ICS 35.240
CCS L70

团 体 标 准

T/ISCXXX—XXXX

安全可信智能云平台能力分级方法

Capability Grading Method for Secure and Trustworthy Intelligent Cloud Platforms

（征求意见稿）

xxxx-xx-xx 发布 xxxx-xx-xx 实施

中 国 互 联 网 协 会 发布

目次

前　言

本文件按照GB/T1.1—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。

本文件由中国互联网协会提出并归口。

本文件起草单位：

本文件主要起草人：

——

引　言

随着人工智能技术与云计算深度融合，智能云平台已成为数字经济时代的关键基础设施。在技术快速迭代与应用深度拓展的同时，安全、可信、可控、可靠成为智能云服务发展的核心诉求。当前，我国智能云产业发展面临核心技术自主水平不足、安全风险防范体系不完善、产业生态协同性不高等挑战，亟需建立统一、规范的能力要求标准，引导产业健康有序发展。

本文件立足于国家网络安全与信息化发展战略，结合信创产业发展实际，从安全可信与技术能力双维度构建评价体系。通过明确智能云平台在基础资质、设计研发、安全合规、运营保障等方面的可信要求，以及在基础设施、基础服务、安全运维和性能指标等方面的技术要求，为相关产品研发、服务提供与能力评价提供系统性指导。

本文件注重与GB/T 46350—2025《信息技术 云计算 智能云服务通用要求》等国家标准的衔接，同时聚焦安全可信特性，强化了安全可信技术栈兼容、数据安全、模型安全等关键能力要求，旨在推动构建安全可信的智能云服务生态。

本文件明确了安全可信智能云平台的能力分级方法。

对本文件中的具体事项，法律法规另有规定的，需遵照其规定执行。

安全可信智能云平台能力分级方法

1 范围

本文件从安全可信特性、智能云基础能力和性能三方面规范了安全可信智能云平台能力分级方法。

本文件适用于指导安全可信智能云平台能力评价。

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 46350—2025 信息技术 云计算 智能云服务通用要求

3 术语和定义

GB/T 46350—2025界定的以及下列术语和定义适用于本文件。为了便于使用，以下重复列出了GB/T 46350—2025中的某些术语和定义。

3.1

云计算 cloud computing

通过网络将可伸缩、弹性的共享物理和虚拟资源池以按需自服务的方式供应和管理的模式。

注：资源包括服务器、操作系统、网络、软件、应用和存储设备等。

[来源：GB/T 46350—2025，3.1]

3.2

云服务 cloud service

通过云计算（3.1）已定义的接口提供的一种或多种能力。

[来源：GB/T 32400—2015,3.2.8]

3.3

智能云服务 AI cloud service

支撑或提供人工智能能力的云服务。注：人工智能能力指模型开发、模型服务、人工智能应用开发等。

[来源：GB/T 46350—2025，3.3]

3.4

MLOps machine learning operations

MLOps 是一组用于自动化和简化机器学习工作流程与部署的实践，旨在统一机器学习系统开发（Dev）与运营（Ops），实现模型的快速试验、部署、质量保证与端到端追踪。

3.5

LMOps language model operations

LMOps 是面向大语言模型（LLM）的 MLOps 实践，专注于管理由 LLM 驱动的应用程序全生命周期，包括提示工程、微调、检索增强、推理优化与内容治理等。

4 符号和缩略语

下列符号和缩略语适用于本文件。

AI： 人工智能(ArtificialIntelligence)
API： 应用程序接口(Application Programming Interface)
ASIC： 专用集成电路(Application SpecificIntegrated Circuit)
CCIX： 缓存一致性加速器互联(Cache Coherent Interconnect for Accelerators)
CPU： 中央处理器(Central ProcessingUnit)
DPO： 直接偏好优化(Direct Preference Optimization)
FPGA： 现场可编程门阵列(Field Programmable GateArray)
GPU： 图形处理器(Graphics Processing Unit)
IDE： 集成开发环境(Integrated Development Environment)
LMOps： 大模型运维(Large Model Operations)
MLOps： 机器学习运维(Machine Learning Operations)
P2P： 点对点(Peer-to-Peer)
PCIe： 高速外围部件互连(Peripheral Component Interconnect Express)
RDMA： 远程直接内存访问(Remote Direct Memory Access)
MFU： 模型算力利用率(Model FLOPs Utilization)
P99： 百分位延迟(99%的请求低于该值)
RDMA： 远程直接内存访问(Remote Direct Memory Access)
TTFT： 首token延迟(Time To First Token)
TPS： 生成吞吐(Tokens Per Second)
QPS： 请求并发处理能力(Queries Per Second)
AllReduce：集合通信操作(常用于梯度同步)
CKPT： 模型检查点(Checkpoint，模型训练状态快照，用于容错恢复)

5 安全可信要求

5.1 基础资质

5.1.1 企业资质

a） 组织应为在中华人民共和国境内注册的企业法人或事业法人，应能够提供真实有效的法人资质证明文件，企业成立36个月以上，企业组织机构代码证经营范围覆盖申请产品。
b） 组织最高管理者，如法定代表人、主要负责人、实际控制人、董（监）事会人员、公司高层管理人员，应具有中华人民共和国国籍。
c） 组织的资本构成应符合通过间接方式投资的外方投资者及其一致行动人的出资比例最终不超过20%。

d） 组织应具备相关产品研发与制造经验，并至少提供近3年（自然年）内的1项研发证明（如产品认证证书、设计开发资料等）及1项制造证明（如出货记录、订购合同、发票等）。

