CAICT 中国信通院

集智 蓝皮书 No.202510

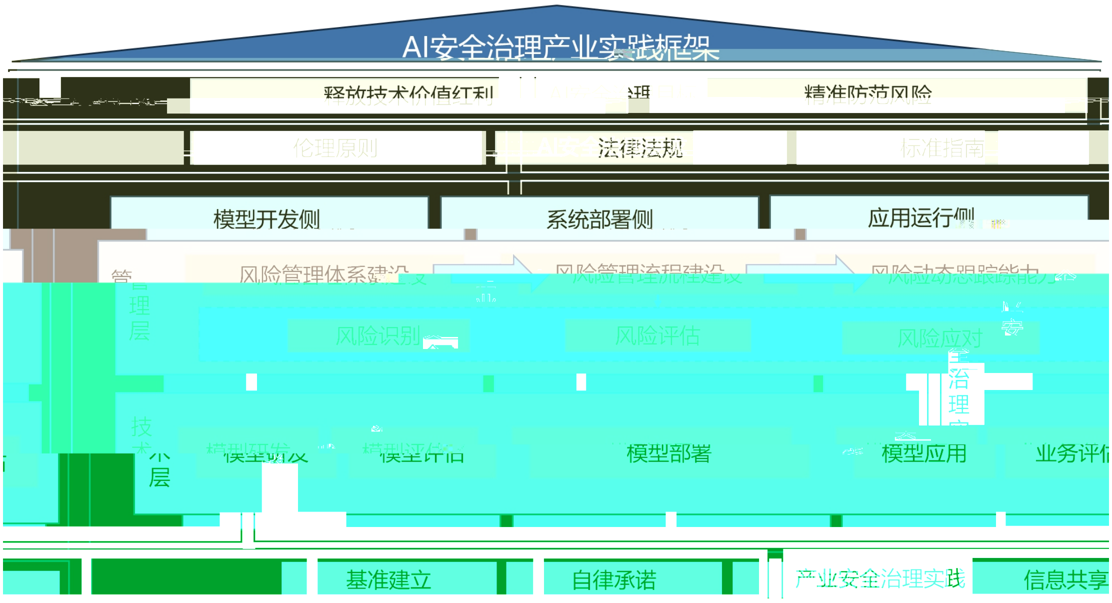
中国信息通信研究院
2026年1月

版权声明

（一）宏观规范要求迈向务实新阶段

（二）技术应用发展凸显风险新态势

（三）产业实践需要安全治理新框架



（一）国际合作层面，全球加强安全治理深化交流

（二）监管政策层面，各国推动安全治理体系化落地

（三）产业实践层面，构建安全治理务实举措成为核心议题

（一）技术发展扩大内生安全新敞口

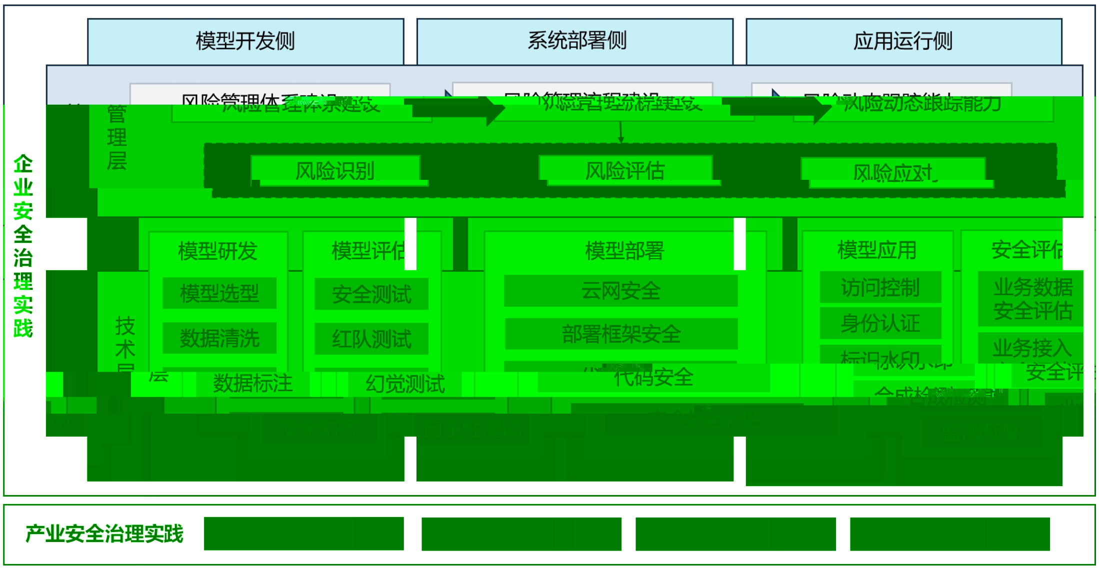
模型在数学推理、代码生成等创造性任务上的突破，往往伴随着事实准确性的下降，形成能力跃升与幻觉率攀升的“双刃剑”效应。从技术原理来看，当前大语言模型的创造力本质上是基于海量训练数据的模式挖掘与概率性生成，而非真正意义上的理解与思考，模型幻觉问题难以根除。

