ICS 35.030
CCS L 80

中华人民共和国国家标准

GB/T 45958—2025

网络安全技术
人工智能计算平台安全框架

Cybersecurity technology—Security framework for artificial intelligence computing platform

2025-08-01 发布　　2026-02-01 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前言 …… Ⅲ
1　范围 …… 1
2　规范性引用文件 …… 1
3　术语和定义 …… 1
4　安全框架 …… 2
5　安全功能 …… 3
5.1　资源层安全功能 …… 3
5.2　调度层安全功能 …… 4
5.3　应用支撑层安全功能 …… 4
6　安全管理 …… 5
6.1　身份管理 …… 5
6.2　密码管理 …… 5
6.3　日志管理 …… 5
6.4　安全监测 …… 5
6.5　安全审计 …… 6
6.6　风险管理 …… 6
6.7　个人信息保护 …… 6
7　角色安全职责 …… 6
7.1　概述 …… 6
7.2　平台提供方 …… 6
7.3　数据提供方 …… 6
7.4　模型提供方 …… 6
7.5　应用提供方 …… 7
附录 A（资料性）　平台组成与安全风险示例 …… 8
A.1　平台组成 …… 8
A.2　安全风险示例 …… 8
附录 B（资料性）　角色描述与典型活动 …… 11
参考文献 …… 12

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由全国网络安全标准化技术委员会(SAC/TC 260)提出并归口。

本文件起草单位：华为技术有限公司、中国电子技术标准化研究院、北京中关村实验室、中国电信股份有限公司、公安部第三研究所、国家信息技术安全研究中心、中国移动通信集团有限公司、四川大学、中国科学院软件研究所、中电长城网际系统应用有限公司、中国科学技术大学、西安电子科技大学、北京火山引擎科技有限公司、北京数字认证股份有限公司、蚂蚁科技集团股份有限公司、中国科学院信息工程研究所、南湖实验室、启明星辰信息技术集团股份有限公司、贝壳找房(北京)科技有限公司、中国电力科学研究院有限公司、浙江大华技术股份有限公司、科大讯飞股份有限公司、国网新疆电力有限公司电力科学研究院、西安交通大学、上海商汤智能科技有限公司、国家工业信息安全发展研究中心、华控清交信息科技(北京)有限公司、北京远鉴信息技术有限公司、北京天融信网络安全技术有限公司、云从科技集团股份有限公司、上海市信息安全测评认证中心、北京快手科技有限公司、国家信息中心、北京数安行科技有限公司、郑州信大捷安信息技术股份有限公司、北京交通大学、华中科技大学、美的集团(上海)有限公司、公安部第一研究所、北京眼神科技有限公司、深圳市洞见智慧科技有限公司、北京银联金卡科技有限公司、国网区块链科技(北京)有限公司、杭州安恒信息技术股份有限公司、奇安信科技集团股份有限公司、中电科网络安全科技股份有限公司、上海燧原科技股份有限公司、中国联合网络通信有限公司广东省分公司、北京山石网科信息技术有限公司、深圳大学、江苏保旺达软件技术有限公司、北京数牍科技有限公司、中国软件评测中心(工业和信息化部软件与集成电路促进中心)、浪潮(北京)电子信息产业有限公司、武汉东湖大数据科技股份有限公司。

本文件主要起草人：葛小宇、严敏瑞、许晓耕、刘勇、张宇光、谷红勋、张侃、徐浩、陆臻、郭晓雷、江为强、张峰、张瑞、陈兴蜀、张立武、闵京华、左晓栋、李兴华、陈妍、袁静、郭建领、张永强、林冠辰、王蕊、荆丽桦、陈治宇、张磊、唐文、王子恬、蒋发群、李陟、胡月、马梦娜、刘栋、郝沁汾、陈炯、王晓辉、李道兴、郑碧煌、张宇、文良君、王宏、谢铮涵、卞超轶、吴晓杰、陈剑波、梅瑞、叶波、杨慧婷、张严、蔺琛皓、邱云翔、俞锦浩、陆一凡、杨韬、王超、王雨晨、吴庚、朱倩倩、王运帷、靳晨、沈超、王雪强、佟剑辉、雷晓锋、杨显森、张俊彦、任永攀、落红卫、谷晨、章恒、王和俊、李昂、刘玉红、刘为华、王伟、刘敬楷、李瑞轩、蔡亚森、王号召、刘军、杨春林、姚明、何浩、彭宇翔、胡师阳、杨波、王栋、杨珂、刘博、李剑锋、安锦程、曹占涛、梅敬青、王思善、曾楚轩、程伟、吴疆、何伊圣、刘伟丽、谢江、葛颂、金银玉、刘云龙、高国鲁、杜乐。

