通用人工智能 AGI 等级保护白皮书

（2023 版）

网络安全等级保护与安全保卫技术国家工程研究中心

2023 年 7 月

人工智能产业链联盟

星主： AI产业链盟主

知识星球

微信扫描预览星球详情

引言

党的二十大提出，要建设现代化产业体系，推动战略性新兴产业融合集群发展，构建人工智能等一批新的增长引擎。2023 年 4 月 28 日，中共中央政治局召开会议，指出要重视通用人工智能发展，营造创新生态，重视防范风险。

当前，通用人工智能已成为继移动互联网技术之后最大的一波技术浪潮，而预训练大模型作为人工智能从专业智能走向通用智能的关键技术，在全球范围内引发科技巨头争相布局、掀起创业热潮的链式反应，同时也在技术合规、网络安全、隐私保护及社会伦理法律等方面带来了新的风险和挑战。只有有效应对这些风险和挑战，才能使通用人工智能真正服务于社会和产业，确保国家安全、企业安全和个人安全不受威胁。

为贯彻落实国家发展新一代人工智能的决策部署，全面贯彻总体国家安全观，公安部第三研究所充分发挥在网络等级保护领域国家队的优势，并依托国家级创新平台——网络安全等级保护与安全保卫技术国家工程研究中心，成立通用人工智能安全工作组，旨在重点研究通用人工智能发展现状、安全风险、网络安全等级保护合规需求以及赋能等级保护场景等，并初步形成《通用人工智能等级保护白皮书（2023 版）》，全面介绍当前公安部第三研究所在通用人工智能等级保护工作方面研究进展情况，分享相关工作经验，为推动我国人工智能高质量发展和全方位各领域高水平应用保驾护航。

术语和定义

1、人工智能 Artificial Intelligence，AI

人工智能是一门研究如何使计算机能够模拟、模仿和执行人类智能活动的科学与技术。它涉及构建智能系统，使其能够感知、理解、学习、推理、决策和交互，以解决复杂的问题和执行各种任务。

2、通用人工智能 Artificial General Intelligence，AGI

通用人工智能是一种能够像人类一样拥有智能、学习、推理、解决问题和适应新环境的人工智能系统，也称为强人工智能。AGI 的目标是实现人类智能的所有方面，包括感知、认知、思考、学习和创造等。

3、狭义人工智能 Artificial Narrow Intelligence，ANI

狭义人工智能是一种专注于解决特定问题或执行特定任务的人工智能系统，也称为弱人工智能。ANI 的能力通常是基于预定义的算法和规则，无法像人类一样适应新环境或解决未知问题。

4、超级人工智能 Artificial Super intelligence，ASI

超级人工智能是一种超越人类智能的人工智能系统，也称为强人工智能。目标是实现比人类智能更高的智能水平，能够解决人类无法解决的问题。

5、预训练大模型 Pre-trained Large Mode，PLM

预训练大模型是指在大规模数据集上进行预训练的模型。预训练是指在一个任务或数据集上进行初始训练，以学习通用的特征表示或模型参数。

6、模型即服务 Model as a Service，MaaS

模型即服务是一种基于云计算的 AI 服务模式，它提供了一种将机器学习和深度学习模型转换为可重复使用的服务的方式。

目 录

1 通用人工智能发展现状

本章主要介绍通用人工智能的基本情况，包括其发展历史、现状以及组成架构等内容。本文还将通过从技术角度出发来分析通用人工智能在网络安全领域的应用趋势。

人工智能（Artificial Intelligence, AI）的出现和发展具有一个漫长的历史过程。简单来说，从 20 世纪 50 年代开始，科学家们就开始了对人工智能的研究，并提出了一些理论和算法，但由于技术和计算能力的限制，最初的人工智能应用相对较为简单，甚至受到了一些挫折和停滞。到了 21 世纪初，随着计算机硬件技术的快速发展，加之大量数据的可用性，以及机器学习、深度学习等新技术的兴起，人工智能得以迎来了全面的发展。

当前，人工智能大致可分为狭义人工智能（ANI）、通用人工智能（AGI）和超级人工智能（ASI）。

狭义人工智能（ANI）指的是一种专注于解决特定问题或执行特定任务的人工智能系统，也称为弱人工智能。ANI 的能力通常是基于预定义的算法和规则，无法像人类一样适应新环境或解决未知问题。狭义人工智能的实现基础是通过收集和分析大量的数据，利用机器学习、深度学习等技术来训练网络模型，从而使得机器可以从输入的数据中自动提取特征和算法，进而给出相应的输出结果。其目前应用非常广泛，如语音识别、图像识别、自然语言处理、推荐算法、智慧城市、智能家居、机器人等领域中都有大量的应用案例。

通用人工智能（AGI），也称为强人工智能或者深度人工智能，是一种能够像人类一样拥有智能、学习、推理、解决问题和适应新环境的人工智能系统。与目前较为常见的 ANI 相比，AGI 具有思考、理解、学习和应用其智能来解决复杂问题的能力。目前，由于技术的限制和人工智能领域的研究方向问题，实现 AGI 仍然是一项挑战性任务。如果能够实现 AGI，可以为解决许多关键问题提供极大帮助，例如医疗和健康、环境保护、教育、金融和安全等领域。但是，要实现 AGI，需要克服诸如数据获取、算法设计、计算性能、伦理和法律等多种难题。

超级人工智能（ASI）是一种超越人类智能的人工智能系统，也称为强人工智能。它具备极高的智慧水平和创造力，能够在不同领域中做出复杂的判断、解决问题并创造出新的知识。其不仅能理解人类的情感和经历，还能唤起人类的情感、信念和欲望。与狭义人工智能和通用人工智能不同，超级人工智能更像是一种人工生命形态，而非纯技术的产物。

人工智能的发展也带来了许多挑战和讨论，包括数据隐私、伦理道德、人机关系等方面的问题。然而，人工智能的应用前景仍然广阔，它有望在各个领域带来更多的创新和突破，为人类生活和社会发展带来积极的影响。

1.1 通用人工智能的组成架构

本小节将主要介绍通用人工智能的组成架构，包括平台层、模型

层、数据层和应用层等，并阐述每一层的作用和主要内容。本文还将介绍各层之间的关系以及它们的合作模式，以便把握 AGI 的整体运行机制。

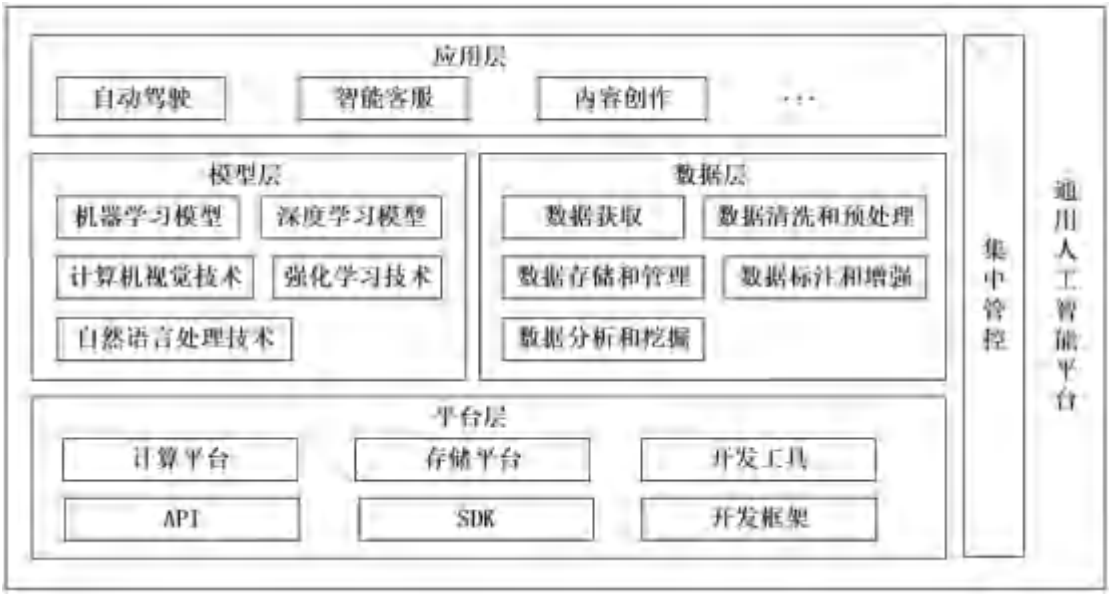

图 1 通用人工智能平台框架

1.1.1 平台层

平台层是指用于搭建、训练和部署 AGI 模型的基础设施和工具集合。这一层的目标是为开发者提供一个高效、灵活、可扩展的环境，以支持不同类型的 AGI 应用程序和服务。

通用人工智能的平台层通常包括计算平台、存储平台、开发工具和框架、API 和 SDK，这些平台可以是云端的、本地的或是混合的。

其中计算平台用于处理 AGI 模型训练和推理所需的计算任务，通常采用 GPU、TPU 等专门加速器进行计算。

存储平台用于存储训练数据、模型参数、日志和其他文件等，支持高可用性、可扩展性和安全性。

开发工具和框架用于创建和管理 AGI 模型的代码和资源，可支

持不同的编程语言和开发框架，例如 Python、TensorFlow、PyTorch 等，开发者可以利用这些工具和框架来创建、训练和部署自己的 AGI 模型。API（应用程序接口）和 SDK（软件开发工具包）用于简化 AGI 服务的集成和使用，可提供对 AGI 功能的访问界面，例如语音识别、图像识别、自然语言处理等，可为开发者快速搭建 AGI 应用程序和服务。

平台层为 AGI 开发者提供了高效、灵活、可扩展的环境，帮助他们更好地搭建、训练和部署 AGI 模型，从而支持不同类型的 AGI 应用程序和服务。

1.1.2 模型层

模型层是指用于解决各种问题的机器学习和深度学习模型、算法和技术，其目标是探索和研究不同类型的模型，以更好地解决各种问题。模型层通常由机器学习模型、深度学习模型、计算机视觉技术、自然语言处理技术和强化学习技术几部分组成。

机器学习模型用于处理结构化数据的模型，例如回归分析、决策树、支持向量机等。这些模型适用于各种业务场景，例如金融预测、客户关系管理等。

深度学习模型用于处理非结构化数据的模型，例如图像、语音、文本等。这些模型使用神经网络来解决复杂的问题，例如图像分类、自然语言处理等。

计算机视觉技术包括各种用于处理图像和视频的技术，例如图像识别、物体检测、人脸识别等。

自然语言处理技术包括各种用于处理文本和语音的技术，例如文本分类、情感分析、机器翻译等。

强化学习技术用于训练 AGI 模型进行智能决策的技术，例如棋类游戏、机器人控制等。

常见的模型层有适用于图像分类、目标检测、自然语言处理等任务的浅层神经网络模型；适用于语音识别、文本生成、机器翻译等任务的深度神经网络模型；由一系列决策节点组成的树状结构，适用于分类和回归任务的决策树模型；将多个决策树进行集成的集成学习模型；可将高维空间映射到低维，通过超平面最大化间隔距离来进行分类的支持向量机模型；此外还有聚类模型和生成对抗网络模型等。

不同的模型和技术可以被应用于各种不同的场景，例如自动驾驶、医学影像诊断、语音助手、金融预测等。AGI 开发者可以根据需求和场景选择适合的模型和技术，从而构建更加精准、高效和智能的 AGI 应用程序和服务。

1.1.3 数据层

数据层是指用于获取、存储和处理数据的技术和工具。在通用人工智能中，数据是非常重要的资源，越是高质量、多样化、具有代表性的数据，就能够让 AGI 模型拥有更好的学习效果，从而提升其性能和效率。数据层通常包括数据获取、数据清洗和预处理、数据存储和管理、数据分析和挖掘、数据标注和增强等。

数据获取包括各种方式和渠道的数据采集和收集，例如网络爬虫、传感器、数据库和数据集等，其需要根据具体的业务需求和场景进行

规划和设计，以确保获取到的数据具有足够的多样性和质量。

数据清洗和预处理是指借助各种工具和技术对采集到的数据进行筛选、整合、去噪、归一化等处理，以使其更加适合用于机器学习和深度学习，例如 Python 的数据处理库 NumPy 和 Pandas、机器学习框架 Scikit-learn 等。

数据存储和管理是将处理后的数据存储到数据库或其他数据存储系统中，以便于后续使用和分析，常见的数据存储系统包括关系型数据库、NoSQL 数据库、分布式文件存储系统等。

数据分析和挖掘是借助各种工具和技术对各种统计学和机器学习算法获得的数据进行分析和挖掘，以发现数据中的规律和趋势，例如 Python 的数据分析库 Pandas 和可视化库 Matplotlib、机器学习框架 TensorFlow 和 PyTorch 等。

数据标注和增强是对数据特征进行显示标记和扩充的手段，以提高 AGI 模型的学习效果和泛化能力，数据标注和增强需要注意数据的准确性和可用性。

数据层是构建 AGI 模型的基础，其中的每个组件都需要花费大量的时间和精力进行设计、优化和调试。

1.1.4 应用层

应用层是最终服务于用户的层次，它主要负责将通用人工智能的算法以及数据运用到实际解决问题的应用场景中，以更好地为用户提供实用和优质的智能服务。在实际应用中，通用人工智能可以涉及多个领域、多个垂直行业和不同的应用场景，并且随着智能技术自身的

