ICS 35.240
CCS L 70

中华人民共和国国家标准化指导性技术文件

GB/Z 246—2026

人工智能　算力平台通用技术要求

Artifical intelligence—General technical requirements for the computing platform

2026-08-27 发布

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　缩略语 …… 1
5　总体技术框架 …… 2
6　通用要求 …… 3
6.1　算力底座层 …… 3
6.2　平台服务层 …… 3
6.3　模型使能层 …… 3
6.4　行业应用层 …… 4
6.5　集成与接口 …… 4
6.6　安全 …… 5
参考文献 …… 6

前　言

本文件为规范类指导性技术文件。

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委员会(SAC/TC 28)提出并归口。

本文件起草单位：山东港口青岛港集团有限公司、青岛港国际股份有限公司、中国电子技术标准化研究院、青岛大学、中国海洋大学、华为技术有限公司、中国交通信息科技集团有限公司、华东师范大学、先进计算与关键软件(信创)海河实验室、山东大学、交通运输部科学研究院、青岛市人工智能产业协会、青岛科技大学、河北交投智能科技股份有限公司、浙大宁波理工学院、交通运输部天津水运工程科学研究所、浦江国家实验室、博大视野(厦门)科技有限公司、苏州数智科技集团有限公司、青岛前湾联合集装箱码头有限责任公司、科大讯飞股份有限公司、青岛工业软件研究所、青岛前湾西港联合码头有限责任公司、青岛前湾联合集装箱码头有限责任公司、山西省智慧交通实验室有限公司、青岛新前湾集装箱码头有限责任公司、青岛图灵科技有限公司、国家信息中心、北京博雅睿视科技有限公司、青岛前湾集装箱码头有限责任公司、中国移动通信集团山东有限公司、山西省智慧交通实验室有限公司、厦门供应链数智创新有限公司、方舟智能港航口岸服务(青岛)有限公司、赛飞特工程技术集团有限公司、北京宝兰德软件股份有限公司、上海壁仞科技股份有限公司。

本文件主要起草人：吴宇震、魏志强、郭乙运、聂婕、徐洋、马殿光、殷波、马乐、肖圣龙、沈芷月、贾东宁、高雪松、冯栋、艾明飞、王晶、李广合、苏建光、毕涛、王晋、王涛、徐斌、张琪、舒振宇、张杰、白露、张润宇、张翰林、孙骁、刘长辉、张立峰、聂简荻、吕长虹、陈敏鸣、郭立峰、张珂谨、代伟军、赵蓉、申振、孙秀良、赵嘉璇、陈晓明、韩巍、魏建超、苏亮、陈瑞恒、蔡峰、陈乾坤、李志桐、魏雪铭、苏晓丹、黄莉莉、陈军、纪昌秀、姜代楠、贾磊、黄俊、郑志平、李迪、张林梁、丁瑞全、陆仲达、邢冯、李洪巍。

人工智能　算力平台通用技术要求

1　范围

本文件确立了人工智能算力平台技术框架，规定了平台的通用技术要求，包括算力底座层、平台服务层、模型使能层和行业应用层，并规定了各层的通用技术要求。

本文件适用于指导人工智能赋能行业时算力平台的建设和共享。

2　规范性引用文件

本文件没有规范性引用文件。

3　术语和定义

下列术语和定义适用于本文件。

3.1

模型　model

对系统、过程、现象或对象的抽象表示，可用于描述、模拟、预测或决策。

[来源：GB/T 41867—2022，3.1.15]

3.2

推理　inference

从已知前提导出结论的推理方法。

[来源：GB/T 5271.28—2001，28.03.01]

3.3

算力平台　computing power platform

支撑人工智能研发、训练、推理及应用，实现算力、数据、模型、应用等资源汇聚、调度与共享的智能计算服务平台。

3.4

智能体　agent

〈人工智能领域〉能够与环境交互，感知环境、处理信息、决策行动，并根据环境反馈调整自身参数或策略实现进行学习优化和行动的计算机程序或系统。

4　缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

AI4S：人工智能用于科学（AI for Science）

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

CRM：客户关系管理（Customer Relationship Management）

CV：计算机视觉（Computer Vision）

DB：数据库(Database)
ERP：企业资源计划(Enterprise Resource Planning)
GIS：地理信息系统(Geographic Information System)
ID：标识(Identifier，ID)
NLP：自然语言处理(Natural Language Processing)
OS：操作系统(Operating System)
SLA：服务等级协议(Service Level Agreement，SLA)