网络安全技术
人工智能计算平台安全框架

1 范围

本文件确立了人工智能计算平台的安全框架，规定了安全功能、安全管理和角色安全职责。

本文件适用于人工智能计算平台的设计、建设、应用和运维。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本(包括所有的修改单)适用于本文件。

GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求

GB/T 31168—2023 信息安全技术 云计算服务安全能力要求

GB/T 35273—2020 信息安全技术 个人信息安全规范

GB/T 41479—2022 信息安全技术 网络数据处理安全要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能计算平台 artificial intelligence computing platform

为人工智能应用开发与运行提供高效、可扩展的资源环境和支撑组件的软硬件系统。

注：人工智能计算平台通常应用于数据中心及边缘计算场景。

3.2

人工智能服务器 artificial intelligence server

为人工智能应用提供高效能计算处理能力的服务器。

注：人工智能服务器通常集成人工智能加速处理器(3.3)、人工智能加速卡(3.4)或人工智能加速模组(3.5)，以符合人工智能应用的加速计算需求。

[来源：GB/T 42018—2022，3.5，有修改]

3.3

人工智能加速处理器 artificial intelligence accelerating processor

具备适配人工智能算法的运算微架构，能完成人工智能应用加速运算处理的集成电路。

注：典型的人工智能加速处理器有图形处理器、神经网络处理器和张量处理器。

[来源：GB/T 42018—2022，3.8，有修改]

3.4

人工智能加速卡 artificial intelligence card

专为人工智能计算设计、符合人工智能服务器(3.2)硬件接口，集成了人工智能加速处理器(3.3)或其中的运算微架构的扩展加速部件。

[来源:GB/T 42018—2022,3.6,有修改]

3.5

人工智能加速模组 artificial intelligence accelerating module

专为特定领域人工智能计算设计的,集成了人工智能加速处理器(3.3)或其中的运算微架构的扩展加速部件。

注1:人工智能加速模组可用于智能摄像机、无人机和人工智能服务器等不同场景。

注2:本文件关注的是用于人工智能服务器中的人工智能加速模组。

[来源:GB/T 42018—2022,3.7,有修改]

3.6

机器学习模型 machine learning model

一种基于输入数据或信息生成推理或预测的计算结构。

注:本文件在不引起误解的语境中,将机器学习模型简称模型。

[来源:GB/T 41867—2022,3.2.11,有修改]

3.7

训练 training

利用数据,基于机器学习算法,建立或改进机器学习模型参数的过程。

[来源:GB/T 42018—2022,3.11]

3.8

推理 inference

根据已知信息进行分析、分类或诊断,做出假设,解决问题或者给出推断的过程。

注:人工智能领域的推理包括逻辑推理、机器学习推理等,本文件关注的是机器学习推理。

[来源:GB/T 42018—2022,3.12,有修改]

4 安全框架

人工智能计算平台的组成与安全风险示例见附录A,本文件基于平台组成和安全风险确立了人工智能计算平台的安全框架,包括安全功能、安全管理和角色安全职责,见图1。安全功能分为资源层安全功能、调度层安全功能和应用支撑层安全功能,资源层安全功能包括计算资源安全、存储资源安全、网络资源安全和虚拟资源安全,调度层安全功能包括资源调度安全和任务调度安全,应用支撑层安全功能包括数据处理安全、模型训练安全和模型推理安全。安全功能侧重通过技术手段降低安全风险,实际应用中,一个人工智能计算平台不必具备所有的安全功能,而根据参与方对安全风险的接受程度,选择部署相应的安全功能。安全管理侧重通过管理手段降低安全风险,包括身份管理、密码管理、日志管理、安全监测、安全审计、风险管理和个人信息保护七个方面。角色安全职责描述了平台提供方、数据提供方、模型提供方和应用提供方如何分工协作共同保护模型与数据。