深化，其应用场景还将不断扩展。

在整个通用人工智能系统中，平台层、模型层、数据层和应用层之间相互协作。平台层提供计算资源和 AGI 库，帮助模型层快速训练和优化；数据层提供数据监控和清洗功能，确保模型层使用数据的质量和安全性；模型层提供各种 AGI 算法和模型实现，帮助应用层快速部署和使用；应用层根据业务需求和场景设计开发具有个性化特点的应用，并通过模型层为用户提供增加价值的功能。

1.2 通用人工智能参与方

本小节将主要介绍通用人工智能的参与方，包括通用人工智能服务提供方和通用人工智能服务用户方。在此基础上，本文结合网络安全责任制进行阐述。

图 2 通用人工智能服务的责任主体和组成部分

1.2.1 通用人工智能服务提供方

通用人工智能服务提供方包括 AGI 平台提供方、AGI 模型提供方、AGI 数据提供方和 AGI 应用提供方，分别对应通用人工智能平台的平台层、模型层、数据层和应用层。

AGI 平台提供方、AGI 模型提供方和 AGI 应用提供方是通用人工智能服务的重要参与方。其主要目的是为 AGI 的数据、算法以及应用提供强有力的基础设施支持，用于支持大规模、复杂的深度学习神经网络模型的训练和部署，并为广大技术人员和机构提供智能计算资源。

目前，国内外的云计算平台和超算平台已经是支撑通用人工智能的主要力量之一，包括百度 AI、腾讯 AI Lab、OpenAI 的 GPT、亚马逊 AWS、微软 Azure 等。

通用人工智能服务平台是实现 AGI 技术商业化的重要基础设施之一，它们可以帮助企业和开发者快速构建、训练和部署强大的 AGI 解决方案，促进 AGI 技术的发展和应用。同时这些平台使用强大的计算力、大数据存储和数据处理等技术，为大众提供了通用人工智能开发的基础环境。

AGI 数据提供方是指为通用人工智能服务用户方提供数据的机构或个人，他们提供的数据包含训练样本、测试样本、标注信息等。这些数据是大型模型训练和应用的关键要素，可以帮助用户构建更加准确、有效的模型。通用人工智能的数据提供者可以是政府机构、大型互联网公司、科研机构、开源社区等。这些数据提供者在推动 AGI

技术发展和商业化方面扮演着重要的角色，他们的数据为各种 AGI 应用提供了支持和基础。

这些角色之间相互合作，形成一个生态系统，推动通用人工智能的发展和应用。AGI 数据提供方为通用人工智能服务用户方提供数据，通用人工智能服务用户方通过通用人工智能服务平台构建和训练模型，并将模型应用于实际问题中。

1.2.2 通用人工智能服务用户方

通用人工智能服务用户方是指使用 AGI 技术解决实际问题的用户。用户可以是任何需要通过处理大规模、复杂数据，以及需要运用深度学习技术提升业务能力和效率的企业、政府机构、科研院所或个人。

通过使用通用人工智能技术来解决各种实际问题，包括自然语言处理、图像识别、机器视觉、文本分类、预测模型等。用户通过这些需求，与通用人工智能生态圈的其他参与方持续协作，推动通用人工智能技术应用的发展。

1.3 通用人工智能网络安全应用趋势

本小节将主要介绍目前 AGI 未来在网络安全领域的应用趋势。我们将列举 AGI 在网络渗透、漏洞扫描、预警监测、态势感知、入侵检测、风险评估、网络安全管理和等级保护方面的应用，并分析其优劣势，以期对未来的 AGI 应用趋势进行预测和研判。

1.3.1 网络渗透领域

利用通用人工智能技术，可以对网络中存在的恶意行为、异常流量和攻击事件进行实时监测和分析，提高网络安全水平；可以通过分析网络拓扑结构和行为模式，自动发现和优化攻击路径，有效提高攻击难度和安全性，减少网络被攻击的可能性和损失；可以识别恶意代码，分析其行为和特征，并自动清除感染的恶意代码，提高网络安全保障能力。此外，基于人工智能技术的用户身份识别和认证系统可以通过多种生物特征识别技术，如面部识别、声纹识别、指纹识别等，实现用户身份识别和认证，有效提高网络安全水平。

1.3.2 网络监测预警

利用通用人工智能技术，可对网络安全状态进行实时监测和分析，及时发现和预警网络安全事件，提高网络安全保障能力；可实现对大量安全情报数据的分析和挖掘，自动发现和预警新的安全威胁和漏洞，帮助机构及时采取相应的安全措施；可对网络风险进行评估和指标分析，帮助企业和机构了解其安全风险状况，并提供相应的优化建议。

1.3.3 网络态势感知

利用通用人工智能技术，可对网络中各种数据进行实时监测和分析，快速发现网络异常行为、攻击行为和安全隐患等，提高网络安全态势感知能力；可根据历史数据和模型算法，自动化预测和预警网络安全事件，帮助企业提前做好安全防范和应对措施，降低网络被攻击的概率和风险；可对大量的网络日志和数据进行挖掘和分析，从中提

取有价值和关键的信息，帮助企业了解当前网络安全状况和未来的风险趋势。

1.3.4 网络漏洞挖掘

利用通用人工智能技术，可对网络系统进行全面的漏洞扫描和分析，自动发现和分类各种漏洞，并生成相应的漏洞报告，提高网络漏洞挖掘的效率和准确性；可模拟各种不同的漏洞攻击方式，对网络安全进行评估和测试，帮助企业了解其网络安全弱点和漏洞，并提供相应的修复建议；可根据历史数据和模型算法，自动分析网络中的漏洞和弱点，并提供相应的优化建议和修复措施，帮助企业提高其网络安全防范能力。

1.3.5 网络等级保护

利用通用人工智能技术，可自动化完成网络安全风险评估和等级保护定级评估，提高等级保护执行的效率和准确性；可实时监测和预警网络威胁，包括网络攻击、流量异常和安全漏洞等，帮助企业及时发现和处理安全威胁，并自动进行阻断和修复，提高等级保护的实时性和有效性；可自主学习和优化网络安全策略和措施，不断提高等级保护的精度和强度，同时还可以实现自主学习和优化网络安全防护方案，提升等级保护的合规水平。

综上，通用人工智能在网络安全领域的应用前景十分广阔，在漏洞扫描、监测预警、态势感知、入侵检测、风险评估、网络安全管理和等级保护等各个方面均可以大显身手。

2 通用人工智能安全风险分析

2.1 传统安全风险

近年来，全球重大网络安全事件如 DDoS 攻击、勒索软件、数据泄露和供应链攻击等持续频繁发生，且变得更加难以防护和更具危害性，全球网络安全攻击态势日趋严峻。

随着通用人工智能的高速发展，资本的大量进入，人工智能资产的价值也将逐步凸显，人工智能基础设施、平台框架、算法模型和人工智能应用也逐渐成为攻击者进行网络攻击的重点目标，但由于通用人工智能仍处于发展初期，安全投入相对较少，通用人工智能系统的保密性、完整性和可用性时刻受到威胁。

当前，通用人工智能系统面临的安全风险主要包括基础设施安全风险、模型与数据层安全风险和应用层安全风险。

2.1.1 平台基础设施安全风险

通用人工智能基础设施主要由计算服务、存储服务、网络服务、容器服务、加速服务等底层服务支持层组成。跟其他领域关键信息基础设施类似，通用人工智能基础设施同样面临物理攻击、网络攻击、计算环境被篡改和运维安全等各种的安全风险。

（1）物理攻击方面。由于 IDC（Internet Data Center）机房的管控不到位，允许未经授权的物理访问，容易出现硬件被破坏、篡改、禁用、窃取的安全风险。

（2）网络攻击方面。针对通信网络的攻击形势依然相当严峻，

一方面攻击者可通过窃听、劫持等手段入侵网络，造成通信故障或直接窃取机密信息等；另一方面攻击者也可以直接发动针对基础设施网络的 DDoS 攻击，作为关键信息基础设施之一的人工智能基础设施必定首当其冲，进而导致整个人工智能系统服务不可用。

（3）计算环境的安全风险。风险一方面主要来自基础设施自身存在安全漏洞导致的系统入侵；另一方面也可能是基础设施本身与其他系统未进行严格的网络隔离，从而导致面临来自内部其他系统的横向渗透风险。如果攻击者在入侵到基础设施系统后，注入后门、木马等恶意程序，将导致整个系统后期将面临严重的安全风险。

此外，通用人工智能系统是一个极度依赖算力驱动的系统，其他如运维、变更等流程不完善的问题导致的 SLA（服务等级协议）服务水平低，也将影响到人工智能系统本身的服务性能。

2.1.2 模型与数据层安全风险

通用人工智能模型与数据层主要包括各种算法与技术框架、数据处理的技术与工具集，以及各种由该层提供的人工智能服务。模型与数据层面临的主要安全风险有供应链攻击、漏洞攻击，以及 API 安全、运维安全等风险问题。

（1）供应链攻击。当前供应链攻击日益普遍，已成为当前互联网应用最为严重的安全威胁之一。根据高德纳咨询公司 Gartner 预测，到 2025 年，全球大约 45% 的组织将遭受一次或多次软件供应链攻击。由于供应链攻击本身具有极强的隐蔽性，再加上 AGI 开发框架、大模型的普遍使用，将进一步扩大供应链攻击在人工智能系统开发中的

广泛性。

（2）漏洞攻击。随着通用人工智能的爆发式发展，传统安全厂商、专业研究机构也将更多的资源、精力投入到了对通用人工智能安全技术的研究，随着通用人工智能相关框架、算法方面的日益丰富，其安全漏洞挖掘也必然呈现逐步增长的态势。可以预见，针对通用人工智能模型与数据层的漏洞，例如框架类漏洞、算法漏洞的攻击将日益严重。

（3）API 安全问题。API 的开放一方面为企业发展提供了强有力的支撑，但同时也因为对外暴露了更多的资产而导致面临更多的外部威胁。近年来，针对 API 攻击导致的入侵事件和数据泄露事件屡屡发生，加上企业对自身发布的 API 治理不善，对发布了多少服务、有哪些服务对外开放、服务的访问控制是否到位等问题未给予足够的重视，恶意用户或攻击者可利用通用人工智能系统对外暴露的 API 接口，发起对通用人工智能系统的攻击或者窃取机密信息。

（4）运维安全问题。同通用人工智能平台基础设施一样，模型与数据层同样面临运维、变更等流程不完善导致的安全问题。

2.1.3 应用层安全风险

通用人工智能应用是基于人工智能底层服务提供的 API，以实现在各种实际场景下的落地。除了一些大企业、研究机构具备人工智能全栈能力外，更多的企业依赖这些大企业、研究机构提供的底层服务能力来开发人工智能应用。企业在开发具体人工智能应用时，除了面临着供应链攻击的风险外，也面临自身应用安全架构不合理、应用层

漏洞以及各种合规风险。

2.2 数据泄露和滥用风险

2.2.1 模型和训练数据作为核心资产保护难度提升

随着通用人工智能深度学习的网络层次加深，模型的训练成本不断提高，大模型训练数据的采集以及算法模型的设计开发已经逐渐变成了企业的重资产投入，特别是性能优良的算法模型、参数以及训练数据已经成为人工智能企业、研究机构的核心资产。另一方面，企业、高校和研究机构基于推广和使用的考虑，往往倾向于将大模型部署在云端，并以 API 的方式将大模型的相关功能开放给其他用户，以实现模型即服务（MaaS），这些算法模型、训练数据势必面临着来自互联网恶意用户的攻击、窃取的风险。

除了传统的因流程管理不善和基础设施平台被入侵导致的数据泄露风险外，大模型由于其自身的数据规模性、算法复杂性等原因同样面临着新型的数据泄露风险。恶意用户或攻击者通过模拟大量输入请求得到大量模型输出，从而逆向还原模型功能，达到窃取模型或者训练数据的目的，例如攻击者可以通过推断攻击，进而判断某个个体的数据是否存在于某个医疗诊断模型的数据训练集中，从而进一步推断该个体的隐私信息。这些算法模型、训练数据的泄露都将大大削弱企业大模型自身的商业竞争力。

2.2.2 用户数据用于训练泄露隐私、机密信息

大模型在训练过程中需要采集大量且多样性的数据，但在实际模

型开发过程当中，数据采集方往往并未遵循最小化原则，而是过多的采集了用户的隐私信息，例如在传统互联网应用当中，应用方采集用户的手机号、姓名、性别、身份证号码等个人信息以及操作记录、消费记录等行为信息，而通用人工智能应用除上述信息外，还广泛采集其他具有强个人属性的唯一生物特征信息，如声纹、虹膜、指纹等。

这些个人隐私信息都有可能被恶意用户或攻击者窃取，也有可能面临内部训练环节流程管控不到位，致使内部泄密的风险。另外，随着交互式人工智能应用的发展及普及，用户往往在不知情的情况下输入了敏感信息或者机密信息，例如员工在办公环境中使用智能问答平台时，容易将公司的商业机密信息输入平台寻找答案，而平台在获取到该机密信息后又用于自学习过程，继而导致机密信息在其他场景下泄露。

