ICS 35.240
CCS L 70

中华人民共和国国家标准

GB/T 45288.3—2025

人工智能　大模型
第3部分:服务能力成熟度评估

Artificial intelligence—Large-scale model—
Part 3:Service capability maturity assessment

2025-01-24 发布　　2025-01-24 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

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。

本文件是GB/T 45288《人工智能　大模型》的第3部分，GB/T 45288已经发布了以下部分：

——第1部分：通用要求；

——第2部分：评测指标与方法；

——第3部分：服务能力成熟度评估。

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：中国电子技术标准化研究院、华为技术有限公司、浪潮云信息技术股份公司、清华大学、华为云计算技术有限公司、中国科学院自动化研究所、北京百度网讯科技有限公司、深圳市腾讯计算机系统有限公司、菲特(天津)检测技术有限公司、北京奇虎科技有限公司、北京航空航天大学、国能信息技术有限公司、麒麟合盛网络技术股份有限公司、上海市人工智能行业协会、上海燧原科技股份有限公司、阿里云计算有限公司、平头哥(上海)半导体技术有限公司、上海计算机软件技术开发中心、浙江大华技术股份有限公司、青岛海信电子技术服务有限公司、上海人工智能研究院有限公司、南方电网人工智能科技有限公司、航天信息股份有限公司、广东电网有限责任公司、北京大学长沙计算与数字经济研究院、北京大学、中国科学院软件研究所、蚂蚁科技集团股份有限公司、中国移动通信集团有限公司、马上消费金融股份有限公司、深圳云天励飞技术股份有限公司、深圳思谋信息科技有限公司、北京格灵深瞳信息技术股份有限公司、中国南方电网有限责任公司超高压输电公司、北京软件产品质量检测检验中心有限公司、中国电力科学研究院有限公司、上海文鳐信息科技有限公司、浪潮软件科技有限公司、浪潮电子信息产业股份有限公司、浪潮软件集团有限公司、中电科大数据研究院有限公司、上海商汤智能科技有限公司、中国电信集团有限公司、科大讯飞股份有限公司、中国电信股份有限公司北京研究院、中移(苏州)软件技术有限公司、中国科学院新疆理化技术研究所、杭州海康威视数字技术股份有限公司、上海文鳐信息科技有限公司、西北工业大学、云知声智能科技股份有限公司、北京工业大学、北京智芯微电子科技有限公司。

本文件主要起草人：徐洋、马珊珊、于超、王莞尔、董建、陶建华、曹晓琦、鲍薇、黄现翠、马骋昊、郑佳佳、郑子木、朱贵波、王金桥、刘静、汪群博、杨旭、马同森、靳伟、刘海涛、曹彬、张向征、任海峰、刘祥龙、刘艾杉、张旭、陈曦、赵春昊、蒋燕、梅敬青、彭骏涛、张艺伯、陈敏刚、孔维生、刘微、刘常昱、宋海涛、任正国、邵彦宁、刘佳宁、周昊、杨超、孟令中、孙曦、金镝、李宽、王志芳、吕江波、胡全一、王宁、王志刚、孔昊、莫文昊、仲凯韬、王珂琛、刘璐、张天霖、蒋慧、刘敬谦、刘威辰、高建清、孟建、舒珏淋、商兴宇、李旭东、杨雅婷、钟凯伦、仲凯韬、张涛、梁家恩、刘峥、郑哲、武姗姗。

引　言

大模型已成为人工智能发展的重要技术手段，在引领产业变革中发挥重要作用，国内外人工智能相关机构相继研究开发百余种大模型产品和评测榜单，导致用户难以有效评价人工智能产品的技术水平和服务能力。GB/T 45288 旨在规定通用大模型的技术要求、评测指标和服务能力，拟由五个部分构成。

——第 1 部分：通用要求。目的在于确立大模型的参考架构，规定通用技术要求。

——第 2 部分：评测指标与方法。目的在于确立大模型的评测指标，描述评测方法。

——第 3 部分：服务能力成熟度评估。目的在于给出大模型服务能力成熟度等级及评估方法。

——第 4 部分：计算机视觉大模型。目的在于定义计算机视觉大模型的概念和功能，规定技术要求和测试方法。

——第 5 部分：多模态大模型。目的在于定义多模态大模型的概念和功能，规定技术要求和测试方法。

人工智能　大模型
第3部分:服务能力成熟度评估

1　范围

本文件给出了大模型服务能力框架和评估指标,描述了大模型服务能力成熟度等级划分及评估方法。

本文件适用于服务提供方和需求方对大模型平台、模型定制及推理运营服务的能力进行全面评估,也适用于指导大模型服务能力的规划、设计和实现。

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 42018—2022　信息技术　人工智能　平台计算资源规范