（二）应用延展引发衍生安全新难题

参见国家网络安全通报中心：《大模型工具 存在安全风险，情况通报》，载北京日报，

（三）组织管理体系构建面临新卡点

（四）多元共治协同机制尚待健全完善



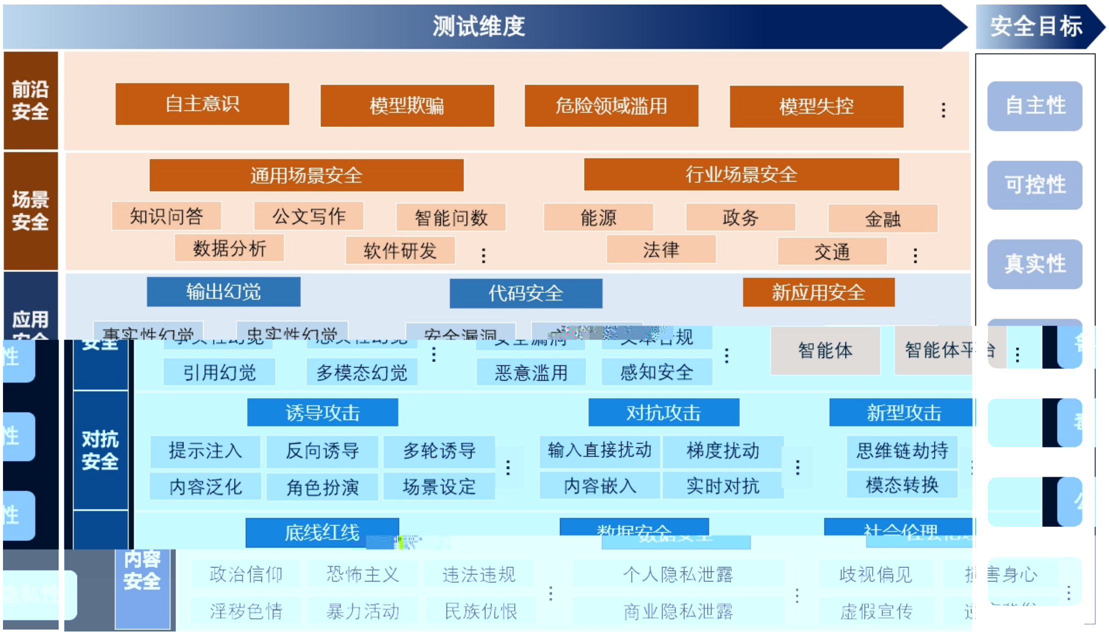
（一）风险管理：构建闭环的人工智能风险管理体系

（二）模型研发：筑牢人工智能开发安全源头根基

（三）系统部署：构建人工智能部署安全防护屏障

（四）应用运行：强化人工智能应用安全动态评估

（五）产业生态：共建基准测试体系与协同治理机制


测试维度
安全目标
前沿安全
自主意识
模型欺骗
危险领域滥用
模型失控
自主性
场景安全
通用场景安全
行业场景安全
知识问答
公文写作
智能问数
能源
政务
金融
数据分析
软件研发
法律
交通
可控性
应用安全
输出幻觉
代码安全
新应用安全
事实性幻觉
忠实性幻觉
安全漏洞
引用幻觉
多模态幻觉
恶意滥用
感知安全
智能体
智能体平台
真实性
对抗安全
诱导攻击
对抗攻击
新型攻击
提示注入
反向诱导
多轮诱导
输入直接扰动
梯度扰动
思维链劫持
内容泛化
角色扮演
场景设定
内容嵌入
实时对抗
模态转换
内容安全
底线红线
数据安全
社会伦理
政治信仰
恐怖主义
违法违规
个人隐私泄露
歧视偏见
淫秽色情
暴力活动
民族仇恨
商业隐私泄露
虚假宣传

《人工智能安全承诺》实践披露

Disclosure of Practices on the Artificial Intelligence Security and Safety Commitments

关于《人工智能安全承诺》

- 2024年12月，中国人工智能产业发展联盟（AIIA）发布《人工智能安全承诺》（简称《承诺》），首批17家领军企业签署，展现守护AI安全、推动智能向善的产业共识。
- 2025年2月，AIIA在巴黎人工智能行动峰会上宣介《承诺》，获国际积极反响。同时，AIIA发起自律披露行动，倡议签署企业披露安全实践。
- 2025年7月，[illegible]安全、可靠、可控、可持续[illegible]

22家 签署企业

18家 披露企业

参与《承诺》企业名单（按单位首字母排序）

签署企业（22家）　披露企业（18家）

（一）金融行业打造人工智能风险管控方案

参见《中国互联网金融协会金融科技发展与研究专委会、中国信息通信研究院人工智能研究所：《金融大模型合规应用研究报告》。

（二）医疗行业多维度提升模型输出准确性

（三）交通行业结合应用场景探索安全方案

自动驾驶是人工智能在交通领域中，最大的智能原生应用场景之一。为

。

（四）能源行业加强基础设施的安全防护

（五）通信行业加强“以技治技”的安全治理体系