大模型安全研究报告

2024 FOUNDATION MODEL SAFETY RESEARCH REPORT

阿里云 aliyun.com CAICT 中国信通院

COVER GENERATED BY: WANXIANG 2.0

阿里云计算有限公司
中国信息通信研究院安全研究所

2024年9月

版权声明

LEGAL NOTICE

阿里云计算有限公司与中国信息通信研究院共同拥有本报告的版权，并依法享有版权保护。任何个人或机构在转载、摘录或以其他形式使用本报告的文字内容及观点时，必须明确标注“资料来源：阿里云计算有限公司与中国信息通信研究院”。对于任何未经授权的转载或使用行为，我们将依法追究其法律责任。

前言

FORWORD

当前，由 ChatGPT 引发的全球大模型技术竞赛正推动人工智能由专用弱智能向通用强智能迈进，这不仅标志着智能水平的显著提升，也预示着人机交互方式和应用研发模式的重大变革。大模型在各行各业的广泛应用，为第四次工业革命的爆发提供了蓬勃动力和创新潜力。

然而，随着大模型商业化应用和产业化落地加速，大模型技术局限和恶意使用不仅加剧了原有人工智能安全风险，也引入了模型“幻觉”、指令注入攻击、网络攻击平民化等新型风险。面对这些挑战，国际组织和世界主要国家正通过制定治理原则、完善法律法规、研制技术标准等方式，积极开展大模型安全治理。同时，大模型在逻辑推理、任务编排等方面的卓越能力，为解决网络空间安全瓶颈问题带来了新的机遇。

为有效防范和消减大模型的安全风险，并促进其在安全领域的应用，阿里云计算有限公司联合中国信息通信研究院等三十余家行业单位共同编制《大模型安全研究报告（2024 年）》。本报告凝聚业界专家共识，聚焦当前大模型突出安全风险和网络空间安全瓶颈问题，从大模型自身安全和大模型赋能安全两个维度，提出涵盖安全目标、安全属性、保护对象、安全措施四个方面的大模型自身安全框架，以及大模型赋能安全框架。期待这些框架能为社会各方提供有益参考，共同推动大模型技术产业的健康发展。

目录

大模型安全概述

大模型自身安全

大模型赋能安全

大模型安全展望

编制说明

大模型安全概述

1. 大模型技术演进

2. 大模型面临严峻安全挑战

3. 大模型带来新安全机遇

4. 大模型安全研究范围

1

1. 大模型技术演进

2012 年，杰弗里·辛顿（Geoffrey Hinton）课题组提出的卷积深度神经网络 AlexNet 在计算机视觉权威比赛 ImageNet 中以压倒性优势获得第一名，拉开了全球深度神经网络研究浪潮。2020 年，OpenAI 推出了 GPT-3，标志着以“标注数据监督学习”和服务特定任务为特点的小规模深度神经网络（即小模型），正式向以“大规模数据集无监督预训练 + 有监督微调”和服务多任务的大规模预训练深度神经网络（即大模型）转变。大模型以其庞大的无标注训练数据、巨大的模型参数、智能“涌现”现象和多任务处理能力，被业界认为是实现通用智能的可行路径。整体看，从小模型向大模型的演进经历了如下三个时期。

1.1 探索期：预训练语言模型（2017 年 -2021 年）

2017 年，谷歌提出了基于自注意力机制的深度神经网络结构——Transformer，奠定了此后大模型发展的算法架构基础。2018 年，基于 Transformer 的 GPT-1 和 BERT 的成功应用，标志着预训练模型成为自然语言处理领域的主流。2020 年，OpenAI 推出了模型参数规模高达 1750 亿的 GPT-3，因其在多类语言任务上的性能大幅提升获得了广泛关注和认可。这个阶段，预训练语言模型在多任务领域内生成语义连贯的类人文本方面展现出了极强潜力，全球为不断提高大模型性能不遗余力扩大模型的参数规模。

1.2 爆发期：语言大模型（2022 年 -2023 年）

2022 年末，OpenAI 发布的 ChatGPT 引爆了全球大模型技术竞赛。此后，谷歌的 PaLM、Meta 的 LLaMA、Anthropic 的 Claude、阿联酋技术创新研究所的 Falcon 和 NOOR、阿里云的通义千问、百度的文心一言等语言大模型争相发布，全球呈现“千模大战”态势。这个阶段，大模型拥有了对自然语言的理解、生成、记忆和推理能力，实现了与人类的顺畅交流。与此同时，全球开始对大模型的经济性和安全性给予更多关注，研究焦点正从单纯扩大模型参数规模和提升模型智能水平，转向追求模型参数效率和确保模型与人类价值观的一致性。

1.3 提升期：多模态大模型（2024- 至今）

2024 年，OpenAI 发布的 Sora 和 GPT-4o 凭借强大的视频语义理解和高质量的文生视频能力震惊全球，开启了全球多模态大模型研发和应用热潮。谷歌的 Gemini Ultra、阿里云的 Qwen-VL Max、百度的 Ernie-ViLG 3.0、

华为云的 MindSpore 等多模态大模型快速涌现，进一步推动了这一领域发展。区别于语言大模型，多模态大模型能同时处理来自语言、图像、声音等不同感知通道的信息，极大提高了场景理解准确度，促使大模型初步拥有了类似人类的感知和理解物理世界的能力。

此外，得益于大模型强大的泛化、自适应和持续学习能力，研究人员在语言、多模态等基础大模型之上，通过使用行业专有数据进行微调，形成适用于金融、医疗、交通等特定行业和任务场景的定制化大模型。基础大模型的智能和安全水平，是影响面向特定行业和任务场景的定制化大模型性能表现的关键因素。

2. 大模型面临严峻安全挑战

随着各类大模型与经济社会的深度融合，其技术局限和潜在恶意使用不仅威胁大模型系统自身的安全稳定运行，也可能为使用大模型的各行各业带来非预期安全影响。

为尽可能全面应对大模型领域的基础共性安全挑战，本报告优先对语言、多模态等各类基础大模型系统的安全风险进行系统梳理。与此同时，参考 ISO/IEC 5338-2023《人工智能系统生命周期过程》国际标准，将基础大模型系统抽象为训练数据、算法模型、系统平台和业务应用四个重要组成部分，并通过描绘这四个组成部分面临的重要和一般安全风险，形成大模型安全风险地图，如图 1 所示。其中，重要风险是发生概率高和影响程度大的风险，一般风险则反之。

2.1 训练数据安全风险

在训练数据部分可能存在训练数据泄露等一般风险，其重点风险包括：

（1）训练数据违规获取：通过不正当手段或未经授权的方式获取训练数据，可能违反法律法规、数据质量受损和发生安全事故。

（2）训练数据含有违法不良信息：训练数据中可能包含违法不良、涉及商业机密或个人隐私等信息。

（3）训练数据投毒：攻击者可能在训练数据中植入恶意样本或对数据进行恶意修改，影响模型的准确性和安全性。

（4）训练数据质量低下：训练数据集中可能存在错误或噪声数据，影响模型训练的效果。

（5）训练数据缺乏多样性：数据来源、特征和分布可能过于单一，不能全面覆盖各种实际应用场景。

图1 大模型安全风险地图

2.2 算法模型安全风险

在算法模型部分可能存在测试验证不充分等一般风险，其重点风险包括：

（1）模型鲁棒性不足：主要体现在分布外鲁棒性不足和对抗鲁棒性不足两个方面。分布外鲁棒性不足主要指模型在遭遇实际运行环境中的小概率异常场景时，未能展现出预期的泛化能力，从而生成非预期的结果。而对抗鲁棒性不足则主要指模型面对攻击者利用精心设计的提示词或通过添加细微干扰来构造对抗样本输入时，模

型可能无法准确识别，影响输出的准确性。

（2）模型“幻觉”现象：模型在回答用户问题时，可能产生看似合理但包含不准确、虚构或违背事实的信息，这种现象被称为模型“幻觉”。

（3）模型偏见和歧视：模型在处理数据时可能表现出某种偏好或倾向，这可能导致不公平的判断或生成带有歧视性的信息。

（4）模型可解释性差：模型的决策过程和结果难以被详细准确地解释，使得用户难以理解模型输入如何影响输出，以及模型产生特定结果的原因。

2.3 系统平台安全风险

在系统平台部分可能遭受非授权访问和非授权使用等一般风险，其重点风险包括：

（1）机器学习框架安全隐患：流行的机器学习框架（如 TensorFlow、Caffe、Torch）可能存在漏洞，攻击者可能利用这些漏洞发起攻击，造成系统受损、数据泄露或服务中断。

（2）开发工具链安全风险：大模型系统开发过程中使用的工具链（如 Langchain、Llama-Index、pandas-ai）可能存在安全漏洞，例如 SQL 注入、代码执行或命令注入等，攻击者利用这些漏洞可能导致数据损坏、信息泄露或服务中断。

（3）系统逻辑缺陷风险：大模型系统可能存在数据权限和隔离、访问控制、业务逻辑等方面的缺陷，这些缺陷可能使得系统容易受到未授权访问、API 滥用、数据窃取或滥用、越权访问等攻击，进而可能导致法律纠纷和监管处罚。

（4）插件相关安全风险：大模型的插件可能存在缺陷，在与大模型服务交互时可能引发敏感信息泄露、提示词注入、跨插件请求伪造等安全问题，这可能导致系统遭受攻击、数据泄露或服务中断。

2.4 业务应用安全风险

在业务应用部分可能存在测试验证数据更新不及时等一般风险。其重点风险包括：

（1）生成违法不良信息：大模型可能产生包含对国家安全、公共安全、伦理道德和行业规范构成威胁的内容。

（2）数据泄露问题：存在攻击者通过逆向工程、成员推理攻击或提示词注入等手段窃取训练数据的风险，这些数据可能包含敏感的个人信息或商业机密，可能导致隐私泄露、知识产权侵权和经济损失。此外，用户在与大模型互动时，也可能由于疏忽或不熟悉相关风险，无意中泄露自己的隐私或保密信息。

（3）用户恶意使用风险：在大模型应用中，存在一些用户或实体不遵守道德规范和法律法规，使用模型进行恶意活动的风险。

总体来说，大模型在人工智能的发展中引入了模型“幻觉”、提示注入攻击、大模型插件缺陷等新风险，并加剧了数据泄露、模型偏见、系统缺陷等传统人工智能技术已有风险。

3. 大模型带来新安全机遇

当前网络空间安全面临攻击隐蔽难发现、数据泄露风险高和违法信息审核难等挑战。大模型展现出强大的信息理解、知识抽取、意图和任务编排等能力，为网络空间安全瓶颈问题提供了新的解决思路和方法。与此同时，大模型发展也催生了恶意软件自动生成、深度伪造等新型攻击方式，已有安全措施无法有效检测和防御，亟待利用大模型技术创新保护机制抵御新型威胁。

