

中华人民共和国国家标准

GB/T 46350—2025

信息技术 云计算 智能云服务通用要求

Information technology—Cloud computing—General requirements of
AI cloud service

2025-10-05 发布

2026-02-01 实施

国家市场监督管理总局 发布
国家标准化管理委员会

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 缩略语 1

5 能力框架 2

6 基础设施要求 3

 6.1 资源 3

 6.2 资源管理 4

 6.3 加速 5

 6.4 任务管理和调度 6

7 模型开发要求 6

 7.1 数据处理 6

 7.2 模型构建 8

 7.3 模型管理 9

 7.4 模型部署 11

 7.5 模型推理 11

 7.6 数据闭环 12

 7.7 大模型提示工程 12

 7.8 大模型指令调优 12

 7.9 MLOps 工作流 12

 7.10 LMOps 工作流 13

8 模型服务要求 13

 8.1 计算机视觉类 13

 8.2 语音类 13

 8.3 自然语言处理类 14

 8.4 多模态类 14

9 人工智能应用开发要求 14

 9.1 开发环境 14

 9.2 应用开发框架 14

 9.3 应用组件 14

 9.4 应用模板 14

 9.5 知识管理 15

 9.6 应用管理 15

参考文献 16

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。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：北京百度网讯科技有限公司、中国电子技术标准化研究院、华为云计算技术有限公司、浪潮云信息技术股份公司、天翼云科技有限公司、中移(苏州)软件技术有限公司、广州市品高软件股份有限公司、中国人民解放军国防科技大学、阿里云计算有限公司、深信服科技股份有限公司、北京华胜天成科技股份有限公司、中兴通讯股份有限公司、济南浪潮数据技术有限公司、中移雄安信息通信科技有限公司、中移(杭州)信息技术有限公司、北京天融信网络安全技术有限公司、湖南麒麟信安科技股份有限公司、京东科技信息技术有限公司、上海商汤智能科技有限公司、浪潮电子信息产业股份有限公司、中国移动通信有限公司研究院、北京银信长远科技股份有限公司、安超云软件有限公司、中国烟草总公司湖南省公司、电子科技大学、中国科学院信息工程研究所、杭州谐云科技有限公司、荣联科技集团股份有限公司、哈尔滨工业大学、清华大学、北京国信网联科技有限公司、甘肃同兴智能科技发展有限责任公司、中航材导航技术(北京)有限公司、广东粤电信息科技有限公司、深圳聚瑞云控科技有限公司、广东车卫士信息科技有限公司、深圳市易连汇通科技有限公司、宝德计算机系统股份有限公司、北京大道云行科技有限公司、北京厚普医药科技有限公司、广州赋安数字科技有限公司、深圳市五兴科技有限公司、海天地数码科技(北京)有限公司、桂林理工大学、江西云眼视界科技股份有限公司、北京金智达管理顾问有限公司、澳克多普有限公司、广脉科技股份有限公司、北京智慧易科技有限公司、四川汉唐云分布式存储技术有限公司、苏州明逸智库信息科技有限公司、广东博今信息技术集团有限公司。

本文件主要起草人：宋飞、陈行、贾超、王志鹏、杨丽蕴、玄凌博、谢永康、刘增志、查丽、赵华、张敏、郑佳佳、杨明、苗子聪、郭旻、徐军、林冬艺、黄继承、包涵、王意洁、吴涛、朱松、刘晨、魏胜平、梁钢、李响、亓开元、张坤、姚飞、宁溯、张风雷、喻涵、张朝璐、雷晓锋、杨鹏举、邓旺波、刘辰、韩冬、吴庚、代文斌、鲁璐、叶丰华、范亚梅、毛浣渠、季寒德、鞠万里、郭晓、许刚、邹瞰、李益文、罗光春、陈爱国、边鹏旭、王翔宇、王向东、王静、刘奇旭、吴晓清、谭儒、李伟泽、徐冠群、乔霖、陈志峰、王程安、王瑞敏、李幸、闫纪红、黄双喜、杨思维、高宇、曹建农、詹越、孙沛、王林军、胡周达、林禧桐、郭文艺、曾俭、张云霞、刘爱贵、张芳、梁华、赵亚云、董广明、刘刚、罗军、王文迪、谢晓兰、唐涛、赵本金、王巍、赵淑飞、关涛、聂琨林、王亮、彭黎。