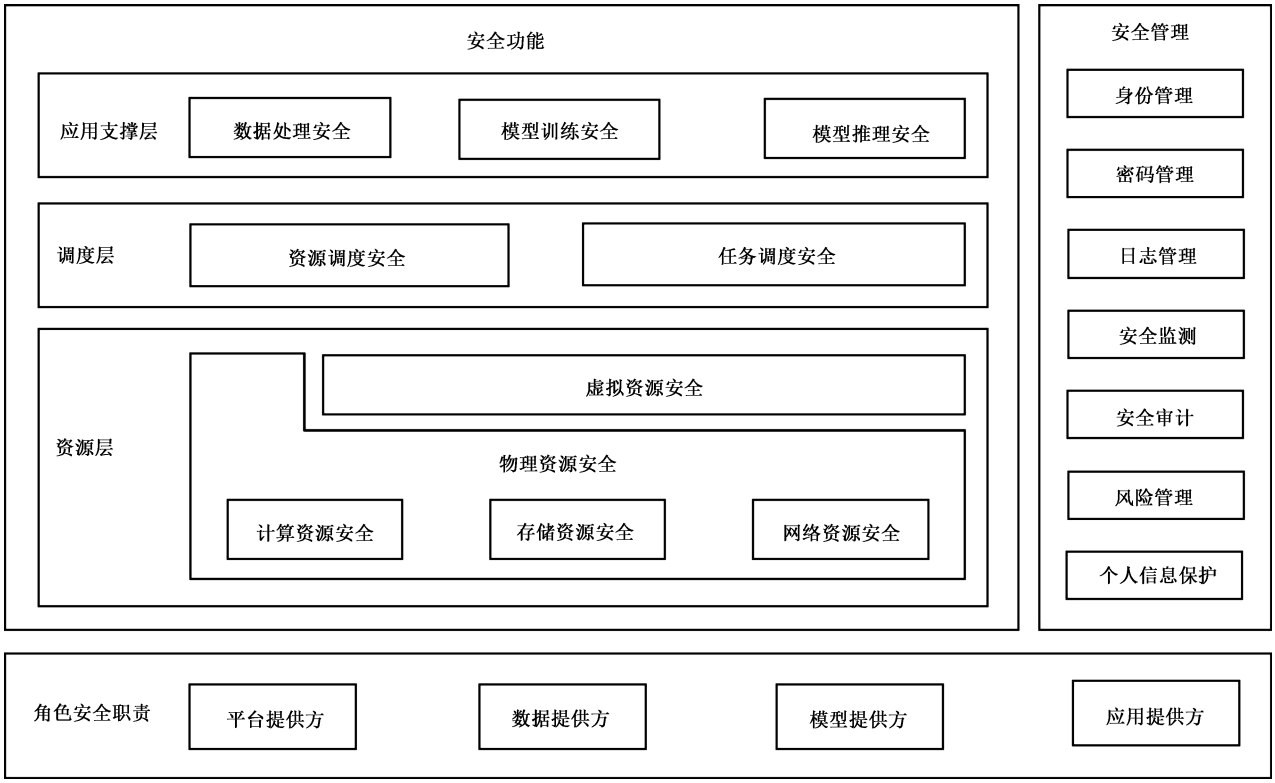

图 1　人工智能计算平台安全框架

5　安全功能

5.1　资源层安全功能

5.1.1　计算资源安全

人工智能计算平台的计算资源安全应符合 GB/T 22239—2019 中 8.1.4.1、8.1.4.2 和 8.1.4.6 规定的要求，还包括下列内容：

a）　对人工智能加速计算资源固件等进行完整性和真实性校验，校验不通过则停止启动或产生报警；

b）　验证人工智能加速计算资源固件的版本正确性，防止被回退到有安全隐患的版本；

c）　为相关参与方提供验证人工智能加速计算资源完整性的参数和接口等，支撑其判定人工智能加速计算资源能否以符合预期的方式工作；

d）　为管理员或相关进程提供管理与配置人工智能加速计算资源的接口，同时保护其中的模型与数据不被非授权访问；

e）　监测并处置人工智能加速计算资源故障，处置措施包括但不限于隔离或替换故障的人工智能加速处理器、人工智能服务器等；

f）　对用于数据处理、模型训练和模型推理等不同目的的人工智能加速计算资源进行安全隔离与管控；

g）　对于高安全需求场景，通过构建中央处理器与人工智能加速处理器安全互联的可信执行环境等，保护模型与数据加载进内存运算以及在多个处理器间传输过程中的保密性和完整性。

5.1.2 存储资源安全

人工智能计算平台的存储资源安全应符合 GB/T 22239—2019 中 8.1.4.8 b)、8.1.4.9 和 8.1.4.10 规定的要求，还包括下列内容：

a) 冗余存储模型训练过程中记录的检查点内容，包括模型参数和优化器状态信息等，防止因设备故障等造成检查点内容丢失；
b) 通过分布式存储架构等提升存储资源的可扩展性和可用性，防止因存储空间耗尽，导致任务中断，模型与数据丢失等；
c) 通过多重备份、安全隔离等措施，防止模型与数据被恶意加密、勒索等；
d) 采用密码技术等保护模型与数据存储过程中的保密性和完整性。