大模型可显著提升网络威胁识别防御响应的精准度和时效性。在威胁识别阶段，大模型通过整合威胁情报、挖掘零日漏洞、执行代码审计和网络攻击溯源，可有效识别系统业务风险，提供针对性防御措施。在安全防御阶段，大模型通过对安全策略进行动态推荐与调整，强化安全防御效果。在安全检测阶段，大模型通过告警分析、报文检测、钓鱼邮件识别和未知威胁检测，深度识别攻击意图，研判攻击样本，提升攻击识别准确度。在安全响应阶段，大模型针对实际攻击行为提供自动化响应策略与处置流程，并撰写事件分析报告。在安全恢复阶段，基于运营目标执行全面的模拟演练，为安全恢复提供最佳实践指导。

大模型能有效提升数据安全技术的普适性和易用性。大模型通过深度学习和自然语言处理技术，能够理解和分析复杂的非结构化数据，自动识别并提取关键特征。这种智能分析能力极大地降低了对人工分析的依赖，使得数据分类分级等数据安全技术更加易于普及。同时，大模型的自学习能力意味着它可以不断从新的数据安全标准及样例集中学习并优化其安全策略，无需频繁的人工干预。这不仅提高数据安全的适应度和响应速度，也提高了数据安全技术的易用性。

大模型能有效提升内容安全技术的鲁棒性和准确性。与传统的内容安全技术相比，大模型在多模态数据处理上具有显著优势，正在成为提升内容安全技术鲁棒性和准确性的关键。大模型的鲁棒性体现在其对不同格式、风格和质量的内容均能保持稳定的检测性能，即使面对攻击者采用隐喻、漫画等形式刻意规避检测手段，也能保持较高的识别率。大模型的自学习能力，使其能够不断从新数据中学习，适应不断变化的虚假信息、深度伪造（Deepfake）等网络威胁，从而提高检测的准确性。

4. 大模型安全研究范围

大模型安全由大模型自身安全及大模型赋能安全两个核心要素构成。前者包含大模型安全目标、安全属性，保护对象及安全措施四个方面，后者则为发挥大模型对网络空间安全的赋能作用提供技术方向指引。

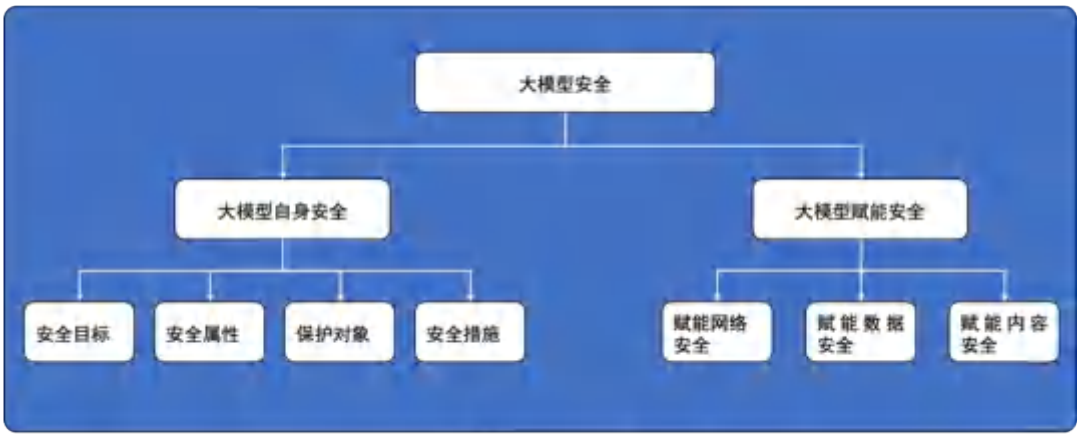

图2 大模型安全研究范围

4.1 大模型自身安全

大模型自身安全是指在训练数据、算法模型、系统平台、业务应用这四个重要层面执行安全措施，以确保模型的安全、可靠、可控，并保障其伦理性、合规性、可靠性、可控性、鲁棒性等安全属性。同时，对大模型的系统、数据、用户、行为四个对象进行严格保护，确保大模型系统提供服务时的安全性。

4.2 大模型赋能安全

大模型赋能安全是指在网络安全、内容安全、数据安全等领域，利用大模型的信息处理、知识抽取、意图识别等能力，增强网络安全防御能力、数据安全保护能力、内容安全检测过滤能力，提高安全事件处理的效率和准确性，提升安全技术的智能化水平，促使安全防护更加主动、智能和高效。

大模型
自身安全

1. 大模型自身安全框架

2. 训练数据安全措施

3. 算法模型安全措施

.4. 系统平台安全措施

2

1. 大模型自身安全框架

大模型赋

图3 大模型自

能安全


系统平台安全稳定
业务应用安全可控
准确性
机密性
可问责任
公平性
透明性
可解释性
可控性
鲁棒性
用户
行为
现
数据集安全检测
数据增广与数据合成
安全对齐数据集构建
模型"幻觉"级解
模型偏见缓解
模型可解释性提升
生成信息标识
用户协议和瑰私致策

自身安全框架

本报告从消减大模型面临的基础共性安全风险出发，构建如图 3 所示大模型自身安全框架。大模型自身安全框架涵盖安全目标、安全属性、保护对象、安全措施四个层面，这四个层面基于自顶向下、层层递进的方式提出了构建大模型自身安全的整体解决方法。

1.1 安全目标

目前，欧盟、美国、中国等世界主要地区和国家以及微软、谷歌等科技企业均提出大模型伦理准则。本安全框架在充分借鉴国内外大模型伦理准则要求的基础上，在我国战略层面提出的“确保大模型安全、可靠、可控”总体目标基础上，基于大模型面临的安全风险和挑战，根据大模型应用的实际需要，提出以下四个方面安全目标。

训练数据安全可信：训练数据是大模型的基石，大模型从训练数据中汲取知识的同时，也面临着数据泄露、数据偏见、数据投毒等诸多安全隐患。因而，应确保大模型的训练数据不被窃取，不会泄露用户隐私，且未被篡改，能够真实反映物理世界和人类社会情况。

算法模型安全可靠：大模型技术正逐渐应用于智慧医疗、无人驾驶等安全关键性场景，大模型算法模型的鲁棒可靠愈加重要。然而，大模型存在的鲁棒性不足、模型“幻觉”、可解释性差等自身技术局限，以及指令注入、对抗样本、算法后门等新型安全攻击方式，都可能使大模型应用产生非预期的错误输出。因而，应当确保大模型的算法模型在规定的运行条件和时间周期内始终产生预期的行为和结果，且一直处于可控状态。系统平台安全稳定：构建大模型系统是各行各业使用大模型技术解决实际问题的主要方式，同时，大模型研发平台承担着为设计研发人员提供高效、便捷的研发环境的重要作用。然而，大模型系统和研发平台自身安全漏洞被攻击者利用也将带来应用系统被控制、敏感数据泄露、智能服务中断等严重安全问题。因而，应及时检测发现并修复大模型系统和研发平台的安全漏洞，做好大模型与插件间交互的安全监测与认证。

业务应用安全可控：大模型已在交通、医疗等领域展现出了强大的能力。滥用或恶意使用大模型应用将会给物理世界和国家社会带来巨大的负面影响。因此，首先应确保大模型系统应用目标符合国家法律法规和社会伦理的规范要求。

1.2 安全属性

参考 ISO/IEC 22989:2022《信息技术 人工智能 人工智能概念和术语》国际标准、GB/T 41867-2022《信息技术 人工智能术语》国家标准等标准文件，将大模型安全属性概括如下：

真实性：训练数据能真实反映物理世界客观规律和人类社会实际运转情况的性质。

多样性：训练数据应覆盖尽可能多的样本，以确保大模型能对不同情况进行泛化的性质。

准确性：针对所规定的各项安全要求，大模型展现其正确实现这些要求的性质。

机密性：确保大模型的参数、架构和训练过程的信息对未授权的个人、实体或过程不可用或不泄露的性质。

可问责性：大模型及其利益相关方对其行动、决定和行为负责任的状态。

可预测性：大模型满足利益相关方所提出输出做出可靠假设的性质。

公平性：尊重既定事实、社会规范和信仰，大模型产生的行为或结果不受偏袒或不公正歧视影响的性质。

透明性：大模型系统与利益相关方交流关于该系统适当信息的性质。

可解释性：大模型系统以人能理解的方式，表达影响其执行结果的重要因素的能力。

合规性：用户对大模型系统的应用方式以及大模型系统自身行为和输出信息满足法律法规和规章要求的性质。

可靠性：大模型实施一致的期望行为并获得结果的性质。

可控性：大模型被人类或其他外部主体干预的性质。

鲁棒性：大模型在任何情况下都保持其性能水平的性质。

1.3 保护对象

保护对象包括系统、数据、用户、行为。

系统：系统即大模型系统，一般由服务器、传感器等硬件，数据库、操作系统等基础软件，基于大模型的算法模型，以及支持大模型研发运营的框架平台等主要组件组成。

数据：数据是大模型系统的核心。从大模型系统生命周期视角来看，主要包括训练数据、测试数据和运行时输入数据三类。从类型来看，主要包括文本、图像、视频、音频以及来自于数据库的结构化数据等。

用户：用户指使用大模型系统的组织或实体，可以是自然人和法人，也可以是账户、软件、网络设备等具有唯

一性身份的与大模型系统进行交互的信息收发源。

行为：行为指用户与大模型系统的交互过程，这不仅包括用户对大模型的日常操作，还包括大模型系统与其他系统间的调用操作。

1.4 安全措施

在国家法律法规、各行业监管政策以及社会伦理规则指引下，从训练数据、算法模型、系统平台、业务应用等层面提出相应的安全保护措施。

训练数据安全措施：训练数据安全措施指对大模型训练数据部署的安全防御能力。训练数据安全措施主要包括数据合规获取、数据标注安全、数据集安全检测、数据增广与数据合成、安全对齐数据集构建五个方面。

算法模型安全措施：算法模型安全措施指针对大模型算法模型部署的安全防御能力。算法模型安全措施主要包括模型内生安全评测、模型鲁棒性增强、模型“幻觉”缓解、模型偏见缓解、模型可解释性提升等五个方面。

系统平台安全措施：系统平台安全措施指针对大模型框架平台部署的安全防御能力。系统平台安全措施主要包括系统安全加固保护、大模型插件安全保护两个方面。

业务应用安全措施：业务应用安全措施指在大模型业务层部署的安全防御能力。业务应用安全措施主要包括输入输出安全保护、生成信息标识、账号恶意行为风控、用户协议和隐私政策四个方面。

