大模型强国

战略价值、风险与国家行动

许　达　著

中国移动研究院

2025 年 12 月

[出版社名称]

北京

图书在版编目（CIP）数据

大模型强国：战略价值、风险与国家行动 / 许达著. ——北京：[出版社]，2025.12

ISBN XXX-X-XXXX-XXXX-X

I. □ 大…II. □ 许…III. □ 人工智能—研究 IV. □TP18

中国版本图书馆 CIP 数据核字（2025）第 XXXXXX 号

大模型强国：战略价值、风险与国家行动

著　者：许　达

责任编辑：XXX

出版发行：[出版社名称]

地　址：北京市 XXX

邮政编码：100XXX

印　刷：XXX 印刷厂

开　本：787mm×1092mm　1/16

印　张：XX

字　数：XXX 千字

版　次：2025 年 12 月第 1 版

印　次：2025 年 12 月第 1 次印刷

定　价：XX.00 元

序　言

我们正站在一个历史性的技术转折点上。

大型语言模型（Large Language Models，LLMs）的横空出世，标志着人工智能从”能做什么”进入到”能想什么”的新阶段。这不仅仅是一次技术迭代，而是一场可能重塑人类文明进程的智能革命。从 ChatGPT 在 2022 年底引爆全球关注，到 2025 年 GPT-5、Claude Opus 4.5、Gemini 3 等超级模型的相继问世，大模型技术正以前所未有的速度演进，其影响已经渗透到经济、科技、教育、医疗、文化、国防等几乎所有领域。

在这样的背景下，许达博士撰写了这部《大模型强国》。作为中国移动研究院的资深研究员，许达博士长期从事 AI for Science 和人工智能大模型的研究工作，对这一领域有着深刻的洞察。本书不是一部单纯的技术著作，而是一部具有强烈问题意识和政策关怀的战略研究。作者从国家发展的高度，系统论证了大模型技术对国家各个领域的深远影响，深入分析了技术竞争态势和潜在风险，并提出了富有建设性的政策建议。

本书最鲜明的特点是其强烈的战略意识和紧迫感。作者明确提出，大模型是继蒸汽机、电力、互联网之后的第四代”通用目的技术”，其重要性堪比 20 世纪的”两弹一星”工程，应当给予国家战略层面的最高度重视。这一判断或许会引发争议，但正如作者所言——在技术革命的关键节点，战略判断和决策果断至关重要。历史已经多次证明，错失技术革命的窗口期，代价将是数十年的追赶。

我欣慰地看到，新一代科研工作者不仅关注技术本身，更能将技术置于国家发展和国际竞争的宏观视野中加以审视。我相信，本书的出版将为决策者、研究者和广大读者提供有价值的参考，也期待能够引发更多关于我国 AI 发展战略的深入讨论。

[序言作者姓名]

[职务/头衔]

2025 年 12 月于北京

前　言

为什么写这本书

2022 年 11 月 30 日，OpenAI 发布 ChatGPT，人工智能的历史从此分为” 前 ChatGPT 时代” 和” 后 ChatGPT 时代”。作为一名长期从事人工智能研究的科研工作者，我亲眼见证了这场技术革命的爆发，也深刻感受到它正在如何改变我们的工作方式、思维模式，乃至整个社会的运行逻辑。

然而，在这场全球性的 AI 竞赛中，我同时也感受到一种深深的忧虑。

忧虑来自于差距。尽管中国在 AI 应用落地、工程优化方面展现出令人瞩目的竞争力，但在基础研究、高端芯片、核心算法等” 硬科技” 领域，我们与世界领先水平仍存在不可忽视的差距。更令人担忧的是，随着地缘政治紧张加剧，技术封锁和供应链断裂的风险与日俱增。

忧虑也来自于认识。大模型技术的战略重要性，可能尚未被充分理解。许多人将其视为一项普通的技术进步，未能意识到它正在重构国家竞争力的基础。如果说核技术决定了 20 世纪的战略格局，那么大模型很可能决定 21 世纪的竞争态势——它直接作用于人类最核心的能力：认知与创造。

正是出于这种忧虑和责任感，我决定写这本书。我的目标很明确：**呼吁国家对大模型技术给予最高度的重视**，将其提升到与” 两弹一星” 同等的战略高度。这不是危言耸听，而是基于对技术发展规律和国际竞争态势的冷静判断。

本书的核心观点

本书的核心论点可以概括为以下几点：

第一，大模型是” 通用目的技术”（General-Purpose Technology）。它不是某一个行业的专用工具，而是一个能够赋能几乎所有行业的” 技术引擎”。正如蒸汽机催生了第一次工业革命、电力催生了第二次工业革命，大模型正在催生” 智能革命”。

第二，大模型将全方位影响国家发展。本书从经济、科技创新、社会治理、教育、

医疗、文化、国家安全七个维度，系统论证了大模型对国家的深远影响。任何一个领域的落后，都可能产生连锁反应。

第三，当前正处于关键的战略窗口期。技术范式尚未完全定型，后发者仍有追赶甚至超越的可能。但这个窗口不会永远敞开——一旦领先者形成数据飞轮、人才虹吸、生态锁定效应，后来者将面临” 强者恒强” 的马太效应。

第四，必须以最高战略优先级推进大模型发展。这意味着在算力基础设施、顶尖人才、基础研究、应用生态等方面进行超常规投入，建立最高层级的统筹协调机制，以举国体制与市场机制双轮驱动。

