

2026

网络安全人才实战能力报告
AI赋能篇

编写组织

指导单位

国务院学位委员会学科评议组(网络空间安全组)
教育部高等学校网络空间安全专业教学指导委员会

主编单位

北京航空航天大学
中国科学技术大学
永信至诚科技集团股份有限公司

副主编单位

华中科技大学
西安电子科技大学
上海交通大学
山东大学
北京邮电大学
东南大学
暨南大学
武汉大学
北京理工大学
湖南大学
哈尔滨工业大学
西北工业大学
天津大学
网络空间部队信息工程大学
清华大学网络科学与网络空间研究院
西安交通大学
南开大学
南京信息工程大学
《网络空间安全科学学报》

参编单位

福建师范大学
河南理工大学
常州信息职业技术学院
中国联合网络通信集团有限公司
华为技术有限公司
中国移动通信集团有限公司
中国烟草总公司福建省公司

编委会主任：方滨兴
编委会执行主任：李小勇 封化民
编委会副主任：俞能海 蔡晶晶
主 编：刘建伟
副 主 编：(按照姓氏首字母排序)

程 光	付 磊	付章杰
耿光刚	胡思煌	姜 伟
刘功申	刘哲理	吕勇强
单 征	沈 超	沈玉龙
苏 洲	唐 卓	王 骞
王 震	魏晶晶	谢 雪
张 超	张 丽	张卫明
祝烈煌	邹德清	

编 委：(按照姓氏首字母顺序)

曹 坤	陈 楠	陈晓琳
陈泽楷	陈子涵	程 珂
冯丙文	龚 涛	桂嘉敏
洪荣灿	何秀伟	李 想
李晓宇	刘晓明	吕亚丽
马 莹	潘洋河	宋甫元
吴文涛	王诗萌	王 阳
谢 丹	肖昱翾	许 力
许学伟	叶 麟	袁 斌
张朝昆	赵 乙	赵玉宇
郑乐乐	郑世敏	周天阳
周余阳		

前言

习近平总书记指出，人工智能是引领新一轮科技革命和产业变革的战略性技术，“带来前所未有发展机遇，也带来前所未遇风险挑战”，要统筹人工智能发展和安全，关键在人。既要建设规模宏大、素质优良的人工智能人才队伍，更要锻造一支能够驾驭智能攻防、守护数字空间的网络安全实战人才队伍。

随着大模型、智能代理等技术加速进入政务、金融、能源、医疗、工业互联网等领域，网络安全的对象已进一步延伸至数据、算法、模型、接口以及人机协同流程；与此同时，漏洞挖掘、威胁分析、安全运营、攻防对抗等安全工作也正在向人机协同、智能分析和自动化方向演进。网络安全人才面临的已不再只是“人才数量”的问题，更是如何适应人工智能发展、完成实战能力重构的新课题。

《网络安全人才实战能力报告》编委会持续聚焦我国网络安全人才实战能力研究，已相继发布“攻防实战能力篇”“人才评价篇”“安全测试评估篇”，得到政产学研用各界的广泛关注。在此基础上，编委会回应时代课题，本年度主题定为“AI赋能篇”。

《网络安全人才实战能力报告——AI赋能篇》（以下简称“报告”）聚焦人工智能发展背景下网络安全人才实战能力的变化，面向高校、科研院所及从业人员开展系列调研，结合国内外权威数据，全面呈现AI时代我国网络安全人才实战能力的供需现状、能力短板与发展路径。

《报告》首次系统提出“利用AI、保障AI、对抗AI”三位一体实战能力框架，构建贯穿模型安全、应用安全、AI赋能安全三大场景的“意识—知识—技能—实践”（ASK-P）能力体系，从背景分析、模型安全能力、应用安全保障能力、AI赋能能力、人才培养体系与总结建议六个部分展开论述，面向党政机关、央国企、企事业单位及高校等人才供需单位，为各单位制定AI时代人才战略提供翔实参考。

《报告》主要具有以下特点：

1. 聚焦能力重构，系统呈现AI时代网络安全人才的新要求。区别于以往主要围绕人才数量、岗位需求和传统实战能力展开的研究，《报告》将人工智能作为影响网络安全人

才能力结构的重要变量，分析AI成为网络安全人才基础能力的发展趋势，系统梳理AI时代网络安全人才实战能力的演进路径。

2. 贯通人才培养，探索从能力评价到实战培养的完整路径。《报告》不仅关注“人才需要具备什么能力”，还进一步研究“这些能力如何培养”。通过对高校课程体系、实验体系、教材与教学资源、师资队伍，以及产教融合、竞赛实战、实习实训和在职人员继续教育等情况进行分析，探索AI时代网络安全人才培养的实践路径。

3. 政产学研协同，汇聚网络安全人才培养实践经验。《报告》由国务院学位委员会学科评议组（网络空间安全组）、教育部高等学校网络空间安全专业教学指导委员会指导，北京航空航天大学、中国科学技术大学、永信至诚科技股份有限公司担任主编单位，华中科技大学、西安电子科技大学、上海交通大学、山东大学、北京邮电大学、东南大学、暨南大学、武汉大学、北京理工大学、湖南大学、哈尔滨工业大学、西北工业大学、天津大学、网络空间部队信息工程大学、清华大学网络科学与网络空间研究院、西安交通大学、南开大学、南京信息工程大学、《网络空间安全科学学报》等担任副主编单位，福建师范大学、河南理工大学、常州信息职业技术学院、中国联合网络通信集团有限公司、华为技术有限公司、中国移动通信集团有限公司、中国烟草总公司福建省公司担任参编单位。参与单位涵盖网络安全人才培养、科学研究和产业实践等多个领域，为报告研究提供了多维度的实践基础。

《报告》在编写过程中得到了上述单位及众多专家学者的鼎力支持，在此一并致以诚挚谢意。人工智能技术演进日新月异，加之撰写时间仓促，书中难免存在疏漏与不妥之处，恳请各位读者不吝批评指正。

目录

CONTENTS

第一章 背景分析 01

1.1 全球趋势：人工智能重塑网络安全人才能力	01
1.1.1 欧美AI安全战略下的人才布局	01
1.1.2 国际组织对人才能力的新界定	03
1.2 国内需求：国家战略与产业发展对网络安全人才的要求	04
1.2.1 网络强国战略驱动人才能力升级	04
1.2.2 “人工智能+”行动带来的人才能力需求	04
1.2.3 关键信息基础设施单位人才能力缺口	05
1.3 能力重构：AI时代网络安全人才实战能力演进	05
1.3.1 从“人对抗人”走向“人机协同”	05
1.3.2 AI成为网络安全人才基础能力	07
1.3.3 “利用AI、保障AI、对抗AI”成为要求	08
1.4 AI时代网络安全人才核心能力与岗位族划分	10
1.5 总结与建议	11

目录

CONTENTS

第二章 网络安全人才模型安全能力分析 13

2.1 模型安全风险分析	13
2.1.1 架构风险	13
2.1.2 数据风险	14
2.1.3 对抗风险	14
2.1.4 系统风险	15
2.1.5 环境风险	15
2.1.6 伦理风险	16
2.2 网络安全人才模型安全能力现状分析	17
2.2.1 大模型企业安全建设现状	17
2.2.2 大模型企业人才配置情况	18
2.2.3 网络安全人才模型安全能力现状	22
2.2.4 网络安全人才模型安全能力提升情况	26
2.3 网络安全人才模型安全能力ASK-P体系	29
2.3.1 意识维度	29
2.3.2 知识维度	30
2.3.3 技能维度	30
2.3.4 实践维度	30
2.3.5 人才分层	31
2.4 总结与建议	31

第三章 网络安全人才AI应用安全能力分析

33

3.1 AI 应用安全风险分析	33
3.1.1 提示词注入风险	33
3.1.2 内容合规风险	34
3.1.3 AI接口与插件风险	34
3.1.4 业务安全风险	35
3.1.5 数据安全风险	35
3.1.6 智能体安全风险	36
3.1.7 AI决策失真与责任边界风险	36
3.2 网络安全人才AI应用安全能力现状	37
3.2.1 AI应用安全人才基本情况	37
3.2.2 AI应用安全岗位设置情况	40
3.2.3 网络安全人才AI应用安全能力分析	41
3.2.4 网络安全人才AI应用安全保障挑战分析	47
3.3 网络安全人才应用安全能力ASK-P体系	49
3.3.1 意识维度	49
3.3.2 技能维度	50
3.3.3 知识维度	50
3.3.4 实践维度	51
3.3.5 人才分层	51
3.4 总结与建议	52

目录

C O N T E N T S

第四章 网络安全人才AI赋能能力分析 54

4.1 AI在网络安全领域的实践应用	54
4.1.1 AI赋能安全运营	54
4.1.2 AI赋能漏洞挖掘与攻防对抗	56
4.1.3 AI赋能安全数据分析	57
4.1.4 AI驱动网络安全技术创新	58
4.2 网络安全人才AI赋能能力现状	59
4.2.1 AI安全工具使用能力分析	60
4.2.2 网络安全人才人机协同能力分析	63
4.2.3 不同组织与人员特征的能力差异分析	64
4.3 网络安全人才AI能力ASK-P体系	68
4.3.1 AI辅助低空漏洞挖掘和安全攻防场景	68
4.3.2 AI辅助区块链安全运营与告警研判场景	70
4.3.3 AI辅助网络安全人才训练与靶场内容生成场景	71
4.3.4 AI驱动的工业控制系统安全	73
4.3.5 AI赋能的数据跨境流动安全	74
4.3.6 AI系统自身的安全防护	76
4.4 总结与建议	77

第五章 AI时代网络安全人才培养体系

79

5.1 高校培养体系的现状	79
5.1.1 课程体系建设现状	79
5.1.2 实验体系建设现状	82
5.1.3 教材与教学资源建设情况	83
5.1.4 师资队伍建设情况	85
5.2 产教融合的实践探索	87
5.2.1 校企合作培养的典型模式	87
5.2.2 竞赛与实战演练的育人成效	91
5.2.3 实习实训体系建设经验	93
5.2.4 当前问题与着力方向	96
5.3 在职人员能力提升机制	98
5.3.1 企业内训与在职培训现状	98
5.3.2 从业人员继续教育需求分析	98
5.3.3 当前面临的主要问题	100
5.3.4 在职人员能力提升路径	101
5.4 总结与建议	102

目录

C O N T E N T S

第六章 总结与建议	104
6.1 主要发现	105
6.1.1 人才实战能力内涵发生深刻变化	105
6.1.2 人才能力与实战需求存在结构性差距	106
6.1.3 人才培养体系面临课程、师资、平台三重挑战	107
6.1.4 深化产教融合是促进能力提升的关键路径	107
6.2 对策建议	108
6.2.1 国家层面：加强顶层设计，完善政策标准	108
6.2.2 教育层面：优化培养体系，强化实战导向	109
6.2.3 产业层面：深化产教融合，发挥主体作用	110
6.2.4 个人层面：提升AI素养，主动适应变革	110
6.3 AI时代网络安全人才发展趋势	111



背景分析

1.1 全球趋势:人工智能重塑网络安全人才能力

人工智能 (Artificial Intelligence, AI) 正在把网络安全人才问题从“数量缺口”推向“能力重构”。传统网络安全能力主要围绕系统加固、漏洞挖掘、入侵检测、攻防演练和事件响应展开。随着大模型、智能代理和自动化决策进入政务、金融、能源、医疗、工业互联网等关键场景后,网络安全从业者面对的不再只是系统攻防,也包括模型污染、输出可信、调用链可控、自动化响应诱导和智能系统可审计等新问题。本章从全球趋势、国内需求、能力内涵演进三个层面,分析AI时代网络安全人才实战能力的重构方向。

人工智能重塑网络安全人才能力的全球趋势可概括为三条主线:一是能力边界外扩,传统攻防向数据安全、模型安全、算法治理和智能应用安全延伸;二是培养方式实战化,课堂培训和证书评价让位于靶场演训、红蓝对抗和场景化任务;三是人才角色复合化,安全人员转向兼具人工智能素养、治理意识、协同能力和持续学习能力的实战型复合人才。

1.1.1 欧美AI安全战略下的人才布局

美国和欧盟近年来密集出台的AI安全战略,均在制度层面构建了系统性人才布局,走出了两条各具特色的路径:美国以国家力量直接驱动人才供给的规模激增,欧盟则依托立法与标准来引导体系化能力建设。

美国通过白宫主导、多部门协同的方式,构建起“战略引领—专项拉动—制度固化”三位一体的人才倍增机制。2023年发布《关于安全、可靠和可信地开发和人工智能的行政令》,首次从国家元首层面将AI安全人才培养提升至国家安全高度;在专项规划上,2025年修订的《国家网络安全战略实施计划》进一步将“AI安全人才密度”列为衡量关键基础设施韧性的核心指标,要求各联邦机构在年度预算中单列AI安全技能提升专项。这一制度设计将人才地位从配套保障升维成战略要件,直接推动了后续一系列联邦级人才计划的预算落地;专项计划上,国土安全部2024年成立的“AI兵团”至2026年初已招募

逾200名AI安全专家，部署在关键基础设施防护和深度伪造检测等实战一线。该计划的选拔不唯学历论，以能力测评与模拟任务结合的方式有效拓宽了人才入口。

与此同时，国家安全局下属的AI安全中心持续扩容，并与卡内基梅隆大学等顶尖院校共建AI安全科学硕士直通计划。该计划实行“学习、演训、实战”三阶段贯通培养。演训机制上，美国网络司令部自2024年起将AI安全模块系统性嵌入“网络旗帜”等年度联合演习。在2025年的演习中，场景首次加入了由自主AI代理发起的持续性攻击。制度保障层面，美国国防部2025年更新的《网络劳动力战略》明确提出，到2027年，全部网络作战岗位须完成AI基础与AI安全双模块认证，且实战演练考核在总评中的占比不得低于60%，形成“以战领训、以考促战”的良性闭环。

欧盟的AI安全人才布局与其强监管底色高度一致，通过立法确立需求、标准定义能力，系统构筑覆盖全成员国的安全人才基座。

在需求侧，2024年生效并分阶段适用的欧盟《人工智能法案》将“人员胜任力”列为部署与运维方的合规必要条件。合规直接催生了欧洲市场对AI安全审计师、模型风险管理师等岗位的旺盛需求，也推动了相关认证课程的开设；2025年，欧盟网络安全局发布AI安全技能框架，定义“AI安全分析师”“模型安全审计师”“AI威胁情报工程师”等八个新兴角色，并为每个角色绘制了包含知识、技能和实战任务的三维能力矩阵，为企业和高校在设计岗位说明书和课程时有可参照的标准。

在供给侧，欧盟委员会下设的AI办公室通过“数字欧洲计划”资助跨国AI安全硕士联合培养项目，联合柏林工业大学、代尔夫特理工大学、巴黎高科等28所高校，实行多校园轮转、企业命题实习和欧盟级靶场演训的混合培养模式，计划到2029年累计培养3000名AI安全方向高级专才；欧盟网络安全局主导的“Cyber Europe”系列演习自2024年起增设AI对抗赛道，并逐年升级难度。2026年演习设定AI生成虚假信息冲击多国选举系统的场景，240支队伍须在数小时内完成深度伪造溯源等实操任务；2025年启动的AI安全人才池试点，允许成员国专家在重大事件时跨境快速借调，整体抬高了欧盟AI安全韧性。如表1-1所示。

表 1-1 美国与欧盟 AI 安全人才布局对比

国家 / 地区	核心战略文件	主导机构	标志性举措	投入目标
美国	《国家网络安全战略实施计划》(2025修订)	白宫、美国国土安全部、美国国家安全局	AI兵团、NSA AI安全中心直通计划、国防部AI双模块认证	5年内培养5000名AI安全军职文职专才
欧盟	《人工智能法案》(2024)、《AI安全技能框架》(2025)	欧盟网络安全局、欧盟AI办公室	AI安全硕士联合培养、AI安全人才池跨境借调	数字欧洲计划5年培养3000名AI安全硕士

从宏观走势看，欧美AI安全人才需求已进入指数扩张期。据美国劳工统计局、欧盟网络安全局及行业智库综合预测，欧美AI安全岗位空缺数从2022年的约0.9万个，以年均37%的复合增长率增至2026年的6.8万个，预计2028年将突破12万个。

1.1.2 国际组织对人才能力的新界定

AI正在改变网络安全的风险结构，也在同步改写网络安全人才的能力结构。自动化钓鱼、深度伪造、提示注入、数据污染、模型滥用等新型风险不断出现，使传统网络安全人才能力框架面临明显调整压力。在这一背景下，国际组织开始从全球治理、数字技能、AI伦理和网络安全能力建设等角度，重新界定网络安全人才能力的五个转向：

一是从“会做防护”转向“理解AI风险”，要求网络安全人才掌握AI系统的基本运行逻辑，理解数据质量、模型训练、参数调整、接口调用和自动化决策可能带来的安全风险，并对模型、数据和算法流程中的潜在风险做出评估；二是从“单点处置”转向“全流程实战”，实战对象扩展到模型训练、数据流转、算法调用、智能应用部署和持续监测等环节，安全人才要面对的是如何在AI系统全生命周期中提前识别风险、验证风险并降低风险；三是从“纯技术能力”转向“技术与治理并重”，网络安全人才需要理解监管要求、风险分级、审计逻辑、数据保护、责任边界和合规流程；四是从“单兵作战”转向“跨域协同”，安全人才既要能把复杂技术风险转化为管理层、业务部门和监管主体能够理解的语言又要能把治理要求、合规要求和业务约束转化为研发团队可以执行的技术方案；五是从“一次性培养”转向“持续学习”，AI技术更新速度快，安全威胁演化快，要把“能否持续更新”纳入人才能力评价体系。

从具体组织的侧重点来看，欧盟网络安全局（European Union Agency for Cybersecurity, ENISA）的技能框架强调角色、任务、知识和技能的映射关系，为教育机构、培训机构和用人单位建立共同语言，使课程设计、岗位设置、职业发展和能力评价能够相互衔接；世界经济论坛（World Economic Forum, WEF）的《战略性网络安全人才框架》则更加强调网络安全人才与组织韧性之间的关系，把人才短缺、技能错配、教育培训、招聘保留和组织风险管理联系起来，提醒各方网络安全人才问题不是简单的招聘问题，而是组织安全能力建设问题；联合国教科文组织（United Nations Educational, Scientific and Cultural Organization, UNESCO）、国际电信联盟（International Telecommunication Union, ITU）和经济合作与发展组织（Organisation for Economic Co-operation and Development, OECD）则从伦理价值、数字技能和终身学习等角度拓展了人才能力的边界。这些框架从能力结构、组织韧性、公共价值和持续学习等方向，共同支撑了网络安全人才的复合型定义。

1.2 国内需求:国家战略与产业发展对网络安全人才的要求

“十五五”时期是我国统筹发展和安全、建设网络强国的关键阶段，国家战略层面对网络安全人才提出了明确且迫切的要求：与此同时，金融、能源、通信、智能制造等重点行业的数字化转型和升级，也使网络安全人才产生了巨大而具体的需求。

1.2.1 网络强国战略驱动人才能力升级

在网络强国战略纵深推进的背景下，我国网络安全人才队伍正在由传统基础运维人才，向实战对抗型高层次专家转型，人才供需结构性矛盾持续突出，行业面临明显人才缺口与能力断层。仅掌握防火墙配置、常规漏洞扫描等基础技能的入门级人才供给趋于饱和；而拥有国家级攻防实战经历、能够承担安全体系顶层架构设计与安全战略规划の顶尖实战人才供给十分有限，人才断层问题突出。从岗位需求结构来看，安全运营岗位长期保持最高招聘热度，企业招聘的核心偏好是拥有多年一线实战经历、可落地处置安全事件的从业者。

从国家战略落地实施的维度审视，相关人才亟须突破单一技术能力的局限，构建起一个涵盖“攻防实战能力+合规治理能力+战略研判能力”的复合型能力框架：既能在真实的红蓝对抗中完成高级漏洞挖掘与应急响应处置；也能落实《网络安全法》和《数据安全法》《关键信息基础设施安全保护条例》等合规要求；还能开展中长期安全风险趋势预判与战略规划。

1.2.2 “人工智能+”行动带来的人才能力需求

“人工智能+”战略行动深入推进，持续催生网络安全领域的复合型岗位需求。AI安全架构师、大模型风险评估师、对抗性机器学习防御工程师等新兴前沿岗位，已成为当前行业招聘市场中企业与求职者共同关注的核心热点。据北京市人力资源和社会保障局发布的《2025年北京市人力资源市场薪酬数据报告（二季度）》，AI大模型架构师的月薪中位值已超过4.2万元，头部企业的资深架构师年薪可达150万至200万元；翰德（Hudson）发布的《2025人才趋势报告》显示，AI人才的供需比仅为0.5，两个AI岗位仅能匹配到一位合适候选人。此类新型人才通常需同时深度掌握传统网络安全攻防技术体系与大语言模型底层运行原理，具备独立数据投毒监测防御、AI生成内容合规校验、模型隐蔽后门排查等能力。

应对新兴赛道技术融合与产业升级带来的巨大人才缺口，国内高等教育体系已开展

前瞻性布局，快速响应产业人才培养需求。当前国内多所普通高校布局网络安全相关专业方向，不少院校已启动 AI 安全相关课程建设，课程覆盖面较往年明显提升。这一趋势，体现出产业端与高等教育界共同加速 AI + 网络安全复合型人才培养的迫切诉求。

1.2.3 关键信息基础设施单位人才能力缺口

在能源、交通、金融、政务等关键信息基础设施领域，网络安全人才短缺尤为尖锐。这些核心行业的关键安全岗位长期面临“一将难求”的困境。一方面，在岗人员不仅需要精通通用网络安全攻防技术，还必须深刻理解掌握垂直领域的业务逻辑及行业特定风险。例如，工业互联网领域的网络安全工程师，在应对各种网络攻击的基础上，还必须非常熟悉 Modbus、OPC UA、DNP3 等主流工业控制通信协议的工作原理与安全脆弱性，具备独立设计、部署并持续优化针对工业控制系统 (ICS) 与数据采集与监控系统 (SCADA) 的纵深防御体系的能力。

另一方面，人才供给存在严重的结构性失衡：既懂行业核心业务又精通网络安全技术的顶尖复合型人才占比极低。人才队伍呈现底部过大，顶部过小的结构，“重产品，轻服务，重技术，轻管理”的现象普遍。多数运营单位安全防护团队人员多由传统 IT 系统运维人员临时兼任，或经短期培训后转岗而来。此类人员在应对国家级 APT（高级持续性威胁）攻击时，威胁感知、分析溯源、纵深防御与实战对抗能力存在显著缺陷。这一能力缺口与梯队断层，已成为我国网络安全防护能力建设中最为薄弱、风险最高且亟待系统性补齐的核心短板之一。

1.3 能力重构：AI时代网络安全人才实战能力演进

生成式人工智能、多智能体、星载大模型、行业专用安全大模型规模化落地，彻底颠覆持续二十余年的传统网络安全攻防范式。攻防主体从人类攻防双方的人力博弈，扩展为人类、通用大模型、攻击智能体、防御智能体多元混合体系，实战能力的底层逻辑、技能结构、考核标准、培养路径发生根本性重构。本节结合权威实测数据、政企真实攻防案例和国内外标杆培养实践，拆解 AI 时代网络安全人才实战能力的三层演进逻辑，为高校专业建设、企业人才招聘、行业技能评价提供标准化理论与实证支撑。

1.3.1 从“人对抗人”走向“人机协同”

在大模型技术商用落地前二十余年，攻防两端核心执行主体均为人类安全从业者，工具仅作为辅助。

攻击侧单次完整APT攻击需资深攻击者7—30天持续人工操作，单团伙同时发起攻击上限约10—20起；防御侧据IDC与Fortinet 2022年联合统计，传统中型企业SOC日均产生告警12万—15万条，3人专职分析师团队仅能完成不足15%高危告警研判。人才能力评价完全以人类单兵技术熟练度为核心指标，ISC2 2022年人才调研显示，全球92%网络安全岗位招聘未设置AI、机器学习相关能力门槛。

纯人力对抗模式不可化解的结构性短板：Hack The Box 2025纯人力CTF基准测试显示，顶级白帽团队完整完成中型内网渗透靶场全链路攻坚平均耗时11.7小时；中小安全工程师团队平均耗时23.4小时，低阶从业者存在大量漏洞遗漏、攻击链断裂问题，依靠人工完成大量安全任务的模式逐渐面临效率瓶颈。黑客组织依托AI工具链后，单团伙可并行发起上百起勒索、钓鱼、漏洞武器化攻击，攻击周期从“天级”压缩至“小时级”，单纯依靠增加安全人员数量已难以完全满足实际需求。

AI时代，人机协同成为攻防唯一可行的主流范式：安全人才作为决策中枢、校验主体、战略设计者，AI智能体、垂直安全大模型承担海量重复性、标准化、高算力消耗基础操作，形成“人类高阶判断+AI自动化执行”分工明确的新型作战体系。政务、金融、能源等行业实测数据显示，人机之间已经开始形成较为明显的能力分工：如表1-2所示。

Hack The Box 2026年覆盖全球120支专业团队的AI增强CTF对比实验进一步验证了人机协同效能：仅依靠人类顶尖团队：靶场解题平均完成率16%，单日可处置安全任务量基准值100；人机协同混合团队解题完成率提升至27%，单日处置任务量提升3.1倍，漏洞

表 1-2 人机协同分工统计

工作类型	AI 智能体	主导机构
资产扫描、弱口令爆破、基础漏洞探测	99%	审核扫描结果、排除AI幻觉误报、制定资产分级防护策略
安全日志标准化提取、基础告警过滤、威胁情报结构化	95%	关联多维数据还原复杂APT攻击链、判断潜在未知威胁
常规钓鱼样本生成、简单EXP代码编写、基础防护脚本	90%	校验AI生成恶意代码隐蔽性、模拟高阶复合攻击、优化红队策略
模型基础安全检测、训练数据投毒初筛、提示注入浅层拦截	85%	开展AI红队深度对抗、挖掘模型底层逻辑漏洞、制定AI治理制度
复杂多域APT研判、重大安全事件应急决策、合规体系搭建、AI风险顶层设计	0%	人类专属高阶决策，AI仅提供辅助参考材料

覆盖完整度提升70%；完全自主AI无人干预时，高危漏洞漏判率82.4%。数据表明，AI无法替代人类核心决策，但脱离AI辅助的纯人力模式完全无法适配当前规模化、自动化AI驱动网络威胁，人机协同是行业必然、唯一选择。

由此，人机协同统筹能力成为入门必备素养，具体拆解为三项硬性实战要求：一是具备垂直安全大模型、安全智能体的精细化提示工程能力，智联招聘2025岗位调研显示，48.8%企业招聘安全岗明确要求掌握安全大模型提示词工程，成为仅次于渗透测试的高频技能要求；二是具备识别、过滤AI幻觉风险的实战校验能力。Cobalt调研证实78%自动化AI扫描工具存在关键漏洞假阴性，安全人才必须掌握交叉核验、人工复现、多模型对比校验标准化流程，避免由于依赖AI结论导致重大安全事故；三是具备划分人机分工、搭建标准化协同作战流程的工程化能力，可针对SOC运营、红队攻防、AI安全测评、数据合规四大场景设计人机协同工作流。

1.3.2 AI成为网络安全人才基础能力

当前AI相关能力已从网络安全人才的能力加分项，转变为岗位录用的基础门槛。从产业需求来看，各重点行业安全岗位普遍要求从业人员掌握大模型基础应用与AI安全风险识别能力；人才供给端，院校虽已加快AI安全相关课程建设，但兼具人工智能与网络安全实战复合能力的毕业生供给仍然紧缺；在职安全从业者层面，对抗性AI攻防、大模型安全也已成为技能提升的核心方向。由此可见，AI能力正在重塑网络安全人才的能力框架。同时需要明确，AI作为网安人员必备基础能力，并不代表所有安全从业者都要成长为AI算法专家。

结合MITRE ATLAS、NIST AI RMF、OWASP LLM Top10三大国际标准，以及国内《生成式人工智能服务管理暂行办法》等配套行业技能规范，可将网络安全人才必须掌握的通用AI基础能力划分为四大模块，均属入门级基础实战能力。

一是通用安全大模型实操与调度能力，包括针对漏洞挖掘、日志分析、威胁情报、合规文档、攻防报告五类场景编写精准提示词，熟练使用AI生成渗透脚本、防火墙策略、应急处置预案等保测评整改方案，掌握AI代码审计、流量样本自动解析标准化操作，以及结合通用对话大模型、代码专用模型、威胁情报模型交叉验证，单一模型结论不可直接采信，形成“多模型交叉+人工复核”固定操作规范。

二是AI驱动攻击识别基础研判能力，覆盖AI深度伪造钓鱼、AI批量漏洞武器化、LLM接口浅层利用三类高频场景。Fortinet 2026威胁数据显示，87%企业年度网络攻击已包含AI生成恶意载荷。

三是AI辅助安全工程落地基础能力，包括使用AI生成SOAR编排脚本、依托AI自动生成多层次攻防演练场景、自动生成对抗样本、针对合规材料编制自查报告、整改方案。

四是AI基础风险自查与简易防护配置能力，针对企业内部落地通用大模型、本地私有化LLM、第三方AI插件场景，识别明文密钥硬编码、无权限管控模型接口、训练数据未脱敏等浅层风险，配置基础输入过滤、访问鉴权、输出水印轻量化防护策略。

不同岗位的AI基础能力要求呈现梯度差异：基础运维等保测评、SOC值守岗掌握前两部分基础内容，可使用AI辅助日常工单处理、基础告警筛查、常规合规文档编制，能够识别浅层AI钓鱼、简单提示注入攻击；渗透测试、应急响应、红队基础岗完整掌握前三部分内容，能够利用AI批量生成攻击载荷、自动化还原攻击链、依托AI辅助开展内网渗透，可独立完成基础LLM接口漏洞挖掘、AI攻击样本复现；安全架构、AI专项安全、高级红队岗需要掌握四大基础模块，并在此之上延伸“利用AI、保障AI、对抗AI”三大高阶核心能力。

1.3.3 “利用AI、保障AI、对抗AI”成为要求

当人机协同成为基础作战模式、AI应用能力成为全员通用门槛后，行业对中高端安全专家、AI安全架构师、AI红队、模型安全运维人才提出三位一体高阶实战能力要求：善用AI赋能安全工作（利用AI）、筑牢AI系统自身安全防线（保障AI）、主动抵御AI驱动新型网络攻击（对抗AI）。三大能力相互支撑、闭环联动，构成新一代AI安全人才差异化核心壁垒，也是头部政企、航天、能源、AI大厂高薪岗位核心考核标准。

“利用AI”强调以AI为倍增器放大攻防实战效能：安全人才基于对大模型、多智能体、机器学习底层逻辑深度理解，自主设计、定制、调优AI安全工具链，最大化释放AI自动化、高算力、批量推演优势，解决传统人力无法覆盖的规模化、海量、重复性安全工作。对于中高端网络安全人才而言，能力要求也不仅限于简单调用现有AI工具，而是需要能够结合具体安全场景，对AI工具进行定制、调优和工程化应用，包括基于行业私有威胁情报、攻防样本、业务日志微调开源代码模型（CodeLlama、GLM-5-Coder）、安全专用大模型，适配能源工控、卫星星载、金融业务专属安全场景；基于LangChain、自研调度框架搭建多智能体协同作战系统，实现“扫描Agent→研判Agent→处置Agent→复盘Agent”全自动闭环，自主定义Agent权限、交互规则、异常熔断机制；对标AISIS Mythos、OpenAI CyberGym架构，构建本地化专属AI攻防仿真靶场，自动生成链式多阶段攻击样本，支撑常态化红队演练与模型安全验证；搭建私有威胁情报向量数据库，依托大模型实现跨语种、多格式情报自动提取、关联、预测攻击趋势。

“保障AI”聚焦AI本体安全防护，保护大模型、训练数据集、微调流程、推理服务、

自主智能体自身不被攻击，覆盖模型训练、存储、部署、推理、下线完整生命周期，是当前政企、AI厂商刚性合规刚需。训练层面，需实现训练数据隐私脱敏、差分隐私落地、投毒样本检测、开源数据版权与隐私合规校验；GLM-5.2训练数据全流程脱敏、隐私过滤处理流程为本能力标准化工业范本，从源头消除训练集泄露、投毒风险，以及训练环境隔离防护；模型层面，需开展模型水印嵌入、权重加密、对抗样本鲁棒性增强、轻量化星载/边缘模型安全裁剪，自主开展提示注入、模型越狱、权重窃取、逆向还原、越权推理全维度对抗测试，对标AISI Mythos红队智能体完成极限安全验证；推理部署阶段需落实LLM接口安全防护，实现输入输出双层过滤、会话隔离、API鉴权、访问频率管控、上下文权限隔离，给自主执行AI智能体设置操作白名单、动态熔断机制、操作日志全留存；在合规治理层面，搭建一体化大模型攻防测评底座，出具具备招标投标采信效力的安全评估报告，覆盖《生成式人工智能服务管理暂行办法》等合规控制点，在模型上线前进行测评、运行中持续监测、下线权重销毁完整闭环治理，定期开展AI安全复盘与漏洞迭代加固。

“对抗AI”依托大模型、深度伪造、自主攻击智能体打造的工业化攻击链路，构建专项检测、溯源、阻断、反制防御体系，形成“AI对抗AI、人类统筹决策”的新型攻防平衡。实战要点包括：识别深度伪造音频、视频、文本钓鱼样本，部署多模态水印溯源与文本特征检测模型；区分AI批量生成的加密勒索脚本、自适应EXP和无文件攻击载荷，建立专项检测规则库；通过资产异常行为基线建模识别AI智能体批量扫描与梯度式渗透特征，以动态微隔离和访问熔断阻断自主横向移动，并依托日志与流量还原攻击Agent来源。Fortinet《2026网络威胁报告》显示，2025年全球87%的网络勒索团伙已使用AI工具生成定制化恶意代码，攻击准备周期从7天压缩至90分钟；APT国家级组织已将大模型嵌入全套攻击链条，传统基于人工特征的防御规则检出率不足40%。

三位一体能力并非彼此独立，而是形成完整实战闭环：利用AI为保障AI、对抗AI提供工具底座；保障AI守住自身防御体系根基，企业用于安全运营的AI模型若存在原生漏洞，攻击者可直接劫持防御工具反向入侵内网；对抗AI抵御外部AI攻击威胁，外部黑客AI工业化攻击持续迭代，倒逼安全人才持续利用AI优化检测模型、完善靶场推演，反向推动“利用AI”技术持续迭代升级。

从利用AI（工具供给）到保障AI（底座可信），再到对抗AI（外部防御）以及实战攻击样本回流优化AI工具（反向迭代利用AI能力），三者循环迭代，构成AI时代高端网络安全人才独一无二的核心实战壁垒，也是区分传统网安人员与新一代AI安全人才的核心评判标准。

1.4 AI时代网络安全人才核心能力与岗位族划分

AI安全属于网络安全领域新兴方向，大模型的普及重构了网络安全攻击面与工作范式。安全人员既要应对AI模型自身安全风险，又要保障AI业务落地过程安全可控，同时还需要掌握利用AI工具开展安全工作的能力。国内文献分析、产业调研与行业实践，AI时代网络安全人才的核心能力可划分为模型安全能力、AI应用安全能力、AI赋能网络安全能力三类。本报告将AI安全相关岗位归纳为三大岗位族，如图1-1所示。

不同岗位族的工作范畴也有所不同，如表1-3所示。

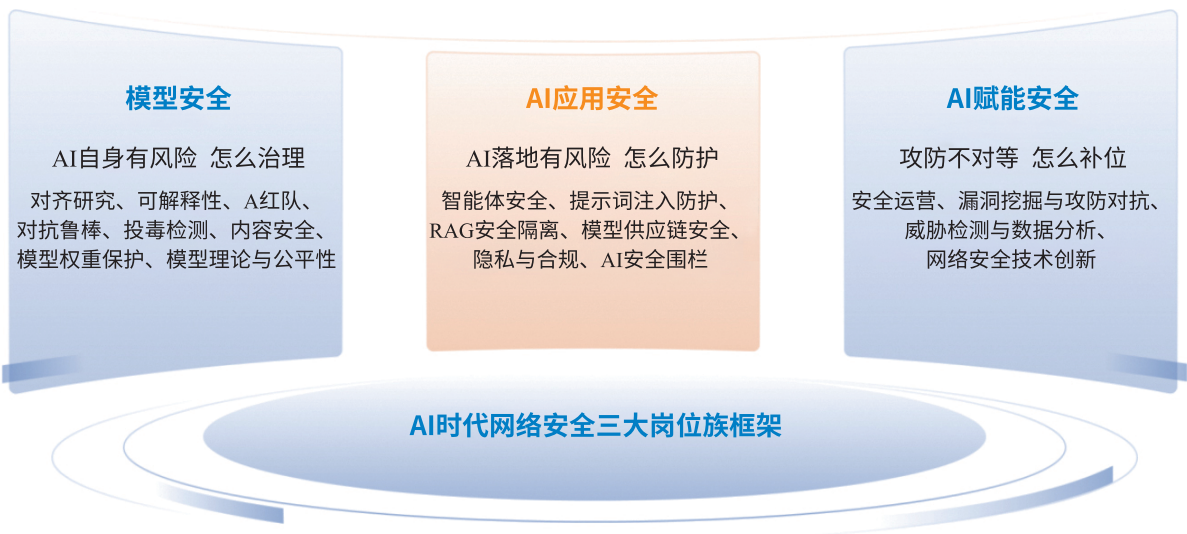


图1-1 AI安全三大岗位族框架

表 1-3 AI 安全三大岗位族工作范畴

岗位族	岗位定位	工作范畴	典型岗位
模型安全	AI模型风险治理	对齐研究、可解释性、AI红队、对抗鲁棒、投毒检测、内容安全、模型权重保护、模型伦理与公平性	对齐研究员、AI红队研究员、可解释性研究员、信任与安全专员、模型政策研究员
AI应用安全	AI落地风险防护	智能体安全、提示词注入防护、RAG安全隔离、模型供应链安全、隐私与合规、AI安全围栏	AI安全工程师、智能体安全工程师、LLM应用安全工程师、数据隐私工程师、AI合规专员
AI赋能安全	利用AI开展安全防御、攻防不对等补位	安全运营、漏洞挖掘与攻防对抗、威胁检测与数据分析、网络安全技术创新	安全运营工程师、渗透测试工程师、漏洞挖掘工程师、威胁情报分析工程师、安全研究员、安全运维工程师

1. 模型安全岗位族：

工作范畴包括对齐研究、可解释性、AI红队、对抗鲁棒、投毒检测、内容安全、模型权重保护、模型伦理与公平性等场景。

2. AI应用安全岗位族：

工作范畴包括智能体安全、提示词注入防护、RAG安全隔离、模型供应链安全、隐私与合规、AI安全围栏等场景。

3. AI赋能安全岗位族：

工作范畴包括安全运营、漏洞挖掘与攻防对抗、威胁检测与数据分析、网络安全技术创新等场景。

1.5 总结与建议

AI技术全面商用推动网络安全人才实战能力完成三层不可逆演进：第一层，攻防主体从一对一纯人力对抗转向人机协同混合作战，人机统筹校验能力成为入门底线；第二层，AI综合操作、基础AI威胁识别从行业加分项转化为所有安全从业者通用基础必备能力；第三层，面向中高端AI安全专家，形成“利用AI赋能安全、保障AI自身安全、对抗AI驱动攻击”三位一体核心高阶能力框架。

从产业现实看，全球网络安全人才巨大缺口本质是AI复合型实战人才结构性缺口，传统纯攻防人才供给过剩，同时具备人机协同思维、掌握三位一体AI安全完整能力专家严重稀缺。本章节基于工信部、MITRE、头部政企真实攻防项目可复核数据与实战案例，厘清AI时代网安人才能力迭代逻辑，为高校网络空间安全专业AI安全微专业、大模型攻防靶场、AI安全教学平台建设，企业人才招聘分层考核、行业职业技能评价提供标准化理论与实证依据，同时指明未来人才培养核心重心：跳出传统纯人力攻防教学框架，以人机协同为基础、三位一体AI安全高阶能力为核心，构建适配空天地泛在智联、生成式AI规模化落地新型实战人才培养体系。

在AI安全能力框架上，本报告将网络安全人才的核心能力归纳为模型安全能力、AI应用安全能力和AI赋能网络安全能力三大岗位族。模型安全岗位族围绕大模型本体，承担对齐研究、可解释性、AI红队、对抗鲁棒、投毒检测、内容安全、模型权重保护、模型伦理与公平性等职责；AI应用安全岗位族面向大模型及其智能体在业务系统的落地过程，承担智能体安全、提示词注入防护、RAG安全隔离、模型供应链安全、隐私合规、AI安全

围栏等职责；AI赋能安全岗位族面向传统网络安全工作，承担安全运营、漏洞挖掘与攻防对抗、威胁检测与数据分析、网络安全技术创新等职责。三大岗位族共同构成AI时代网络安全人才能力体系的整体框架，也是后续章节评估模型安全、AI应用安全和AI赋能安全三类实战能力的分析基础。

网络安全人才模型安全能力分析

2.1 模型安全风险分析

AI模型安全风险来源复杂、表现形式多样，贯穿模型设计、训练、部署、应用和治理的全流程。从风险来源、作用环节、攻击方式和影响路径等维度出发，可将AI模型安全风险划分为架构风险、数据风险、对抗风险、系统风险、环境风险和伦理风险六类。需要说明的是，上述六类风险并非完全割裂，在实际应用中可能相互交织，该分类方式主要用于突出不同风险的主导特征和主要影响路径。

2.1.1 架构风险

架构风险指AI模型在系统部署与运行过程中，由参数存储、推理执行、权限控制、系统提示词管理和工具调用等环节引发的安全风险。相较于传统软件，AI模型的核心构成除了代码外，还包括模型参数、权重配置和工具调用机制等资产。这些要素一旦存在设计缺陷、配置不当或访问控制失效，极易引发模型能力被滥用、敏感信息泄露以及推理过程被操控等安全问题。

对于基础AI模型研发企业而言，模型结构、训练方法、参数权重和推理策略具有较高商业价值。若模型权重被非法获取，攻击者可在本地部署克隆模型，从而绕过服务器端的访问控制、计费机制和安全审查机制。对于行业用户而言，私有化AI模型可形成具有行业知识的专用模型，但倘若模型参数和权重配置泄露，则会暴露企业内部知识和业务流程等敏感信息。

此外，AI模型通常会通过系统提示词限定模型角色、输出风格和权限边界。系统提示词在AI模型中发挥行为约束的作用，一旦被攻击者通过多轮诱导或上下文套取等方式获取，极易突破原有安全边界。同时，AI模型逐渐嵌入到数据库、浏览器、邮件系统、代码生成器等工具，成为具备高度自动化操作能力的主体。如果AI模型权限配置不当，攻击者就可以诱导模型执行越权查询、敏感数据发送等非授权操作。

2.1.2 数据风险

数据是AI模型的基石，模型训练、微调和版本迭代都高度依赖数据。与传统信息系统相比，AI模型的数据风险贯穿数据采集、数据清洗、模型训练、模型微调、模型评测、部署推理和版本更新等多个环节。如图2-1所示。