2.2.3 训练数据的滥用加大企业的合规风险

通用人工智能的发展一方面极大的促进了数字内容的繁荣，另外一方面又依赖互联网数据的反馈输入，特别是大模型的训练依赖海量的数据输入，如果通用人工智能系统的训练数据不正确、不完整，或者带有偏见性数据，将会产生大量的不良信息、歧视性信息、偏见性输出。

目前全球互联网流量每秒已超 PB 级别，传统的内容审核模式在应对如此海量数据时显得捉襟见肘。而机构在训练模型时，如果为了节省成本，对从互联网采集的数据不加清洗而全部用于模型的训练；或者虽然存在清洗环节，但是由于员工缺乏数据安全意识、员工本人

含有特定价值倾向等原因，滥用误用低质量、不完整的数据对模型进行训练，都会导致大模型输出不良信息、不符合社会主流价值观的信息等，从而引发有关机构面临监管处罚或者应用下架的风险。

2.3 模型可靠性风险

2.3.1 数据源污染导致训练效能低下、决策偏离预期

大模型的判断与决策能力来自于海量数据的训练过程。数据是大模型的养料，对模型训练过程至关重要。客观公正、完整的数据集在很大程度上影响了模型判断的准确性。

如果模型的训练数据不可控，例如采集的数据来自公共网络空间但未做清洗，或者直接使用从不可信的第三方数据源采集的数据进行模型训练，将有可能导致模型面临推理效能低下或者决策偏离预期的风险。

另一方面，由于大模型神经网络训练和推理需要使用高耗能的GPU（Graphics Processing Unit，图形处理器）和 TPU（Tensor Processing Unit，张量处理器）等加速硬件，如果用于训练的数据集是一些特殊的样本，例如海绵样本（Sponge Sample），可让大模型在极低的性能下进行推理，从而导致该服务的推理性能下降，而运行成本大幅提高，进而影响大模型的使用效率。

此外，攻击者可在污染数据源中植入毒化数据、后门数据，从而导致大模型的决策偏离预期，甚至攻击者可在不破坏模型原来准确率的同时入侵模型，使得大模型在后续应用过程中对攻击者的数据做出

符合攻击者预期的决策。

2.3.2 模型鲁棒性缺乏无法应付异常场景决策

模型的鲁棒性要求大模型有抗干扰、应付异常场景的能力，对于异常或者微小扰动的输入样本，能持续做出稳定、准确的判断与决策。模型鲁棒性缺乏主要由训练数据集不完整、算法自身的复杂性导致，其风险来自于真实物理世界的环境多变、攻击者的对抗攻击、伪造攻击等。

大模型的训练需要完整的数据集，而训练数据往往无法覆盖到真实世界的各种异常场景，因此在模型训练阶段表现良好的模型，在投入实际运营后，针对训练中未出现的、真实世界的各种异常输入无法做出准确的判断与决策，例如光照强度、视角角度距离、图像仿射变换、图像分辨率等环境因素会对模型产生不可预测的影响，这些都导致了 AGI 图像识别技术在真实场景下无法准确识别出图像。鲁棒性的缺乏可使某些场景导致严重的安全事故，例如医疗机器人在面对异常情况下做出的错误决策，可能严重威胁到病人的生命健康安全。

攻击者在对抗攻击中，可在输入样本中添加细微的人眼无法识别的干扰噪声，从而在不引起人们注意的情况下，导致通用人工智能系统给出偏离预期的错误决策，例如在自动驾驶领域，若攻击者给路标加入细微的恶意噪声，会导致无人驾驶汽车无法正确识别路标并做出错误的决策，从而引发交通事故。

此外，攻击者还可通过伪造具备个体唯一性特征的信息（指纹、虹膜、面容等），并将该信息作为智能认证系统的输入，用以通过身

份认证，实现伪造攻击的目的。例如攻击者通过收集目标对象多个场合下的视频信息，并裁剪成匹配视频认证服务的视频，借此达到通过认证的目的。

2.3.3 算法的黑箱性导致过程难解释、事件难追溯

通用人工智能大模型核心基础为深度学习，其算法结构中存在多个隐层，导致算法的输入与输出之间存在人类难以理解的因果关系、逻辑关系。对于用户来说，大模型的算法过程是一个技术黑箱，用户无法完全理解其计算推理过程，只能被动接受由模型给出的结果。此外，通用人工智能算法模型还有自适应、自学习等特性，其复杂程度更是超过了人类大脑的理解范畴。由于通用人工智能模型的不可解释性，给人工智能安全事件的溯源分析带来了严峻挑战。

2.3.4 算法的偏见性导致网络空间意识形态面临新风险

人工智能模型算法追求的是统计的最优解，本身并不具备客观公正的判断能力。模型对于价值的判断完全依赖于训练数据，而伦理、道德、政治等复杂问题本身具有地域、文化特性。

符合一个地区人群价值观念判断的人工智能，很有可能会与另外区域人群的价值观念判断相冲突。若在模型训练过程中，由于开发者自身价值观、世界观的影响，或者数据源的质量不到位，误将不符合社会主流价值观的数据作为训练数据输入到模型中，经过模型的强化学习，可能会做出带有歧视或偏见的决策，算法输出的带偏见性的价值观甚至会对我国的网络空间意识形态造成冲击。

2.4 大模型滥用误用风险

大模型的兴起加速了通用人工智能时代的来临，助推人工智能在安防、客服、教育、医疗等诸多领域当中加速落地。

人工智能深度赋能产业迭代升级，加速社会发展、提高了生产效率。但随着通用人工智能应用场景越来越多，应用被滥用误用、恶意使用的现象也愈发频繁，例如大数据杀熟、基于深度伪造的电信诈骗等时有发生，日益显著的威胁到社会、企业、人身、网络空间安全。

2.4.1 大模型滥用误用危害社会稳定

人工智能在社交媒体中的滥用、误用可扰乱正常的社交秩序，甚至可能激化社会矛盾。社交机器人是网络媒体空间中传播社交信息的一类智能机器，已广泛参与到舆论信息的讨论与扩散环节中，成为人类网络社交活动的重要参与者。

机器人运营者通过注册大量社交媒体账号，用虚假的个人身份同真实用户建立社交关系，传播运营者的诉求并力图影响网络舆论。随着人工智能的发展，社交机器人已经摆脱以往简单点赞和转发的低级模式，当前社交机器人能借助大模型挖掘用户数据并做出深入分析，如果生成的是虚假信息、负面信息，可混淆视听、左右公众舆论，甚至改变热点事件、政治事件的舆论走向，给社会带来不稳定因素。

此外，恶意用户利用深度伪造技术可以制作政治人物的虚假负面音频、视频信息，严重扰乱社会的正常公共秩序，例如国外就多次出现过利用 AGI 技术伪造政治人物负面新闻抹黑政治人物的事件。另

外，深度伪造技术也可能被一些不法分子用于欺诈、诈骗等违法犯罪活动。

2.4.2 大模型滥用降低企业创造积极性

随着生成式人工智能的发展，大量的智能生成物如小说、视频、歌曲等由人工智能自动创建，且可在网络自由复制、传播。

然而，由于部分大模型的训练数据本身采集自互联网的公开数据，再加上版权保护法律的普遍滞后，新的人工智能创造物得不到法律的有力保护，将严重影响着企业投入人工智能创造的积极性。

2.4.3 大模型滥用误用侵犯个人基本权益

通用人工智能的滥用、误用对于个人基本权益的侵犯主要表现在三个方面：

首先，通用人工智能应用使用场景的泛滥容易导致个人的隐私权益被侵犯。各种人工智能应用使用场景并未得到严格的规范与限制，应用运营方只考虑自身的功能需求，滥用人工智能技术大量采集用户的隐私信息，如指纹、人脸、虹膜等信息，严重侵害到了个人的隐私权益。

其次，通用人工智能应用的滥用，可能侵犯个人人格尊严。恶意用户使用通用人工智能技术来创建虚假的人工智能聊天机器人，用于进行网络欺诈或社交工程攻击，获取个人敏感信息或诱导用户进行不恰当的行为。这种滥用行为侵犯了用户的个人隐私和人格尊严，可能导致经济损失、心理伤害或社交破坏。

最后，算法的偏见也会影响到个人基本权益。大模型的性能依赖训练数据的质量，训练数据的完整性、算法设计者主观的情感偏向、训练数据本身包含的人类社会的固有偏见和不同地区文化差异等各方面因素，都有可能影响到大模型的输出。若在大模型设计、训练之初误用或滥用包含非客观公正、带歧视性的训练数据，大模型的偏见性输出很有可能给用户带来情感上的伤害。

2.4.4 大模型滥用危害网络空间安全

随着大模型应用功能的不断增强和普及，这些技术也逐渐成为网络安全攻击的一种新型手段。

目前，基于大规模语言的预训练模型广泛应用于网络攻击各个环节当中，例如，ChatGPT 可用于快速收集目标资产信息，生成并发送大量钓鱼邮件，也能基于目标资产指纹快速发现 NDay 漏洞，甚至可以通过扫描开源代码、泄露的代码自动检测到 0Day 漏洞。

此外，基于大规模语言的预训练模型还能根据漏洞原理自动构建漏洞利用代码，从而实现快速入侵目标系统。大模型的使用大大提高了攻击效率、降低了攻击门槛，使更多人加入到攻击者队伍当中，进一步加剧了网络空间安全风险。

3 通用人工智能等级保护合规需求

依据 GB/T 22239-2019《信息安全技术 网络安全等级保护基本要求》第三级通用要求内容，本节主要从安全物理环境、安全通信网络、安全区域边界、安全计算环境、安全管理中心、安全管理制度、安全管理机构、安全管理人员、安全建设管理和安全运维管理十个层面，对通用人工智能的等级保护合规落地需求进行阐述和分析。

3.1 安全物理环境

3.1.1 物理位置选择

通用人工智能大模型的训练、推理和应用依赖于算力大规模集中的数据中心。在物理位置选择方面，机房环境选择在雨水不易渗漏，门窗不因风雨造成尘土严重，机房周围建筑，如：屋顶、墙体、门窗和地板等设施完整良好，无破损开裂，且建筑物具有抗震审批文档。

重要行业单位的主机房或灾备机房，尽量选择在自然灾害较少的环境中，机房的位置不宜选择在顶层或地下室。

3.1.2 物理访问控制

通用人工智能大模型的训练、推理和应用依赖于算力大规模集中的数据中心。在物理访问控制方面，机房出入口是进入机房的第一道屏障，电子门禁系统的部署可以避免无关人员随意进入机房，增强机房的安全性和可靠性。

此外，电子门禁系统的另一个作用是可以对进出机房的人员进行

记录，系统一旦遭受到破坏，可以进行事件追溯。

3.1.3 防盗窃和防破坏

通用人工智能大模型的训练、推理和应用依赖于算力大规模集中的数据中心。在防盗窃和防破坏方面，为保证机房各系统正常工作，防止不法分子对机房内的设施进行盗窃和恶意破坏，机房内应装置防盗报警系统或安排专人负责监控并响应视频监控系统。

在将设备固定后，需在明显处注明标识，方便对机房内的设备进行管理和维护。机房内的通信电缆铺设在隐蔽安全处，防止通信线缆破坏后影响系统的正常运行。

3.1.4 防雷击

通用人工智能大模型的训练、推理和应用依赖于算力大规模集中的数据中心。在防雷击方面，微电子设备内部由众多微小的电子元器件组成，具有高密度、高响应、低时延和低功耗等特性，使其对雷电、过压等电力系统的影响非常敏感。为消除雷电、过压等带来的风险隐患，机房的配电柜内须装置性能好的电源二级防过电压防雷保护系统。机房的供电系统至少二级防浪涌处理，关键设备或重要负载末端也要进行防浪涌处理，机房内采用综合接地手段，相线、零线、PE 线分色应符合国家标准。

3.1.5 防火

通用人工智能大模型的训练、推理和应用依赖于算力大规模集中的数据中心。在防火方面，机房发生火灾事故通常是由电线短路、过

载发热和其他违规操作等问题造成的。如恶意破坏、纵火、自然灾害、闪电和电压不稳会极大提升机房发生火灾事件的风险，一旦发生火灾响应慢、扑救困难和破坏性大等，给单位和人员造成难以估量损失，因此采取相关措施对火灾事故进行预防和自动消防。配备符合国家相关部门标准的火灾自动消防系统，如七氟丙烷等自动灭火设备，并对设备进行定期巡检，机房内安装应急照明装置、维修工具，墙体、顶棚、壁板和隔断采用耐热不易燃的建筑材料。

3.1.6 防水和防潮

通用人工智能大模型的训练、推理和应用依赖于算力大规模集中的数据中心。在防水和防潮方面，机房一旦发生渗水、漏水、水蒸气结露等现象，会对设备造成损坏和信息丢失，电缆线路在潮湿的地方会生锈或是引起短路，给人员带来触电的风险，因此需要加强机房内的防水防潮措施，机房顶棚、地面和墙体四周要加强防水和防潮措施，窗户禁止开启，安装防水防潮检测报警系统，实时监控机房内湿度信息。

3.1.7 防静电

通用人工智能大模型的训练、推理和应用依赖于算力大规模集中的数据中心。在防静电方面，机房静电具有隐蔽性、潜在性和随机性特点，静电不仅会对电路形成干扰，导致电路中的电子元器件发生击穿和毁坏，甚至还会造成设备运行出现故障，使系统服务中断带来重大损失，因此需要对静电采取相关防护措施。机房内可铺设防静电地