5.1.3 网络资源安全

人工智能计算平台的网络资源安全应符合 GB/T 22239—2019 中 8.1.2 和 8.1.4.8 a)规定的要求，还包括下列内容：

a) 采用密码技术等保护模型与数据在不同安全域间传输过程中的保密性和完整性；
b) 模型训练与推理过程中，涉及多个人工智能服务器间高速同步数据时，通过轻量化的访问控制校验或密码技术等措施保护通信数据的保密性和完整性；
c) 对用于数据处理、模型训练和模型推理等不同目的的网络资源进行分区隔离。

5.1.4 虚拟资源安全

人工智能计算平台的虚拟资源安全应符合 GB/T 31168—2023 中 7.11、7.12 和 7.13 规定的要求，还包括下列内容：

a) 对虚拟机、容器等进行安全加固和最小化安装，减少风险暴露面；
b) 对不同的虚拟化人工智能加速计算资源进行安全隔离，且隔离机制无法被突破或绕过；
c) 多任务复用人工智能加速计算资源时，清除相应存储空间中残留的模型与数据。

5.2 调度层安全功能

5.2.1 资源调度安全

人工智能计算平台的资源调度安全包括但不限于下列内容：

a) 对物理资源和虚拟资源按策略进行统一管理调度与分配；
b) 对不同用户和不同任务占用的资源进行安全隔离，防止不同任务间的相互干扰；
c) 监测不同任务的资源占用情况，对异常的资源占用情况，产生报警。

5.2.2 任务调度安全

人工智能计算平台的任务调度安全包括但不限于下列内容：

a) 故障处置过程中，自动保存模型参数与任务上下文等，在故障处置完成后，自动恢复任务运行状态；
b) 在资源缺乏时，及时对承载任务的虚拟机、容器等进行安全迁移，符合任务的高可用需求。

5.3 应用支撑层安全功能

5.3.1 数据处理安全

数据处理安全应符合 GB/T 41479—2022 第 5 章规定的要求，还包括下列内容：

a) 对数据集生成物料清单文件，随同数据集流转，支撑验证数据集的完整性、真实性、版本正确性和安全缺陷追溯等；
b) 保护国家重要或核心数据、商业秘密和个人信息等数据的保密性与完整性。

5.3.2 模型训练安全

模型训练安全包括但不限于下列内容：
a) 对模型生成物料清单文件，随同模型文件流转，支撑验证模型的完整性、真实性、版本正确性和安全缺陷追溯等；
b) 对模型训练过程留存不可篡改的日志记录，支撑合规审计与问题回溯等。

5.3.3 模型推理安全

模型推理安全包括但不限于下列内容：
a) 在不更改推理任务工作流程的条件下，对模型进行授权控制或加解密保护；
b) 在推理任务执行之前，对模型的完整性、真实性和版本正确性进行验证；
c) 对推理请求进行安全检测，避免推理请求者诱骗获取模型或训练数据中的敏感信息；
d) 对推理结果进行安全检测，避免模型或训练数据中的敏感信息被诱导输出。

6 安全管理

6.1 身份管理

人工智能计算平台的身份管理应符合 GB/T 22239—2019 中 8.1.5.1 和 8.1.5.3 规定的要求。

6.2 密码管理

人工智能计算平台的密码管理应符合 GB/T 22239—2019 中 8.1.10.9 规定的要求，还包括下列内容：
a) 采用安全强度等符合密码相关国家标准与行业标准相关要求的密码算法，保护模型与数据的保密性、完整性和真实性；
b) 对模型和数据加密等的密钥进行安全管理，其中，密钥生存周期管理见 GB/T 39786—2021 的附录 B。

6.3 日志管理

人工智能计算平台日志管理应符合 GB/T 22239—2019 中 8.1.10.6 e)、f)和 g)规定的要求，还包括下列内容：
a) 维护并管理数据处理、模型训练和模型推理等全流程的日志记录；
b) 对日志访问请求进行访问控制，并保护日志的完整性；
c) 提供按有关参数条件查阅和审计日志的功能，符合数据处理、模型训练和模型推理过程的透明性审计需求。