3　术语和定义

下列术语和定义适用于本文件。

3.1

大模型平台　large-scale model platform

为开发或使用大模型提供各类资源的软硬件平台。

注:大模型平台不包含大模型。

3.2

大模型服务　large-scale model service

开发、应用大模型及大模型系统的服务,以及以此为手段提供支持需方业务活动的服务。

注:大模型系统是大模型与大模型平台的集成,是与大模型服务相关的活动、过程等的集合。

3.3

工具链　toolchain

用于支撑大模型开发、定制及应用的软件集合。

3.4

指令　instruct

由大模型输入和输出信号构成的二元组。

注:如自然语言处理的大模型,指令以提问文本和回答文本成对呈现。

4　缩略语

下列缩略语适用于本文件。

AI:人工智能(Artificial Intelligence)

API:应用程序编程接口(Application Programming Interface)
CPU:中央处理器(Central Processing Unit)
GE:千兆位以太网(Gigabit Ethernet)
GPU:图形处理器(Graphic Process Unit)
NFS:网络文件系统(Network File System)
PEFT:参数高效微调(Parameter-Efficient Fine-Tuning)
PFLOPS:每秒千万亿次浮点运算(Peta Floating Point Operations Per Second)
POSIX:可移植操作系统接口(Portable Operating System Interface)
REST:强化自训练(Reinforced Self-Training)
RLHF:基于人类反馈的强化学习(Reinforcement Learning from Human Feedback)
SDK:软件开发工具包(Software Development Kit)
S3:简单存储服务(Simple Storage Service)

5 概述

5.1 大模型服务类型

基于大模型服务的功能特性、流程阶段、服务所面向的用户需求及技术实现的不同层面等维度,大模型服务分为大模型平台服务、大模型开发定制服务、大模型推理及运营服务,见表1。

表1 大模型服务类型说明

序号	服务类型	供方	需方	说明
1	大模型平台服务	基础设施提供者	模型提供者、应用服务者	供方利用软硬件基础设施,为需方提供大模型开发定制的技术环境,包括硬件资源、软件及工具链、平台综合性能等,不包含大模型本身
2	大模型开发定制服务	模型提供者、应用服务者	模型应用集成者、应用消费者	通过一系列的活动(模型设计、开发、微调等),向需方交付符合环境限制及性能要求的模型
3	大模型推理及运营服务	模型应用服务者、应用集成者	模型应用消费者(如终端用户、业务系统开发者等)	供方利用大模型处理需方的特定输入,交付推理结果、提供技术支持及开展应用活动,以满足需方在性能、效率等方面的要求

5.2 服务能力框架

大模型服务能力框架能力域包括大模型平台能力域、大模型开发定制能力域、大模型推理及运营能力域,见表2。

a) 大模型平台能力域:为模型开发定制、模型推理及运营等提供计算资源、软硬件基础设施平台的能力,包括硬件、软件及工具链、平台综合3个能力子域。
b) 大模型开发定制能力域:提供大模型设计、开发及定制服务的能力,包括数据资源、模型生产定制2个能力子域。
c) 大模型推理及运营能力域:基于大模型及其系统提供推理或运营服务的能力,包括模型推理、平台应用、运营赋能3个能力子域。

表 2　能力域和能力子项

能力主域	能力子域	能力子项
大模型平台	硬件	计算资源
		网络资源
		存储资源
	软件及工具链	训练工具
		数据处理工具
		微调工具
		模型压缩工具
		监测及分析工具
		模型评估工具
	平台综合	兼容性
		可靠性
		分布式并行优化
		易用性
大模型开发定制	数据资源	数据接入
		数据处理
	模型生产定制	模型设计
		模型训练
		模型微调
		模型压缩
		模型验证
		模型部署
大模型推理及运营	模型推理	推理服务部署
		推理服务效率
		推理服务稳定性
	平台应用	行业应用
		交互应用
	运营赋能	数据回流
		生态建设