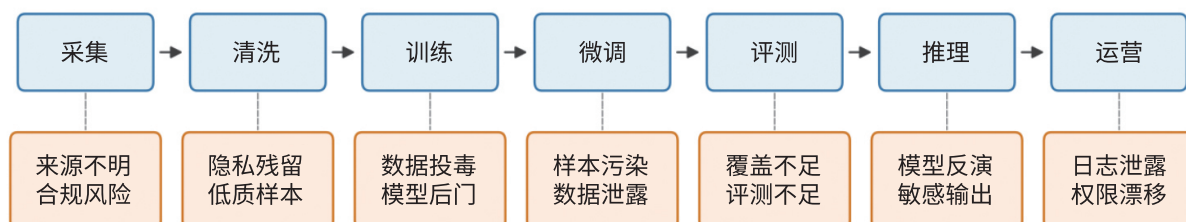


图2-1 AI模型的数据风险链条

如图2-1所示，在数据采集和清洗环节，数据来源不明、敏感信息残留以及低质量样本混入等问题会导致后续模型的质量受显著影响。在训练阶段，数据投毒、样本污染和后门植入可导致模型的行为发生偏移。在推理阶段，还可能被诱导出一些内部成员的敏感信息。进入运营阶段后，用户输入、模型输出、调用日志会不断累积，权限配置也会动态调整，若缺乏有效的数据脱敏、日志审计、访问控制和留存管理机制，还可能引发日志泄露、权限漂移、越权访问等风险。

2.1.3 对抗风险

对抗风险主要指攻击者利用精心构造的提示词或交互诱导，使AI模型产生误判、越权或违背人类预期的行为。此类风险主要表现为提示词注入攻击和越狱攻击，并可通过多轮诱导、上下文污染、任务分解、语义伪装和跨模态隐藏指令等方式增强攻击的隐蔽性和持续性。攻击者的目标不再局限于让模型判断错误，还试图绕过安全策略使模型行为失控。

提示词注入攻击是最具代表性的对抗攻击之一，分为直接提示词注入攻击和间接提示词注入攻击两类。在直接提示词注入攻击中，由于AI模型通常会将用户当前输入和上下文一起作为生成条件，因此攻击者可以尝试插入一些恶意指令来破坏模型对于原任务的执行。在间接提示词注入攻击中，当模型处理外部内容时，如果缺乏内容来源识别、指令隔离和边界判断，就可能将外部内容中夹带的恶意内容误识别为待执行的指令，从而引发越权输出、敏感信息泄露、非授权操作等问题。

越狱攻击主要是针对AI模型的安全对齐机制发动攻击。攻击者通过角色扮演、假设场景、编码转换、多语言混淆等方式，诱导AI模型弱化或绕过原有的安全策略，生成原本

会被拒绝的有害内容、违法指南或敏感信息。与直接要求模型输出违规内容的单轮攻击相比，越狱攻击还可能通过多轮诱导和上下文污染逐步实现。攻击者通过在多轮对话中逐步构建语境、分解敏感任务或弱化安全约束，使得模型在长上下文中逐渐偏离原安全边界。在具备长期记忆或外部知识库写入能力的系统中，相关攻击信息还可能被写入记忆模块或知识库，进而影响后继任务的安全执行。

2.1.4 系统风险

系统风险是指AI模型系统在开发、部署和运行过程中，由底层框架、运行环境、插件工具和软件依赖等带来的安全风险。AI模型并不是孤立运行的算法，而是依赖深度学习框架、推理引擎、容器平台、模型仓库、开源依赖和云服务基础设施共同支撑的复杂系统。因此，AI模型系统风险主要体现在运行环境、接口暴露、基础组件、软件依赖和供应链等方面。

深度学习框架和推理引擎是AI模型开发与部署过程中的基础组件。在实际应用中，相关框架和工具链可能存在反序列化漏洞、内存安全漏洞、软件依赖漏洞以及远程代码执行漏洞等问题。例如，加载不可信来源的文件可能触发恶意代码执行，推理服务若暴露未经授权访问的接口，则可能导致模型探测、批量调用、数据泄露和推理资源被恶意消耗等风险。

此外，AI模型系统的供应链风险日益凸显。目前，AI模型系统的开发广泛依赖开源模型、开源数据集、开源框架、第三方插件和模型托管平台。开发者可以直接从互联网下载预训练模型并进行微调，以快速获得面向特定任务或行业场景的模型能力。

这种开放性降低了AI模型系统的开发门槛，但也带来模型来源不明、权重配置被篡改、软件依赖被投毒、插件携带恶意逻辑等风险。

2.1.5 环境风险

环境风险是指AI模型系统进入真实业务场景后，由组织流程、用户行为、业务数据、权限配置和自动化执行链路等运行环境因素引发的安全风险。不同于实验环境，生产环境下AI模型往往接触到真实的业务数据、接口及自动化工作流，并能将输出结果用于权限判断、工具调用和业务系统的执行。

如图2-2所示，在AI模型的应用过程中，业务安全风险往往是从用户的自然语言请求出发，经历模型理解、权限判断、调用相应的工具或接口、业务系统执行等过程后，最后进入日志和审计阶段。以上任一环节的设计缺陷，都有可能造成风险的扩大化，比如用户请求表达不清导致模型误判了任务的目标，没有进行权限判断会使得低权限的用户可以

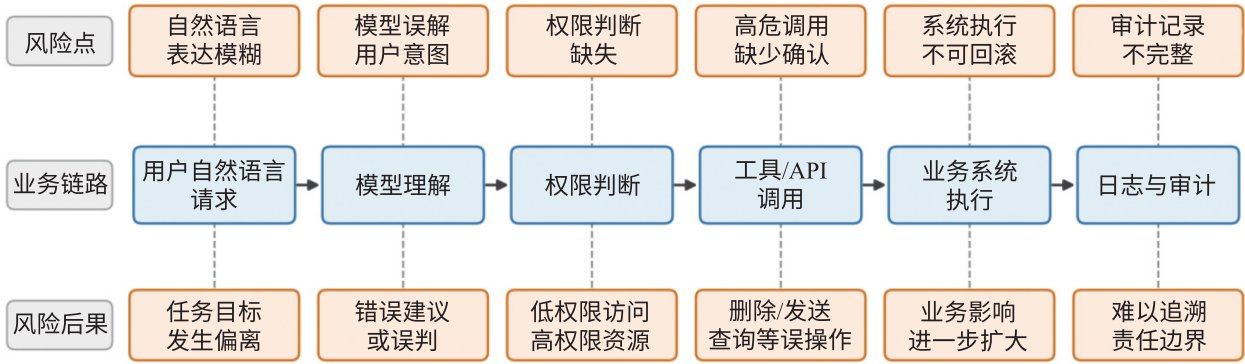


图2-2 AI模型应用中的业务安全风险传导路径

间接地访问高权限的资源，对一些工具或接口的使用缺乏二次确认可能会造成模型执行删除、发送、查询等高风险的操作。

2.1.6 伦理风险

AI模型伦理风险是指模型在训练、部署和应用过程中，因数据偏差、价值对齐不足、输出不可解释、责任边界不清或使用方式不当，对个人权益、组织决策、社会公平和公共信任造成负面影响的风险。随着AI模型广泛应用于教育、医疗、金融、政务等领域，其生成内容正日益影响个人决策、组织管理和社会公共信任。在这些高频应用场景中，一旦模型输出内容存在偏差、失真或责任边界不清等问题，就可能造成不公平决策、错误信息扩散、责任追溯困难，从而引发伦理层面的社会争议。

偏见是AI模型面临的主要伦理风险之一。AI模型在生成内容或辅助决策时，因训练数据的结构性失衡、标注规则偏差或人为设定，极易偏向或排斥某一类群体、文化或观点，从而固化刻板印象，甚至产生歧视性输出。例如，一些文本生成模型在不同语言和文化背景下，将性别、职业、地域或族群相关的刻板印象当作一般性知识进行输出；部分图像生成模型也可能在涉及特定地区、职业或群体的提示下，生成具有标签化特征的图像。

责任边界不明也属于AI模型伦理风险范畴。当AI模型发生误判、泄露、歧视、违规输出或者越权时，涉及到的责任主体包括基础模型提供商、二次开发方、数据提供方、模型部署方、终端使用人员等。如果缺乏完善的AI模型使用规范、审计机制和责任划分制度，就难以界定和追溯各方责任。因此，在高风险应用场景中，有必要明确模型输出的使用范围、日志留存要求和异常处置流程。对于涉及个人权益、组织决策和公共服务的模型输出，应明确其辅助性定位，避免将其作为唯一决策依据。对于敏感数据处理和外部内容发布等关键环节，应设置人工复核机制，以便在发生异常时及时定位问题、追溯责任并完成

风险处置。

2.2 网络安全人才模型安全能力现状分析

2.2.1 大模型企业安全建设现状

从国际市场看，Anthropic、OpenAI和Google DeepMind三家公司代表了当前大模型安全治理的三种典型路径。

美国Anthropic公司的特点是将安全治理前置到模型研发环节。2022年12月15日，Anthropic发布学术论文《宪法式人工智能：来自AI反馈的无害性》，系统提出宪法人工智能方法；2023年9月19日，正式发布责任扩展政策（RSP），提出人工智能安全等级（ASL）分级框架；2024年10月15日发布RSP更新版；到2026年4月29日，RSP已更新至3.2版。其路径特点是把对齐训练、风险分级和内部治理机制持续嵌入前沿模型研发流程。

作为AI大模型的开创者之一的OpenAI公司则更强调把安全评估和上线治理做成工程体系。2023年10月26日，OpenAI发布《OpenAI应对前沿风险的方法》，并宣布组建Preparedness团队；2023年12月18日发布Preparedness Framework（Beta）；2025年4月15日发布更新版Preparedness Framework。该框架围绕网络安全、化学与生物风险、模型自主性等高风险能力建立评估与缓释机制，体现出较强的工程化治理特征。

Google DeepMind公司更侧重基础研究与前沿风险框架建设。2024年5月17日，Google DeepMind发布前沿安全框架；2025年2月4日发布更新版，并明确表示已将其用于Gemini 2.0等前沿模型的安全与治理流程。其代表性在于将安全研究、内部治理与前沿模型评估体系化结合。

从国内市场看，主流厂商更强调监管适配、业务落地和与既有安全体系融合。大模型厂商强调从源头解决模型自身的安全问题，同时也会在MaaS服务平台层面配套安全防护能力。

从国内市场看，大模型厂商强调从源头解决模型自身的安全问题，同时也会在MaaS服务平台层面配套安全防护能力。

智谱为GLM-5.3构建了三层风险审查系统：外层滥用分类器用轻量模型标记并拦截大规模滥用请求；推理监控器在模型推理过程中实时审查任务意图，判断是否存在潜在危害；深度安全对齐，让模型从根本上自主识别并拒绝攻击性请求。同时在全面推出前，智谱对进行了尤为密集的安全性评估与广泛的红队测试，保障大模型在风险防护与能力释放之

间取得良好平衡，相比于前代模型显著提高了未授权恶意滥用的门槛。

阿里云千问大模型创新式构建三层防护体系，覆盖模型研发训练、部署应用、运行上线全流程，通过内置算法安全设计、模型安全训练、内容安全引擎和推理在线防护，确保大模型技术可控、内容合规、服务可靠。

腾讯基于安全专家能力，深度剖析大模型原理，建立了从训练到推理的全链路防护体系。该体系通过接口输入、内容预处理、模型识别、策略辅助、平台调度分析、人工标注运营共6个维度为模型提供支撑。基于腾讯混元大模型的实际运营与测试基准，安全体系在漏洞检测覆盖、资产识别能力、攻击防御能力、合规与标准等指标上表现明确。

华为盘古大模型的安全性主要从四个方面考虑，在数据安全和隐私保护方面，覆盖数据全生命周期管理，运用加密、审计、脱敏、隐私计算等手段保护敏感数据，防范隐私泄露；在内容安全方面，通过价值观对齐训练与内容审核，过滤各类有害信息；在模型安全方面，借助模型动态混淆，防护模型权重与结构泄露；在系统安全方面，依托网络隔离、身份鉴权、Web安全能力，抵御外部攻击。

火山引擎豆包大模型，基于飞书企业级协作平台的安全合规体系，豆包工作提供严格的全链路Agent防护与治理能力；同时旗下火山方舟上线基于安全沙箱的大模型安全互信计算方案，覆盖链路加密、数据高保密、环境隔离和操作审计，目标是把模型和数据资产的风险控制在源头。

网络安全厂商则更强调在企业部署大模型时提供外挂式安全保护。网络安全厂商侧重外挂式保护，即在不改造大模型内部架构的前提下，在其外围部署独立的安全能力层。核心措施包括：过滤违规敏感内容、防御提示词注入和越狱攻击、防止数据泄露、提供合规评估与加固，并延伸至智能体行为管控和幻觉抑制。此类方案以围栏或防火墙形态旁路部署，覆盖输入输出、访问及运行全链路，技术上不侵入模型原生架构，可独立部署和快速接入，是当前国内大模型安全市场中的主要方案之一。

2.2.2 大模型企业人才配置情况

当今，国内外主流大模型企业的安全人才配置也呈现差异化特征。国外呈现重模型安全、辅以应用安全、AI赋能安全多通过云业务或合作伙伴释放的整体特征。其内部组织通常将Safety（对齐与内容）和Security（系统与权重）作为两条相对独立的线，分别由首席科学家或独立的首席安全官与首席信息安全官双轨负责。如表2-1所示。

1.OpenAI已设立Security Preparedness Lead, Coding Agents 岗位，标志着模型权重与内部AI代理已被作为国家安全级资产单独编制。

表 2-1 国外主流大模型企业人才配置

企业	岗位重心	标志性团队 / 产品	代表性公开岗位
OpenAI	模型安全为主，应用与赋能补充	Safety Systems、Preparedness	MTS Safety、Researcher Automated Red Teaming
Anthropic	模型安全旗舰	Alignment Science	Frontier Red Teamer
Google / DeepMind	三岗位齐全	Frontier Safety、Google Cloud AI Threat Defense	Frontier Safety Researcher、Threat Analyst
Microsoft	应用与赋能并重	AI Red Team、Security Copilot	AI Red Teamer、Security Researcher
Meta	开源工具/分类器	Purple Llama Llama Guard	Responsible AI Engineer

2.Anthropic 形成了Alignment Science、Frontier Red Team、Safeguards Research、Threat Intelligence相对完整的研究闭环。其中，Alignment Science团队的核心任务是训练、评估并监控高能力模型，于2026年初推出Claude Code Security面向企业客户开放。

3.Google DeepMind是少数在三象限均有大编制的企业；Microsoft的人才重心则偏向AI应用安全与AI赋能安全两端，AI Red Team出品的开源框架PyRIT已成为业界主流的红队工具，2026年5月再度开源RAMPART和Clarity，团队构成由红队研究扩展为Agent安全工程。

4.Meta团队规模相对精简，主力投入在工程化的安全分类器与评测工具上，公开成果集中体现在PurpleLlama项目，包含Llama Guard、Prompt Guard、CyberSecEval等开源组件。

与国外不同，国内大模型企业的人才配置存在一项明显的合规约束。需要说明的是，相关企业均未公开披露安全人才编制数据，本节内容基于官网、开源项目和行业报告等可见信息整理：

1.阿里巴巴旗下安全团队，在AI安全方向持续投入深入研究，研究方向包括AI安全评测、外层安全防护、模式内生安全加固、红队及攻防体系、合规化内容溯源与标识等，兼顾模型本体治理与业务应用防护需求。

2.百度旗下百度安全团队，针对AI模型安全、AI数据安全、AI能力滥用、AI硬件系统安全和AI软件实现安全开展研究，支撑基座模型安全能力建设。

3.字节跳动成立AI与数据安全一级部门，为Seed基座模型及AI原生应用提供跨模态数据服务，覆盖数据生产全流程，包含模型评估标准的制定、数据规模化生产、数据飞轮

搭建，不断提升数据质量，支持模型快速迭代。旗下火山引擎通过大模型安全测评平台、智能体安全防护平台等，对外释放安全服务。

4.腾讯旗下朱雀实验室，专注AI安全领域的实战攻防与前沿技术研究，研究方向涵盖大模型安全、AI智能体安全、AI赋能安全与AI生成检测等领域。玄武、科恩等安全实验室矩阵也在传统网络安全业务的基础上，拓展大模型安全、智能体安全、AI攻防等研究方向。

本节重点呈现我国模型安全保障人才的整体现状。调研覆盖两类主体：一类是承担基础模型研发与服务的AI大模型企业，包含云厂商旗下的大模型业务单元；另一类是以安全产品、攻防服务为主业，并在大模型方向有显著投入的网络安全企业。基于以上样本，本节系统刻画了模型安全保障人才基本情况、人才实战能力水平、人才能力提升情况等。

从模型安全保障人才从业年限分布可以看出，人才队伍呈现中短从业年限人员占主体、资深人员占比偏低的特征。从业35年人员占比33%，从业3年以下人员占比32%，两者合计达到65%，说明当前模型安全保障人才多数是从从业年限较短的从业人员，大量人员由网络安全、AI相关领域转入，队伍整体偏年轻化。从业5—10年的人员占27%，构成队伍的中坚力量；而10年以上资深人员占比仅为8%，具备长期安全工程积累的资深人才供给相对稀缺。整体来看，模型安全属于新兴交叉领域，行业缺少大批量原生资深人才，人才更多依靠现有安全与AI从业者转型成长。如图2-3所示。

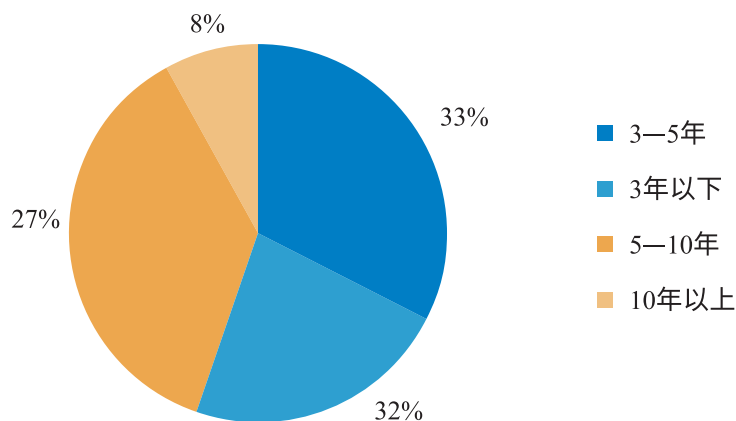


图2-3 模型安全保障人才从业年限分布

从模型安全保障人才的岗位设置来看，国内模型安全专职化组织建设尚不充分，兼职代管仍是主流，部分企业存在治理责任主体缺位的问题。仅23%的单位设立模型安全专职团队或实验室；37%的单位依靠现有安全团队兼任，27%的单位依靠研发团队自行承担相关工作；另有13%企业暂无明确岗位承接模型安全工作。如图2-4所示。

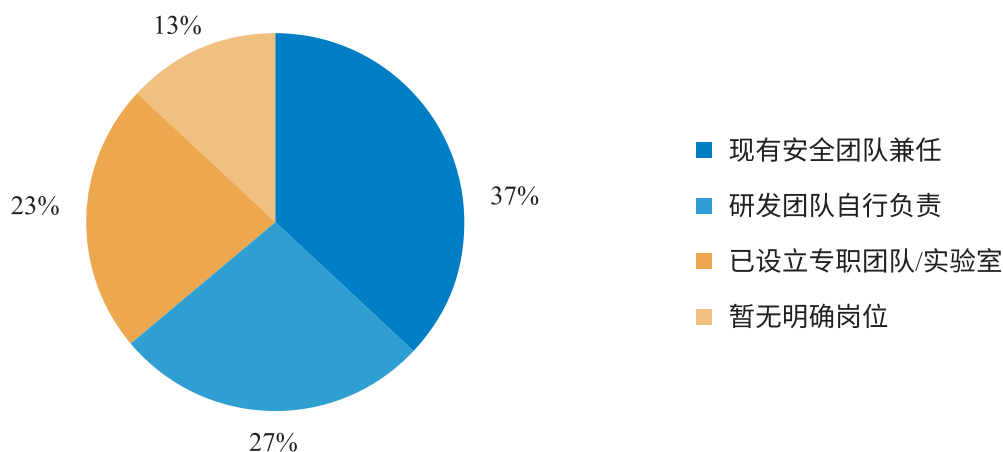


图2-4 模型安全人才岗位设置分布

从模型安全保障人才来源结构来看，算法或人工智能背景、传统网络安全背景转型是人才两大主要输入渠道，占比均超过50%，说明当前模型安全人才主要依靠AI技术人员与传统安全人员跨界转型而来；校园招聘占比约32%，是人才补充的重要途径；依托外部厂商或专家支持的占比仅约10%。整体可以看出，模型安全作为交叉新兴领域，原生成熟人才供给不足，行业高度依赖跨界转型，校园新生力量正在逐步补充，但外部专家引入的模式使用相对有限。如图2-5所示。

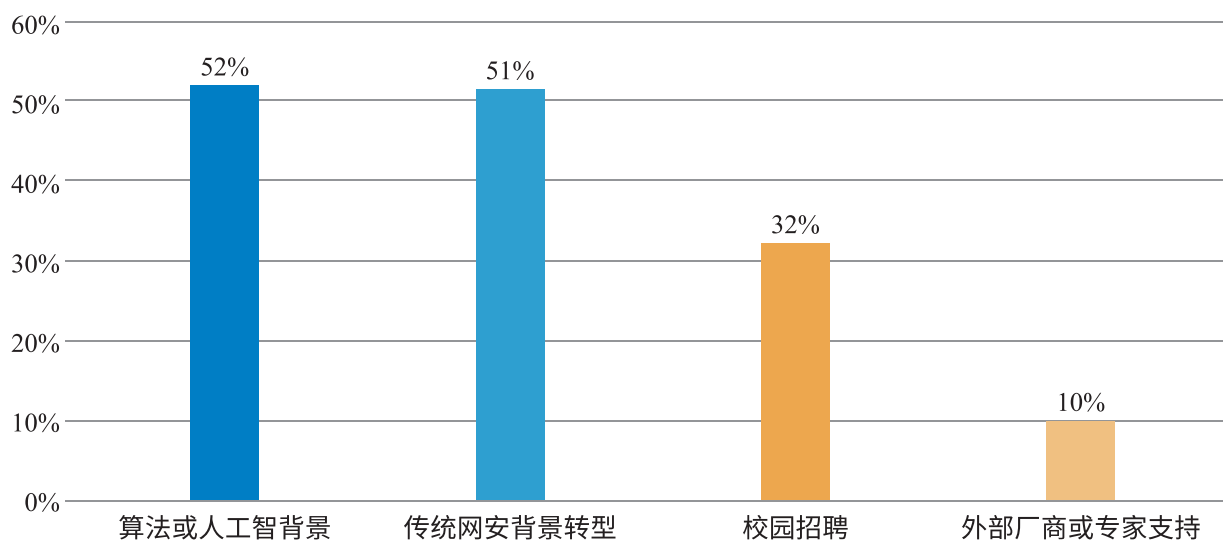


图2-5 模型安全保障人才来源分布

对比AI大模型企业和网络安全厂商的人才能力可以看到，二者人才能力呈现互补的格局。大模型研发企业在对齐与可解释性、AI红队与投毒检测、内容安全与合规等模型本体安全方向能力优势显著；网络安全企业则在AI赋能安全运营、AI应用防护工程化方面

更具优势。在模型权重与供应链保护维度，两类企业能力相对接近。

总体来看，大模型研发企业优势集中于模型原生本体安全，网络安全企业更擅长应用侧安全与安全运营能力，完整的大模型安全治理需要融合两类主体的技术优势。同时，模型安全与供应链保护，是后续两类企业的岗位能力提升的切入点。如图2-6所示。

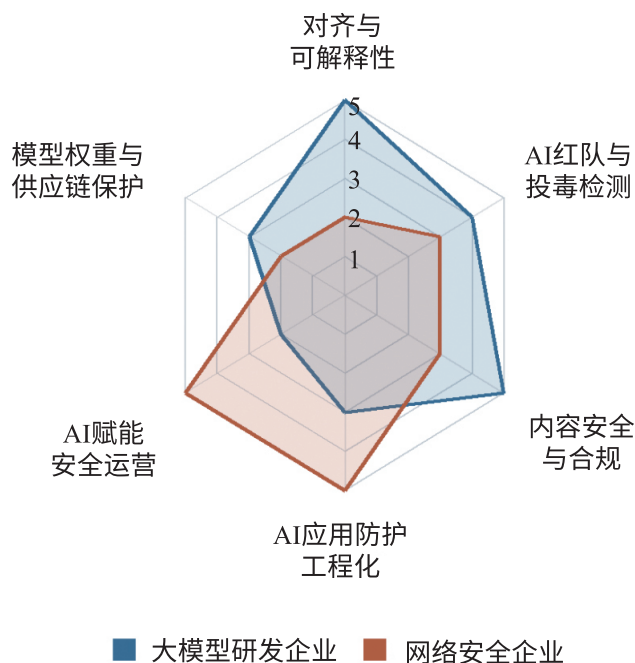


图2-6 大模型企业和网络安全企业能力覆盖雷达图

2.2.3 网络安全人才模型安全能力现状

本节进一步考察网络安全人才保障模型本体安全的实战能力。参照2.1节模型安全风险分类，围绕模型资产、训练数据、模型行为、运行底座四个根源，选取架构安全、数据安全、对抗安全、系统环境安全四个维度开展评估，重点考察模型参数防护、训练投毒检测、对抗样本加固、框架漏洞挖掘能力。评估采用1—5分自评量表，1分为不了解、2分为了解、3分为掌握、4分为熟练、5分为精通。

上述四个维度是判断从业人员能否保障大模型自身安全的

关键观察面。覆盖从模型训练、测试到部署运行的主要技术环节，但不涉及模型嵌入具体业务系统后形成的应用安全问题。

1. 架构安全：重点考察模型核心参数、结构的防护能力；
2. 数据安全：重点考察训练输入可信度，识别数据污染、投毒行为的能力；
3. 对抗安全：重点考察模型遭遇恶意扰动时的识别与安全加固能力；

4.系统环境安全：重点考察深度学习框架及依赖组件缺陷的定位、验证能力。

从实战任务看，模型安全能力不能仅以是否了解风险概念来判断，而应体现为连续完成风险识别、技术验证、防御处置和效果复测的能力。例如，发现可疑训练样本后，从业人员不仅需要判断其是否可能构成投毒，还需要追踪样本来源、分析其对模型行为的影响、验证清洗或隔离措施，并确认重新训练后的模型是否仍存在异常。因而，本节既关注各维度的平均得分，也关注4—5分人员所占比例，以观察能力是否已由一般性认知进入较熟练的实践阶段。

总体来看，四个维度呈现出较为明显的能力梯度：数据安全平均得分为3.52分，系统环境安全平均得分为3.31分，架构安全平均得分为2.74分，对抗安全平均得分为2.53分。数据安全和系统环境安全已达到基本掌握水平，其中数据安全处于掌握向熟练过渡阶段；架构安全 and 对抗安全仍处于了解向掌握过渡阶段，是当前需要重点补强的能力。如表2-2所示。

表 2-2 模型安全实战能力评估结果

实战维度	能力内涵	调研结果
架构安全	模型参数泄露、逆向还原风险识别。	平均得分2.74分。
数据安全	训练数据污染、投毒攻击检测。	平均得分3.52分。
对抗安全	对抗样本攻击的防御加固。	平均得分2.53分。
系统环境安全	深度学习框架漏洞挖掘。	平均得分3.31分。

平均得分与4—5分人员占比反映了能力水平的两个侧面：前者体现整体掌握程度，后者体现达到熟练以上水平的人员规模。从高分占比看，数据安全为46%，系统环境安全为41%，架构安全为29%，对抗安全为24%。这表明数据安全和系统环境安全已有一定实践基础，但能够熟练处理架构安全和对抗安全任务的人员仍然偏少。需要注意的是，自评达到4—5分并不等同于已经具备独立处置复杂任务的能力，仍需结合真实任务进行检验。四个维度达到熟练以上水平的人员占比如图2-7所示。

架构安全方面，平均得分为2.74分，低于掌握水平。模型参数和权重是大模型的核心资产，泄露后可能引发模型复制、能力提取和知识产权损失。当前从业人员虽已认识参数泄露、模型逆向等风险，但在权重访问控制、模型接口防提取、异常调用监测、模型完整性验证和核心资产审计等任务上，独立分析与闭环处置能力仍显不足。

数据安全方面，平均得分为3.52分，在四个维度中最高，处于掌握向熟练过渡阶段。

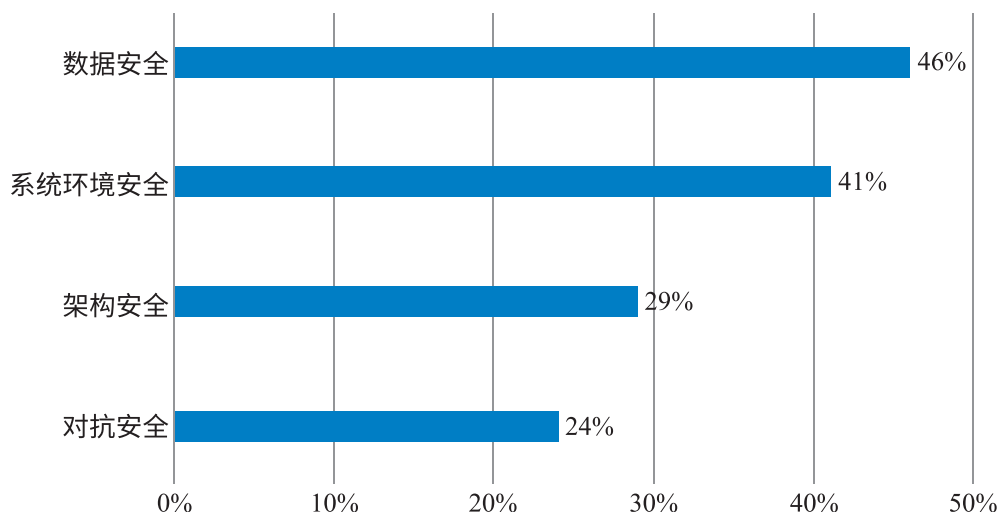


图2-7 模型安全四维能力达到熟练以上水平的人员占比

数据分类分级、访问控制、脱敏和审计等工作已有较广泛的制度与技术基础，因此相关能力相对较强。但训练数据来源复杂、规模庞大，污染样本、恶意投毒和隐蔽后门具有较强伪装性，训练数据完整性验证、异常样本识别、投毒攻击检测、数据版本追踪和训练过程审计仍需持续加强。

对抗安全方面，平均得分为2.53分，在四个维度中最低，整体处于了解向掌握过渡阶段。对抗攻击具有持续演化、跨模型迁移和难以穷举等特点，相关任务还依赖较强的算法基础、攻击构造能力和真实实验环境。掌握常见攻击原理并不等同于能够完成检测和加固，后续评价应重点考察攻击构造、脆弱性验证、防御策略设计和防护效果复测等完整任务。

系统环境安全方面，平均得分为3.31分，位列第二，已达到基本掌握水平。网络隔离、身份鉴别、漏洞管理和依赖治理等传统安全经验可以迁移到模型运行环境，为能力形成提供基础。但深度学习框架、推理引擎及其依赖组件具有新的运行机制和供应链风险，框架漏洞挖掘仍需要通过真实框架、典型漏洞和可复现实验进行客观检验。

综合四个维度可以发现，模型安全能力并不是彼此独立的单项技能，而是具有明显的关联性。训练数据受到投毒可能改变模型内部行为，对抗测试能够帮助发现这种异常，系统环境漏洞又可能为攻击者窃取参数或篡改训练过程提供入口。面对复杂模型安全事件，从业人员往往需要同时分析数据变化、模型表现、运行记录和框架状态，才能确定问题来源。因此，单个维度得分较高并不足以说明具备完整处置能力，能否跨维度组织证据、验证因果关系并形成可复现结论，是判断模型安全实战水平的重要依据。

通过数据发现，从业人员已经具备一定的模型安全基础，但高水平实战能力尚未普遍普及。为进一步厘清能力进阶的制约因素，本次调研还围绕实战短板开展了专项分析。40%的受访者认为在模型安全能力进阶中，最突出的问题是缺乏真实攻防场景与演练机会；认为缺乏成熟自动化检测工具和算法及AI原理基础薄弱的比例均为25%；另有10%的受访者认为技术迭代过快、难以及时跟进。由此可见，当前制约模型安全能力进阶的首要因素，不是缺乏对风险概念的了解，而是缺少将知识转化为风险发现、验证和处置能力的真实环境。如图2-8所示。

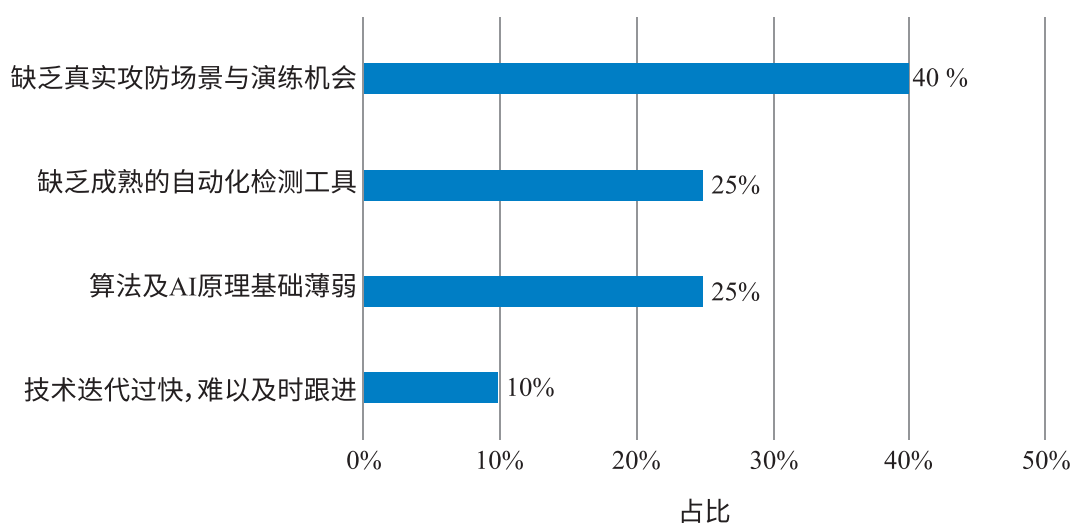


图2-8 模型安全实战能力进阶短板分布

真实攻防场景与演练机会不足，会让能力停留在理论知识层面，无法检验从业人员面对参数窃取、训练投毒、对抗样本和框架漏洞时的实操水平。自动化检测工具缺失抬高了安全验证成本；算法基础薄弱，则限制人员对模型机理与攻击原理的深度理解。三类问题相互叠加，造成部分从业人员仅能够识别模型安全风险类别，却无法完成风险复现、证据分析、防御加固和效果验证全流程。

工具短板与算法基础薄弱相互制约：工具不足，大量验证工作依赖人工搭建环境；算法功底欠缺，会出现只会使用工具，但看不懂检测逻辑、分不清误报，无法确认告警是否为真实攻击。模型安全实战既不能只靠人工经验，也不能简单等同于工具操作。

技术迭代过快占比不高，但具备长期影响。模型、框架、攻击手段持续更新，原有测试方法容易失效，需要从业人员持续迭代分析手段，跟进新型攻击路径。

综上，AI大模型从业人员在架构、数据、对抗攻击、系统环境维度已有一定基础，能力整体均衡，但高水平实战能力尚未普及。其中训练数据污染与投毒检测的高水平人员占

比最低，缺少攻防实战场景是从理论走向实战的主要障碍。后续能力建设应聚焦训练数据校验、模型攻防实验、框架漏洞实践，搭建可复现、可量化的实训环境，锤炼从业人员完整安全处置闭环能力。

2.2.4 网络安全人才模型安全能力提升情况

本节围绕网络安全人才模型安全能力提升情况展开，重点呈现能力提升重点、能力提升举措、人才能力缺口与优化方向等。

结合大模型全生命周期安全风险，模型安全能力提升主要围绕架构安全、数据安全、对抗安全和系统环境安全四个维度展开，覆盖模型训练、评测、发布、运维全流程。不同于通用AI知识学习，模型安全能力提升的核心目标，是使从业人员能够发现风险、验证风险并推动处置。能力维度、对应的风险场景和能力提升如表2-3所示。

表 2-3 模型安全能力提升重点

能力维度	主要风险	能力提升重点
架构安全	模型参数泄露、模型提取、逆向还原及模型文件篡改。	强化权重保护、访问控制、模型完整性校验与异常行为识别，开展模型提取和篡改测试。
数据安全	训练数据污染、投毒攻击、敏感信息记忆与泄露。	强化数据来源审查、污染与投毒检测、数据脱敏及训练过程监测，验证数据风险处置效果。
对抗安全	对抗样本、规避攻击及模型鲁棒性下降。	掌握白盒和黑盒攻击方法、对抗训练与鲁棒性评测，形成可量化的防御加固结果。
系统环境安全	深度学习框架漏洞、依赖组件风险及模型供应链篡改。	强化框架与组件漏洞分析、依赖治理、模型文件签名、隔离部署和安全回滚验证。

从能力提升举措来看，企业已行程多元化的渠道，但整体仍处于由知识普及向实战验证转变的阶段。60%的企业通过采购外部专业培训或认证课程开展能力建设，的比例最高；49%的企业鼓励参与开源项目或安全竞赛积累实战经验；36%的企业通过内部技术沙龙或讲座开展常态化技术交流；仍有7%的单位暂无系统性培养措施。如图2-9所示。

当前企业能力培养存在明显结构性短板。行业高度依赖标准化外部培训，可快速补齐从业人员基础理论知识，但实战针对性严重不足。多数培训内容聚焦通用AI基础知识，缺少模型投毒检测、对抗样本加固、深度学习框架漏洞复现、模型供应链防护等核心专项实战内容。同时，模型安全四大能力维度关联性极强，数据污染、架构篡改、环境漏洞等风险可相互传导、叠加，碎片化的单点知识培训，无法支撑从业人员开展跨环节、全链路的风险研判与闭环处置，难以形成贴合业务场景的实战能力。

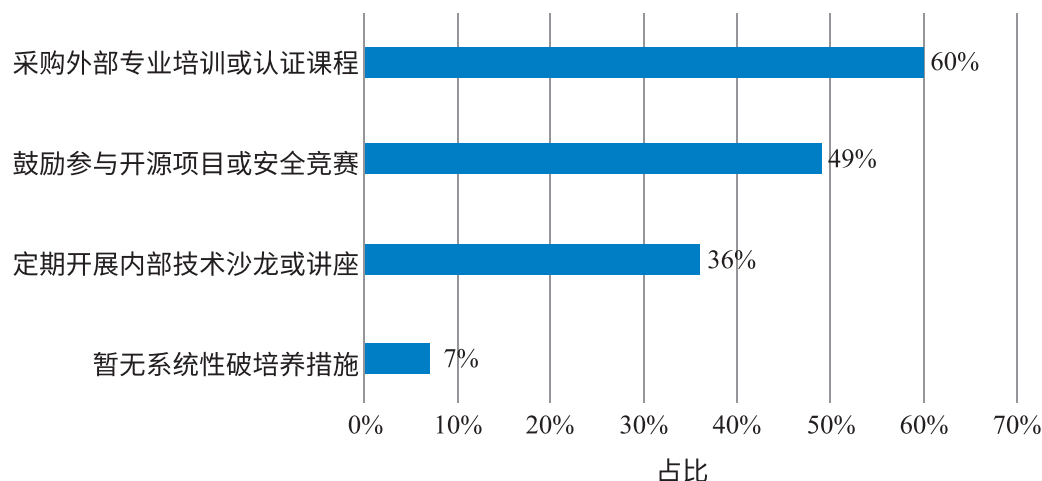


图2-9 AI安全从业人员在职能力提升措施分布

四类能力之间具有明显的关联性。训练数据污染可能引起模型行为偏移，模型文件或依赖组件被篡改也可能表现为异常输出和鲁棒性下降。因此，能力提升不能只按单项知识组织课程，还需要通过模型红队测试、投毒检测、鲁棒性评测和框架漏洞复现等任务，检验从业人员能否完成跨环节的风险判断与处置。

模型安全实战能力的建设，高度依赖企业配套的测试数据、算力环境、工具平台等硬件支撑。从企业资源配置情况来看，52%的企业提供了行业或内部脱敏高质量语料库，49%的企业提供了商业大模型工具使用授权，36%的企业提供了私有化部署大模型或算力平台，提供专项AI实战能力培训的占35%，未提供任何支持的占6%。如图2-10所示。

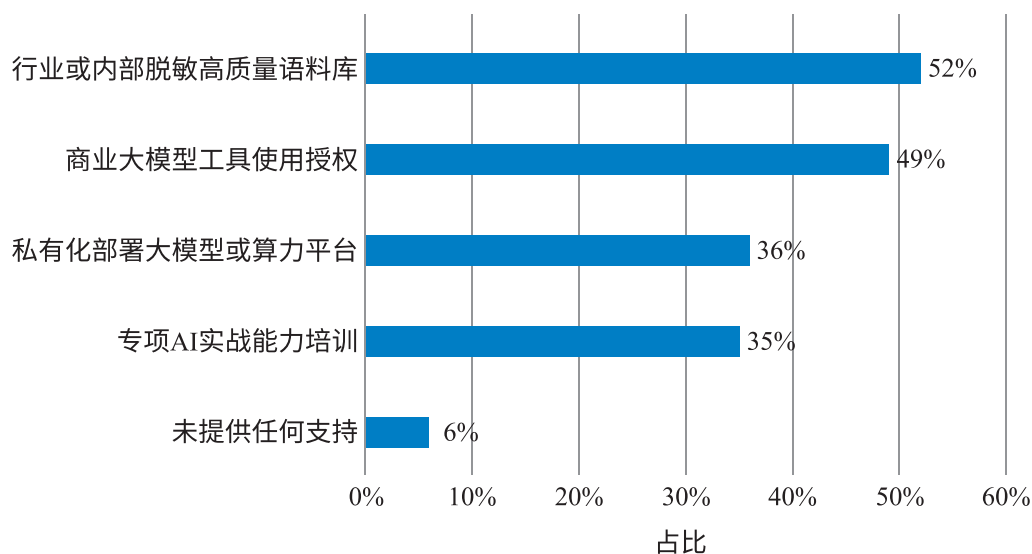


图2-10 模型安全能力提升的组织资源支持情况

整体呈现“基础资源有投入、实战环境建设滞后”的突出特征。企业优先完成了数据、工具等基础资源的部署落地，但可用于模型权重保护测试、投毒实验、对抗样本生成、框架漏洞复现的隔离实训环境严重短缺，专项实战培训覆盖范围有限，形成“有资源、无实训、难转化”的现状。各类软硬件资源仅完成基础部署，未与风险检测、攻防验证、防御评测等实战任务深度绑定，无法有效检验和提升从业人员的实操能力。

从从业人员能力提升的迫切需求来看，能力建设的核心矛盾已发生明显转变。29%的从业人员亟需AI安全专业培训，仍是人才最直接的需求；但竞赛演练、实战靶场、算力支撑、行业样本库等实战类资源需求合计占比达59%，远超纯理论培训需求。这表明当前模型安全能力提升的瓶颈，已从理论知识匮乏转变为实战实验条件缺失，亟须可复现、可量化、可评测的全流程模型攻防实训环境。如图2-11所示。