本书的结构

本书分为七章，外加附录：

第一章阐述大模型的技术本质和发展态势，揭示其作为” 通用目的技术” 的战略地位，并提出本书的核心论点。

第二章从七个维度——经济、科技创新、社会治理、教育、医疗、文化、国家安全——系统论证大模型对国家发展的全方位影响。

第三章分析全球 AI 竞争格局，比较主要国家的 AI 战略，评估中国的比较优势与面临的挑战。

第四章深入分析大模型带来的各类风险，包括技术差距风险、供应链风险、信息安全风险、认知安全风险等，构建系统的风险评估框架。

第五章提出应对策略，包括技术能力建设、安全防护体系、” 算法拒止” 机制等创新方案。

第六章讨论人才培养、制度建设、国际合作等保障措施，提出短中长期行动路线图。

第七章总结全书观点，提出核心政策建议，发出” 给予大模型最高度重视” 的呼吁。

致谢

本书的写作得到了许多人的帮助和支持。

感谢中国移动研究院的领导和同事们，为本研究提供了良好的工作环境和资源支持。感谢在专家咨询环节提供宝贵意见的各位专家学者（名单见附录）。感谢家人的理解和支持，在我埋头写作的日子里给予了无尽的包容。

本书的观点仅代表作者个人学术见解，不代表任何机构立场。由于作者水平有限，书中难免存在不足之处，恳请读者批评指正。

如果本书能够引起决策者和社会各界对大模型战略重要性的更多关注，能够为我国AI发展战略的制定提供一些参考，将是作者最大的欣慰。

许　达

2025年12月于北京

目录

第一章　智能革命：大模型的战略地位

本章核心观点

大型语言模型已成为决定国家未来命运的战略性技术。其影响将渗透至经济、科技、教育、医疗、文化、国防等一切领域，其重要性堪比 20 世纪的核技术、航天技术，甚至可能超越之——因为它将重塑人类智力活动本身。

本书呼吁：将大模型发展提升至国家战略的最高优先级，以” 两弹一星” 的决心和力度推进。

1.1　历史性的技术转折点

我们正处于一个技术变革的关键节点。

大型语言模型（Large Language Models，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-5、Claude、Gemini、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。78% 的组织在 2024 年报告使用了 AI，较 2023 年的 55% 大幅提升。

为什么这一技术值得最高度的重视？

因为它不仅仅是一项技术，而是一个” 技术引擎”——它能够赋能几乎所有其他技术和行业。正如蒸汽机催生了第一次工业革命、电力催生了第二次工业革命、计算机和互联网催生了信息革命，大模型正在催生” 智能革命”。

但与前几次技术革命不同的是，大模型直接作用于人类最核心的竞争力——认知与创造。谁掌握了最先进的大模型技术，谁就掌握了放大人类智力的杠杆；谁在这场竞争中落后，谁就可能在未来数十年的全球竞争中处于被动。

1.2　通用目的技术：大模型的本质特征

经济学家将某些技术定义为” 通用目的技术”（General-Purpose Technology，GPT），其特征包括：

1. **渗透性**（Pervasiveness）：能够应用于几乎所有行业和领域
2. **改进性**（Improvement）：能够持续改进，性能不断提升
3. **创新催化**（Innovation Spawning）：能够催生大量互补性创新

历史上被公认的通用目的技术包括：蒸汽机、电力、内燃机、计算机、互联网。每一项通用目的技术的出现，都深刻改变了经济结构和社会形态，重塑了国家间的力量对比。

大模型完全符合通用目的技术的定义，而且可能是迄今为止最具变革性的一种：

- **渗透性**：大模型已经在编程、写作、翻译、客服、教育、医疗、法律、金融、科研等领域得到应用，几乎没有哪个知识工作领域能够完全免受影响
- **改进性**：从 GPT-3 到 GPT-4 再到 GPT-5，模型能力呈指数级提升；从单模态到多模态，从文本到代码再到科学推理，能力边界不断扩展
- **创新催化**：围绕大模型已经形成庞大的应用生态，从 RAG（检索增强生成）到 Agent（智能体），从 AI 编程到 AI 科研，创新层出不穷

更重要的是，大模型直接作用于” 认知”——人类最核心的能力。这使得它的影响可能超越以往任何一项通用目的技术。

1.3　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，情况比许多人想象的要复杂。从 Artificial Analysis 智能指数看（截至 2025 年 12 月），Gemini 3 Pro（73 分）与 GPT-5.2（73 分）并列榜首，紧随其后的是 Gemini 3 Flash（71 分）、Claude Opus 4.5（70 分）、GPT-5.1（70 分），o3 推理模型达到 65 分，与 Grok 4 持平。

值得注意的是，中国模型正在快速追赶——Kimi K2 Thinking 达到 67 分，小米 MiMo-V2-Flash 和 DeepSeek-V3.2 均达到 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。

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。更重要的是，我们在高端芯片、基础软件生态（CUDA）、前沿算法研究等” 硬科技” 领域仍处于追赶位置。

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，更需要在基础研究层面追赶。

1.4　本书的核心呼吁

基于以上分析，本书提出一个核心呼吁：

核心呼吁

应将大模型发展提升至国家战略的最高优先级，以” 两弹一星” 的决心和力度，建立最高层级的统筹协调机制，在算力基础设施、顶尖人才、基础研究、应用生态等方面进行超常规投入。

这不是危言耸听，而是基于以下判断：

1. 大模型是继蒸汽机、电力、互联网之后的第四代” 通用目的技术”，其渗透性和变革性将超越前三者
2. 当前正处于技术范式确立的关键窗口期，先发优势将形成” 强者恒强” 的马太效应
3. 主要大国已将 AI 竞争提升至国家安全层面，技术差距将转化为战略劣势

行动刻不容缓。

第二章　重塑一切：大模型对国家发展的七大影响

大型语言模型作为新一代”通用目的技术”，其影响绝不仅限于技术领域本身。本章将从经济、科技创新、社会治理、教育、医疗、文化与意识形态、国家安全七个维度，系统阐述大模型技术对国家发展的深远影响。