6　评估指标

6.1　大模型平台

6.1.1　硬件

6.1.1.1　计算资源

大模型平台的计算资源指标包含：

a)　用于训练任务的物理计算资源，应符合 GB/T 42018—2022 中 6.1.1.1 b)～6.1.1.1 u)和 6.1.2.1 的规定；

b) 用于推理任务的物理计算资源，应符合 GB/T 42018—2022 中 6.1.1.2 b)～6.1.1.2 k)和 6.1.2.2 的规定；
c) 具备预处理模块，实现大模型推理硬件加速，支持并行解码(如多路视频、图像或音频等)；
d) 具备算子库，能构造大模型，实现自然语言处理、计算机视觉、多模态生成功能；
e) 用于训练任务的计算设备算力理论值求和，百亿模型不低于 50 PFLOPS(FP16 精度)；
f) 平台提供多台计算设备组成集群用于训练任务时，用于百亿参数模型训练任务的集群线性度不低于 80%；

注：集群线性度的计算方法为：组成集群的多台计算设备总吞吐率(测试集群中 2 个～3 个代表性计算设备的吞吐，求平均)与各计算设备单独工作时吞吐率之和(测试 2 个～3 个代表性计算设备的吞吐，求平均)的比值。

g) 具备训练 128 K 序列长度的模型；
h) 用于推理的计算资源，具备 128 K 序列长度的推理任务；
i) 用于推理的计算资源，具备多模态模型的推理；
j) 用于训练任务的计算设备算力理论值求和，千亿、万亿模型不低于 1 000 PFLOPS(FP16 精度)；
k) 平台提供多台计算设备组成集群用于训练任务时，用于千亿、万亿参数模型训练任务的集群线性度不低于 70%。

6.1.1.2 网络资源

大模型平台的网络资源指标包含：

a) 用于大模型训练的参数交互网络具备 200 GE 或 400 GE 或 800 GE 总体带宽；
b) 具备多租户访问隔离能力，可使用安全传输协议；
c) 具备链路聚合能力；
d) 具备跨设备链路聚合能力，提高链路可靠性；
e) 实现运行计算任务时，能通过图形界面、面板等方式实现计算设备间通信关系的可直观显示；
f) 用于大模型训练的参数交互网络，具备网络负载均衡。

6.1.1.3 存储资源

大模型平台的存储资源指标包含：

a) 具备多种存储协议，如 NFS、S3、POSIX 等；
b) 具备多种存储媒体，支持数据流通；
c) 具备多层级数据安全保护能力，如防病毒、数据备份、数据加解密等；
d) 能承载动态混合负载，百亿参数模型训练平台存储带宽不小于 1 TB/s；
e) 具备数据副本或纠删码冗余策略、用户认证域、业务网段隔离等数据安全保护能力；
f) 具备人工智能加速卡直通存储能力；
g) 千亿、万亿参数模型训练平台具备元数据检索，元数据如数据所有者、处理者或格式等；
h) 千亿、万亿参数模型训练平台具备千亿个文件的存储和管理；
i) 千亿、万亿参数模型训练平台存储读写带宽不小于 10 TB/s。

6.1.2 软件及工具链

6.1.2.1 训练工具

大模型平台的训练工具指标包含：

a) 提供损失函数、优化器等函数调用或组件；
b) 实现训练过程监控和日志记录，跟踪训练或微调过程；
c) 实现训练数据安全及隐私保护机制；

d) 具备混合精度训练，包含自动混合精度、手动混合精度；
e) 具备自动并行模式，能根据模型和集群的情况，配置模型中专家模块的并行执行策略。

6.1.2.2 数据处理工具

大模型平台的数据处理工具包含以下指标。
a) 提供指令处理工具，具备数据处理流程的定制或可直观编排。
b) 数据清洗功能：
   1) 提供数据转换算子，包含表情符、网页标签、特殊符号、无字符、间隔符等字符或字符串转换；
   2) 实现数据过滤功能，包含乱码文本、汉字比率过滤、参考文献过滤等；
   3) 实现去重、脱敏、敏感词过滤、正则替换等功能；
   4) 具备自定义数据清洗算子；
   5) 具备数据集备份；
   6) 具备数据拆分。
c) 数据标注功能：
   1) 具备文本、视频、图像的标注；
   2) 具备人工、团队标注；
   3) 具备标注结果的核验和重标注；
   4) 具备自动化标注。
d) 实现训练数据处理加速机制，具备数据加载和预处理多步并行流水线。
e) 实现多级缓存机制，加速数据处理过程。

6.1.2.3 微调工具

大模型平台的微调工具指标包含：
a) 提供带有预置权重的预训练模型，支持自动加载模型；
b) 具备配置微调超参，如学习率、批次大小、迭代次数等；
c) 具备 PEFT 机制，如 LoRA、QLoRA、P-Tuning 等；
d) 具备低代码微调；
e) 实现 RLHF 机制。