5.1.2 产品信息

a） 应具备产品相关软件著作权或专利证书（已在国家知识产权局备案）等知识产权证明材料。
b） 应具备安全可信环境下的适配测试报告（内容涵盖产品主要功能和基础性能）。

5.2 设计研发

5.2.1 研发团队

a） 应具备不少于100人规模的自有研发团队，核心成员应具备3年以上相关经验。

5.2.2 研发管理

a） 应具备覆盖需求分析、设计、开发、测试、运维、退役的全生命周期研发管理体系。
b） 应实现安全测试全生命周期覆盖，如静态代码扫描、动态渗透测试等。
c） 宜采用安全可信的代码签名工具，确保发布包完整性与来源可信性。

5.2.3 环境配套

a） 应至少具备2套基于不同芯片基础架构的安全可信技术栈环境，用于产品的兼容性测试；

5.3 安全合规

5.3.1 安全管理

a） 应建立安全事件分级响应机制，明确高危事件（如数据泄露）的处置流程与时限；
b） 应提供漏洞扫描报告，且不存在中、高危漏洞；
c） 宜具备漏洞扫描机制，扫描规则库版本更新周期不超过一周；
d） 宜具备威胁情报共享机制，实时同步行业最新漏洞信息。

5.3.2 技术合规

a） 应具备清晰的软件物料清单（SBOM），SBOM应覆盖全部物料，确保组件来源透明化且无许可证风险；
b） 应支持数据本地化存储，确保用户数据仅存储在可控数据中心。

5.4 运营保障

5.4.1 服务保障

a） 应提供7×24小时技术支持服务，一般问题响应时间应不超过4小时，问题解决时间不超过8小时；
b） 应提供政企客户上门服务保障；
c） 应提供产品使用培训及典型问题解决说明材料。

5.5 生态合作

5.5.1 生态适配

a） 应兼容至少2种不同技术架构的安全可信技术栈；
b） 宜通过权威机构组织的生态适配认证。

5.5.2 生态共建

a） 应为信创相关生态平台或组织的成员单位，并参与推进信创生态建设；
b） 宜加入至少1个国内开源社区，并贡献代码或文档；
c） 宜参与安全可信相关标准制定，年度参与数量不少于2项；
d） 宜参与开源社区漏洞众测计划，主动公开修复方案并提交漏洞库备案。

5.6 可持续性

5.6.1 可持续发展

a） 年度研发费用占比应不少于15%；
b） 应具备产品版本迭代计划，年度至少发布2版次更新；
c） 应具有核心技术演进路线图；
d） 应承诺产品支持周期不低于5年；
e） 宜建立代码混淆与防逆向工程机制，保护核心算法与知识产权。

6 平台基础能力要求

6.1 基础环境要求

a） 应基于兼容至少一款符合安全可信要求的CPU、GPU芯片；
b） 应兼容至少一款符合安全可信要求的操作系统、数据库；
c） 应兼容符合安全可信要求的中间件产品；

6.2 基础设施

6.2.1 计算资源增强

6.2.1.1 计算资源通用要求

a） 应支持基于国产架构(如ARM、LoongArch等)的裸金属、虚拟机和容器；
b） 应支持不少于2种主流国产AI芯片的接入、识别与驱动管理；
c） 应支持国产AI芯片的切分、显存隔离与算力聚合；
裸金属、虚拟机、容器应支持：
d） 不同类型和数量的人工智能加速处理器，包括GPU、ASIC、FPGA等；
e） 神经网络计算，如卷积计算、向量计算、标量计算等；
f） 人工智能加速处理器的多卡P2P通信；
g） 人工智能加速处理器间互联，如CCIX、PCIe等互联方式。
注：裸金属指未预装操作系统和虚拟化软件，提供物理服务器资源的云服务。
h） 虚拟机应支持人工智能加速处理器在单卡分片模式下多实例间计算单元、显存单元的隔离；
i） 虚拟机宜支持人工智能加速处理器在单卡分片模式下多实例间编解码的隔离；
j） 容器应支持不同数量的人工智能加速处理器，如1卡、2卡、4卡、8卡等；

k） 容器应支持单卡和多卡人工智能加速处理器共享和隔离。

6.2.1.2 计算资源管理要求

a） 应支持对裸金属、虚拟机的生命周期管理，如创建、关机、启动、重启、释放等；
b） 应支持容器的生命周期管理，如创建、删除等；
c） 应提供包含多种人工智能加速处理器驱动程序的操作系统、虚拟机、容器镜像；
d） 应支持虚拟机、容器配置变更，如调整挂载的显卡数量；

6.2.1.3 计算加速

a） 应支持基础算子加速，如ElementWise类算子、Tensor与矩阵计算类算子、Tensor变换类算子、池化类算子等；
b） 应支持融合算子加速，如ConvolutionBiasRelu等；
c） 宜支持多种开源模型的训练和推理加速；
d） 宜支持采用多种异构计算芯片进行训练和推理加速；
e） 可提供根据环境配置自动匹配并行策略的工具；
f） 可提供模型权重格式转换与切分工具，在不同训练框架中对模型权重进行格式转换与切分。

6.2.2 存储资源增强

6.2.2.1 存储资源

a） 应支持块存储服务；
b） 应支持并行文件存储服务；
c） 应支持对象存储服务；
d） 宜支持对象存储数据挂载到并行文件存储；
e） 应支持面向AI的大带宽、低延时并行文件存储；
f） 应支持多级缓存策略(内存/SSD/HDD)，优化国产环境下的IO瓶颈。