2. 训练数据安全保护措施

2.1 数据合规获取

数据获取渠道主要包括从互联网或用户处直接获取数据、通过交易或合作方式获取数据、通过自研业务收集或生成数据三种。针对三种渠道，安全保护要点如下。

直接获取方式的安全措施。直接获取数据包括直接获取互联网公开数据和用户输入数据两类。为保护直接从互联网公开获取的数据，需在采集时遵循三点原则：一是爬虫仅获取开放数据，技术非侵入性，且基于正当目的；二是需建立违法不良数据源清单，标识含有威胁的数据源；三是标记溯源数据来源，记录数据的来源、获取时间、获取记录，同时需标记、识别、记录数据中的违法不良信息。为保护直接从用户输入获取的数据，需在采集前明确告知用户此次数据收集的目的、使用方式以及存储期限，获得用户同意。

交易或合作方式的安全措施。通过数据交易或合作方式获取数据的，需签署商业合同或合作协议。

自研业务方式的安全措施。通过自研业务获取的数据包括企业在自身经营活动中产生的数据和在为客户服务过程中产生的数据。针对此种方式，应区分并根据数据权利归属，严格按照约定的数据使用用途、范围和目的进行处理。

除上述保护要点，数据合规获取还需对相关人员进行定期培训，帮助涉及训练数据获取的人员了解相关法律法规要求，明确合规标准与违规风险，提升合规意识和自觉性。

2.2 数据标注安全

数据标注安全包括标注任务制定、标注人员管理及培训、标注执行过程安全以及标注结果的检查与纠错四个方面。

制定清晰的标注任务。首先，标注任务要充分考虑实际要求；其次，提供明确的标注任务目标、标注规则、标注要求等；最后，还需在标注规则中提供参考示例，以帮助标注人员更好地执行任务。

对标注人员进行管理及培训。需设置不同的标注人员角色，包括标注执行人员、标注审核人员、仲裁人员、监

督人员等。需根据标注任务对标注人员进行培训，并要求标注人员必须在通过考核后方可执行标注任务。

标注执行过程安全。执行标注任务前，需检查标注工具是否存在安全漏洞并及时对漏洞进行修复，同时还需建立标注工具访问控制机制，实施身份验证和授权访问控制，确保只有授权人员才可使用标注工具。执行标注任务时，需提供安全的标注环境，并对标注数据的访问和操作进行权限管理和记录。

标注结果的检查与纠错。需对标注结果进行质量审查，可选择全量或抽样、人工或自动核验方式。对于发现的标注错误问题需及时纠正并再次复核，确保质量达标。

2.3 数据集安全检测

数据集安全检测包括违法不良数据检测、数据多样性检测、数据投毒污染检测以及数据隐私检测。

违法不良数据检测。参照《网络信息内容生态治理规定》中规定的 11 类违法信息和 9 类不良信息，对文本、图像、视频、音频、代码等各类训练数据进行安全检测和过滤。文本类训练数据可采用关键词匹配、自然语言处理（NLP）、小模型和大模型语义识别检测等技术。多媒体类训练数据可采用图像识别、语音识别及小模型检测等技术。代码类训练数据检测可采用特征码扫描、行为分析和沙箱检测等技术。

数据多样性检测。对训练数据来源、特征、分布等维度的多样性进行检测。其中，来源多样性检测通过计算来自不同数据源的比例、计算数据来源的地域分布、内容分类等方法进行多样性检测；特征多样性检测通过对数据进行特征统计、特征相关性分析、特征重要性评估、聚类分析等方法进行多样性检测；分布多样性检测通过 KDE 核密度估计、KL 散度、K-S 检验、聚类分析等方法进行多样性检测。

数据投毒污染检测。数据污染投毒检测需对预训练和内部微调数据进行检测。检测算法通过比较投毒数据与正常数据在样本、特征和标签层面的差异，以及模型在两者上的训练过程和神经元响应差异，来进行区分。仅利用数据差异的检测可在黑盒条件下实施，而利用模型性能差异的检测需获得算法模型内部信息及研发者的支持。

数据隐私检测。数据隐私检测是指识别与检测数据中的敏感信息，常用技术包括数据标识符、正则表达式和关键词匹配等。数据标识符检测准确率高，而正则表达式和关键词匹配可能存在漏报和误报。结合上下文分析和库表字段注释，可提升检测准确率。

2.4 数据增广与数据合成

数据增广和数据合成可以扩充数据集规模，并增加数据样本的多样性，从而有效解决数据量不足，以及多样化不足带来的泛化能力弱等问题。

数据增广是在保持原数据集不变的前提下，通过一系列的变换操作，生成新的数据集，且新生成的数据集一般与原数据集保持一定程度的关联，包括基础数据增广技术和高级数据增广技术。

一是基础数据增广技术。在计算机视觉领域，基础数据增广技术的应用尤为成熟，包括几何变换（如旋转、平移、缩放、裁剪）和像素变换（如噪声注入、颜色抖动）等常见技术。

二是高级数据增广技术。随着技术的发展，基于网络架构搜索（NAS）的动态数据增广等新方法逐渐出现，并被应用于图像分类、目标检测、语音识别、自然语言处理（NLP）等任务中。在语音应用领域，通过添加噪声等手段进行数据增广也取得了良好效果。这些先进技术在提升数据多样性和模型性能方面展现出了巨大潜力，但同时也带来了更高的技术复杂度和计算成本。

数据合成是在不使用原数据集的情况下生成数据。合成数据基于算法或模型生成，包括生成对抗网络（GAN）、变分自编码器（VAE）、基于物理仿真、基于统计模型或者基于机器学习等方法。合成数据作为真实数据的一种替代，现阶段虽然在预训练占比不高，但未来发展潜力巨大，可作为一个“新物种”密切关注。

在大模型预训练阶段，合成数据将在多模态和领域知识生成中发挥重要作用。合成数据的生成需要确保数据的真实性和有效性，以避免对模型的训练和测试产生负面影响。可行的应用方案是按照一定比例将合成数据与真实数据进行混合，用于模型性能优化，提升泛化能力。

2.5 安全对齐数据集构建

安全对齐数据集旨在降低大模型产生不真实、有偏见、不道德等风险，确保模型的输出符合人类规则和道德准则。安全对齐数据集包含有监督正样本、恶意样本及外部检索对齐数据。

一是有监督正样本数据构建。该数据集为人类标注的正样本，符合人类价值观，旨在模型微调时更好地学习和对齐。借助专家知识和经验标注数据，确保符合所定义的价值观。微调时以人类价值观为原则调整标注数据的排序方式，对有帮助性、无害性以及基于事实的优质问答打高分，指导奖励模型学习更符合人类价值观的策略，

从而发挥价值对齐技术的优势。

二是恶意样本数据构建。该数据集包含各种经过标注的针对大模型的对抗性攻击提示词和违法不良信息样本，旨在帮助开发人员构建评测数据集，测试模型的内生安全性以及生成内容的安全性，了解模型在面对异常样本、提示注入攻击、数据窃取攻击时的表现，有助于开发人员增强模型鲁棒性、缓解决策偏见等问题。

三是外部检索对齐数据构建。在面向特定的问题时，用于检索要对齐的价值观并作出合适的回复的数据基准，适用于法律、法规、制度文件等比较定制化的价值维度，即为模型建立法律和道德标准，对回复进行约束。该方法可以有效提升检索生成增强的效果，进一步缓解模型幻觉现象。

安全对齐数据集的构建需考虑数据多样性、攻击复杂性和安全评估科学性。同时，为保持有效性和实时性，需定期更新数据集以应对变化中的攻击手段。

3. 算法模型安全保护措施

3.1 模型内生安全评测

模型内生安全评测主要包括模型鲁棒性评测、模型“幻觉”评测和模型偏见性评测。

模型鲁棒性评测。该评测旨在全面客观定量评价模型在面对小概率异常场景、提示注入攻击场景以及恶意添加扰动的对抗样本输入时仍产生正确输出的概率。目前，大语言模型的鲁棒性测评较为成熟，多模态大模型的评测仍处于研究初期。针对大语言模型，分布外鲁棒性评测数据集主要包括 Flipkart、DDXPlus 等，对抗鲁棒性评测数据集主要包括 AdvGLUE、ANLI、PromptBench 等。鲁棒性评测指标主要包括模型预测的准确性、性能下降率等。

模型“幻觉”评测。目前，模型“幻觉”评测基准主要集中在大语言模型，多模态模型的“幻觉”评测方法仍较为初级。大语言模型“幻觉”评测主要评估大语言模型生成内容与输入信息或者事实知识的内容一致性及相关性程度，目前主要包括基于事实度量、基于分类器度量、基于问答系统度量、不确定性估计以及基于大模型的度量方法。主流的幻觉评测数据集包括 TruthfulQA、HalluQA、UHGEval 等。

模型偏见性评测。该测评旨在全面客观定量评价大模型在训练阶段和推理阶段的偏见歧视程度。模型偏见性评测流程可分为偏见风险分析、评测任务选择、评测指标选择和数据集构建。主流的偏见性评测数据集包括WINOGENDER、BOLD等。通常，评测人员会统计生成内容中的关键属性和词语的概率，来反映模型的偏见程度。

3.2 模型鲁棒性增强

模型鲁棒性增强以对抗性训练为主，通过模拟提示注入攻击场景和对抗样本，支撑算法模型从数据中学习到相关特征以提升算法鲁棒性。

提示词安全增强，包括提示词语义增强和提示词结构增强。提示词语义增强的核心是在提示词中增加鲁棒性任务描述以及对模型进行提示注入攻击少样本学习。鲁棒性任务描述方法，通过在提示词中额外添加鲁棒性任务描述，用于提升模型对原有用户任务的执行度。例如，可在用户输入提示词中强调原有任务的执行力度并忽略任何非原任务意图的指令。少样本学习方法，通过在训练数据中增加多项添加了提示注入攻击指令的提示词和正确回复的示例，对模型进行专项训练，从而指导模型正确识别提示注入攻击。

提示词结构增强的核心是提示词位置调整和特殊符号标记。提示词位置调整方法，是通过更改原有用户输入信息和任务指令的位置，使攻击提示词部分失效，从而降低模型被提示注入攻击的概率。例如，可将原有任务指令置于用户输入信息之后，可以使大模型不执行“忽略下列指令”等诱导性指令。特殊符号标记方法，是通过特殊符号增强用户输入信息和任务指令的差异性，减少模型将诱导性用户输入信息误解为任务指令进行执行的情况，有效提升模型抵御指令注入攻击的能力。