6.1.2.4 模型压缩工具

大模型平台的模型压缩工具指标包含：
a) 具备百亿级参数量模型的压缩；
b) 提供加速优化选项配置接口，如剪枝、量化、蒸馏 API，配置加速选项；
c) 具备至少 2 种量化精度；
d) 具备 2 种以上量化方案的选择或组合，如二值化、线性量化、对数量化、训练后量化、量化感知训练和动态量化；
e) 具备 2 种以上蒸馏方案，如软蒸馏、硬蒸馏等；
f) 具备千亿、万亿级参数量模型的压缩。

6.1.2.5 监测及分析工具

大模型平台的性能监测及分析工具包含以下指标。
a) 具备计算环境的资源监测及状态查询，包括：

1） 计算设备部件状态监测和查询，如处理器、内存、网络等资源利用率监测和统计；
2） 监测模型训练任务执行过程，计算资源用量变化；
3） 计算资源故障检测及自动恢复；
4） 网络资源监测及状态查询，如用于训练或推理的网络的即时流量、延迟等其他网络设备运行状态指标；
5） 存储资源监测及状态查询，如用于训练或推理的内存用量、带宽。

b） 具备模型开发过程的性能监测，包括：
1） 计算过程溢出检测（如寄存器、内存等）；
2） 监测模型训练过程中的性能指标变化，如准确率、损失函数等；
3） 集群性能可直观显示分析，包括集群迭代间隙、计算时间、通信时间、链路带宽、集群节点数据处理性能。

6.1.2.6 模型评估工具

大模型平台的模型评估工具指标包含：
a） 具备自动化的客观评估，能针对至少 1 种下游任务实施，并配备测试指标；
b） 具备模型可信赖评估，包含对抗样本和可解释性评估；
c） 具备主观评估，提供图形化界面，使评分者能根据输出质量给出评分或反馈。

6.1.3 平台综合

6.1.3.1 兼容性

大模型平台的兼容性指标包含：
a） 具备 REST 接口对接管理平台；
b） 能兼容不同版本的模型，实现加载、卸载等模型使用机制；
c） 兼容 2 种以上深度学习框架；
d） 兼容 2 种以上大模型分布式计算加速库，跟随版本支持最新特性，如 DeepSpeed、MegatronLM、AscendSpeed 等；
e） 兼容 2 种以上模型算法库，如 OpenMMLab、Hugging Face 等；
f） 具有向量数据库。

6.1.3.2 可靠性

大模型平台的可靠性保障指标包含：
a） 具备节点或通信不可用时，重调度新节点及配置集合通信，继续计算任务；
b） 具备集群训练任务的断点续训，具备自动检测、隔离故障资源；保存故障时的断点信息（如 checkpoint），从故障断点恢复训练；
c） 具备故障报告通道，检视通道和控制通道解耦，容错控制仅依赖所检视资源的状态，避免多控制源；
d） 具备弹性调整资源：具备能够根据工作负载的变化，动态调整计算资源，如 CPU、GPU、内存等；
e） 百亿参数大模型负载的训练任务无故障或中断运行时间超过 7 d；
f） 千亿参数、万亿参数大模型负载的训练任务无故障或中断运行时间超过 24 h；
g） 大模型的训练任务故障（如 AI 加速处理器故障、内存故障、电源故障）恢复时间不超过 1 h。