板，机柜接地且插口采用三角插口，在机房门口装置人体静电消除仪器，操作人员须穿戴防静电服和防静电手环，按相关流程进行操作。

3.1.8 温湿度控制

通用人工智能大模型的训练、推理和应用依赖于算力大规模集中的数据中心。在温湿度控制方面，机房内的设备大多采用的是微电子电路、精密机械设备等组合而成，而这些设备由于自身属性，易受到温度、湿度的影响，影响设备自身的功能及寿命，因此，需采用相关措施保证设备在机房内运行的稳定性和可靠性。

机房内装置可调节温湿度的空调设备，空调设备数量及部署须根据机房内的实际情况安装。保持机房内温度不宜超过 25℃，相对湿度范围应在 45%-65%。如果设备自身能够在所处环境的温湿度中正常运行，可以不使用温湿度调节措施。

3.1.9 电力供应

通用人工智能大模型的训练、推理和应用依赖于算力大规模集中的数据中心。在电力供应方面，机房内电气设备繁多，如果供电系统不能正确、稳定的运行，可能会造成严重的后果，轻则业务系统中断，重则造成生命财产损失。因此，需采取相关措施对供配电进行安全管理。通过配备稳压器和过压防护等电气设备，以此保证线路电压突然发生变化时不影响设备正常运行。通过配备 UPS 设备保证在突然断电情况下提供不间断电源，维持系统在一定时间内的正常运行。机房电源系统采用冗余方式，由两路不同形式的供电线路进行供电。

3.1.10 电磁防护

通用人工智能大模型的训练、推理和应用依赖于算力大规模集中的数据中心。在电磁防护方面，为消除线路之间的电磁干扰，机房内电源线和通信线缆进行隔离铺设。此外，计算机、显示器等电子设备在正常运行时，不可避免的会产生电磁波，该电磁波可能携带计算机正在处理的数据信息，存在会被专用设备窃取其信息的风险，因此关键设备（核心网络设备、涉密设备等）需要进行电磁防护。

强电和弱电需要进行隔离铺设，通常情况下强电走地板下，弱电走高架桥，或强弱电均走高架桥，但不同桥架分开，或均走地下，但需要采用不同线槽，线路之间的距离应不小于 15cm。铁质线槽可有效防护电磁干扰且可以保护线缆安全。机柜柜门应使用防外界电磁干扰的屏蔽门。

3.2 安全通信网络

3.2.1 网络架构

通用人工智能大模型的训练和推理依赖于算力大规模集中的数据中心，对算力硬件设备的高速互联要求较高，要求存储分层、计算分层和弹性容错。在网络系统中，通常网络架构可以划分为不同的区域，各区域之间采用技术措施进行网络隔离，且在网络出口处配备防火墙、隔离网闸等安全设备进行防护。根据实际业务需求特点和安全防护要求划分为不同的安全域，并根据方便管理及控制的原则分配地址，在安全域之间设置防火墙等安全设备进行防护。

对可用性和实时性要求较高的系统。网络区域中关键业务网络节点满足高峰期业务处理能力和网络带宽。服务器对外提供数据信息的，要增强对协议的访问控制。重要的网络层级要保证通信线路和网络设备的硬件冗余，DCS 设备、重要通信设备等应实现双网冗余或环网功能。

3.2.2 通信传输

通用人工智能大模型的训练和推理依赖于算力大规模集中的数据中心，容易发生窃听、劫持手段入侵网络，造成通信故障或信息泄露，或直接对基础设施进行 DDoS 攻击。在通信传输过程中，采用密码技术或校验技术来保护数据的完整性，采用密码技术来保证数据的保密性，防止关键信息被第三方窃取、篡改、破坏。采用提供通信加密和通信校验的网络系统，开启加密的通信协议。对访问请求进行完整性及权限合法性校验，仅允许通过校验的请求访问本地安全隔离环境中的数据等。

通信过程中的数据可以使用奇偶校验、累加校验或 CRC 校验等方式进行数据校验。对于关键数据，如鉴别信息等应采取密码技术对其保护，WEB 服务可以采用 HTTPS 协议进行防护，当采用广域网对网络系统中的设备进行数据或指令通信时，应采用加密认证技术措施完成身份认证、访问控制和数据加密传输，例如 VPN 等方式。

3.2.3 可信验证

通用人工智能大模型的训练和推理依赖于算力大规模集中的数

据中心，对算力硬件设备的高速互联要求较高，要求存储分层、计算分层和弹性容错。通信设备如交换机、路由器等具备可信验证功能，基于可信根，构建可信链，对通信设备的系统引导程序、系统程序、参数、应用等进行验证。对应用程序业务关键调用等重要操作进行实时动态可信验证，对违规操作进行记录、告警，并生成日志，发送至安全管理中心。

3.3 安全区域边界

3.3.1 边界防护

通用人工智能大模型的训练和推理需要大规模数据交换，网络边界需要具备足够带宽来支持高速数据传输，需要采取措施保证不会对模型的训练和推理过程造成威胁。网络系统中不同层级和安全域通常位于不同的网络区段，采用防火墙或网络隔离设备进行安全防护。跨越边界的通信需要通过受控接口，并以最小化开放原则，只允许采用专有协议通过，并对协议内容规范进行深度检测。未授权的设备通常带有未知病毒等恶意代码，任意连接内网会造成病毒传播或扩散的风险，因此需要采取技术手段对此进行检查或限制。

3.3.2 访问控制

通用人工智能大模型的训练和推理需要大规模数据交换，网络边界需要具备足够带宽来支持高速数据传输，需要采取措施保证不会对模型的训练和推理过程造成威胁。设定访问控制策略向不可信设备拒绝一切通信，这是一种比较安全的防护方法。对于受控接口内的通信，

还需根据源地址、目的地址、源端口、目的端口、应用协议和应用内容等方面进行检查，提供允许/拒绝访问的能力。核查设备中访问控制策略，删除多余或无效的访问控制规则，增强网络隔离设备的实时性，提高工作效率。

3.3.3 入侵防范

通用人工智能大模型的训练和推理需要大规模数据交换，网络边界需要具备足够带宽来支持高速数据传输，需要采取措施保证不会对模型的训练和推理过程造成威胁。入侵防范是一种可识别潜在的威胁并迅速地做出应对的网络安全防范办法。入侵防范技术作为一种积极主动的安全防护技术，提供了对外部攻击、内部攻击和误操作的实时保护，在网络系统受到危害之前拦截和响应入侵。入侵防范在不影响网络和主机性能的情况下能对网络和主机的入侵行为进行监测，在发生严重入侵事件时提供报警。

3.3.4 恶意代码和垃圾邮件防范

通用人工智能大模型的训练和推理需要大规模数据交换，网络边界需要具备足够带宽来支持高速数据传输，需要采取措施保证不会对模型的训练和推理过程造成威胁。通用人工智能大模型的应用可以带来批量生产钓鱼邮件，还能伪造证书、凭据、身份等风险。在网络系统的关键网络节点处部署恶意代码防范产品时要保证系统的业务通信性能和连续性，可采用白名单形式的恶意代码防范产品，并且应以

最小化原则，只允许网络系统中使用的专有协议通过。

3.3.5 安全审计

通用人工智能大模型的训练和推理需要大规模数据交换，网络边界需要具备足够带宽来支持高速数据传输，需要采取措施保证不会对模型的训练和推理过程造成威胁。安全审计是通过一定的安全防护策略，利用记录、系统活动和用户活动等信息，检查、审查和检验操作事件的环境及活动，从而发现系统漏洞、入侵行为或改善系统性能的过程，也是审查评估系统安全风险并采取相应措施的一个过程，是提高系统安全性的重要举措。

安全审计不但能够监视和控制来自外部的入侵，还能够监视来自内部人员的违规和破坏行动，为网络违法与犯罪的调查取证提供有力支持。在第三级系统中，安全审计加强了对于远程连接到互联网用户行为的审计日志单独审计，记录其操作和访问日志，并进行数据分析。

3.3.6 可信验证

通用人工智能大模型的训练和推理需要大规模数据交换，网络边界需要具备足够带宽来支持高速数据传输，需要采取措施保证不会对模型的训练和推理过程造成威胁。可信验证是基于可信根，构建可信链，一级度量一级，一级信任一级，把信任关系进行传递的过程，从而保证设备运行过程和启动过程的可信。针对可信验证过程中产生的可信性被破坏的行为需与安全管理中心建立报警和审计机制。

3.4 安全计算环境

3.4.1 身份鉴别

当访问控制或身份验证机制未正确实施时，会出现访问控制不足，从而允许未经授权的用户与通用人工智能大模型交互并可能利用漏洞。未能对访问通用人工智能大模型执行严格的身份验证要求，基于角色的访问控制(RBAC)实现不充分允许用户执行超出其预期权限的操作，以及未能为通用人工智能大模型生成的内容和操作提供适当的访问控制都是常见的例子。通用人工智能大模型在身份鉴别方面应该满足下述要求：

（1）身份鉴别是网络业务应用系统（软件系统）、应用服务器和终端以及计算机网络系统中确认操作者身份的过程，为确保安全，确定该用户是否具有对某种资源（包括操作系统和数据库系统等）的访问和使用权限，需要对登录的用户进行身份标识和鉴别。

（2）对于在网络系统、网络应用服务器和终端以及计算机网络系统中登录的用户，为确保用户身份是可信的，对操作系统、应用程序登录失败的用户应启用结束会话、限制非法登录次数等措施。

（3）在网络应用服务器和终端以及计算机网络系统以远程管理方式进行用户登录的过程中，应采取传输信息加密的方式确保鉴别信息的保密性。

（4）应使用两种或两种以上组合的鉴别认证方式加强身份鉴别的可靠性和安全性，通过统一的认证接口，采用认证加固功能和密码

策略，在保证便捷性的同时加强认证环节，实现组合的鉴别技术对用户进行身份鉴别。

3.4.2 访问控制

通用人工智能大模型应用同样存在对用户的访问进行控制。在访问控制方面，对于默认账户，通常是具备一定权限的管理账户，必须要重命名并修改默认密码，不可以使用系统默认名和默认口令，这些都是高风险项。对于多余、无效、长期不用的账户要及时删除，建议定时（每周、每月、每季度等）针对无用账户进行清理。不能存在共享用户，即一个账户多人或多部门使用，这样不便于审计，出现事故无法准确定位故障点，不能进行追责。

实施严格的输入验证和清理流程，以防止通用人工智能大模型处理恶意或意外提示。确保充分的沙盒隔离并限制通用人工智能大模型与底层系统交互的能力。防止攻击者通过自然语言提示利用通用人工智能大模型在底层系统上执行恶意代码、命令或操作时，发生未经授权的代码执行。

3.4.3 安全审计

通用人工智能大模型的应用应进行审计和记录。对用户进行安全审计，主要安全目标是为了保持对系统用户行为的跟踪，以便事后追溯分析。

应关注用户操作日志和行为信息的安全审计，审计覆盖范围应覆盖到服务器、防火墙、交换机等每个系统资源相关用户。对用户登录

信息进行审计，能够及时准确地了解和判断安全事件的起因和性质，并对审计记录必须提供有效的安全保护措施。

3.4.4 入侵防范

通用人工智能大模型的训练和推理依赖于大模型的框架和各种组件，面临着注入后门、恶意程序等攻击风险，需要为不同用户的不同训练、推理任务建立 GPU 侧安全隔离环境，包括但不限于基于容器的隔离、基于虚拟机的隔离等。

网络安全计算环境的入侵防范，主要安全目标是防止服务器、终端等设备和其它网络系统计算资源由于自身存在的操作系统漏洞、应用程序后门、系统服务、默认端口等安全风险遭受外来非法入侵，从而导致生产数据被破坏、操作系统崩溃、设备宕机、损坏甚至危及人身安全等安全事件此外，还可以通过对抗样本工具箱来进行测试等技术手段，帮助开发者和测试人员发现模型的薄弱点和漏洞，从而针对性地进行加固和优化，提高模型的健壮性和安全性。

（1）企业或集成商在进行系统安装时，遵循最小安装原则，仅安装业务应用程序及相关的组件；

（2）企业或集成商进行应用软件开发时，需要考虑应用软件本身对数据的符合性进行检验，确保通过人机接口或通信接口收到的数据内容符合系统应用的要求；

（3）企业或集成商在选择主机安全防护软件时除了要考虑主机安全防护软件的安全功能以外，还要考虑与实际业务场景结合的问题，能够有效的帮助业主解决实际问题。主机安全防护软件应可以通过最

简单的配置来满足等级保护的要求；

（4）解决安全漏洞最直接的办法是更新补丁，企业可委托第三方安全厂家对系统进行漏洞的扫描，发现可能存在的已知漏洞，根据不同的风险等级形成报告，企业或集成商根据报告在离线环境经过测试评估无误后对漏洞进行修补。