信息技术 云计算 智能云服务通用要求

1 范围

本文件确立了智能云服务能力框架,规定了智能云服务基础设施、模型开发、模型服务、人工智能应用开发等方面的要求。

本文件适用于指导智能云服务的设计、开发、部署、使用和测评。

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 32400—2015 信息技术 云计算 概览与词汇

3 术语和定义

GB/T 32400—2015 界定的以及下列术语和定义适用于本文件。

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 32400—2015,3.2.5,有修改]

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。

4 缩略语

下列缩略语适用于本文件。

AI:人工智能(Artificial Intelligence)

API:应用程序接口(Application Programming Interface)

ASIC:专用集成电路(Application Specific Integrated Circuit)

CCIX:缓存一致性的加速器互联(Cache Coherent Interconnect for Accelerators)

CPU:中央处理器(Central Processing Unit)

DPO:直接偏好优化(Direct Preference Optimization)

FPGA:现场可编程门阵列(Field Programmable Gate Array)

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)

5 能力框架

智能云服务通过云服务方式为用户提供 AI 算力资源和人工智能能力,支持人工智能应用,帮助用户高效、便捷、低成本获取和使用人工智能云服务。智能云服务能力框架见图 1,包括基础设施、模型开发、模型服务、人工智能应用开发等能力,智能云服务提供其中一种或多种组合能力,各层功能如下:

- a) 基础设施:提供面向人工智能的基础设施功能,进行创建和管理人工智能算力集群,提交和调度任务,基于加速库使用异构计算资源(包括裸金属、虚拟机、容器等),支撑人工智能负载运行;
- b) 模型开发:提供模型开发工具链,包括数据处理、模型构建、模型管理、模型部署、模型推理、数据闭环等模型开发通用功能,和提示工程、指令调优等大模型开发功能,以及 MLOps、LMOps 工作流,支撑人工智能模型在云端的训练、优化、管理和部署;
- c) 模型服务:提供多种模型的云服务功能,包括计算机视觉类、语音类、自然语言处理类、多模态类等,满足用户在云端使用预训练好的模型;
- d) 人工智能应用开发:提供 AI 应用开发工具链,包括开发环境、应用开发框架、应用组件、应用模板、知识管理、应用管理等,降低人工智能应用开发门槛,提升人工智能应用开发效率,支撑用户在具体业务场景中应用人工智能技术。