6.4 安全监测

人工智能计算平台的安全监测应符合 GB/T 22239—2019 中 8.1.5.4 c)规定的要求及其他相关标准要求。

6.5 安全审计

人工智能计算平台的安全审计应符合 GB/T 22239—2019 中 8.1.4.3 和 8.1.5.2 规定的要求，还需对不同任务的资源占用情况进行监测与审计，识别恶意资源占用和容器逃逸等攻击行为。

6.6 风险管理

人工智能计算平台的风险管理应符合 GB/T 22239—2019 中 8.1.10.5 规定的要求，还包括下列内容：

a) 对人工智能计算平台相关固件、系统软件、数据处理组件、模型训练组件和模型推理组件等进行安全更新和漏洞修复；
b) 接收到人工智能计算平台相关漏洞时，及时进行漏洞验证和修复，并发布补丁。

6.7 个人信息保护

人工智能计算平台中的个人信息处理应符合 GB/T 35273—2020 中 6.1 和 6.3 规定的要求，还应符合 GB/T 22239—2019 中 8.1.4.11 规定的要求。

7 角色安全职责

7.1 概述

人工智能计算平台的参与角色分为平台提供方、数据提供方、模型提供方和应用提供方。各参与角色的描述和典型活动见附录 B。

7.2 平台提供方

平台提供方的安全职责包括但不限于：

a) 提供安全功能，在人工智能应用开发与运行过程中，保护模型与数据安全；
b) 实施安全管理，在人工智能应用开发与运行过程中，进行用户身份管理与日志记录，并持续开展安全监测、安全审计与风险管理；
c) 以符合相关国家标准与行业标准要求的方式，开展密码应用与个人信息保护。

7.3 数据提供方

数据提供方的安全职责包括但不限于：

a) 结合人工智能计算平台的安全功能与安全管理，为模型提供方和应用提供方提供安全的数据集；
b) 支撑模型提供方和应用提供方验证数据集的完整性、真实性、版本正确性和安全缺陷追溯等；
c) 支撑模型提供方和应用提供方对数据处理过程进行透明性审计等。

7.4 模型提供方

模型提供方的安全职责包括但不限于：

a) 涉及非自有数据的，符合数据提供方的安全要求，基于数据集开展模型训练；
b) 涉及非自有预训练模型的，符合预训练模型提供方的安全要求，基于预训练模型开展微调训练；
c) 结合人工智能计算平台的安全功能与安全管理，为应用提供方提供安全的模型或模型推理接口；

d） 支撑应用提供方验证模型的完整性、真实性、版本正确性和安全缺陷追溯等；

e） 支撑应用提供方对模型训练过程进行透明性审计等。

7.5 应用提供方

应用提供方的安全职责包括但不限于：

a） 符合应用开发与运行的通用安全要求，包括编码安全、测试安全、部署安全和运维安全等；

b） 基于非自有数据或模型开发与运行人工智能应用时，符合数据提供方或模型提供方的安全要求，使用模型或数据；

c） 基于自有数据或模型开发与运行人工智能应用的，符合 7.3 或 7.4 规定的要求；

d） 结合人工智能计算平台的安全功能与安全管理，保护模型与数据安全；

e） 通过标识与鉴别等措施，保护推理请求接收与响应等接口安全；

f） 通过规则匹配和语义分析等措施识别和阻断提示注入等攻击；

g） 在为最终用户提供服务时，记录并留存日志，发生安全事件时应可追溯。

附 录 A
（资料性）
平台组成与安全风险示例

A.1 平台组成

人工智能计算平台的组成见图 A.1，其中。

a） 资源层：包括物理资源和虚拟资源，物理资源包括计算资源、存储资源和网络资源，特别是专为人工智能应用开发与运行设计的人工智能加速计算资源、高速存储资源和高速网络资源。其中，人工智能加速计算资源是指一台或多台人工智能服务器以及与其配套的软件等，高速存储资源或高速网络资源是指基于高带宽、低时延的接口和通信协议等设计的存储设备或网络设备，以及与其配套的软件。虚拟资源是物理资源的虚拟化形态。

b） 调度层：包括资源调度和任务调度组件，将人工智能加速计算资源、高速存储资源和高速网络资源纳入统一管理和调度，支撑人工智能应用开发与运行涉及的诸多任务有序、高效执行。

c） 应用支撑层：包括数据处理、模型训练和模型推理组件，高效实现任务切分，提升任务运行并行度及资源利用率等，进而提升人工智能应用开发与运行效率。其中，数据处理和模型训练组件作用于人工智能应用开发阶段，模型推理组件作用于人工智能应用运行阶段。