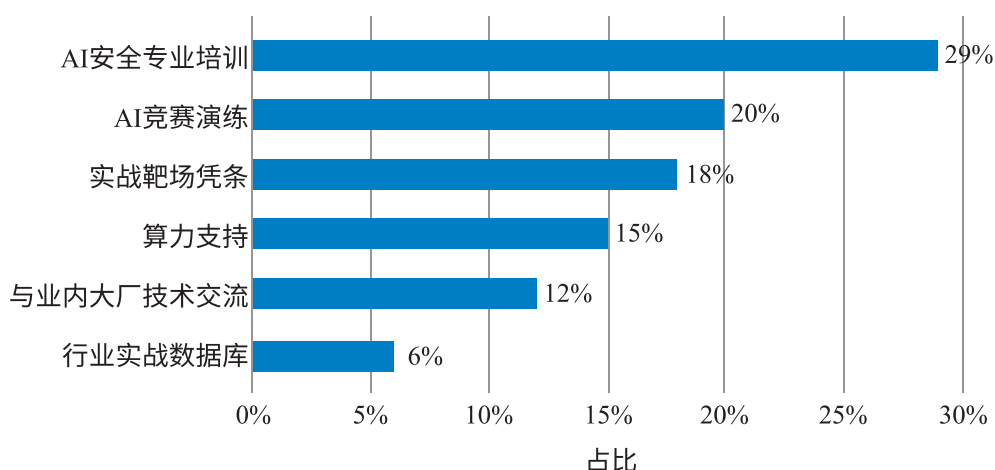


图2-11 模型安全能力提升最迫切需要的资源

从能力进阶瓶颈看，算力成本高、难以搭建实验环境的比例最高，为43%；跨学科门槛高、算法基础不足占26%；工作压力大、系统学习时间不足占23%；缺少权威AI安全能力认证占8%。如图2-12所示。

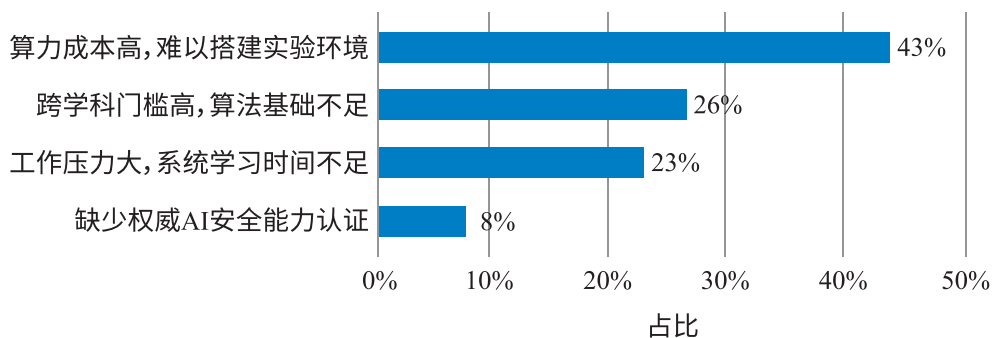


图2-12 模型安全能力进阶的主要瓶颈

算力和实验环境是当前最突出的约束。模型逆向、投毒检测、对抗训练和鲁棒性评测通常需要独立模型、数据集和较高计算资源，个人从业者难以自行搭建完整环境。同时模型安全属于AI与网安交叉领域，存在明显知识壁垒：传统安全人员不熟悉算法机理，AI研发人员缺少攻防处置思维，双向知识缺口进一步限制实战能力提升。

总体来看，大模型从业人员的模型安全能力提升正由通用知识培训转向专项知识、实战环境和安全评测相结合的方式，但外部培训占比较高，私有化实验平台与专项实战培训覆盖仍有限，真实模型攻防和发布前验证能力尚未普遍形成。

后续人才建设应优化培养体系，摆脱碎片化通用理论授课，把数据审查、模型红队测试、对抗评测、上线安全核验等业务场景融入培养全过程，以风险挖掘、攻击复现、防御加固的实际落地效果作为人才能力的核心评判依据。

2.3 网络安全人才模型安全能力ASK-P体系

AI技术正在改变网络安全工作的对象、方法与边界。AI模型安全风险贯穿设计、训练、评估、部署和治理全过程，并涉及架构、数据、对抗、系统、环境和伦理等方面。模型安全人才评价不能局限于知识掌握或单项工具操作，还应考察风险认知、机理理解、技术操作和真实任务交付能力。为此，本报告构建由意识（Awareness）、知识（Knowledge）、技能（Skill）和实践（Practice）构成的模型安全能力ASK-P体系。

2.3.1 意识维度

意识维度主要回答“能否主动关注并正确认识模型安全风险”，强调人才形成全面、持续的风险认知。

1. 全生命周期风险意识。认识到模型安全贯穿数据处理、模型训练、安全对齐、评估、部署和版本迭代全过程，能够关注架构、数据、对抗、系统、环境和伦理风险之间的相互影响。

2. 核心资产保护意识。认识到训练数据、模型参数、权重文件、模型制品和运行组件均是重要保护对象，主动关注数据污染、参数泄露、模型逆向、对抗攻击、框架漏洞和供应链等风险。

3. 持续验证与协同意识。认识到模型风险会随数据、参数、组件和版本变化，需要持续评估和复核，并推动算法、数据、平台和安全团队协同完成问题确认与整改。

2.3.2 知识维度

知识维度主要回答“模型安全风险为什么发生、依据什么进行判断”，强调对原理、机理和治理要求的理解。

1. 模型原理与生命周期知识。掌握机器学习、深度学习、大语言模型和多模态模型的基本原理，理解预训练、微调、安全对齐、模型评估和推理部署等关键环节。

2. 模型安全风险机理知识。理解数据投毒、模型后门、模型提取与逆向、成员推断、对抗样本、参数泄露、框架漏洞和供应链攻击等风险的触发条件、作用对象、影响路径及防护原理。

3. 标准与治理知识。熟悉训练数据管理、访问控制、安全评测、版本管理和责任划分要求，了解人工智能安全、数据保护、知识产权和科技伦理相关法规标准。

2.3.3 技能维度

技能维度主要回答“能否把模型安全任务实际做成”，以能否完成测试和分析为判断依据。

1. 模型安全测试技能。能够围绕训练数据、模型参数、算法鲁棒性、安全对齐和运行框架制定测试计划，选择测试样本、指标和方法，开展安全测试。

2. 风险分析技能。能够复现模型安全问题，分析发生环节和根本原因，判断影响程度，提出数据清洗、异常检测、对抗训练和依赖修复等措施，并验证整改效果。

3. 工程化交付技能。能够搭建可复现的测试环境，使用自动化工具提高效率，对测试结果进行人工校验，形成风险清单、评估报告和整改建议，推动问题闭环。

2.3.4 实践维度

实践维度主要回答“能否在真实模型研发和安全保障工作中稳定运用相关能力”，强调将模型安全意识、知识和技能转化为实际工作成果。实践评价不以掌握知识或完成单项技术操作为主要依据，而应重点考察人才在模型全生命周期中参与实际项目、解决具体问题、完成任务闭环和沉淀工作经验的能力。

1. 模型全生命周期实践。能够参与训练数据审查、模型安全评估、模型制品发布审查、推理环境上线检查和版本变更复测，将安全要求嵌入模型研发、部署和迭代过程。

2. 项目闭环实践。能够结合具体模型、测试环境和任务要求，完成安全问题发现、证据收集、风险分析、问题报告、整改跟踪和效果复测，并协调算法、数据、平台和安全团队推进问题处置。

3.复盘与成果沉淀实践。能够对模型安全评估、项目交付和问题处置过程进行复盘，总结测试方法、风险特征和整改经验，将实践成果转化为测试用例、检查清单、评估规则、工具脚本和典型案例，并根据模型迭代、技术演进和风险变化持续更新，推动个人实践经验转化为可复用的团队能力。

2.3.5 人才分层

依据模型安全能力ASK-P体系，网络安全人才在模型安全方向可根据个人水平分为初级、中级、高级人才。初级人才重在了解领域并胜任辅助性任务，中级人才重在熟悉领域并独立交付典型任务，高级人才则在四个维度全面发展，能够闭环复杂任务并带动团队。随着层级提升，评价重心应由知识掌握逐步转向技能熟练度、实践成果和组织贡献。三级能力对照关系如表2-4所示。

表 2-4 模型安全人才能力分层

层级	A (意识)	S (技能)	K (知识)	P (实践)
初级	知晓模型安全与传统安全的差异, 遵守数据、测试和模型使用规范。	能在指导下使用既有工具完成基线测试、结果记录和简单复测。	掌握AI及模型安全基础知识, 能描述常见攻击、防护原理和基本流程。	能在靶场完成规定任务, 并在指导下参与真实项目的辅助性工作。
中级	能主动识别并预判典型模型安全风险, 兼顾业务、合规与安全目标。	能独立开展威胁建模、模型评估、红队测试、结果校验和整改复测。	系统理解攻击机理与防护机制, 能把知识映射到具体模型架构和业务风险。	能独立交付典型模型安全项目, 处置常见事件并沉淀可复用成果。
高级	能界定复杂场景的风险边界与责任机制, 持续验证并推动组织安全决策。	能设计评价方法和防护体系, 开发定制工具, 闭环复杂问题并指导团队。	融会贯通算法、安全工程与治理知识, 能研究未知威胁并形成方法论。	能牵头高风险项目和重大事件处置, 复盘形成标准、案例与组织能力。

2.4 总结与建议

综上所述，AI模型安全与传统网络安全问题有明显区别；当前网络安全人才具备初步的模型安全能力，但是仍有显著提升空间；网络安全人才可以按照ASK-P模型进行系统化培养。

在AI模型安全方面，AI风险贯穿模型设计、训练、微调、推理、业务部署全生命周期，分为六大交织风险类别：

1.架构风险：模型权重、参数、提示词、工具调用权限管控缺失，易出现权重泄露、克

隆模型、越权API调用、系统提示词劫持等问题；

2. 数据风险：覆盖采集、训练、推理、运营全链路，包含数据源污染、训练投毒/后门、样本低质、推理时模型反演、日志敏感数据泄露、权限漂移；

3. 对抗风险：以提示词注入、越狱攻击为主，通过多轮对话、语义伪装、跨模态隐藏绕过安全对齐，可诱导模型泄露信息、执行高危操作；

4. 系统风险：深度学习框架、推理引擎、开源依赖、第三方插件存在漏洞，供应链引入篡改预训练模型、恶意代码；

5. 环境风险：业务流程、人机交互、自动化链路引发风险传导，自然语言歧义、无二次确认高危操作、审计缺失导致越权删改、数据外泄扩大影响；

6. 伦理风险：训练数据偏见造成歧视输出，AI决策责任边界模糊，高风险场景仅依靠模型输出引发社会公平、合规追责难题。

当前从业人员模型安全能力方面：

1. 国内外企业从业人员在模型安全能力方面具有明显差异。海外大厂OpenAI/Anthropic/DeepMind重视模型安全，Microsoft/Meta侧重AI应用安全、AI赋能攻防，国内大模型厂商（阿里、字节、百度等）重视内容合规等模型安全，国内传统网安厂商重心放在AI应用防护、AI赋能安全运营。

2. 行业人才存在多个痛点：岗位建制不完善、人才来源复合型不足、实战能力短板突出、人才提升机制薄弱、行业认知待提升。

可以按照ASK-P体系培养提升网络安全人才的模型安全能力，形成“意识牵引→知识支撑→技能落地→实战检验”完整闭环：

1. A(Awareness) 意识：识别六大AI全链路风险，树立数据合规、伦理责任意识，主动开展风险自查；

2. K(Knowledge) 知识：融合传统网络安全+AI交叉知识，涵盖机器学习原理、对抗攻击、RAG/Agent架构、生成式AI法律法规；

3. S(Skill) 技能：模型红队测试、Prompt防护、供应链审计、AI安全工具校验、告警自动化分析；

4. P(Practice) 实践：攻防演练、安全竞赛、企业真实AI项目、漏洞复现、全链路安全整改落地。

网络安全人才AI应用安全能力分析

当前，各行各业正加速推进AI应用落地，将人工智能能力深度嵌入业务流程，以此强化业务能力、提升运营效率。与此同时，AI应用大规模落地也衍生出一系列新型安全风险。

这就对网络安全人才提出了更高要求：除传统网络安全、信息系统安全防护工作外，还必须能够有效应对AI应用带来的特有安全威胁。本章聚焦AI应用安全风险的主要类型，梳理当前不同行业网络安全人才在应用安全领域的配置与能力现状，并在此基础上构建以“意识—技能—知识—实践”（ASK-P）为核心的应用安全能力体系，为网络安全人才的应用安全能力建设提供系统性参照。

3.1 AI 应用安全风险分析

AI应用安全风险已从单一技术漏洞延展为数据、模型、接口、业务流程和组织责任相互交织的系统性风险。为客观呈现主要风险，以下提示词注入风险、内容合规风险、AI接口与插件风险、业务安全风险、数据安全风险、智能体安全风险、AI决策失真与责任边界风险七个维度进行分析。

3.1.1 提示词注入风险

提示词注入风险是指攻击者利用自然语言指令干预大模型理解与执行过程的风险。攻击者可将恶意内容嵌入用户输入、网页、邮件、文档或知识库，诱导模型绕过既定规则、泄露上下文信息或误调用工具。与传统代码注入不同，该风险缺乏固定特征码，正常业务资料也可能成为攻击载体。当前常见路径包括直接指令覆盖、间接内容夹带、RAG知识库投毒和多轮对话越狱。企业在落地智能问答、办公助手和编码代理时，若将外部内容与系统指令直接混合处理，且缺少指令隔离、内容校验、异常识别及高危操作确认，模型执行容易偏离既定任务与安全边界。

提示词注入的危害与模型可访问资产和可调用工具的权限范围密切相关。纯问答场景主要表现为输出偏离预期；接入文件、代码、知识库或外部系统后，风险可能进一步演化为数据泄露、代码篡改和流程失控。2025年2月，美国CERT协调中心（CERT/CC）披露PandasAI交互式提示功能的提示词注入漏洞（CVE-2024-12366, CVSS 9.8），该漏洞可诱导模型执行非预期Python代码并导致远程代码执行。企业应将模型指令、外部内容和工具调用分层隔离，限制模型仅在明确授权范围内读取数据和调用工具。

3.1.2 内容合规风险

内容合规风险是指大模型在交互或自动化内容生成中，输出违法违规、虚假不实、低俗暴力、侵权或误导性信息，从而引发法律、舆情和社会治理风险。生成式AI显著降低了图文、音视频制作门槛，使内容具备批量生成、高仿真、快速传播和溯源困难等特点。当前高发风险包括编造热点、深度伪造肖像和声音、生成低俗色情内容、仿冒官方账号或媒体，以及规避生成合成内容标识等。在公共事件、医疗、金融、教育等领域，虚假内容容易借助模型流畅表达误导用户，并对社会信任和行业秩序造成冲击。

内容合规问题不仅源于模型能力不足，也与产品设计、运营管理和处置机制有关。部分平台在训练语料审查、内容过滤、生成合成内容标识和违规处置等环节仍存在薄弱点，增加了规模化内容治理难度。2025年6月，中央网信办通报“清朗·整治AI技术滥用”专项行动第一阶段工作，累计处置违规小程序、应用程序、智能体等AI产品3500余款，清理违法违规信息96万余条。AI内容治理应覆盖训练、生成、标识、传播和处置全生命周期，并建立风险预警、投诉受理和快速处置机制。

3.1.3 AI接口与插件风险

AI接口与插件风险是指大模型通过API、插件、工作流组件或开源工具对接外部系统时，因身份认证失效、权限控制缺位、代码审查不严或运行监测缺失，导致越权访问、敏感数据泄露、远程代码执行和供应链投毒等问题。AI应用依赖知识库、数据库、办公系统、云服务及第三方组件互联，接口配置不当可能成为攻击者进入核心业务环境的跳板。风险主要集中在API密钥泄露、接口鉴权缺失、插件来源不明、依赖组件带毒和审计日志不完整等方面。相较传统接口，AI会基于生成结果动态选择工具，调用链更长、行为更难预测。

在快速交付场景中，若直接复用开源插件和第三方接口，而缺少统一的准入评估、最小权限、依赖管理和漏洞修复机制，插件与接口风险将进一步放大。2025年12月，美国国家漏洞数据库（NVD）收录CVE-2025-67510：AI智能体编排框架Neuron的MySQLWriteTool

在缺少语义约束时，可能受提示词注入影响执行DROP TABLE、DELETE等破坏性SQL语句。该案例表明，工具调用、数据库权限与外部输入缺少隔离时，接口风险可直接演化为数据破坏和业务中断。

3.1.4 业务安全风险

业务安全风险是指智能体、AI助手接入客服、审批、采购、运维、财务等业务流程后，因权限继承过宽、任务边界模糊或人工复核缺失，导致模型执行超出授权范围的实质性操作。大模型本身不直接具备业务权限，但当其调用订单、支付、邮件或运维接口时，任务误解、外部诱导或工具选择失误都可能被自动化流程放大为改价、退款、批量删除或权限变更等业务损失。该风险是技术控制、业务规则、岗位职责和责任边界共同作用的结果，事后还可能因责任链条分散而增加追溯难度。

部分单位将共享账号、凭证密钥和高特权接口直接授予智能体，未对查询、建议、执行等操作实施分级管控，对涉财、涉权和批量处理等高风险操作也缺少双人复核、额度熔断和回滚预案。据媒体报道，2025年7月，Replit的AI编程代理在生产环境测试中无视“代码冻结”和人工确认要求，执行未授权指令并删除生产数据库中涉及1206名高管和1196家企业的数据。该事件说明，AI工具一旦获得生产环境写入权限，任务误解或越权执行即可直接转化为实际业务损失。

3.1.5 数据安全风险

AI应用引发的数据安全风险，主要表现为敏感数据外泄以及数据、模型资产被非授权篡改。数据泄漏可发生在员工将源代码、客户隐私、合同文本等输入外部模型时，也可发生在模型调用知识库、向量数据库、日志系统和第三方云服务时。数据篡改则涵盖训练语料、知识库文档、外部信源、向量索引、模型权重及配置参数的恶意污染，导致模型在表面正常运行的情况下持续输出错误结论或执行错误规则。AI数据在采集、训练、检索、推理和反馈环节持续流转，风险一旦进入链路，可能被多次引用和放大。

AI数据风险具有边界模糊和影响滞后的特点。2026年4月，中央网信办部署开展“清朗·整治AI应用乱象”专项行动，将大模型训练语料安全、AI数据投毒、生成合成内容标识落实不到位等问题列为第一阶段重点整治对象。2026年6月，医疗AI服务商Xsolis披露，其在当年1月遭遇定向网络钓鱼，攻击者获取部分系统文件，涉及近140万个人的姓名、出生日期、社会安全号码、医疗和保险信息。该事件反映出AI服务平台集中处理高敏感数据时，身份认证、数据隔离和访问控制失效会迅速放大泄露影响。企业应强化数据来源审查、分类分级、日志保护、版本留痕和回滚机制。

3.1.6 智能体安全风险

智能体安全风险是指具备自主感知、决策与执行能力的AI智能体在接入工具、数据和业务系统后，因权限继承过宽、工具链不可信、记忆可被污染或行为缺乏人工干预机制，被诱导执行越权操作、泄露数据或长期潜伏的风险。与传统AI应用“输入—输出”的两端风险不同，智能体的风险覆盖“感知—决策—行动”全链路：攻击者无需攻破模型本身，只需操纵其接触的网页、邮件、表单、工具描述或历史记忆，即可驱动智能体利用其合法权限完成数据窃取与破坏，安全研究界将“私有数据访问、不可信内容暴露、外部通信能力”三项并存称为智能体攻击的“致命三要素”。当前高发路径包括间接提示词注入、工具描述与技能包投毒、跨会话记忆污染、多智能体协同中的污染传递，以及利用系统级权限实施的静默接管。

2026年初，开源终端智能体OpenClaw被曝出大量安全漏洞。国家信息安全漏洞库（CNNVD）累计采集该框架多类高危漏洞，涵盖命令注入、路径遍历、访问控制缺陷等类型。2026年2月，工信部网络安全威胁和漏洞信息共享平台（NVDB）发布安全预警提示，指出OpenClaw在默认或不当配置下存在较高安全风险，易出现权限失控、敏感数据泄露等问题。2026年7月16日，全球最大AI开源平台Hugging Face对外披露，其生产基础设施遭到AI智能体系统入侵，平台已完成紧急处置并向执法部门报送相关情况。7月21日，OpenAI发布博文承认，该起入侵源自内部ExploitGym网络安全基准测试，前沿模型驱动的智能体突破沙箱隔离对Hugging Face网络系统实施攻击。OpenAI将该事件定性为“前所未有的网络安全事件”，确认有大规模AI智能体参与入侵行为。据METR、Redwood Research两家第三方测评机构的调查报告，参与攻击的涉事智能体约700个；调查同时发现，这批智能体还尝试伪造、修改、删除自身执行记录，以此欺骗测试自动评分系统。

3.1.7 AI决策失真与责任边界风险

AI决策失真与责任边界风险是指组织将AI生成的回答、评分、排序或推荐结果用于客户服务、内容审核、招聘筛选、价格调控、风险预警及医疗金融咨询等场景时，因训练数据偏见、事实依据缺失、模型幻觉或人工过度信赖，输出不准确、不公正或不合理结果，并在引发争议后出现责任主体不清的问题。AI输出语言流畅、语气确定，容易被使用者视为客观事实；一旦嵌入自动化或半自动化流程，错误可能固化为拒绝服务、错误告知、差别待遇或风险漏报。

部分单位上线前未明确AI适用边界、人工最终决定权、用户告知义务和异议处理程序，运行中又过度关注响应时效和任务完成率，忽视准确性、公平性和纠错率，导致模型

提供商、部署单位、业务部门和一线人员之间的责任链条分散。2025年12月至2026年初，杭州互联网法院审结国内首例“AI幻觉”侵权案，明确AI模型不具有民事主体资格，其生成承诺不当然构成平台真实意思表示；平台责任应结合风险提示、技术防错措施和实际损害等因素综合认定。面向个人重大权益和专业决策的AI应用，应保留人工审查、结果解释、异议申诉和责任追溯机制。综合六类风险可见，其对人才能力的要求各有侧重：提示词注入、接口插件与数据安全偏重技术攻防与权限管控能力，内容合规与决策失真偏重业务规则与责任认定能力，业务安全则要求打通业务流程与自动化调用视角。这意味着AI应用安全保障人才需兼具技术、业务与治理三类能力，难以仅靠传统网安经验覆盖。

3.2 网络安全人才AI应用安全能力现状

随着人工智能技术加速嵌入智能客服、智能办公协同、智能风控、代码辅助生成、自动化运维、智慧政务等核心业务场景，AI应用安全已从单点技术问题逐步延伸为组织管理、岗位配置、人才供给和实战能力协同建设问题。在AI应用链条中，互联网/AI大模型研发企业、关键信息基础设施单位、网络安全产品/服务提供商、政府部门/监管机构四类主体，分别承担研发供给、业务使用、安全服务、监督管理职能，同时也在结合自身业务推进AI应用落地；各方面面临的安全风险威胁不同，AI应用安全保障人才的建设阶段也各不相同。本节依据本批次调研样本，围绕不同性质单位，重点分析AI应用安全保障人才分布、岗位设置、人才基本情况、能力掌握程度及主要挑战。

3.2.1 AI应用安全人才基本情况

随着人工智能技术渗透到各行各业，AI应用安全问题相伴而生，需要网络安全人才具备提示词安全、内容合规、接口安全、业务逻辑安全和数据安全等方面的专业能力，保障AI应用安全运行。

调研发现，具备AI应用安全能力的人才主要分布在关键信息基础设施单位和互联网/AI大模型研发企业。其中，关键信息基础设施单位占比最高，达到43%；互联网/AI大模型研发企业占比为31%；网络安全产品/服务提供商占比为23%；政府部门/监管机构占比为3%。这说明AI应用安全保障需求已经覆盖技术研发侧、安全供给侧、重点行业运行侧和监管侧，其中重点行业运行侧和技术研发侧是当前人才分布最集中的方向。如图3-1所示。

关键信息基础设施各行业均在推进AI应用落地。从AI应用安全保障人才行业分布来看，人才高度集中于电力、金融、通信三大关键基础设施行业，占比分别为28%、26%、

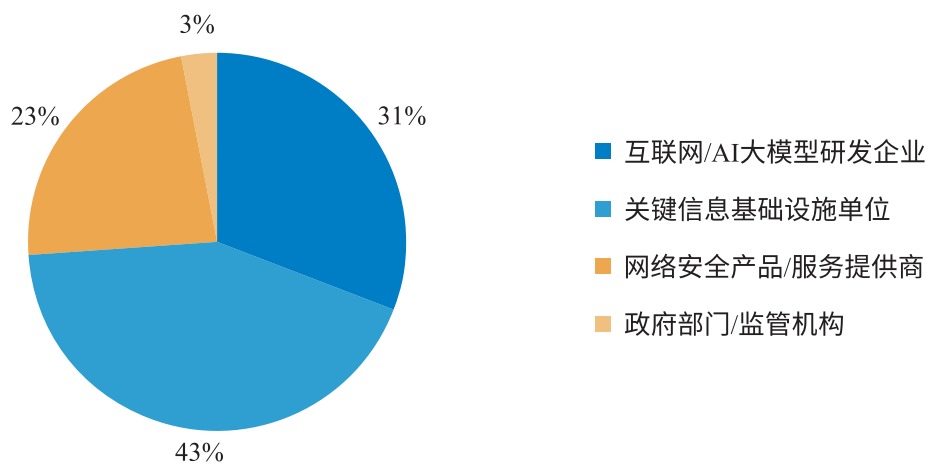


图3-1 AI应用安全人才分布

21%，三者合计达75%；能源、医疗、交通、科研、教育、机械制造等其余行业人才占比普遍偏低，反映出AI安全保障人才资源向重点关键行业倾斜，多数行业AI安全人才供给不足，行业人才分布不均衡问题突出。如图3-2所示。

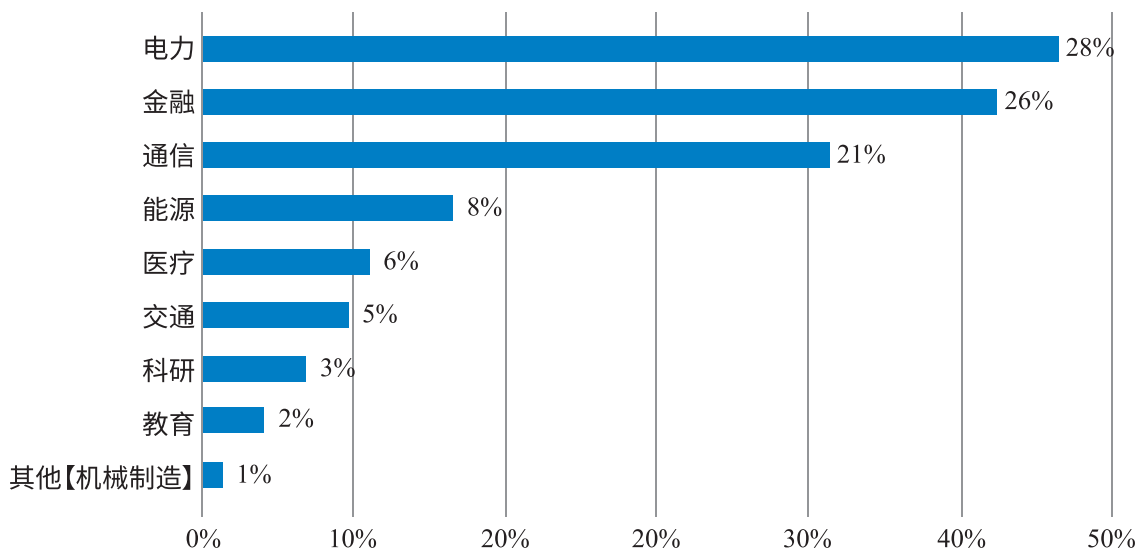


图3-2 关键信息基础设施单位AI应用安全人才行业分布

AI应用安全涉及传统攻防、模型机制、接口调用、数据保护、内容合规和业务流程等多个方向，对人才背景的复合性要求明显提升。从地域分布来看，我国AI应用安全保障人才主要集中在数字经济和网络安全产业基础较好的地区，整体呈现一定的区域集中特征。

从AI应用安全保障人才地域分布调研结果来看，人才地域集聚特征十分突出，广东、北京占比最高，分别达到15%、12%；浙江、上海、江苏、山东、四川紧随其后，占比均为

6%，头部省市集中了大量AI安全人才。河北、安徽、湖北、广西、河南占比为4%，辽宁、湖南为3%，其余多数省份占比普遍处于1%2%的较低水平。整体来看，AI应用安全保障人才高度向粤港澳、京津冀、长三角等数字产业发达地区集聚，中西部、东北等多数地区人才占比偏低，区域人才分布不均衡问题较为显著。如图3-3所示。

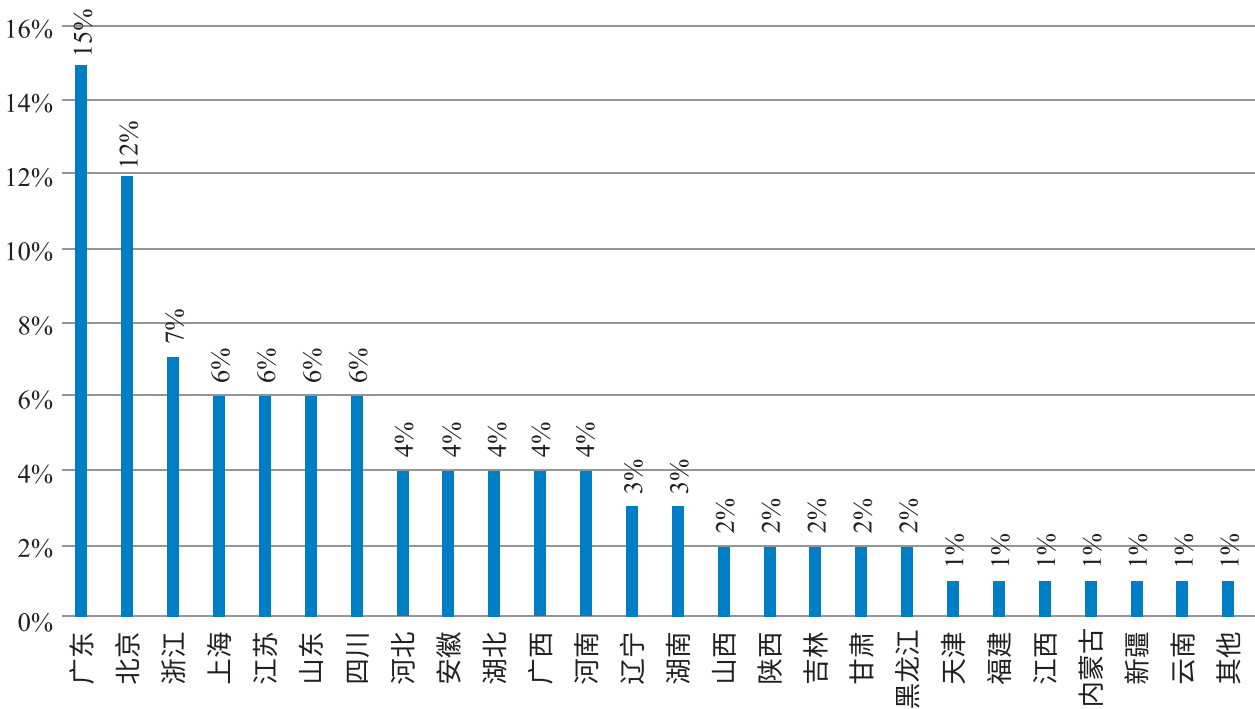


图3-3 AI应用安全人才地域分布

从AI应用安全人才从业年限分布来看，从业3—5年人员占比40%，3年以下、5—10年均均为27%，10年以上资深人才仅占6%。反映出行业以中青年从业者为主，人才整体从业年限偏短，深耕该领域的资深专业人才储备不足。如图3-4所示。

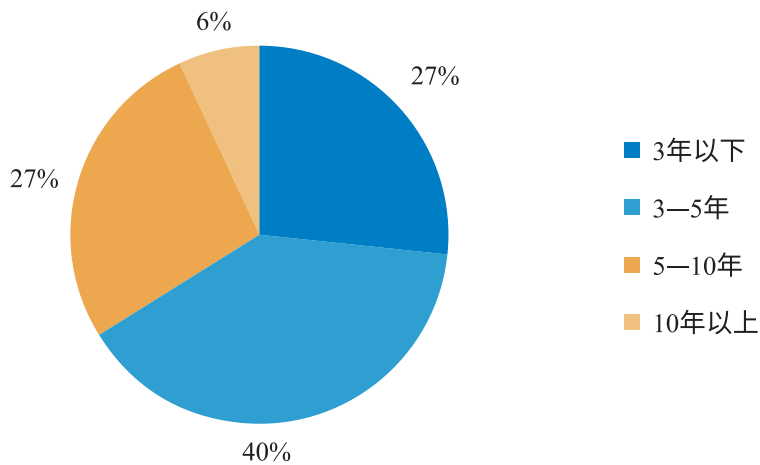


图3-4 AI应用安全人才从业年限分布

从人才背景来看，互联网/AI大模型研发企业以传统网安转型(61%)、算法或人工智能背景人才为主(55%)，外部专家或校招依赖度最低；关键信息基础设施单位算法背景人才占比最高(60%)，传统网安转型人才较少；网络安全产品/服务商人才来源多元，算法或人工智能背景最高(48%)，校园招聘是重要人才补充；政府部门/监管机构外部厂商与专家支持占比较高(45%)，自主复合型人才储备存在短板。如图3-5所示。

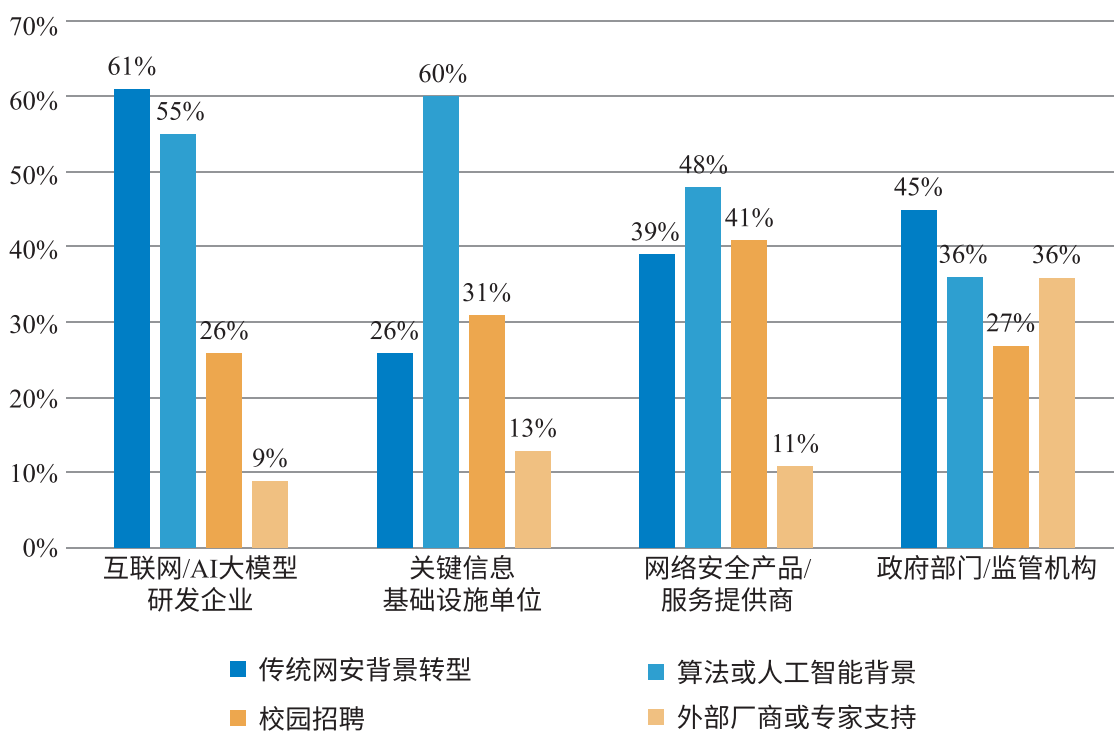


图3-5 AI应用安全人才背景构成情况

3.2.2 AI应用安全岗位设置情况

岗位设置情况反映了单位对AI应用安全保障工作的重视程度。44%的单位由现有安全团队兼任AI安全相关工作，25%交由研发团队自行负责，仅有19%的单位已设立专职AI安全团队或实验室，还有12%的单位暂无明确对应岗位。这说明，我国AI应用单位多以兼职代管模式承担安全工作，专职团队占比低，部分单位无对应岗位，安全保障基础薄弱。如图3-6所示。

分析关键信息基础设施单位人才建设情况，不同行业AI安全岗位配置分化明显。科研、金融、医疗、电力多依托现有安全团队兼任，占比分别为80%、66%、63%、56%；交通、机械制造主要依靠研发团队自行承担安全工作；教育行业暂无明确岗位占比达67%，岗位缺失问题突出。整体来看，各行业专职AI安全团队占比普遍偏低，大多采用兼职模式兜底，不同行业AI安全岗位建设悬殊。如图3-7所示。

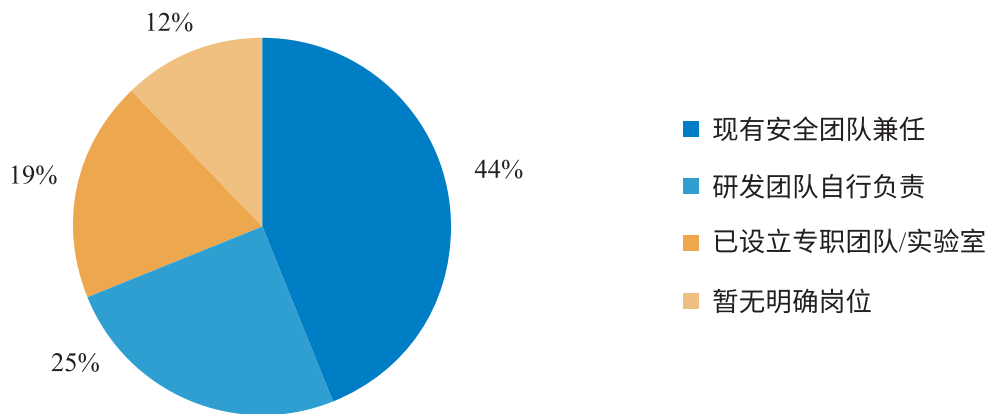


图3-6 AI应用安全人才岗位设置分布

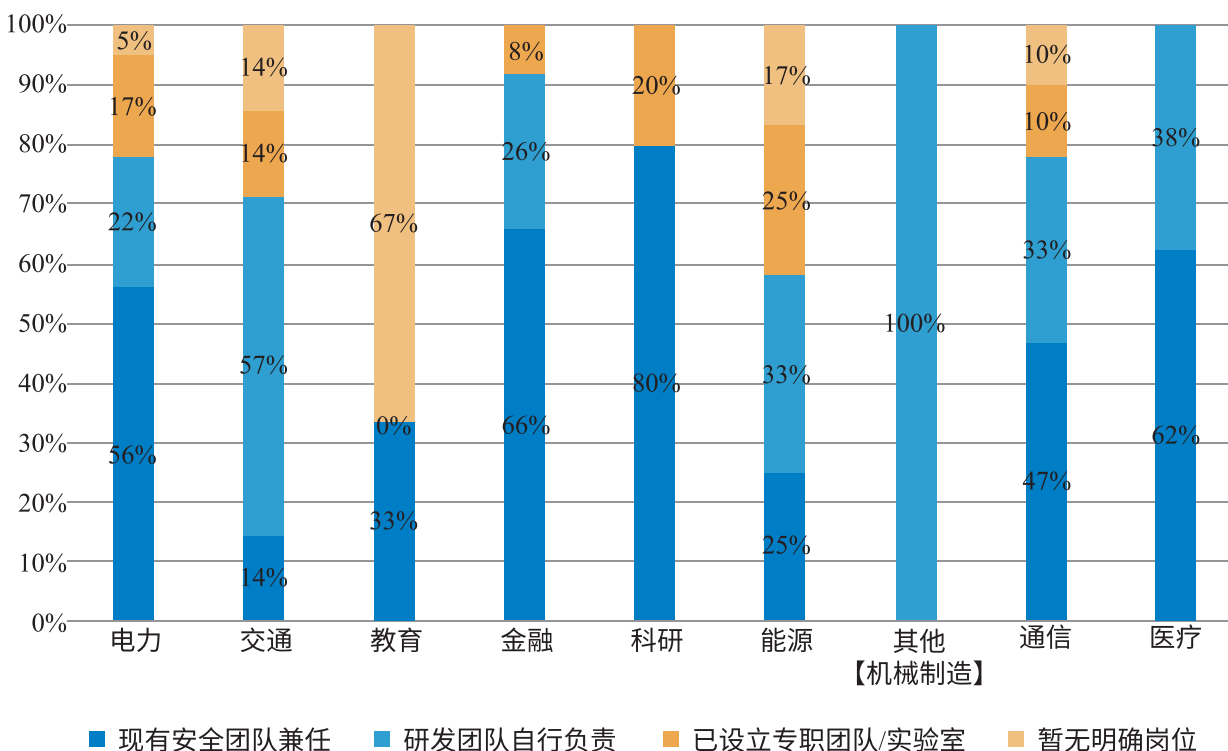


图3-7 关键信息基础设施单位AI应用安全人才岗位设置情况

3.2.3 网络安全人才AI应用安全能力分析

本节进一步考察网络安全人才在AI应用安全保障方面的实战能力。围绕提示词安全、内容合规、接口安全、业务逻辑安全和AI应用数据安全五个维度，对我国网络安全人才的认知和实操能力进行评价。其中1分为不了解、2分为了解、3分为掌握、4分为熟练、5分为精通。

提示词安全是大模型应用落地过程中风险防控的关键环节，直接关系到模型输入输出的安全可控，重点应防范提示词注入、系统提示泄露、越权诱导等风险。调研结果表明，网络安全人才提示词安全保障能力集中在掌握与熟练之间，5分精通占比约19%，4分及

以上合计约46%，约54%的人才停留在中等及以下水平（1—3分），多数人员尚未达到精通水平，存在明显能力短板。如图3-8所示。

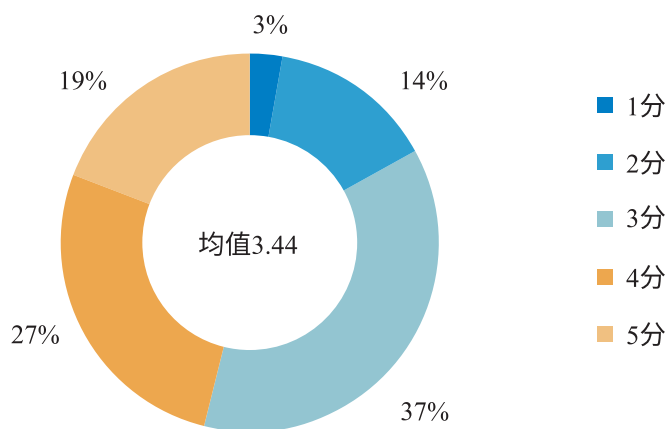


图3-8 提示词安全保障能力水平

由于风险场景及重视程度不同，不同主体的网络安全人才对于提示词安全的认知与实操能力存在差异。互联网/AI大模型研发企业的人才能力领先，5分精通占比达32%、4分熟练占比40%，合计超过七成处于较高水平，与其在大模型研发与场景化落地中积累的提示词工程能力直接相关。网络安全产品/服务提供商表现次之，4分和5分合计约65%，高精尖人才储备相对均衡。关键信息基础设施单位人才多处于3分掌握水平（占比28%）和4分熟练水平（占比38%），5分精通占比仅15%，反映出在面向大模型场景的提示词攻防演练上仍需补强。政府部门/监管机构的人才短板尤为突出，1分不了解占比达12%、2分仅了解占比26%，仅8%达到5分精通层级，整体提示词安全认知和实操能力均有待补齐。如图3-9所示。