6.2.2.2 存储资源管理

a） 应支持块存储的创建、删除、权限管理、访问和查询；
b） 应支持并行文件存储的挂载、卸载、权限管理、访问和查询；
c） 应支持对象存储的创建、删除、权限管理、访问和查询；
d） 宜支持对象存储的挂载管理。

6.2.2.3 存储加速

a） 应支持对接不同类型的数据源，如对象存储、文件存储、块存储等；
b） 应支持多种缓存结构，如内存缓存、磁盘缓存等；
c） 应支持配置数据缓存策略，如动态加载、预加载、缓存准入、缓存替换等；
d） 宜支持以层级命名空间访问数据源；
e） 宜支持人工智能加速处理器读写数据；
f） 可提供检查点异步加速能力，减少计算任务的阻塞时长。

6.2.3 网络资源增强

6.2.3.1 网络资源

a） 计算节点间和容器间应支持RDMA网络，如采用InfiniBand、基于融合以太网的RDMA(RoC等网络协议；注：计算节点包括裸金属和虚拟机。
b） 应支持多租户间RDMA网络隔离；
c） 应支持多租户间性能隔离，如带宽隔离、包速率隔离等；
d） 应支持RoCEv2或InfiniBand网络。

6.2.3.2 网络资源管理

a） 网络资源管理应支持配置RDMA网络隔离。

6.2.3.3 网络加速

a） 应支持集合通信的语义，如Send、recv、all-gather等；
b） 应支持节点内的拓扑感知；
c） 应支持节点间的拓扑感知；
d） 应支持感知网络慢速的节点；
e） 宜支持通信算子卸载到硬件（如交换机、智能网卡等）。

6.2.4 任务调度

6.2.4.1 任务管理

a） 应支持任务生命周期管理，如创建、删除、启动、暂停、恢复等；
b） 应支持多种任务提交方式，如命令行、控制台、API等；
c） 应支持多种类型的任务，包括训练任务、推理任务等；
d） 应支持多种场景的任务资源配额配置；
e） 应支持查看任务信息，如运行状态信息、监控和日志信息等；
f） 应支持任务使用计算资源的弹性扩缩容；
g） 可支持多层级资源配额设置。

6.2.4.2 任务调度

a） 应支持拓扑感知调度，如人工智能加速处理器间拓扑感知调度、节点间网络拓扑感知调度等；
b） 应支持同一任务中多个实例的资源批量调度；
c） 应支持按人工智能加速处理器的类型调度；
d） 应支持共享与独占任务调度；
e） 应支持任务优先级调度；
f） 应支持抢占式调度，如队列内抢占、队列间抢占等；
g） 应支持任务调度时，本队列在资源不足时向其他队列借用资源；
h） 应支持多种调度策略，如Binpack、Spread等；
i） 宜支持训练与推理任务混合调度；
j） 宜支持单一任务同时调度多种人工智能加速处理器进行混合训练。

6.2.4.3 任务容错

a） 应支持多种故障判断，如任务异常退出、假死等；
b） 应支持多种任务恢复方式，如本地重启恢复、自动重调度恢复等。

6.3 基础服务

6.3.1 模型开发

6.3.1.1 数据处理

6.3.1.1.1 数据接入

a） 应支持接入多种来源的数据，如块存储、文件存储和对象存储等；
b） 应支持接入结构化和半结构化数据，如csv、tsv、txt、parquet等数据类型；
c） 应支持接入非结构化数据，包括图片、语音、文本等数据类型；
d） 应支持接入压缩包文件；
e） 宜支持数据同步，并支持设置数据同步策略；
f） 宜支持接入加密数据。

6.3.1.1.2 数据预处理

a） 应支持结构化数据的清洗，如数据拆分、异常值检测、缺失值填充等；
b） 应支持非结构化（多模态）数据的清洗，根据特定规则剔除不符合要求的非结构化数据，如内容去重等；
c） 应支持自定义数据预处理功能，如用户自定义预处理算法等；
d） 宜支持自动预处理。

6.3.1.1.3 数据标注

a） 应支持多种数据类型的标注工具或模板，如文本类、表格类、图片类、音视频类等；
b） 应支持对标注标签、标注属性等标注信息的管理，如编辑、删除和查询等；
c） 应支持可视化标注，标注信息在原始数据直观呈现；
d） 应支持团队标注的管理，如任务管理、人员管理等；
e） 应支持对标注的评估，如准确性、有效性等；
f） 应支持对标注数据、标签等标注结果导出；
g） 宜支持智能标注，如调用算法或外部服务自动标注数据、通过训练算法自动标注等。

6.3.1.1.4 数据管理

a） 应支持数据集的生命周期管理，如创建、删除、导入、导出、发布等；
b） 应支持统一管理图片、文本、音频、视频、表格等类型数据；
c） 应支持有标注数据和无标注数据的导入、导出、查看；
d） 应支持数据集信息的展示和查询，如原始数据、数据标注信息、标签信息等；
e） 应支持数据集的管理，如权限管理、版本管理等；
f） 应支持数据集的共享。

6.3.1.1.5 数据分析

a） 应支持结构化数据的预览；
b） 应支持非结构化数据的预览，如文本、图片、视频、音频等类型数据；
c） 应支持数据集的分析，如数据集的统计特征分析、质量特征分析等；
d） 应支持数据分析的可视化，如数据分布可视化、标签分布可视化等；
e） 宜支持数据集重构，如通过清洗、集合、填充、过滤等操作形成新的数据集；
f） 宜支持多种维度的数据质检，如样本的数量、完整度、分布等。