6.1.3.3 分布式并行优化

大模型平台的分布式并行优化指标包含：

a) 提供异构物理设备、虚拟设备用于训练或推理，具备虚拟化等云化资源池的能力；

b) 实现分布式并行优化组件，包括分布式并行训练或推理框架、深度学习模型编译器、集合通信库、虚拟化与调度组件；

c) 具备单机多卡、多机多卡环境下的多维混合分布式并行训练，包括但不限于数据并行、模型并行、张量并行、流水线并行、优化器并行等；

d) 具备单机多卡、多机多卡环境下的分布式推理。

6.1.3.4 易用性

大模型平台的易用性保障包含以下内容。

a) 平台至少具备 2 种使用方式，如命令行、API、SDK 或网络界面。

b) 能提供可直观显示工具，提供以下功能：

1) 模型迭代信息溯源；

2) 训练等任务状态检查。

c) 提供数据增强功能接口，通过接口配置数据增强策略。

d) 提供推理工作流，具备从数据处理、训练、部署、推理过程的流程化处理。

e) 提供可直观显示的拖拽布局编程服务，组合各种数据源、组件、算法、模型和评估模块。

f) 具备下列可直观显示的图形表达方式，包括：

1) 至少 2 种图形，如等值线图、地形图、3D 图、表格、指示灯、曲线图、热力图等；

2) 模型性能指标对比分析，如使用条形图或曲线图对比验证集准确率、损失函数值等；

3) 能为不同的数据类型和格式提供特定的可直观显示方式，如时间序列数据线图、分类数据饼图、空间数据地图等；

4) 具备对特定预测结果的解释，解释要素包含如特征重要性、激活图或热力图等。

6.2 大模型开发定制

6.2.1 数据资源

6.2.1.1 数据接入

大模型开发定制的数据接入指标包含：

a) 具备从不同外部系统接入数据的能力；

b) 提供数据集版本管理功能，实现版本标识、数据变更记录；

c) 能记录数据过程，包含数据特征、标签、数据预处理方法等；

d) 具备多源异构数据存储，如结构化数据与非结构化文本、图像、语音等数据的存储。

6.2.1.2 数据处理

大模型开发定制的数据处理包含以下指标。

a) 具备 3 种以上数据增强方式。

注：常见的数据增强方法包括但不限于：文本数据增强（如同义词替换、随机插入、随机删除、句子乱序），图像数据增强（如翻转、旋转、缩放、裁剪、平移、色彩变换），音频数据增强（如重采样、变声、音频切割、语速调整、噪音调整、音频拼接）。

b) 监督微调阶段：

1） 具备至少 2 种构建指令数据的方法，如基于大模型自动生成、人工标注；
2） 能根据任务需求，明确定义指令微调任务的标签信息；
3） 具备使用单轮和多轮对话指令数据；
4） 提供指令数据清洗方法，如去重等；
5） 提供数据处理加速功能。

c） 具备数据集处理后的发布（如平台用户工作空间，模型库等）。
d） 具备数据胶囊机制。

6.2.2 模型生产定制

6.2.2.1 模型设计

模型设计指标包含：

a） 具备模型架构的设计，包括对模型图结构设计、算子设计；
b） 具备模型架构搜索，包含模型骨架以及特定模块等；
c） 具备模型架构探索，实现机制缩小模型搜索空间（如启发式模型放缩策略）；
d） 能匹配模型结构，提供数据前、后处理功能模块的搜索，支持多目标的网络模型结构设计等。

6.2.2.2 模型训练

模型训练包含以下指标。

a） 具备多种模型切分策略，综合生成分布式并行训练策略。
b） 具备模型的版本控制和管理，确保在迭代训练中能够追溯和比较不同版本的模型性能。
c） 具备以下内存优化机制：
1） 重计算，减少并行计算过程中内存占用；
2） 优化器状态分组和参数分组；
3） 混合精度训练；
4） 梯度累积。

d） 具备利用超参数优化等算法，完成自动化的参数调整。
e） 具备将相同源及目的节点通信算子打包同时执行，避免多个单算子执行带来的额外开销。
f） 具备动态学习率调整策略，能根据模型的表现和收敛情况调整学习率。
g） 具备多方协作模型开发，包含工作空间协同、资产协同及管理。

6.2.2.3 模型微调

模型微调指标包含：

a） 具备多种微调方式，如适配器微调、指令微调、持续学习或 RLHF 等；
b） 具备模型微调过程显示，便于用户创建、查看、管理微调过程；
c） 具备全量参数微调和部分参数微调；
d） 具备在线微调，微调时平台自动化分配所需物理资源；
e） 具备管理和保护模型微调过程中产生的数据，如用户数据的隐私保护、模型数据的存储安全；
f） 具备对微调后模型在下游任务中的自动化性能评测；
g） 在设备功能允许时，具备边缘侧或端侧模型的本地微调。

6.2.2.4 模型压缩

模型压缩指标包含：

a) 具备模型剪枝(如基于幅度的剪枝、结构化剪枝和基于梯度的剪枝)、模型量化、模型蒸馏等机制;
b) 提供工具,定量评估小型化后模型的性能损失;
c) 具备小型化过程的可直观显示和监控,能实时跟踪模型的处理过程和处理后模型大小、推理速度和精度损失等关键指标;
d) 具备针对特定任务和资源限制条件相应进行定制优化;
e) 具备面向首字返回的量化,如量化子图融合、有效解码流式传输。

6.2.2.5 模型验证

模型验证指标包含:
a) 具备模型效果评估,支持用户在线上传测试集进行效果评估并能生成评估报告;
b) 提供大模型测试工具或能使用平台之外的测试基准(含数据集和验证指标)验证模型效果;
c) 具有选用模型评估指标,如准确率、召回率等评估指标;
d) 具备提供模型评估中发现预测错误的用例,供用户优化模型。