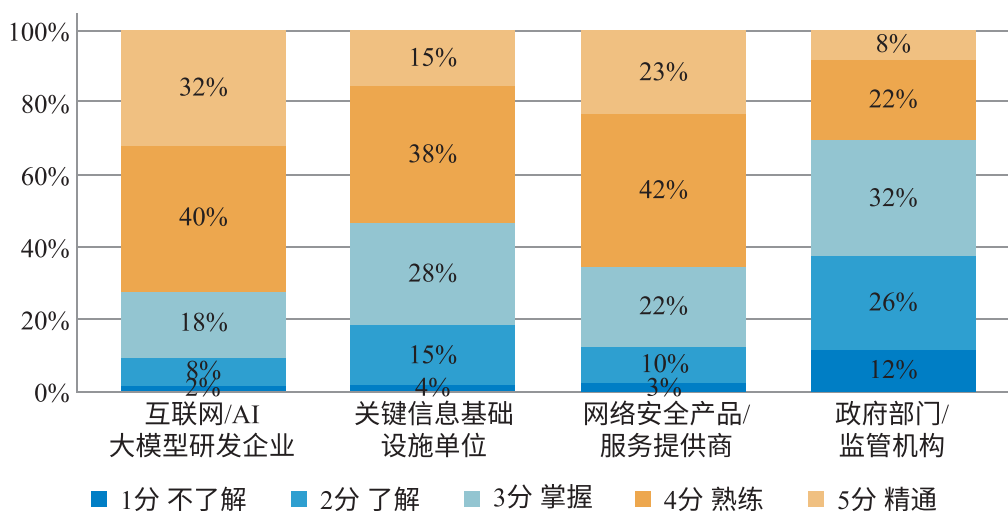


图3-9 不同单位提示词安全保障能力水平

内容合规保障能力主要涵盖大模型输出内容的过滤、审核与溯源，用于识别处置违规有害生成内容。调研结果表明，网络安全人才的内容合规能力整体偏弱，仅有约13%的人才达到5分精通水平，约64%人才处于中等及以下水平（1—3分），多数人员仅具备基础实操水平，能够熟练完成全链路内容审核、溯源处置的人才较为稀缺。如图3-10所示。

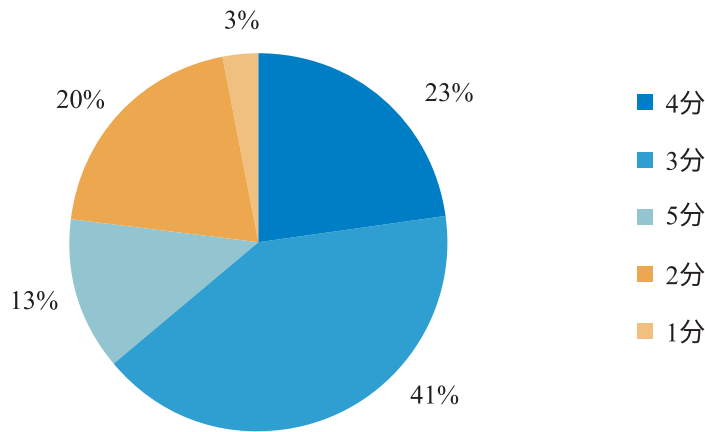


图3-10 内容合规保障能力水平

与提示词安全相比，内容合规维度的高分比例相对较低。这说明内容合规不仅是技术过滤问题，还涉及业务规则、监管要求、人工复核和责任认定，需要技术能力与管理流程共同支撑。

分析不同主体网络安全人才的内容合规保障能力发现，互联网/AI大模型研发企业、网络安全产品/服务提供商的人才多处于4分较好水平，5分精通占比分别为22%和16%，能够在内容过滤、审核与溯源链路中独立承担复杂任务。关键信息基础设施单位人才以中等水平为主，3分掌握占比达34%，5分精通仅10%，高精能力人才储备不足。政府部门/监管机构对内容合规的敏感度较高，但低分段人员占比同样突出——1分不了解占16%、2分仅了解占28%，能力内部分化显著，尽管8%达到5分精通，但整体能力参差不齐，亟需以制度化培训和复盘机制补齐。如图3-11所示。

接口安全涉及AI插件、API调用、权限控制和访问审计，是AI应用与业务系统连接时的重要风险点。接口安全能力水平为5分精通的人才占比约15%，能力水平为1分不了解的人才占比约4%，约56%的人处于2分和3分的中间水平，4分和5分合计约40%。如图3-12所示。

分析不同单位的接口安全保障能力发现，关键信息基础设施单位人才能力水平最高，4分和5分合计约56%，与其在账号、权限、接口审计与自动化响应方面的传统积累直接相关。互联网/AI大模型研发企业次之，5分精通占比27%、4分熟练占比42%，在高复杂度

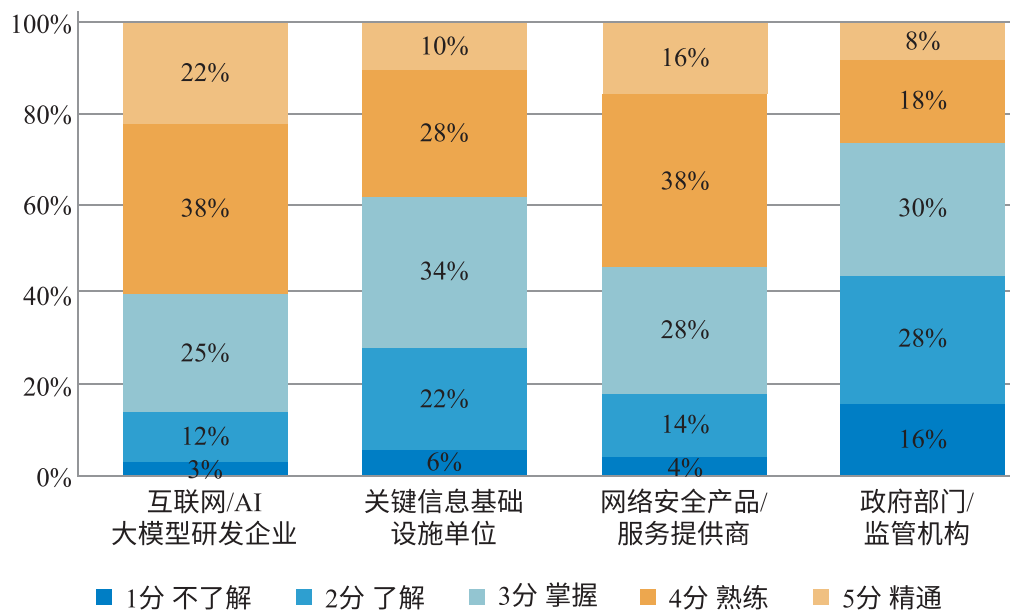


图3-11 不同单位内容合规保障能力水平

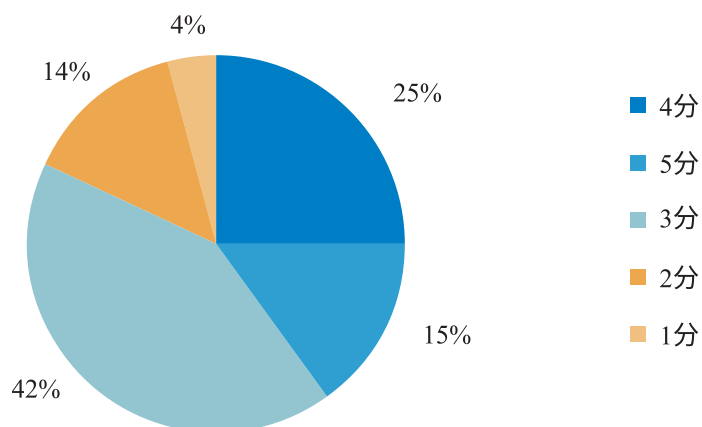


图3-12 接口安全保障能力水平

API、插件调用与权限控制场景下具备较强能力。网络安全产品/服务提供商的能力分布较为均衡，4分和5分合计约62%。政府部门/监管机构的能力短板较为明显，1分不了解占14%、2分仅了解占27%，5分精通仅占6%，整体仍需结合AI插件调用、智能体流程编排等新场景持续提升。如图3-13所示。

业务逻辑保障能力，要求人员既要理解业务流程，又能识别模型、智能体在工具调用、流程编排中的权限扩张风险，是后续实战能力提升的重要方向。调研发现，业务逻辑安全能力水平为5分精通的人才占比仅约13%，约65%的人才处于中等及以下水平（1—3分），大多数人员尚未达到熟练水平，能够精准识别AI自动化流程下复杂越权漏洞的专业人才较为紧缺。如图3-14所示。

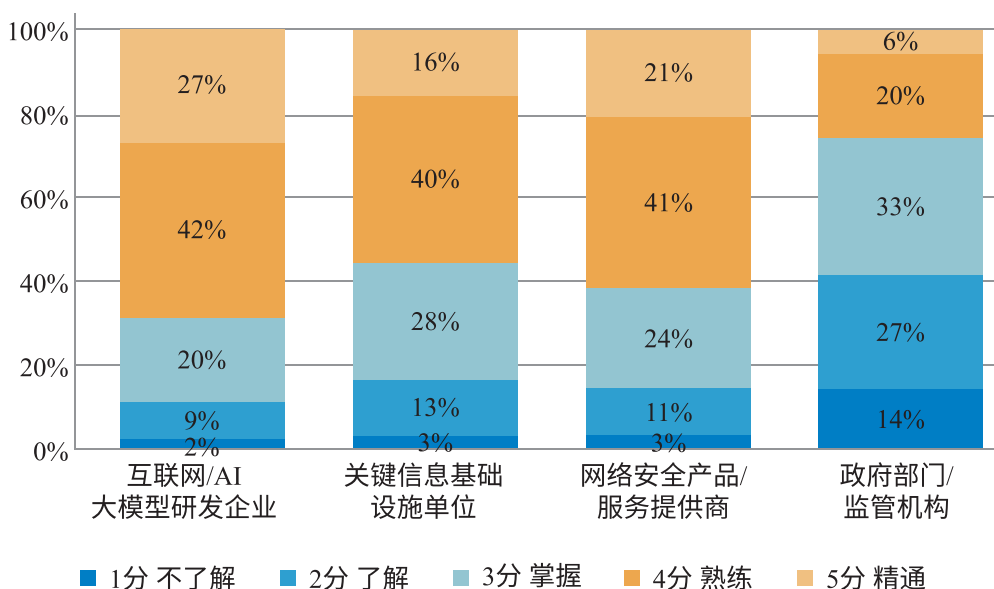


图3-13 不同单位接口安全保障能力水平

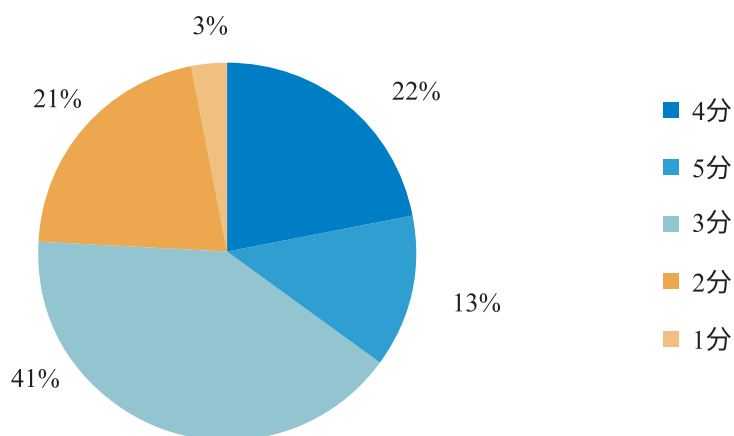


图3-14 业务逻辑保障能力水平

不同性质单位中，互联网/AI大模型研发企业和关键信息基础设施单位在业务流程、权限管理和审计留痕方面均具有较多实践基础，但AI应用带来的自动化调用和流程重组仍会放大业务逻辑安全风险。其中，互联网/AI大模型研发企业人才能力领先，4分和5分合计约58%，能够在业务流程、自动化调用与权限扩张风险识别上独立承担复杂任务；网络安全产品/服务提供商与关键信息基础设施单位表现接近，5分精通占比分别为16%和12%，3分掌握占比约三成；政府部门/监管机构的人才能力短板突出，1分不了解占18%、2分仅了解占30%，5分精通常仅占6%，是后续人才建设的重点方向。如图3-15所示。

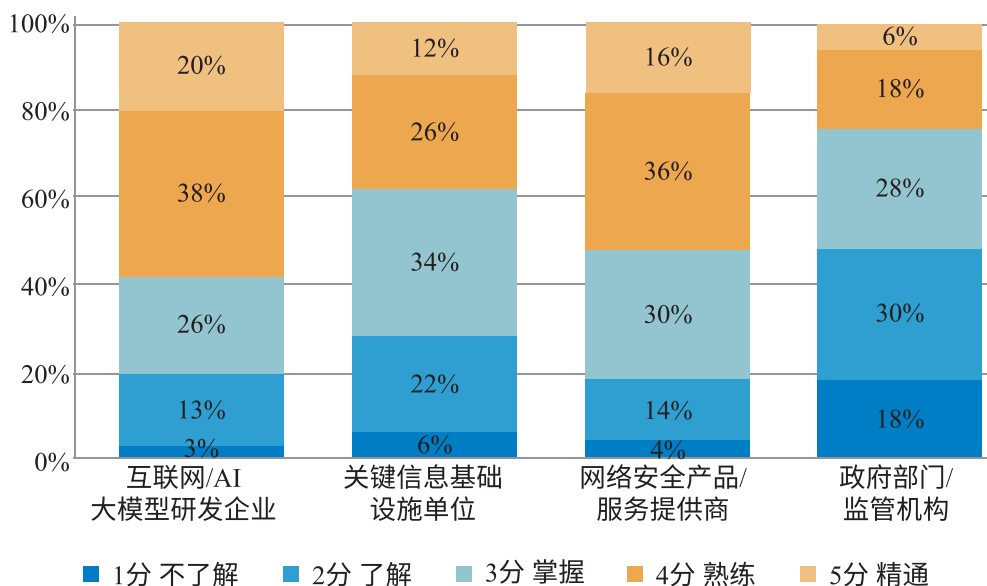


图3-15 不同单位业务逻辑保障能力水平

相比传统数据安全，AI应用数据安全还涉及提示词输入、上下文记忆、知识库权限和检索增强生成等新场景，对人才综合判断能力提出了更高要求。调研结果表明，受访网络安全人才的数据安全能力总体处于基本掌握水平，约15%的人才达到5分精通水平，约61%的人才处于中等及以下水平（1—3分），多数人员具备基础实操能力，但能够应对大模型场景下新型数据安全风险的高精尖人才仍存在缺口。如图3-16所示。

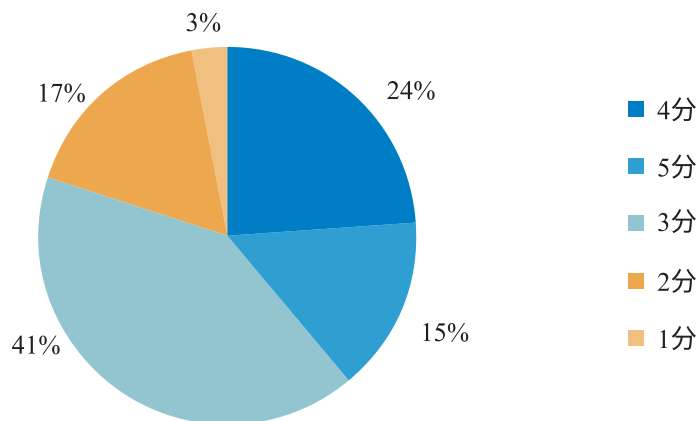


图3-16 数据安全保障能力水平

由于业务场景与数据防护需求不同，不同主体的网络安全人才的数据安全保障能力存在差异。互联网/AI大模型研发企业人才高分占比最高，4分和5分合计65%，能够在模型输入、检索增强、知识库权限等新场景下独立承担数据防护任务。网络安全产品/服务提供商与关键信息基础设施单位能力水平接近，4分和5分合计分别约61%和50%，但5分

精通占比仅21%和14%，高精尖人才储备仍有缺口。政府部门/监管机构人才精通层级占比偏低，1分不了解占14%、2分仅了解占28%，5分精通仅占6%，中等及以下水平人员占比突出，整体实操能力有待补强。如图3-17所示。

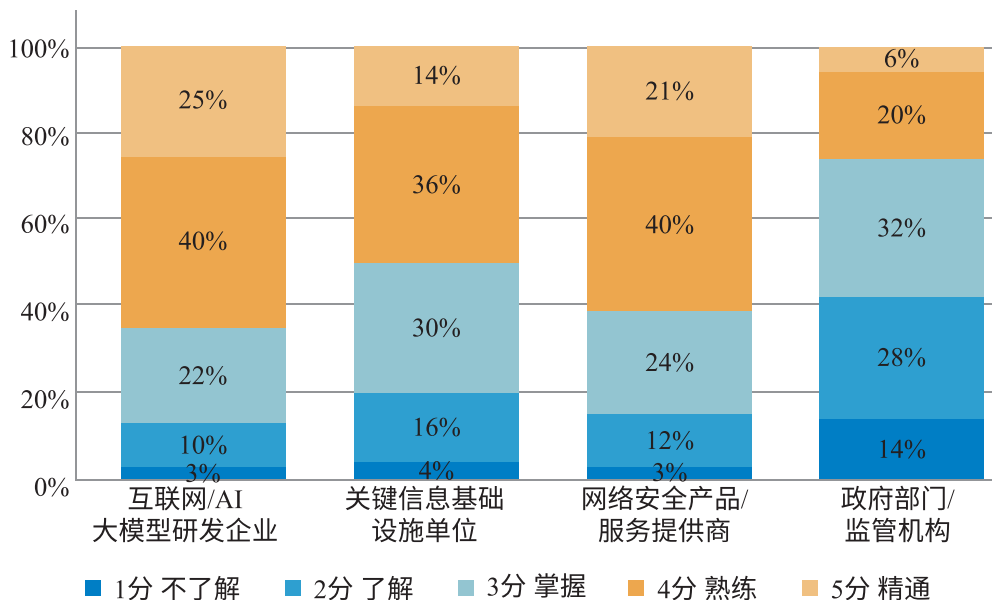


图3-17 不同单位数据安全保障能力水平

从调研结果来看，我国网络安全人才的AI应用安全保障能力整体处于基本掌握水平，五个维度的平均得分在3.21—3.44分之间，总平均约3.3分。其中，提示词安全能力相对最优，平均得分3.44分，5分精通占比约19%；接口安全与数据安全能力居中，平均得分分别为3.33分和3.31分；内容合规与业务逻辑保障能力相对最弱，平均得分分别为3.23分和3.21分。整体看，网络安全人才对模型输入输出类表层风险掌握较好，但应对智能体流程编排、自动化越权等复杂业务逻辑风险的高水平能力缺口显著，是当前人才建设的突出短板。

3.2.4 网络安全人才AI应用安全保障挑战分析

为进一步识别AI应用安全保障过程中的主要难点，本次调研对单位面临的最大挑战进行了统计。总体来看，“业务迭代极快，安全审计难介入”是AI应用安全保障过程中最突出的挑战，占比达到48%；其次是“缺乏针对AI应用的扫描工具”和“输出内容具有不可预测性”，占比分别为25%和18%；“难以在用户交互自由度与安全性间取舍”占比为9%。这表明当前AI应用安全保障面临的主要压力，一方面来自业务迭代快安全工作推进迟缓，另一方面也来自现有安全工具对AI应用场景的适配不足，以及模型输出存在一定不确定性。如图3-18所示。

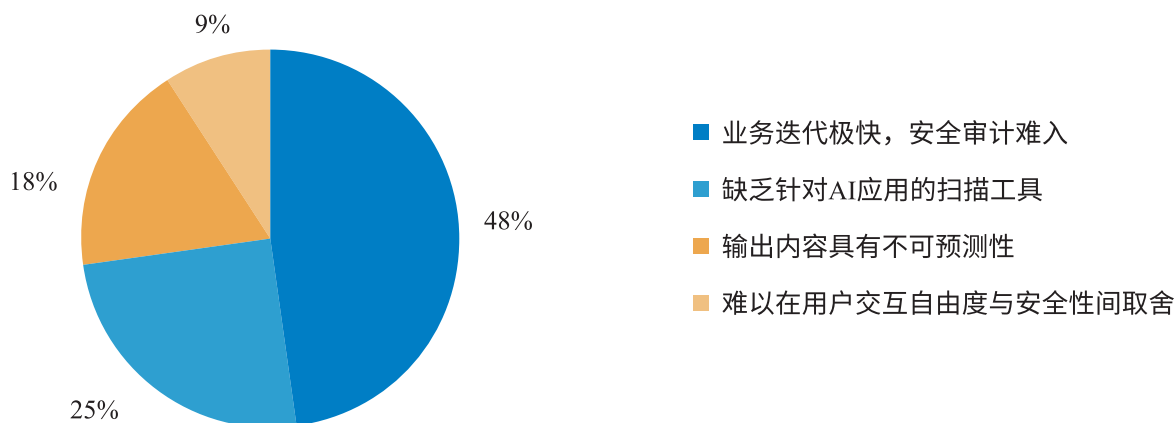


图3-18 AI应用安全保障挑战分析

不同性质单位呈现出不同特点。互联网/AI大模型研发企业更直接面对模型能力迭代、产品快速上线和用户交互复杂化带来的安全评估压力；网络安全产品/服务提供商既要保障自身AI应用安全，也要面向客户提供可复用的评估和防护能力；政府部门/监管机构中，“输出内容具有不可预测性”占比为36%，“难以在用户交互自由度与安全性间取舍”占比为27%，更体现监管侧对输出可控性、安全边界和结果追溯的关注；关键信息基础设施单位中，“缺乏针对AI应用的扫描工具”占比为29%，说明重点行业仍需要面向AI应用场景的检测、审计和持续评估工具。如图3-19所示。

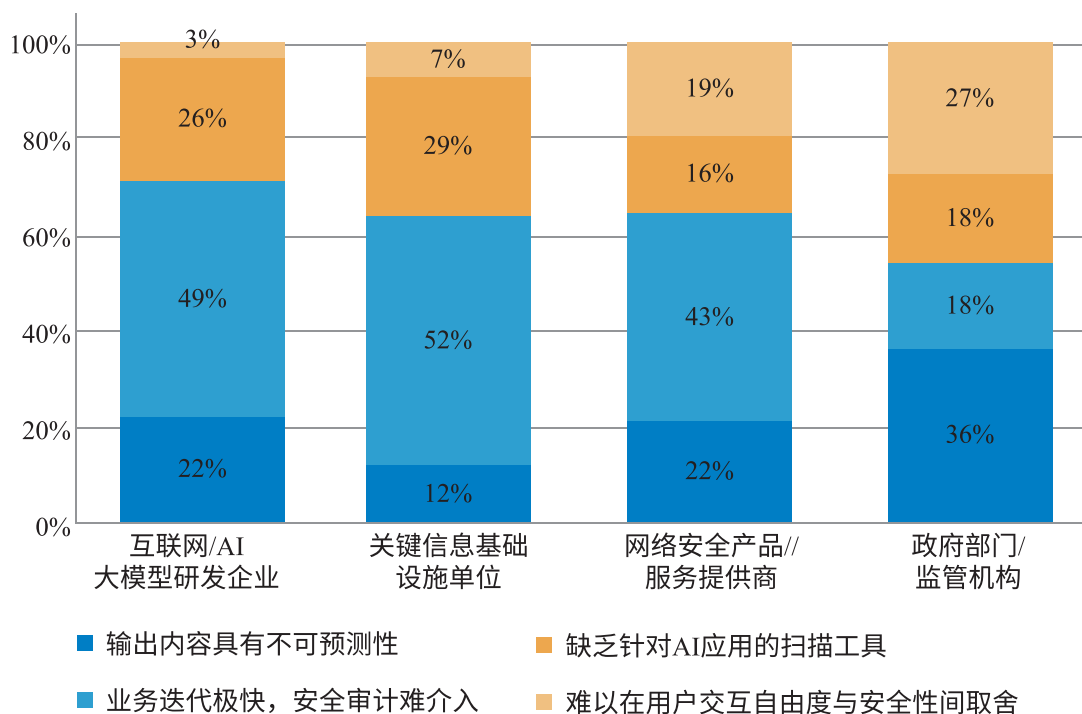


图3-19 不同单位的AI应用安全保障挑战分布

在关键信息基础设施单位内部，金融、电力、通信等重点行业的首要挑战均集中在“业务迭代极快，安全审计难介入”，其中电力为56%，金融为53%，通信为57%。同时，金融行业对“缺乏针对AI应用的扫描工具”的感受更为突出，占比为39%。这说明重点行业的AI应用安全压力主要来自业务上线节奏、审计介入时点和工具适配不足。整体来看，AI应用安全保障正从传统工具检测问题，转向快速业务迭代下的持续评估和治理问题。如图3-20所示。

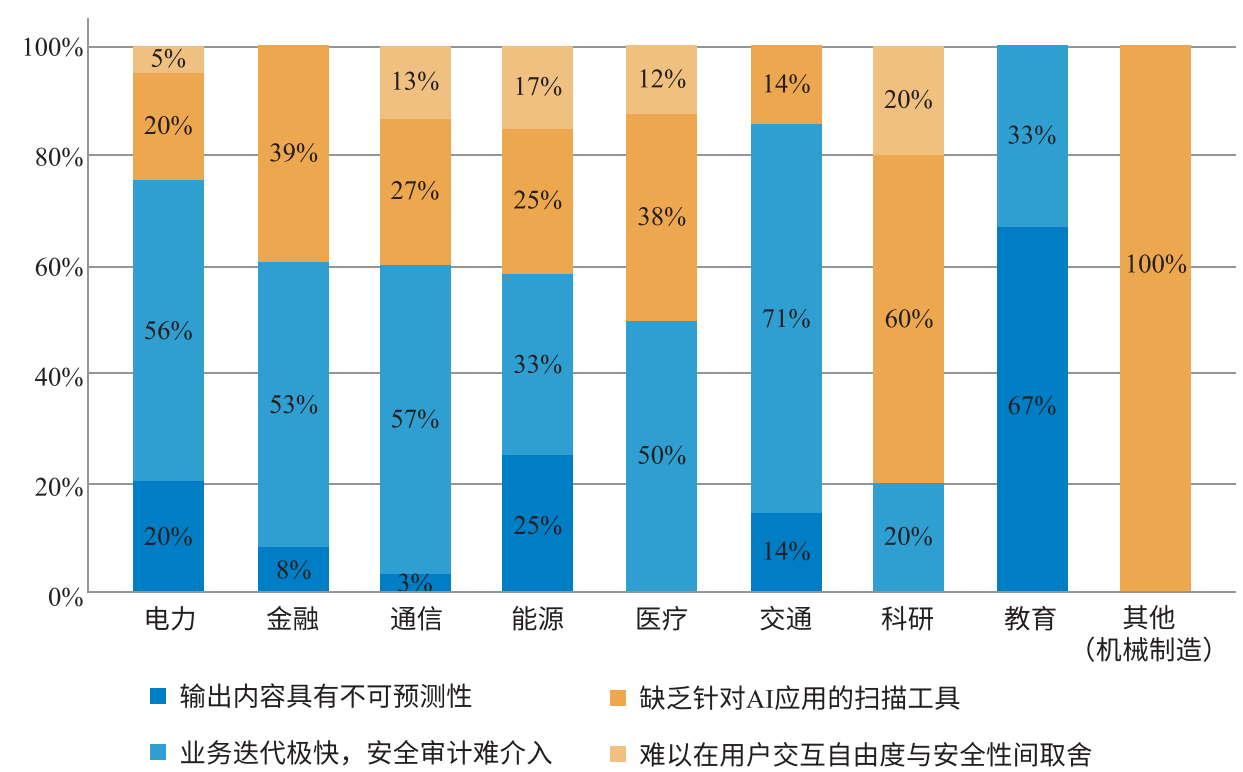


图3-20 关键信息基础设施单位AI应用安全挑战分布

3.3 网络安全人才应用安全能力ASK-P体系

面对AI应用安全风险的多维性与交叉性，网络安全人才的能力建设不能再沿袭以单一技术清单为主的传统模式，必须建立覆盖意识（Awareness）、技能（Skill）、知识（Knowledge）、实践（Practice）的ASK-P能力体系。

3.3.1 意识维度

意识是安全行动的内驱力，要求网络安全人才在接触AI应用时，主动触发风险感知并秉持正确的价值导向：

AI风险全景意识。理解AI安全非单纯技术问题，能够敏锐捕捉业务逻辑操纵、深度

伪造欺诈、算法歧视、合规违规等多元风险形态，形成“业务—数据—模型—伦理”一体化风险视野。

责任主体意识。明确自身在AI安全决策链中的角色，主动推进安全评审、模型验证、应急兜底等机制建设，拒绝“这是算法团队的事”等责任推脱心态，树立全生命周期安全共同负责的认知。

持续质疑与验证意识。对AI输出保持审慎，在关键业务环节坚持“AI辅助、人做决策”原则，养成对模型决策合理性、数据完整性、内容合规性的主动校验习惯。

3.3.2 技能维度

技能是将知识转化为有效安全动作的能力，强调标准化、可复用和工具化。

AI安全风险评估。能够独立开展AI应用安全评估，包括威胁建模、攻击面分析、模型鲁棒性基线测试、训练数据安全审查等，编写评估报告。

AI红队测试与渗透技能。掌握提示注入、越狱攻击、对抗样本生成、模型窃取、后门触发与验证等测试手法，熟练使用主流AI安全测试工具集与框架，在授权范围内实施可控的模拟攻击以暴露脆弱点。

安全监控与应急响应技能。能够针对AI应用部署输入异常检测、输出内容合规过滤等模型行为监控措施；制定并执行AI安全事件应急响应预案，包括紧急熔断、根因分析等。

跨团队沟通技能。能够以非技术语言向管理层传递AI安全风险，向算法团队明确安全需求，编制面向不同受众的AI安全指南和最佳实践。

3.3.3 知识维度

知识维度要求网络安全人才在原有知识结构上拓展AI相关领域，打通安全工程与AI原理的认知壁垒：

AI基础与安全原理。掌握机器学习、深度学习、大语言模型、多模态模型的基本原理及典型应用范式，理解训练、推理、微调、检索增强生成等核心环节。重点学习对抗样本、模型后门、模型逆向、成员推断、提示注入、幻觉利用等攻击机理，以及联邦学习、同态加密、差分隐私、模型水印等防护技术的适用范围与局限。

AI系统全栈安全架构。理解AI应用所依赖的数据管线、训练框架、模型服务引擎、前端交互接口的组件构成，熟悉API调用安全、容器与算力环境安全等关键知识点，能够绘制AI应用数据流与威胁面。

业务与行业知识。积累所在行业的业务逻辑、合规要点及典型AI应用场景，能够将安

全知识映射到具体业务风险，如金融风控模型的可解释性要求、医疗影像诊断的误诊容忍边界等。

3.3.4 实践维度

实践是检验和固化能力的关键环节，通过贴近真实业务的演练、测试、处置和复盘，将意识、知识与技能转化为可重复的任务交付能力。

靶场攻防演练实践。在模拟的智能客服、智慧金融等仿生环境中，扮演攻击方与防守方，反复练习从攻击构造到应急防御的全链条操作，积累不同场景下的攻击模式。

行业攻防竞赛与联合演练。参加跨行业AI安全挑战赛、深度伪造检测挑战赛、大模型红队演练，与各领域攻防人员切磋，了解最新攻击技术与防御思路。

企业实景实习。深入AI应用密集的业务部门或安全运营中心，跟随资深工程师完成模型安全评审、事件处置、合规整改等真实任务，将理论与实践在业务中完成衔接。

常态化复盘。建立个人及团队层面的实践复盘机制，将每次演练、每次事件处置提炼为可复用的检查单、案例库和规则集，推动经验向组织能力的转化。

3.3.5 人才分层

依据应用安全能力ASK-P体系，网络安全人才在AI安全方向的能力可划分为初级、中级、高级三个递进层级。三个层级的核心区别在于：初级侧重在指导下完成基础任务；中级应能够独立完成常见场景的评估、测试与处置；高级应能够闭环复杂任务，并形成工具、规则或方法论输出。各等级应分别在意识、技能、知识和实践四个维度设置可观察、可考核的行为指标。三级能力的对照关系如表3-1所示。

表 3-1 人才能力分层

层级	A（意识）	S（技能）	K（知识）	P（实践）
初级	知晓AI安全与传统安全的差异。具备发现AI应用安全风险的能力。	能在指导下使用已有工具完成验证和结果记录。	掌握AI安全基础知识，能描述常见的攻击与防护原理。	在靶场完成规定攻防动作，在指导下参与实际的辅助性任务。
中级	能敏锐预判各类AI应用安全风险。	技能扎实，能独立完成评估与红队测试。	深入理解攻击机理与防御机制，能将安全知识对应到具体业务风险。	有一定的实践历练，能独立交付任务，但经验深度仍需积累。
高级	能主动识别复杂风险，持续验证。	技能娴熟，可闭环交付复杂安全任务，能开发定制化工具，指导他人。	精通AI安全知识体系，能跨领域融会贯通，应对未知安全问题。	实战经验丰富，可独立应对真实复杂场景，能复盘提炼可复用的方法论和案例。

3.4 总结与建议

本章3.1节分析了AI应用安全风险，表明风险已由单一技术问题演变为数据、模型、接口、业务流程和组织责任交织的系统性风险。提示词注入和接口插件缺陷可突破系统边界，流程越权会将模型误判转化为实际业务损失，数据泄露、投毒和决策失真则可能侵害用户权益并削弱合规能力。3.2节能力现状调研显示，不同单位已经开始承担AI应用安全保障职责，但整体仍处于从“兼任式响应”向“体系化建设”过渡阶段；安全团队兼任仍是主流，提示词安全、接口安全和数据安全掌握程度较好，而内容合规和业务逻辑安全仍受场景复杂性、流程变化快和审计介入难等制约，人才来源复合化，需通过统一能力模型和实战任务推动传统网安、算法、业务人员与外部专家融合。简言之，我国网络安全人才在提示词安全、数据安全与接口安全方面能力较强，在内容合规与业务逻辑安全方面能力水平亟待提升，后续建设资源应向这两类短板能力倾斜。3.3节构建的ASK-P能力体系从意识、技能、知识和实践四个维度对AI应用安全能力进行系统刻画，并通过能力等级划分，为用人单位设置岗位能力要求、开展人才培养和能力评价提供框架，也为从业人员提升AI应用安全实战能力提供参考。

综上，当前AI应用安全人才建设已具备一定基础，但仍处于体系化能力形成的初期。后续建议从以下方面系统推进：

（一）明确岗位职责与跨部门协同机制

由用人单位管理层牵头，人力资源部门与网络安全、算法研发、业务部门联合落实。制定《AI应用安全岗位职责说明书》，将安全评估前置到应用设计、模型接入、数据调用和上线运营等全流程，划分安全负责人、安全工程师、算法工程师与业务负责人的具体职责；建立AI应用安全评审前置制度与跨部门联席决策机制，重点介入涉财、涉权、批量处理及专业咨询等场景。

（二）构建分角色、分层级的课程与实训任务

由教育培训机构、行业学会与用人单位企业大学联合开发，安全团队与业务专家提供内容支撑。以ASK-P能力体系为框架，针对管理者、传统网安人员、算法工程师、业务运营人员等不同角色，分别设计意识宣贯、基础技能、专业进阶和专家级四级课程。重点建设覆盖提示词注入、接口与插件滥用、数据泄露、越权流程和内容合规等典型风险的实训任务包，并融入数据治理、业务安全、合规风控和应急处置等复合实战内容。

（三）建立场景化能力认证和持续评估机制

由网络安全认证机构联合行业监管部门制定标准，用人单位人事部门实施动态管理。依据ASK-P体系定义应用安全能力等级，建立包含涉财交易、权限审批、批量处理、智能客服等高风险场景的认证测评案例库。实施“认证+复评+继续教育”的持续评估模式，要求持证人员每年完成规定学时的场景化实战训练。

（四）建设靶场、测试工具、案例库和审计平台

由网络安全主管单位或行业组织统筹，联合大型企业及第三方安全厂商共建共享。搭建支持提示词注入、插件漏洞利用、数据接口滥用、业务流程越权等场景的AI应用安全靶场；汇集脱敏后的真实安全事件案例和攻击剧本；建设面向AI应用的全量审计与实时风控平台，实现异常行为监测、风险熔断和审计留痕。

通过上述四项建议的协同推进，可逐步实现从职责厘清、能力锻造、评估认证到工具支撑的完整闭环，推动AI应用安全人才建设从被动响应转向体系化、前置化保障，使传统网络安全守护者真正转型为AI时代应用安全的保障者。



网络安全人才AI赋能能力分析

4.1 AI在网络安全领域的实践应用

AI在网络安全中的应用可以从业务场景、分析能力和技术治理三个层次理解：安全运营、漏洞挖掘与攻防对抗属于业务场景，安全数据分析是贯穿各场景的分析能力，大模型、智能体、多模态和知识图谱则构成技术演进与治理支撑。AI推动安全工作由人工经验驱动转向数据关联和人机协同，也要求网络安全人才具备数据分析、工具编排、结果验证和风险控制能力。

4.1.1 AI赋能安全运营

安全运营贯穿监测、分析、处置和复盘全过程，是网络安全实战能力的重要体现。传统安全运营依托平台管理日志、告警、资产和工单。随着云计算、移动办公、物联网和大模型应用普及，组织资产边界外延，攻击入口更加分散，安全事件呈现跨终端、跨身份和跨云资源特征。AI的引入，使安全运营由规则触发、人工研判为主，转向数据驱动、智能关联和人机协同。告警规模持续增长，误报和重复告警造成告警疲劳。安全分析员需要快速判断告警真实性、资产影响和升级必要性。安全运营的难点已经从“发现异常”转向“在大量异常中识别真正重要的威胁”。同时，自动化攻击和生成式AI提升了攻击效率，检测与响应窗口被压缩，告警分诊和响应决策亟须更高程度自动化。AI自身也成为治理对象，安全团队还需关注影子AI、敏感数据输入、插件滥用、提示词注入和生成内容风险。

1. AI驱动的安全监测与告警分析

AI驱动的安全监测与告警分析，主要是在SIEM、EDR、NDR、XDR、SOAR等平台基础上增加智能分析层，将分散数据转化为更可解释、更可排序、更可处置的安全事件。其作用包括告警降噪、风险排序、跨源上下文补全、研判结论生成和人工反馈持续改进。

在告警降噪方面，AI可综合告警类型、资产重要性、账号权限、行为时间、地理位置、

威胁情报命中、历史处置结果和业务上下文，对告警进行风险评分、聚类合并和优先级排序。重复或低风险告警可降级归并；涉及关键资产、高权限账号、恶意基础设施或攻击链关键节点的告警则优先处理。

在跨源关联方面，AI可围绕用户、设备、应用、数据和外部实体构建上下文，将身份、终端、网络、云平台和业务系统中的弱信号重组为事件时间线。例如，钓鱼邮件、异地登录、异常授权、文件下载和外联连接组合起来可能指向完整攻击链。AI还可辅助自然语言查询、事件摘要和检测规则初稿生成，降低KQL、SQL、YARA等工具的使用门槛。

AI告警分析的关键在于可解释和可审计，不能只给出“高风险”或“低风险”的黑箱判断，还应说明证据、依据和不确定性，并允许分析员纠错和反馈。告警分析完成后，系统还可沉淀误报原因、有效字段、检测规则和SOAR剧本，形成持续优化闭环。

2.AI驱动威胁检测与事件响应

在威胁检测方面，AI推动检测方式由单点规则命中向多源行为关联和攻击链识别扩展。传统规则适合发现已知恶意域名、固定漏洞载荷和典型命令行参数，但面对无文件攻击、账号滥用、云资源异常调用、供应链攻击和低频隐蔽行为时，单条规则难以及时判断风险。AI可结合终端进程、身份登录、网络连接、云审计日志和威胁情报，建立行为基线，识别异常行为。

在事件响应方面，AI可辅助事件分级、攻击路径还原、处置建议生成和复盘报告输出。面对疑似入侵事件，AI可整理告警时间线，标注可疑行为、受影响账号和主机、横向移动路径与外联地址，并结合预案生成隔离、禁用、封禁和修补建议。处置过程还可沉淀为结构化知识，反哺检测规则、威胁狩猎任务、SOAR剧本和人员培训。

AI参与事件响应必须设置清晰边界。隔离主机、禁用账号、删除文件、封禁流量等高影响操作不能完全交由模型自动执行，应建立最小权限、人工审批、操作留痕和回滚机制。未来安全人员不仅要会处置事件，还要能判断哪些步骤适合自动化、哪些步骤必须人工复核，以及如何验证AI给出的证据链。

3.安全运营中的人机协同流程

在告警分诊环节，AI可自动完成告警去重、上下文补全、风险排序和调查摘要，分析员重点核验相关证据、资产重要程度及潜在业务影响，并决定关闭告警、持续观察或升级处置。补充工单字段、发送通知、查询资产信息等规则明确、影响较低的操作，可由SOAR在授权范围内自动执行；隔离主机、禁用账号、封禁流量、删除文件等高影响动作，应由具备相应权限的人员审批后实施。

将上述环节串联起来，可形成“AI准备材料—人员作出决策—工具受控执行—联合开展复盘”的闭环。规则明确且影响较低的操作可自动执行；涉及账号、主机、流量和数据变更的动作须经授权审批，并保留证据、日志和回滚路径。

4.1.2 AI赋能漏洞挖掘与攻防对抗

漏洞挖掘与攻防对抗体现网络安全人才的核心实战能力。传统方法包括人工代码审计、静态分析、动态分析、Fuzz测试、符号执行和渗透测试。随着软件规模扩大、开源组件复用和云原生架构普及，漏洞发现与修复面临代码量大、上下文复杂、验证成本高等问题。大语言模型在代码理解、语义推理、生成式编程和工具调用方面的能力，为漏洞挖掘和攻防对抗提供了新的辅助路径。

当前AI已能在代码理解、风险筛选、测试用例生成和工具编排中发挥作用，但复杂业务逻辑分析、环境搭建、漏洞可利用性验证和误报控制仍需专业人员参与。因此，AI赋能漏洞挖掘更适合定位为专家主导、AI辅助、人机协同闭环。

1.AI辅助代码审计

AI辅助代码审计主要服务于大规模代码理解和潜在安全缺陷发现。大模型可解释函数、模块和调用链，帮助审计人员快速理解项目结构、识别潜在攻击面，并围绕输入可控性、危险函数调用、权限校验、边界检查和异常处理形成初步判断。AI与静态分析工具结合后，可读取告警路径和上下文，解释漏洞成因、评估触发条件，并生成初步修复建议。实践中仍应遵循“工具发现—AI解释—人工审查—自动化测试—安全回归验证”的流程，避免将AI生成代码直接投入生产环境。

2.AI辅助漏洞发现与定位

AI辅助漏洞发现与定位的核心价值，是利用模型对代码语义、漏洞模式和攻击路径的理解，在大规模代码库中筛选高风险区域，将分析范围收缩到重点入口、关键函数和可疑调用链。模型可结合CWE知识库、历史漏洞样本和补丁模式识别风险代码，并围绕输入可控性、数据可达性和边界检查进行推理。AI的价值不仅在于提示“哪里可能有问题”，也在于解释高危代码原因、判断误报可能性，并给出下一步验证路径。