6.3.1.1.6 数据增强

a） 应支持文本类数据的多种数据增强策略，如回译、同义词替换、非核心词替换等；
b） 应支持图片类数据的多种数据增强策略，如几何变换、扭曲图像、加噪声、色彩抖动等；
c） 宜支持音频类数据的多种数据增强策略，如加噪声、调整音量、调整混响等时域增强策略，或调整音高、调整播放速度、频谱交换等频域增强策略；
d） 可支持多种类型数据的自动增强，如文本类数据、图片类数据、音频类等；
e） 可支持对指令数据集和多轮会话数据集的数据扩充和增强；
f） 可支持对提示词数据集的转换和增强。

6.3.1.1.7 数据闭环

a） 应支持对模型请求数据日志和业务操作日志的管理，如对原始请求输入和模型输出数据及链路日志的查询、存储检索等；
b） 应支持对调用请求数据的分析筛选，如失败案例提取、日志分析等；
c） 应支持请求数据和日志筛选结果与训练数据对接；
d） 应支持数据加密、脱敏等安全策略；
e） 宜支持多维度的请求数据和调用日志管理，如时间维度、业务应用维度等；
f） 宜支持定制的打点采集，如采集用户对模型返回结果评价等；
g） 宜支持对请求数据日志和业务操作日志运行自定义脚本，如执行定时脚本等。

6.3.1.2 模型构建

6.3.1.2.1 算法仓库

a） 应支持多种机器学习算法的存储和查询，如分类、回归、聚类等；
b） 应支持多种深度学习算法的存储和查询，如卷积神经网络、循环神经网络等；
c） 应支持多种计算机视觉类算法的存储和查询，如目标检测、图像分类、文字识别等；
d） 应支持多种语音类算法的存储和查询，如声音分类等；
e） 应支持多种自然语言处理类算法的存储和查询，如文本分类、文本实体抽取、情感分析等；
f） 应支持自定义算法的存储和查询，如自定义名称、唯一标识、算法组件等；
g） 宜支持迁移学习和强化学习等算法的存储和查询；
h） 宜支持时序状态数据处理算法的存储和查询，如时序预测等。

6.3.1.2.2 算法管理

a） 应支持算法的生命周期管理，如设计、开发、测试、部署等；
b） 应支持算法的版本管理，如版本号管理、版本发布等；
c） 应提供算法相关信息的说明，如算法效果、性能等；
d） 宜支持管理同一算法的多种语言版本或运行环境；
e） 宜提供算法适用场景的使用示例和说明。

6.3.1.2.3 特征工程

a） 应提供特征提取，如按照给定的定义提取特征；
b） 应支持多种特征选择方式，如Gini增益、信息增益、信息值等；
c） 应支持特征组合，如将多个特征组织组合或衍生为新的特征等；
d） 应支持特征转换，如数据归一化、标准化、分箱、数值替换等；

e） 应支持数据降维，如主成分分析、线性判别分析等；
f） 宜支持自定义特征工程方法；
g） 宜支持特征分析可视化，如特征指标的图表可视化等；
h） 宜支持特征异常评估，如基于统计方法、距离方法或谱方法等方法进行异常点检测；
i） 宜支持特征库的管理，如特征存储、分享、特征库接入等；
j） 宜支持多种自动特征工程，如自动特征选择、自动特征衍生等。

6.3.1.2.4 开发环境

a） 应支持交互式编码环境；
b） 应支持对代码的增加、删除、修改和查看；
c） 应支持线上IDE环境；
d） 应支持多种机器学习框架，如Scikit-learn、XGBoost等；
e） 应支持多种深度学习框架，如TensorFlow、PyTorch；
f） 应支持开发环境的管理，如增加、删除、查看、修改等；
g） 宜支持自定义开发环境，如以镜像方式提供自定义的开发环境。

6.3.1.2.5 模型训练

a） 应支持单机多卡、多机多卡和跨地域多节点等分布式训练；
b） 应支持配置训练资源，如CPU核数、GPU个数、内存、GPU显存等；
c） 应支持自定义训练参数，如算法参数、运行参数、训练数据、验证数据等；
d） 应支持训练任务的生命周期管理，如创建、删除、重启等；
e） 应支持对训练任务的状态进行定时检查点保存；
f） 应支持查看训练任务的信息，如训练状态、训练进度、训练结果、训练失败原因等；
g） 应支持多种模型再训练方式，如基于预训练模型微调、算法选择和参数调优等；
h） 应支持模型训练过程的可视化，如训练参数、训练指标、模型图的可视化；
i） 宜支持自定义代码的模型训练；
j） 宜提供自动调参工具，根据模型及数据量自动设定参数；
k） 宜支持人工智能加速处理器单卡虚拟化后多模型训练方式；
l） 宜支持对大规模无监督数据的预训练任务，如千亿规模语言模型的预训练。