对抗性样本输入增强，可根据鲁棒性评测结果，针对性构建含有字符级、单词级、句子级以及语义级干扰信息的训练数据集，用于缓解含有干扰信息的提示词对模型鲁棒性的影响。

3.3 模型“幻觉”缓解

模型“幻觉”缓解主要包括检索增强生成、有监督微调、思维链技术以及价值对齐技术。

一是检索增强生成（Retrieval Augmented Generation, RAG）

该技术是一种将检索器与生成式大模型相结合的技术。在大模型生成过程中，通过检索器从外部源或向量数据

库检索知识，并由大模型根据原始输入信息和检索器获得的知识合成所需的回答。目前检索增强生成主要包括一次性检索、迭代检索和事后检索。一次性检索通过将一次检索获得的外部知识直接添加到输入提示词中，可持续提高大模型生成信息的准确性。迭代检索是为了解决应对复杂问题时一次性检索能力限制问题，该方法允许在整个信息生成过程中多次检索收集知识，可有效减少推理链中事实性错误。事后检索通过使用检索获得知识从而对大模型已生成的信息进行修正，可有效增强大模型生成信息的准确性。

二是有监督微调（Supervised Fine-Tuning，SFT）

有监督微调是一种通过微调数据集提升大模型理解和生成能力的技术。该技术的优点是可在现有模型知识水平基础上进一步提升模型的信息理解和生成能力。例如，针对多轮对话中上下文不一致的模型“幻觉”问题，使用含有多轮提示词及正确回复的微调数据对模型进行安全性微调，可有效提升模型在多轮对话后的注意力，增强上下文一致性。

三是思维链技术（Chain-of-thought，CoT）

该技术是一种可增强大模型生成信息逻辑性的技术。通过向大模型展示少量包含详细推理过程的样例，帮助大模型在生成信息时不仅给出结果还提供推理过程。该方法在提升大模型推理过程透明度的同时，可显著提升生成信息的准确性。

四是价值对齐技术（Value Alignment）

该技术是一种确保大模型系统的目标和行为与人类的价值观和利益保持一致的技术和理念。目前，价值对齐主要包括基于人工反馈的强化学习、基于人工智能反馈的强化学习两类。

基于人工反馈的强化学习（Reinforcement Learning from Human Feedback, RLHF）。RLHF 是一项通过人工反馈大模型生成信息好坏排序以指引大模型价值观与人类对齐的强化学习技术。RLHF 适用于对已经微调的大模型进行改进，使其更加符合人类偏好。由于 RLHF 性能受人类标注的数据质量和时效性影响较大，且奖励模型存在通过学习欺骗式奖励策略实现“欺骗式”对齐的风险，因此需要进一步探索高可靠性价值对齐技术。

基于人工智能反馈的强化学习（Reinforcement Learning from Artificial Intelligence Feedback，RLAIF）。RLAIF 是一种结合人工反馈和人工智能反馈的强化学习方法。在强化学习阶段，RLAIF 通过人工智能模型部分取代人类标注员对大模型生成信息好坏进行排序，并将其与人类标注员排序结果进行融合，共同用于奖励模型的训练。目前该项技术尚处于研究初期，主要以 Anthropic、OpenAI 和 Google 等公司的实践为主。

3.4 模型偏见缓解

大模型的偏见缓解措施主要用于缓解训练阶段和推理阶段的偏见问题。

训练阶段的模型偏见缓解措施。通过优化模型训练过程和模型结构对模型进行偏见缓解，包括构建偏见性样本进行对抗性训练、优化损失函数、选择性冻结部分模型参数、移除偏见歧视相关的神经网络节点等。

推理阶段的模型偏见缓解措施。基于预训练模型或者微调后的模型，在不进行进一步微调的前提下控制偏见内容的输出，以提升预训练或微调模型的公平性，包括调整输入的关键词类别、分布以及模型权重等。

3.5 模型可解释性提升

针对大模型的可解释性提升可分为局部可解释和全局可解释。

局部可解释性方法。该方法主要包括特征属性分析和 Transformer 结构分析。特征属性分析旨在识别和评估哪些输入特征对模型生成信息造成影响及其影响程度，主要包括干扰分析法、梯度分析法、向量分析法等，目前实践以 SHAP 和 LIME 等方法为主。Transformer 结构分析旨在研究 Transformer 自注意力层和多层感知机层的机理，通过分析注意力权重了解模型如何对输入分配注意力，从而理解模型在文本生成中关注的输入信息的关键部分。例如，OpenAI 正在尝试使用 GPT-4 模拟解释 GPT-2 神经元与生成信息的映射关系。

全局可解释性方法。该方法主要包括基于探针的方法和机制可解释。基于探针的方法旨在分析和理解大模型生成信息的高层次表征，这些表征有助于从宏观角度理解大模型生成信息的行为，如研究人员采用神经元热力度的方法、观察模型输出信息是否真实等。机制可解释旨在通过类比复杂计算机程序的逆向工程思路探索神经元的提取特征与大模型生成信息的映射关系。例如，Anthropic 正在研究通过字典学习等方法分解神经元，尝试解释神经元提取的单一特征与生成信息之间的映射关系。

4. 系统平台安全措施

4.1 系统安全加固保护

系统安全加固保护主要应对上述的机器学习框架、大模型系统开发工具链、系统缺陷三方面安全风险，包括建立良好的安全开发机制、加强供应链安全管控、实施多层次的安全测试、构建有效的安全响应机制、定期开展漏洞检查工作、建立严格的访问控制机制等六方面内容。

建立良好的安全开发机制。首先，在大模型系统开发过程中建立安全开发标准，制定详细的安全编码规范、架构设计指南、安全配置模板等，确保开发人员在各个环节有明确的安全操作准则；其次，规范安全开发流程，将安全管控活动嵌入到大模型系统开发流程中，包括需求分析、设计、编码、测试、部署和运维阶段，实现安全与开发的深度融合；再次，加强安全开发培训，定期举办安全培训课程，提升安全设计、安全编码、安全开发的意识、素养；同时，持续跟进安全开发技术，加强大模型系统安全开发实践；最后，定期进行内部或第三方安全审计，评估安全开发机制的有效性，识别改进点。

加强供应链安全管控。首先，对大模型系统相关的机器学习框架、大模型系统开发工具链、大模型插件进行供应链安全管控，对大模型系统使用了哪些开源组件以及组件之间的依赖关系进行分析，评估这些开源组件的安全性并识别它们带来的潜在风险；其次，跟踪大模型系统所使用组件的更新和维护情况，确保获取最新的安全补丁和更新；最后，关注大模型系统相关供应商的信誉和安全实践，选择有良好声誉和专业的供应商以减少潜在的安全风险。

实施多层次的安全测试。首先，使用静态代码分析工具检查大模型系统模型代码、服务端代码、客户端代码等，查找常见的编程错误、安全漏洞和不符合安全编码规范之处；其次，对大模型系统进行功能安全测试，验证其在正常操作下能否正确执行权限控制、数据过滤、输入验证等功能；再次，对大模型系统进行接口安全测试，测试其 API 接口的认证、授权、数据加密、速率限制等安全特性；同时，对大模型系统进行模糊测试，检测其对异常或边界条件的处理能力；最后，对大模型系统进行渗透测试，验证其防御措施是否有效。

构建有效的安全响应机制。第一，制定详细的大模型系统应急响应预案，涵盖安全事件分类分级、事件响应、责任人分配、通信渠道保障、决策流程高效等环节；第二，组建大模型系统应急响应小组，包括安全专家、开发人员、系统管理员等，明确各自职责与协作方式；第三，持续监控大模型系统运行状态、访问行为、数据流动、异常日志等，及时发现潜在的安全威胁；第四，设置大模型系统警报阈值与告警机制，确保在安全事件发生时能够迅速通知相关人员；第五，实施大模型系统修复措施，如打补丁、更新配置、调整安全策略、强化访问控制等；第六，定期开展应急响应演练，提升团队协同作战能力和对应急预案的熟悉度。

定期开展安全漏洞检查工作。首先，在大模型系统全生命周期中，明确安全漏洞检查频率（例如，每月、每季度或每年），并制定安全漏洞检查流程和责任分配机制；其次，综合运用静态分析、动态分析和渗透测试等技术手段，识别潜在的安全漏洞；再次，建立大模型系统漏洞报告和修复机制；同时，记录所有发现的漏洞（包括详细信息、风险评估和修复状态）；最后，定期复审安全漏洞检查流程，评估其有效性，并根据需要进行改进。

建立严格的访问控制机制。首先，所有请求访问大模型系统的用户必须通过身份验证，确保只有授权用户才能访问模型，防止未授权访问和潜在的恶意使用；其次，用户通过身份验证后，根据用户的权限级别授予相应的访问权限，避免权限过度集中或滥用；再次，通过统一的 API 安全措施进行访问控制、流量管理、认证授权、速率限制、请求转换等，增强大模型系统接口安全性；最后，对已知恶意用户或 IP 地址设置访问黑名单，阻止其对系统的任何访问。

4.2 大模型插件安全保护

大模型插件安全保护包括加强对大模型插件输入内容的检测、大模型插件功能 “最小化” 、有效管控大模型插件的安全权限、建立重要功能的人工审核机制、增强供应链安全审核等五方面内容。

加强对大模型插件输入内容的检测。第一，插件开发人员应根据 OWASP ASVS 的建议，进行有效的输入验证和参数净化；第二，插件应尽可能强制执行参数化输入，并对输入数据的格式、类型和范围进行检查，对于不符合规范的输入，应拒绝处理并返回适当的错误信息；第三，应检查输入数据是否包含敏感信息，如个人身份信息、密码等，以防止潜在的隐私泄露风险；第四，当因应用程序语义而必须接受自由格式的输入时，应仔细检查以确保没有调用潜在威胁的方法，包括可能会引起提示词注入攻击的输入；第五，对大模型插件输入进行记录和监控，记录所有接收到的输入数据，以便于事后分析和追踪潜在的安全风险。

大模型插件功能“最小化”。首先，限制可调用插件的功能，仅限于必要的最小化功能；其次，限制插件与第三方系统交互权限至最小集合，并对其使用情况进行审计，记录异常调用；最后，插件应当只访问完成其功能所必需的数据，不应无故收集或存储额外信息。

有效管控大模型插件的安全权限。一方面，在大模型插件上线后对大模型插件的访问权限进行管理，包括哪些用户或系统具有访问大模型插件的权限，以及具体的权限范围（例如读取、写入、执行等）；另一方面，大模型插件应当只请求其运行所必需的最低权限。