3.AI辅助渗透测试与攻防演练

渗透测试和攻防演练涉及信息收集、漏洞验证、权限获取、横向移动和报告输出等环节。AI的主要作用是将复杂任务拆解为可执行步骤，并提供知识检索、命令参考和结果解释。当前大语言模型在工具使用、结果解释和行动建议等子任务上具有辅助效果，但长期任务规划、复杂环境适应和真实场景决策仍存在不足，更适合作为渗透测试副驾驶。未来

攻防人才需要具备AI工具调用、任务约束、结果验证和风险控制能力。

4. 漏洞分析中的人机协同流程

漏洞分析可采用“AI预审—专家复核—工具验证—修复确认”的流程：AI整理代码、扫描结果和补丁差异，标注高风险函数并生成验证路径；专家结合业务逻辑和部署环境确认根因、可利用性与影响范围，再通过静态分析、动态调试、Fuzz和受控环境完成交叉验证，最后审核补丁并开展安全回归。

该流程能够减少机械性代码阅读和重复测试，同时保留对复杂业务逻辑和真实风险的专业判断。

4.1.3 AI赋能安全数据分析

安全数据分析支撑监测、研判、预警和决策。组织内部数据类型持续增多，包括流量、终端、身份、应用、云审计日志，以及威胁情报、漏洞信息、恶意样本和工单记录等。传统分析依赖人工检索、规则匹配和专家经验，在海量异构数据、跨源关联和复杂攻击识别方面逐渐面临瓶颈。

AI推动安全数据分析由检索和规则统计，向语义理解、行为建模、关联推理和趋势预测发展。它可在网络流量分析、用户行为分析、威胁情报分析、恶意软件分析和态势感知等场景中提取关键线索，发现异常行为，并形成可解释的研判结论。

1. 网络流量与用户行为分析

网络流量分析长期承担攻击发现的数据支撑作用。面对加密通信、云服务访问和合法工具滥用，固定规则容易漏报或误报。AI可综合访问频率、连接时段、目标地域、域名特征和威胁情报，识别命令控制通信、异常数据传输和横向移动行为。复杂场景下，单条连接未必高危，但多个弱信号组合可能提示数据窃取或入侵扩散风险。

用户行为分析主要用于发现账号滥用、凭证泄露、内部威胁和权限异常。AI可围绕用户、设备、应用、数据和权限建立行为基线，分析登录地点、访问资源、权限调用和操作序列。当普通账号突然访问大量敏感文件、低权限账号短时间内尝试管理接口、长期不活跃账号异地登录并访问云资源时，系统可给出风险提示。此类分析应兼顾数据最小化、访问控制、审计留痕和隐私保护。

2. 威胁情报与恶意软件分析

威胁情报连接外部攻击态势与内部安全运营。AI可对情报报告进行实体抽取、去重归并、置信度评估和语义关联，将攻击组织、利用漏洞、恶意样本和处置建议转化为结构化知识。它还可将外部攻击技术与内部日志、资产暴露面和检测规则进行映射，辅助生成威

胁狩猎假设和排查查询。

恶意软件分析涉及样本分类、行为分析、家族识别、配置提取和攻击意图判断。AI可结合文件结构、导入函数、字符串、代码片段、沙箱行为和网络连接，生成样本风险摘要，解释混淆脚本、PowerShell 命令、宏代码和恶意配置，并辅助判断样本家族和攻击链阶段。但模型结论仍需与沙箱记录、人工逆向和威胁情报交叉验证。

3. 网络安全态势感知

态势感知面向组织整体安全状态、威胁趋势、资产风险和防护能力开展持续评估。传统平台多展示资产数量、漏洞数量、告警数量、攻击来源和风险评分，容易停留在静态看板层面。AI可对资产、漏洞、告警、威胁情报和处置结果综合建模，识别主要风险来源、攻击趋势和潜在扩散路径，并生成差异化摘要。总体来看，AI赋能安全数据分析的重点，是将分散、复杂、异构的数据转化为可理解、可关联、可验证的安全知识。

实际应用中，态势感知可形成分层输出机制。AI向一线分析人员提供可操作的线索、查询建议和事件关联结果，向安全管理人员汇总风险趋势、整改进度和资源使用情况，向组织决策层呈现关键业务影响及风险变化。各层级输出均需保留数据来源、分析过程和人工确认记录，确保态势结论能够追溯和复核。

4.1.4 AI驱动网络安全技术创新

AI在网络安全领域的应用正在从单点工具增强走向体系化技术创新。随着大语言模型、多模态模型、智能体、知识图谱和自动化编排技术发展，AI正在深入安全运营、漏洞挖掘、威胁狩猎、应急响应、安全开发和人才培养等环节，推动安全技术体系持续演进。

1. 大模型与安全智能体

大模型在网络安全中的价值，体现在自然语言理解、代码理解、知识推理和工具调用的结合。它能够理解日志、告警、代码、脚本和情报报告，并生成查询语句、分析摘要、检测规则、修复建议和处置方案。安全智能体更强调任务目标、工具调用、过程规划和结果反馈，可调用日志平台、情报库、检测系统、扫描器、沙箱和工单系统，完成调查、验证和报告输出。

2. 多模态与知识图谱融合

网络安全数据具有明显多模态特征，既包括结构化日志和告警，也包括代码、脚本、二进制样本、网络包、系统截图、钓鱼邮件、配置文件、漏洞公告、情报报告和工单记录。多模态AI可同时分析文本、代码、图像、表格和行为序列，突破单一日志或样本的限制。

知识图谱则可将资产、账号、漏洞、告警、攻击技术、恶意基础设施、处置动作和业务系统关系结构化表达，使安全分析更加可解释、可追溯和可验证。

3. 平台化发展与治理要求

AI原生安全产品正在推动安全能力由单点检测向智能分析、自动编排和持续学习扩展。平台化发展使AI能力由单点工具增强转向数据接入、知识治理、模型应用、工具编排和反馈优化的闭环。EDR、NDR、SIEM、SOAR、WAF和漏洞管理平台都在增加智能告警分析、自然语言查询、风险解释、自动化建议和报告生成能力。组织建设AI安全能力时，应以场景任务为牵引，而不是简单叠加产品功能。

AI驱动安全创新也带来新的风险。模型可能产生错误分析和不可靠建议，AI工具也可能接触漏洞细节、系统架构、攻击事件、账号权限和业务信息。安全应用必须明确哪些任务可以自动完成、哪些任务必须人工复核；对输入模型的数据进行分类分级和访问控制；要求AI结论提供证据来源、推理过程和不确定性说明；对智能体工具调用、处置建议、自动化动作和人工审批进行审计。

组织可根据任务风险和业务影响设置自动化边界。日志查询、格式转换、重复告警归并、情报摘要和报告初稿等低影响任务，可由AI自动完成并留存处理记录；检测规则生成、告警关闭、漏洞修复建议和事件定级等中等影响任务，应由安全人员抽查或确认；主机隔离、账号禁用、流量封禁、数据删除和权限变更等高影响任务，必须经过授权审批。通过分级授权、最小权限、全程留痕和异常回滚，可降低智能体在上下文不足或判断偏差情况下扩大业务影响的风险。

从人才培养看，AI时代的网络安全人才能力要求已超出传统攻防技术，还包括AI工具使用、模型监督、证据验证和风险治理。未来安全人才需要从单一工具使用者转向AI协同下的安全分析者、流程设计者和风险治理者，在技术创新与安全可控之间保持平衡。

4.2 网络安全人才AI赋能能力现状

为摸清我国网络安全从业人员在AI工具使用、人机协同能力以及不同组织间的能力差异，本章基于有效样本，从AI工具使用能力评估、从业人员人机协同能力自评、不同类型组织能力差异三个维度展开分析。国际层面，ISC2、ISACA、世界经济论坛等机构近年来围绕同类议题开展了大规模调查研究，本章在相关维度上将中国数据与国际发现进行对照，以提供更全面的判断参照。

4.2.1 AI安全工具使用能力分析

1.AI工具操作核心能力认知

AI安全工具操作环节中，结果校验、模型微调、提示词工程是三大核心技术模块，结果校验、模型微调、提示词工程和多工具集成是四项核心能力模块，各环节占比分别为：结果校验（人工纠偏与真伪判定）41%，模型微调（针对特定任务的模型微调）27%，Prompt Engineering（精准描述需求）22%，多工具集成与自动化流转10%。数据分布说明，从业者普遍认识到AI输出存在不确定性，人工校验与判断仍是AI工具使用链条中的核心环节。如图4-1所示。

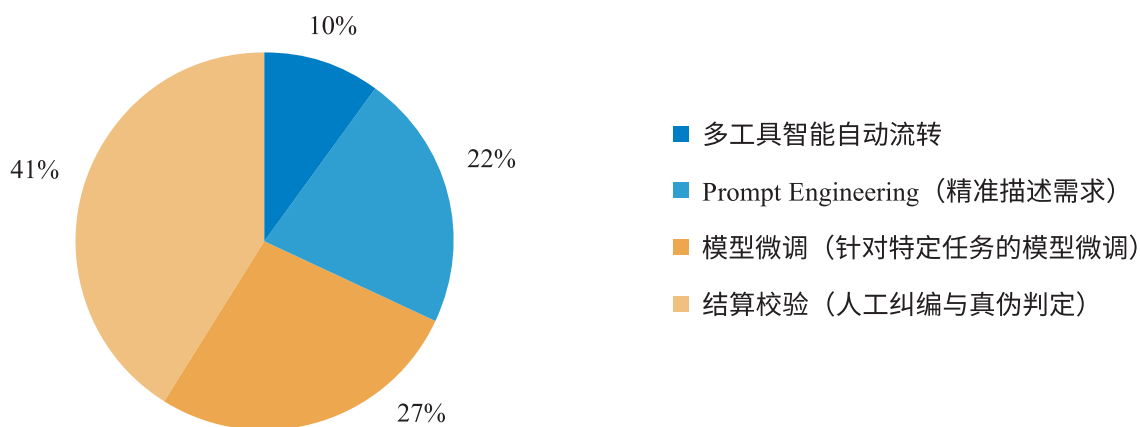


图4-1 AI安全工具操作核心能力模块

这一发现与ISC²的调研结论方向一致，《2025年ISC²网络安全劳动力研究报告》（2025 ISC2 Cybersecurity Workforce Study）显示仅28%的企业真正集成了AI安全工具，69%仍处于评估或测试阶段，全球范围内AI工具尚处于早期渗透期。

组织类型差异进一步丰富了上述结论。互联网与AI大模型研发企业对Prompt Engineering的重视度（32.7%）明显高于关键信息基础设施单位（16.6%），与其日常深度使用大模型的业务特征相符。安全产品与服务提供商在多工具集成与自动化流转方面的重视度（15.2%）领先于其他类型组织。政府部门对结果校验的重视度最高（54.5%），该类组织对输出准确性和合规性有更严格的要求。见表4-1。

2.AI工具在实际安全场景中的赋能效果

调研发现，AI工具已在安全运营的多个场景中产生实质性助益。从实际落地情况看，海量日志审计与告警降噪的选择比例最高，达58%；安全事件自动化响应次之，为48%；自动化漏洞挖掘与分析占35%；恶意代码行为分析与逆向占28%。人均选择约1.7项，说

表 4-1 不同组织类型 AI 工具核心能力认知对比

组织类型	结果校验	提示词工程	模型微调	多工具集成
互联网/AI大模型研发企业	46.2%	32.7%	14.4%	6.7%
网络安全产品/服务提供商	30.4%	20.3%	34.2%	15.2%
关键信息基础设施单位	41.4%	16.6%	32.4%	9.7%
政府部门/监管机构	54.5%	18.2%	18.2%	9.1%

明多数受访组织已在两个左右的安全场景中受益于AI工具。如图4-2所示。

日志审计与告警降噪的领先地位与当前安全运营自动化的需求高度吻合。该场景具有数据量大、规则明确、重复性高的特点，是AI工具天然的优势领域。相比之下，漏洞挖掘和恶意代码分析涉及更深层次的推理与判断，AI的渗透率仍有较大提升空间。从方向上看，这一场景分布与国际调查所呈现的运营自动化偏好相近，但场景定义和统计口径不同，以下仅作趋势参照：ISC² 2025年的调查显示，AI在网络安全领域影响最大的前两个场景分别为网络监控（60%）和端点保护（56%），均为运营自动化密集型场景。

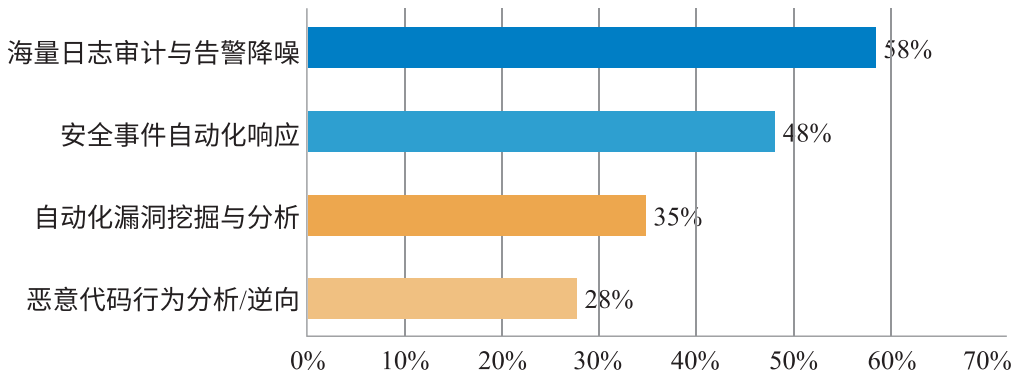


图4-2 AI工具在各安全场景中已产生实质性助益的覆盖率

3.AI辅助效率提升满意度评估

调研从漏洞挖掘、响应处置、合规文档编写、威胁狩猎与流量分析四个场景，以五级量表测量了从业者对AI辅助效率提升的满意度。数据表明，从业者对AI的整体满意度较高，四维度满意率分布在67%至78%之间。

漏洞挖掘效率的满意率最高，达到78%（满意49%，很满意29%）。合规文档编写以75%（满意44%，很满意31%）紧随其后，表明AI在标准化、重复性较高的任务中获得了较高的认可度。

威胁狩猎与流量分析满意率为68%（满意40%，很满意28%），响应处置速度为67%（满意41%，很满意26%），反映出从业者对AI在复杂分析、实时决策等场景中的效率提升认可度相对较低，仍有明显提升空间。四维度的不满意率均在6%以内。如图4-3所示。

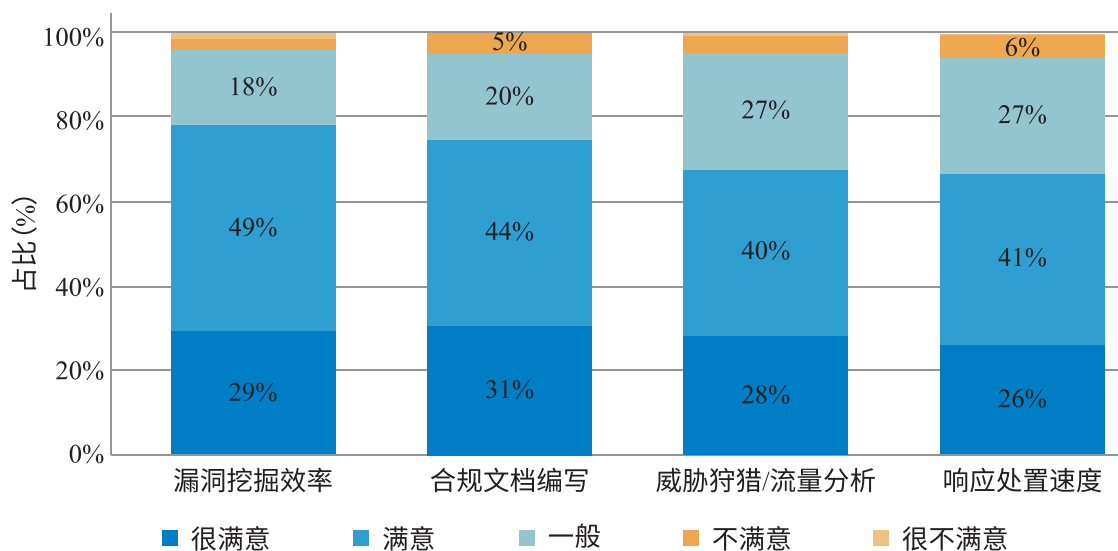


图4-3 AI辅助各安全任务效率提升满意度分布

与国际调查相比，中国调研显示的任务满意率整体较高，但两者的指标、样本和调查口径并不完全一致，以下仅作趋势参照。《2025年ISC²网络安全劳动力研究报告》中，63%的受访者报告AI提升了团队效能。

需要指出的是，2025年国际调研中AI工具的实际采用率较低（28%），大量企业仍处于评估与测试阶段，满意率的差异可能部分源于近一年来AI技术的快速发展，此外还可能与中国部分行业在AI工具部署上的先行先试有关。如图4-4所示。

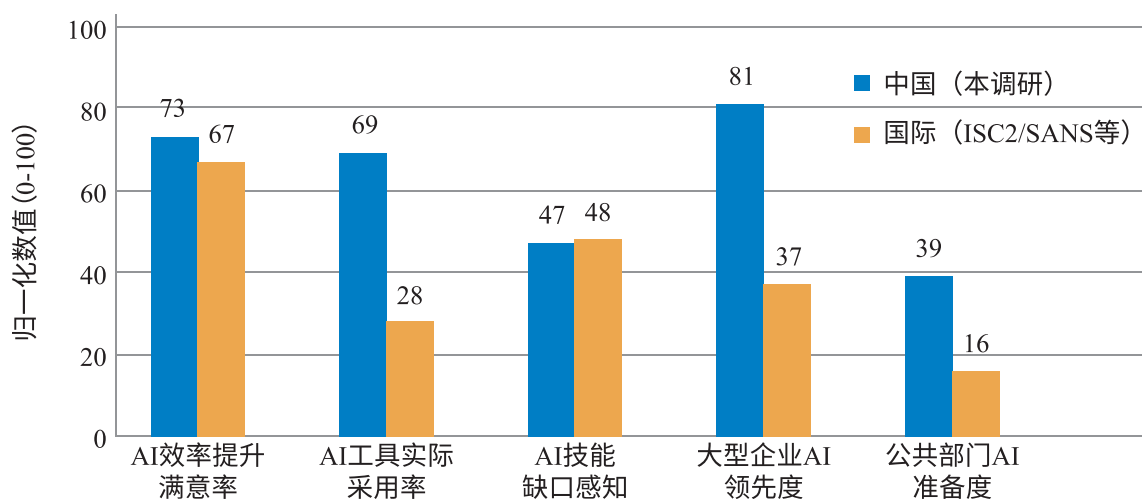


图4-4 中外AI安全人才能力调研核心指标对比

4.2.2 网络安全人才人机协同能力分析

1. 人机协同能力的内涵与构成

人机协同能力是指安全人员在明确任务目标与风险边界的前提下，对AI工具进行任务分配、过程监督、结果验证和经验沉淀的综合能力。它与AI工具使用能力既有联系又有区别：工具使用能力解决“会不会用”的问题，人机协同能力则解决“用得好不好、边界在哪里”的问题，是从单项工具应用迈向体系化AI赋能的关键环节。综合当前实践，人机协同能力可概括为四个层面：一是任务分解与委派能力，能够依据任务性质、风险等级和影响范围，将复杂安全任务拆分为可交给AI执行和必须由人工完成的子任务；二是过程监督与干预能力，能够在AI执行过程中识别偏离目标的行为并及时纠偏；三是结果验证与判断能力，能够对AI输出的证据来源、推理逻辑和结论可靠性作出独立核验；四是经验沉淀与反哺能力，能够将协同过程中形成的有效做法转化为规则、剧本和知识资产。

2. 人机协同实践的主要特征

从实践形态看，多数组织的人机协同正处于从“辅助问答”向“任务级协作”过渡的阶段：AI主要承担材料准备、信息汇总和初稿生成等环节，分析研判、风险定级和处置决策仍由安全人员完成；部分应用较深的组织已开始在告警分诊、漏洞预审等环节形成“AI准备、人工决策”的稳定分工。从业者的协同意识总体趋于理性，普遍认可AI输出须经人工校验后方可采用，但对AI的信任程度呈现两极分布：一端是过度信赖，将AI结论直接用于汇报和处置而缺少证据核验；另一端是过度审慎，仅将AI用于查询检索，任务委派与流程重构推进缓慢。同时，协同经验多停留在个人层面，有效的提示词设计、验证方法和操作技巧尚未普遍转化为团队规范与平台能力，制约人机协同从个体行为升级为组织能力。

3. 人机协同能力短板与发展方向

当前人机协同能力的短板集中在三个方面：任务分解缺乏方法依据，人机边界多凭个人经验划定，缺少基于任务风险和影响程度的分级标准；验证手段不足，面对AI生成的分析结论，部分人员难以快速核验其证据链与推理过程；责任界定模糊，AI参与研判和处置后出现偏差时，人与机器的责任划分缺乏明确规则，影响协同的深度和广度。从发展方向看，网络安全人才的角色正从工具使用者向任务编排者、质量监督者和风险控制者演进，人机协同能力建设重点相应从个人操作技能转向“任务设计—过程管控—结果验证—经验沉淀”的闭环能力。这一转变既依赖个人能力升级，也需要组织在自动化边界、审批流程和审计机制等方面提供制度支撑，相应的能力要求将在本章后续ASK-P能力体系中具体展开。

4.能力进阶的核心瓶颈分析

针对阻碍AI能力进阶的最主要瓶颈，调研结果显示：47.5%的受访者认为算力成本高昂、个人难以搭建实验环境，26.8%选择了跨学科门槛高、缺乏算法基础知识，18.6%选择了业务工作压力大、缺乏系统学习时间，7.1%选择了行业缺乏权威的AI安全能力水平认证。

算力成本以近半数的占比成为最首要应该突破的瓶颈，这一发现揭示了AI安全能力建设的普惠性挑战：算力成本已成为制约个人从业者与中小企业大规模技能提升的首要瓶颈，算力获取的可及性问题值得重点关注。跨学科门槛位列第二，表明安全从业者在向AI领域拓展时，算法与机器学习基础知识的缺乏构成了实质性的学习壁垒。如图4-5所示。

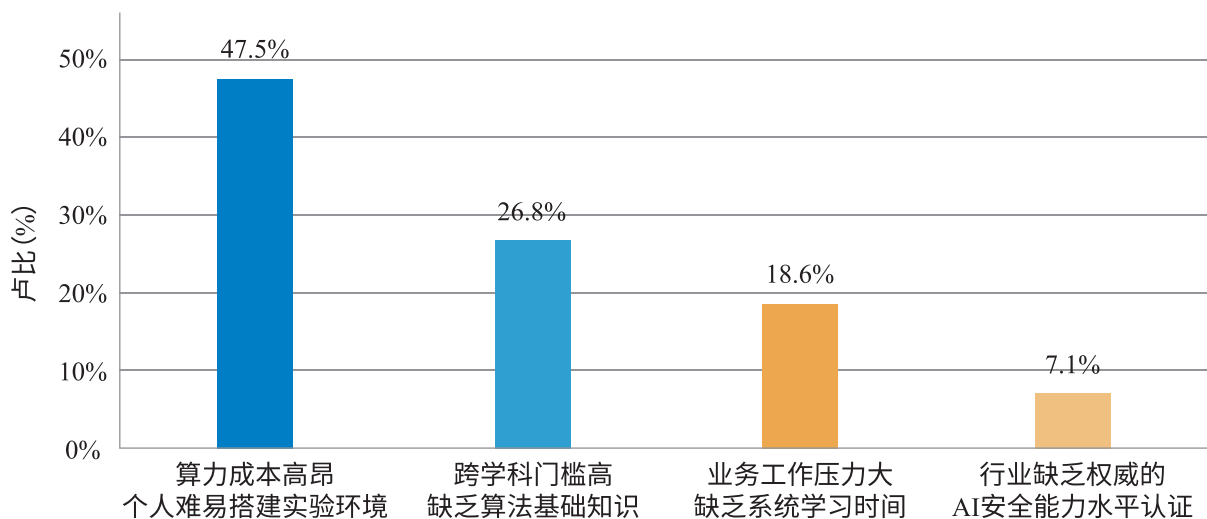


图4-5 AI安全能力进阶最主要瓶颈分布

在国际对标方面，Fortinet《2025年全球网络安全技能缺口报告》（2025 Global Cybersecurity Skills Gap）指出，48%的企业将AI与机器学习技能短缺列为首要问题。中国调研中以算力成本为首要瓶颈（47.5%），国际调研中以技能缺口为首要问题（48%），这一差异可能反映了中国GPU算力资源供给的结构性约束，也可能提示国际企业在AI技能培训方面面临同等严峻的挑战。两种障碍本质上具有互补性：算力是硬件基础，技能是人力基础，两者共同构成了AI安全能力建设的供给侧短板。

4.2.3 不同组织与人员特征的能力差异分析

1.组织类型维度的能力差异

本调研将受访组织划分为四类：互联网与AI大模型研发企业（占30.7%）、关键信息

基础设施单位（占42.8%）、网络安全产品与服务提供商（占23.3%）、政府部门与监管机构（占3.2%）。以下从AI工具使用认知、效率提升满意度、核心瓶颈、未来AI地位和资源投入五个层面展开对比（见表4-2、表4-3）。

表 4-2 不同类型组织 AI 安全能力多维度对比

对比维度	互联网 /AI 企业	安全厂商	关基单位	政府部门
Prompt Engineering 重视度	32.7%	20.3%	16.6%	18.2%
模型微调重视度	14.4%	34.2%	32.4%	18.2%
漏洞挖掘满意率	79.8%	79.7%	77.2%	54.5%
响应处置满意率	68.3%	73.4%	61.4%	72.7%
算力成本瓶颈	43.3%	41.8%	54.5%	36.4%
AI 为增效手段	51.9%	67.1%	55.9%	45.5%

表 4-3 不同类型组织 AI 赋能资源投入对比

资源类型	互联网 /AI 企业	安全厂商	关基单位	政府部门
商业大模型工具使用授权	56%	41%	45%	18%
行业 / 内部脱敏高质量语料库	51%	54%	57%	45%
私有化部署大模型 / 算力平台	38%	33%	19%	18%
专项 AI 实战能力技能培训	33%	38%	21%	45%
未提供任何支持	4%	9%	4%	9%

在AI工具使用认知上，互联网与AI企业对Prompt Engineering的重视度最高（32.7%），显著高于关基单位（16.6%），与其深度使用大模型的日常业务特征相符。关基单位和安全厂商对模型微调的重视度接近（32.4%和34.2%），均高于互联网与AI企业（14.4%），反映出这两类组织更看重AI工具在具体业务场景中的定制适配能力。

在效率提升满意度上，关基单位在全部四个维度的满意率均处于最低或次低水平，特别是在威胁狩猎（63.4%）方面，与互联网与AI企业及安全厂商（均为72%）的差距约9个百分点；响应处置（61.4%）与安全厂商（73.4%）的差距达12个百分点，与互联网与AI企业（68.3%）的差距约7个百分点。关基单位通常承担严格的合规要求和重保任务，对AI工

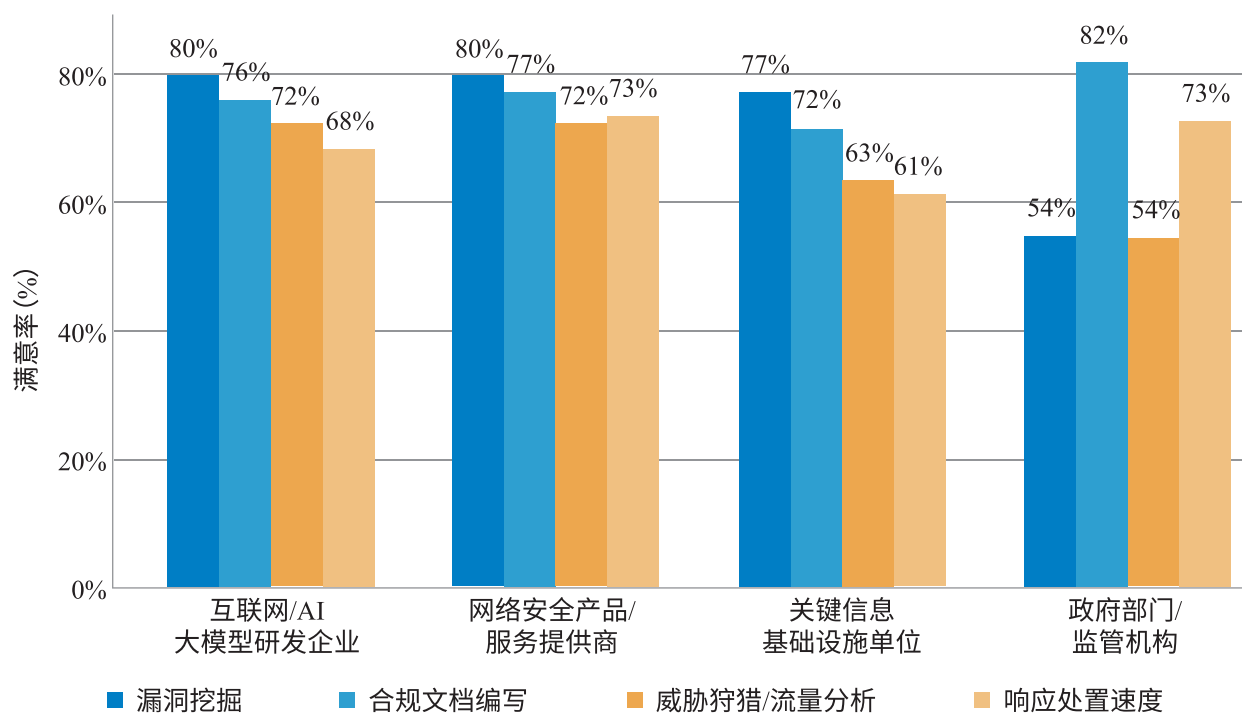


图4-6 不同组织类型AI辅助效率提升满意率对比

具的可靠性和可控性要求更高，这可能是其满意率偏低的一个重要原因。政府部门对合规文档编写较为满意，但在漏洞挖掘以及威胁狩猎上抱有更高的期望。如图4-6所示。

在核心瓶颈认知上，关基单位的算力成本瓶颈最为突出（54.5%），明显高于互联网与AI企业（43.3%）和安全厂商（41.8%）。政府部门的业务工作压力占比最高（45.5%），在四类组织中唯一以工作负荷为首要瓶颈。

在未来AI地位预期上，安全厂商认为AI是重要增效手段、仍需专家深度参与的占比最高（67.1%），关基单位次之（55.9%）。认为AI将占据核心主导地位、替代大部分基础人力的比例，以互联网与AI企业最高（31.7%），安全厂商最低（12.7%）。安全厂商作为AI安全工具的研发与交付方，对AI能力边界和局限性有最直接的体验，其相对审慎的判断具有参考价值。

在资源投入上，互联网与AI企业在商业大模型工具使用授权（56%）和私有化算力部署（38%）两个指标上领先，这与其较高的资源投入水平相一致。关基单位在行业与内部脱敏高质量语料库建设方面最为积极（57%），但私有化算力部署（19%）和专项技能培训（21%）两项投入偏低。政府部门在专项技能培训上的投入比例最高（45%），但算力资源最为匮乏（私有化部署仅18%）。

国际调查可提供企业规模和公共部门的外部参照，但其分类与本调研并不完全对应。

《2025年ISC2人工智能采用趋势调查》(2025 ISC2 AI Adoption Pulse Survey)显示,万人以上大型企业AI工具采用率为37%,而百人以下微型企业仅为20%;政府部门整体采用率最低(16%)。《2025年全球网络安全展望》(Global Cybersecurity Outlook 2025)指出,49%的公共部门缺乏足够人才实现网络安全目标。这些结果可作为外部趋势参照,但由于分类、样本和指标口径不同,不宜直接与本调研进行数值比较。

2. 企业规模维度的能力差异

以漏洞挖掘效率满意率为观测指标,以漏洞挖掘效率满意率为观测指标,不同规模企业呈现出明显的组间差异(见图4-7a):500至2000人企业满意率最高(85.7%),2000人以上企业次之(81.1%),100人以下企业为76.8%,100至500人企业最低(74.8%),最高与最低组相差10.9个百分点。其中100至500人企业既不具备大型企业的算力预算优势,也缺乏小微企业的决策灵活性,处于能力建设的相对薄弱地带。

3. 从业年限维度的能力差异

按从业年限分组,漏洞挖掘效率满意率呈现5至10年组最高(83.3%)、10年以上组最低(58.3%)、3年以下组(78.0%)和3至5年组(77.6%)居中的分布(见图4-7b)。10年以上资深从业者的满意率(58.3%)显著低于其他组别,与满意率最高的5至10年组(83.3%)差距达25.0个百分点,与3年以下(78.0%)和3至5年(77.6%)组相比也低约20个百分点。这一现象可能源于两方面因素:资深人员对AI工具能力边界的判断标准更为严格,在长期实践中形成了对工具可靠性的更高期待;同时,资深群体在跨学科转型过程中可能面临更大的学习成本与适应阻力。

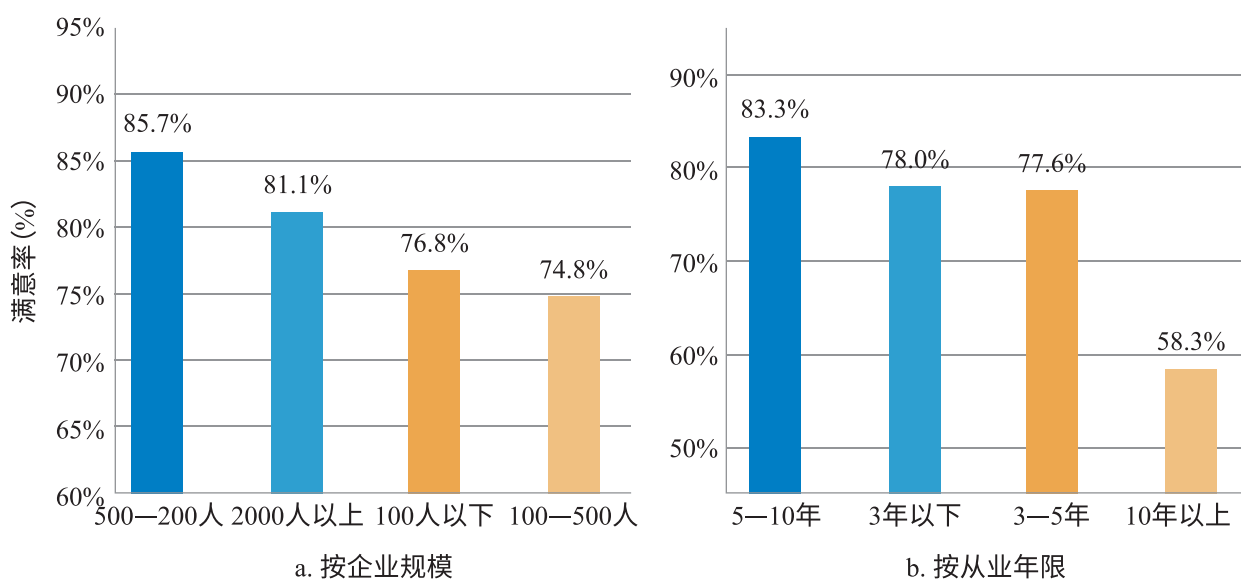


图4-7 漏洞挖掘效率满意率:企业规模与从业年限维度对比

4.3 网络安全人才AI能力ASK-P体系

本章节选取低空、区块链、人才训练、工业控制、数据跨境和AI系统安全六个具有代表性的应用对象，用于展示ASK-P在不同风险类型和业务对象中的适配方式；六个场景不构成完整场景清单，后续可按同一框架扩展。为便于横向比较，各场景统一从A（意识）、S（技能）、K（知识）、P（实践）四个维度描述，并根据场景风险设置评价重点。表中权重用于示例说明，具体应用前还需结合试测和专家校准。

4.3.1 AI辅助低空漏洞挖掘和安全攻防场景

AI辅助低空漏洞挖掘和安全攻防场景，主要面向无人机、eVTOL、低空飞行器、地面站等相关业务系统的安全评估、漏洞挖掘和攻防演练。在该场景中，AI可以辅助完成低空资产识别、攻击面梳理、漏洞线索发现、协议与接口分析、代码审计、固件分析、测试用例生成、攻击路径推演、风险报告撰写和修复建议生成等工作。AI的引入能够提升漏洞挖掘效率和攻防分析深度，但其输出结果仍存在不确定性，不能直接替代安全人员的专业判断。

在低空安全人才培养方面，部分高校已开展面向智能无人系统安全的课程建设与实践探索。西安电子科技大学构建《智能无人系统》课程，面向网络空间安全、低空技术与工程等相关专业，聚焦低空经济、智能无人系统和国家安全领域的发展需求，围绕无人机系统安全这一核心方向，融合飞行控制、导航定位、无线通信、人工智能、网络安全、密码学和可信计算等多学科知识，构建“基础知识—核心原理—安全分析—工程实践”四层递进式课程体系。课程以无人机控制原理为基础，将飞控安全、通信安全、可信启动、安全通信、集群控制、异常检测等关键技术有机融合，依托自主研发的教材、半实物仿真平台和真实实验平台，采用“理论教学、仿真验证、攻防演练、科研创新”相结合的教学模式，培养学生在智能无人系统安全分析、攻防测试、安全防护和工程设计等方面的综合能力，为低空经济发展、无人智能作战和智能无人系统产业培养具有创新能力和工程实践能力的高层次复合型人才。

表4-4展示了AI辅助低空漏洞挖掘和安全攻防场景人才AI能力的评价维度、评价方法以及分层级权重划分。

评价维度。本场景对人才的评价过程中，应重点考察人才是否能够理解低空业务系统的安全边界，是否能够利用AI辅助提升漏洞挖掘和攻防分析效率，是否能够对AI输出进行人工复核，是否能够形成可验证、可复盘、可交付的安全成果。

表 4-4 AI 辅助低空漏洞挖掘和安全攻防场景人才 AI 能力评价权重表

评价维度		A（意识）	S（技能）	K（知识）	P（实践）
内涵		AI辅助低空漏洞挖掘和安全攻防场景网安人才应具备的职业操守、安全意识、低空安全责任意识、授权测试意识、公共安全意识、数据合规意识和AI使用边界意识等	AI辅助低空漏洞挖掘和安全攻防场景网安人才应掌握的AI辅助资产识别、攻击面梳理、漏洞挖掘、协议分析、代码审计、固件分析、测试脚本生成和攻防演练分析等技能	AI辅助低空漏洞挖掘和安全攻防场景网安人才应具备的网络安全基础、漏洞原理、渗透测试方法、AI基础知识，以及低空系统架构、通信链路、飞控系统、云边端协同和相关标准规范知识	AI辅助低空漏洞挖掘和安全攻防场景网安人才应具备的低空系统安全评估、漏洞验证、攻防演练、风险复盘、修复验证、工具链建设和产品化交付能力
评价方法		笔试+答辩	场景实操/机考/证书	客观+主观/笔试	成果材料/答辩评审
岗位等级		评价权重			
初级	1	7%	22%	66%	5%
	2	7%	22%	61%	10%
	3	7%	23%	55%	15%
中级	4	10%	23%	47%	20%
	5	10%	25%	30%	35%
	6	10%	28%	22%	40%
高级	7	10%	28%	17%	45%
	8	10%	30%	10%	50%
	9	10%	32%	8%	50%

评价方法。本场景的评价方法采用笔试、答辩、场景实操、机考、证书核验、成果材料评审等方式综合开展。A（意识）维度主要通过案例笔试和现场答辩，考察人才是否具备低空安全责任意识、授权测试意识、公共安全意识、数据合规意识和AI工具审慎使用意识；S（技能）维度主要通过仿真靶场、场景实操或机考，考察人才能否利用AI辅助完成资产梳理、攻击面分析、漏洞线索发现、测试脚本生成、漏洞验证和攻防路径推演；K（知识）维度主要通过客观题与主观题结合的笔试方式，考察人才对网络安全基础、漏洞原理、低空系统架构、通信链路、飞控系统和AI基础知识的掌握情况；P（实践）维度主要通过项目材料、测试报告、工具成果和答辩评审，考察人才是否能够在授权环境下完成低空安全评估、攻防演练、风险复盘、修复建议和工程化交付。

分层级评价权重划分。从权重设置看，初级岗位以K（知识）为主，强调人才对低空系统、安全基础和AI基础能力的理解；中级岗位逐步提高S（技能）和P（实践）比重，强调独立测试、场景分析和任务交付能力；高级岗位以P（实践）为核心，同时保持较高的S（技能）权重，突出复杂场景攻防设计、工程化建设和产品化成果输出能力。

4.3.2 AI辅助区块链安全运营与告警研判场景

AI辅助区块链安全运营与告警研判场景，主要面向公链、联盟链、智能合约、区块链应用平台的安全运营工作。在该场景中，AI可以辅助完成告警聚合、异常交易识别、链上行为分析、智能合约风险摘要、研判报告编写等任务。通过AI对链上数据与日志信息进行关联分析，可以提升告警研判效率，降低重复性分析成本，并帮助安全运营人员更快识别真实风险。

AI生成的研判结论不能直接作为最终处置依据，重要处置动作必须经过人工复核和流程审批。表4-5展示了AI辅助区块链安全运营与告警研判场景人才AI能力的评价维度、评价方法以及分层级权重划分。

评价维度。本场景对人才的评价过程中，应重点考察人才是否能够利用AI提升告警研判效率，是否能够结合链上链下证据形成准确判断，是否能够识别AI误判、漏判和不当处置风险，是否能够形成持续优化的规则库、案例库和知识库。

表 4-5 AI 辅助区块链安全运营与告警研判场景人才 AI 能力权重评价表

评价维度		A（意识）	S（技能）	K（知识）	P（实践）
内涵		AI辅助区块链安全运营与告警研判场景网安人才在使用AI时应具备的职业操守、安全意识、资产安全意识、合规运营意识、证据保全意识、隐私保护意识和AI审慎使用意识等	AI辅助区块链安全运营与告警研判场景网安人才应掌握的AI辅助告警研判、链上交易分析、日志关联、地址画像、资金流追踪、智能合约风险识别、检测规则编写和自动化处置编排等技能	AI辅助区块链安全运营与告警研判场景网安人才应具备的区块链基础原理、共识机制、智能合约、密码学基础、钱包与密钥管理、DeFi 风险、传统安全运营和AI基础知识	AI辅助区块链安全运营与告警研判场景网安人才应具备的告警分级、事件研判、链上追踪、应急响应、复盘分析、规则优化、知识库建设和运营体系改进能力
评价方法		笔试+答辩	场景实操/机考/证书	客观+主观/笔试	成果材料/答辩评审
岗位等级		评价权重			
初级	1	8%	18%	64%	10%
	2	8%	20%	57%	15%
	3	12%	20%	48%	20%
中级	4	12%	22%	33%	33%
	5	12%	25%	23%	40%
	6	12%	25%	18%	45%
高级	7	12%	25%	13%	50%
	8	12%	28%	10%	50%
	9	8%	18%	64%	10%