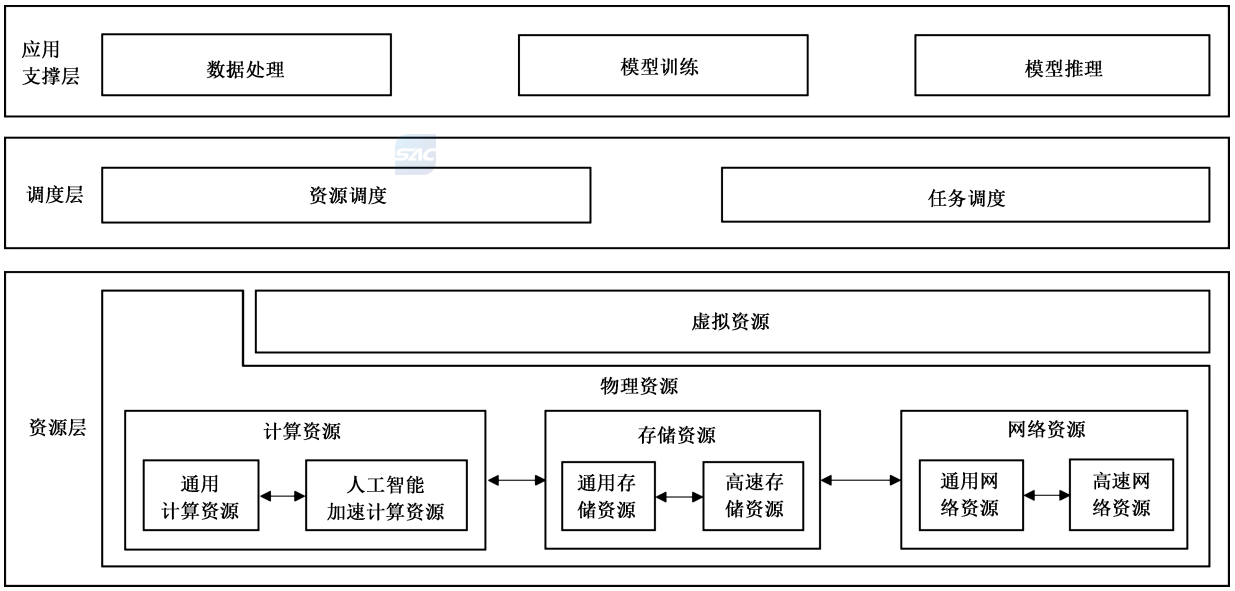

图 A.1 人工智能计算平台组成

A.2 安全风险示例

网络和数据安全风险在人工智能计算平台中依然存在。本文件主要关注以下三个方面的安全风险，典型的安全风险示例见表 A.1，其中：

a） **资源层安全风险**：人工智能计算平台引入人工智能加速计算资源、高速存储资源和高速网络资源，进而引入新的风险暴露面（例如，人工智能加速处理器的固件和内存等），一旦被攻击或破坏，将导致模型与数据在资源层传输、存储和运算过程中，遭窃取或篡改等；

b） **调度层安全风险**：人工智能计算平台支撑数据处理、模型训练与模型推理，需实现大规模算力

资源高效、并行工作，资源调度和任务调度组件的组成与工作模式随之发生变化，相应的调度组件一旦被攻击或破坏，将导致任务中断、模型与数据丢失等；

c) **应用支撑层安全风险**：首先，人工智能计算平台的数据处理、模型训练和模型推理相关组件可能被非授权访问，导致模型与数据遭窃取或篡改等，其次，模型与数据可能由多个参与方处理并在多个资源环境之间流转，模型与数据流转过程中，可能遭窃取或篡改等，而模型与数据本身的偏见、歧视与后门等脆弱性被利用导致的安全风险，则不在本文件范围内。