此外，通用人工智能大模型结合提示工程可以用于创建更智能、交互式的自动化助手，这就衍生出提示词注入风险。提示词注入是指绕过过滤器或使用精心设计的提示词来操纵通用人工智能大模型，使模型忽略先前的指令或执行意外操作。这些漏洞可能导致意想不到的后果，包括数据泄露、未经授权的访问或其他安全漏洞。常见的即时注入漏洞包括通过使用特定的语言模式或标记来绕过过滤器或限制，利用通用人工智能大模型标记化或编码机制中的弱点，以及通过提供误导性上下文误导通用人工智能大模型执行意外操作。预防措施可包括对用户的提示词实施严格的输入验证和清理、使用上下文感知过滤和输出编码来防止提示词操纵、定期更新和微调通用人工智能大模型，以提高其对恶意输入和边缘情况的理解。

3.4.5 恶意代码防范

通用人工智能大模型的训练和推理依赖于大模型的框架和各种组件，面临着注入后门、恶意程序等攻击风险。在恶意代码防范方面，服务器和终端操作系统易遭受主机病毒威胁，木马和蠕虫泛滥，防范恶意代码的破坏尤为重要，应采取病毒查杀、系统防御、系统加固、日志告警、统计分析等避免恶意代码攻击的技术措施，或者以白名单

方式构建安全可靠的服务器、终端等主机设备运行环境，有效阻断恶意代码感染路径或运行条件。

除了可信验证技术手段外，可以采用恶意代码防护工具。可以采取平台化管理，安全设备集中防护；也可以分区域，前置安全设备（NGFW、WAF、IDPS 等）。同时也要做好主机层面的恶意代码防护工作，安装必要的杀毒软件或杀毒模块。

3.4.6 可信验证

可信验证是基于可信根，构建可信链，一级度量一级，一级信任一级，把信任关系扩大到计算节点，从而保证计算节点可信的过程。可信根内部有密码算法引擎、可信裁决逻辑、可信存储寄存器等部件，可以向节点提供可信度量、可信存储、可信报告等可信功能，是节点信任链的起点。可信固件内嵌在 BIOS 之中，用来验证操作系统引导程序的可信性。可信基础软件由基本信任基、可信支撑机制、可信基准库和主动监控机制组成。

3.4.7 数据完整性

通用人工智能大模型训练和应用的过程中会产生用户鉴别数据、模型参数数据、数据集数据、用户问答数据、调优参数数据等。

数据完整性是指重要数据在传输和存储过程中，应采用校验技术或密码技术确保信息或数据不被未授权用户非法篡改或在被篡改后能够迅速识别，以保证其完整性。重要数据包括但不限于鉴别数据、重要业务数据、重要审计数据、重要配置数据、重要视频数据和重要

个人信息等。

其他服务的重要数据，例如备份、数据上送等可以使用相关安全产品或服务进行数据完整性安全防护。通常采用具备主机加固、主机防护功能的系统进行服务器、交换机等计算资源中数据存储过程中的数据完整性安全防护。

3.4.8 数据保密性

通用人工智能大模型的训练过程中需要采集大量的数据，存在采集用户敏感信息问题，而且在实际应用中也会给用户带来敏感信息泄露问题。

数据保密性安全要求实现的安全目标主要是指采用密码技术保证重要数据在传输和存储过程中的保密性。重要数据包括但不限于鉴别数据、重要业务数据和重要个人信息等。安全隔离环境生命周期结束时，能清除运行在环境中的 AGI 核心资产。

通常应由网络应用软件自带数据保密安全功能，或采用具备数据防泄漏功能（DLP）的系统进行服务器、交换机等计算资源中数据存储过程中的数据保密性安全防护。

通常网络系统、网络应用软件应具有数据加密安全功能，或采用具备认证加密、VPN 功能的安全设备（使用 SSH、VPN、TLS、HTTPs 等协议）进行网络应用服务器、交换机等计算资源之间数据传输过程中的数据加密性安全防护。

需要注意的是敏感信息泄露、数据库加密、弱口令管理、系统与数据库不使用同样的账户、备份的数据由专人管理等基本安全工作。

实施严格的输出过滤和上下文感知机制，以防止通用人工智能大模型泄露敏感信息。在通用人工智能大模型的训练过程中使用差分隐私技术或其他数据匿名化方法来降低过拟合或记忆的风险，并定期审计和审查通用人工智能大模型的回复，以确保敏感信息不会被无意中泄露。

3.4.9 数据备份恢复

通用人工智能大模型训练的算法、参数以及训练数据已经成为机构的核心资产。在数据备份恢复方面，网络系统生产数据和运行数据重要程度不言而喻，因此必须要做好业务系统、数据服务器数据的备份容灾。在进行数据备份时，必须根据实际需要配置备份策略。

对重要的业务应用系统应采用本地容灾方式保障业务不中断，对相关的本地生产数据进行定时、实时的数据备份，并提供异地数据备份功能，通过通信网络将重要数据定时、实时备份至备份场地，从而保证一旦发生故障，损坏或者丢失的数据可快速恢复，不会对用户造成大的损失。此外，重要数据处理系统初始设计时，要考虑系统硬件设施、功能组件配置的热冗余，保证重要数据处理系统的高度安全可用。

3.4.10 剩余信息保护

通用人工智能大模型应用过程中，需要对用户使用产生的敏感数据妥善处理。剩余信息保护主要针对主机和应用系统层面，不涉及网络和安全设备。鉴别信息应该是指用户身份鉴别的相关信息，敏感数据就是个人信息或企业重要信息。

网络安全等级保护的第三级系统中要求对这两类信息的存储空间再次分配前，应得到完整的释放。对于操作系统、内存和磁盘存储，可采取多次删除后覆盖的手段。对于应用系统，在设计时就应将这项功能集成在系统中。

3.4.11 个人信息保护

通用人工智能大模型的训练过程中需要采集大量的数据，存在采集用户隐私信息问题，而且在实际应用中也会给用户带来隐私信息泄露问题，例如用户的问答记录、上网行为习惯和消费记录等。

个人信息的保护应根据个人信息的使用场景不同采取不同的管控措施。对于个人信息在数据生命周期的不同阶段以及数据所处不同状态，数据面临的安全风险及需要实现的安全目标都会有所不同；同时，在不同状态下与数据相关的系统元素（主体、应用、存储、网络）也都有不同。所以需要采用不同安全机制和技术，对可能存在的安全风险进行控制。能够采用的安全机制包括认证、授权、控制、加密和审计；实施的对象包括人员、设备、应用和网络。

3.5 安全管理中心

3.5.1 系统管理

通用人工智能大模型依靠平台来进行训练和提供服务。在系统管理方面，安全管理中心的系统应当设置系统管理员账户，同时对系统管理员账户的使用进行身份鉴别。安全产品或组件具备一种或多种身份鉴别方式。具有审计管理权限的用户可以对系统管理员用户的各类

操作进行审计。

系统管理员可以对安全管理中心系统的资源、运行状况进行配置和管理，其功能包括但不限于用户身份配置管理、系统资源管理、系统加载和启动、系统运行异常处理、数据和设备的备份和恢复等。

3.5.2 审计管理

通用人工智能大模型依靠平台来进行训练和提供服务。在审计管理方面，安全管理中心的系统应当设置审计管理员账户，同时对审计管理员账户的使用进行身份鉴别。安全产品或组件具备一种或多种身份鉴别方式。具有审计管理权限的用户可以对审计管理员用户的各类操作进行审计。对审计管理员用户的审计需要由另一个审计管理员用户来进行。

审计管理员可以通过安全产品自身或额外的审计工具对审计记录进行分析，并根据分析结果进行处理，包括但不限于根据安全审计策略对审计记录进行存储、管理和查询等。

3.5.3 安全管理

通用人工智能大模型依靠平台来进行训练和提供服务。安全管理中心的系统应当设置安全管理员账户，同时对安全管理员账户的使用进行身份鉴别。安全产品或组件具备一种或多种身份鉴别方式。

具有审计管理权限的用户可以对安全管理员用户的各类操作进行审计。安全管理员可以对系统中的安全策略进行配置，包括安全参数的设置，主体、客体进行统一安全标记，对主体进行授权（例如网

络的访问控制策略等），配置可信验证策略等。

3.5.4 集中管控

通用人工智能大模型依靠平台来进行训练和提供服务。在集中管控方面，网络系统中划分出特定的网络区域作为安全管理中心的管理区域，对分布在网络中的安全设备或安全组件进行管控。

各安全产品应具有将各类运行状况和审计数据主动发送至安全管理平台的功能，包括但不限于网络链路、安全设备、网络设备和服务器等的运行状况。安全管理平台在汇总并进行集中分析后，向用户提供分析报告和相关告警、处置建议等。并依据《网络安全法》要求将审计记录留存六个月以上。

3.6 安全管理制度

3.6.1 安全策略

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的安全管理能力也是保障大模型平台安全运行的重要环节。网络安全工作的总体方针和安全策略的制定，应明确网络安全的总体目标、安全管理工作的范围、网络安全工作的原则和机构的安全管理框架。

总体方针和安全策略应是最高层的安全文件，以文件方式发布，可以是单一文件，也可以是一套文件。安全策略文件中应涵盖网络安全责任机构、职责和网络安全工作运行的模式。

网络安全责任机构应是网络系统中保护对象全生命周期中关键

的网络安全管理活动的责任部门。安全管理框架应包含安全组织机构、岗位职责划分、人员安全管理、环境安全管理、资产安全管理、系统建设安全管理、系统运行安全管理、安全事件处置和安全事件应急响应等方面内容。

3.6.2 管理制度

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的安全管理能力也是保障大模型平台安全运行的重要环节。安全管理制度应覆盖到网络安全等级保护对象的全生命周期，包括设计、建设、开发、测试、运维、升级和改造等各个环节。也应包含安全物理环境、安全管理机构、安全管理人员、安全建设和安全运维等方面内容。安全管理制度是一套文档体系，可划分若干个制度，若干个分册。

安全操作规程是各项安全管理活动的具体操作步骤和操作方法，可以是策略文档、操作手册、记录表单或规范方法，但应涵盖使用范围、目的、具体的管理活动、具体的规范方式等内容。

安全管理制度体系应当覆盖网络安全等级保护测评对象的全部方面，从顶层的总体方针策略，到具体执行的管理制度，以及日常的操作规程和各类记录表单文档等。

3.6.3 制定和发布

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的安全管理能力也是保障大模型平

台安全运行的重要环节。安全管理制度的制定应该规范化，应该在机构内部相关的部门或专门的人员的负责和指导下进行，并覆盖安全管理制度的起草、审定、论证和发布等主要环节，且制度文档的格式、编号和要求应该统一。

安全管理制度的发布应通过内部机构认可（审批通过、签名或盖章等），并通过机构允许的有效发布渠道进行发布，发布范围应仅覆盖到该制度所涉及的相关部门即可，如部门公告栏发布、内部办公系统发布、机构电子邮箱发布、企业办公终端 APP 发布和正式文件发布等方式。安全管理制度的发布应进行文件版本控制，版本控制信息中应涵盖文档编号、受控范围、文档版本号、发布日期和生效日期等基本信息。

3.6.4 评审和修订

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的安全管理能力也是保障大模型平台安全运行的重要环节。在评审和修订方面，安全管理制度的定期并没有限定期限，一般来说每年组织一次较好，也可以进行约定。论证和审定的范围主要根据安全管理制度实际落实情况，对该制度存在的不合理内容或缺失内容进行修订，例如某类操作规范是否合理、管理制度体系是否完备等方面。

3.7 安全管理机构

3.7.1 岗位设置

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的自身管理能力也是保障大模型平台安全运行的重要环节。在岗位设置方面，应设立指导和管理网络安全工作的委员会或领导小组，负责单位的网络安全管理的全局工作，最高负责人应由单位的主管领导担任、委任或授权，并应以文件形式明确其组成机构的工作职责。

网络安全领导委员会或领导小组的主要职责包括对机构安全管理制度体系的适用性和合理性进行审定、对机构内部关键的网络安全管理活动进行授权和审批，统筹机构的网络安全管理的全局性工作等。

3.7.2 人员配备

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的自身管理能力也是保障大模型平台安全运行的重要环节。在人员配备方面，应避免拥有关键操作权限的人员因操作失误或渎职问题导致安全管理活动中断，应配备一定数量的安全管理人员，如系统管理员、审计管理员和安全管理员等，这里的一定数量建议每种管理员配备至少 2 位或 2 位以上，其中的安全管理员与其他管理员岗位不能兼任，并且应提供管理人员任职名单。

3.7.3 授权和审批

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平

台依靠企业自身的研发能力，企业的自身管理能力也是保障大模型平台安全运行的重要环节。在授权和审批方面，为了保证系统发生安全问题时能够被溯源，机构对每个部门和岗位的职责应明确授权审批事项、审批部门和批准人等信息，并以文件形式确定授权和审批的制度，对审批程序和范围等内容进行规范。各项审批活动应进行记录，并与相应的职责文件能够对应。

审批程序中必须涵盖系统变更、重要操作、物理访问、系统接入等事项，如在变更管理制度、机房管理制度和网络管理制度中明确审批流程等，其中对重要的安全活动还要求建立逐级审批制度。审批活动应具备审批记录、审批程序、审批部门以及批准人应与审批制度文档相应。