6.2.2.6 模型部署

模型部署指标包含:
a) 具备部署多种精度(如 FP16、INT8 等)的模型;
b) 具备在虚拟环境(如容器、虚拟机)部署;
c) 具备设置动态批尺寸;
d) 具备与框架无关的模型部署;
e) 具备根据不同的硬件环境和应用需求,选择运行效率最佳的压缩模型版本;
f) 具备边缘侧或端侧模型的升级或更新,支持灰度更新;
g) 具备边缘侧或端侧多版本(如新旧版本)模型的同时运行和访问分流;
h) 对于部署在边缘侧或者端侧的模型,具备根据模型的安全防护要求提供相应的安全保护措施。

6.3 大模型推理及运营

6.3.1 模型推理

6.3.1.1 推理服务部署

模型推理部署指标包含:
a) 具备在分布式运行环境中部署,在云边和多云推理环境中部署;
b) 具备通过 API 或 SDK 的方式提供模型服务;
c) 具备自动化构建,根据服务版本及依赖资源环境自动构建为容器镜像并发布部署;
d) 具备服务与模型本身解耦,实现配置文件与模型部署解耦,灵活进行模型效果更新;
e) 具备服务编排能力,如模型动态可直观显示编排;
f) 具备模型服务间的顺序调用、条件判断、分支选择等配置功能;
g) 具备自定义推理逻辑接入分布式推理引擎;
h) 具备知识库的集成,实现知识库的创建、配置、编辑和删除。

6.3.1.2 推理服务效率

模型推理服务效率优化指标包含:
a) 具备服务上线、下线、更新、回滚,支持灰度更新;

b) 具备统一集群部署、调度、资源调度策略配置管理等；
c) 具备弹性伸缩机制，资源池能根据负载情况和规模实施扩容或缩容；
d) 具备推理流量的负载均衡，根据入口流量智能地均衡到不同的服务节点；具备亿级日流量的推理服务负载均衡；
e) 具备在正式上线前根据实际场景，进行自定义比例的灰度节点发布部署；
f) 具备异构资源管理和调度，如 GPU、CPU 协同计算和调度；
g) 用于对实时性要求高的推理任务，平均 Token 输出时延(不含首 Token)不超过 50 ms；
h) 用于对实时性要求低的推理任务，平均 Token 输出时延(不含首 Token)不超过 125 ms。

6.3.1.3 推理服务稳定性

模型推理服务稳定性保障指标包含：
a) 具备模型上线服务前的测试，测试内容包括功能、性能等；
b) 具备模型服务日志记录、清洗、检视，支持日志文件的本地和云端存储；
c) 具备模型系统运行的可直观显示监控，包括服务器监控、容器监控、调用成功率监控等；
d) 具备实时告警，对于服务器异常、授权不足、成功率超阈值等异常场景，自动触发预警通知的功能，告警包括邮件/短信/微信等方式；
e) 具备流水线化的集成验证，在正式部署前，根据模型使用场景，自动化生成验证用例和自动化触发模型验证；
f) 持跨地域和跨机房集群调度，避免因某地或某机房故障而导致服务异常，保障服务高可用性；
g) 具备现网实时拨测，拨测时间间隔可配置，主动探测生产环境服务是否存在异常。

6.3.2 平台应用

6.3.2.1 行业应用

大模型的行业应用包含：
a) 应在至少 5 个行业使用大模型构建行业解决方案，行业包含如医疗、电力、交通、金融、政务、工业、气象等；
b) 具备内容安全审核，对用户的提问及大模型的回答进行安全审核，例如涉黄、涉恐、涉政等不安全言论进行拦截，并进行后台统计分析，对敏感词、违禁策略、违禁标签等进行编辑和管理；
c) 具备大模型反作弊系统，具备黑白名单管理，对作弊用户进行封禁操作及对风险用户进行信息管理；
d) 具备行业知识库接入和使用，根据不同行业需求构建专属的行业知识库或知识图谱，提升模型和问答效果；
e) 具备插件服务能力，如面向行业应用的插件注册、制作、编辑、管理和数据统计，开发者发布垂直领域个性化插件。