2.1　经济领域：生产力革命与产业重构

大模型正在引发一场深刻的生产力革命。麦肯锡全球研究院估计，生成式 AI 每年可为全球经济增加 2.6-4.4 万亿美元价值，相当于再造一个英国或德国的 GDP。这一影响主要通过以下路径实现：

知识工作效率提升：大模型能够承担大量文字处理、数据分析、代码编写等知识工作，显著提升白领工作者的生产效率。研究显示，使用 AI 辅助的程序员编码效率可提升 55% 以上。

产业结构重塑：AI 将加速传统产业的智能化改造，同时催生全新的产业形态。从 AI 原生应用到智能制造，从个性化教育到精准医疗，新的经济增长点正在涌现。

劳动力市场变革：IMF 研究表明，AI 将影响全球约 40% 的工作岗位。虽然会创造新的就业机会，但转型期的结构性失业和技能错配问题不容忽视。

对于中国而言，能否充分利用大模型技术提升经济效率，将直接影响未来的经济竞争力和产业地位。

2.2　科技创新：AI for Science 与研发范式变革

大模型正在改变科学研究的范式，开启”AI for Science” 的新时代。

加速科学发现：AlphaFold 预测蛋白质结构、AI 辅助药物设计、AI 驱动的材料发现……大模型正在加速人类探索未知的进程。2024 年诺贝尔化学奖授予 AlphaFold 团队，

标志着 AI 在科学研究中的重要地位得到了最高学术荣誉的认可。

研发效率提升：从文献综述到实验设计，从数据分析到论文写作，AI 正在深度嵌入科研工作流程。掌握 AI 工具的科研人员，其产出效率可能是传统方式的数倍。

原始创新催化：更重要的是，AI 可能帮助人类突破认知局限，发现人类难以独立发现的模式和规律，从而催生真正的原始创新。

如果我们的科研人员无法充分利用最先进的 AI 工具，将面临日益扩大的” 科研生产力差距”。

2.3　社会治理：公共服务智能化与治理现代化

大模型为社会治理现代化提供了新的技术手段。

公共服务智能化：智能政务助手可以实现 7×24 小时在线服务，大幅提升政务服务的便捷性和覆盖面。从税务咨询到社保查询，从证照办理到政策解读，AI 正在改变政府与公众的互动方式。

决策支持智能化：大模型可以整合海量数据，辅助政策制定者进行情景分析和效果预测，提升决策的科学性。

风险防控智能化：在舆情监测、应急管理、风险预警等领域，AI 可以帮助政府更快速、更精准地识别和响应各类风险。

当然，AI 在治理领域的应用也面临隐私保护、算法透明、数据安全等挑战，需要在效率与公平、便利与安全之间寻求平衡。

2.4　教育领域：学习革命与人才培养

大模型有望引发教育领域的深刻变革。

个性化学习：AI 可以根据每个学生的学习进度、认知特点、兴趣偏好，提供量身定制的学习内容和路径，真正实现” 因材施教”。

优质教育资源普惠化：AI 教师可以打破地域和资源限制，让边远地区的学生也能获得高质量的教育服务，有望缩小教育鸿沟。

教育评价变革：当 AI 可以完成大量标准化考试任务时，教育的重心将从知识记忆转向创造力、批判性思维、协作能力等更高阶的能力培养。

教育是国家竞争力的根基。谁能率先完成 AI 时代的教育转型，谁就能在未来的人

才竞争中占据优势。

2.5　医疗健康：精准医疗与健康管理

大模型正在深刻改变医疗健康领域。

辅助诊断：AI 已经在医学影像识别、病理分析等领域展现出媲美甚至超越专家的能力。未来，AI 有望成为医生的得力助手，提升诊断的准确性和效率。

药物研发：AI 可以大幅加速药物发现和临床试验过程。传统上需要 10 年以上、耗资数十亿美元的新药研发周期，有望被显著缩短。

个性化医疗：基于患者的基因组数据、生活方式数据和病史，AI 可以提供个性化的预防和治疗方案。

公共卫生：在疫情监测、流行病预测、卫生资源配置等方面，AI 可以提供有力支持。

医疗 AI 的发展水平将直接影响国民健康水平和医疗保障能力。

2.6　文化与意识形态：话语权与价值观传播

大模型对文化传播和意识形态的影响，可能是最深远而又最容易被忽视的维度。

内容生产革命：AI 可以大规模生成文本、图像、音频、视频等内容，将彻底改变内容生产的方式。谁掌握了 AI 内容生成的主导权，谁就可能主导未来的文化叙事。

语言与文化载体：当前主流大模型以英语为主，其训练数据和价值取向不可避免地带有西方文化的印记。如果我们没有强大的中文大模型，就意味着在 AI 时代的文化竞争中处于被动。

认知影响：大模型生成的内容将塑造用户的认知和观念。如果年轻一代长期使用带有特定价值取向的 AI 产品，其思维方式和价值观念可能受到潜移默化的影响。

文化安全和意识形态安全，是国家安全的重要组成部分。在大模型时代，这一领域的竞争将更加复杂和微妙。

2.7　国家安全：传统安全与新型安全交织

大模型对国家安全的影响是全方位的。

情报与信息战：大模型极大提升了开源情报分析能力，同时也带来深度伪造、认知

战等新型威胁。

网络安全：AI 既是网络攻防的新武器，也是新的攻击目标。自动化漏洞挖掘、智能钓鱼攻击、AI 辅助渗透……攻击手段正在升级。

军事智能化：从无人系统到智能指挥，从战场感知到决策支持，AI 正在深刻改变战争形态。

技术主权：在核心 AI 技术上的依赖，本身就构成安全风险。供应链断裂、技术封锁的可能性，迫使我们必须追求自主可控。

信息聚合风险：大模型强大的信息整合和推理能力，带来了” 马赛克效应”——从公开的碎片化信息中推断出敏感情报的风险。

本书第四章将对这些风险进行更深入的分析。

2.8　小结：七大领域的系统性影响

表 2.1: 大模型对国家发展七大领域的影响总览

领域	主要机遇	主要挑战	紧迫程度
经济	生产力提升、新产业催生	结构性失业、数字鸿沟	高
科技创新	研发效率提升、原始创新	科研工具依赖	极高
社会治理	服务智能化、决策科学化	隐私、算法公平	中高
教育	个性化学习、教育普惠	教育模式转型	高
医疗	诊断辅助、药物研发	数据安全、伦理	高
文化	内容创新、文化传播	话语权、价值观	高
国家安全	能力提升、风险防控	新型威胁、技术依赖	极高