需定期对相关审批事项进行审查活动，对发生变更的事项，例如审批部门、审批项目和审批人等信息进行更新，并留存更新记录以便查证。

3.7.4 沟通和合作

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的自身管理能力也是保障大模型平台安全运行的重要环节。在沟通和合作方面，对机构内部的涉及多个业务部门的等级保护对象的运行，需要各部门的配合与协调，建议采取定期例会和不定期召开会议的形式进行协商处理。应以文件形式确立各类管理人员、组织内部机构和网络安全管理部门的合作与沟通机制，并对每次协商活动进行记录，记录应该涵盖会议内容、会议时间、

参会人员和会议结果等信息。对外部单位应建立联系列表，包含外联单位名称、合作内容、联系人和联系方式等信息，并应对该表进行实时更新。

3.7.5 审核和检查

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的自身管理能力也是保障大模型平台安全运行的重要环节。在审核和检查方面，应建立定期常规安全检查机制，并以制度文件形式落实。

常规的安全检查范围包括系统日常运行、系统漏洞和修复以及系统备份与恢复等情况，一般是以月度、季度、半年或一年为期限开展，需留存安全检查记录。定期进行全面的安全检查，可由机构内部组织或通过第三方机构进行。检查范围包含现有的安全措施是否有效及落实情况、安全配置与安全策略是否一致、安全管理制度的执行情况等，并对全面检查活动进行文档记录。

3.8 安全管理人员

3.8.1 人员录用

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的人员管理能力也是保障大模型平台安全运行的重要环节。在人员录用方面，对安全管理人员的录用需要指定或授权专门的部门或人员来负责，并应在员工的聘用合同中明确安全职责范围和责任，保证安全管理人员录用过程的规范性。

对重要安全岗位的人员应进行身份、安全背景、专业资格或资质方面的审查，并对技术人员的专业技术水平进行考核，需留存审查和考核的相关记录或文档。

与安全管理岗位所录用的所有人员签署保密协议，明确保密的范围、责任、违约责任、协议的有效期和责任人等内容。对关键岗位还应签署安全岗位责任协议，明确岗位安全责任、协议的有效期和责任人等内容。

3.8.2 人员离岗

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的人员管理能力也是保障大模型平台安全运行的重要环节。在人员离岗方面，对以任何原因（离职、退休、合同到期、辞职和解雇等）离岗的人员都应在人员离岗前办理离岗手续，终止该人员拥有的全部权限，包括系统及物理环境，软硬件等，如技术文档、操作软件、操作账户口令、工作证、密钥及计算机设备等，并留存人员离岗记录文档，并记录离岗人员的归还资产清单。调离的保密承诺可单独签署也可在保密协议中有相关条款说明，需要有调离人员的签字文档。

3.8.3 安全意识教育和培训

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的人员管理能力也是保障大模型平台安全运行的重要环节。在安全意识教育和培训方面，对各类人员（普

通用户、运维用户、管理用户和机构负责人等）进行安全教育、岗位技能和安全技能培训，并对培训过程记录，明确培训周期、方式、内容和考核方式等内容。相关的安全责任和惩戒措施制度文档中应包括具体的安全责任条款和惩戒方式方法等内容。

对不同岗位的培训分别制定培训计划，培训内容应涵盖网络安全的基础知识、岗位操作规程等内容，并对留存培训记录，对参加培训的人员、培训的内容和培训的考核结果等内容进行描述。

3.8.4 外部人员访问管理

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台依靠企业自身的研发能力，企业的人员管理能力也是保障大模型平台安全运行的重要环节。在外部人员访问管理方面，对外部人员访问建立管理文档，明确外部人员能够访问的物理环境范围、进入的条件和访问控制措施等内容，并进行记录，包括外部人员访问重要区域的进入时间、离开时间、陪同人员和访问的区域等内容。

对外部人员接入受控网络的情况进行管理，以管理文档的形式确定外部人员接入受控网络的申请和审批流程，记录外部人员的访问权限、受控批准人、访问账户和时间周期等内容。

对获得授权的外部访问人员要签署保密协议，对操作的范围进行控制，如不得进行非授权操作、不得复制和泄露敏感信息等。

获得访问权限的外部人员离场后，按流程清除访问权限，并进行登记记录。

3.9 安全建设管理

3.9.1 定级和备案

通用人工智能大模型依靠平台来进行训练和提供服务，大模型平台应该依据国家法律法规要求进行定级备案活动。《网络安全等级保护定级指南》明确了定级的方法和理由，《等级保护对象安全等级保护定级报告》是全国各类等级保护对象定级报告的通用模板。

定级结果的合理性和准确性需要安全技术专家的论证评审。定级结果为第三级的，可组织本行业和网络安全行业专家进行评审。定级结果需由上级部门或本单位相关部门的批准。有主管部门的，备案材料需向主管部门和公安机关备案；没有主管部门的，备案材料需向相应公安机关备案。

3.9.2 安全方案设计

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台安全方案设计方面，确定安全保护等级后，安全规划设计需要依据安全保护等级的要求、风险分析的结果来补充、调整、确定基本安全保护措施，安全设计方案应当包含保护对象安全保障体系的总体安全策略、安全技术框架、安全管理策略、总体建设规划、详细设计方案等内容，并经过相关部门和网络安全行业专家的论证、审定和批准。

保护对象是整个单位等级保护对象的一部分，其安全方案应作为单位整体安全规划的一部分；保护对象的安全性可能与其它系统存在依赖、共享、支撑关系，需要系统性地规划与设计；对于涉及访问控

制、数据保护等设计内容，应采用适配保护等级的密码技术。

3.9.3 产品采购和使用

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台产品采购和使用方面，网络安全产品采购应遵循国家的管理要求，比如《网络安全法》《密码法》《网络安全审查办法》《网络关键设备和网络安全专用产品相关国家标准要求（征求意见稿）》等，从已获得《计算机等级保护对象安全专用产品销售许可证》的产品系列中选择。

涉及商用密码产品的，应当按照《密码法》中有关商用密码的要求，从网络关键设备和网络安全专用产品目录中选择。

3.9.4 自行软件开发

通用人工智能大模型依靠平台来进行训练和提供服务。在自行软件开发方面，为避免开发过程中对系统的影响，要保证开发环境与实际运行的环境（如办公网）物理隔离，测试数据和测试结果完全可控。并制定软件开发的管理制度，明确开发过程的控制措施和开发人员的行为准则，制定相应的编程语言编写规范，以便于代码的阅读、理解、维护、修改、跟踪调试、整合。

开发人员需要编制软件从需求分析、概要设计、详细设计、编码、测试、交付、验收、维护全流程的相关文档和使用指南，并对这些系统开发文档的存档、使用、交流等进行严格控制，便于指导相关技术人员对程序资源库的访问、维护、更新进行严格的控制。软件交付用

户前，应当通过工具测试和人工确认的方式进行软件的代码安全审计，以发现其中的漏洞或恶意代码。

3.9.5 外包软件开发

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台外包软件开发方面，与自行软件开发一样，对于外包软件，在交付前同样需要进行恶意代码检测，以保证软件的安全性。可要求外包方进行检测或机构内部自行检测。

软件开发完成之后，外包方应当按照协议约定提交软件需求分析、初步设计、详细设计、编码、测试、使用指南等系列规范的文档。后门和隐蔽信道的审查可通过专业的机构进行，若外包方无法提供该类审查报告，则需要提供书面材料保证软件源代码中不存在后门和隐蔽信道。

3.9.6 工程实施

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台工程实施方面，等级保护对象工程实施应当指定或授权专门的部门或人员负责工程实施过程的管理，以保证实施过程的正式有效性。

工程实施过程的控制需要事先制定实施方案，以明确工程时间限制、进度控制和质量控制等内容。对于外包实施项目，需要第三方工程监理的参与，以控制项目的实施过程，对工程进展、时间计划、控制措施、工程质量等进行把关。

3.9.7 测试验收

通用人工智能大模型依靠平台来进行训练和提供服务。大模型平台测试验收方面，测试验收包括两种情形，一种是外包单位项目实施完成后的测试验收，另一种是机构之间的内部开发部门移交给运维部门的验收。无论是哪种形式，均需要严格的测试验收方案，对验收的条件、标准、结果进行详细说明。

为保证系统建设工程按照既定方案和要求实施，并达到预期要求，工程交付后正式运行之前，应当指定或授权专业机构依据安全方案进行安全性测试。安全性测试不局限于抗 DDoS、敏感数据泄露、缓冲区溢出漏洞等典型脆弱点，特别对密码应用，应测试其保密性、完整性和可用性。

3.9.8 系统交付

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台系统交付方面，系统在工程实施并验收完成之后，需要根据协议有关要求，对照交付清单完整交付相应设备、软件、文档。交付单位或部门应当提交系统建设的过程文档以及系统运行维护的详细技术和管理文档，并对运维和操作人员进行必要的培训，方便后期的运营维护。

3.9.9 等级测评

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台等级测评方面，对等级保护对象进行等级测评是检验系统是否达

到相应等级保护要求的主要途径，也是发现系统安全隐患的重要途径。选择有资质的测评机构对系统进行定期的测评，有助于及时发现系统的问题并进行整改。对定为三级的保护对象，应当每年至少进行一次等级测评。

系统的重大变更，是指网络结构调整、大范围的设备更换、系统功能的较大变化。在这类重大变更之后，必须重新进行等级测评。同时，应重新评估系统的级别是否发生变化，若有变化，则需要按照新的安全保护等级要求进行测评。

3.9.10 服务供应商选择

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台服务供应商选择方面，对各类供应商的选择，均应符合国家的相关管理要求（如服务资质、产品资质（商用密码产品的检测认证）、运维资质、信息系统集成资质等）。

为了防范服务商引入的新的安全问题，在选择服务商时除了考虑其具有相应的服务资质外，还需要全面审查其相关背景、服务经历，同时以协议或合同的方式明确其职责以及后期的服务承诺。

对供应商的要求主要基于与其签订协议或合同中约定的网络安全相关条款和条件，检验其所提供服务与约定的符合程度，通过定期评审其工作服务报告，确保其有足够的能力，可按既定的工作计划履行其服务职责。

3.10 安全运维管理

3.10.1 环境管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台环境管理方面，良好的机房环境管理可以避免机房环境发生不可控变化引起的安全风险，极大提高机房的安全性和可靠性。

指定部门或人员负责机房安全管理工作，对机房出入进行管理，对基础设施进行定期维护。机房出入登记表记录来访人员、来访时间、离开时间、携带物品等信息。基础设施维护记录包含维护日期、维护人、维护设备、故障原因、维护结果等方面内容。机房管理制度覆盖物理访问、物品进出和环境安全等方面内容，明确来访人员的接待区域。

3.10.2 资产管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台资产管理方面，资产管理是为了对资产进行分类标识，依据标识的资产价值选取适当的管理手段和措施来保证资产安全。

资产清单需要包括资产类别（含设备设施、软件、文档等）、资产责任部门、重要程度和所处位置等内容。资产管理制度明确资产的标识方法（一般依据资产的重要程度等），不同类别的资产选取不同的管理措施。信息分类文档规定分类标识的原则和方法（如根据信息的重要程度、敏感程度或用途不同进行分类），信息资产管理办法规定不同类信息的使用、传输和存储等要求。

3.10.3 介质管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台介质管理方面，介质需要存放在符合其存放条件的环境中，并由专人管理。介质管理记录包括介质归档、查询、使用和定期盘点等情况。在介质进行物理传输时，对人员选择环节、打包环节、交付环节等制定规范化管理要求。

3.10.4 设备维护管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台设备维护管理方面，通过部门或人员岗位职责文档明确指定专人或专门部门对各类设备、线路进行定期维护。设备维护管理制度明确维护人员的责任、维修和服务的审批、维修过程的监督控制等方面，还包括设备带离的审批流程。

通过信息分类文档明确重要数据、敏感数据和授权软件的定义，在设备维护管理制度中明确含有重要数据的设备带离时的加密手段和措施；含有敏感数据和授权软件的设备报废或重用时，进行完全清除或完全覆盖的手段和措施。

设备维护管理制度包括明确维护人员的责任、维修和服务的审批、维护过程的监督控制等方面内容。任何设备均面临维护，设备维护管理制度需要涉及所有层次。

3.10.5 漏洞和风险管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型

平台漏洞和风险管理方面，识别安全漏洞和隐患的方式多种多样，通过网络漏洞挖掘系统，对网络系统所使用的专用控制设备进行未知漏洞挖掘；通过网络漏洞扫描系统，对网络系统所使用的专用系统进行漏洞扫描；通过渗透测试对网络系统所使用的专用系统进行渗透；通过网络安全相关部门的安全通报，识别网络系统所使用的专用系统的安全隐患。针对安全漏洞和隐患识别的结果，经过评估可能的影响后进行修补。

通过定期的安全测评活动，以安全测评报告的形式，记录发现的安全漏洞和隐患，并对这些安全漏洞和隐患进行评估和修补，最终增强系统的安全性。

3.10.6 网络和系统安全管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台网络和系统安全管理方面，网络和系统安全管理文档中管理员划分不同角色（至少划分为网络管理员和系统管理员），并定义各个角色的责任和权限；明确账户管理人员及相关审批流程；网络和系统安全管理文档覆盖网络和系统的安全策略、账户管理（用户责任、义务、风险、权限审批、权限分配、账户注销等）、配置文件的生成及备份、变更审批、授权访问、审计日志管理、日常操作、升级与打补丁、登录设备和系统的口令更新周期等方面。