6.3.1.2.6 模型调优

a） 应支持机器学习类模型的效果和性能调优，如分类模型、聚类模型、回归模型、序列预测模型等；
b） 应支持生成类模型调优评估，对模型效果进行对比；
c） 应支持增加、删除或修正训练数据集样本，实现模型调优和效果对比；
d） 应支持设置多种模型超参数实现模型调优和效果对比，如全量数据迭代数、批量样本数量等；
e） 应支持多种自动模型调优和效果对比的策略，如EarlyStopping方法、超参网格搜索等；
f） 宜支持异常样本检测，如通过修正数据标签、挖掘潜在噪声样本等优化模型进行检测；
g） 宜支持设置模型权重，实现模型调优和效果对比，如boosting模式等；
h） 宜支持通过编辑神经网络层结构实现模型调优和效果对比，如隐藏层节点数、数据块大小、优化方法等；
i） 宜支持生成式模型的有监督精调、DPO对齐等优化方法。

6.3.1.3 模型部署

6.3.1.3.1 云端部署

a） 应支持模型服务的生命周期管理，如启动、停止、测试等；
b） 应支持模型服务的模型信息查询及展示，如模型基本信息、推理方式、推理状态等；
c） 应支持模型服务的接口信息查询及展示，如版本、实例数、接口格式等；
d） 应支持模型服务的状态信息查询及展示，如运行状态、调用量、调用成功率等；
e） 应支持部署多种人工智能模型推理加速库和面向硬件适配的推理加速库；
f） 应支持自定义模型服务使用的资源规格，如计算资源类型、资源数量等；
g） 应支持模型服务的手动资源调度；
h） 宜支持模型服务的自动资源调度，如按CPU占比、内存占比、显存占比等自定义策略进行自动扩缩容；
i） 宜支持多模型动态编排；
j） 宜支持A/B测试，如按照不同分组策略对不同版本的服务流量进行精准分配和统计分析。

6.3.2 模型运维

6.3.2.1 MLOps 工作流

a） 应支持工作流的生命周期管理，如创建、停止、删除等；
b） 应支持工作流的全链路可观测性与调试能力，支持对流程中任意原子任务执行实例进行追溯；
c） 应支持查询任务单元的输入参数快照、输出数据制品、实时运行日志、性能指标及异常信息；
d） 应支持多种工作流编排方式，如可视化编排、代码编排等；
e） 应支持模型开发全流程工作流，包含数据处理、模型构建、模型管理、模型部署等；
f） 宜支持定制化执行工作流节点，如一键运行、定时执行、信号文件触发执行等；
g） 宜支持自定义工作流算子；
h） 宜支持工作流执行实例的对比，如对比同一任务在不同工作流实例中模型性能、模型效果等；
i） 宜提供工作流模板，如模型训练、模型评估、模型发布等服务流程的模板；
j） 宜支持自动化工作流，如模型的自动化训练、服务自动化发布、漂移监控并触发告警；
k） 宜提供从数据到部署的全流程自动化工作流
l） 宜支持基于历史执行记录的断点重跑与参数修正。

6.3.2.2 LMOps 工作流

a） 应符合6.3.2.1中的MLOps工作流要求；
b） 应支持将大模型的提示工程作为工作流节点，使工作流具备如提示模板、提示词调试环境等能力；
c） 应支持将大模型的指令调优作为工作流节点，使工作流具备如全量参数监督微调(SFT)、部分参数高效率调优低秩自适应（LoR）等能力；
d） 宜支持将大模型生成内容作为工作流节点，使工作流具备如基于人类反馈的强化学习(RLHF)、DPO等能力；
e） 宜支持将大模型效果评估作为工作流节点，使工作流具备如针对大模型的多种基准评估数据集、多种基准评估指标、多种效果评估方法等效果评估能力。

6.3.3 推理服务

6.3.3.1 模型推理

a） 应支持判别式模型的推理，如文本分类、目标检测等；
b） 应支持生成式模型的推理，如文本续写、问答、摘要、文生图等；
c） 应支持对模型推理服务的状态查询和展示，如内存、显存、I/O等指标；
d） 应支持保障数据隐私和安全的模型推理，如对请求数据加密等；
e） 应支持基于国产芯片指标的自动扩缩容；
f） 宜支持流式推理；
g） 宜支持多模态模型推理；
h） 宜支持分布式高可用推理；
i） 宜支持离线批量推理；
j） 宜支持屏蔽底层国产芯片差异，提供统一的推理API接口；
k） 应支持推理服务的多副本负载均衡与故障切换。

6.3.4 模型管理

6.3.4.1 模型仓库

a） 应支持模型仓库的管理和配置，如模型存储、模型版本控制等；
b） 应支持多种模型文件格式，如pdparams、infer.model、lite等；
c） 应支持对模型文件的管理，如导入/导出、删除、排序、分类等；
d） 应支持模型文件的格式转化，如以开放神经网络交换(ONNX)模型格式为中介进行格式转换；
e） 应支持模型可视化，如模型结构、网络层级、网络权重等的可视化；
f） 应支持多种模型部署及测试策略，如滚动更新、灰度测试、A/B测试等；
g） 应支持面向业务场景的多模型编排；
h） 应支持查看模型部署的信息，如部署状态、失败信息、日志等；
i） 宜支持可视化的模型编排交互方式；
j） 宜支持模型热更新，如根据模型评估结果更新模型版本；
k） 宜支持模型信息溯源，如查看模型与数据集、算法间的关系等。

6.3.4.2 模型适配

a） 应支持用户自定义推理服务使用的资源规格，如计算资源类型、资源数量等；
b） 应支持模型的不同操作系统、依赖环境和容器环境适配；
c） 宜支持模型压缩，如模型量化、模型剪枝、模型蒸馏、自动加速等；
d） 宜支持模型与云端、边缘端、终端等多种基础设施适配。注：云端指云服务系统环境，终端指客户端系统环境，边缘端指介于云端和终端之间，与两者相互通信的服务节点的系统环境。