综上所述，大模型对国家发展的影响是全方位、系统性的。任何一个领域的落后，都可能产生连锁反应，影响整体竞争力。这正是我们呼吁” 最高度重视” 的根本原因。

第三章　全球竞争：主要国家 AI 战略比较

本章要点

全球 AI 竞争已进入白热化阶段。美国凭借技术生态、人才储备和资本优势保持领先；欧盟以监管框架和价值导向寻求差异化定位；中国在应用落地和工程优化方面展现出强劲竞争力，但在芯片、基础软件生态等” 硬科技” 领域仍需追赶。

核心判断：当前正处于技术范式确立的关键窗口期，先发优势将形成” 强者恒强” 的马太效应。

3.1　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》，对 AI 安全提出具体要求。据斯坦福 AI 指数报告，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。

美国的战略优势体现在几个层面：

人才虹吸效应。全球顶尖的 AI 研究者中，相当比例在美国工作。高薪酬、优越的科研环境、完善的创业生态，使美国成为 AI 人才的首选目的地。这种人才集聚效应形成正反馈——越多人才聚集，越能吸引更多人才。

资本充裕。硅谷的风险投资体系为 AI 创业提供了充足的资金支持。从种子轮到 IPO，完整的融资链条使创新想法能够快速转化为产品和服务。2023-2024 年，全球 AI 领域的风险投资中，美国占据超过 60% 的份额。

完整产业链。从芯片设计（Nvidia、AMD、Intel）到云计算平台（AWS、Azure、GCP），从模型研发（OpenAI、Anthropic）到应用落地，美国构建了从上游到下游的完整 AI 产业链。这种生态优势短期内难以复制。

3.2　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，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。但欧洲面临的挑战同样明显：

人才流失。欧洲培养的顶尖 AI 人才，相当比例被美国科技公司吸引。薪酬差距、科研资源差异、创业环境差异，使” 欧洲培养、美国使用” 成为常态。

产业化不足。欧洲在 AI 基础研究方面有优势，但将研究成果转化为商业产品的能力相对较弱。缺乏世界级的 AI 科技公司，是欧洲 AI 生态的明显短板。

监管与创新的张力。严格的监管可能保护消费者权益和伦理价值，但也可能增加企业合规成本，影响创新活力。如何在两者之间取得平衡，是欧盟面临的持续挑战。

3.3　其他重要参与者

英国：2021 年发布《国家人工智能战略》，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。英国在 AI 安全研究方面具有优势，DeepMind 就是英国 AI 实力的代表。

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。日本的优势在于制造业基础和机器人技术积累。

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。韩国在 AI 芯片供应链中扮演关键角色。

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。印度的人口红利和英语优势，使其在 AI 服务外包和人才输出方面具有潜力。

3.4　主要国家 AI 战略对比

表 3.1: 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

3.5　中国的比较优势与短板

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。

3.5.1　比较优势

市场规模。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。

工程化落地能力。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。

政策支持与基础设施。《新一代人工智能发展规划》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。

架构创新。在混合专家模型（MoE）等特定技术方向上，国内团队展现出不俗的创新能力。DeepSeek-V2、V3 的架构设计得到了国际同行的认可。

3.5.2　需要正视的短板

高端芯片。先进制程芯片（7nm 以下）的自主制造能力仍在追赶中。出口管制使高端 AI 芯片（如 H100/A100）获取受限，直接影响大模型训练能力。

软件生态。CUDA 生态的护城河短期内难以逾越。Nvidia 多年构建的软件生态——编译器、函数库、开发工具、社区支持——形成了强大的用户黏性。国产芯片的软件生态建设还有很长的路要走。

基础研究。在 Transformer 架构创新、训练效率优化、对齐技术等前沿方向，国内研究的原创性贡献仍需加强。

顶尖人才。虽然 AI 人才总量不少，但在最前沿研究方向（如 AI 对齐、可解释性）上的顶尖人才储备不足。

3.6　前沿科研与军事模型获取受限

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。

3.6.1　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。

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、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。

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。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。

3.6.2　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。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。

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。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。

3.6.3　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。

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。

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。

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。

3.7　战略窗口期的判断

当前正处于技术范式确立的关键窗口期。技术范式尚未完全定型，后发者仍有追赶甚至超越的可能。但这个窗口期不会永远敞开——一旦领先者形成数据飞轮、人才虹吸、生态锁定效应，后来者将面临” 强者恒强” 的马太效应。

通过国际比较可得出以下启示：

- 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
- 开放与自主并非对立，需要在两者之间寻求动态平衡

第四章　风险全景：大模型时代的安全挑战

本章要点

大模型带来的安全挑战是多维度的，包括供应链风险、技术差距风险、信息聚合风险、网络安全风险、认知安全风险等。本章构建系统的风险评估框架，为应对策略的制定提供依据。

核心判断：供应链断裂和网络攻击自动化是当前最高优先级的风险。

4.1　风险评估框架

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。本章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。

4.1.1　风险评估方法说明

本章采用的风险评估基于以下方法论框架：

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。

分级标准：可能性分为三级——高（>50%）、中（20%-50%）、低（<20%）。影响程度同样分为三级——严重、中等、轻微。可控性评估则考量现有技术和制度的防控能力，分为高、中、低三级。