表 A.1　人工智能计算平台安全风险示例

平台层组成		安全风险示例
资源层	计算资源	人工智能加速处理器配套固件可能被篡改、替换或恶意回退到有安全缺陷的版本等
		管理人员等恶意利用人工智能加速处理器通信接口，非授权访问人工智能加速处理器内存中的模型与数据等
		人工智能加速处理器或人工智能服务器等可能出现故障，导致任务中断，模型与数据丢失等
		中央处理器与人工智能加速处理器协同工作，甚至多个人工智能服务器之间的内存借用时，人工智能加速处理器内存可能被非授权访问（内存转储等多种方式），导致模型与数据遭窃取或篡改等
		内存（包括主存与人工智能加速处理器内存）等被恶意转储、非授权或越权访问等，导致模型与数据在推理过程中遭窃取或篡改等
	存储资源	存储资源可能出现单点故障，导致模型训练过程记录的检查点内容丢失等
		存储空间耗尽，导致任务中断，模型与数据丢失等
		存储系统被非授权访问，导致其中的模型与数据等被窃取、篡改或恶意加密与勒索等
	网络资源	通信网络被非法入侵或非授权访问等，导致模型与数据遭窃取或篡改等
		数据传输信道被非法监听或干扰破坏等，导致模型与数据泄露、丢失等
		人工智能高速网络（例如，基于 InfiniBand 协议的高速网络）中，多个人工智能服务器之间数据互访时，出现非授权或越权的数据访问
		运维管理人员通过流量镜像或监听接口等技术手段，窃取处理器间或服务器间互联信道中的模型与数据等资产
	虚拟资源	人工智能加速计算资源虚拟化机制不安全，导致并行运行的多任务之间，出现资源劫持控制、数据窃取或故障扩散等
调度层	资源调度	模型推理任务遭海绵样本攻击或被植入挖矿程序等，导致计算资源被恶意消耗，正常的模型推理请求无法得到响应
	任务调度	模型训练过程中，因为人工智能加速处理器、人工智能服务器、通信网络或参数面互连线路等故障等，导致模型训练任务中断、模型与数据丢失等
		模型训练或推理任务执行过程中，因为资源短缺等，导致任务中断，模型与数据丢失等

表 A.1　人工智能计算平台安全风险示例（续）

平台层组成		安全风险示例
应用支撑层	数据处理	数据处理相关组件可能存在安全缺陷等，导致数据提供者基于相关组件进行数据处理时，数据处理结果不可靠、数据集遭篡改或替换等
		可能因为内部人员作恶或攻击者入侵等因素，导致私有化数据集等被篡改，无法训练出符合预期目标的模型
	模型训练	模型训练相关组件可能存在安全缺陷等，导致模型提供者基于相关组件进行模型训练时，模型训练结果不达预期、模型文件遭篡改或替换等
		可能因为内部人员作恶或攻击者入侵等因素，导致模型训练参数、中间版本模型文件或数据集等被篡改，导致无法训练出符合预期目标的模型
	模型推理	模型推理相关组件可能存在安全缺陷等，导致应用提供者基于相关组件进行模型推理时，模型文件可能遭篡改或替换等，导致无法获得正确的推理结果等
		推理阶段使用了版本错误的、被篡改、伪造或替换的模型文件或推理缓存数据等，导致模型推理结果出错等
		因为内部人员作恶或攻击者入侵等因素，导致模型文件被非授权复制到其他设备中运行

附 录 B
（资料性）
角色描述与典型活动

人工智能应用开发与运行过程中涉及的角色与典型活动见图 B.1，其中：

a） 平台提供方：设计和建设、管理和运维人工智能计算平台的参与方，典型活动包括平台设计和建设、管理和运维；

b） 数据提供方：为人工智能应用开发与运行提供数据的参与方，典型活动包括数据处理和数据供给；

c） 模型提供方：为人工智能应用开发与运行提供模型或模型推理接口的参与方，典型活动包括模型搭建、模型训练、模型验证、模型测试、模型更新和模型退役；

d） 应用提供方：开发与运行人工智能应用的参与方，典型活动包括应用开发、模型部署、模型评估、模型推理、应用运行和模型销毁，其中人工智能应用运行核心依赖的是模型推理，基于模型推理结果实现业务目标。