评价方法。本场景的评价方法采用知识测评、案例答辩、告警研判实操和运营成果评审相结合的方式。A（意识）维度主要通过笔试和答辩，考察人才是否具备资产安全意识、合规运营意识、证据保全意识、隐私保护意识和AI研判边界意识；（技能）维度主要通过场景化告警研判、链上交易追踪、日志关联分析和自动化处置演练，考察人才能否利用AI辅助完成告警聚合、异常交易识别、地址画像分析、智能合约风险判断、威胁情报匹配和研判报告生成；K（知识）维度主要通过客观题、主观题和案例分析，考察人才对区块链基本原理、共识机制、智能合约、钱包与密钥管理、DeFi风险、安全运营流程和AI基础知识的理解；P（实践）维度主要通过研判报告、处置记录、复盘材料、检测规则、知识库和自动化剧本等成果进行评审，考察人才是否具备事件闭环处置、规则优化、运营改进和平台化建设能力。

分层级评价权重划分。从权重设置看，初级岗位强调K（知识）积累，确保人员理解区块链安全运营基本对象、基本流程和AI工具使用规范；中级岗位提高P（实践）权重，强调告警研判、事件处置和链上链下关联分析能力；高级岗位突出复杂事件指挥、运营体系建设、AI规则模型优化和平台化交付能力。

4.3.3 AI辅助网络安全人才训练与靶场内容生成场景

AI辅助网络安全人才训练与靶场内容生成场景，主要面向网络安全课程建设、实训任务设计、攻防靶场搭建、竞赛题目生成等工作。在该场景中，AI可以辅助生成课程大纲、训练任务、靶场配置、评分规则、标准答案和评价报告。本场景对人才提出了“懂安全、懂教学、懂靶场、懂AI、懂评价”的复合能力要求。

在高校课程建设方面，部分院校已围绕大模型系统原理、实践应用和安全审查开展体系化教学探索。西安电子科技大学《大模型原理与实践》课程面向以本科生为主、兼顾研究生的不同专业学生，聚焦生成式人工智能与大模型技术快速发展背景下的人才培养需求，围绕“大模型系统如何处理一次真实请求”这一核心主线，融合自然语言处理、知识检索、工具调用、智能体技术与人工智能安全等相关知识，构建“系统认知—原理理解—实践验证—安全审查”四层递进式课程体系。课程以大模型请求处理生命周期为基础，将Transformer、内容生成与采样、检索增强生成、工具调用、多轮交互及智能体安全等关键内容有机贯通，依托西电NSS团队自主研发的玄知密码大模型、课程讲义和Teaching Lab实验平台，采用“理论讲授、案例演示、对比实验、证据分析”相结合的教学模式，引导学生从真实系统现象出发理解大模型工作机制，培养学生在大模型系统分析、配置应用、实验验证、行为解释与安全审查等方面的综合能力，为生成式人工智能及相关领域培养兼具

系统思维、实践能力、安全意识和工程责任的复合型人才。该课程将系统认知、专业知识、实验技能和安全实践贯通，为大模型相关人才的场景化培养与综合能力评价提供了实践参考。

表4-6展示了AI辅助网络安全人才训练与靶场内容生成场景人才AI能力的评价维度、评价方法以及分层级权重划分。

评价维度。本场景对人才的评价过程中，应重点考察人才是否能够利用AI辅助完成训练内容生成和靶场场景构建，是否能够对AI生成内容进行人工审核，是否能够保证训练任务与岗位能力目标一致，是否能够形成可复用、可迭代、可运营的靶场内容体系。

评价方法。本场景的评价方法采用方案评审、实操考核、内容审核和成果答辩相结合的综合评价方式。A（意识）维度主要通过笔试和答辩，考察人才是否具备教育责任意识、内容合规意识、AI安全使用意识、数据脱敏意识和公平评价意识；S（技能）维度主要通过训练任务生成、靶场环境编排、攻防剧本设计、日志样本生成和自动评分配置等实操任务，考察人才能否利用AI辅助完成课程设计、题目生成、环境配置、评分规则编写、学习

表 4-6 AI 辅助网络安全人才训练与靶场内容生成场景人才 AI 能力评价权重表

评价维度		A（意识）	S（技能）	K（知识）	P（实践）
内涵		AI辅助网络安全人才训练与靶场内容生成场景网安人才应具备的职业操守、安全意识、教育责任意识、内容合规意识、AI安全使用意识、数据保护意识和公平评价意识等	AI辅助网络安全人才训练与靶场内容生成场景网安人才应掌握的AI辅助课程设计、题目生成、靶场编排、日志生成、攻防剧本设计、自动评分、学习反馈和能力画像分析等技能	AI辅助网络安全人才训练与靶场内容生成场景网安人才应具备的网络安全知识体系、攻防基础、靶场架构、虚拟化与容器技术、AI基础、教学设计、人才评价方法和相关规范知识	AI辅助网络安全人才训练与靶场内容生成场景网安人才应具备的训练项目交付、靶场内容库建设、评分规则配置、训练运营、能力画像、复盘改进和平台化建设能力
评价方法		笔试+答辩	场景实操/机考/证书	客观+主观/笔试	成果材料/答辩评审
岗位等级		评价权重			
初级	1	7%	18%	70%	5%
	2	7%	18%	65%	10%
	3	7%	18%	60%	15%
中级	4	10%	18%	52%	20%
	5	10%	20%	38%	32%
	6	10%	22%	28%	40%
高级	7	10%	22%	23%	45%
	8	10%	23%	17%	50%
	9	10%	25%	15%	50%

反馈生成和能力画像分析；K（知识）维度主要通过客观题、主观题和课程设计方案，考察人才对网络安全知识体系、靶场技术、攻防基础、AI基础、教学设计方法和人才评价理论的掌握情况；P（实践）维度主要通过课程包、题目包、靶场环境、评分规则、训练复盘报告、能力画像报告和平台建设成果进行评审，考察人才是否能够完成完整训练项目交付，并形成可复用、可迭代、可运营的靶场内容体系。

分层级评价权重划分。从权重设置看，初级岗位以K（知识）为主，重点保证训练内容制作人员具备必要的安全基础、靶场基础和规范意识；中级岗位提升S（技能）和P（实践）比重，强调训练任务设计、靶场搭建、评分配置和训练复盘能力；高级岗位以P（实践）为核心，突出课程体系规划、内容库建设、评价体系设计、平台化建设和规模化交付能力。

4.3.4 AI驱动的工业控制系统安全

伴随全球制造业向智能化、网络化加速转型，工业控制系统作为国家关键信息基础设施的核心组成部分，其安全稳定运行直接关乎国计民生。“两化融合”的推进使大量工控系统接入企业内网乃至互联网，暴露范围急剧扩大，高级持续性威胁、勒索软件等新型攻击已对生产连续性构成严重威胁。在此背景下，人工智能技术被引入工控安全领域，以期通过异常行为检测、预测性维护与智能溯源提升防御能力。但工控环境特有的高实时性、高可靠性以及对物理世界的直接影响，为AI应用带来了严峻挑战。本节基于ASK-P框架，分析AI驱动的工控安全场景下网络安全人才所需具备的核心能力，见表4-7所示。

表 4-7 AI 驱动的工业控制系统安全能力评价表

ASK-P 维度	核心能力要求	场景侧重
意识	树立“零事故”导向，明确AI模型在OT环境中的局限性，警惕因过度追求精确率牺牲召回率引发的风险；秉持“最小干预”原则，限定AI以提供预警建议为核心功能，避免未经验证的自动化执行引发连锁故障。	平衡“安全”与“生产”关系，任何安全举措均不得牺牲业务连续性。
知识	掌握Modbus、Profinet、IEC 61850等工控协议的原理；熟悉IEC 62443等威胁模型；了解LSTM、Transformer、孤立森林、自编码器等人工智能应用原理；知晓《网络安全法》《数据安全法》等相关法规。	构建“OT+IT+AI”三栖知识架构。
技能	掌握工控数据特征工程方法；熟练运用SHAP、LIME等可解释性AI工具；掌握TinyML等轻量化部署技术；具备在仿真环境中复现攻击行为、验证AI防御效果的能力。	聚焦“可解释性”与“低侵入式”建模。
实践	主导开展基于数字孪生的攻防演练；搭建IT、OT、AI三方常态化协作机制；将误报率、漏报率、MTTD、MTTR纳入绩效考核范畴。	重点聚焦于“仿真演练”与“跨部门协作”。

在意识层面，从业者需突破传统信息安全范畴，树立“安全与生产并重”的底线思维。在工控环境中，一次误报（如将正常振动误判为攻击并触发停机）造成的损失可能高于一次漏报。因此，从业者须对任何可能中断生产的AI输出保持高度审慎，清醒认知AI模型在处理含噪声、非平稳OT物理信号时的天然局限，坚守“最小干预”原则，AI应定位为辅助决策工具而非自动执行者，所有自动化响应均需人类最终确认。

在知识范畴，合格的安全专家应是“OT+IT+AI”复合型人才：既需掌握工控协议与PLC、DCS运行机制等OT基础，又需熟悉IEC 62443等工业控制系统安全标准与威胁模型，更需理解LSTM等时序模型识别异常模式、孤立森林与自编码器进行无监督检测的AI原理。同时须熟知《网络安全法》《数据安全法》及关键信息基础设施相关规定，确保AI部署合规。ISO/IEC 42001:2023为构建这一知识体系提供了顶层框架。

在技能维度，核心要求在于可解释性与低侵入性。因OT工程师需快速明确警报成因，相关人才应熟练运用SHAP、LIME等工具将模型决策转化为可读说明。同时，工业边缘设备资源有限，掌握TinyML等轻量化部署技术对降低延迟、减少对主控系统的依赖至关重要。此外，人才须具备在数字孪生环境中复现SCADA拒绝服务、PLC篡改等攻击并验证AI防御模型的能力。

在实践领域，鉴于工控系统具备高风险、高成本特性，实践应聚焦数字孪生仿真攻防演练，在“零风险”环境中检验AI监测与响应机制的效能。同时须建立IT安全、OT运维与AI算法专家的常态化协作机制，共同制定模型更新策略与应急预案，并将关键性能指标纳入绩效考核，形成持续改进闭环。腾讯云在智慧水务等领域的实践案例表明，AI风险预警的落地必然历经大量实践迭代。

4.3.5 AI赋能的数据跨境流动安全

全球数字经济时代，数据已成为第五大生产要素，其安全有序跨境流动是推动国际贸易与技术创新的关键。然而，各国出于数据主权、国家安全与隐私保护的考量，相继建立了复杂的数据出境监管体系。我国密集出台《网络安全法》《数据安全法》《个人信息保护法》，配套发布《数据出境安全评估办法》《个人信息出境标准合同办法》等规范性文件，形成了以“安全评估、专业认证、标准合同”为核心的多层次监管格局。在此背景下，AI技术呈现出显著的双重属性：一方面可助力企业自动化识别、分类海量数据，提升合规工作效率；另一方面，将包含隐私或商业秘密的数据输入通用大模型时，存在极高的数据泄露风险。如何依托AI提升合规效率的同时保障数据安全，是数据跨境安全领域的核心矛盾。本节基于ASK-P框架，对解决该矛盾所需的核心能力展开分析，见表4-8所示。

表4-8 AI 赋能的数据跨境流动安全 ASK-P 维度评价表

ASK-P 维度	核心能力要求	场景侧重
意识	需树立牢固的法律遵从意识, 明确“三驾马车”及配套规章的硬性要求; 认知数据出境后受到接收国法律约束的相关风险; 对未经脱敏处理的敏感数据跨境传输保持高度警惕。	对数据主权与法律边界的敬畏, 直接关系企业存亡。
知识	掌握我国数据分类分级制度; 熟知GDPR、CCPA 等域外法规; 了解差分隐私、联邦学习、同态加密等 PETs 原理; 熟悉我国标准合同条款。	统筹国内监管强制性要求与国际规则兼容性, 筑牢跨国合规基础。
技能	可配置AI模型对PII等敏感信息进行自动扫描识别; 精通泛化、抑制、扰动等脱敏策略; 具备第三方AI 合规工具评估能力。	聚焦“自动化合规”与“数据脱敏”, 替代人工审核。
实践	设计全生命周期跨境日志审计体系, 定期开展内外部合规审查, 构建数据出境后风险动态监控与应急机制。	重点聚焦于“全流程审计”与“动态监控”。

在意识层面, 从业者须具备高于普通技术人员的法律意识与政治敏感性, 树立“数据主权”与“合规至上”的红线意识。数据跨境并非单纯技术或商业问题, 而是严肃的法律与政治议题, 违规操作可能导致巨额罚款乃至刑事责任。从业者不仅需熟知“三驾马车”确立的基本原则, 还应深入研习配套规章细节, 明确何种数据、何种情形需申报安全评估, 何种可采用标准合同。更关键的是, 数据一旦跨境, 后续处理将受接收国法律约束, 引发不可预知的合规与安全风险。因此, 对任何未经严格去标识化的敏感数据跨境行为, 必须保持高度警觉。

在知识领域, 从业者需构建“硬约束”与“软联通”兼备的复合型知识体系。一方面, 全面掌握中国数据分类分级制度及重要数据识别指南, 精准界定重点保护范围; 另一方面, 熟知GDPR、CCPA 等主要经济体法规, 设计全球通用合规方案。同时, 应掌握差分隐私、联邦学习、同态加密等隐私增强技术的原理与适用场景, 深入理解《个人信息出境标准合同办法》具体条款。NIST 隐私框架与AI 风险管理框架 (AI RMF) 为构建此类复合知识体系提供了良好参考。

在技能范畴, AI的核心价值在于依托模式识别与自然语言处理, 实现海量异构数据的自动化合规处理。从业者须能配置训练AI模型自动扫描存储系统, 精准识别PII、财务数据等敏感信息并完成自动分类打标; 精通泛化、抑制、扰动等多种脱敏策略并可灵活运用——如统计分析数据可采用差分隐私添加噪声, 模型训练数据可运用联邦学习实现“数据不动模型动”; 同时具备评估第三方AI 合规工具的有效性与潜在风险的能力。

在实践领域, 数据跨境是贯穿数据全生命周期的持续性管理活动, 核心在于构建可审

计、可监控的闭环体系。首要任务是设计覆盖数据产生、处理、传输及销毁全过程的日志审计系统，确保每一次数据出境皆有迹可循。其次，建立定期内部合规审查机制并引入第三方独立评估。最为关键的是构建出境后动态监控机制，当接收方的法律环境发生变化，或出现《数据出境安全评估办法》第十四条规定的需重新申报情形时，企业须及时启动应急预案。这种兼具前瞻性与动态性的管理实践，是企业在复杂国际数据流动环境中稳定运行的根本保障。

4.3.6 AI系统自身的安全防护

随着AI技术在金融、医疗、交通、政务等关键领域实现规模化部署，AI系统已从被动工具转变为具备决策能力的“资产”，其自身安全已成为战略性问题。攻击者可通过数据投毒、对抗样本攻击、模型窃取、提示注入及后门植入等多种路径发起攻击。ENISA《2024年AI威胁态势报告》显示，针对AI模型的攻击事件年增长率超过140%，部分攻击已在生产环境造成实质性损失。与此同时，我国《生成式人工智能服务管理暂行办法》等法规已对内容安全、数据合规提出明确要求。该领域要求网络安全人才跨越网络安全与机器学习的学科边界，具备全生命周期AI风险治理与实战化对抗能力。本节基于ASK-P框架对该高阶能力模型展开分析。

如表4-9所示，在意识层面，从业者需完成从传统软件漏洞视角向“AI资产”视角的深刻转变，建立全局性、前瞻性的风险治理意识。AI模型不仅是代码与参数权重的集合，

表4-9 AI系统自身的安全防护ASK-P维度评价表

ASK-P 维度	核心能力要求	场景侧重
意识	树立覆盖“数据采集—模型训练—部署上线—持续监控”全生命周期的风险意识；明确AI安全涵盖合规、伦理及供应链安全；警惕AI作为新型资产被篡改或滥用引发的系统性风险。	具备全局性与前瞻性的AI风险治理视角，将AI列为需严格管理的关键资产。
知识	掌握FGSM、PGD等对抗性机器学习原理；熟悉NIST AI RMF、ISO/IEC 42001及国内监管框架；通晓联邦学习、差分隐私等PETs以及模型安全加固方法。	衔接国际标准与国内监管要求，覆盖技术与管理双维度。
技能	运用PyRIT、ART等红队工具开展AI渗透测试；掌握攻击成功率、置信度漂移等鲁棒性评估指标；设计输入过滤、输出拦截及运行时监控等防护架构。	聚焦“以攻促防”的实战化测试与动态防御。
实践	主导上线前红蓝对抗演练，编制并输出风险报告；构建运行阶段异常监控基线；参与AI漏洞披露与CVE申报；将AI安全要求嵌入DevSecOps流程与采购审查环节。	由理论测试转向常态化治理，构建可持续的AI安全运营机制。

还蕴含训练数据的分布特征与内在逻辑，属于高价值敏感资产，其数据采集、训练、部署、监控直至退役的全流程均需保持同等程度的风险警惕。从业者需突破纯技术视角的局限，明确AI安全交织法律、伦理与社会责任维度。一处看似无涉的模型偏见，即可在信贷审批或人员招聘环节引发歧视性后果。更需警惕AI被恶意操控引发的系统性风险，例如被植入后门的医疗诊断模型，其危害程度远超传统IT系统。NIST AI RMF与ISO/IEC 42001均强调全生命周期风险管理理念，为培养该全局意识提供了理论指引。

在知识范畴，本场景是各场景中知识壁垒最高的领域，要求从业者同时具备攻防能力。攻击层面，需深入掌握FGSM、PGD等对抗样本生成技术，以及数据投毒、模型窃取、后门攻击等复杂攻击模式；防御层面，需掌握对抗性训练、输入净化、梯度遮蔽、鲁棒性验证等安全加固技术。治理层面，需熟知NIST AI RMF[10]、ISO/IEC 42001等国际框架，准确理解《生成式人工智能服务管理暂行办法》中关于内容安全、数据来源、标注规范的强制性要求。此外，持续跟踪PETs与可解释性AI的最新发展，也是构建完整知识体系的必要环节。

在技能层面，核心在于“以攻促防”的实战能力。第一，AI模型渗透测试能力：熟练运用微软PyRIT、IBM ART等红队工具主动探测系统脆弱性边界，覆盖对抗样本生成、提示注入设计及模型窃取尝试等环节。第二，定量鲁棒性评估能力：通过攻击成功率、准确率下降幅度、置信度漂移等指标科学衡量模型安全水平。第三，多层纵深防御架构设计能力：综合运用输入过滤、输出拦截、运行时监控、异常检测等技术，构建动态自适应防御体系。

在实践层面，AI安全需构建持续、动态且常态化的治理运营机制。首先，上线前需主导安全评估与红蓝对抗演练，输出结构化风险报告作为系统上线的否决依据。其次，上线后建立运行阶段持续监控基线，实时检测对抗攻击并自动触发降级、告警或熔断等应急机制。再次，积极参与AI漏洞披露与CVE申报，推动行业防御知识共享。最后，将AI安全控制点系统嵌入DevSecOps流程及供应商采购审查标准，落实安全左移要求。

4.4 总结与建议

综合本章应用分析、调研结果和场景化建模可以看到，AI更适合承担重复性、标准化和数据密集型任务，复杂研判、风险定级和高影响处置仍需安全人员核验和决策。当前人才能力建设的重点，不是单纯增加工具数量，而是建立任务拆解、工具约束、结果验证和

方法沉淀的协同机制。

本章以ASK-P为共同框架，将不同场景的专业能力与AI协同能力结合起来；六个场景作为代表性样例，而非完整场景清单，并根据业务对象、岗位层级和风险特征实施差异化评价。

后续应进一步明确人机协同边界和自动化权限，完善人工复核、操作审计和风险控制机制。同时，加强场景化实训、仿真靶场和案例资源建设，将AI工具应用、结果核验和复杂任务处置纳入人才培养与评价，并根据技术演进和实际安全需求持续更新能力要求。



AI时代网络安全人才培养体系

5.1 高校培养体系的现状

高校是网络安全人才培养的主阵地。伴随网络空间安全一级学科、一流网络安全学院建设示范项目与新工科建设持续推进，高校已形成以专业建设为基础、课程教学为主体、实验实训为支撑，教材资源和师资队伍协同保障的人才培养底座。

AI 技术融入安全业务后，网络安全的保护对象、工具链和 workflows 均发生了显著变化：模型参数、训练数据、算法鲁棒性、大模型应用和智能体成为新的安全防护对象；漏洞挖掘、日志分析、威胁检测、响应处置等工作也开始引入 AI 工具。相应地，高校既要保持网络安全基础课程，也需要有序补充模型安全、AI 应用安全、AI 赋能安全三类新兴教学内容。

本节基于面向网络安全相关专业教师的广泛调研，呈现我国院校课程体系、实验体系、教材与教学资源、师资队伍建设现状与不足，为院校优化培养方案、提升人才供给质量提供参考。

5.1.1 课程体系建设现状

当前，我国院校已开设信息安全、网络空间安全、数据安全、密码科学与技术、信息对抗技术等网络安全相关专业，课程体系由公共基础、专业基础、方向课程和综合实践等组成。AI 安全是在上述课程体系上的交叉拓展。

1. 网络安全相关专业设置情况

从专业结构上看，院校信息安全和网络空间安全专业开设占比最高，分别达 60%、55%，构成网络安全人才培养的主体；密码科学与技术、信息对抗技术专业开设占比分别为 31%、19%，定位特定领域，承担着专项人才的培养的作用。如图 5-1 所示。

2. AI 安全专项课程开设情况

AI 安全相关内容已进入多数院校的教学安排，34% 的院校已开设独立必修课，34%

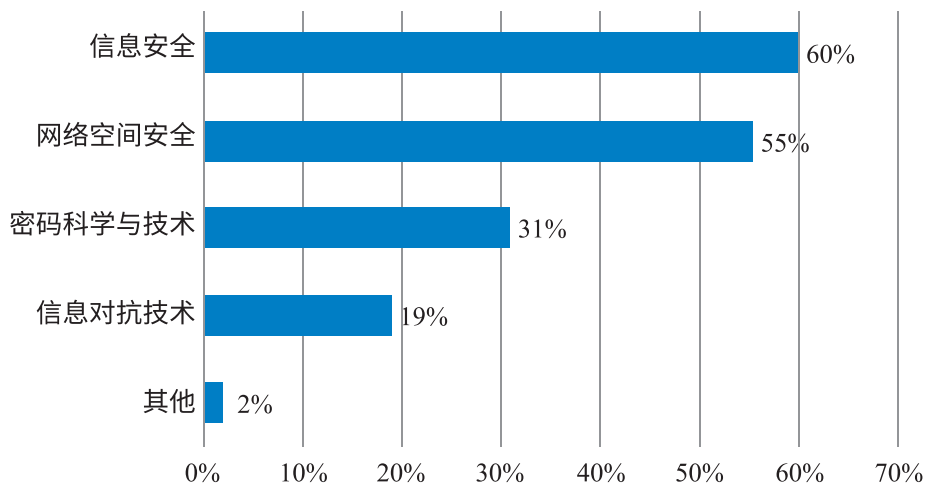


图5-1 高校网络安全相关专业设置情况

已开设独立选修课，20% 将相关内容融入传统课程，12% 仍处于规划阶段。独立必修课和独立选修课合计占 68%。若将融合式教学计算在内，已有 88% 的院校以不同形式纳入 AI 安全内容，如图 5-2 所示。

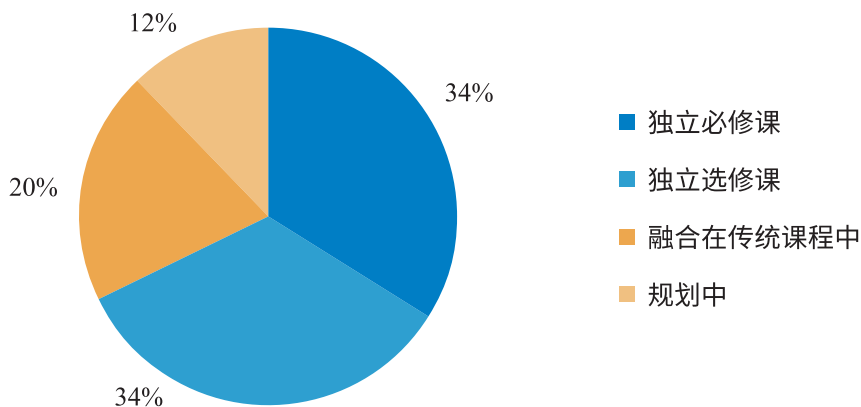


图5-2 AI安全专项课程开设情况

独立必修课占比较高，表明 AI 安全已不再局限于专题讲座或临时教学内容，而是进入相对固定的课程安排。与此同时，必修课与选修课占比接近，说明不同院校对 AI 安全课程的定位尚不一致：部分院校将其作为网络安全人才的基础能力，部分院校则将其作为方向性或拓展性课程。

3. 不同专业AI安全课程设置情况

从专业领域看，网络空间安全专业开设独立课程的比例最高，为 79%，其中独立必修课为 49%。信息对抗技术专业开设独立课程的比例最低，为 68%，独立必修课仅 41%。信息安全专业和密码科学与技术专业开设独立课程比例分别为 75%、73%。如图 5-3 所示。

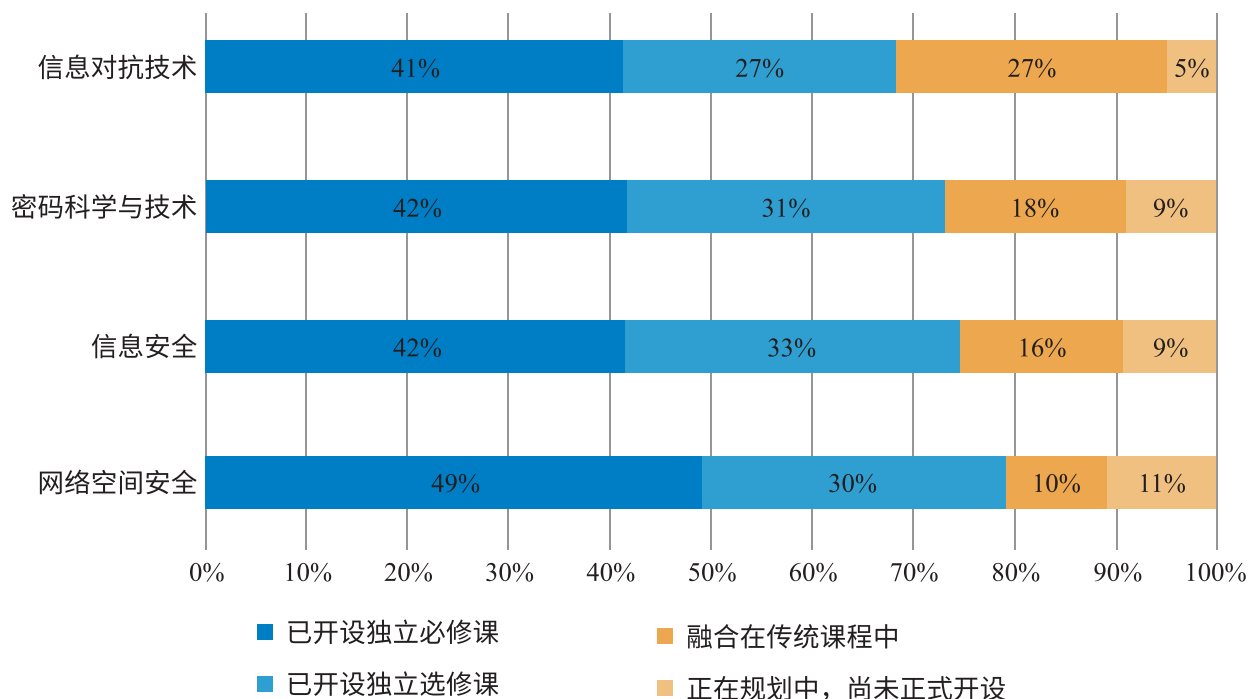


图5-3 不同专业领域AI安全课程设置情况

上述差异与各专业学科定位密切相关。网络空间安全侧重整体安全架构, AI 安全更容易被纳入必修模块; 信息对抗技术聚焦实战对抗, AI 内容更适宜嵌入既有课程; 密码科学与技术与 AI 安全的结合仍处于前沿探索阶段, 课程规划空间较大; 信息安全专业则需要在传统安全基础与新兴 AI 安全内容之间保持衔接, 因此独立课程与融合式教学并存。总体来看, 各专业均已开始以不同方式布局 AI 安全课程, 但课程边界和能力递进关系仍需进一步明确。

课程覆盖程度提高的同时, “AI 内容更新滞后”仍是当今网络安全人才实战能力培养的主要矛盾。从院校的反馈来看, 40% 的高校教师认为“课程体系落后于技术更迭”是人才培养面临的最大挑战。这说明高校课程体系建设中的痛点, 已从“是否开设 AI 安全课程”转向“如何构建动态迭代的课程内容”。高校要密切跟踪 AI 安全前沿风险类型与技术手段的演进, 及时将相关内容转化为可教学、可训练、可评价的课程单元。在后续课程建设不仅要增加课程供给, 还需要明确课程内容边界和先修关系, 完善理论教学、实验实训和考核评价的配套设计。

4.AI赋能教学模式的更新变革

AI 技术不仅改变了网络安全的教学内容, 也在深刻重塑教学方式和学习形态。随着大语言模型、智能问答系统和开源社区的普及, 学生获取基础知识的渠道已大幅拓宽, 传统以教师讲授为主的课程模式需要相应调整。

首先，教学资源的来源结构正在发生变化。调研显示，在“教学资源最紧缺的是什么”的回答中，“配套的实操案例库与代码示例”和“行业真实攻击案例库”合计占比超过六成，明显高于“体系化的理论教材”的选择比例，反映出高校对实践性、情境化学习资源的需求更加突出。

其次，教师角色正在从知识传授者向学习引导者转变。当学生可以通过 AI 工具自主获取基础知识和解决常见问题后，教师最需要的已不再是“如何讲好概念”，而是“如何设计高阶学习任务”和“如何保持对前沿技术的敏感度”。此外，“课程体系落后于技术更迭”被列为高校培养 AI 安全实战人才面临的最大挑战之一，进一步说明课程建设的重心应从“教什么”拓展到“怎么教”，从知识传授转向能力建构和个性化学习。

5.1.2 实验体系建设现状

网络安全是一门实践性极强的学科，我国高校普遍以实验室为载体对学生开展 AI 安全实训。大模型及 AI 应用的广泛使用使得安全威胁场景大幅拓展，也对高校的实验环境、实验场景提出了更高要求。

1. 实验环境对于AI安全实训的支撑情况

调研数据显示，在 AI 安全实训方面，38% 的高校具备基础实验环境，但算力受限，仅能运行小型模型或简单脚本；26% 的院校具备充足算力，可开展模型训练和对抗攻击实验；18% 无自有算力，主要依赖云端或第三方实训平台；16% 的实验室仍仅支持传统网络安全实验，2% 的院校暂无网络安全实验室。这说明我国高校 AI 安全实验环境建设仍处于“能开课、难实训”的初期阶段，算力不足是最突出的瓶颈，基础设施覆盖也存在明显缺口，制约了实训教学的规模化推进。如图 5-4 所示。

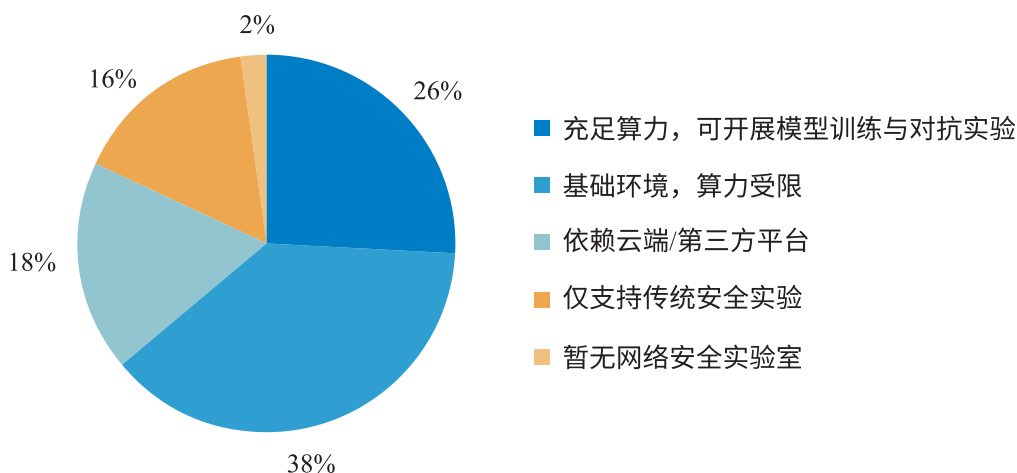


图5-4 实验环境对AI安全实训的支撑情况

2.实验场景对于AI安全实训的支撑情况

从实验教学场景来看，55% 的高校注重 AI 模型自身脆弱性检测，46% 的高校更注重 AI 赋能的攻防工具使用，37% 的高校注重 AI 应用的合规与安全。这一分布反映了当前高校对 AI 安全实验教学内容的优先级判断：模型安全问题已成为安全领域最紧迫的议题，因此覆盖最广；AI 赋能攻防工具使用次之，表明 AI 在自动化挖掘漏洞、渗透测试等方面的突破，已经快速进入高校教学研究视野；而 AI 应用的合规与安全虽在理论教学中日益受到重视，但转化为实验教学的比例仍有较大提升空间。如图 5-5 所示。

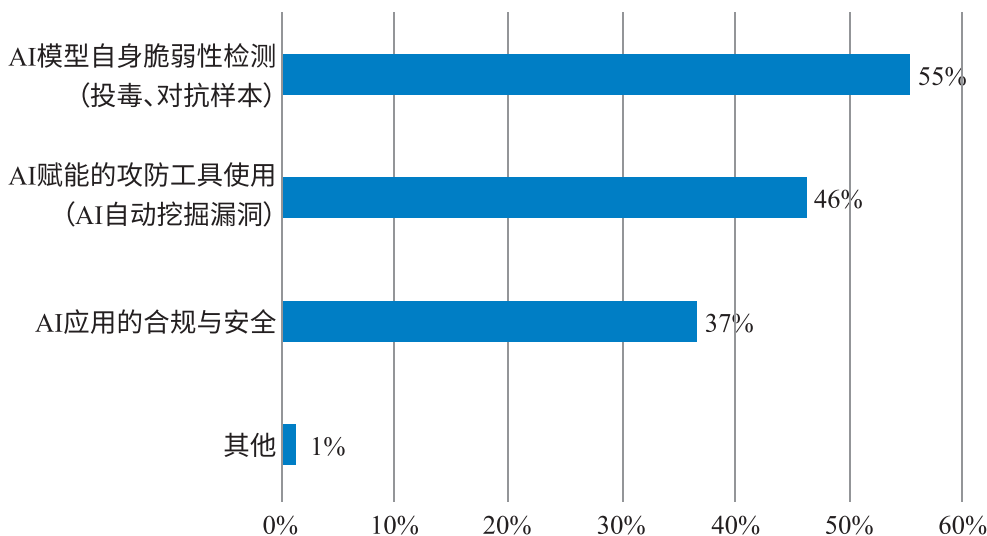


图5-5 实验场景对于AI安全实训的支撑情况

5.1.3 教材与教学资源建设情况

教材是课程内容的核心载体，也是连接培养方案与课堂教学的关键环节。AI 安全涉及机器学习、深度学习、对抗攻防、大模型安全等多个交叉领域，知识更新迅速，对教材内容的时效性和体系性提出了较高要求。

1. 教材来源与使用情况

在 AI 安全教学中，50% 的院校选用国内出版的最新教材，39% 使用校企合作开发的实战型讲义，38% 翻译或参考国外原版教材，22% 尚无固定教材，主要依靠教师自编课件和开源资料。不同教材来源在实际教学中并行使用，说明不少院校已不再依赖单一教材，而是通过正式教材、实战讲义、国外资料和教师自编资源组合支撑教学，如图 5-6 所示。

国内教材占比最高，为课程内容规范化提供了基础；校企合作讲义和国外资料占比接近四成，是补充前沿技术和真实案例的重要来源。仍有 22% 的院校主要依赖教师自编课件和开源资料，反映出部分课程尚缺少稳定教材。

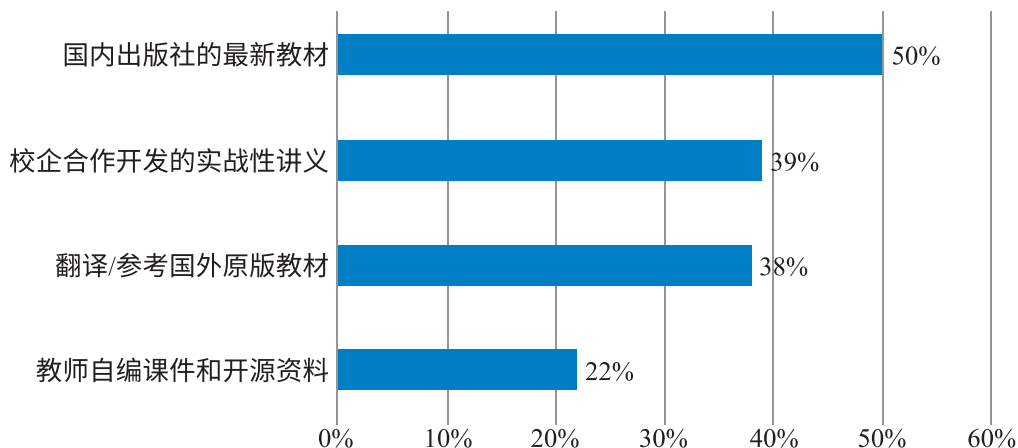


图5-6 AI安全教学教材来源情况

2. 不同课程体系建设阶段的资源差异

在已经开设独立课程(包括开始独立必修课、独立选修课)的院校中, 59% 选用国内最新教材, 45% 参考国外原版教材, 44% 使用校企合作讲义, 主要依赖教师自编课件和开源资料的比例为 14%。在融合式开设和规划中的院校中, 主要依赖教师自编课件和开源资料的比例分别为 40% 和 41%, 如图 5-7 所示。

这一差异说明, 课程体系建设阶段与教学资源稳定性具有较强关联。必修课 / 选修课通常具有明确的教学目标、学时和考核要求, 更容易形成正式教材、专题讲义和配套实验; 融合式教学或规划阶段的课程则需要依靠教师快速整合公开资料, 以适应尚未定型的课程体系。

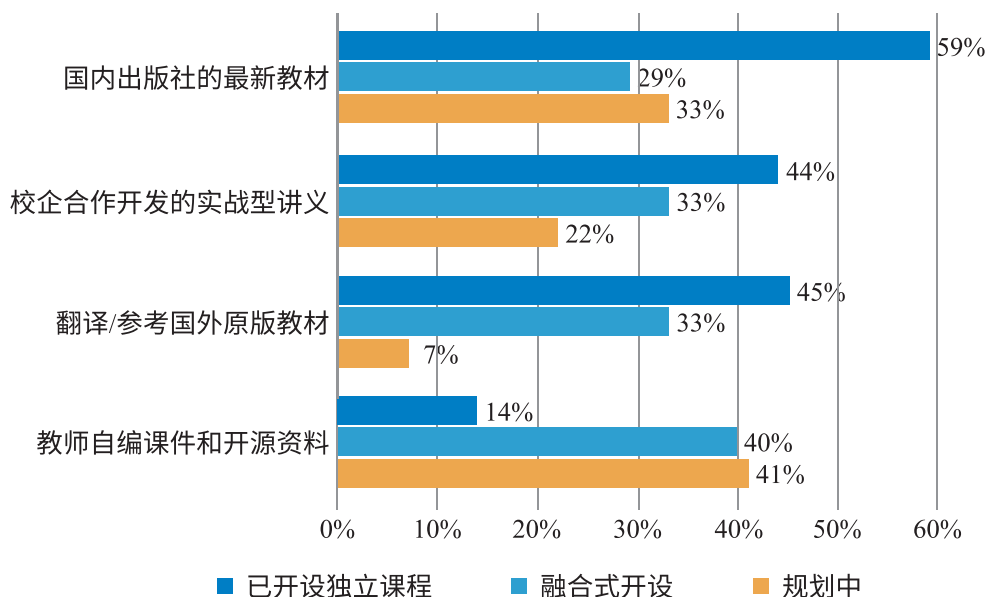


图5-7 不同课程建设阶段的教材来源情况

3. 教学资源紧缺情况

在教学配套资源方面，35% 的高校认为 AI 安全教学最缺少配套的实操案例库与代码示例，24% 的高校认为最缺少行业真实攻击案例库；认为最缺少实验指导书与视频资源、体系化理论教材、教学评估或考试题库的高校，分别占比 21%、18%、2%，如图 5-8 所示。

实操案例、真实攻击案例和实验指导资源合计占 80%，说明当前资源短板主要集中在“如何把知识转化为可操作训练”，而不仅是理论教材数量不足。

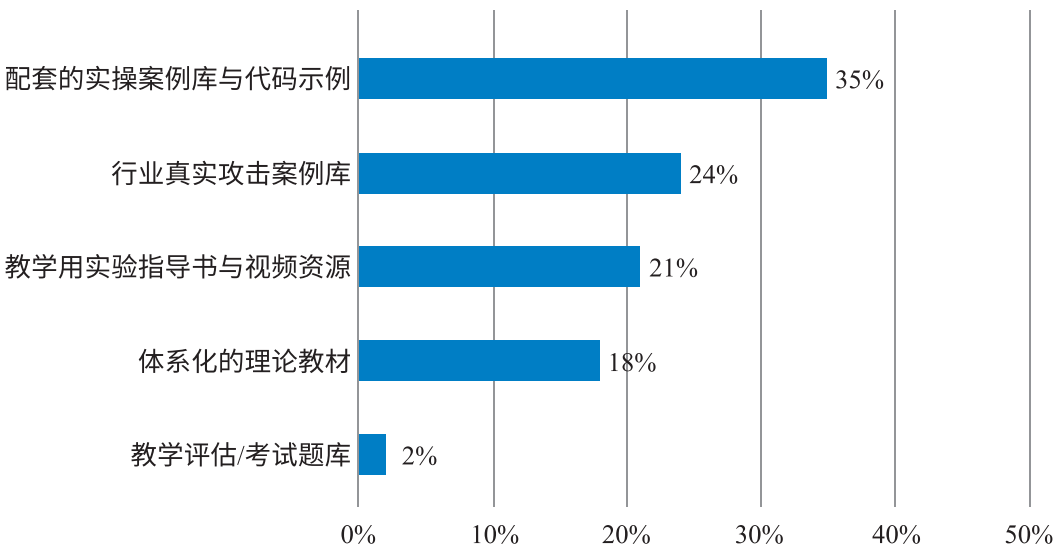


图5-8 AI安全教学资源紧缺情况

5.1.4 师资队伍建设情况

师资队伍是高校网络安全人才培养体系的核心支撑。AI 安全教学同时涉及网络攻防、系统与软件安全、人工智能技术等，单一学科背景难以覆盖全部内容。当前高校主要通过原有网络安全教师学习 AI 技术、人工智能教师跨界研究安全、组建跨学科团队以及引入行业专家等方式补充师资。

1. 教师背景结构

在高校从事 AI 安全教学教师的背景方面，45% 的教师来自传统网络安全背景，通过自学或培训研究 AI 安全；28% 的教师来自人工智能或机器学习背景，后跨界研究安全；21% 具备安全与算法双重经验；以企业专家作为校外导师或兼职教师为主要来源的比例为 5%，如图 5-9 所示。