3.10.7 恶意代码防范管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型

平台恶意代码防范管理方面，采取培训和告知等方式提升员工的防恶意代码意识，通过恶意代码防范管理制度明确对外来计算机或存储设备接入系统前进行恶意代码检查。

网络运维负责在数据中心统一部署防火墙、入侵检测系统等防范设备，实现接入业务的恶意代码防范，网络运维负责组织内部所有信息处理设施防病毒软件的安装、自动扫描设置和定期升级。负责对所使用的操作系统进行补丁升级。对电子邮件接收或下载软件开启病毒实时防护，进行检查。

3.10.8 配置管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台配置管理方面，网络系统中的配置信息包含网络设备接口、IP 地址、MAC 地址和掩码等信息，也包括网络拓扑结构、各个设备安装的软件组件、软件/固件组件的版本和补丁信息、各个设备或软件组件的配置参数等。

3.10.9 密码管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台密码管理方面，密码管理过程中需要遵循《中华人民共和国密码法》。密码相关产品需要获得有效的国家密码管理主管部门规定的检测报告或密码产品认证证书。

3.10.10 变更管理

通用人工智能大模型依靠平台来进行训练和提供服务。大模型平

台变更管理方面，变更方案包含变更类型、变更原因、变更过程、变更前评估等内容。在变更控制制度的申报和审批程序规定需要申报的变更类型、申报流程、审批部门、批准人等方面内容。变更实施方案中明确变更中止或失败后的恢复程序（恢复流程）、工作方法和职责，必要时对恢复过程进行演练。

3.10.11 备份与恢复管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台备份与恢复管理方面，应制定定期备份的重要业务信息、系统数据、软件系统的列表或清单，明确备份方式、频度、介质、保存期等内容，制定数据的备份策略和恢复策略、备份程序和恢复程序等。

3.10.12 安全事件处置

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台安全事件处置方面，《网络安全法》中明确要求应当按照事件发生后的危害程度、影响范围等因素对网络安全事件进行分级，并规定相应的应急处置措施。

网络安全事件分类分级制度，可以有效的提高安全事件响应速度，及时采取相应的处理和报告制度。安全事件的事后总结可以吸取经验教训，防止同类事件的再次发生。

3.10.13 应急预案管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台应急预案管理方面，《网络安全法》中明确要求建立健全网络安

全风险评估和应急工作机制，制定网络安全事件应急预案，并定期组织演练。

针对机房、系统、网络等各个方面制定应急预案，包括应急处理流程、系统恢复流程等内容，定期对相关人员进行应急预案培训和演练，定期评估并修订完善应急预案。

3.10.14 外包运维管理

通用人工智能大模型依靠平台来进行训练和提供服务。在大模型平台外包运维管理方面，应选取外包运维服务商时需要符合国家有关规定，外包运维服务协议明确约定外包运维的范围、工作内容、安全要求（包含可能涉及对敏感信息的访问、处理、存储要求，对基础设施中断服务的应急保障要求等内容）等。

4 通用人工智能赋能等级保护

通用人工智能大模型赋能是指利用开发出的超大规模深度学习模型，为其他自然语言处理任务和应用提供基础支持和指导，从而提高它们的性能表现。这些大模型通常采用预训练的方式，在海量数据上进行自监督学习，从而学习到深度和丰富的语言知识和语义信息。它们的优势在于，可以处理更加复杂多样的自然语言任务，并且可以应对一些新颖的任务，实现更加智能的语言理解和生成。

通用人工智能大模型赋能的应用范围非常广泛，例如机器翻译、语音识别、情感分析、文本生成等。通过对大模型的分析、解构和优化，可以为这些应用提供更好的训练方法、模型选择、超参数设置等技术支持，从而提高它们的表现。此外，在大模型赋能的过程中，还可以挖掘和发掘更加丰富的语言规律和知识，为后续的语言处理任务和应用提供更好的发展方向和思路。

本章节介绍几个通用人工智能大模型赋能的典型场景，包括等级保护知识赋能、等级测评赋能和等级保护培训赋能等。总的来说，随着大模型的不断发展和应用，各个领域都可以基于大模型进行建模、学习和智能决策，为人类提供更加快速、准确和智能的服务和应用，这是大模型赋能的一个重要意义。

4.1 网络安全等级保护知识赋能场景

4.1.1 等级保护政策法规知识赋能

近年来，网络安全得到了党和国家的高度重视，相关法律法规和标准不断出台，网络安全等级保护成为了重要的工作内容。通过遵循法律法规和政策标准，依法开展网络安全等级保护工作，可以有效地保障国家网络空间安全。

（1）国家法律法规政策知识辅助

国家先后颁布了《中华人民共和国网络安全法》《中华人民共和国密码法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》《关键信息基础设施安全保护条例》和《商用密码管理条例》等网络安全相关国家法律法规，重要信息系统和关键信息基础设施已经成为国家网络安全中的重要组成部分，维护国家网络空间主权和国家安全成为新时代的工作重点。

这些法律法规都和等级保护领域有着直接或间接的关系，再加上各部门依据法律法规出台的政策文件，如《关于落实网络安全保护重点措施 深入实施网络安全等级保护制度的指导意见》（公网安〔2022〕1058号）等。这些法律法规和政策文件的出台，意味着等级保护从业者需要承担更加繁重的责任和挑战，从业者需要不断更新自己的知识和技能，以适应新的法律法规和政策文件的要求，同时需要不断加强对等级保护领域的业务学习，不断改进和完善自己的技术和方法，以确保国家网络安全的稳定和可靠，为网络安全事业做出更大的贡献。

通过引入通用人工智能大模型，可以解决这个问题。一方面，通过分析和学习大量的等级保护领域相关的法律法规和政策文件知识数据集，通用人工智能可以为等级保护从业者提供准确、全面的法律法规和政策文件知识辅助，帮助他们更好地理解和应对法律法规和政策文件的要求和变化，从而提高等级保护领域法规和政策方面的服务质量。另一方面，通过分析和学习等级保护领域相关的法律法规和政策文件之间的逻辑关系，通用人工智能可以为等级保护从业者提供更加精准、高效的法律法规方面的专业支持和解决方案，提升等级保护领域法律法规和政策文件的专业水平和应对能力。

（2）标准知识辅助

在等级保护标准体系中，除了 GB 17859-1999《计算机信息系统安全保护等级划分准则》，经历了 GB/T 22239-2008 《信息安全技术 网络安全等级保护基本要求》和 GB/T 22239-2019《信息安全技术 网络安全等级保护基本要求》两代核心标准，包括与之相关的 GB/T 22240-2020 《信息安全技术 网络安全等级保护定级指南》、GB/T 28448-2019 《信息安全技术 网络安全等级保护测评要求》、GB/T 25058-2019 《信息安全技术 网络安全等级保护实施指南》和 GB/T 25070-2019 《信息安全技术 网络安全等级保护安全设计技术要求》等系列标准。

这些标准包含大量的技术术语和专业术语，需要等级保护从业者具备高度的专业知识和技能才能够理解和应用。此外，这些标准之间还可能存在着相互依存、相互制约的关系，需要行业耕耘者对标准之

间的逻辑关系进行深入的分析和理解，才能够正确地理解和执行标准。

同时，由于标准的更新和变化比较频繁，等级保护从业者还需要不断地学习和更新自己的知识和技能，以适应新的标准要求和变化。这需要他们具备高度的学习能力和适应能力，不断地更新自己的知识和技能，这是等级保护从业者需要面对的一个重要的挑战，他们需要花费极大的时间和精力来应对。

引入通用人工智能大模型能够有效地解决上述标准之间复杂的逻辑关系所带来的问题。具体来说，通用人工智能大模型可以从海量的标准文本中抽取相关的知识和信息，自动分析和理解标准之间的逻辑关系，形成参数集，为等级保护从业者提供精准和全面的标准解析和执行指导。

这样一来，等级保护从业者就不需要花费大量时间和精力去理解标准相互之间的关系，大模型会简化标准间复杂逻辑关系，提供逻辑连贯的解析和执行指导，从而提高标准执行的准确率和效率。此外，大模型还可以通过自动问答和用户交互等方式，为等级保护从业者提供实时和智能的标准知识咨询服务。当标准发生变化或新标准出台时，大模型还可以自动分析和提取最新的标准要求和变化，以便等级保护从业者及时更新自己的知识和技能，保持与标准同步。

除此之外，大模型还可以进行多语言处理和跨文化交流，在全球范围内帮助等级保护领域耕耘者更好地理解和应用标准，提升他们的国际化标准知识理解和执行能力。

4.1.2 等保定级备案知识赋能

国家要求信息系统的信息系统运营使用单位开展网络安全等级保护定级备案工作。信息系统运营使用单位确定信息系统的安全保护等级后，组织专家评审后到公安机关办理备案手续，由公安机关网安部门向备案单位颁发信息系统安全保护等级备案证明文件。

（1）系统定级知识辅助

2007 年 7 月，公安部、国家保密局、国家密码管理局和国务院信息化工作办公室联合发布《关于开展全国重要信息系统安全等级保护定级工作的通知》（公通字〔2007〕861 号）。该通知为全国重要信息系统等级保护开展定级工作给出要求顶层设计。随后，2008 年 6 月国家标准发布《信息系统安全保护等级定级指南》（GB/T22240—2008），为全国定级工作给出方法指导。该标准与 2020 年修订为《信息安全技术 网络安全等级保护定级指南》（GB/T 22240-2020）。

信息系统定级工作涉及到信息系统的数量、分布、业务类型、应用或服务范围、系统结构等基本情况，确定定级对象和安全保护等级并形成定级报告，对于一定安全保护等级的信息系统，还需要进行专家评审环节。信息系统定级是一个复杂的过程，需要一定的专家知识和专业知识，尤其是本身就复杂的信息系统，如何确定定级对象和安全保护等级，一直是等级保护系统定级中的重点和难点。

等级保护从业者需要理解系统定级工作的复杂性和专业性，但由于个人专业知识的限制，很多从业者无法完全理解和掌握系统定级工

作的核心要点和关键流程。

这种情况下，引入通用人工智能大模型的整合能力，训练时引入大量的信息系统定级备案数据，通过对历史数据和规律的分析，并结合预测算法，大模型可以给出针对一个对新信息系统的定级指引，辅助信息系统运营者制定出科学合理的定级方案，从而提高信息系统定级的效率和水平，从业者也可以通过这些定级方案，不断的学习和实践，提高自己的能力水平，从而更好地理解和应用信息系统定级工作，为等级保护工作的顺利开展做出更大的贡献。

甚至，相关机构还可以利用大模型赋予的能力，加强对等级保护从业者的培训和指导，提高等级保护从业者的整体素质和能力水平，以推动等级保护定级工作的发展和进步。

（2）系统备案知识辅助

根据《信息安全等级保护管理办法》，信息系统安全保护等级为第二级以上的信息系统运营使用单位或主管部门到公安部网站下载《信息系统安全等级保护备案表》和辅助备案工具，持填写的备案表和利用辅助备案工具生成的备案电子数据，到公安机关办理备案手续，提交有关备案材料及电子数据文件。

隶属于中央的在京单位，其跨省或者全国统一联网运行并由主管部门统一定级的信息系统，由主管部门向公安部办理备案手续。跨省或者全国统一联网运行的信息系统在各地运行、应用的分支系统，向当地设区的市级以上公安机关备案。

信息系统的备案也是一个复杂的过程，首先需要准备《信息安全

等级保护备案表》，将需要备案的信息系统的信息填入，然后按照《信息系统等级保护备案实施细则》的要求进行备案。目前国家并没有提供专门的信息系统备案咨询服务，信息系统运营者需要额外聘请等级保护领域专家或者专业机构来进行处理。

引入通用人工智能大模型后，大模型可以加速整个备案的流程。大模型通过预训练过程中学习的等级保护信息系统备案知识参数，自动抽取和整理出系统的关键信息和规则，提高备案数据的质量和准确性。

甚至，大模型可以从海量数据中自动抽取出与系统等级保护有关的项和细节，结合系统运营者的系统和资产信息等，自动填充到相应的表格中。通过自动填表，大模型能够避免手动操作的错误和漏填，同时也能提高信息系统备案表填写的效率和精确度。另外，大模型还拥有语义理解和模式识别能力，通过对备案表数据的智能分析和有关规则的学习，它能够探索和发现备案数据规律，并推出适用于不同系统的信息系统备案方案，帮助系统运营者更加科学、规范地开展等级保护备案工作。

4.1.3 等保测评知识赋能

等级测评工作是指测评机构依据国家网络安全等级保护制度规定，按照有关管理规范和技术标准，对未涉及国家秘密的信息系统安全等级保护状况进行检测并评估其是否达到相应等级要求的活动，是网络安全等级保护制度的重要环节。等级测评过程包括测评准备活动、方案编制活动、现场测评活动和报告编制活动四个基本测评活动。

在现场测评中，需要测评人员具备复杂的判断能力和等级保护专业技能，以便能够快速高效的获取信息系统全面的信息，形成一套可以落地实施的测评方案。这需要测评人员在现场进行复杂的工具测试和证据获取，涉及到众多的系统资产、证据表单，以及测试结果，需要细致入微地分析和整理这些数据，非常耗费时间和精力。