建立重要功能的人工审核机制。在大模型插件重要功能执行时引入人工审核，如在调用插件执行特权操作（例如删除电子邮件）时，应要求用户批准该操作。这将减轻间接提示注入的风险，以防止用户在不知情或未经同

意的情况下执行危险操作。

增强供应链安全管理。一方面，仔细审查大模型插件供应商（包括服务条款和隐私政策），尽量使用可信赖的插件供应商，确保采取足够的、经过独立审核的安全措施；另一方面，在进行大模型插件开发的时候，采用 SCA 代码组件成分分析工具对用到的第三方组件进行漏洞检测和分析，维护一个最新的软件物料清单（SBOM）以便对组件版本进行跟踪，避免使用过时和存在漏洞的第三方组件。

5. 业务应用安全措施

5.1 输入输出安全保护

大模型的输入输出信息安全需构建输入输出信息的护栏，对输入输出内容进行风险检测，对敏感问题进行安全回复，并对输出内容进行安全改写，该系统的框架流程图如图 4 所示。

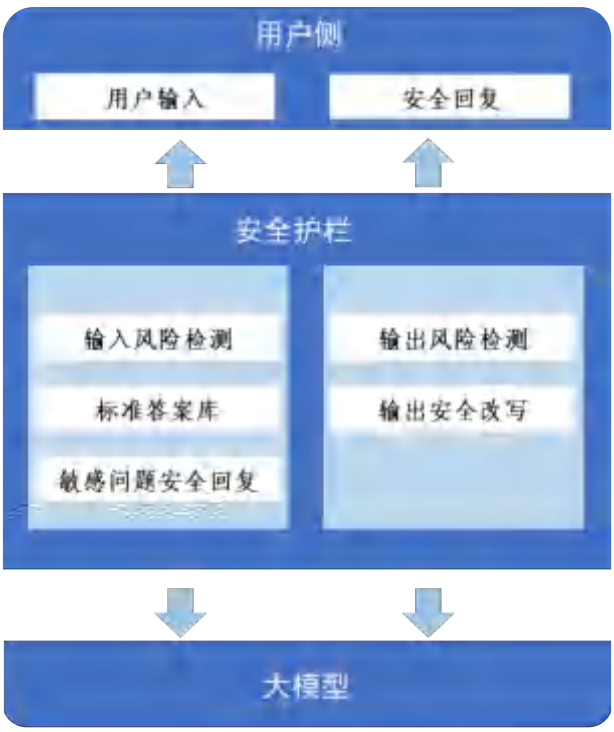

图4 安全护栏架构流程图

输入输出风险检测：通过人工运营的敏感词库和分类模型识别出用户请求和大模型生成内容中是否含有违法不良信息，提前制止或者即时阻断大模型生成不良信息。

敏感问题安全回复：对于一些敏感而又难以回避的话题，通过人工干预的方式配置一个安全回复答案，也可以通过专有数据微调的方式训练一个安全回复大模型，专门为敏感问题生成安全回复。

违规输出安全改写：如果在输出侧检测到不安全的信息，除了中断会话，还应提供重新改写的方式，在保证安全的情况下提升服务体验。

5.2 生成信息标识

在大模型业务应用中，对 AIGC 生成内容添加标识已经是有明确规定的必要措施。对大模型生成内容添加标识的目的主要有下列几种：1）用来标记该内容是由人工智能生成；2）用于声明版权或所有权等；3）用于追溯生成内容的起源；4）用于标记生成内容的分发或传播渠道等。为大模型业务应用添加上述的一种或多种标识，对于模型生成内容的分类识别、版权保护、来源追溯和定责等方面有着重要的意义，从实现方法上看，添加 AIGC 标识可以采取下列解决方案：

显式水印标识。一般指人类可以直接感知且识别的标识内容，例如文字、Logo、背景音等形式。通常来说，显式水印标识宜添加在生成内容的适当位置，例如图像的四角。显式水印标识的添加应尽量保证标识足够明显，且避免影响生成内容的正常使用。常用的显式水印解决方案有：

（1）可在交互区域显著位置，以文字、透明图等形式提示服务所提供的内容由人工智能生成，提示内容包含服务提供者名称、内容生成时间等其他信息；

（2）在大模型生成的文字类内容的开头和结尾添加提示类文本内容；

（3）在大模型生成的图像、视频画面的合适位置以文字、Logo、纹理等形式添加所需的标识信息；

（4）在大模型生成的音频的适当位置，或是以背景音、独立提示音的形式添加提示。

隐式水印标识。一般指人类无法直接感知或识别，但可通过技术手段从内容中提取的标识信息。隐式水印标识的技术种类繁多，根据不同类型的生成内容、不同的标识长度、不同的提取要求等因素，所用的实现方式差别较大。相比于显式标识，隐式水印标识有着难以感知、安全性较高、鲁棒性较好、容量较高等众多优势。隐式水印也同样要避免影响生成内容的正常使用。常用的隐式水印解决方案有：

（1）对于大模型生成的图像内容，可通过变换域水印、模板水印、压缩域水印、直方图水印、最低有效位水

印等技术手段嵌入隐式水印标识信息；

（2）对于大模型生成的视频内容，除上述列举的适用于图像的隐式水印标识技术外，还可通过时序水印等技术手段嵌入隐式水印标识信息；

（3）对于大模型生成的音频内容，可通过变换域水印、模板水印、压缩域水印、直方图水印、最低有效位水印、回声水印、扩频水印等技术手段嵌入隐式水印标识信息；

（4）对于大模型生成的图像、音频、视频以文件形式输出时，在其文件头中写入提供者名称、生成时间等标识信息。

5.3 账号恶意行为风控

在大模型应用过程中，账号行为风控是确保平台安全、保护用户利益的关键环节。对于风险账号的挖掘、异常行为的预警和及时干预，可以采取以下解决方案：

一是建立环境感知分析机制。环境感知分析机制，通过收集和分析用户的设备环境信息来识别和防范风险。常用的技术包括：

（1）设备指纹：收集设备硬件和软件信息，生成唯一指纹以识别重复或伪造的设备。

（2）环境感知：检测设备运行环境，比如操作系统版本、浏览器类型和插件，以识别可能的异常。

（3）接口防刷：防止恶意攻击者通过高频率调用接口消耗服务器资源。可以通过控制接口调用频率、自动限制或延迟请求处理、人机验证码等方式进行防范。

二是建立账号安全机制。账号安全主要是为了防止账号被盗用、信息泄露等风险。常用的技术包括：

（1）多因素认证（MFA）：除了用户名和密码外，还需要第二种或多种认证方式，如短信验证码、电子邮件验证码、生物识别等；

（2）风险评分：通过用户行为分析，结合登录地点、设备、时间等因素，动态评估登录风险；

（3）账号异常行为监测：通过分析用户行为，识别异常登录（如异地登录、频繁登录失败等）并采取相应措施。

三是建立风控机制。根据环境感知分析机制收集整理的数据，结合风控规则 / 模型 / 特征识别，提供给业务风控建议，业务根据实际需要选择放行、拦截或者人机识别处理的决策。其中，人机识别技术用于区分操作是由人类还是机器人执行，以防止自动化工具进行恶意操作。常用的技术包括：

（1）行为识别：分析用户的操作习惯，如敲键力度、速度和节奏，来识别用户。

（2）生物识别：利用指纹、虹膜、面部识别等生物特征来确认用户身份。

（3）图灵测试：通过一些只有人类能回答的问题来区分人与机器。

（4）语音和图像识别：分析声音的音调和图像中微表情，以区分真实用户和机器人。

5.4 用户协议和隐私政策

大模型应用场景下，处理敏感个人信息和保护用户隐私至关重要，为了更好地明确用户和服务提供者之间数据使用、存储和共享的界限，大模型服务提供者需要在用户使用服务前向用户明示用户协议和隐私政策。

用户协议。定义服务使用的规则和条件，包括服务描述、用户行为规范、账户管理、知识产权归属、服务终止条件以及责任限制。这些条款指导用户如何安全合规地使用服务，并明确了在违规行为发生时的后果。

隐私政策。阐释个人信息的收集、使用和保护方式。详述哪些数据被收集，使用的目的，与第三方的数据共享条件，用户对自己数据的权利，以及数据保护措施。政策也会定期更新，确保用户了解其隐私权的最新保护措施。

大模型
赋能安全

1. 大模型赋能安全框架

2. 大模型赋能网络安全

3. 大模型赋能数据安全

4. 大模型赋能内容安全

3

1. 大模型赋能安全框架

大模型赋

图5 大模型赋

赋能安全


数据安全
内容安全
安全恢复
• 智能应急策略制定L1
生成L4
• 自动化数据分类分级 L2
• 自动化APP (SDK) 违规处理个人信息检测 L3
• 智能文本内容安全检测L2
• 智能图像视频内容安全检测L2
• 智能音频内容安全检测L2
模型在某一类或几
较为广泛的部署
数量的稳定用户
L4:较多应用，大模型在多个安全领域实现广泛应用，被众多企业，机构采用，成为主流的安全技术手段之一。
L5:全面应用，大模型在安全行业的几乎所有重要场景中实现深度渗透与全面覆盖，成为不可或缺的基础技术设施。

赋能安全框架

当前，随着新一代信息技术和数字经济的飞速发展，网络空间范围快速膨胀，安全保护对象的复杂性和数量急速增长，攻击者亦加速利用新技术增强威胁破坏行为的隐蔽性和精准性。网络空间安全面临攻击行为愈加隐蔽难以精准发现、数据资产散落各处难以有效管理、高逼真虚假不良信息难以精准识别、安全人才培养难周期长难以满足实际需要等诸多挑战。

大模型具有的自然语言理解、知识抽取整合、意图识别判断、任务生成编排和代码理解生成等能力，为网络空间安全瓶颈问题解决带来了新思路和新方法。例如，自然语言理解能力能帮助大模型准确理解专业人员下达的安全指令含义；知识抽取整合能力可助力大模型自动化精准生成威胁情报；意图识别判断能力能帮助大模型从海量安全告警信息快速准确识别出潜藏的攻击意图；任务生成编排能力可助力大模型面向复杂网络安全问题形成全面可行的解决方案和执行步骤。大模型强大能力的有效应用将大幅提升现有网络空间安全保护技术体系效能，大模型赋能安全框架如图 5 所示。

结合行业实践情况，本报告重点阐述大模型在网络安全、数据安全、内容安全三个领域的探索应用情况。网络安全领域，大模型可应用于安全威胁识别、保护、检测、响应、恢复等多个保护环节中的关键场景。数据安全领域，大模型可应用于数据分类分级、APP（SDK）违规处理个人信息检测等场景。内容安全领域，大模型可应用于文本内容安全检测、图像视频内容安全检测和音频内容安全检测等场景。