5 总体技术框架

人工智能算力平台是面向人工智能技术快速演进和大规模创新应用的服务平台，实现算力、数据、模型和应用等资源的汇聚和共享，其总体架构分为算力底座层、平台服务层、模型使能层和行业应用层。



图 1 算力平台总体技术框架

架构组成说明如下。

a) 算力底座层：提供计算、网络、存储、异构计算架构、操作系统(OS)、数据库(DB)等的一体化基础设施。

b) 平台服务层：基于计算、网络、存储等资源池化，由统一的云平台底座屏蔽底层软硬件差异，提供统一接口服务，软件定义资源调度解决算力效率，服务化的标准接口对接开放的模型使能层，支持来源多样的模型(比如开源大模型)；满足大模型训练和推理开发功能要求，提供模型开发、模型训练、模型评测、推理部署、模型发布等服务。

c) 模型使能层：提供数据资产管理、模型资产管理、智能体(Agent)资产管理等功能，支持模型开发中的数据工程、模型工程和智能体(Agent)工程。

d) 行业应用层：根据行业需求，行业应用引擎可提供知识图谱、GIS 等辅助功能。

6 通用要求

6.1 算力底座层

算力底座层支持自主可控的能力，主要组成部分应包括如下。

a) 机房配套：包含一体化机房、制冷、供配电等子系统，为上层算力底座提供空间、供电、散热、环境等功能，实现高可靠、绿色节能、易扩容易演进等能力。

b) 算、网、存基础设施：通过一体化的算网存基础设施层内部的有机配合，实现集群算力最优，支持存算网协同、软硬协同等能力。其功能包括但不限于：

1) 智能计算：包括训练算力和推理算力，训练算力提供面向大规模基础训练、二次微调训练的底层计算能力，推理算力提供面向中心、端边缘侧场景的人工智能基础设施方案；

2) 通用计算：适配各行业多样性计算、绿色计算需求，提供通用算力平台，可用于部署数据处理、存储、云平台、业务应用、高性能计算等业务；

3) 存储：面向训练与推理任务提供高效读写性能的存储能力；

4) 网络：包含参数面、业务面（含存储业务面）、管理面交换机，需要满足人工智能算力底座对网络高吞吐大带宽、低时延、高可靠等要求，支持面向大规模分布式训练和推理场景的专用网络拓扑优化能力。

c) 其他：支持异构计算架构、操作系统、数据库以及存算网等资源统一管理和运营运维的能力；具备对计算、网络、存储资源的智能监控与动态调配能力，以适应不同业务场景的需求。

6.2 平台服务层

平台服务层应支持各类开/闭源模型，支持自然语言、多模态、视觉、预测、科学计算等多种能力模型，面向未来模型发展可以持续演进，应包括算力资源管理和人工智能模型开发两个子层。

a) 算力资源管理：整合分散的物理资源，对算力基础设施物理资源进行池化和集中式管理，实现训练、推理资源共池管理，支持算力资源优化调度，向上层人工智能开发平台提供包括计算、存储、网络、数据库服务在内的服务。从功能上划分为资源池、基础服务、高阶服务：

1) 资源池：统一管理计算、存储、网络等基础设施，形成虚拟计算、虚拟存储和虚拟网络资源池，并提供池化后资源的管理能力，支持分布式池化部署；支持基于业务需求的弹性扩缩容，灵活调整计算、存储、网络等资源分配；