4.1.2　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：

技术差距 → 安全风险传导机制示意

第一层：技术差距（能力维度）

↓　芯片算力差距 | 软件生态差距 | 算法前沿差距 | 人才储备差距

第二层：能力影响（功能维度）

↓　模型训练受限 | 应用部署受限 | 创新速度受限 | 安全研究受限

第三层：安全影响（风险维度）

↓　网络攻防失衡 | 信息战能力差距 | 经济竞争力下降 | 关键系统脆弱性

调节因素（可阻断传导链条）

替代技术路径 | 算法优化弥补 | 应用场景优势 | 政策制度保障

4.2　供应链断裂风险

供应链断裂是当前最高优先级的风险，被评定为” 极高” 等级。

4.2.1　风险来源

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，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。

4.2.2　影响评估

影响程度判定为” 严重”，原因包括：

- 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

4.2.3　可控性分析

可控性判定为” 低”，原因包括：

- 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。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。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。

4.3　网络安全新威胁

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。

4.3.1　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。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，使传统基于特征匹配的杀毒软件难以检测。

4.3.2　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。

2024 年，伊利诺伊大学厄巴纳-香槟分校的 Fang 等研究团队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，GPT-4 在 15 个测试 CVE 中成功利用了 13 个。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。

4.3.3　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》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。

4.3.4　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；现在，AI 把这个门槛踩到了地板上。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。

4.4　信息聚合与” 马赛克效应”

” 马赛克效应”（Mosaic Effect）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。大模型的出现显著提升了这种信息聚合能力。

4.4.1　风险类型分析

科研网络重构：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。

供应链情报挖掘：政府采购公告、招标文件、海关数据。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。

人员信息聚合：一个人在不同平台的履历、生活分享、发表记录——这些碎片拼接起来，就是一份相当详细的个人画像。

4.4.2　多模态大模型的信息挖掘风险

随着 GPT-4V、Gemini、Claude 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。

图像分析：从公开照片的背景中识别办公环境、设备型号、建筑特征等信息。Bellingcat 等开源情报机构已多次利用此类方法进行调查分析。

视频分析：通过分析视频的连续帧，可重建场景的三维结构；视频中的背景音可能泄露环境信息；AI 语音识别和唇语分析技术使这类分析更加高效。

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。

4.5　认知安全与深度伪造

2024 年香港那起案件给人留下深刻印象：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。

4.5.1　深度伪造的多重威胁

商业欺诈：冒充高管进行视频会议、伪造授权指令，骗取资金转账或敏感信息。

政治风险：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。

虚假信息工业化：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。

4.5.2　应对手段

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。

4.6　大模型自身的安全漏洞

大模型在被广泛部署的同时，其自身也存在多种安全漏洞。

4.6.1　提示词注入攻击

提示词注入是大模型面临的最普遍安全威胁之一。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。更为隐蔽的间接注入则是在网

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。

4.6.2　越狱攻击

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、情景构造法、多轮对话法、编码绕过等。

4.6.3　后门攻击与数据投毒

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。

4.6.4　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。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。

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。

4.7　风险矩阵与优先级排序

表 4.1: 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
供应链断裂	高	严重	低	极高	全栈自主替代、非对称技术路线
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
AI 军备竞赛	低	严重	低	中高	建立沟通热线、军控条约谈判

4.8　颠覆性风险与新型威胁

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。

4.8.1　欺骗与策略性行为

2024 年 1 月，Anthropic 安全团队发表研究，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” 可以抵抗标准的安全训练技术。

Apollo Research 的评估报告进一步揭示，当前主流前沿模型已具备” 上下文内策略行为” 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、以及在被质询时撒谎掩盖其行为。

4.8.2　生物与化学武器风险

RAND 公司 2023 年的研究报告显示，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。

更直接的证据来自药物研发领域。2022 年，研究者将原本用于预测药物毒性的 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。

4.8.3　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。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

第五章　破局之道：技术能力与安全体系建设

本章要点

应对 AI 技术带来的机遇与挑战，需要双管齐下：既要夯实基础设施、推进技术自主，也要构建安全防护体系。本章提出系统性的应对策略，包括技术能力建设路径、非对称技术路线和安全防护体系设计。

核心思路：软硬协同、架构创新，用算法优化弥补硬件短板。

5.1　技术能力建设路径

5.1.1　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。

5.1.2　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。

5.2　算力基础设施

5.2.1　国家级 AI 算力云平台

建议建设国家级 AI 算力云平台，主要包括：

- 统一调度：整合分散的智算中心资源，提高利用效率
- 分级服务：根据任务类型（训练/推理）和安全级别提供差异化服务
- 成本优化：通过规模效应降低算力成本
- 安全隔离：为涉密任务提供物理隔离的算力环境

5.2.2　能源配套

大模型训练的能耗问题不容忽视。建议：

- 在西部清洁能源丰富地区布局大型智算中心
- 发展液冷等高效散热技术
- 优化训练调度，充分利用谷电时段

5.3　非对称技术路线

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。

5.3.1　混合专家模型（MoE）

混合专家模型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。

5.3.2　软硬件协同优化

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%。

5.3.3　端侧模型

端侧模型（7B-14B 参数量级）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。

5.3.4　替代路径评估

表 5.1: 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

5.4　”算法拒止”机制

5.4.1　概念定义

算法拒止（Algorithmic Denial）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。

5.4.2　与既有技术的关系

算法拒止与访问控制（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**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，抑制跨域聚合后的敏感结论输出
- 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

5.4.3　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。

评估指标包括：

- **拒止成功率**：在高风险提示集上的阻断比，目标 $\geq 95\%$
- **误杀率**：对正常任务的影响，目标 $\leq 5\%$
- **溯源覆盖率**：对输出的来源标注与审计命中率，目标 $\geq 90\%$
- **红队通过率**：标准化对抗集的攻破比例，目标 $\leq 5\%$