2. 大模型赋能网络安全

目前，大模型已在网络安全领域展现出了巨大应用潜力，有望显著提升网络安全整体防护水平和安全事件处置效率。从安全能力框架 IPDRR（Identify Protect Detect Respond Recover, IPDRR）看，大模型在各环节均已开展试点商业化应用。

2.1 风险识别 (Identify)

大模型在风险识别环节拥有显著应用潜力。本报告重点介绍大模型在智能威胁情报生成整合、自动化漏洞挖掘、自动化代码审计、智能网络攻击溯源等场景的商业化应用情况。

2.1.1 智能威胁情报生成整合

威胁情报旨在为面临威胁挑战的资产所有者提供全面、精确、高度针对性的威胁知识与信息，以辅助资产所有

者制定有效的安全保护决策。但是，目前高质量威胁情报生成整合领域缺乏能从各类威胁情报来源中准确抽取关键信息的自动化工具。

大模型拥有信息提取能力、自然语言理解能力和情报生成能力，可以准确便捷的从 CVE 漏洞信息、安全论坛讨论、暗网聊天记录等各类公开和私有的安全信息中，准确提炼出恶意 IP 地址、恶意 URL、恶意文件哈希值等各类高价值威胁指标进而生成威胁情报，供安全人员后续分析。而且，大模型具有关联分析能力和数据整合与可视化能力，能对多类数据源进行关联分析，将看似无关的信息片段拼接成完整的威胁全景图。例如，大模型能将 IP 地址、域名、文件哈希值、攻击签名等散乱数据点关联起来，从而揭示出隐藏的攻击链路和攻击者意图。

在行业应用方面，大模型在智能威胁情报生成整合方面成熟度达到 L3 级别。已有部分安全厂商推出了基于大模型的威胁情报生成整合产品，可支持用户以自然语言方式进行威胁情报查询，显著提高了威胁识别和应对效率。

2.1.2 自动化漏洞挖掘

漏洞挖掘旨在识别尚未被软件开发商或安全研究者发现并公开披露的软件漏洞。但是，目前漏洞挖掘面临着严重依赖安全专家经验、缺乏自动化工具的挑战。

大模型在此领域展现了强大的代码和文本理解分析能力，能自动审查海量源代码、二进制文件和系统日志，并通过运用模式识别与异常检测技术来发现未公开的零日漏洞。例如，在实际运行环境中大模型可监控程序的行为特征，通过检测识别出显著偏离正常行为模式的异常行为，来预测零日漏洞的存在。此外，大模型还可依据对程序内部结构的理解、通过已知漏洞特征来推测未知漏洞特征，并通过生成高质量测试数据集达成有效触发和识别潜在零日漏洞的目标。

在行业应用方面，大模型在自动化漏洞挖掘方面成熟度达 L1 级别。尽管大模型在漏洞挖掘领域展示出巨大潜力，但实际应用中仍面临误报率高、解释性不足以及对新型攻击手法适应性等问题，目前处于早期研究阶段。

2.1.3 自动化代码审计

代码审计旨在检查程序源代码中是否存在编码错误、逻辑错误等安全缺陷，并提供相应的修复方案与改进建议。但是，目前代码审计面临自动化工具误报漏报率高难以实用的挑战。

大模型通过学习海量的优质代码和编程错误案例，可学习掌握各种编程语言的语法、库函数用法及常见问题解

决策略。大模型凭借强大的上下文理解能力，可精准识别代码的功能意图和逻辑流程，并准确发现编码错误、调用错误、逻辑错误等多类型的已知和未知安全漏洞。在检测识别出安全漏洞后，大模型利用其代码生成能力，提供漏洞修复建议，可帮助开发者快速定位并解决问题，减少人为错误率。

在行业应用方面，大模型在自动代码审计方面成熟度已达 L2 级别。当前一些安全厂商在代码审计工具中尝试应用大模型技术，并取得一定效果，能够有效发现代码问题并提出实用的修复建议。

2.1.4 智能网络攻击溯源

网络攻击溯源旨在通过技术手段追踪与分析网络攻击的源头及其发起者。但是，目前网络攻击溯源的主要挑战在于，已有自动化工具难以满足对高隐蔽性网络攻击行为溯源的及时性和准确性要求。

大模型凭借意图识别、信息整合等技术能力，可在攻击路径重建、攻击者画像等多个关键溯源环节发挥关键作用。一是攻击路径重建方面。大模型能够利用事件日志、防火墙记录、终端遥测等数据，复原攻击者从初始突破点到目标系统的完整攻击链，展示攻击者如何绕过安全防御、进行权限提升并在系统中扩散的详细过程。二是攻击者画像方面。大模型通过综合分析攻击手法、攻击工具、IP 地址、域名、注册邮箱等信息，能推测出攻击者技术水平、组织归属、攻击偏好等关键信息，进而建出攻击者的详细画像。三是恶意基础设施追踪方面。大模型通过分析 C&C (Command and Control) 通信流量、DNS 查询记录、IP 信誉数据库等信息，追踪攻击者所使用的 C&C 服务器、恶意域名和僵尸网络节点等恶意基础设施。

在行业应用方面，大模型在网络攻击溯源方面成熟度已达 L1 级别。部分安全厂商已开始将大模型技术集成于安全产品中用于增强攻击溯源能力，但这一应用仍严重依赖于情报库的支持，整体上还处于探索阶段。

2.2 安全防御 (Protect)

目前，大模型在生成安全决策和执行处置行为时的准确性及可靠性方面尚有欠缺，未能满足安全防御环节的实际业务需求，尚未在安全防御环节发挥显著作用。本报告重点介绍大模型在动态策略管理场景的商业化应用情况。

2.2.1 动态策略管理

策略管理旨在根据本机构安全目标制定、实施、监控和持续优化安全政策的过程。但是，目前策略管理面临着

主要依赖专家经验，自动化工具难以满足策略制定精准性和及时性要求等挑战。

大模型凭借其突出的自然语言理解和意图识别能力，能深刻洞察实际应用场景中的安全需求，并能结合时刻变化的安全威胁和风险演变情况，动态地推荐和调整安全策略，以强化安全防御能力。例如，当检测到特定 IP 地址发起可疑扫描活动时，大模型能够即时推荐添加对该 IP 的封锁策略；或者在检测到某个端口被频繁用于攻击时，大模型能够设置临时关闭该端口访问的策略。而且，对于复杂的安全策略集，大模型能深入理解每一条策略蕴含的安全意图，进一步对策略集进行梳理、筛选、整理与合并，达到优化精简安全策略的目标。

在行业应用方面，大模型在动态策略管理方面成熟度已达 L2 级别。目前大模型已在防火墙、入侵防御等设备的策略管理上得到实际应用，并确定一定效果。

2.3 安全检测 (Detect)

大模型在安全检测环节拥有广阔应用前景。本报告重点介绍大模型在自动化告警分析、智能报文检测、智能钓鱼邮件检测、智能未知威胁检测等场景的商业化应用情况。

2.3.1 自动化告警分析

告警分析旨在从海量告警中快速识别和响应真正的安全威胁。但是，目前告警分析面临海量告警、误报率高等挑战。

大模型凭借多源信息融合、关联分析等技术能力，可在攻击路径还原、告警过滤与降噪等多个关键告警分析环节发挥关键作用。一是攻击路径还原方面，通过整合告警信息，大模型能精准识别并关联同一攻击事件中的告警，构建出黑客的攻击轨迹，从而直观呈现其从入侵、权限提升、横向移动直至目标达成的连贯行动链。二是告警过滤与降噪方面，大模型深入分析疑似误报，融合上下文和历史数据，精确辨识真实威胁，有效降低安全团队对无效告警的响应。三是告警解释方面，大模型综合告警详情与相关上下文信息，如环境、用户行为和工具使用情况，能够生成详尽的解释报告，助力分析师快速把握告警本质并采取恰当响应措施。四是警情评估方面，大模型有助于实现动态的警情评估，当告警与高危威胁情报匹配，例如 APT 活动或零日漏洞，大模型将提升告警级别并触发应急响应，指导安全团队迅速采取防护措施，确保及时应对潜在威胁。

在行业应用方面，大模型在告警分析方面成熟度已达 L3 级别。当前，大模型在安全告警分析领域的应用已经相对成熟，常被深度集成至扩展检测与响应平台、态势感知系统等安全运营平台中，提高了告警分析的效率和

准确性。

2.3.2 智能报文检测

报文检测旨在通过监控与深度分析网络中传输的数据包发现潜在的恶意活动、异常流量、漏洞利用或其他安全威胁。但是，目前报文检测面临着从网络流量中识别安全攻击的准确率低等挑战。

大模型凭借其强大的自学习能力，能够从海量数据中自动提取关键特征，有效识别出异常报文，例如，它能通过语义分析出看似正常的 JavaScript 代码中隐藏的 SQL 注入攻击。结合威胁情报，大模型还能对网络流量进行深度包检测，识别出与 APT 攻击相关的报文，如发现伪装成合法通信的 C&C 通信，揭示正在进行的高级持久威胁活动。此外，大模型通过分析报文中新颖或未知的特征，结合机器学习算法预测潜在的零日攻击，如在大规模扫描活动中识别出可能利用未公开漏洞的探测性攻击。

在行业应用方面，大模型在报文检测方面成熟度已达 L3 级别。当前，许多国内外安全企业正积极探索将大模型应用于深度报文检测，已取得市场的积极反馈。尽管如此，该领域仍面临模型解释性、数据隐私与安全等挑战。

2.3.3 智能钓鱼邮件检测

钓鱼邮件检测旨在识别并拦截那些含有欺诈信息、企图盗窃用户敏感信息或诱使用户执行恶意操作的电子邮件。但是，目前钓鱼邮件检测面临着难于准确识别出高隐蔽性钓鱼邮件等挑战。

大模型凭借其强大的自然语言理解能力，深入解析电子邮件内容，从邮件标题和正文抽取关键信息，并结合上下文进行深入分析，以精准识别出钓鱼邮件。例如，大模型能够识别邮件中紧迫的语气、逻辑上的矛盾、链接与邮件内容的不一致以及使用同音异形词构造的 URL 等典型的钓鱼邮件特征，从而判断邮件的真实意图。此外，大模型的文本生成能力可以清晰地呈现钓鱼邮件的判断逻辑，帮助用户提升对钓鱼邮件的认知理解，有助于他们在未来遇到类似情况时做出更准确的判断。

在行业应用方面，大模型在钓鱼邮件检测方面成熟度已达 L2 级别。当前，大模型在钓鱼邮件检测中已展现出一定的成效，其提供的详细判断依据显著增强了用户体验和安全意识。