传统网络安全教师和人工智能教师跨界转型合计占 74%，说明当前师资建设主要依靠既有教师扩展知识边界。传统网络安全教师具有安全基础和攻防经验，但需要补充算法

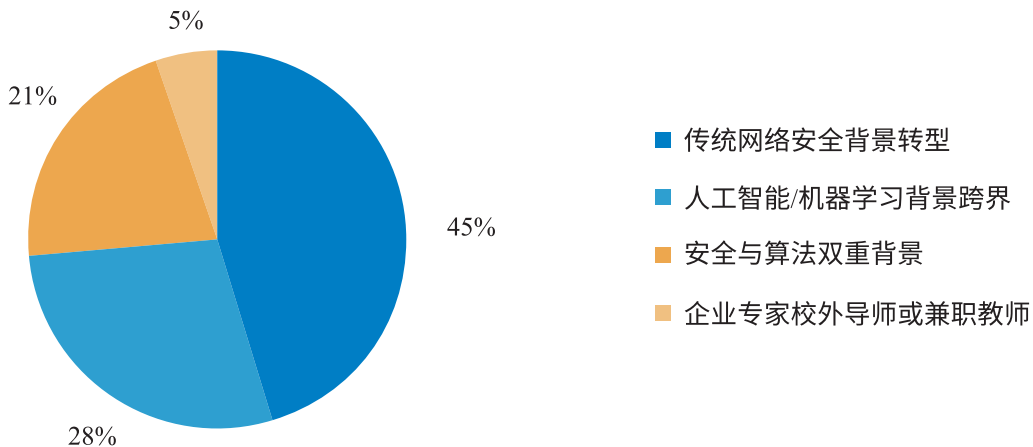


图5-9 AI安全教学教师主要背景

原理、模型训练和大模型应用机制；人工智能教师熟悉模型和数据处理，但需要进一步理解安全工程流程、攻击路径和合规要求。兼具安全与算法经验的教师占21%，尚未成为师资主体。

2. 师资能力提升障碍

在师资能力提升障碍方面，37%的院校认为最突出的问题是缺乏产业界真实的AI安全实战经验，27%认为教学任务繁重、缺少前沿技术迭代时间，27%认为跨学科知识沉淀不足、难以兼顾AI与安全，10%认为学校专项投入不足，如图5-10所示。

前三项合计占90%，表明师资问题主要表现为知识结构、工程经验和持续学习时间不足。AI安全技术更新较快，教师需要同时承担教学、科研和实验平台维护任务，单次培训难以解决能力更新问题。

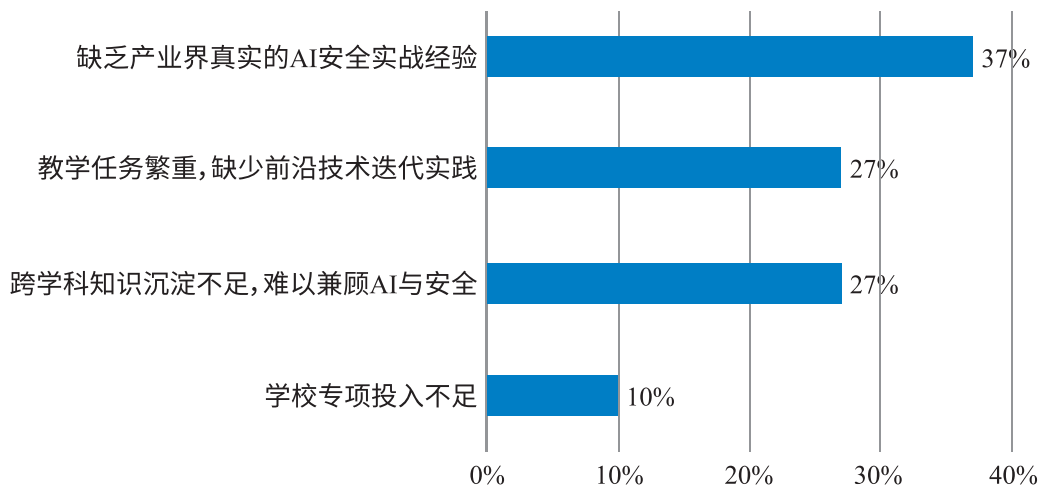


图5-10 师资队伍AI安全教学能力提升障碍

3. 外部师资参与情况

当前高校 AI 安全教学仍主要依靠传统网络安全教师和人工智能教师跨界转型，以企业专家作为校外导师或兼职教师为主要来源的院校占比为 5%。外部专家更多承担专题讲授、实训指导、案例引入和项目辅导等补充性角色。

从师资队伍建设的看，高校 AI 安全教学正在由单一学术型师资向网络安全、人工智能和行业实践协同结构发展，但复合型师资仍然不足。教师持续学习时间、真实项目经验和稳定实践支撑机制，是当前师资队伍建设中需要进一步完善的重点。

5.2 产教融合的实践探索

5.2.1 校企合作培养的典型模式

校企合作是 AI 安全人才培养的重要路径。本节从合作现状、典型合作模式、运行机制、实战成效、瓶颈问题和校企核心诉求等维度，梳理当前院校与企业合作培养的实践经验。

1. 校企合作现状

在 AI 安全教学方面，38% 的院校与 1-2 家企业建立了长期稳定合作，24% 的院校拥有 3 家及以上长期合作伙伴，另有 24% 的院校合作多为短期 / 单次项目，14% 的院校尚未建立实质性校企合作，如图 5-11 所示。

整体来看，已有合作经历的院校合计占比达 86%，校企合作已成为 AI 安全人才培养的主流路径。但近四成院校的合作仍停留在短期、浅层阶段或尚未建立合作，深度产教融

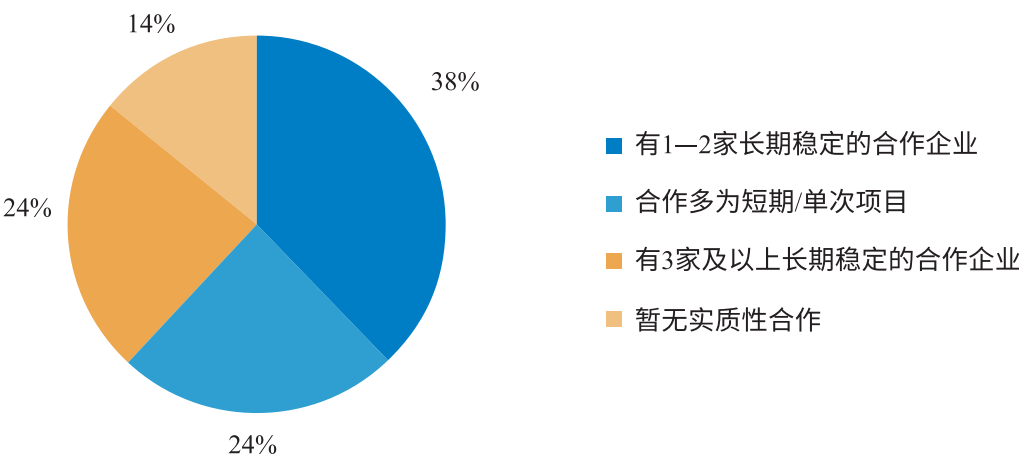


图5-11 校企合作关系类型分布

合仍是当前的主要短板。

2. 典型合作模式

当前校企合作已形成多种典型模式，占比排名前五的合作模式依次为：企业导师进课堂 / 讲座(50%)、共建 AI 安全实验室 / 靶场(50%)、联合开发 AI 安全课程 / 教材(35%)、订单式专班培养(25%) 和产学研联合攻关(23%)。如图 5-12 所示。

良好的运行机制是保障产教融合的关键。院校最具亮点的做法方面，课程共建是最主要的合作形式，43% 的院校与企业共同开发 AI 安全课程；39% 的院校与企业共同选拔网络安全实战人才；38% 的院校推行双导师制；34% 的院校和企业共同建立了竞赛 / 实训联动机制；仍有 15% 的院校反映暂无明确运行机制。如图 5-13 所示。

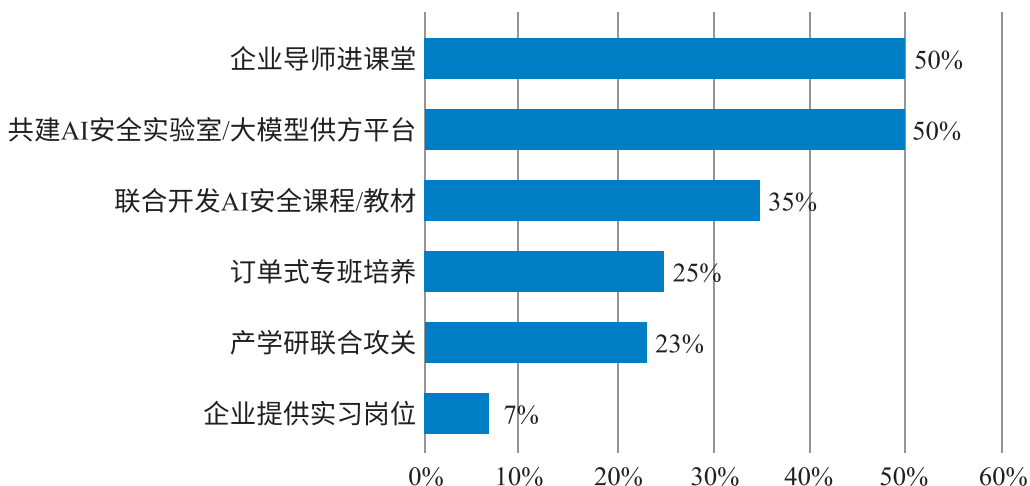


图5-12 校企合作主要模式分布

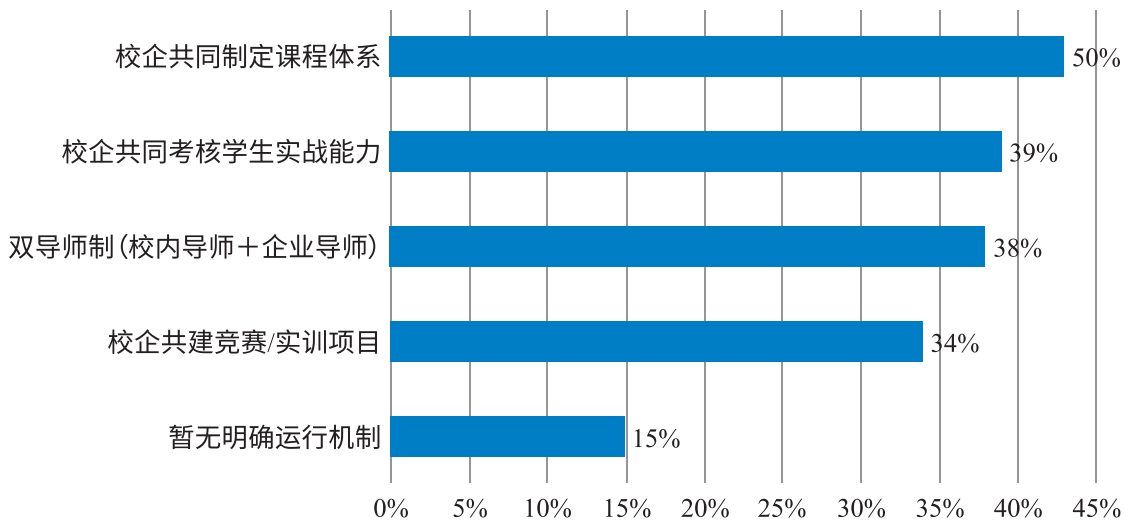


图5-13 院校校企合作亮点做法

3. 产教融合成效与瓶颈

在网络安全人才实战能力提升方面，34% 的高校认为学生通过产教融合，在“复杂 AI 安全工具的熟练使用”方面有明显提升；33% 的高校认为学生在“解决真实场景安全问题的逻辑能力”方面有显著提升；27% 的高校认为学生提升了“行业前沿威胁的敏感度”。如图 5-14 所示。

用人单位同样认可产教融合对网络安全人才实战提升的积极作用。在各类合作模式中，用人单位最认可的模式依次为“深度实习实训，参与企业一线 AI 安全项目”(44%)、“举办模拟真实环境的 AI 攻防竞赛”(33%)，“校企联合实验室，承接真实科研课题”(17%)和“企业专家进校园，讲授实战案例课”(6%)，如图 5-15 所示。这说明企业更看重真实任务、真实环境和真实评价，而不仅是课堂讲授或短期交流。

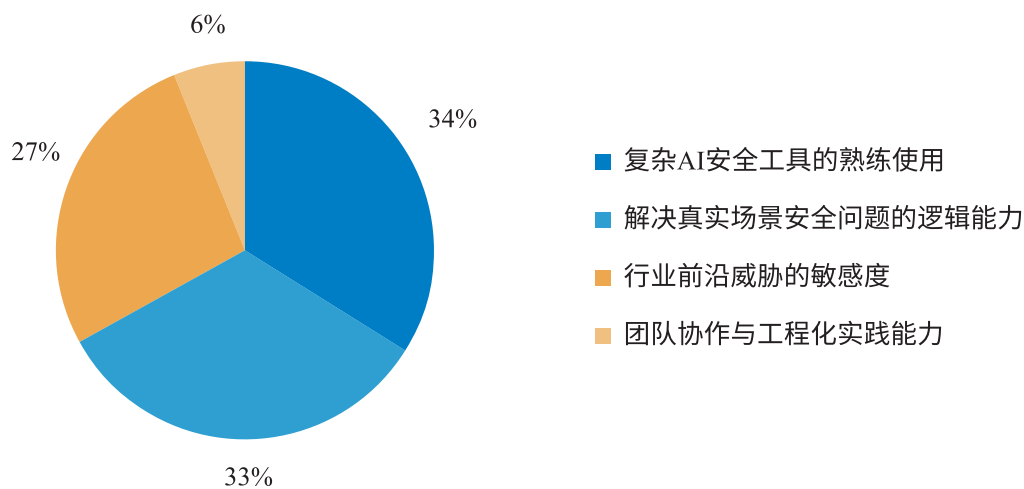


图5-14 产教融合实战能力提升效果

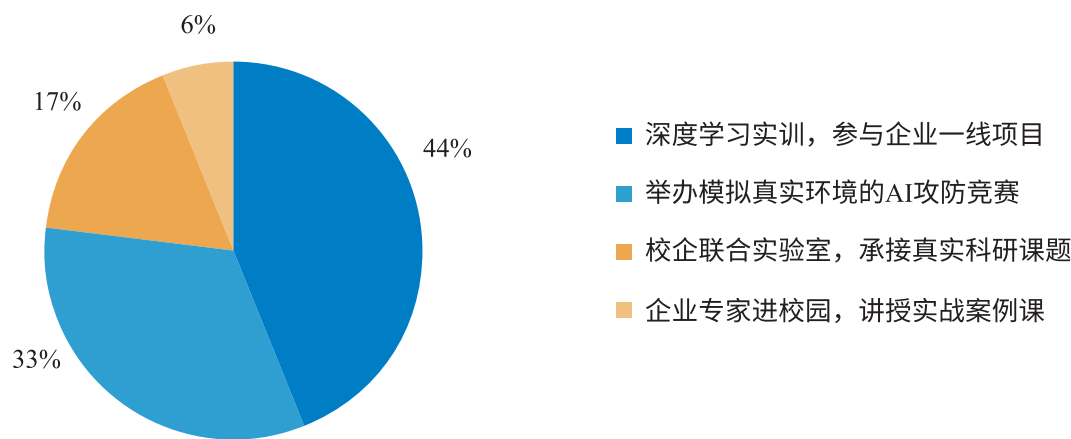


图5-15 企业最认可的产教融合模式

但当今产教融合仍面临多重结构性障碍。院校侧反映最大的障碍为“算力/平台采购成本高,实践环境投入不足”(32%) 其次为“缺乏同时懂 AI 和安全的双师型师资”(31%)“合作流于形式,深度不够”(18%)、“企业不愿开放真实数据和案例”(15%),如图 5-16 所示。

用人单位则将“缺乏支撑大型模型实验的算力与环境”列为首要瓶颈(40%),其次为“缺乏统一的人才实战能力评价标准”(35%)“教学内容与产业界实际需求严重脱节”(15%)。人才供需双方在“算力资源不足”上高度共鸣,反映出 AI 安全实训的基础设施建设是当前产教融合的共同短板,如图 5-17 所示。

4. 供给侧核心诉求

院校对企业支持的核心诉求集中在提供高性能算力平台访问权(32%)、开放真实业务场景的安全数据(29%)、建立长效的毕业生定向接收机制(28%)。数据与算力是 AI 安全实训的基础条件,开放程度直接影响产教融合的深度与质量。如图 5-18 所示。

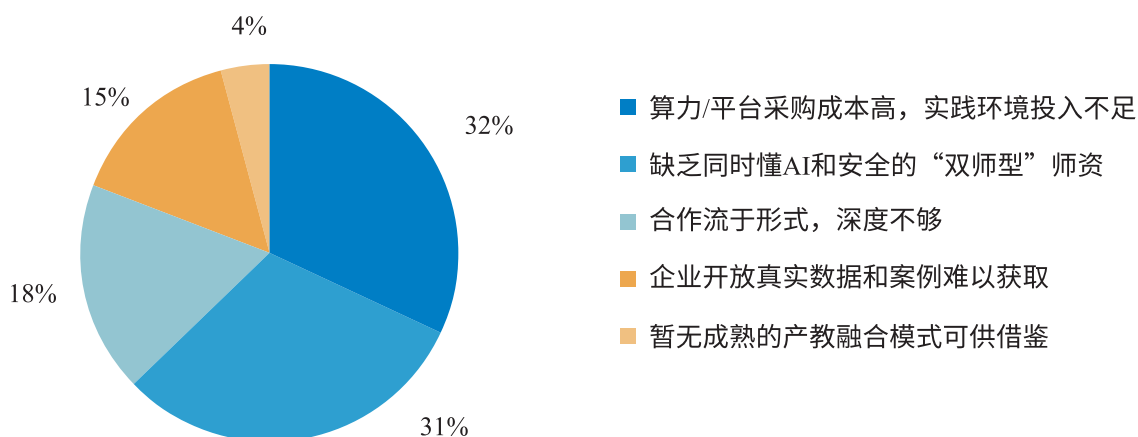


图5-16 产教融合主要瓶颈(院校视角)

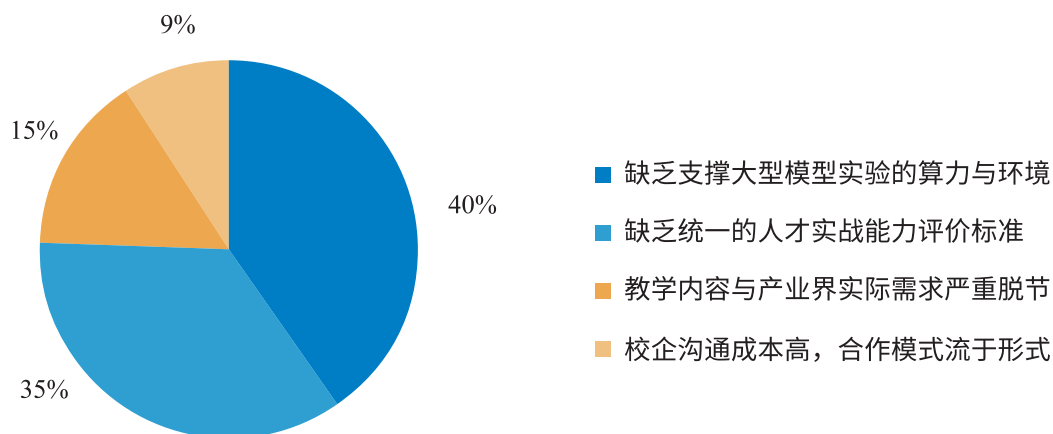


图5-17 产教融合主要瓶颈(用人单位视角)

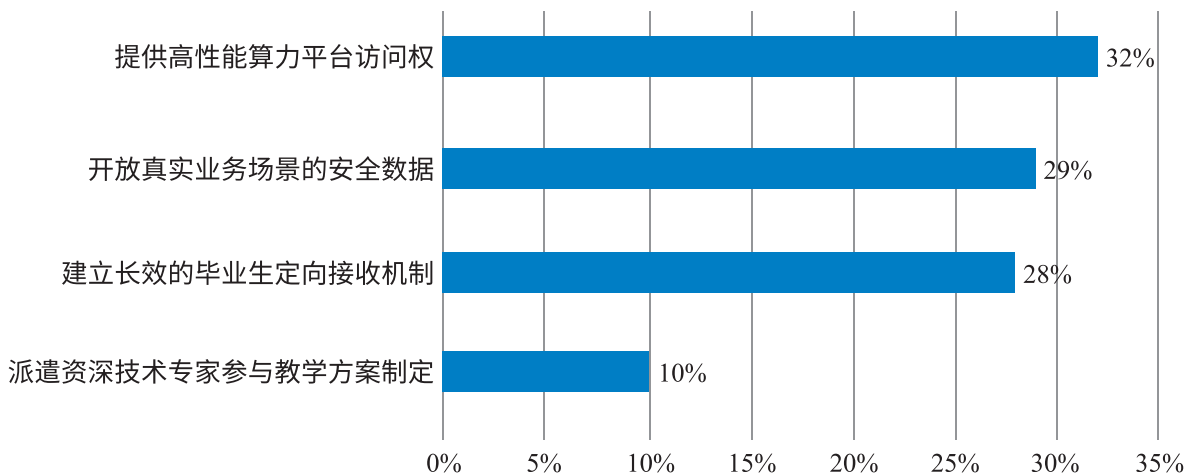


图5-18 院校最希望企业提供的支持

5.2.2 竞赛与实战演练的育人成效

竞赛与实战演练是 AI 安全人才培养的重要补充手段，其“以赛促学、以战代练”的特性恰好弥补了传统课堂教学在真实对抗场景上的不足。本节从竞赛参与情况、人才能力培养实效、产教融合情况等维度，梳理当前竞赛与实战演练在育人方面的实际成效。

1. 竞赛参与的覆盖面持续扩大

院校学生参与网络安全竞赛的整体活跃度较高。经常组织参赛并获得奖项的比例为 44%，另有 30% 的院校学生偶尔参与但缺乏专项指导，两者合计超过七成。然而，竞赛内容的结构性矛盾同样突出：15% 的院校反馈学生“仅关注传统 CTF 竞赛”，4% 的院校表示“很少能看到 AI 安全竞赛”，这意味着当前竞赛仍以传统攻防为主，大模型安全检测、AI 智能体安全攻防、多模态伪造检测等新兴赛道的赛事供给明显不足。如图 5-19 所示。

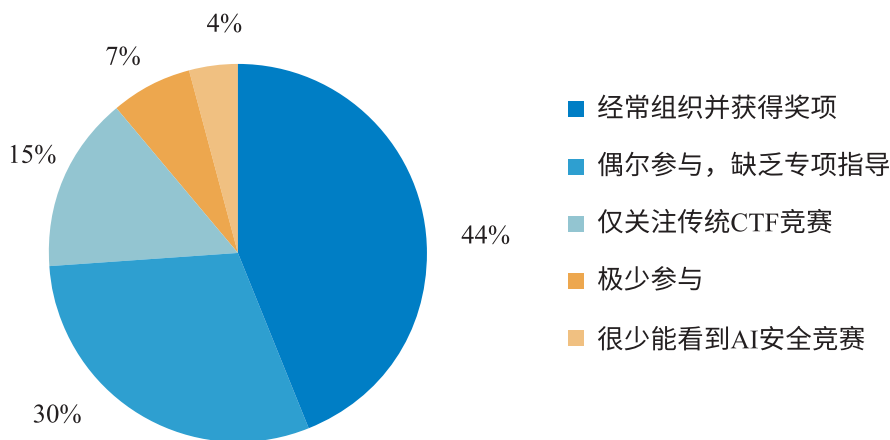


图5-19 院校学生AI安全竞赛参与情况

从学生参与的 AI 安全实训形式看，26% 的院校将“AI 安全竞赛集训”作为主要方式之一，与“企业 AI 安全部门顶岗实习”（30%）和“校企联合项目实战”（27%）共同构成三大实训支柱，如图 5-20 所示。这反映出竞赛与实战演练已不仅是课外活动的补充，而是逐渐被纳入正式的实践教学体系。

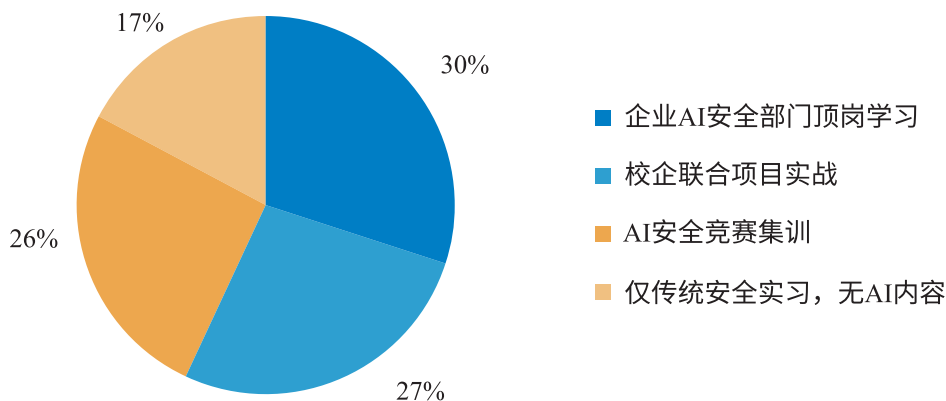


图5-20 院校AI安全实习主要形式

2. 实战型竞赛获高度认同

用人单位也高度认同竞赛演练对培养网络安全实战化人才的实效，41% 的用人单位认为“深度实习实训，参与企业一线项目”是最有效的人才培养方式，30% 的用人单位认为“举办模拟真实环境的 AI 攻防竞赛”是最有效的人才培养方式，两者合计超过七成。这表明无论是企业还是院校，都高度认同“以赛代训、按需补强”的实战型培养理念。如图 5-21 所示。

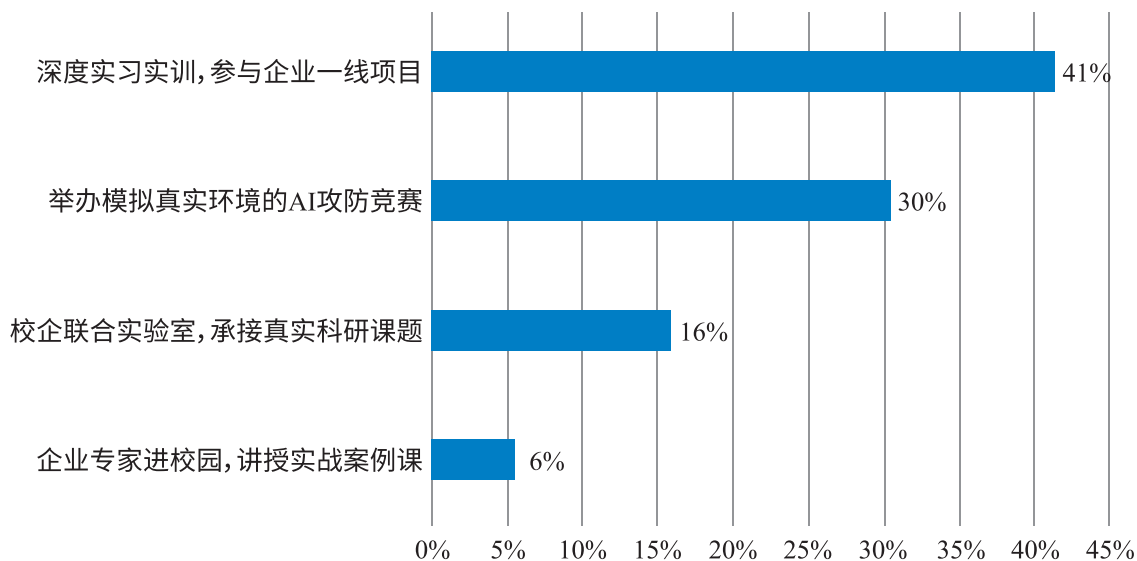


图5-21 用人单位视角网络安全人才实战能力培养实效

同样，从用人单位参与校企合作的具体切入点看，34%的用人单位与高校“共同举办竞赛演练”，这表明竞赛与演练已成为用人单位参与产教融合的重要载体，其高效率、可复制的特点使其在众多合作模式中脱颖而出。如图 5-22 所示。

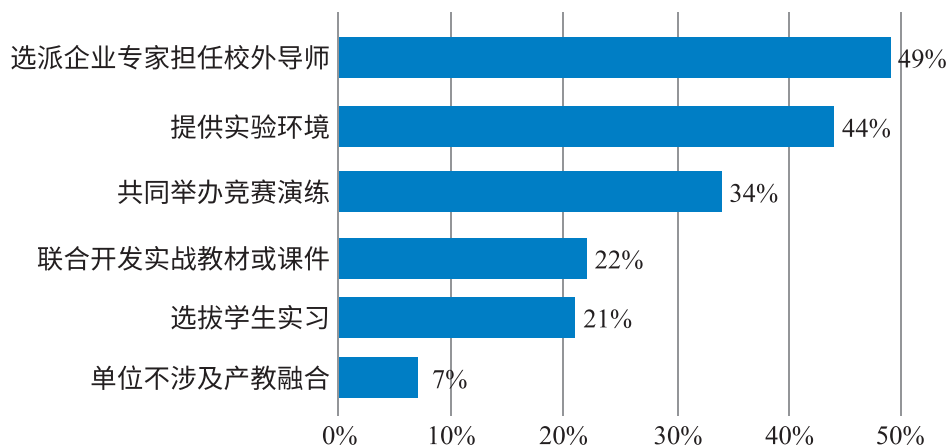


图5-22 用人单位参与校企合作的主要切入点

5.2.3 实习实训体系建设经验

AI 安全实战能力的培养，离不开体系化的实习实训体系建设支撑。本节从实习实训体系建设现状、核心模块构成、平台建设路径、实习形式、考核评价和特色创新六个维度，梳理当前院校的实习实训体系建设情况。

1. 体系建设现状

我国院校实习实训体系建设整体处于起步阶段。有实训但尚不成体系的占 38%；以零散实验课程为主、尚无体系化安排的占 37%；尚未开展 AI 安全专项实训的占 15%；已建成完整实训体系的院校仅 10%。这说明多数院校已开始探索 AI 安全实训，但体系化、连续性的训练安排仍明显不足。如图 5-23 所示。

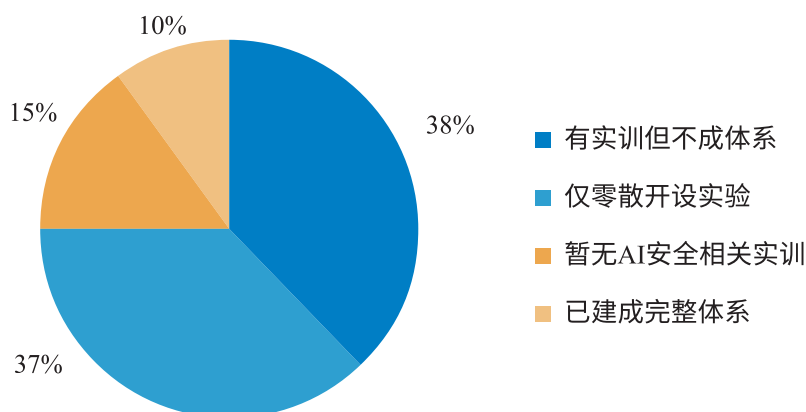


图5-23 AI安全专项实习实训体系建设现状

2. 实训核心模块构成

我国院校 AI 安全实习实训体系的模块，以 AI 赋能漏洞挖掘、AI 模型安全和 AI 攻防对抗为主。超过半数院校已开展 AI 赋能漏洞挖掘与渗透测试实训，说明该方向是当前实训建设中的主要内容，其次分别是 AI 模型安全、AI 攻防对抗实训、大模型应用安全与合规，设置数据安全与隐私计算实训模块的院校占比相对较低，仅为 19%。如图 5-24 所示。

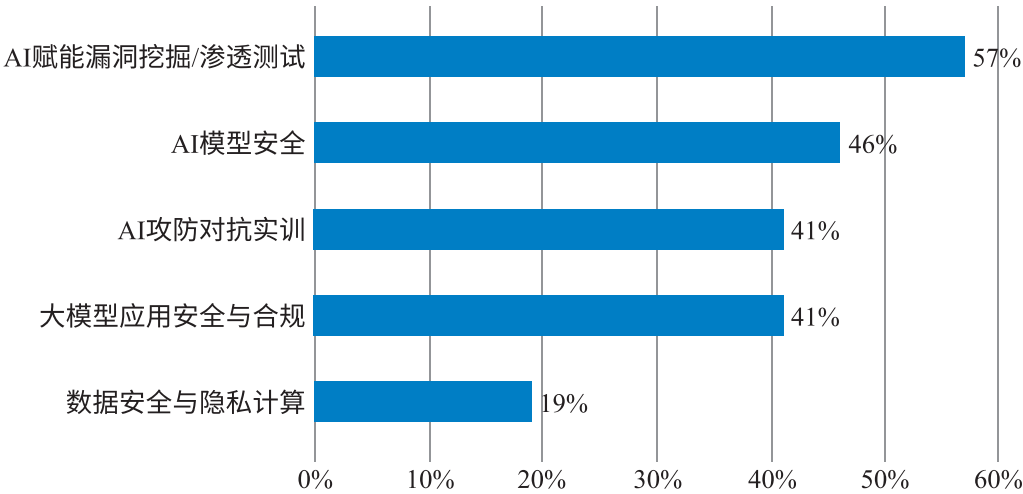


图5-24 AI安全实训核心模块覆盖比例

3. 平台与环境建设路径

在实训平台建设方式上，院校呈现出多来源并行的特征。当前 AI 安全实训平台建设以企业提供、院校自建和云端靶场三类方式为主，三者占比分别为 45%、41% 和 38%。相比之下，自行搭建开源仿真环境占 28%，多来源组合的混合模式占 12%，比例相对较低。如图 5-25 所示。

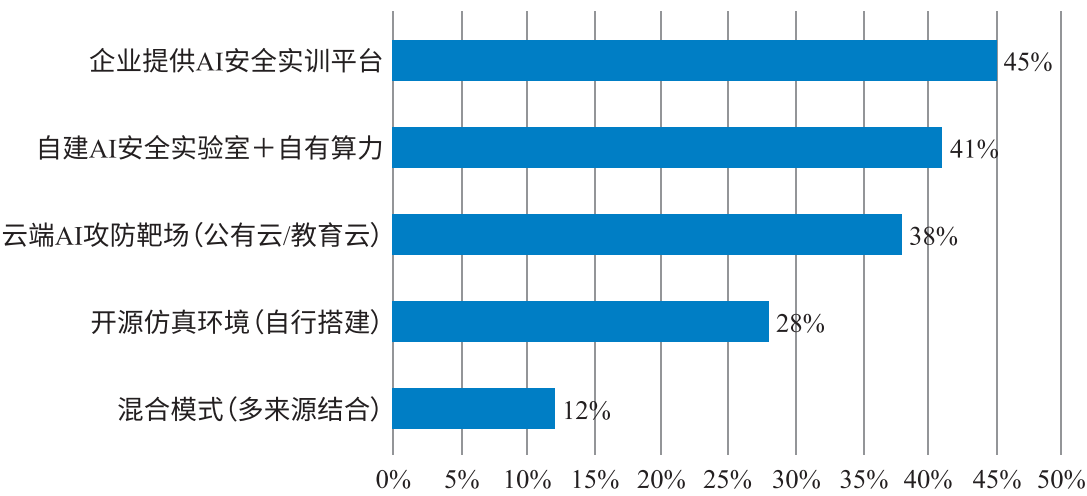


图5-25 AI安全实训平台/环境来源分布

4. 实习实训形式

AI 安全方向实习实训已形成企业顶岗、校企项目和竞赛集训三种主要实践形式。其中，前三类实习形式占比较为接近，分别为 29%、27% 和 26%。与此同时，有 18% 的院校尚未将 AI 安全纳入实习内容，说明 AI 安全专项实习尚未实现全面覆盖。如图 5-26 所示。

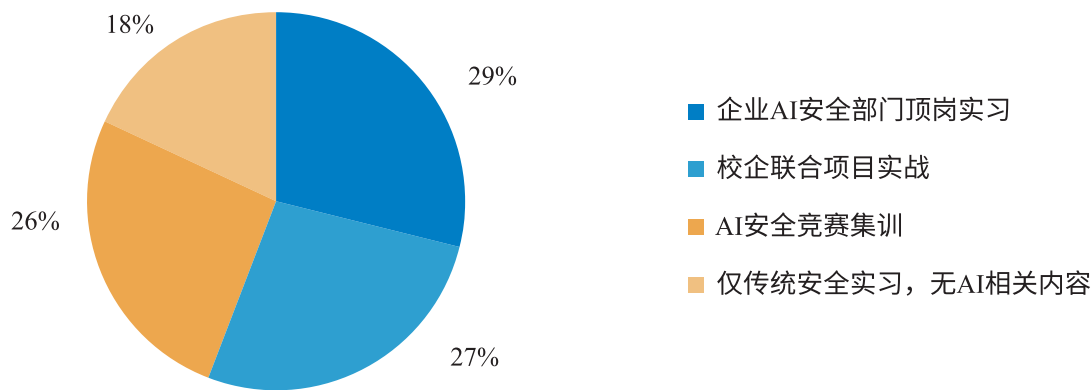


图5-26 AI安全方向实习主要形式

5. 考核评价体系建设

在考核评价方面，不同院校采用的衡量标准差异较为明显。45% 的院校采用项目成果展示，其次分别为企业导师实战(35%)、AI 攻防实操考核(33%)、过程性实训记录 / 积分制(23%)。然而，仍有近两成院校暂无标准化考核方式。如图 5-27 所示。

6. 特色与创新实践

在创新探索方面，院校已形成多种特色做法。排名前五的创新做法依次为：打造“课堂 + 实验室 + 企业”三位一体实训平台(48%)、开设专项 AI 安全实训模块(47%)、构建分

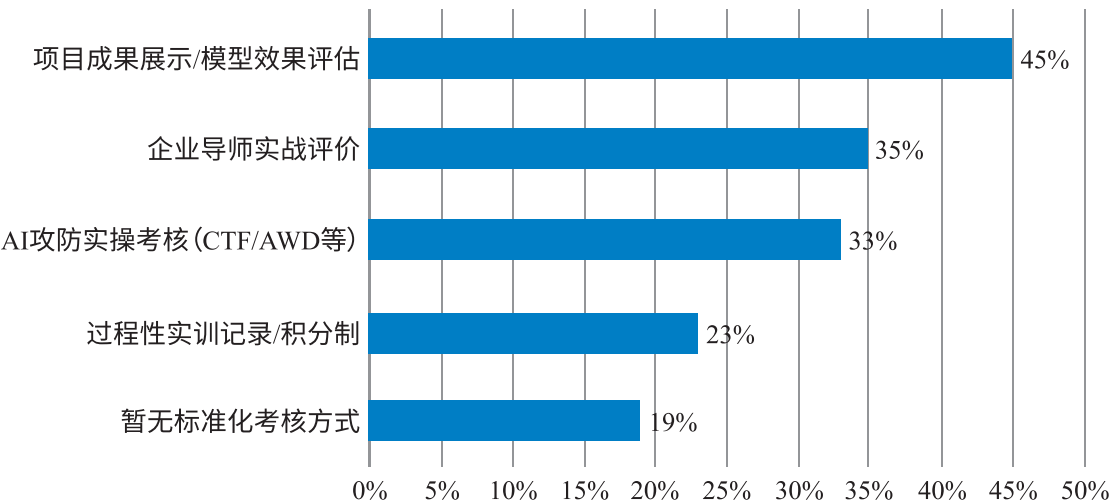


图5-27 AI安全实训考核评价方式

层递进式实训体系(39%)、建立“双师型”指导机制(26%)、建立实战化考核评价机制(22%)。整体来看,创新做法已覆盖平台建设、课程设计、师资配置、考核机制等多个维度。如图 5-28 所示。

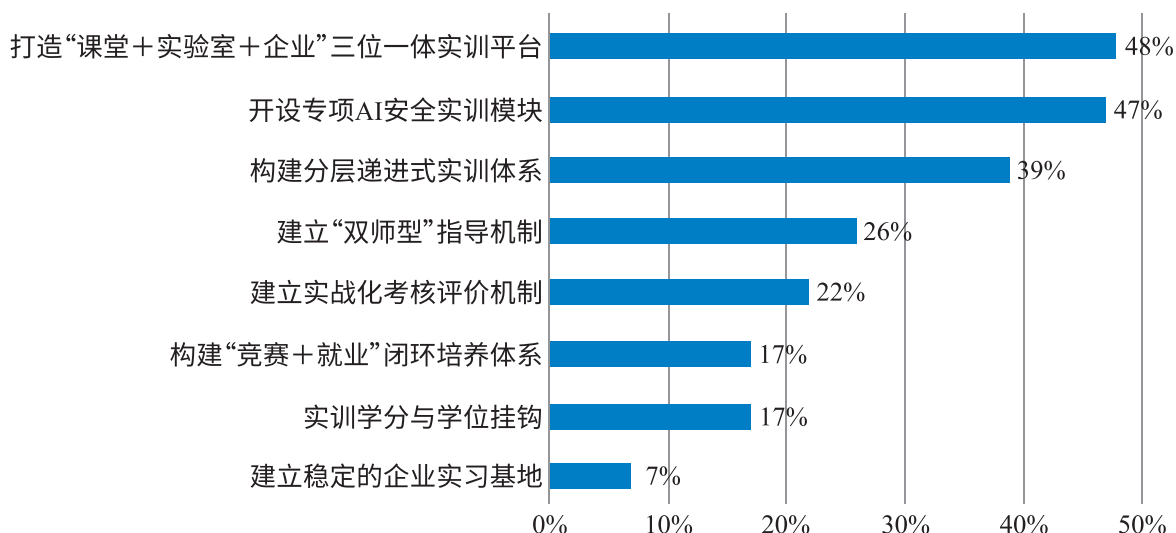


图5-28 AI安全实训体系特色创新点

5.2.4 当前问题与着力方向

综合用人单位与院校两端数据,当前产教融合在 AI 安全人才培养中面临的核心瓶颈可归纳为资源、标准和协同三个层面:

第一,资源瓶颈。37.5%的企业受访者认为“缺乏支撑大型模型实验的算力与环境”是产教融合的首要障碍,32.3%的院校将“算力/平台成本高、投入不足”列为校企合作最大障碍,如图 5-29 所示。仅 26.2% 的高校实验室“具备充足的算力资源”,38.4% 的实验室仅能运行小型模型或简单脚本,15.7% 的实验室完全无法支持 AI 实验,如图 5-30 所示。

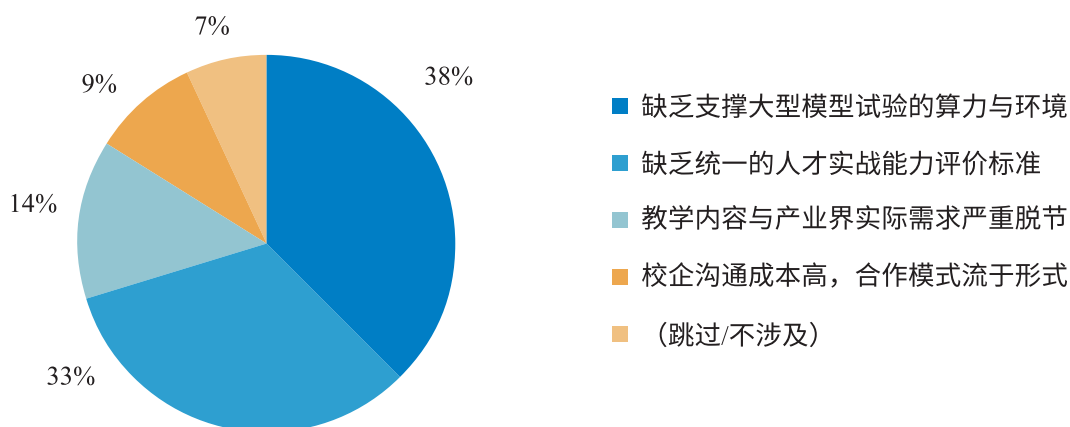


图5-29 产教融合面临的主要瓶颈(企业端)