6.3.4.3 模型加速

a） 应支持对多种类型的模型进行加速，如计算机视觉类、语音类、自然语言处理类、多模态类；
b） 应支持多种机器学习或深度学习框架的模型加速；
c） 应支持多种模型加速策略，如量化压缩、多种机器学习或深度学习框架间的转换、模型蒸馏、模型剪枝、模型网络精简等；
d） 应支持适配多种架构的人工智能模型推理加速库的模型加速，如x86、ARM等；
e） 应支持查看模型加速任务详情，如加速状态、任务日志、自动评估结果等；
f） 宜支持模型加速评估，对比加速前后模型的效果、性能；
g） 宜支持面向请求特征的加速策略，如对流式序列生成的加速等；

h） 宜支持面向模型算子与人工智能加速处理器I/O特性结合的加速策略，如FlashAttention、PagedAttention等。

6.3.4.4 模型评估

a） 应支持模型效果评估，如准确率、召回率、F1分数等；
b） 应支持模型性能评估，如CPU占用率、功耗、显存占用率等；
c） 应支持多种类型的模型评估，如计算机视觉类、语音类、自然语言处理类、多模态类等；
d） 应支持配置模型评估环境，如数据集、CPU、人工智能加速处理器、操作系统等；
e） 应支持展示模型评估任务信息，如任务名称、状态、模型类型等信息；
f） 宜支持模型可解释性评估，如PDP、特征重要性等；
g） 宜支持模型评估信息可视化，如模型效果指标、模型性能指标、评估环境配置信息等；
h） 宜支持多模型间的效果对比评估。

6.3.4.5 资产管理

a） 应支持模型相关资产的生命周期管理，如订阅、发布、上架、下架等；
b） 应支持模型相关资产的权限配置，如读写、增加、删除、更新等权限；
c） 宜支持模型相关资产的关联使用；
d） 宜支持模型相关资产的共享。注：模型相关资产指数据、模型代码、参数配置等具有知识产权的数字化要素。

6.4 运维与安全

6.4.1 数据与加密安全

a） 应实现显存、算力的隔离;
b） 镜像仓库需具备国产软件成分分析（SCA）能力，识别非信创组件漏洞;
c） 镜像仓库的基础镜像应经过漏洞扫描且源头可信。
d） 数据传输、存储加密应支持SM2/SM3/SM4国密算法，且支持调用国产加密卡硬件加速。

6.4.2 模型安全

a） 应具备模型鲁棒性检测能力，能够识别并防御针对国产模型的对抗样本攻击。
b） 应支持面向生成式内容的全链路治理能力，支持对用户输入请求与模型输出结果进行多层级安全管控。具体包括：可配置的内容审核策略引擎（支持关键词匹配、语义识别、多模态检测）、合规规则包管理（适配法律法规与行业价值观要求）、人工复核与干预通道、操作审计与效果评估机制，确保内容生成过程可控、可管、可追溯;
c） 应支持与外部工具（API、插件）的交互过程进行审计与风险控制。

6.4.3 异构资源运维

a） 屏蔽不同芯片的底层指标差异，统一展示AI加速卡温度、ECC错误、带宽利用率、显存碎片等关键指标。
b） 从底层驱动到上层框架再到应用的端到端日志聚合与检索。
c） 针对芯片“掉卡”、“慢节点（Straggler）”等现象，应建立亚健康检测机制。
d） 监测到硬件故障时，应自动隔离故障节点，并基于最新的Checkpoint自动重启训练任务，无需人工干预。

6.4.4 资源运营审计

a） 应支持多层级的部门/项目资源配额管理。
b） 应对平台内的所有资源申请、模型下载、数据访问行为进行不可篡改的审计记录。

7 性能要求

7.1 高性能 IaaS 基础设施

7.1.1 算力性能及规模

a） 应记录芯片基础算力指标；
定义：包括但不限于 FP32、FP16、BF16、INT8 等精度下的理论算力（TFLOPS）与实测算力。
b） 应记录显存带宽与缓存性能；
定义：GPU/NPU 显存带宽（GB/s）及其饱和度；
L2/L3 缓存容量、命中率；
内存、显存间数据搬运效率（PCIe/NVLink 带宽利用率）。
c） 应记录支持AI加速卡数量级；
定义：包括但不限于单集群/单任务最大支持的AI加速卡数量（如256/1024/万卡级）。

7.1.2 高性能网络性能

a） 应记录节点互联带宽与延迟（测试 RDMA 协议（InfiniBand / RoCEv2）并提供底层通信性能基准）；
定义：GPU/NPU 节点间的通信带宽（GB/s）、P2P 延迟（μs）、PPS（每秒包数）。
b） 应记录节点互联带宽与延迟；
定义：测试 AllReduce、AllGather、AllToAll、Broadcast 等集合通信操作的吞吐与延迟。
c） 应记录集群线性加速比（Linearity）；
定义：多机多卡扩展时，总吞吐随设备数量增长的比例。

7.1.3 存储与 IO 性能

a） 应记录存储吞吐与时延；
定义：大文件写入带宽（Checkpoint 持久化，单位 GB/s）；
小文件读取 IOPS（用于数据集加载，尤其是图像、文本样本）；
元数据操作性能（open/stat/read/close 等调用延迟）。
b） 应记录数据加速性能；
定义：数据预处理速度（Samples/sec）；
分布式 DataLoader QPS（每秒加载样本数）；
数据解码、增强 pipeline 的 CPU/GPU 利用率平衡。