2.3.4 智能未知威胁检测

未知威胁检测旨在主动识别和分析那些尚未被明确定义、分类或广泛认知的潜在安全威胁，以便及早采取预防和应对措施，减少未知攻击可能造成的损害。但是，目前该技术主要面临高隐蔽性、复杂性、多变性攻击难以被准确检测等挑战。

大模型凭借代码理解、意图识别等技术能力，可在新型恶意软件检测、零日漏洞利用检测等多个关键未知威胁检测环节发挥重要作用。一是新型恶意软件检测方面，大模型能够分析网络流量中的异常文件下载行为，即使这些文件未被传统反病毒软件标记，也能够通过其网络行为，如隐蔽通信、自我复制、加密数据交换，识别出潜在的新型恶意软件。二是零日漏洞利用检测方面，当监测到系统进程异常崩溃，大模型能够通过分析发现崩溃前的内存访问和系统调用序列，基于模式识别技术，预测可能存在的零日漏洞利用。三是内部威胁预警方面，通过分析员工账号的行为模式，大模型能够发现与常规行为显著偏离的活动，如在非工作时间的异地登录和异常数据导出，即使这些行为不违反任何明确的策略，也会触发内部威胁预警。四是供应链攻击检测方面，大模型监控软件供应链环节，能够识别出软件更新包的数字签名微小差异，通过深度学习模型判断签名伪造的可能性，并进一步分析确认该更新包是否携带后门。五是网络隐身攻击识别方面，在网络流量分析时，大模型能够识别出看似正常但具有微妙差异的 TCP 连接，揭露利用网络协议特性进行隐身的新型攻击。

在行业应用方面，大模型在未知威胁检测方面成熟度已达 L1 级别。当前，大模型在未知威胁检测领域展现出一定潜力，但其效果受到安全知识数据质量的显著影响，实战效果有待进一步验证和观察。

2.4 安全响应 (Response)

由于大模型在调用安全工具的准确性、时效性上尚有欠缺，无法满足在安全响应环节的实际业务需求，未发挥出显著作用。本报告重点介绍大模型在智能响应、智能事件报告生成等少量场景的商业化应用情况。

2.4.1 智能响应

旨在及时检测和应对网络威胁、安全违规行为或攻击，其目标是在威胁造成影响前进行有效预防，并最大限度降低攻击导致的成本损失与业务中断。但是当前智能响应面临着着高度依赖于专家经验，难以快速形成联动应对方案等挑战。

大模型利用其决策能力，根据当前网络风险状况，为安全专家提供自动化的响应策略与处置流程建议。它能自

动生成响应脚本，并与多种安全工具（如防火墙、入侵防御系统、终端安全等）集成，直接调整设备策略或执行必要的修复操作，如隔离受感染设备、阻断恶意流量、更新防火墙规则等。通过与各种安全工具的集成，大模型能够跨工具进行任务编排，确保整个安全体系的响应和处置既快速又协调，极大提升了安全事件的响应效率。

在行业应用方面，大模型在智能响应方面成熟度已达 L3 级别。当前，已有多家安全厂商将大模型与安全编排自动化及响应平台相结合，针对部分安全事件实现了智能化的决策与处置。

2.4.2 智能事件报告生成

旨在迅速记录、报告、分析和处理可能影响资产安全、运营连续性、员工安全或组织声誉的意外事故、违规行为、系统故障或潜在威胁。当前事件报告面临着高度依赖专家撰写、报告内容不够全面等挑战。

大模型凭借数据理解、摘要总结、文本生成等能力，可在自动化数据收集与初步分析、攻击过程可视化等方面发挥重要作用。一是自动化数据收集与初步分析。大模型自动搜集来自防火墙、入侵检测系统和日志服务器的相关数据，通过初步的关联分析、识别异常行为、可疑 IP 地址、恶意文件等关键信息，为报告编写提供基础素材。二是攻击过程可视化。大模型通过攻击矢量图、系统状态变迁图等图表或图形方式，直观地呈现攻击者的活动、受害系统的响应、安全防护措施的触发等，使读者快速把握事件的全貌。三是根源分析与风险评估。大模型深入分析攻击成功的根本原因，并量化评估事件对业务、数据和系统安全等方面的潜在影响。四是应对措施总结与教训提炼。大模型总结应急响应、系统恢复、漏洞修复等措施，评估有效性，并从事件中提取安全运营、风险管理、员工培训等方面的教训和改进建议。五是合规性评估。大模型确保报告内容满足法律法规的相关的要求，包括事件通报时限、数据泄露通知义务和记录保存标准等，并提出改进建议。

在行业应用方面，大模型在事件报告生成方面的成熟度已达 L4 级别。当前，市场上多数已发布的安全大模型都已集成或支持通过智能问答功能生成事件报告，这一功能已成为安全运营人员广泛采用的大模型功能之一。

2.5 安全恢复 (Recovery)

目前安全恢复在网络安全运营中实际需求较少，行业内对大模型在安全恢复环节应用的尚处于起步阶段。本报告重点介绍大模型在智能应急策略制定等少量场景的商业化应用情况。

2.5.1 智能应急策略制定

智能应急策略制定是一种先进的安全恢复方法，结合自动化工具和大模型技术，旨在当网络遭受故障或攻击导致非正常状态时，迅速采取行动恢复网络的正常运行。当前应急策略制定面临着过度依赖已有恢复方案，难以根据复杂安全事件快速生成定制化的有效恢复策略。

大模型利用其丰富的安全知识库与最佳实践案例库，为制定应急策略提供了坚实的理论基础。通过持续的学习与优化，大模型能够及时捕捉最新的威胁动态与技术进展，保证应急策略的时效性与针对性。面对安全威胁，大模型利用其卓越的数据洞察、语言理解和推理能力，根据组织的特定环境和业务需求，智能生成定制化的应急策略，并协助执行。在紧急安全事件发生时，大模型能够迅速制定应急响应策略，涵盖隔离受影响系统、封锁攻击源、恢复关键服务、收集证据等关键步骤，确保响应措施的及时性和有效性。

在行业应用方面，大模型在智能应急策略制定方面的成熟度已达 L1 级别。当前，大模型在智能应急策略制定领域的应用仍处于实验室技术攻关阶段，市场上成熟的应用案例仍然较少。

2.6 其他

大模型除了在安全威胁的识别、保护、检测、响应和恢复各环节应用外，还可在智能安全问答等对基础场景中发挥重要作用。

2.6.1 智能安全问答

安全问答作为网络安全领域的一项重要业务形态，旨在通过人机交互界面或智能机器人，帮助开发、安全服务与运维人员快速获取所需知识与数据，从而提升工作效率。目前智能安全问答面临着过度依赖于已有问答库，知识更新慢、扩展能力不足，用户交互体验不佳等挑战。

大模型凭借文本理解、文本生成等能力，可在精准理解与解答、上下文感知与个性化推荐等方面发挥重要作用。一是精准理解与解答。大模型能够准确把握用户提出的网络安全相关问题，无论是技术细节、政策法规、最佳实践还是特定场景下的应对策略，都能提供精确且有针对性的答案。二是上下文感知与个性化推荐。能够理解用户的提问背景、角色（如管理员、开发者、普通用户）、关注点以及历史交互记录，提供高度匹配用户需求的答案，并推荐相关的学习资源、解决方案或专家意见。三是实时更新与热点追踪。通过采用增强检索生成（Retrieval-Augmented Generation,RAG）、知识图谱等技术，大模型能够从外部知识库中检索关联信息，

如网络安全资讯、漏洞公告和威胁情报等，加快知识更新速度，确保生成的结果更契合用户的实际需求，有效避免产生不实或偏离事实的信息。四是多轮对话与引导式咨询。支持与用户进行多轮交互，通过追问、澄清和引导等方式，逐步深入理解用户问题的本质，提供更精细、全面的咨询服务。

在行业应用方面，大模型在安全问答方面的成熟度已达 L4 级别。当前，安全问答已成为用户与大模型交互的主要方式，几乎所有市面上发布的安全大模型都无一例外地整合了这一功能。此外，安全问答功能功能易于通过 API 接口或插件形式无缝集成到现有的安全产品生态系统中。

3. 大模型赋能数据安全

由于数据安全技术保护体系尚处于构建完善中。本报告重点介绍大模型在数据分类分级、APP（SDK）违规处理个人信息检测等少量场景的商业化应用情况。

3.1 自动化数据分类分级

数据分类分级是一种必备的数据治理方法，旨在依据数据的性质、内容、来源、用途等属性将其归入相应的类别，同时根据数据的敏感性和安全风险级别进行分级。目前，该技术面临着难以准确识别非结构化数据、难以自学习分类分级规则等挑战。

大模型通过自动化学习行业数据安全标准及已有分类分级的样例数据，或依据人工设置的规则提示，能够从海量非结构化数据源中准确识别并提取关键特征，实现数据的自动化分类分级。例如，大模型通过学习医疗数据，能自动化学习到应将患者病历归类为“健康信息 - 极高敏感”，医生处方归类为“医疗处方 - 高敏感”，患者满意度调查问卷归类为“非诊断数据 - 低敏感”。而且，对于结构化数据，大模型通过学习行业规范、标准及人工标注数据，能实现对数据库表名、字段名、注释和示例等信息的精确解读，可大幅提高数据分类分级的准确度。

在行业应用方面，大模型在数据分类分级方面的成熟度已达 L2 级别。目前，大模型在分类分级标准自学习、非结构化数据识别等方面已显示出显著成效。

3.2 自动化 APP（SDK）违规处理个人信息检测

APP（SDK）违规处理个人信息检测技术旨在识别 APP、软件开发工具包（Software Development Kit, SDK）、小程序中是否存在违反个人信息保护法规的行为。通过沙箱、深度包检测等技术，检测并报告个人信息的违规收集、使用和共享情况，并根据相关法律法规与标准进行评估。目前该领域面临着难以准确理解和自动适应复杂的合规要求等挑战。

大模型可在智能问答、个人信息识别、隐私政策分析、潜在问题发现及检测报告生成等方面为 APP（SDK）违规处理个人信息检测提供有力支持，能帮助开发者更好遵循个人信息保护原则。一是智能问答服务。大模型通过学习大量法律法规和标准规范文能够提供易于理解的法规解读，针对个人信息保护相关的政策疑问提供指导，促进合规开发。二是个人信息识别。利用其在文本、图像和音频中识别个人信息的能力，大模型能够快速扫描 APP、SDK 和小程序，准确定位并提示存在个人信息。三是隐私政策分析。大模型能够理解和评估隐私政策的合规性，包括政策的透明度、完整性以及用户知情同意等方面是否符合法律法规和标准规范。四是潜在问题发现。基于对大量安全案例的学习，大模型能够识别 APP、SDK 和小程序中的潜在隐私问题，例如个人信息的过度收集或未经同意的使用等。五是自动化检测报告生成。大模型能够自动编制详细的检测报告，明确列出问题、问题类型、严重程度以及建议的解决方案等，帮助开发者快速识别并解决 APP 中的隐私问题。