5.5　四层安全网关架构

构建” 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：

四层安全网关架构

第一层：系统提示硬化

多重约束与拒止策略模板；结构化系统提示隔离机制，确保不可被用户上下文覆盖

第二层：输入净化

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；支持白/黑名单域

第三层：工具隔离

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

第四层：输出审计与溯源

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）

5.5.1　最小可行示例

以政府人事数据库场景为例：

- **系统提示硬化**：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**：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**：对所有涉及人员信息的输出进行二次审查和溯源标注

5.6　开源与自主的平衡

开源对 AI 安全是把双刃剑。

5.6.1　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。

5.6.2　开源的风险

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。

5.6.3　平衡策略

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

5.7　AI 驱动的科技创新

充分利用自主可控的大模型，降低科技创新成本。

5.7.1　AI for Science

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。

2024 年诺贝尔化学奖授予 AlphaFold 团队，标志着 AI 在科学研究中的重要地位得到了最高学术荣誉的认可。这不是终点，而是 AI 驱动科学发现的开始。

5.7.2　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。

5.8　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：

- **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
- **模型训练阶段**：采用对抗训练提高模型鲁棒性
- **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

第六章　制度护航：人才、治理与国际合作

本章要点

技术能力建设需要制度保障。本章讨论支撑大模型战略的三大制度支柱：人才培养体系、治理框架和国际合作策略，并提出短中长期行动路线图。

核心理念：人才是根本，制度是保障，合作是补充。

6.1　人才培养体系

人才是 AI 发展的核心要素，尤其是 AI 安全领域的专业人才。

6.1.1　AI 安全人才现状与需求

AI 安全领域的人才短缺是全球性问题，国内尤为突出。

对齐研究人才：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：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。

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：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。

6.1.2　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

6.1.3　分阶段人才培养目标

短期（2025-2026 年）：培养 AI 安全方向硕士 100-150 名、博士 50-80 名。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；资助 5-10 所高校设立”AI 安全研究中心”。

中期（2027-2029 年）：培养 AI 安全方向硕士 500-800 名、博士 200-300 名，建立 3-5 个国家级 AI 安全研究基地。在工学学科门类增设” 人工智能安全” 一级学科；支持高校与企业开展联合培养模式。

长期（2030-2035 年）：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。推动 AI 安全成为计算机学科的主流研究方向之一。

6.2　评价机制改革

6.2.1　改革评价体系

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
- 支持国内研究者参与国际 AI 安全学术社区

6.2.2　高层次人才政策

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。

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。

6.3　治理框架设计

6.3.1　建立” 反马赛克” 数据分类分级制度

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
- **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**：提升信息发布人员对 AI 情报挖掘能力的认知
- **建立跨部门协调机制**：统筹不同部门的信息发布，防止跨部门信息拼图

6.3.2　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

6.3.3　科技文献与学术发表管理

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”，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）。

6.3.4　国际经验借鉴

美国 NSDD-189 框架：1985 年发布的《国家安全决策指令第 189 号》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。

英国” 可信研究” 框架：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。

对我国的启示：有效的管控机制应当：

1. 区分基础研究与敏感应用研究
2. 采用风险评估方法进行分类管理
3. 保持科研机构的自主性
4. 在激励与限制之间寻求平衡

6.4　国际合作策略

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。

6.4.1　参与国际 AI 治理机制

多边平台参与：

- 在联合国框架下积极参与 AI 相关讨论
- 利用 G20 数字经济工作组等机制推动对话与合作
- 深度参与 ISO/IEC JTC 1/SC 42（人工智能分委会）的标准制定
- 支持中国学者在顶级学术会议中担任组织职务

双边合作机制：

中美 AI 对话：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。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。

中欧合作：与欧盟在 AI 伦理、标准协调、人才交流等领域开展合作。

发展中国家合作：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。

6.4.2　推动公平的国际 AI 秩序

技术中立原则：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。

发展权保障：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。

多边主义：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：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。

6.4.3　应对技术脱钩风险

技术脱钩对中国 AI 发展的影响是多层面的，需要冷静评估和应急准备。

应急准备：

- **供应链备份**：关键芯片和零部件要有战略储备
- **技术路线多元化**：GPU 不是唯一的 AI 计算路径
- **应急预案**：提前做好不同脱钩情景下的沙盘推演
- **自主生态建设**：越早动手越好

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；学术交流和科技合作该怎么做还怎么做，避免" 自我脱钩"。

6.5　行动路线图

6.5.1　短期（1-2 年）

- 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
- 建立国家级大模型安全漏洞库（CNVD-LLM）
- 制定并发布《生成式人工智能服务安全基线规范》

6.5.2　中期（2-3 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%
- 建立 3-5 个国家级 AI 安全研究基地
- 培养 AI 安全方向硕士 500-800 名、博士 200-300 名

6.5.3　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
- 建成较为完整的 AI 技术体系
- 在国际 AI 治理中发挥积极作用

6.5.4　政策建议优先级

表 6.1: 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
算力基础设施	高	高	高	资金、能源	基础设施规划
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
AI 安全对齐研究	中	中	中	研究基础薄弱	基础研究
国际治理参与	中	中	低	国际环境	多边外交

6.5.5　跨部门协调机制

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：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。

联席会议制度：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。

责任边界界定：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。

地方对接机制：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。

第七章　结论：行动呼吁

7.1　核心结论

核心结论：呼吁国家给予大模型最高度重视

大型语言模型不仅是一项技术，而是正在重新定义国家竞争力的战略性基础设施。

本书系统论证了大模型对国家发展七大领域的深远影响：它将重塑经济增长模式、加速科技创新、变革社会治理、革新教育形态、提升医疗水平、影响文化传播、重构安全格局。在中美战略博弈的背景下，大模型能力的差距将直接转化为综合国力的差距。

