ICS 35.040
CCS L 80

T

团　　体　　标　　准

T/ISEAA 006—2024

大模型系统安全测评要求

Security evaluation requirements for large model system

2024-04-30 发布　　2024-06-01 实施

中关村信息安全测评联盟　发布
中国标准出版社　出版

目　　次

前言 …… Ⅲ
引言 …… Ⅳ
1　范围 …… 1
2　规范性引用文件 …… 1
3　术语和定义 …… 1
4　概述 …… 1
4.1　大模型系统架构 …… 1
4.2　安全测评方法 …… 2
5　通用安全测评要求 …… 3
5.1　物理环境 …… 3
5.2　网络架构 …… 3
5.3　边界防护 …… 4
5.4　身份鉴别 …… 5
5.5　访问控制 …… 6
5.6　安全审计 …… 7
5.7　入侵防范 …… 9
5.8　恶意代码防范 …… 10
5.9　集中管控 …… 10
5.10　供应链管理 …… 11
5.11　个人信息保护 …… 12
5.12　数据安全保护 …… 14
5.13　备份恢复 …… 15
5.14　数据溯源 …… 15
6　设计开发安全测评要求 …… 16
6.1　数据处理 …… 16
6.2　模型保护 …… 19
6.3　内容安全 …… 21
7　测试安全测评要求 …… 23
7.1　模型评估 …… 23
7.2　模型更新 …… 24
8　部署与运行安全测评要求 …… 25
8.1　模型部署 …… 25
8.2　攻击检测 …… 26

8.3 运行监测 …… 27
8.4 系统管理 …… 28
8.5 变更管理 …… 28
8.6 安全事件处置 …… 29
8.7 应急预案管理 …… 29
9 退役安全测评要求 …… 30
9.1 模型退役 …… 30
9.2 数据删除 …… 31
10 测评结论判定 …… 31
10.1 分类测评结果 …… 31
10.2 风险分析和评价 …… 31
10.3 测评结论判定 …… 31
参考文献 …… 32

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。

本文件由中关村信息安全测评联盟提出并归口。

本文件起草单位：公安部第三研究所、北京百度网讯科技有限公司、蚂蚁科技集团股份有限公司、浙江大学、北京天融信网络安全技术有限公司、启明星辰信息技术集团股份有限公司、快手科技有限公司、上海商汤智能科技有限公司、北京远鉴信息技术有限公司、深圳市网安计算机安全检测技术有限公司、山东新潮信息技术有限公司、金盾检测技术股份有限公司、上海市信息安全测评认证中心、杭州中尔网络科技有限公司、华为技术有限公司、阿里云计算有限公司、深信服科技股份有限公司、科大讯飞股份有限公司、北京百川智能科技有限公司、北京小桔科技有限公司、优刻得科技股份有限公司、浙江君同智能科技有限责任公司。

本文件主要起草人：袁静、刘金鑫、刘楠、张国鹏、文煜乾、陆臻、陈广勇、苏艳芳、刘继顺、郭建领、唐佳伟、李荣昌、杨剑、蒋发群、谷晨、徐浩、郑榕、牛建红、刘智广、高亚敏、朱熹铭、王余、王龑、邵英杰、孙钰娆、张瑶、严敏瑞、孙勇、叶润国、吴晓杰、李建民、樊玉杰、冯明、韩蒙、林昶廷。

引　言

为规范大模型系统安全测评工作的开展，本文件对大模型系统进行安全测评的技术活动提出要求，为评价大模型系统是否符合 T/ISEAA 005—2024《大模型系统安全保护要求》提供了获取证据的途径和方法，用以指导测评人员对大模型系统进行测试评估。

大模型系统安全测评要求

1 范围

本文件规定了大模型系统是否符合 T/ISEAA 005—2024 所进行的测试评估活动的要求。

本文件适用于指导网络安全测评服务机构、大模型开发者、大模型系统运营者对大模型及大模型系统安全保护状况进行安全测试评估，大模型安全监管职能部门依法开展大模型安全保护监督检查参考使用。

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 25069 信息安全技术 术语

T/ISEAA 005—2024 大模型系统安全保护要求

3 术语和定义

GB/T 22239、GB/T 25069 界定的以及下列术语和定义适用于本文件。

3.1

大模型 large model

基于大量数据集训练得到的，具备大参数量、复杂结构的机器学习模型。

[来源：T/ISEAA 005—2024，3.1]

3.2

大模型服务 large model service

基于数据、算法、模型、规则，能够根据使用者提示生成文本、图片、音频、视频等内容的人工智能服务。

注：大模型服务根据行业应用可分为通用大模型服务和行业大模型服务。

[来源：T/ISEAA 005—2024，3.2]

3.3

大模型系统 large model system

由计算机或者其他信息终端及相关设备组成的按照一定的规则和程序，基于大模型进行数据处理或提供大模型服务的系统或平台。

[来源：T/ISEAA 005—2024，3.3]

4 概述

4.1 大模型系统架构

大模型系统可抽象为基础设施层、平台层、模型层和应用层，如图 1 所示。