在行业应用方面，大模型在 APP（SDK）违规处理个人信息检测方面的成熟度已达 L3 级别。目前，中国信息通信研究院已推出“智御”大模型，提供政策标准解读、合规开发指导、公共服务平台使用咨询、常见问题解答等智慧问答服务，以人工智能技术推动 APP 个人信息保护的合规化进程。

4. 大模型赋能内容安全

大模型在内容安全领域具有重要应用价值。本报告重点介绍大模型在智能文本内容安全检测、智能图像视频内容安全检测、智能音频内容安全检测等场景的商业化应用情况。

4.1 智能文本内容安全检测

文本内容安全检测是指对文本信息进行自动化的审查和分析，旨在识别、标记、过滤或阻止文本中可能包含的违法或不良信息。目前该领域面临着文本表述形式复杂多样，违法不良信息变种众多等挑战。

大模型融合了丰富的社会常识、法律法规知识以及伦理道德规范等，能够迅速识别与特定领域或情境相关的不安全文本内容。而且，大模型能深入理解文本的多层次含义，包括字面意义、隐喻、讽刺、暗示等复杂表达方式，以准确判断文本是否存在潜在违规、不良或敏感内容。例如，在论坛或博客平台，用户可能发布看似无害，实则隐含极端政治立场的文章。大模型能够洞察文字背后的深层含义，识别其潜在的煽动性和危害性，触发内容审核机制，防止这类信息误导公众。同样，在直播平台的弹幕评论区，大模型能够实时监控用户发送的每一条弹幕，迅速识别并屏蔽含有谩骂、人身攻击或恶意刷屏等不党内容，以维护健康的直播环境。

在行业应用方面，大模型在文本内容安全检测方面的成熟度已达 L2 级别。鉴于社交媒体平台的特性与监管需求，大模型在过滤社交媒体上的不良信息方面表现卓越，同时其应用也扩展到了电子商务和企业信息安全管理体系中。

4.2 能图像视频内容安全检测

图像 / 视频安全检测通过计算机视觉与深度学习技术对视觉内容进行深入分析，旨在识别并过滤色情、暴力场景等不适宜的内容。目前该领域面临着 AI 生成内容以假乱真、人类和工具难以准确识别等挑战。

大模型利用其强大的数据处理、多模态识别分析能力，能够高效识别异常和伪造内容，显著提升图像视频内容安全检测的准确性和效率。在图像内容方面，大模型通过捕获局部特征以识别违规元素。在视频内容方面，大模型不仅捕捉时间维度上的动态变化，还结合 Transformer 模型的全局注意力机制，以高效追踪潜在的违规行为，理解复杂的视频场景和隐匿信息。例如，在社交媒体平台上，大模型能够准确识别用户上传的图像中是否包含

血腥、裸露、自残等敏感视觉元素，并及时进行标记和限制传播，从而保护未成年人和易感人群。此外，大模型还能够识别图像和视频中的深度伪造痕迹，如换脸、合成人物、篡改场景等，而且能够有效检测 AI 生成的图像和视频。这些内容具有高度逼真性，可能误导公众，威胁公共安全和社会秩序。通过大模型的高级识别技术，可以揭露并防范这些虚假信息的传播，保护社会免受其负面影响。

在行业应用方面，大模型在图像 / 视频安全检测方面的成熟度已达 L2 级别。目前，大模型在社交媒体内容审核、数字媒体合规审查及版权监测等领域已有较为明显的应用效果。

4.3 智能音频内容安全检测

音频内容安全检测通过语音识别与自然语言处理技术，对含有不良言语、仇恨言论或其他不当内容的音频进行有效识别和过滤。目前该领域面临着语音表述方式灵活多样，违法词语占比很少难以准确等挑战。

大模型不仅能深入解析音频数据，直接识别异常语音内容，还能将音频转化为文本进行进一步的深度分析，以精准捕捉攻击性言论或隐晦的暗示。此外，大模型还能够捕捉语音中的语调、语速和情绪等席位特征，并与已知的不良内容和情绪模式进行匹配，从而实现精准过滤。例如，通过分析音频中的说话节奏、音调变化等特征，并结合上下文理解，大模型可以识别潜在的威胁或不当行为，如辱骂或威胁性言论等。同时，它还能够分辨音频中是否包含合成语音，以防止利用语音合成技术进行欺诈、身份冒充或散布虚假信息。

在行业应用方面，大模型在音频内容安全检测方面成熟度已达 L2 级别。目前，大模型在音频内容安全检测方面的应用，在社交媒体平台，尤其是视频直播和在线游戏场景中，表现出了特别显著的效果。

大模型
安全展望

1. 大模型技术产业展望

2. 大模型自身安全展望

3. 大模型赋能数据安全

4. 大模型赋能安全展望

4

1. 大模型技术产业展望

展望未来，大模型技术将从实现与人类社会无障碍交互向跃迁至深刻理解并有效改造数字世界和物理世界的阶段。当前，语言大模型已突破性掌握了人类语言的准确理解和连贯生成，实现了与人类间的无障碍交互。在可预计的未来，多模态大模型将整合图像、视频、音频等多元感知信息，实现其全面理解与精准生成，这标志着大模型对物理世界的认知达到新高度。更进一步，通过融合智能体和具身智能技术，大模型将具备操控软件工具及实体行动的能力，从而在人类塑造数字世界和物理世界的进程中扮演不可获取的角色。

大模型产业正逐步从单一的大模型技术研发焦点，转向全面赋能各行业及催生新兴领域的转型之路。目前，全球大模型企业仍以追求技术性能的领先为核心目标，而大模型的产业实践和商业价值挖掘尚处于探索阶段。然而，未来的图景已然清晰：当大模型的基础性能满足实际应用标准后，针对金融、能源、教育、交通等多元化行业场景，研发定制化大模型并实现规模化部署，将成为业界普遍追求的业务重心。与此同时，大模型与机器人、物联网、汽车等领域的深度融合，将激发产业创新活力，衍生出诸如自主机器人、智能穿戴设备、全自动驾驶汽车等颠覆性产品，开辟全系的经济增长点。

2. 大模型自身安全展望

未来，随着大模型技术能力日益接近人类并在经济社会中得到广泛应用，可能会对国家社会秩序带来严重冲击。首先，当社会信息主要由大模型生成时，获取社会真相的成本将急剧升高。随着多模态大模型技术的成熟和广泛应用，互联网上超 90% 的信息可能由大模型生成，这使得少数不法分子利用大模型进行歪曲事实、操作舆论的行为变得更加隐蔽且难以察觉，导致普通民众越来越难以辨别真相。其次，当社会工作大量由大模型参与完成时，人类自身和物理环境的安全可能面临威胁。智慧金融智能体、自主机器人、全自动驾驶汽车等大模型系统设备的非正常运行可能直接危害人类的生命健康和财产安全。同时，应用于农业、化工、核工业等领域的大模型系统设备如果非正常运行或遭受攻击，可能会对土壤、海洋、大气等环境安全造成破坏。

为全面有效应对大模型安全风险，未来需从构建层次化治理体系和创新安全保护技术两个方面同时发力。在治理体系构建方面，应通过国际、区域和国家三个层面，针对不同层次的问题进行分层解决。在国际层面，以联合国为中心，围绕大模型的突出风险和治理原则等问题，建立全球共识的治理框架，以促进跨国界的威胁信息共享和治理政策的协同。在区域层面，依托区域联盟、经济共同体等国际组织，结合本区域内技术产业发展特点和治理需求，制定相应的区域治理法案或指南。在国家层面，各国政府需根据本国国情，制定本国治理法规和日常监管措施。此外，针对模型弱鲁棒性、模型“幻觉”等大模型安全风险，仍需从改进大模型自身技术机理，发展大模型价值对齐、大模型生成信息检测等安全技术，以确保安全问题的解决。

3. 大模型赋能安全展望

短期来看，大模型将显著提升现有安全技术的性能和智能化水平。得益于大模型在数据理解、意图识别、任务编排等方面的能力，在安全问答、安全运营、数据分类分级、违规处理个人信息检测、音视图文内容安全检测等关键网络安全场景中，大模型能够在大幅减少人工参与的同时，有效提升安全事件处理的效率和准确性。

长期来看，大模型有潜力成为安全防护的核心，从而改变安全的工作模式。当前，大模型主要扮演安全从业人员的辅助工具，用于提高他们的工作效率和效能。未来，随着大模型在自主研判和决策能力方面的提升，它们预期将进化为安全从业人员的合作伙伴，共同应对安全风险的识别、防御、检测、响应和恢复等一系列复杂工作。此外，大模型在数据安全、内容安全等领域也将发挥关键作用。大模型预计将引领安全工作模式的变革，从依赖安全人员调度和使用安全工具，转变为以大模型为核心调度并智能化使用安全工具。

编制说明

本研究报告自 2024 年 2 月启动编制，经历了前期研究、编制启动、框架确定、文稿起草、征求意见、专家评审、修改完善七个阶段，同时面向大模型技术供应方和大模型赋能安全应用方开展了深入调研。本报告由阿里云计算有限公司和中国信息通信研究院安全研究所联合撰写，负责核心章节撰写和报告内容统稿。

参编单位：阿里巴巴（中国）有限公司、阿里巴巴达摩院（杭州）科技有限公司、上海商汤智能科技有限公司、北京数安行科技有限公司、杭州安恒信息技术股份有限公司、北京快手科技有限公司、三六零科技集团有限公司、启明星辰信息技术集团股份有限公司、北京百度网讯科技有限公司、亚信安全科技股份有限公司、北京天融信网络安全技术有限公司、东软集团股份有限公司、深圳市腾讯计算机系统有限公司、山石网科通信技术股份有限公司、蓝象标准（北京）科技有限公司、蚂蚁科技集团股份有限公司、北京东方通网信科技有限公司、天翼安全科技有限公司、天翼电子商务有限公司、浙江省经济信息中心、淘宝（中国）软件有限公司、中电信人工智能科技（北京）有限公司、慧盾信息安全科技（苏州）股份有限公司、上海观安信息技术股份有限公司、中国科学院信息工程研究所、荣耀终端有限公司、合肥高维数据技术有限公司、华信咨询设计研究院有限公司、新华三技术有限公司、绿盟科技集团股份有限公司、奇安信科技集团股份有限公司。

FOUNDATION MODEL SAFETY RESEARCH REPORT