7.2 训练平台性能

7.2.1 分布式训练性能

a） 应记录算力有效性（MFU / HFU）；

定义：硬件利用效率（Model FLOPs Utilization 或 Hardware FLOPs Utilization），即模型训练中实际使用的 FLOPs 占理论峰值 FLOPs 的比例。

b）应记录任务启动时延；

定义：从用户提交训练任务到首个训练步骤开始执行的时间。

c）应记录断点恢复时延；

定义：故障发生后，自动迁移并恢复至最近检查点（Checkpoint）所需时间。

7.2.2 异构与调度性能

a）应记录镜像加速性能；

定义：镜像拉取时间（冷启动 vs 热缓存）；

镜像分发效率（跨节点并发拉取速率）；

是否支持容器镜像去重、分层缓存、P2P 分发（如 Dragonfly）。

7.3 推理服务平台性能

7.3.1 推理核心指标

a）应记录推理首字延迟（TTFT, Time To First Token）；

定义：从接收请求到生成第一个输出 token 的时间。

b）应记录推理吞吐量（TPS, Tokens Per Second）；

定义：单位时间内模型生成的 token 数量（整体系统级 TPS）。

c）应记录端到端延迟（E2E Latency）；

定义：从请求到达 API 网关到完整响应返回客户端的总耗时。

d）应记录并发支持能力（QPS）；

定义：在满足 P99 延迟前提下，系统能稳定承载的最大请求数/秒。

e）最大并发连接数；

定义： 在RT、ttft和成功率等约束的前提下，推理服务支持的最大的并发连接数。

f）应记录 Token 输出间隔时间（TPOT, Time Per Output Token）；

定义：在流式生成（Streaming Generation）模式下，连续两个输出 token 之间的平均时间间隔（单位：毫秒/token）。

7.4 智算平台其他产品

7.4.1 向量库核心性能

a）应记录向量数据库性能；

定义：数据加载速度（Vectors/sec）；

查询 QPS（每秒相似度搜索次数）；

召回成功率（Recall@K，通常 K=10/100）；

查询延迟（P99，毫秒级）。

b）应记录索引构建效率；

定义：完成大规模向量索引训练与构建所需时间。

c）应记录混合检索能力；

定义：结合关键词过滤与向量相似度的复合查询性能。

7.4.2 服务网关性能

a） 应记录网关层最大并发；

定义：API 网关可维持的最大并发连接数及对应的 P99 响应时间。

8 分级方法

8.1 能力等级定义

安全可信智能云平台能力等级定义如下：

等级	名称	定义
1	基础接入级	提供基础的计算基础设施与简单的人工智能推理能力，支持单卡部署与基础数据管理，满足最低限度的云上 AI 资源使用需求。
2	通用推理级	提供多卡异构资源纳管与增强推理能力，支持主流深度学习框架与基础MLOps流程，具备资源隔离与基础监控。
3	增强训练级	提供百卡级分布式训练能力与完成MLOps工作流，支持小模型预训练或大模型微调能力，提供基础任务容错与调度能力。
4	成熟训推级	提供千卡级高效训练与高并发推理服务，具备完整MLOps流程与高级模型治理能力，支持复杂任务调度与自动化运维。
5	训推卓越级	提供万卡级极致性能的全维度大模型推理和训练服务，具备极致算力效率、自治运维、高级安全治理与异构资源智能调度能力，支持。

8.2 能力等级划分方法

标准采用“阶梯式准入”的等级评定逻辑，将评价维度划分为安全可信要求、平台基础能力要求及性能要求三个模块。能力等级由低至高分为1至5级，每一等级的评定均遵循“前提合规、功能达标、性能验证”的原则，即：在满足安全前提下，必须达到对应等级所设定的必选项达标率及性能要求。

8.2.1 第一部分：安全可信要求

安全可信要求是开展能力评定的基础，本模块所有必选项（应支持条款）均须满足，是开展等级评定的前提。

8.2.2 第二部分：平台基础能力要求

平台基础能力是等级划分的重要依据。其中，6.1 a）、b）、c）是各能力等级均须满足的能力项；6.1 d）是能力等级3-5须满足的能力项，此外，6.2-6.4中必选项（应支持条款）达成率与等级划分关系见下表。

等级	1	2	3	4	5
6.2-6.4 中必选项（应支持条款）达成率	50%及以上	65%及以上	80%及以上	90%及以上	95%及以上

8.2.3 第三部分：性能要求

性能指标要求与能力等级划分关系见下表。

章节	用例项（含具体指标）	等级 1	等级 2	等级 3	等级 4	等级 5
7.1 高性能 IaaS 基础设施	**7.1.1 算力性能**					
	a. 芯片基础算力指标 ：					
	1. FP32 理论/实测 TFLOPS	记录	记录	> 65%	> 75%	> 85%
	2. FP16 理论/实测 TFLOPS	记录	记录	> 70%	> 85%	> 90%
	3. BF16 理论/实测 TFLOPS	记录	记录	> 70%	> 85%	> 90%
	4. INT8 理论/实测 TFLOPS	记录	记录	> 80%	> 90%	> 95%
	b. 显存带宽与缓存性能 ：					
	1. GPU/NPU 显存带宽 (GB/s)	记录	记录	记录	记录	记录
	2. 显存带宽饱和度 (%)	记录	记录	> 60%	> 80%	> 90%
	3. L2/L3 缓存容量 (MB)	记录	记录	记录	记录	记录
	4. L2/L3 缓存命中率 (%)	记录	记录	> 60%	> 75%	> 85%
	5. PCIe/NVLink 带宽利用率 (%)	记录	记录	> 60%	> 80%	> 90%
	c. AI 加速卡数量级 ：					
	AI 加速卡数量级	记录	记录	256 级	1024 级	万级
	7.1.2 高性能网络性能					
	a. 节点互联底层通信 (RDMA) ：					
	1. 通信带宽 (GB/s)	记录	记录	记录	> 85%	> 90%
	2. P2P 延迟 (μs)	记录	记录	< 20 μs	< 5 μs	< 1.5 μs
	3. PPS (每秒包数)	记录	记录	记录	记录	记录
	b. 集合通信性能 ：					