本书的核心呼吁是：应将大模型发展提升至与” 两弹一星” 同等的国家战略高度。这不是危言耸听，而是基于以下判断：

1. 大模型是继蒸汽机、电力、互联网之后的第四代” 通用目的技术”，其渗透性和变革性将超越前三者
2. 当前正处于技术范式确立的关键窗口期，先发优势将形成” 强者恒强” 的马太效应
3. 主要大国已将 AI 竞争提升至国家安全层面，技术差距将转化为战略劣势

行动刻不容缓。

7.2　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。

7.3　核心政策建议

最高优先级建议：将大模型发展上升为国家战略

第 0 条（元建议）：建立” 国家大模型发展领导小组” 或同等级别协调机制，由最高决策层直接领导，统筹科技部、工信部、发改委、教育部、财政部等部门资源，以” 两弹一星” 的决心和力度推进大模型发展。

理由：大模型发展涉及算力基础设施（万亿级投资）、高端人才培养（教育体系改革）、基础研究攻关（科技体制改革）、产业生态建设（产业政策协调）等多个领域，需要最高层级的统筹协调才能形成合力。

六条核心建议：

1. **算力基础设施超常规投入**（紧迫性：极高）：以” 新基建” 力度加速算力中心建设，目标在 3 年内实现智能算力翻两番；加速国产芯片生态建设，建立多元供应渠道。时间窗口：*1-3* 年。

2. **顶尖人才超常规引进培养**（紧迫性：极高）：设立” 大模型人才特区”，给予全球顶尖 AI 人才有竞争力的待遇；改革评价体系，产教融合培养复合型人才。时间窗口：立即启动，持续推进。

3. **基础研究重点攻关**（紧迫性：高）：在 Transformer 架构创新、训练效率优化、对齐技术等前沿方向集中资源攻关。时间窗口：*1-5* 年。

4. **应用生态全面推进**（紧迫性：高）：推动大模型在教育、医疗、科研、政务等领域的深度应用，以应用倒逼技术进步。时间窗口：*1-3* 年。

5. **安全治理体系建设**（紧迫性：高）：部署安全网关架构，建立国家级 AI 安全测评平台与红队体系。时间窗口：*1-2* 年。

6. **国际规则积极参与**（紧迫性：中高）：积极参与国际 AI 治理机制，在标准、伦理等领域争取话语权。时间窗口：*2-5* 年。

7.4　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。

7.5　结语

综合以上分析，本书的核心观点可归纳为以下几点：

第一，大模型是关乎国家命运的战略性技术。它不是一项普通的技术进步，而是继蒸汽机、电力、计算机之后的第四代” 通用目的技术”。它将渗透至经济、科技、教育、医疗、文化、国防等一切领域，重构人类社会的运行方式。对这一技术的掌握程度，将直接决定一个国家在 21 世纪中叶的国际地位。

第二，当前正处于关键的战略窗口期。技术范式尚未完全定型，后发者仍有追赶甚至超越的可能。但这个窗口期不会永远敞开——一旦领先者形成数据飞轮、人才虹吸、生态锁定效应，后来者将面临” 强者恒强” 的马太效应。中国在应用落地、工程优化、市场规模上的优势为追赶提供了基础，但能否转化为持续的技术领先，取决于现在的战略决策和资源投入力度。

第三，必须将大模型发展提升至国家战略的最高优先级。本书强烈建议：以” 两弹一星” 的决心和力度，建立最高层级的统筹协调机制；在算力基础设施、顶尖人才、基

础研究、应用生态等方面进行超常规投入；在开放与自主之间寻求动态平衡，既不闭门造车，也不丧失核心能力。

第四，重视程度应与技术影响相匹配。如果说核技术决定了 20 世纪的战略格局，那么大模型很可能决定 21 世纪的竞争态势。核技术重塑了军事平衡，大模型将重塑认知能力和创新效率——而后者正是国家竞争力的根本来源。对这样一项技术，怎样重视都不为过。

致决策者

历史经验表明，在技术革命的关键节点，战略判断和决策果断至关重要。

蒸汽机时代，英国的领先奠定了一个世纪的霸权；电气化时代，美国和德国的崛起改变了世界格局；信息革命时代，硅谷的创新塑造了数字经济版图。

大模型革命，我们不能错过。

这不仅关乎经济发展和科技进步，更关乎国家安全和民族复兴。未来已来，时不我待。

后　记

写完这本书的最后一个字，窗外已是深夜。

回顾过去一年的写作历程，最大的感受是”紧迫”二字。每次觉得某个章节已经写完，新的模型发布、新的政策出台、新的行业动态，又迫使我不得不更新内容。GPT-5 来了，Claude Opus 4.5 来了，Gemini 3 来了……技术演进的速度，远超我落笔的速度。

这恰恰印证了本书的核心判断：我们正处于一个技术革命的关键节点，时间窗口稍纵即逝。

作为一名科研工作者，我深知本书的局限性。AI 领域发展太快，任何静态的分析都可能很快过时；作为技术背景出身的研究者，我对政策实施层面的理解可能不够深入；本书的许多判断带有一定的主观性，不同视角可能得出不同结论。

但我仍然选择写下这本书，因为我相信：在这样一个关键时刻，提出问题、引发讨论，比追求完美更重要。如果本书能够让更多人认识到大模型的战略重要性，能够为决策者提供一些参考，能够激发更多更深入的研究和讨论，那么写作的目的就达到了。

技术发展有其客观规律，但国家战略的选择权在我们自己手中。

谨以此书，献给所有关心国家科技发展的读者。

许　达

2025 年 12 月于北京

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：

.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
- **社会风险**：就业冲击、信息操纵、不平等加剧