2) 基础服务：提供 API 网关、反向代理服务等基础/公共组件能力；

3) 高阶服务：提供容器服务、存储服务、数据库服务、资源池部署及故障诊断服务等。

b) 人工智能模型开发：面向人工智能模型提供一站式开发平台，应具备如下功能：

1) 具备任务调度、资源调度功能；支持主流 AI 开发框架；

2) 支持模型全量训练，模型微调、模型迁移与精度调试、推理部署、性能调优、模型压缩与量化工具链等功能；

3) 提供模型开发过程中的自动化测试与验证功能，确保模型质量与性能符合要求；

4) 支持全周期人工智能工作流管理，涵盖数据处理、模型开发、模型训练、模型部署等人工智能开发的各个环节的功能需求。面向大模型业务场景，提供支持多机分布式场景下的训练能力和推理能力。

6.3 模型使能层

模型使能层专注于人工智能资产的综合管理与 Agent 的智能编排，负责维护关键人工智能资产，包括数据、知识、模型、编排、工具等，确保这些资产的高效利用与迭代演进；其次，通过 Agent 的智

能编排，模型使能层能静态或动态地调整任务执行策略，优化模型的表现，提升模型精准度与泛化能力。

模型使能层应通过高度集成的数据工程、模型工程和 Agent 工程的工具链，实现基于 Agent 交互、编排、规划、执行和持续优化的智能化应用开发。

模型使能应具备三个功能模块：

a) 数据工程模块是智能化应用的基石，负责处理数据的获取、准备、转换和存储，确保数据的质量、丰富性和可用性，支持数据特征开发和数据深度挖掘；支持隐私增强数据处理功能，包括差分隐私、联邦学习数据脱敏等；

b) 模型开发模块是智能化应用的核心，负责模型的训练/微调、压缩、高效对齐和模型治理等；

c) Agent 工程模块是智能化应用的执行层，负责构建能够自主执行任务、学习和适应环境的智能体；Agent 工程模块应支持多 Agent 协作协议和基于历史数据的任务优先级动态识别；应支持建立运行时监控、行为反馈与自我优化机制，支撑任务连续执行与演化。

6.4 行业应用层

应用层是连接技术创新与实际业务场景的桥梁，发挥着至关重要的角色，宜具备以下要求。

a) 根据行业需要，按需提供图谱、GIS 等通用辅助功能，如矿山行业。行业图谱提供行业通识知识和开放的高质量行业数据集；GIS 提供卫星遥感等多维空间地图数据，为城市规划、气候预测、应急管理等行业应用提供支持。

b) 供面向客户和行业伙伴的多租户管理：支持合理分配和管理底层硬件资源，保障多租户环境下的性能需求；应确保不同租户的数据完全隔离（比如：独立数据库、独立数据表单、租户 ID 区分数据等），防止数据泄露或不当访问；支持多租户动态资源配额，保障高价值业务 SLA；应提供租户级别的监控、审计、告警与资源使用统计等运维功能，确保平台稳定运行与服务质量保障。

c) 支持行业应用与外部数据源的实时对接与数据融合，丰富应用的数据来源与功能。

d) 提供协同与开放服务能力：支持提供标准 API 接口与低代码开发，便于行业伙伴快速接入和二次开发。

6.5 集成与接口

6.5.1 概述

在行业与 AI 技术深度融合的背景下，为确保不同系统、平台及应用之间的无缝连接与高效协同，制定行业＋AI 应用基础技术架构标准的集成与接口要求。本章节旨在明确这些要求，以指导行业 AI 应用的技术实施与标准化建设。

6.5.2 集成要求

算力平台的集成要求如下。

a) 模块化设计：要求技术架构采用模块化设计原则，确保各功能模块之间松耦合、高内聚，便于后续的功能扩展、升级与维护。同时，模块间应遵循统一的数据交互协议和接口标准，以实现无缝集成。

b) 标准化接口：明确界定行业 AI 应用与外部系统（如 ERP、CRM 等）之间的标准化接口，包括数据交换格式、通信协议、安全认证机制等。这些接口应确保数据的准确性、完整性和安全性，支持实时或异步的数据传输。

c) 服务化封装：鼓励将 AI 能力以服务化的形式进行封装，提供 RESTful API、gRPC 等标准化的服务接口。服务接口应具备良好的可扩展性、高可用性和容错能力，以支撑复杂的业务场景和大规模并发访问。