引入通用人工智能大模型可以解决这些问题，提升测评人员的工作效率和水平。通用人工智能大模型可以利用其强大的自然语言处理、文本分析、图像识别，以及数据分析和挖掘技术，在现场快速高效地获取信息系统全面的信息，并形成一套可行的测评方案。

同时，大模型可以快速地识别出工具测试的接入点，并获取工具测试结果，以便支撑测评报告的编制工作。此外，大模型还可以通过关联分析和推理，对整个信息系统进行深入分析，准确识别潜在的安全风险和漏洞，帮助测评人员形成准确的分析和判断，并提出对应整改建议和方案。

因此，引入通用人工智能大模型可以解决现场测评中的问题，提升测评人员的工作效率和水平。大模型可以快速获取系统资产、证据表单和测试结果，支撑测评报告的编制；同时，大模型可以通过数据分析和挖掘技术，准确分析和判断信息系统的安全保护状况，帮助测评人员形成整改建议和方案，提升测评人员的专业能力。

4.1.4 等保整改建议知识赋能

等级测评活动过程中，采用风险分析的方法分析等级测评结果中存在的安全问题，并给出针对性的安全整改建议。

在实际的现场测评工作中，测评人员需要具备复杂的判断能力和等级保护专业技能，以便能够快速高效地获取信息系统全面的信息，并形成一套可行的整改方案。这涉及到众多的测试结果，如漏洞扫描工具、渗透测试工具、源代码审计工具和协议分析工具所产生的结果文件等，以及众多的系统资产、证据表单和测试结果，还有复杂的用户系统网络结构和这些证据之间各种逻辑关系等。

测评人员需要快速高效的获取各项测评证据，细致入微地分析和整理这些数据，获取真实可靠的信息系统安全问题，并形成建设整改意见，这都非常耗费测评人员的时间和精力，考较测评人员的专业技能。

通过引入通用人工智能大模型，利用其强大的自然语言处理、文本分析、图像识别、数据分析和挖掘技术，辅助测评人员的现场测评工作。甚至大模型还可以通过关联分析和推理，对整个信息系统进行深入分析，准确识别潜在的安全风险和漏洞，帮助测评人员形成准确的分析和判断，并提出对应整改建议和方案等，提升测评人员的工作效率和水平。

4.2 等级测评赋能场景

4.2.1 系统调研阶段赋能

系统调研阶段对于网络安全等级测评至关重要。在该阶段中，主要是通过收集和了解定级对象的资产信息、文档信息和网络拓扑等相关资料的方式，为下一步的方案编制阶段做准备工作。

然而，在实际工作中，如何能快速、便捷地了解定级系统的全部资产和掌握网络结构是工作中的难点。传统的系统调研活动中，需要填写大量的调查表格，包括物理与逻辑边界、硬件资源、软件资源和信息资源等，这些都是依靠人力来完成的，这不仅需要大量的时间和精力，而且容易因人为因素而出现误差。对于一些比较复杂的信息系统，涉及到的资产信息和网络结构更加复杂，需要更多的人力和时间投入。

因此，引入通用人工智能大模型对测评准备活动进行辅助，可以大大提高工作效率和准确性。例如，利用大模型进行资产信息的快速梳理、自动分析和形成网络拓扑结构等，都可以大大节省人力成本和减少工作的难度。通过大模型对测评准备活动的赋能，可以更加快速、便捷和全方位的了解定级系统，为后续的工作提供准确可靠的数据支撑。

4.2.2 方案编制阶段赋能

方案编制阶段的目标是依据系统调研阶段收集的信息系统各项资料信息，编制为现场测评阶段提供最基本的文档和指导方案。方案编制包括测评对象确定、测评指标确定、测评内容确定、工具测试方法确定、测评作业指导书开发及测评方案编制六项主要任务。

在测评方案制定的实际工作中，确定测评对象和工具测试方法一直是整个阶段中的重点和难点。由于信息系统中的资产众多、网络结构复杂，而测评对象的选择又需要基于一定的原则，如重要性、安全性、共享性和全面性等等，需要测评人员具备极强的等级保护领域的

专业知识素养。

此外，工具测试可能涉及到漏洞扫描、渗透测试、源代码审计和协议分析等工具集，需要直接接入到信息系统的生产环境中，因此需要慎重而又有效地制定工具测试方案，选取合适的接入点，这也考验着等级测评师的智慧和技能。

所以，引入通用人工智能大模型，可以更好地解决这些问题，辅助等级测评师进行测评对象的确定和工具测试方法的制定。通过大模型的各种赋能，例如文生图功能，可以更加快速、准确地确定测评对象，同时还可以有效地制定工具测试方案，提高测评方案的编制效率和准确性。

4.2.3 现场测评阶段赋能

现场测评阶段主要是依据测评方案，进入信息系统的实际生产环境，获取编制测评报告所需要的足够的证据和资料。现场测评包括现场测评准备、现场测评和结果记录、结果确认和资料归还三项主要任务。

现场测评是一项复杂而又艰巨的任务，测评人员往往要面临各种各样的突发情况。例如，无法确定测评对象的位置、网络通信不畅或遭遇到新型号的硬件设备导致作业指导书失效等问题，这些突发状况非常考验测评人员的技术反应能力。特别是在操作工具测试生产环境的信息系统时，稍有不慎就会对系统的正常运行造成一定的影响。因此，在现场测评中，测评人员需要具备专业知识和经验，以快速、准确地应对各种突发情况，确保测评的顺利进行。

引入通用人工智能大模型可以发挥出色的作用，帮助测评人员快速、准确地应对各种突发情况。例如，利用大模型的能力可以快速分析资产结构和网络结构，科学合理的对确定测评对象和工具测试的接入点及测试路径，依靠大模型训练集成的丰富知识体系，可以对测评人员的现场测评进行作业指导，快捷有效的获取测评证据，准确地分析信息系统潜在的安全漏洞，为现场测定活动保护提供更加可靠的技术保障。

4.2.4 报告编制阶段赋能

在现场测评结束后，测评人员对现场测评获得的测评结果（证据资料）进行汇总分析，形成等级测评结论，并编制测评报告。测评人员初步判定单向测评结果后，还需要进行单元测评结果判定、整体测评、系统风险评估，形成安全问题和等级测评结论，并提供整改建议。

在报告编制的实际工作中，大量的表单数据是不可避免的，这些数据需要进行全面、准确的分析，以便形成有效的等级测评结论。传统的测评工具虽然可以自动化地收集和进行一些基本的分析，但其中一些有价值的分析仍需高级测评专家进行人工分析，并对整体测评环节的安全控制点间、层面间进行判断，这对测评人员的技能水平和经验要求很高。

同时，在现场测评中，工具测试的原始结果可能存在大量关联性，需要进行深入关联分析，然而传统的自动化分析仅基于事先制定好的规则，难以及时发现出现的时效性和突发性安全问题。

引入通用人工智能大模型后，过对大量数据和历史信息的学习和

分析，大模型可以为测评人员提供全方位、准确且高效的安全分析，有效地解决传统自动化分析无法解决的复杂问题。

此外，大模型还能够利用自然语言处理技术进行文本挖掘和分析，从而为测评人员提供深入和精确的报告结论。综上所述，通用人工智能大模型为报告编制提供了一种全新的思路，有助于提高测评人员的工作效率和水平，为信息安全保护提供更加可靠的保障。

4.2.5 报告质量审核阶段赋能

在等级测评报告的质量审核阶段，通用人工智能大模型可以发挥重要作用。首先，大模型可以根据报告全文的资产信息、测试结果，结合测评标准和要求，利用其强大的自然语言处理和文本分析技术，对测评报告的各个部分进行深入分析和比对，发现潜在的报告逻辑问题和不符合规范的内容，并提出相应的建议和改进方案，从而提高了审核效率和准确性。

其次，大模型还可以根据学习的测评报告的历史数据和经验，对等级测评报告进行自动批判性评估。大模型可以利用其强大的数据分析和挖掘技术，对历史测评报告的数据进行深入分析和比对，发现测评报告的疏漏和不足之处，快速提出改进方案和建议。这不仅提高了测评报告的质量，同时也提高了测评机构和测评人员的整体水平和能力。

此外，大模型还可以通过引入多模态数据分析，提高测评报告的可解释性和可靠性。大模型可以同时处理多种数据类型，例如文本、图片、视频和音频等，通过各种数据之间的关系和交互，对测评报告

的各个方面进行深入判断和分析，从而提高测评报告的可解释性和可靠性。尤其是在复杂的系统环境下，大模型可以通过分析多种数据类型，对测评报告进行全面的分析和判断，从而提高了测评报告整体的质量水准。

因此，通用人工智能大模型在等级测评报告的质量审核阶段具有重要作用。大模型可以自动分析、自动评估和自动改进测评报告的质量，从而提高了测评报告的准确性、解释性和可靠性，同时也提高了整体测评机构和测评人员的整体水平和能力。

4.3 等级保护培训赋能场景

网络安全等级保护培训旨在帮助等级保护从业者学习和了解相关法律法规和政策标准，从网络安全等级保护的发展历程、正式实施、典型案例分析、工作流程、建设整改、测评实施、监督检查等方面进行全面深入的培训。

通过培训，等级保护从业者将掌握网络安全等级保护领域的基本概念和原理，了解国家相关政策法规和标准规范，掌握等级保护的工作流程和实践方法，提高个人网络安全等级保护的能力和水平。同时，培训还将通过典型案例分析和实践操作，帮助从业者更加深入地理解和掌握网络安全等级保护的实际应用，提高工作的实效性和实用性，提高自身在网络安全领域的竞争力和职业发展前景。

引入通用人工智能大模型后，在网络安全等级保护知识培训方面可以提供以下能力：

（1）知识检索和总结：通用人工智能大模型可以帮助从业者快速检索和总结相关领域的知识，减少搜寻资料的时间，提高学习效率。

（2）自适应学习：通用人工智能大模型可以针对从业者不同的学习进度、习惯、需求和水平，进行个性化的学习方案和推荐资源，提高学习效果。

（3）智能辅导：通用人工智能大模型可以像人类教师一样，为从业者提供问题解答、课程辅导和知识启发，帮助他们理解课程内容和解决学习难题。

（4）等级保护专家支持：通用人工智能大模型拥有网络安全等级保护领域深度和广度的知识，并可以快速学习和掌握新的等级保护领域知识，为从业者提供领域专家辅导。

（5）多模态辅助：通用人工智能大模型还可以将文字、音频、视频等多媒体资料进行融合，使从业者更加直观地了解知识点和应用场景，提高学习表现和兴趣。

总之，通用人工智能大模型可以为网络安全等级保护知识培训和职业发展提供更加智能和个性化的服务，为机构和从业者创造更大的价值。

5 总结

通用人工智能是一种能够模拟人类智能的技术，可以自主地学习和适应新的环境和任务。大模型具备多个场景通用、泛化和规模化复制等诸多优势，被视为实现通用人工智能的重要研究方向。然而，随着通用人工智能技术的快速发展，也带来了一系列的安全风险和挑战。

本白皮书从技术角度出发来介绍了通用人工智能在过去几年中的发展趋势，以及它在未来的应用前景和挑战；分析了通用人工智能的安全风险，包括传统数据安全风险、数据隐私泄露、模型可靠性、模型滥用误用等方面；着重分析了通用人工智能在网络安全等级保护合规落地需求；列举了网络安全等级保护知识赋能场景、等级测评赋能场景和等级保护培训赋能场景，这些场景不仅可以提高等级保护的效率和服务质量，还可以为社会带来更多的便利和价值。

总之，本白皮书旨在为读者提供对通用人工智能在网络安全等级保护领域的全面了解，帮助读者更好地应对通用人工智能大模型带来的安全风险和挑战，同时也为网络安全等级保护在通用人工智能领域的合规落地提供了指导和建议。

参考文献

[1] ISO/IEC 22989 Information technology Artificial intelligence Artificial intelligence concepts and terminology

[2] ISO/IEC 23053 Framework for Artificial Intelligence(AI)Systems Using Machine Learning(ML)

[3] ETSI GR SAI 002 V1.1.1 Securing Artificial Intelligence (SAI); Data Supply Chain Security

[4] ETSI GR SAI 005 V1.1.1 Securing Artificial Intelligence (SAI); Mitigation Strategy Report

[5] OlaGPT： Empowering LLMs with Human-like Problem-Solving Abilities

[6] Scaling Language Models: Methods, Analysis & Insights from Training Gopher[J]. arXiv e-prints, 2021. DOI:10.48550/arXiv.2112.11446.

[7] Neelakantan A , Xu T , Puri R ,et al.Text and Code Embeddings by Contrastive Pre-Training[J]. 2022.DOI:10.48550/arXiv.2201.10005.

[8] Model evaluation for extreme risks.DeepMind.

AI人工智能产业链联盟

#每日为你摘取最重要的商业新闻#

更新 · 更快 · 更精彩

- Zero
  AI人工智能产业链联盟创始人
  河北清华发展研究院智能机器人中心运营经理
- base:北京
- 主要业务:AI商业化答疑及课程应用场景探索，各类AI产品学习手册，答疑及课程

欢迎扫码交流

提供：学习手册/工具/资源链接/商业化案例/行业报告/行业最新资讯及动态