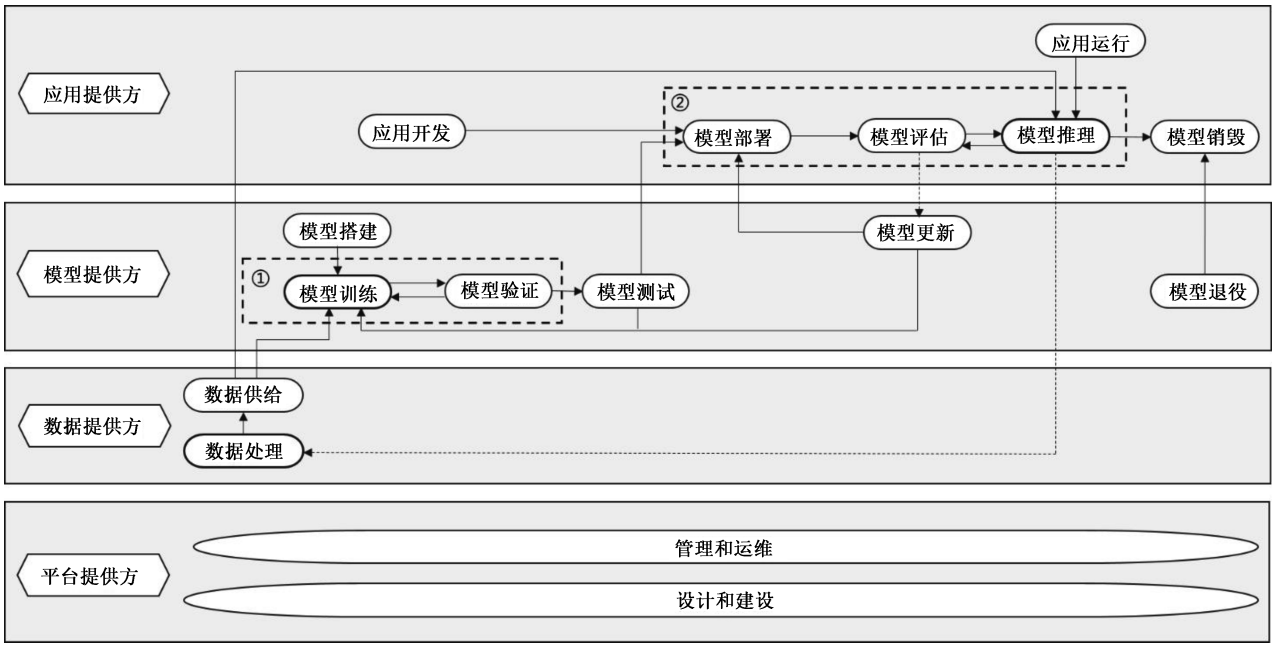

图 B.1 角色典型活动

对于大模型，角色活动会在图 B.1 中①和②虚线框标识处存在区别，大模型的模型训练与模型验证之间还存在诸多优化与微调活动，例如，有监督微调、基于人类反馈的强化学习与指令微调等，见图 B.2。大模型的模型部署与模型评估之间还存在提示工程活动，见图 B.3。

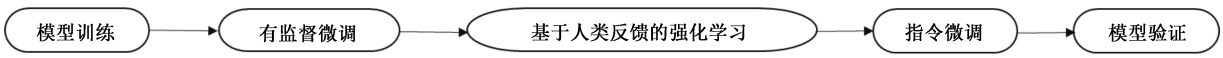

图 B.2 大模型优化与微调活动

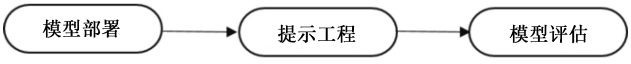

图 B.3 大模型提示工程活动

参 考 文 献

[1] GB/T 18336.2—2024 网络安全技术 信息技术安全评估准则 第2部分:安全功能组件

[2] GB/T 20272—2019 信息安全技术 操作系统安全技术要求

[3] GB/T 36635—2018 信息安全技术 网络安全监测基本要求与实施指南

[4] GB/T 36639—2018 信息安全技术 可信计算规范 服务器可信支撑平台

[5] GB/T 39680—2020 信息安全技术 服务器安全技术要求和测评准则

[6] GB/T 39786—2021 信息安全技术 信息系统密码应用基本要求

[7] GB/T 41867—2022 信息技术 人工智能 术语

[8] GB/T 42018—2022 信息技术 人工智能 平台计算资源规范

[9] GB/T 42570—2023 信息安全技术 区块链技术安全框架

[10] GB/T 42888—2023 信息安全技术 机器学习算法安全评估规范

[11] ISO/IEC 22989:2022 Information technology—Artificial intelligence—Artificial intelligence concepts and terminology

[12] ISO/IEC 23053:2022 Framework for Artificial Intelligence(AI) Systems Using Machine Learning(ML)

[13] PCI-SIG TDISP(TEE Device Interface Security Protocol)

[14] 全国网络安全标准化技术委员会.人工智能安全治理框架 1.0 .2024.

[15] ETSI GR SAI 005 Securing Artificial Intelligence (SAI). Mitigation Strategy Report

[16] CycloneDX Machine Learning Bill of Materials(ML-BOM)

[17] SPDX 3.0 AI profile

[18] OWASP Top 10 for LLM