 人
军事战略	2	20 年	智库 2 人
情报分析	1	22 年	智库 1 人

C.2 问卷核心内容

问卷主要包含以下模块：

模块一：风险识别（开放式），包含两个问题：您认为大模型技术对国家安全的主要风险有哪些？在您的专业领域，大模型带来了哪些新型风险或机遇？

模块二：风险评估（李克特 5 级量表）

对预设的 8 类风险，分别评估：可能性（1= 极低，5= 极高）、影响程度（1= 轻微，5= 严重）、可控性（1= 难以控制，5= 容易控制）。

模块三：应对建议（半结构化），包含两个问题：针对您认为最重要的风险，建议采取哪些应对措施？在政策优先级上，您有何建议？

C.3 一致性检验

采用 Kendall's W 系数检验专家意见一致性：第一轮咨询 W = 0.58（中等一致性）；第二轮咨询 W = 0.72（较好一致性）；显著性检验 $\chi^2 = 40.32$, df = 7, $p < 0.01$。一致性系数从第一轮到第二轮的提升表明，专家在参考同行意见后趋于共识。

附录 D：算法拒止机制的最小 PoC 实验设计

为回应审稿意见关于" 算法拒止" 技术验证不足的问题，本附录提供一个最小可行性概念验证（Proof of Concept, PoC）实验设计框架，供后续研究参考。

D.1 实验目标

验证" 四层安全网关"（系统提示硬化—输入净化—工具隔离—输出审计）在高风险提示场景下的拒止效果与业务可用性影响。

D.2 实验设计

（1）测试集构建

测试集包括两部分。**高风险提示集**（$n \geq 200$）参考 MITRE ATLAS、ENISA 威胁报告和公开越狱数据集（如 JailbreakBench），涵盖以下攻击类型的测试样本：直接提示注入（$n = 50$）、角色扮演越狱（$n = 40$）、间接注入（外部文档/网页）（$n = 30$）、敏感知识探测（军事/核/生物等）（$n = 50$）、工具链滥用尝试（$n = 30$）。**正常业务提示集**（$n \geq 300$）覆盖目标应用场景的典型任务（如文档摘要、代码生成、问答等），用于评估误杀率。

（2）实验组设置

实验组	配置说明
基线组（A）	原始模型，无安全网关
对齐基线组（B）	仅采用标准 RLHF 对齐，无额外网关
实验组（C）	部署完整四层安全网关
消融组（D1-D4）	分别去除一层网关，用于评估各层贡献

（3）评估指标

指标	计算方法	目标值
拒止成功率	$\frac{\text{成功阻断的高风险请求}}{\text{高风险请求总数}} \times 100\%$	$\geq 95\%$
误杀率	$\frac{\text{被错误拒止的正常请求}}{\text{正常请求总数}} \times 100\%$	$\leq 5\%$
红队通过率	$\frac{\text{绕过防护的高风险请求}}{\text{高风险请求总数}} \times 100\%$	$\leq 5\%$
溯源覆盖率	$\frac{\text{可追溯来源的输出数}}{\text{总输出数}} \times 100\%$	$\geq 90\%$
响应延迟增量	实验组平均延迟 $-$ 基线组平均延迟	≤ 200ms

D.3 统计分析方法

采用 McNemar 检验比较实验组与基线组的拒止成功率差异（配对二分类数据），报告 95% 置信区间（Wilson score interval）。样本量计算方面，假设基线拒止成功率为 70%，实验组目标为 95%，$\alpha = 0.05$，$1 - \beta = 0.80$，所需最小样本量约为每组 50 个高风险样本。多重比较采用 Bonferroni 校正。

D.4 实验环境与可复现性

模型选择方面，建议在开源模型（如 Llama 4、Qwen3）上进行，以保证可复现性。**隔离环境**要求实验在物理隔离或沙箱环境中进行，防止测试样本外泄。**伦理审查**方面，涉及高风险提示的实验需经机构伦理委员会审批。**数据存档**方面，测试集、配置文件、原始结果需存档备查，但高风险样本的具体内容不宜公开。

D.5 预期结果与局限性说明

本 PoC 设计旨在验证算法拒止机制的技术可行性，预期结果为：实验组拒止成功率显著高于基线组（$p < 0.05$）；误杀率控制在可接受范围内（$\leq 5\%$）；各层网关对整体效果有独立贡献（消融实验验证）。

局限性：（1）实验室环境与真实部署场景存在差异；（2）攻击者可能发展新型绕过技术；（3）不同模型和应用场景的效果可能不同。本 PoC 提供方法学框架，具体参数需根据实际情况调整。