6.3.2.2 交互应用

大模型的交互应用包含：
a) 具备大模型交互应用开发，以可直观显示产品化应用的方式，支持用户直接获取大模型的效果反馈；
b) 具备用户在交互过程中，可创建行业/用户个性化专属提示词库，并利用提示工程进行定制开发和优化；
c) 具备提示词自动生成、人工撰写、比较、导出等操作；

d) 具备涉及多个模型或处理组件的应用工作流编排，支持用户自定义节点组件(如 python 脚本)；
e) 具备智能体应用的创建、部署、配置与测试(如功能预览)；
f) 具备在工作流中，以拖拽方式编辑大模型(如大语言模型、代码生成模型、视觉模型、意图识别组件、提问器组件等)节点的属性或接续关系；
g) 具备交互提示词自动优化功能，提高大模型回答的准确性和可靠性；
h) 具备单轮或多轮对话或推理的思维链条可直观显示；
i) 具备提示词多轮迭代优化和批量优化；
j) 具备大模型交互快修，对大模型效果不佳的场景，快速准确地检测和修复用户输入/大模型输出的问题。

6.3.3 运营赋能

6.3.3.1 数据回流

数据回流指标包含：

a) 具备运营数据的隐私安全保护，如对于涉及合规监管要求的隐私及敏感数据进行过滤排除；
b) 具备数据回流的追溯，在保障数据隐私信息前提下，进行数据回流并持续用于模型优化，对回流数据进行采集和分析，具备真实用户场景下的持续优化能力；
c) 具备将推理结果和监控信息高效准确反馈给开发者，便于开发者完整得到结果用于效果继续调优。

6.3.3.2 生态建设

生态建设指标包含：

a) 具备大模型相关的数据运营，提供相关数据协议，如.ckpt 文件、.pth 文件等；
b) 具备闭源模型、加速库、数据集等资产的托管，及开源模型的管理和运营；
c) 提供不少于 3 款大模型，供调用或二次训练；
d) 大模型包含自然语言处理、计算机视觉、多模态生成；
e) 大模型包含至少 1 种行业[见 6.3.2.1 a)]的应用功能；
f) 提供不少于 50 款大模型，供调用或二次训练；
g) 提供面向大模型的实践案例、应用参考文档等；
h) 提供面向高校或科研机构的大模型培训、研究、应用、竞技和专家认证；
i) 围绕大模型的生产和应用，主导或参与产业协作，与行业机构合作研究大模型行业解决方案，合作机构数量不少于 10 家。

7 成熟度分级规则

7.1 成熟度等级

大模型服务能力成熟度等级划分为基础应用级、协同优化级、深度赋能级 3 级。

a) 基础应用级具备使用大模型的能力。服务供方能提供基本的大模型平台服务能力，和/或能提供基本的模型开发定制服务能力，和/或能提供基本的模型推理及运营服务能力。
b) 协同优化级具备大模型微调和优化的能力。服务供方能提供较为全面的大模型平台服务能力，和/或能提供较为全面的模型开发定制服务能力，和/或能提供较为全面的模型推理及运营服务能力。
c) 深度赋能级具备模型的预训练、微调和优化的能力。服务供方能提供成熟度相当高的大模型平台服务能力，和/或能提供成熟度相当高的模型开发定制服务能力，和/或能提供成熟度相当

高的模型推理及运营服务能力。

从基础应用级至深度赋能级，大模型服务能力的技术要求逐步提升，服务复杂度逐步提升，定制化能力逐步加强。

7.2 能力要求

不同成熟度等级能力要求见表3。

表3 成熟度等级能力要求

能力主域	能力子域	基础能力级要求	协同优化级要求	深度赋能级要求
大模型平台	硬件	6.1.1.1 a)，b)	6.1.1.1 a)～f)	6.1.1.1
		6.1.1.2 a)，b)	6.1.1.2 a)～d)	6.1.1.2
		6.1.1.3 a)～c)	6.1.1.3 a)～f)	6.1.1.3
	软件及工具链	6.1.2.1 a)	6.1.2.1 a)～d)	6.1.2.1
		6.1.2.2 a)，b)1)～3)，c)1)～3)	6.1.2.2 a)～c)	6.1.2.2
		6.1.2.3 a)～b)	6.1.2.3 a)～d)	6.1.2.3
		6.1.2.4 a)～c)	6.1.2.4 a)～g)	6.1.2.4
		6.1.2.5 a)	6.1.2.5 a)	6.1.2.5
		6.1.2.6 a)	6.1.2.6 a)，b)	6.1.2.6
	平台综合	6.1.3.1 a)，b)	6.1.3.1 a)～d)	6.1.3.1
		6.1.3.2 a)，b)	6.1.3.2 a)～e)，g)	6.1.3.2
		6.1.3.3 a)，b)	6.1.3.3 a)～c)	6.1.3.3
		6.1.3.4 a)，b)	6.1.3.4 a)～e)	6.1.3.4
大模型开发定制	数据资源	6.2.1.1 a)	6.2.1.1 a)～c)	6.2.1.1
		6.2.1.2 a)，b)1)，b)2)	6.2.1.2 a)，b)	6.2.1.2
	模型生产定制	6.2.2.1 a)	6.2.2.1 a)～c)	6.2.2.1
		6.2.2.2 a)	6.2.2.2 a)～d)	6.2.2.2
		6.2.2.3 a)～c)	6.2.2.3 a)～d)	6.2.2.3
		6.2.2.4 a)	6.2.2.4 a)，b)	6.2.2.4
		6.2.2.5 a)，b)	6.2.2.5 a)，b)	6.2.2.5
		6.2.2.6 a)	6.2.2.6 a)～c)	6.2.2.6
大模型推理及运营	模型推理	6.3.1.1 a)～c)	6.3.1.1 a)～f)	6.3.1.1
		6.3.1.2 a)，b)	6.3.1.2 a)～d)	6.3.1.2
		6.3.1.3 a)，b)	6.3.1.3 a)～e)	6.3.1.3
	平台应用	6.3.2.1 a)	6.3.2.1 a)～c)	6.3.2.1
		6.3.2.2 a)～c)	6.3.2.2 a)～e)	6.3.2.2
	运营赋能	6.3.3.1 a)，b)	6.3.3.1 a)，b)	6.3.3.1
		6.3.3.2 a)	6.3.3.2 a)～e)	6.3.3.2