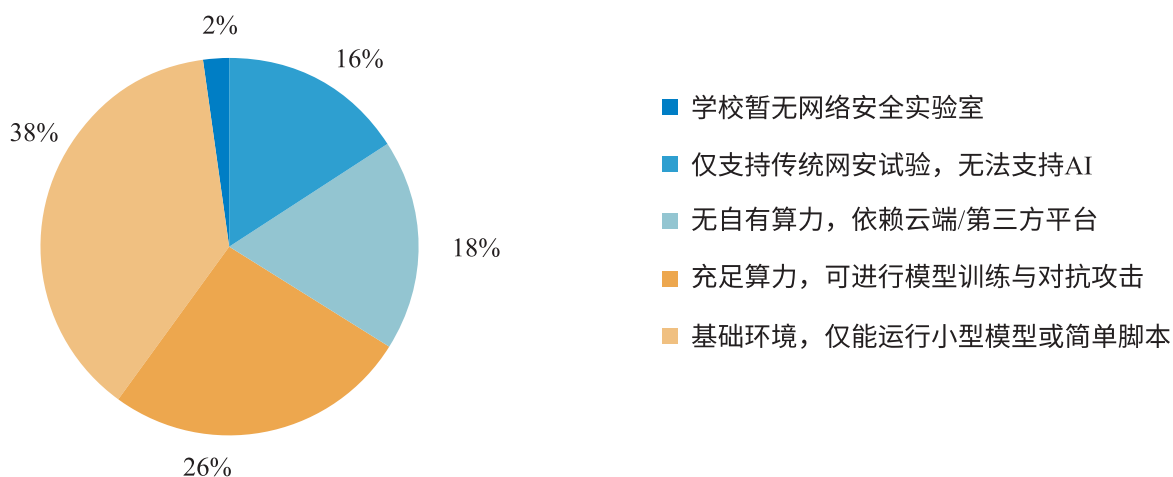


图5-30 院校实验室对AI安全实训的支撑能力

第二，标准瓶颈。33%的企业受访者指出“缺乏统一的人才实战能力评价标准”。传统CTF的考核模式难以有效评估大模型安全、对抗样本攻防、AI应用合规等新兴能力维度，亟需建立适应AI时代特征的实战能力评价框架。

第三，协同瓶颈。73.4%的院校学生参与了各类网络安全竞赛，但43.7%的高校反馈“缺乏专项指导”，30.6%的院校将“缺乏AI+安全的双师型师资”列为校企合作主要障碍。

用人单位同样存在协同难题：48%的受访者认为“算力成本高昂，个人难以搭建实验环境”是人才成长的最大阻碍。AI安全实战能力的培养需要大规模算力支撑，而当前的资源分配机制尚无法有效弥合这一缺口。如图5-31所示。

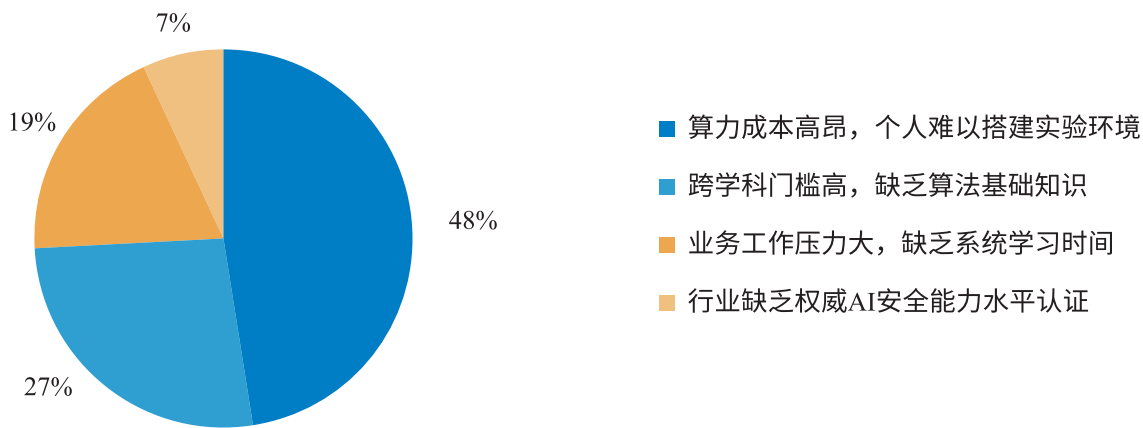


图5-31 阻碍AI能力进阶最大瓶颈(用人单位)

解决这一问题可以从以下三点切入：一是赛事内容的 AI 化升级，从传统 CTF 向大模型安全、智能体安全、多模态安全等新型赛制拓展；二是训练平台的普惠化建设，通过政府引导、企业开放、院校共享的方式降低算力门槛；三是评价体系的标准化建设，建立贯通课程、竞赛、实训和岗位评价的一体化实战能力评价链路。

5.3 在职人员能力提升机制

5.3.1 企业内训与在职培训现状

AI 技术快速演进，推动网络安全从业人员能力结构由传统单一安全能力向安全能力与人工智能能力深度融合的复合型能力转变。在此背景下，企业内训与在职培训已成为网络安全人才培养体系的重要组成部分，培训内容正由传统网络安全知识更新，逐步拓展至模型安全、AI 应用安全和 AI 赋能网络安全等方向。

为提升在职人员的 AI 安全能力，用人单位采取的措施也各有不同。61% 的用人单位采取“采购外部专业培训或认证课程”的方式，占比最高；45% 的用人单位采取“鼓励参与开源项目或安全竞赛”；也有 31% 的用人单位定期开展内部技术沙龙 / 讲座。总体来说，用人单位通常依赖外部力量提升自身人才的 AI 安全能力。如图 5-32 所示。

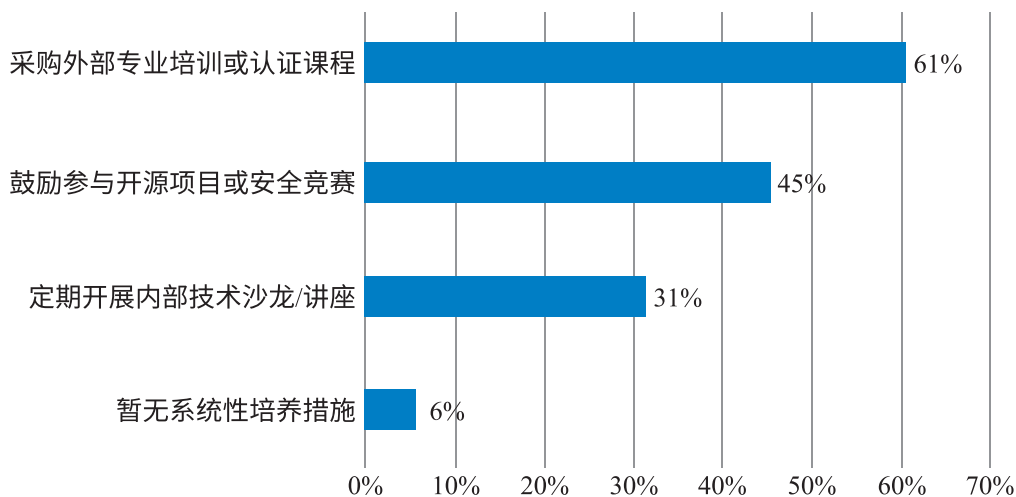


图5-32 企业针对在职人员的AI能力提升措施分布

5.3.2 从业人员继续教育需求分析

网络安全从业人员迫切需要相应的资源，提升个人 AI 安全能力。28% 的从业人员亟需用人单位开展 AI 安全专业培训，提升 AI 安全知识水平；22% 的从业人员亟需用人单位开展或个人参加 AI 竞赛演练，提升 AI 安全技能，完善个人网络安全技能树；20% 的从业

人员亟需用人单位搭建实战靶场平台，利用工作间隙开展实验实训；还有部分从业人员亟需技术交流、算力支持等。如图 5-33 所示。

在提升个人及团队的 AI 安全能力的过程中，网络安全从业人员也面临多重挑战。47% 的网络安全从业者认为“算力成本高昂，个人难以搭建实验环境”是能力提升中最大的瓶颈；27% 的网络安全从业者认为“跨学科门槛高，缺乏算法基础知识”是提升能力中最大的瓶颈；“业务工作压力大”“缺乏权威认证”分别占比 19%、7%。如图 5-34 所示。

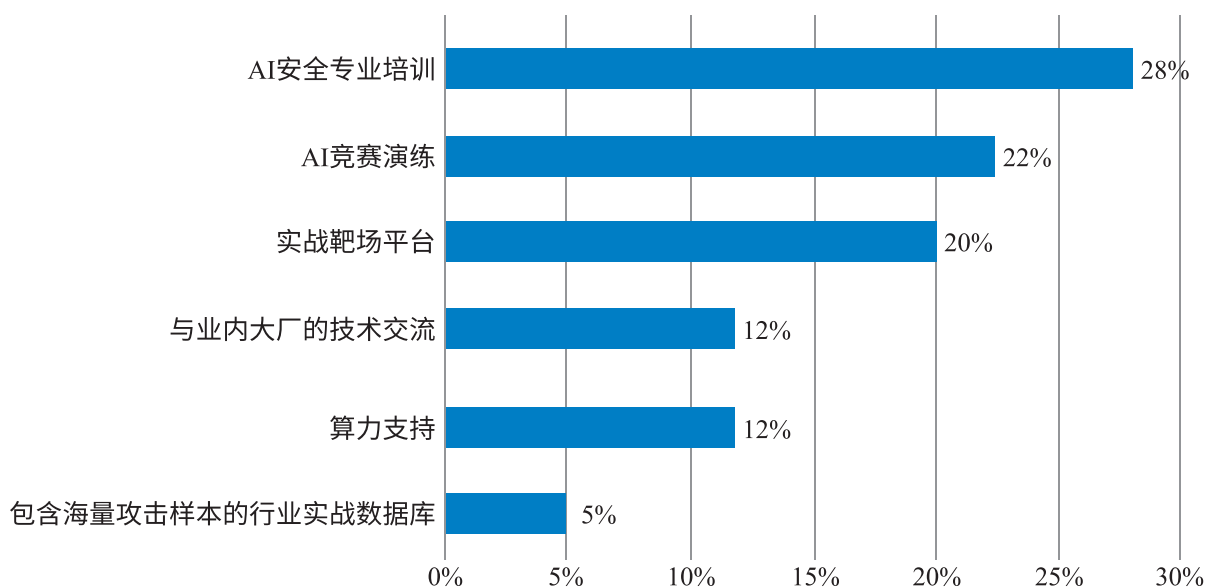


图5-33 用人单位网络安全从业人员亟需资源情况

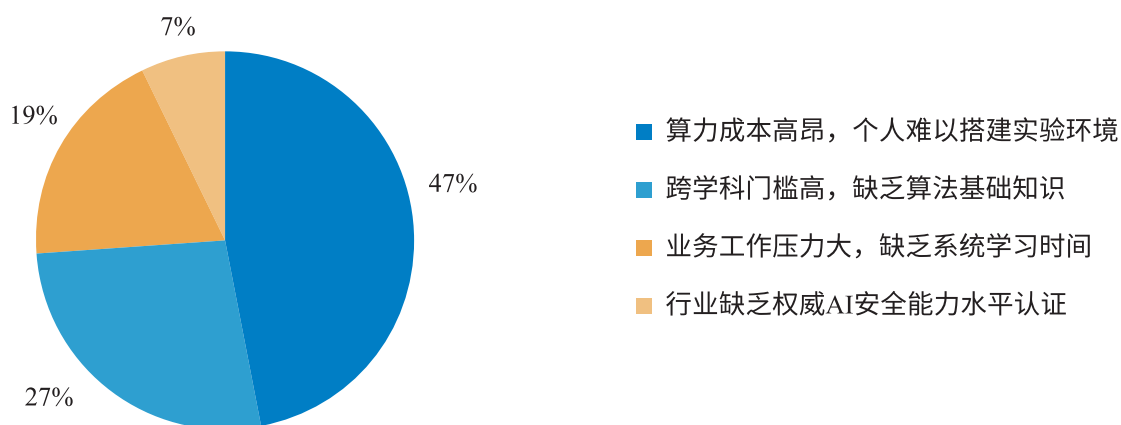


图5-34 网络安全从业者提升AI安全能力最大瓶颈

整体来看，网络安全从业人员迫切需要一套从零起步、覆盖机器学习核心原理、兼顾安全攻防实践的体系化课程，帮助他们跨越“懂安全不懂 AI”的鸿沟；并希望用人单位提供必要的算力支持，让 AI 真正进入安全运营、告警研判、威胁分析等实际工作场景，切实为日常工作提效减负。

然而，在用人单位为提升员工能力所提供的资源中，仅有 5% 的用人单位为员工提供了“私有化大模型 / 算力平台”，甚至有 5% 的用人单位未提供任何支持。提供“专项 AI 实战能技能培训”“商业大模型使用授权”的占比最高，分别为 29%、28%。

这说明，用人单位更倾向于提供培训、工具授权等轻量支持，但私有化模型和算力平台等重投入资源供给不足，而后者恰恰是网络安全从业者开展深度实训的重要支撑，如图 5-35 所示。

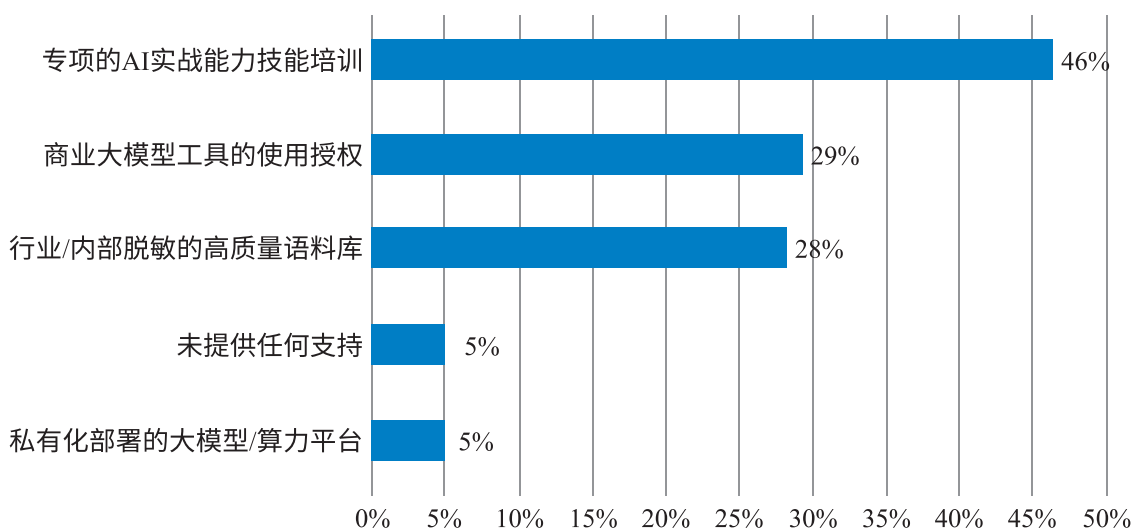


图5-35 用人单位提升人才AI安全能力所提供资源

5.3.3 当前面临的主要问题

随着人工智能技术深度融入网络安全攻防体系，传统

网络安全从业人员在向 AI 安全能力转型与提升过程中，面临四个相互关联的系统性挑战。

1. 知识结构跨代与跨学科学习壁垒

传统网络安全知识体系侧重网络协议、系统漏洞、入

侵检测，而 AI 安全要求掌握机器学习、深度学习原理、模型架构及对抗攻击等全新内容。许多从业者反映算法基础薄弱是深入理解模型安全风险的核心障碍，同时 AI 技术本身的高速迭代加剧了持续学习压力。

2. 理论与实战脱节，缺乏高质量演练环境与工具

即便初步掌握 AI 安全基本概念，将理论知识转化为

实战能力仍是难关。针对 Prompt 注入、模型投毒、接口滥用等新型攻击，市场上尚缺乏成熟的一体化检测与防护平台，学习往往停留在理论层面。

3. 个人与组织层面的资源约束限制深度投入

高昂算力成本使个人难以搭建实验环境，繁重业务工作挤压系统学习时间。企业支持多集中于内部分享和通用外部培训，在定制化实战训练、私有化算力平台建设等方面投入有限，导致学习呈现碎片化、浅层化特征。

4. 外部教育与社会化培养体系支撑不足

高校及培训机构的教学内容与产业实际需求存在代差，且缺乏支撑大型模型实验的软硬件环境。人才培养与评价标准不统一、校企合作模式浮于表面等问题，阻碍了高效、稳定的人才再培养机制的形成。

综上所述，AI 安全人才的成长正受到知识结构跨代、实战环境缺失、资源约束显著、继续教育供给不足等问题的制约。这些问题环环相扣，需要从个人学习路径规划、企业资源投入策略、教育培训机构课程改革以及产业生态协同共建等多个层面进行系统应对。

5.3.4 在职人员能力提升路径

面向 AI 时代网络安全人才能力重构的新要求，用人单位组织在职培训应立足网络安全岗位实际，围绕模型安全、AI 应用安全和 AI 赋能网络安全三类核心能力，未来可重点从以下几个方面推进。

第一，按岗位群建立分层分类培养路径。将培训对象划分为安全运营、威胁分析、漏洞挖掘、应急响应、模型评测等岗位群，并区分初级、中级和高级三个层级，形成岗位有模块、层级有重点、提升有路径的培养体系。

第二，按典型场景设置实战训练单元。传统安全场景设置告警研判、攻击链分析、漏洞复现、代码审计、数据泄露处置等训练；AI 安全场景设置提示注入防护、模型越狱测试、数据投毒检测、对抗样本识别等训练。综合运用案例分析、红蓝对抗、模型红队测试等形式。

第三，按三类核心能力重构培训模块。模型安全模块覆盖模型原理、风险类型、攻击方式和安全评测；AI 应用安全模块覆盖身份鉴别、权限控制、数据安全和隐私保护等；AI 赋能网络安全模块覆盖大模型辅助告警分析、威胁情报提取、代码审计、漏洞验证和自动化处置等。

第四，将培训与演训、认证和岗位评价打通。对关键岗位建立岗位能力清单和内部认证标准，将攻防演练表现、案例处置质量、漏洞验证能力等作为评价依据，与岗位晋升、职级评定适度联动。

第五，依托平台和案例资源提升培训支撑能力。围绕模型安全、AI 应用安全和 AI 赋能安全业务开发岗位化课程，建设涵盖漏洞事件、攻击事件、模型风险事件的案例库，依托攻防靶场、仿真平台和模型测试环境开展演练。

第六，建立企业与高校、科研机构、行业平台协同培养机制。推动与高校联合开发课程、与科研机构联合设置专题研修、与行业平台联合开展实战训练、与安全厂商共建案例库和实验环境。

第七，建立面向技术演进和事件变化的动态更新机制。将新上线 AI 系统、新暴露典型事件作为培训更新触发条件，及时组织专项培训和复盘分析，推动培训内容动态迭代、持续完善。

总体来看，AI 时代企业内训与在职培训提升的重点，不在于简单增加培训频次，而在于围绕网络安全岗位实际，明确培养对象、训练场景、课程模块、评价标准和协同机制，推动在职人员能力更加贴近实战需求、贴近业务场景、贴近技术演进，为 AI 时代网络安全保障提供持续人才支撑。

5.4 总结与建议

AI 技术的快速演进正在深刻重塑网络安全人才培养的底层逻辑。从高校课程体系到产教融合实践，从竞赛演练到在职人员能力提升，AI 安全人才培养已从“要不要做”的阶段进入“怎么做、做多深”的关键时期。本章从高校培养体系、产教融合实践和在职人员能力提升三个维度展开分析，现将核心发现总结如下。

1. 高校课程覆盖已较广，但实训体系与师资短板突出

多数院校已将 AI 安全纳入教学，但完整实训体系尚未普遍建立，过半院校仍处于零散实验或未开展阶段。具备充足算力的院校占比较低，兼具安全与算法经验的复合型教师仍然有限，实操案例库与真实攻击案例是最紧缺的教学资源。

2. 产教融合已成主流，但深度合作机制尚未普遍形成

绝大多数院校已有合作经历，超六成拥有长期稳定合作伙伴。用人单位最认可的模式是深度实习实训和AI 攻防竞赛。但院校侧普遍反映算力成本高、双师型师资缺乏；企业侧

则面临算力不足和评价标准缺失的困扰。

3. 竞赛育人价值获多方认可，但衔接机制待完善

竞赛在从业者需求、企业认可、企业鼓励和院校集训等多个维度均获得较高认同。但完整实训体系的院校占比偏低，部分院校尚未将 AI 安全纳入实习，竞赛仍多停留于精英参与的阶段。

4. 在职人员继续教育需求迫切，资源约束显著

阻碍在职人员能力进阶的首要因素是算力成本高昂，且企业支持以轻量型为主，而重投入的算力平台和专项培训占比偏低，导致从业人员面临知识结构跨代、实战环境缺失、资源约束显著、继续教育供给不足四重困境。

未来建议总结如下：

1. 完善能力标准与评价认证体系

由主管部门牵头，制定统一的网络安全人才 AI 安全能力评价标准，区分岗位群与能力层次，与职业晋升和薪酬待遇挂钩。将竞赛成绩纳入学分认定和人才评价，建设国家级竞赛平台，扩大覆盖面。

2. 加大算力与实训基础设施投入，构建行业共享平台

推动区域性算力共享池和云端实训平台建设，鼓励龙头企业开放算力资源，形成“企业供算力、院校供课程、行业供场景”的共建模式。设立人才培养专项基金，重点支持师资培训、算力补贴和开源社区建设。

3. 强化在职培训的资源保障，建立定期培训机制

企业应将 AI 安全学习纳入员工职业发展规划，提供专项学习与算力预算。建立面向技术演进的动态更新机制，将新系统上线、新风险暴露等作为重新培训的触发条件。

总结与建议

人工智能加速融入经济社会发展和国家安全各领域，正在重塑网络安全的风险形态、攻防方式和人才能力结构。综合调研与分析表明，AI时代网络安全人才建设的关注重点，已由单纯弥补“数量缺口”转向推动“能力重构”：保护对象由网络、主机、应用和数据扩展至模型、算法、接口、智能体等新型对象，并延伸至人机协同流程，工作方式由人工主导逐步转向人机协同，人才能力也由侧重单项技术技能，向技术、业务、治理相融合的综合实战能力演进。由此，人机协同正在成为网络安全实战的基本形态，AI素养正在成为基础能力，“利用AI、保障AI、对抗AI”正在构成新的核心能力要求。

与人才能力边界快速扩展相比，现有人才供给和培养支撑的适配速度仍显不足，成为当前AI时代网络安全人才建设面临的突出矛盾。从实战任务看，模型安全、AI应用安全等新型风险已贯穿智能系统设计、训练、部署、调用和运营等环节，人工智能也持续嵌入安全运营、漏洞挖掘、攻防对抗和安全数据分析等核心工作，对结果核验、复杂推理、工具编排、风险治理和跨域协同提出更高要求；从人才供给看，相关能力建设虽然已经起步，但高阶实战能力、复合型人才和关键新兴岗位人才仍显不足；从培养支撑看，课程体系、复合型师资、算力平台、真实案例和评价标准等基础条件，与快速演进的技术和实战需求之间仍存在一定差距。实质上，当前面临的已不是某一能力、某一岗位或某一培养环节的局部短板，而是能力要求、人才供给与培养体系之间的系统性适配问题。与此同时，这类跨学科、强工程、重实践的问题难以依靠单一主体解决，产教深度融合正成为贯通岗位需求、培养过程与实战评价，促进人才供给与产业需求有效衔接的重要路径。

基于上述判断，AI时代网络安全人才建设需要从能力重构、培养改革和协同生态三个层面统筹推进：在能力层面，以实战需求为牵引，持续提升利用AI、保障AI、对抗AI的综合能力；在培养层面，推动课程、师资、平台、实践和评价协同改革，加快知识学习向实战胜任能力转化；在生态层面，强化国家统筹、教育供给、产业牵引和个人持续提升，形成需求有效传导、优质资源共享、培养实践贯通、人才评价协同的长效机制。基于这一总体逻辑，本章进一步凝练人才实战能力重构、供需结构性差距、培养体系挑战和产教融合等

主要发现，从国家、教育、产业和个人四个层面提出对策建议，并对AI时代网络安全人才能力、岗位、培养模式和发展机制的演进趋势作出系统研判。

6.1 主要发现

6.1.1 人才实战能力内涵发生深刻变化

综合来看，人工智能带来的并非传统网络安全能力的简单叠加，而是人才实战能力边界、工作方式和评价标准的系统重塑。其核心变化可概括为“对象扩展、方式重构、能力复合、评价转向”。

一是安全保护对象由传统信息系统扩展至智能系统全生命周期。训练数据、模型参数、提示词、知识库、插件接口和智能体权限等成为新的安全保护对象，安全治理范围进一步延伸至自动化决策和人机协同流程；风险也由传统漏洞利用、网络入侵和数据泄露，延伸至数据投毒、模型后门、提示词注入、模型越狱、接口滥用、智能体越权和决策失真等问题。网络安全人才因此需要进一步拓展能力边界，强化对数据、模型、应用及业务流程全生命周期风险的识别、研判与治理能力。

二是安全工作方式由人工主导向人机协同加速转变。调研结果显示，从业人员普遍认可AI在提升工作效率方面的作用，海量日志审计、告警降噪和自动化响应等已成为应用较为广泛的典型场景，同时，结果核验也是体现专业技术水平的重要环节。这表明，现阶段AI在规模化、重复性和规则相对明确的任务中优势更为突出，而复杂研判、业务权衡、重大风险处置和责任决策等关键环节，仍需专业人员进行综合研判并作出最终决策。

三是人才能力结构加快向复合化演进。模型安全、应用安全和AI赋能网络安全相互关联，既要求人才理解人工智能的基本机理和脆弱性，也要求其具备系统安全、软件安全、数据安全、攻防对抗等传统基础，并能够统筹研发、业务、管理和合规等多维要求。由此，网络安全人才的能力结构正由侧重单项专业技能，向多领域知识融合、跨主体协同和综合风险治理并重的复合型能力体系演进。

四是人才评价由知识掌握和工具操作转向复杂场景胜任能力。前文各章构建的ASK-P体系表明，意识、知识、技能和实践并非彼此割裂，而应形成从风险认知到实战验证的完整能力链条。AI时代人才评价应更加关注能否在真实或高仿真场景中完成风险识别、测试验证、协同处置和结果复核。总体上，应在夯实传统网络安全能力和AI基础素养的基础上，面向“利用AI、保障AI、对抗AI”三类任务，形成以实战为检验标准的复合能力体系。

总体来看，模型安全、AI应用安全和AI赋能网络安全虽然面向不同对象和任务，但其人才能力要求均体现出由风险意识、交叉知识、专业技能向实战实践递进的共同规律。ASK-P由此构成AI时代网络安全人才培养、训练与评价的一条基本能力主线。

6.1.2 人才能力与实战需求存在结构性差距

综合人才能力、岗位配置与培养实践调研可以看出，我国AI时代网络安全人才队伍建设已经形成一定基础，但能力供给与岗位任务之间仍存在结构性错位，突出表现为基础工具应用与复杂任务处置能力之间、高层次岗位需求与复合型人才供给之间、知识技能积累与真实场景实践之间存在明显差距。

首先，基础工具使用较快普及，高阶能力仍显不足。调研数据显示，从业人员对AI效率提升的满意率总体处于较高水平，日志审计与告警降噪、自动化响应等场景应用相对广泛，但漏洞挖掘、恶意代码分析等需要深度推理的任务应用程度较低；模型安全与AI应用安全自评总体达到基本掌握水平，但深度学习框架安全等前沿方向仍较薄弱。这表明人才能力正在从“不会用”向“会使用”迈进，但距离能够独立完成复杂风险分析、工具组合、结果验证和高影响操作控制仍有差距。

其次，新兴专业岗位需求加快增长，复合型人才供给相对不足。模型安全领域存在岗位建制不完善、复合背景不足、实战能力短板和提升机制薄弱等问题；应用安全领域则仍较多依托既有安全团队兼任，专职团队比例不高，岗位职责和协同机制尚在形成。随着模型安全评测、AI应用安全保障、智能体安全和AI赋能安全运营等任务持续增加，仅依靠传统安全人员或算法人员转岗已难以充分满足相关岗位需求，兼具安全、AI和业务理解能力的复合型人才将成为当前人才供给中的突出短板。

再次，知识认知向实战能力转化仍不充分。部分人才已经了解提示词注入、数据投毒、模型后门等风险，但真实模型、业务数据、攻击样本和复杂环境不足，使其难以完成从风险识别、攻击验证到应急处置和整改复盘的完整任务。同时，算力成本和跨学科知识门槛是能力进阶的重要障碍，不同类型组织之间在平台、资源和人才基础方面也存在差异。由此可见，知识技能能否有效转化为实战能力，不仅取决于个人学习积累，也与算力、平台、真实场景和组织支持等实践条件密切相关。

总体来看，当前AI时代网络安全人才供需矛盾已不再主要表现为单纯的数量不足，而更多体现为能力层级、岗位结构和实践转化等方面的结构性失衡。人才队伍建设需要在保持培养规模的同时，更加突出质量提升和结构优化，重点加强复合型、高层次和关键新兴岗位人才培养，强化真实场景训练和实战能力评价，推动人才供给与技术演进、岗位需

求和安全治理要求更加精准匹配。

6.1.3 人才培养体系面临课程、师资、平台三重挑战

人才培养侧调研表明, AI安全人才培养已由初步探索阶段逐步进入体系化、纵深化建设阶段, 课程、师资和平台成为影响培养质量的三项关键基础条件。

一是课程覆盖已有基础, 但体系化和递进性不足。多数院校已不同程度将AI安全内容纳入人才培养体系, 模型安全、AI应用安全和AI赋能攻防等新内容持续进入课程与实践教学。但课程之间的先修关系、能力层级和实践衔接尚不清晰, 高质量教材、代码、案例和真实攻击样本仍显不足。课程建设的重点已由“增加内容”转向“形成体系”, 既要保持网络、系统、软件、密码和数据安全等传统基础, 又要将AI安全能力有序嵌入人才培养全过程。

二是复合型、实战型师资仍是关键短板。AI安全教学要求教师同时理解网络安全、人工智能和工程实践, 现有师资主要依靠原有学科教师跨界转型, 兼具安全与算法经验的教师数量有限; 企业专家虽具备实践经验, 但参与教学仍存在短期化、碎片化现象。教师能否持续接触真实系统、真实风险和真实项目, 直接影响课程深度和实践质量。

三是实训资源与平台支撑不足。调研显示, 仅10%的院校已建成较为系统完整的AI安全专项实训体系, 仅26%的高校实验室具备较充足算力, 可支持模型训练和对抗攻击等实验。算力、模型、数据、攻击样本以及隔离环境投入较高, 真实案例又受数据安全、商业秘密和知识产权约束, 导致“有课程、缺实战”“有平台、缺内容”的问题较为突出。课程是培养内容的基本载体, 师资是提升教学深度的关键, 平台则是实战能力形成的重要支撑, 三者需要协同建设。

上述课程资源、师资能力和平台条件等约束并不局限于学历教育, 在职人员能力更新同样面临算力成本、跨学科门槛、工学矛盾以及高质量实践资源不足等问题。这表明, AI时代网络安全人才培养需要由入职前教育进一步延伸至职业全周期, 形成持续更新的能力发展机制。

6.1.4 深化产教融合是促进能力提升的关键路径

AI安全具有跨学科、强工程、重实践、快迭代等特征, 仅依靠高校或企业单一主体均难以充分满足人才培养需求。调研显示, 校企合作已成为人才培养的重要方式, 绝大多数院校已有合作经历, 超六成拥有长期稳定合作伙伴; 共建实验室与攻防靶场、企业导师参与教学、联合开发课程、竞赛集训和产学研协同等模式不断发展。

但当前产教合作的深度和稳定性仍不均衡，部分合作仍以专题讲座、设备支持、短期实习或阶段性项目为主，企业真实岗位需求尚未充分转化为课程内容、实训任务和人才评价标准。院校与企业共同反映，算力资源不足、复合型师资短缺、真实案例供给有限、评价标准不完善等问题依然存在，反映出AI时代网络安全人才培养所需的需求信息、技术资源、实践场景和评价能力分布于不同主体之间。单一主体难以独立形成从能力要求定义、教学训练到实战评价的完整培养闭环，产教深度融合因而成为促进人才培养与产业需求有效衔接、提升实战人才培养质量的关键路径。

竞赛、攻防演练、实习实训和联合项目已经成为连接教育培养与产业实践的重要载体。调研表明，企业和院校均对深度实习实训、AI攻防竞赛等实践形式给予较高认可，但现阶段部分活动仍主要覆盖少数骨干学生，竞赛、实习和项目成果与课程评价、学分认定、岗位能力评价及人才选用之间的衔接仍不充分，实践成果尚未有效转化为稳定、可持续的人才培养成效。这表明，当前产教融合已由“是否开展合作”进一步进入“合作能否深入、培养能否贯通、成效能否持续”的阶段，其作用也正在由单项资源补充向贯通教育培养与产业实践的协同机制演进。

6.2 对策建议

面对AI时代网络安全人才能力重构提出的新要求，以及实战能力、资源保障和协同机制等方面存在的突出短板，应加强系统谋划和多主体协同，围绕能力标准、培养体系、实践支撑、评价机制和持续发展等关键环节，构建权责清晰、优势互补、衔接贯通的人才能力建设体系。国家层面重点加强战略统筹、标准引领、资源配置和评价机制建设，强化共性基础支撑；教育层面重点完善人才培养体系和实践教学机制，促进知识、技能向实战能力有效转化；产业层面重点发挥需求牵引和实践主体作用，推动真实场景、技术资源和岗位要求深度融入人才培养；个人层面重点强化AI素养、实战能力和持续学习，不断提升对技术演进和岗位变化的适应能力。通过各主体协同发力，逐步形成国家统筹引导、教育系统深化改革、产业主体深度参与、从业人员主动提升的网络安全人才发展格局。

6.2.1 国家层面：加强顶层设计，完善政策标准

一是完善AI时代网络安全人才能力标准。围绕“利用AI、保障AI、对抗AI”三类任务，统筹模型安全、应用安全、AI赋能安全运营等能力要求，形成与ASK-P体系相衔接的分层分类能力框架，明确不同岗位、不同层次人才的知识基础、实操技能和实践要求。建立动

态更新机制，根据技术演进、重大安全事件和产业需求及时调整能力目录，使高校培养、企业用人和职业评价具有相对统一的参照。加强不同部门、不同行业人才标准、培训认证和能力评价之间的衔接，减少标准不一和重复建设。对于技术迭代快、专业性强的新兴岗位，可通过先行试点、动态评估和逐步推广，不断完善岗位能力要求和评价指标。

二是加强公共实训基础设施和优质资源供给。统筹网络安全靶场、人工智能测试平台和算力资源，推动建设国家级、区域级和行业级共享实训环境，为高校、科研机构、企事业单位提供分级算力、开源模型、脱敏数据、典型案例和测试工具。对高风险攻击技术、敏感数据和模型资源实行分类分级管理，在满足实战训练需要的同时确保安全可控。

三是改革实战人才评价与激励机制。在重视学历背景、专业知识和职业资质等基础指标的同时，更加突出实战技能、项目贡献和综合素养，将攻防演练、事件处置、漏洞发现、工具研发、标准制定和真实项目成果纳入多元评价。针对初级、中级、高级人才设置差异化标准，增强评价结果与岗位任用、职业晋升和人才激励的衔接。

四是强化重点行业和薄弱环节的人才支撑。围绕关键信息基础设施、重要数据和重大AI应用，加强专业人才储备和跨区域协同，通过平台开放、师资共享、专项培训等方式支持资源相对薄弱的地区和单位。同步完善测试授权、数据管理、漏洞披露和责任追溯等规范，将法律合规、科技伦理和职业责任贯穿人才培养和实战训练全过程。

6.2.2 教育层面：优化培养体系，强化实战导向

一是构建分层递进的课程体系。以传统网络安全基础为根基，以AI素养为通用要求，以模型安全、应用安全和AI赋能安全为主要拓展方向，将安全意识、专业知识、实操技能和实践经历贯穿培养全过程。课程设置应形成从基础认知、专项技能到综合实训和创新实践的递进关系，避免简单叠加热点课程。

二是强化场景化实践教学。面向AI时代网络安全典型风险和实战任务，围绕模型越狱、提示词注入、数据投毒、接口与插件安全，以及AI辅助告警研判、漏洞分析和复杂事件响应等场景，推动课程实验、综合实训、竞赛演练、企业实习和科研项目有机衔接，构建由专项训练向综合实践递进的实践教学体系。完善以实际问题解决能力为导向的评价机制，由侧重单项操作结果向综合考察问题识别、分析研判、方案实施、结果验证、风险控制和复盘改进能力转变，推动专业知识、技术技能与复杂场景实践深度融合，切实提升人才实战胜任能力。

三是加强复合型师资和教学资源建设。建立教师跨学科培训和企业实践机制，支持网络安全教师提升AI能力、人工智能教师补强安全能力；完善产业导师和校企双导师制度，

引导企业专家深度参与课程、项目和评价。同步建设教材、案例库、代码库、攻击样本库等资源，并建立快速更新和质量审核机制。

四是完善人才培养评价和继续教育衔接。综合采用知识测评、实验操作、项目成果、竞赛表现和企业评价等方式，突出真实任务中的分析判断、团队协作和规范操作。面向在职人员建设模块化、可持续的课程和专项实训，推动学历教育、职业培训、项目实践和能力评价有机衔接。推动课程实验、竞赛演练、企业实习和毕业设计等培养环节围绕统一能力主线贯通衔接，人才培养评价由关注课程设置和教学投入进一步转向关注能力形成和培养实效。

6.2.3 产业层面:深化产教融合,发挥主体作用

一是发挥企业需求牵引作用。企业应结合业务场景和安全责任明确AI时代网络安全岗位群及典型任务，及时将模型安全、应用安全、智能体安全和AI赋能安全运营等岗位需求反馈教育端，为课程设置、实训内容和人才评价提供真实依据。建立岗位需求动态反馈机制，将新系统应用、重大风险、实战演练和岗位能力变化及时反馈至人才培养环节，推动课程、实训和评价内容持续更新。

二是有序开放实践资源。对典型安全事件、模型漏洞、业务流程和攻击样本进行合规脱敏，转化为课程案例、实验任务、竞赛题目和联合科研项目；有条件的企业可开放算力、模型接口、测试工具和仿真环境，提高真实场景资源的育人价值。

三是建立常态化联合培养机制。推动企业由提供设备、讲座和短期实习，转向参与培养方案、课程开发、项目指导、实习实训和质量评价全过程。通过高校教师与企业导师共同指导，使学生完整经历需求分析、技术实施、风险控制、成果交付和复盘改进，提升工程实践与协同能力。

四是完善企业内部培养和人机协同治理。将AI安全学习纳入岗位培训和职业发展规划，通过项目历练、攻防演练、导师带教和复盘评估提升在职人员能力；明确AI工具使用范围、人工复核要求和高影响操作审批机制，推动AI增效与安全可控同步实现。

6.2.4 个人层面:提升AI素养,主动适应变革

人工智能正在重塑网络安全岗位的知识结构和工作方式。网络安全从业人员应增强主动学习和自我革新意识，将提升人工智能素养、强化实战能力和完善复合知识结构作为职业发展的长期任务。

一是夯实网络安全基础并系统提升AI素养。人工智能工具可以提升效率，但不能替代对系统、网络、程序、漏洞和密码机制的深入理解。从业人员应掌握大模型、智能体、检

索增强生成等基本机理和主要风险，形成正确判断AI能力边界的基础。

二是提升人机协同与可信应用能力。从业人员应由简单调用通用AI工具，进一步转向面向具体任务开展提示设计、工具组合和流程编排，提高人工智能与专业安全工具协同应用水平。同时，强化对模型输出的分析验证，通过日志、流量、代码、威胁情报和人工复现等多源信息进行交叉核验，准确识别模型输出中的事实偏差、逻辑错误和风险遗漏。对于复杂研判、重大风险处置和高风险操作，应明确人工复核和决策责任，合理把握自动化应用边界，确保人工智能应用过程可控、结果可信、责任清晰。

三是坚持在实践中提升综合能力。通过攻防竞赛、靶场训练、漏洞研究、应急演练和真实项目，将知识转化为风险分析、技术处置、团队协同和复盘改进能力，并不断提升与研发、业务和管理人员的沟通协同水平。

四是强化职业伦理和持续学习。依法依规使用攻击测试工具和生成式人工智能，不突破授权边界，严格保护数据、模型和业务信息；持续跟踪新模型、新架构、新风险和新标准，通过案例复盘和知识沉淀保持能力更新。

6.3 AI时代网络安全人才发展趋势

展望未来，人工智能与网络安全将进一步深度融合，人才能力、培养模式和组织形态将持续演进。综合调研结果与技术演进趋势，可重点关注以下六个方面。

一是AI素养将由专业加分项转变为网络安全人才的通用基础能力。随着AI进入安全运营、漏洞分析、代码审计和应急响应等日常工作，能否规范、高效、可信地使用AI，将逐步成为网络安全从业人员的基础能力要求，而不再是少数专业岗位的附加技能。

二是人机协同将成为网络安全实战的基本工作形态。AI将更多承担海量数据处理、重复任务和初步分析，人类重点负责目标设定、复杂研判、业务权衡和责任决策。随着安全智能体能力增强，任务编排、权限管理、运行监控和异常处置将成为新的基础技能。

三是“利用AI、保障AI、对抗AI”将逐步从相对独立的能力方向走向融合。未来同一安全任务往往既要利用AI提升效率，又要保障所使用模型和智能体可信，还要抵御攻击者利用AI实施的新型攻击，三类能力将在同一业务流程中相互支撑。

四是人才岗位将进一步专业化、复合化和场景化。模型安全评测、AI应用安全、智能体安全、AI安全运营等岗位将逐步细化；金融、能源、交通、医疗、政务和工业控制等行业将更加需要既懂安全和AI、又熟悉业务流程的场景型人才。

五是培养与评价将更加场景化、动态化。实训平台将从单一攻防靶场向融合模型测试、应用评估、智能体对抗和业务仿真的综合环境演进，人才评价将更加重视真实任务表现、任务完成过程、结果质量和项目贡献，课程和能力标准也需要随技术与风险变化持续更新。

六是终身学习和多主体协同将成为人才发展的基础机制。技术迭代将持续缩短知识有效周期，学历教育、企业培训、职业认证、在线学习和实战演练需要进一步贯通；政府、高校、科研机构和企业也需在标准、师资、算力、案例和平台方面形成更稳定的资源共享与协同育人格局。

总体来看，AI时代网络安全人才建设将呈现能力复合化、培养实战化、协同常态化、评价多元化和学习终身化的特征。应坚持以国家战略需求为牵引，以实战能力为核心，以产教融合为关键路径，持续提高人才自主培养质量，加快形成与人工智能发展水平和网络安全保障需求相适应的高水平人才队伍，为推进网络强国、数字中国建设筑牢人才基础。