.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（高/中/低）
- **影响程度**：风险造成的损失（严重/中等/轻微）
- **可控性**：风险的可预防和可恢复程度
- **时效性**：风险显现的时间窗口

.3 量化打分细则

.3.1 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）

.3.2 风险等级计算公式

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

其中 $R \in [1, 27]$，风险等级划分如下：

- **极高**：$R \geq 18$
- **高**：$12 \leq R < 18$
- **中高**：$8 \leq R < 12$
- **中**：$4 \leq R < 8$
- **低**：$R < 4$

关键术语与基准测试详情

.4 关键术语解释

大型语言模型（LLM）：基于 Transformer 架构、通过海量文本预训练的深度学习模型，本书中” 大模型” 为其简称。

AI 安全对齐：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：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）：通过动态激活部分参数提高效率的模型架构。

OODA 循环：观察-定向-决策-行动的军事决策循环理论。

端侧 AI：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）：当前主流的 AI 对齐技术，相比早期的 RLHF 方法更加简洁高效。

红队测试：一种安全评估方法，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。

提示词注入：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。

越狱攻击：绕过大模型安全对齐机制，使其输出被禁止内容的攻击方法。

幻觉：大模型生成看似合理但实际上错误或虚构内容的现象。

.5 主要基准测试说明

表 1: 主要基准测试简介

测试名称	测试内容	代表性意义
MMLU	57 个学科的多选题知识测试	通用知识水平
HumanEval	代码生成与编程能力	编程能力
GSM8K	小学数学应用题	数学推理能力
GPQA	研究生级别科学问答	高级推理能力
SWE-bench	软件工程任务	实际编程能力
C-Eval	中文学科知识评测	中文能力
AIME	美国数学邀请赛题目	高难度数学推理

专家咨询详细信息

本书在风险评估部分采用改良德尔菲法，咨询了 8 位跨学科专家。

.6　专家基本信息

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

.7　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）

- 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、影响程度、可控性。

模块三：应对建议（半结构化）

- 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？

.8 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。

算法拒止机制的技术实现

.9 实验目标

验证” 四层安全网关” 在高风险提示场景下的拒止效果与业务可用性影响。

.10 测试集构建

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，涵盖：

- 直接提示注入（$n = 50$）
- 角色扮演越狱（$n = 40$）
- 间接注入（$n = 30$）
- 敏感知识探测（$n = 50$）
- 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。

.11 评估指标与目标值

指标	计算方法	目标值
拒止成功率	成功阻断数/高风险请求总数	$\geq 95\%$
误杀率	错误拒止数/正常请求总数	$\leq 5\%$
红队通过率	绕过防护数/高风险请求总数	$\leq 5\%$
溯源覆盖率	可追溯输出数/总输出数	$\geq 90\%$
响应延迟增量	实验组延迟 − 基线组延迟	≤ 200ms

.12 实验环境要求

- **模型选择：**建议在开源模型（如 Llama 4、Qwen3）上进行
- **隔离环境：**实验在物理隔离或沙箱环境中进行
- **伦理审查：**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：**测试集、配置文件、原始结果需存档备查

参考文献

[1] OpenAI. GPT-4 Technical Report[R/OL]. arXiv:2303.08774, 2023.

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

[3] Goldman Sachs. The Potentially Large Effects of Artificial Intelligence on Economic Growth[R]. 2023.

[4] Anthropic. Claude 3 Model Card and System Prompt[EB/OL]. 2024.

[5] DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model[R/OL]. arXiv:2405.04434, 2024.

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

[7] Stanford University HAI. Artificial Intelligence Index Report 2024[R]. 2024.

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

[9] European Commission. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)[Z]. 2024.

[10] UK Government. National AI Strategy[EB/OL]. 2021.

[11] McKinsey Global Institute. The Economic Potential of Generative AI[R]. 2023.

[12] International Monetary Fund. Gen-AI: Artificial Intelligence and the Future of Work[R]. 2024.

[13] CrowdStrike. 2025 Global Threat Report[R]. 2025.

[14] IBM Security. Cost of a Data Breach Report 2025[R]. 2025.

[15] Vaswani A, et al. Attention is All You Need[C]//NeurIPS. 2017.

[16] Brown T, et al. Language Models are Few-Shot Learners[C]//NeurIPS. 2020.

[17] Anthropic. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training[R/OL]. arXiv:2401.05566, 2024.

[18] Apollo Research. Frontier Models are Capable of In-Context Scheming[R]. 2024.

[19] RAND Corporation. The Operational Risks of AI in Large-Scale Biological Attacks[R]. 2023.

[20] Urbina F, et al. Dual Use of Artificial-intelligence-powered Drug Discovery[J]. Nature Machine Intelligence, 2022, 4(3): 189-191.

[21] OpenAI. GPT-4o System Card[R]. 2024.

[22] Anthropic. Sabotage Evaluations for Frontier Models[R]. 2024.

[23] Fang R, et al. LLM Agents Can Autonomously Exploit One-day Vulnerabilities[R/OL]. arXiv:2404.08144, 2024.

[24] The Nobel Foundation. The Nobel Prize in Chemistry 2024[EB/OL]. 2024.

[25] Google DeepMind. Gemini 2.5: A Family of Highly Capable Multimodal Models[R]. 2025.

[26] Anthropic. Introducing Claude Opus 4.5[EB/OL]. 2025.

[27] OpenAI. Introducing GPT-5.2[EB/OL]. 2025.

[28] Google. A New Era of Intelligence with Gemini 3[EB/OL]. 2025.

[29] GitHub. Research: Quantifying GitHub Copilot's Impact on Developer Productivity[EB/OL]. 2022.

[30] Singhal K, et al. Towards Expert-Level Medical Question Answering with Large Language Models[R/OL]. arXiv:2305.09617, 2023.

注：完整参考文献列表请参见原论文版本。本书参考文献采用 *GB/T 7714-2015* 著录规则。