8 成熟度评估方法

8.1 评分方法

大模型服务能力成熟度分级可采用评分量化的方式进行，能力子域成熟度得分基于大模型服务对能力子域的各个子项下各个技术要求的满足度。

a) 参与基础应用级成熟度评估的服务，满足子域中全部技术要求，该子域为 2 分；不满足的技术要求项仅为 1 项时，该子域为 1 分；其他情况该子域为 0 分。

b) 参与协同优化级成熟度评估的服务，对应各能力域的基础应用级评估都应达到满分 2 分。满足子域中全部技术要求，该子域为 3 分；较上一成熟度等级本子域新增技术要求项满足率过半时，该子域为 2.5 分；其他情况该子域为 2 分。

c) 参与深度赋能级成熟度评估的服务，对应各能力域的基础应用级评估都应达到满分 2 分，对应各能力域的协同优化级评估都应达到满分 3 分。满足子域中全部技术要求，该子域为 5 分；较上一成熟度等级本子域新增技术要求项满足率过半时，该子域为 4 分；其他情况该子域为 3 分。

8.2 评估域权重

大模型服务包括大模型平台服务、大模型开发定制服务、大模型推理及运营服务、同时提供大模型平台服务和大模型开发定制服务，以及提供全部服务。

成熟度评估时，应根据表 4 确定不同的服务及服务组合类型的能力主域及权重，每种类型能力主域权重相同且和为 100%。

每个能力主域中能力子域权重相同且和为 100%。

表 4 大模型服务及服务组合类型的能力主域权重

序号	大模型服务及服务组合类型	能力主域	能力主域权重
1	大模型平台服务	大模型平台	100%
2	大模型开发定制服务	大模型开发定制	100%
3	大模型推理及运营服务	大模型推理及运营	100%
4	同时提供大模型平台服务和大模型开发定制服务	大模型平台	50%
		大模型开发定制	50%
5	同时提供大模型平台服务、大模型开发定制服务、大模型推理及运营服务	大模型平台	100%/3
		大模型开发定制	100%/3
		大模型推理及运营	100%/3

8.3 计算方法

服务能力成熟度等级得分 A 按照公式(1)进行计算。

$$A = \sum_{k}(k \times j) \qquad \cdots\cdots(1)$$

式中：

A ——成熟度等级得分；

k ——能力主域成熟度得分；

j ——能力主域权重。

能力主域成熟度得分 k 按照公式(2)进行计算。

$$k = \sum_m (m \times n) \quad \cdots\cdots(2)$$

式中：

k ——能力主域成熟度得分；

m——能力子域成熟度得分；

n ——能力子域权重。

8.4 成熟度等级判定

当评估对象在某一成熟度等级下的成熟度得分超过评分区间的最低分视为满足该成熟度等级要求，反之，则视为不满足。

根据表5给出的分数与成熟度等级的对应关系表，结合成熟度等级实际得分 A，可判断大模型服务能力当前所处的成熟度等级。

表5 分数与成熟度等级对应关系

等级	成熟度等级得分 A
基础应用级	$1 \leqslant A \leqslant 2$
协同优化级	$2 < A \leqslant 3$
深度赋能级	$3 < A \leqslant 5$