	1. AllReduce/AllGather 吞吐与延迟	—	—	记录	> 75%	> 80%
	2. AllToAll/Broadcast 吞吐与延迟	—	—	记录	> 75%	> 80%
	c. 集群线性加速比 ：					
	1. Linearity (多机多卡扩展比例)	—	—	记录	> 75%	> 80%
	7.1.3 存储与 IO 性能					
	a. 存储吞吐与时延 ：					
	1. 大文件写入带宽 (GB/s)	记录	记录	≥2GB/s	≥10GB/s	≥50GB/s
	2. 小文件读取 IOPS	记录	记录	≥1 万	≥5 万	≥20 万
	3. 元数据延迟 (open/stat 等, ms)	记录	记录	≤5 ms	＜2 ms	＜1 ms
	b. 数据加速性能 ：					
	1. 数据预处理速度 (Samples/sec)	记录	记录	＞1000	> 5,000	> 20,000
	2. 分布式 DataLoader QPS	记录	记录	＞1000	> 5,000	> 20,000
	3. CPU/GPU 利用率平衡度	记录	记录	记录	动态平衡	高效流水无瓶颈
7.2 训练平台性能	**7.2.1 分布式训练性能**					
	a. 算力有效性 ：					
	1. 算力有效性 (MFU/HFU) ：	—	—	> 25%	> 35%	> 40%
	b. 任务启动时延 ：					
	1. 热缓存启动耗时 (min)	记录	记录	≤ 5 min	≤ 3 min	≤ 2 min
	2. 冷启动耗时 (min)	记录	记录	≤ 10 min	≤ 8 min	≤ 5 min
	c. 断点恢复时延 ：					
	1. 故障恢复总时延 (min)	—	—	记录	≤ 10 min	≤ 3 min
	7.2.2 异构与调度性能					
	a. 镜像加速性能 ：					
	1. 镜像拉取/分发时间 (min)	记录	记录	记录	≤ 5 min	≤ 2 min
	2. P2P 分发效率	记录	记录	记录	支持 P2P 并	高效 P2P

					发	秒级
	3. 镜像去重与分层支持	记录	记录	支持去重	支持分层缓存	极致去重分发
	7.3.1 推理核心指标					
	a. 推理首字延迟 (TTFT) ：					
	1. 在线场景 TTFT (ms)	记录	≤ 2000 ms	≤ 1000 ms	≤ 800 ms	≤ 500 ms
	2. Agent/长 Prompt 场景 TTFT (ms)	记录	记录	≤ 1500 ms	≤ 1000 ms	≤ 800 ms
	b. 推理吞吐量 (TPS) ：					
	1. 系统级 TPS	记录	记录	记录	记录	记录
	2. 单卡 TPS (tokens/s/GPU)	记录	记录	≥ 80	≥ 120	≥ 150
7.3 推理服务平台性能	**c. 端到端延迟 (E2E Latency) ：**					
	1. 高优先级 P99 延迟 (s)	记录	≤ 5 s	≤ 3 s	≤ 2.5 s	≤ 2 s
	2. 批量异步 P99 延迟 (s)	记录	记录	≤ 20 s	≤ 15 s	≤ 10 s
	d. 并发支持能力 (QPS) ：					
	1. P99 约束下的最大 QPS	记录	记录	≥ 10	≥ 15	≥ 20
	e. 最大并发连接数 ：					
	1. 满足 RT、ttft/成功率约束连接数	记录	> 500	> 2,000	> 5,000	> 10,000
	f. Token 输出间隔 (TPOT) ：					
	1. 流式输出平均间隔 (ms/token)	记录	≤ 100 ms	≤ 50 ms	≤ 30 ms	≤ 15 ms
	7.4.1 向量库核心性能					
	a. 向量数据库性能 ：					
7.4 智算平台其他性能	1. 数据加载速度 (Vectors/sec)	记录	记录	记录	记录	记录
	2. 查询 QPS	记录	记录	> 1,000	> 5,000	> 20,000
	3. Recall@10/100 召回率	记录	记录	> 95%	> 98%	> 99%
	4. 查询 P99 延迟 (ms)	记录	记录	< 20 ms	< 10 ms	< 5 ms
	b. 索引构建效率 ：					

	1. 大规模索引构建时间 (min)	记录	记录	记录	< 10 min	< 2 min
	c. 混合检索能力 ：					
	1. 复合查询(关键词+向量)性能	记录	记录	记录	记录	满足极低延迟
	7.4.2 服务网关性能					
	a. 服务网关性能 ：					
	1. 最大并发连接数 (万级/百万级)	记录	记录	10 万级	50 万级	百万级
	2. 网关侧 P99 延迟 (ms)	记录	记录	< 200 ms	< 150 ms	≤ 100 ms