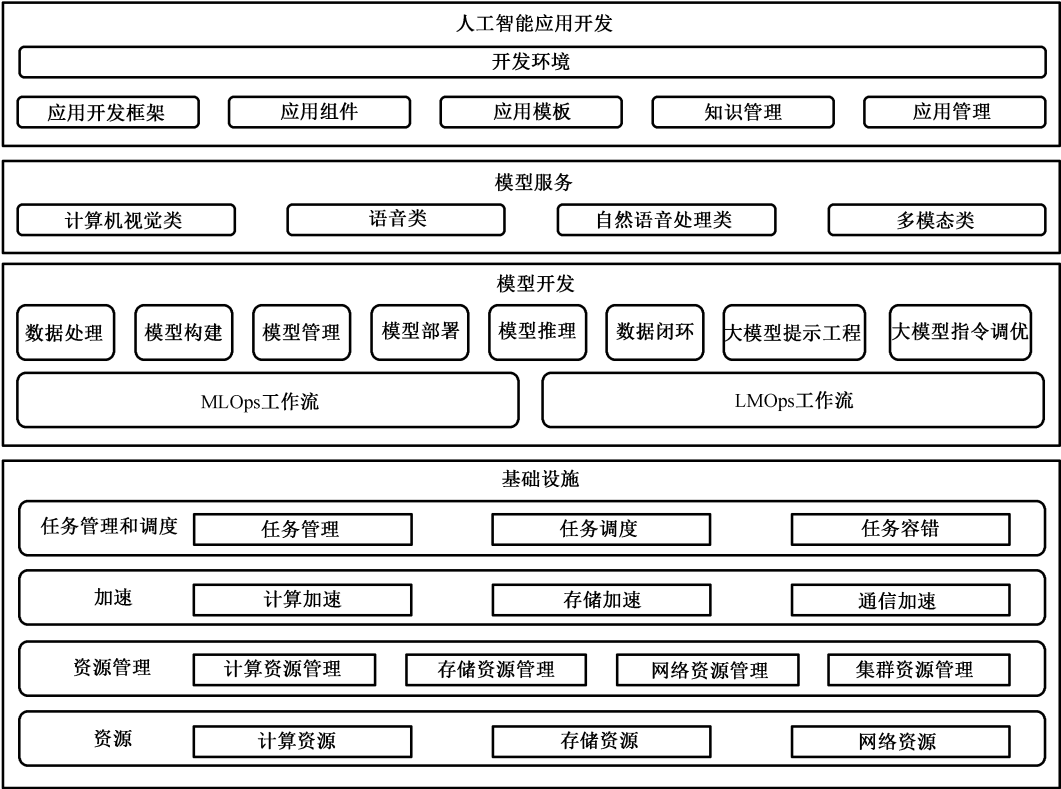


图 1 智能云服务能力框架

6 基础设施要求

6.1 资源

6.1.1 计算资源

6.1.1.1 裸金属

裸金属要求包括：

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。

- a) 应支持不同类型的人工智能加速处理器，包括 GPU、ASIC、FPGA 等；
- b) 应支持不同数量的人工智能加速处理器；
- c) 应支持神经网络计算，如卷积计算、向量计算、标量计算等；
- d) 应支持人工智能加速处理器的多卡 P2P 通信；
- e) 应支持人工智能加速处理器间互联，如 CCIX、PCIe 等互联方式。

6.1.1.2 虚拟机

虚拟机要求包括：

- a) 应支持不同类型的人工智能加速处理器，包括 GPU、ASIC、FPGA 等；
- b) 应支持不同数量的人工智能加速处理器；
- c) 应支持神经网络计算，如卷积计算、向量计算、标量计算等；
- d) 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；

- e) 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；
- f) 应支持人工智能加速处理器的多卡 P2P 通信；
- g) 应支持人工智能加速处理器间互联,如 CCIX、PCIe 等互联方式。

6.1.1.3 容器

容器要求包括：

- a) 应支持不同类型的人工智能加速处理器,包括 GPU、ASIC、FPGA 等；
- b) 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；
- c) 应支持神经网络计算,如卷积计算、向量计算、标量计算等；
- d) 应支持人工智能加速处理器的多卡 P2P 通信；
- e) 应支持单卡和多卡人工智能加速处理器共享和隔离。

6.1.2 存储资源

存储资源要求包括：

- 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。

6.1.3 网络资源

网络资源要求包括：

- 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 (RoCE)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。

6.2 资源管理

6.2.1 计算资源管理

6.2.1.1 裸金属管理

裸金属管理要求包括：

- a) 应支持对裸金属的生命周期管理,如创建、关机、启动、重启、释放等；
- b) 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镜像。

6.2.1.2 虚拟机管理

虚拟机管理要求包括：

- a) 应支持对虚拟机的生命周期管理,如创建、关机、启动、重启、释放等；
- b) 应提供包含多种人工智能加速处理器驱动程序的虚拟机镜像；
- c) 应支持虚拟机配置变更,如调整挂载的显卡数量。

6.2.1.3 容器管理

容器管理要求包括：

- a) 应支持容器的生命周期管理,如创建、删除等;
- b) 应提供包含多种人工智能加速处理器驱动程序的容器镜像;
- c) 应支持容器配置变更,如调整挂载的显卡数量。

6.2.2 存储资源管理

存储资源管理要求包括:

- 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。

6.2.3 网络资源管理

网络资源管理应支持配置 RDMA 网络隔离。

6.2.4 集群资源管理

集群资源管理要求包括:

- a) 应支持集群生命周期管理,如集群创建、删除等;
- b) 应支持集群节点手动、自动扩缩容;
- c) 应支持多种计算资源集群类型,如裸金属集群、虚拟机集群、容器集群等。

6.3 加速

6.3.1 计算加速

计算加速要求包括:

- 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。

6.3.2 存储加速

存储加速要求包括:

- 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工具,减少检查点存储写入时间。

6.3.3 通信加速

通信加速要求包括:

- 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)。

6.4 任务管理和调度

6.4.1 任务管理

任务管理要求包括:

- 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。

6.4.2 任务调度

任务调度要求包括:

- 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。

6.4.3 任务容错

任务容错要求包括:

- 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。

7 模型开发要求

7.1 数据处理

7.1.1 数据接入

数据接入要求包括:

- 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) 宜支持周期性接入数据,如按照设置的时间间隔定期接入数据等;
- f) 宜支持接入加密数据。

7.1.2 数据预处理

数据预处理要求包括:

- 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。

7.1.3 数据标注

数据标注要求包括:

- 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。

7.1.4 数据管理

数据管理要求包括:

- 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 and 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。

7.1.5 数据分析

数据分析要求包括:

- 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。

7.1.6 数据增强

数据增强要求包括:

- 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。

7.2 模型构建

7.2.1 算法仓库

算法仓库要求包括:

- 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。

7.2.2 算法管理

算法管理要求包括:

- 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。

7.2.3 特征工程

特征工程要求包括:

- 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。

7.2.4 开发环境

开发环境要求包括:

- 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。

7.2.5 模型训练

模型训练要求包括：

- 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。

7.2.6 模型调优

模型调优要求包括：

- 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 Early Stopping 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。

7.3 模型管理

7.3.1 模型仓库

模型仓库要求包括：

- 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。

7.3.2 模型适配

模型适配要求包括:

- 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。

7.3.3 模型加速

模型加速要求包括:

- 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。

7.3.4 模型评估

模型评估要求包括:

- 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。

7.3.5 资产管理

模型相关资产管理要求包括：

- 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。

7.4 模型部署

7.4.1 云端部署

云端部署要求包括：

- 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。

7.4.2 云边端协同部署

云边端协同部署要求包括：

- a) 应支持边缘端的模型部署,如边缘服务器、边缘一体机及边缘智算盒等；
- b) 应支持部署云边协同的模型服务,如云端向边缘端、终端设备下发和更新模型服务；
- c) 应支持边缘端的推理数据及日志等记录回流云端；
- d) 应支持模型的多种异构芯片云端和边缘端部署,如 CPU、GPU、ASIC 等；
- e) 宜支持模型的多种操作系统云端和边缘端部署；
- f) 宜支持模型的多种异构硬件终端部署,如 CPU、GPU、FPGA 等；
- g) 宜支持模型的多种操作系统终端部署。

7.5 模型推理

模型推理要求包括：

- 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) 宜支持流式推理；
- f) 宜支持多模态模型推理；

- g) 宜支持分布式高可用推理；
- h) 宜支持离线批量推理。

7.6 数据闭环

数据闭环要求包括：

- 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。

7.7 大模型提示工程

大模型提示工程要求包括：

- a) 应具备提示模板库进行提示管理及可视化展示，如增删改查、存储检索和调用、分类管理等；
- b) 应支持提示评估，通过评估模型对不同提示的响应质量，持续反馈和迭代改进提示语；
- c) 应支持根据历史记录生成新的提示词；
- d) 应支持提示模板定制开发、管理及可视化；
- e) 应支持提示模板可解释，如添加注释、说明、示例等；
- f) 宜支持提示增强，如思维链提示等；
- g) 宜支持提示自动优化；
- h) 宜支持多模态提示设计，如文本、图像或音视频等；
- i) 宜支持个性化提示和专用提示库，如根据用户需求、行为特征等提供个性化提示；
- j) 宜支持提示模板的分享、评价。

7.8 大模型指令调优

大模型指令调优要求包括：

- a) 应具备指令库支持指令集管理，如指令的增删改查、存储检索等；
- b) 应支持对单轮格式的指令数据集进行模型效果调优；
- c) 应支持不同模型间指令格式的自动转换；
- d) 应支持指令调优过程中配置超参数，如学习率、BatchSize 等；
- e) 应支持指令集对应的评估集管理，如评估数据、指标和指标计算脚本等；
- f) 宜支持对指令数据集的增强，如指令数据清洗、自动扩增等；
- g) 宜支持按应用场景分类的多种预置指令数据集；
- h) 宜支持对多轮格式的指令数据集进行模型效果调优；
- i) 宜支持多模态的指令数据集。

7.9 MLOps 工作流

MLOps 工作流要求包括：

- a) 应支持工作流的生命周期管理，如创建、停止、删除等；
- b) 应支持多种工作流编排方式，如可视化编排、代码编排等；

- c) 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;
- d) 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;
- e) 宜支持自定义工作流算子;
- f) 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;
- g) 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;
- h) 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;
- i) 宜提供自动化工作流模板。

7.10 LMOps 工作流

LMOps 工作流要求包括:

- a) 应符合 7.9 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(LoRA)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。

8 模型服务要求

8.1 计算机视觉类

计算机视觉类模型服务能力要求包括:

- a) 应提供多种计算机视觉类模型服务,如图像、视频、文字、人脸的识别、生成或分析等;
- b) 宜支持处理多种内容形态,如图片、视频文件、视频流等;
- c) 宜支持多种服务形式,如在线、离线等;
- d) 宜支持自定义计算机视觉类模型,定制调优模型效果,如自定义图像检索库、自定义人脸库、自定义细粒度图像识别标签等;
- e) 宜提供包括规范汉字在内的不少于 5 种语言文字的计算机视觉类模型服务。

8.2 语音类

语音类模型服务能力要求包括:

- a) 应提供多种语音类模型服务,如语音识别、语音合成或语音转写等;
- b) 宜提供语音唤醒服务;
- c) 宜提供语音实时翻译服务;
- d) 宜支持处理多种内容形态,如音频文件、音频流等;
- e) 宜支持多种服务形式,如在线、离线等;
- f) 宜支持自定义语音类模型,定制调优模型效果,如自训练语音模型、自定义语音合成的语速/音调/音量等;
- g) 宜提供包括规范汉字在内的不少于 2 种语言文字的语音类模型服务。

8.3 自然语言处理类

自然语言处理类模型服务能力要求包括：

- a) 应提供多种自然语言处理类模型服务,如语言理解、语言生成或机器翻译等;
- b) 应支持自定义语言类模型,定制调优模型效果,如自定义专有词表、自训练文本分类、定制情感倾向分析、定制翻译语料训练等;
- c) 宜支持多种内容形态的处理,如关键词、文章、文本、文件等;
- d) 宜支持多种服务形式,如在线、离线等。

8.4 多模态类

多模态类模型服务能力要求包括：

- a) 应支持生成类模型的参数或词表配置,如提示模板、temperature、topK 等;
- b) 宜提供包括图像、语音、文本类等多种模态混合的模型服务,如文生图、文生视频、图生视频等服务。

9 人工智能应用开发要求

9.1 开发环境

开发环境要求包括：

- a) 应支持开发环境的镜像,如应用开发框架、应用组件、应用配置模板等;
- b) 宜支持云上 IDE,具备不少于 2 种代码编辑器,并提供内置的开发镜像。

9.2 应用开发框架

应用开发框架要求包括：

- a) 应支持记忆存储、智能体编排、多轮对话、函数调用等能力;
- b) 应兼容多种开发工具,如 langchain、PromptFlow、LlamaIndex 等;
- c) 应支持多种应用配置,如配置描述、角色指令、组件、知识库、推荐问等;
- d) 应提供应用的调试工具;
- e) 应提供产品文档和演示。

9.3 应用组件

应用组件要求包括：

- a) 应提供多种大模型类基础组件,如问题改写、问题拆解、阅读理解、代码解释器等;
- b) 应提供多种传统模型类基础组件,如文档解析、文档切片、文字识别、图像识别等;
- c) 应提供多种 API 服务类组件,如搜索、天气查询、股票查询等;
- d) 应提供多种云组件,如对象存储、向量数据库、向量检索等;
- e) 应提供检索增强生成(RAG)组件;
- f) 应提供应用组件的调试工具;
- g) 宜支持应用组件的可视化调试。

9.4 应用模板

应用模板要求包括：

- a) 应提供自定义的应用模板;

- b) 应支持依据应用开发需求自动生成模板；
- c) 应提供多种应用模板示例，如客服助手、法律助手、智能问数等；
- d) 应支持复制应用模板。

9.5 知识管理

知识管理要求包括：

- a) 应支持知识库中的知识管理，如知识文档的上传、更新、删除、复制等；
- b) 应支持多种格式的文档管理，如 txt、ofd、docx、xls、pdf 等；
- c) 应支持多种知识抽取方式，如文档结构解析、文档切片、文字识别等；
- d) 应支持多种知识检索方式，如向量检索、倒排索引、全文索引等；
- e) 应支持知识增强，使用大模型对知识进行总结。

9.6 应用管理

应用管理要求包括：

- a) 应支持应用的全生命周期管理，包括创建、调试、发布、下线等；
- b) 应为发布后应用提供统一的接口访问地址和 token 鉴权方式；
- c) 应支持应用的可视化调试和发布；
- d) 宜支持应用的运行数据统计，如访问量、时延、用户数等；
- e) 宜提供应用市场；
- f) 宜支持应用版本管理。

参 考 文 献

- [1] GB/T 37736—2019 信息技术 云计算 云资源监控通用要求
 - [2] GB/T 37938—2019 信息技术 云资源监控指标体系
 - [3] GB/T 41867—2022 信息技术 人工智能 术语
-