d) 流程整合：在技术架构中融入业务流程管理的能力，确保 AI 应用能够无缝嵌入企业的业务流程中。通过流程整合，实现 AI 应用与现有业务系统的深度集成，提升业务处理的自动化水平和智能化程度。

e) 低代码集成：支持 AI 工作流与现有业务系统的低代码集成，提供拖拽式流程编排工具。

f) 容错机制：要求支持各模块间数据交互的容错机制，确保系统在数据传输异常时的稳定性。

g) 日志采集：支持采集各模块的运行日志、审计日志，提供权限管理与归档管理。

6.5.3 接口要求

算力平台的接口要求如下。

a) 数据接口：规定数据接口的数据格式、字段定义、数据传输协议等，确保不同系统间数据的一致性和互操作性。对于敏感数据，应明确加密传输和访问控制要求，保障数据安全；应支持多种数据格式的自动转换，实现不同系统间数据的无缝对接。

b) 控制接口：定义用于控制 AI 应用行为的接口，如启动、停止、参数配置等。这些接口应具备较高的灵活性和可配置性，以适应不同场景下的应用需求。

c) 状态监控接口：提供用于监控 AI 应用运行状态的接口，包括性能指标、错误日志、告警信息等。通过状态监控接口，实现对 AI 应用的实时监控和故障排查，确保应用的稳定性和可靠性。

d) 扩展接口：预留用于未来技术升级和功能扩展的接口，确保技术架构的灵活性和可扩展性。扩展接口应遵循统一的设计规范和命名约定，便于第三方开发者或集成商进行二次开发或集成。

6.6 安全

安全要求应包含如下内容。

a) 数据加密与保护：所有敏感数据在传输和存储过程中采用强加密算法进行加密，确保数据机密性。同时实施访问控制策略，限制对数据的非法访问。

b) 身份鉴别与授权：建立严格的身份鉴别机制，确保用户身份的真实性。实施基于角色的访问控制，确保用户仅能访问其权限范围内的资源。

c) 系统安全加固：对操作系统、数据库、中间件等关键组件进行安全加固，关闭不必要的服务和端口，定期更新补丁，防范已知漏洞。

d) 日志审计与监控：建立完善的日志审计机制，记录系统操作、用户行为、异常事件等关键信息。通过实时监控和告警系统，及时发现并响应安全事件。

e) 应急响应与恢复：制定应急响应预案，明确安全事件的报告、处理、恢复等流程。定期进行应急演练，提升应对安全事件的能力。同时，建立数据备份与恢复机制，确保业务连续性。

f) 数据所有权监管：制定明确的数据所有权协议，确保数据提供者对其数据享有适当的权利和义务。

g) 数据使用合规要求：在行业应用中，确保所有数据使用行为符合行业或地域的数据保护法规，并在数据共享时明确授权与权限。

参 考 文 献

[1] GB/T 5271.28—2001 信息技术 词汇 第28部分:人工智能 基本概念与专家系统

[2] GB/T 41867—2022 信息技术 人工智能 术语

[3] GB/T 42131—2022 人工智能 知识图谱技术框架

[4] GB/Z 42759—2023 智慧城市 人工智能技术应用场景分类指南

[5] GB/T 43782—2024 人工智能 机器学习系统技术要求

[6] JR/T 0221—2021 人工智能算法金融应用评价规范

[7] YD/T 4044—2022 基于人工智能的知识图谱构建技术要求

[8] YD/T 4070—2022 基于人工智能的接入网运维和业务智能化 场景与需求

[9] YD/T 4316—2023 面向智慧城市应用的人工智能服务能力开放技术要求

[10] 工业和信息化部,中央网络安全和信息化委员会办公室,国家发展和改革委员会,国家标准化管理委员会.国家人工智能产业综合标准化体系建设指南(2024版)(工信部联科〔2024〕113号)

[11] ISO/IEC 19770-1:2017 Information technology—IT asset management—Part 1:IT asset management systems—Requirements

[12] IEEE Std 1471—2000 IEEE Recommended Practice for Architectural Description of Software-Intensive Systems