ICS 35.240.99
CCS L70

T/WHCIO

武 汉 企 业 信 息 化 促 进 会 团 体 标 准

T/WHCIO 0007—2026

中小企业大模型应用架构设计技术规范：总则

General Principles for Architecture Design of Large Model Applications for SMEs

2026－1－16 发布　　2026－1－17 实施

武汉企业信息化促进会　　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。

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这些专利的责任。

本文件由武汉市中小企业发展促进中心提出。

本文件由武汉企业信息化促进会归口。

本文件起草单位：武汉市中小企业发展促进中心、中国工业互联网研究院（工业和信息化部密码应用研究中心）、浙江省工业互联网发展研究院、上海工业数字化研究院、北京科技大学、北京航空航天大学、北京机械工业自动化研究所有限公司、中国科学院工程热物理研究所、长飞光纤光缆股份有限公司、武汉高德红外股份有限公司、武汉神动汽车电子电器股份有限公司、信通院（武汉）科技创新中心有限公司、金蝶软件（中国）有限公司武汉分公司、武汉光迅科技股份有限公司、大唐互联科技（武汉）有限公司、隆达铝业（武汉）有限公司、武汉中科通达高新技术股份有限公司、武汉市德发电子信息有限责任公司、中移（上海）信息通信科技有限公司、中移系统集成有限公司、中国联合网络通信有限公司、云镝智慧科技有限公司、中冶南方都市环保工程技术股份有限公司、江苏长江智能制造研究院有限责任公司、美云智数科技有限公司、北京数码大方科技股份有限公司、北京开物数智科技有限公司、苏州登临科技股份有限公司、青岛奥利普奇智智能工业技术有限公司、鼎捷数智股份有限公司、上海展湾信息科技有限公司

本文件主要起草人：凌端新、张智、曹浩、周乐、张宇、张禄、汪越、殷晴青、张洛本、袁帅鹏、陈廷炯、王柏琳、刘义、杨海龙、胡成国、熊冠楚、罗宇、方波、袁帅南、刘泽宇、李丽、刘涛、孙天一、刘侃、李慧文、赵兴龙、计晓军、戴延军、潘艳飞、侯宝存、陈卫东、洪致远、祁贵林、景嘉祥、宋军恒、何宁波、刘趁伟、李敏、王开学、韩振东、唐栎、何鸿曦、周静静、张静、杨延超、石全、王逸伦、郑灵超、董海阔、余丽、李剑飞、苑春秋、王梦婉、黄铭文、宋群灿、刘慧、张文卿

中小企业大模型应用架构设计技术规范：总则

1 范围

本文件规定了中小企业大模型应用架构设计的总体原则、应用架构、部署模式及安全与合规要求，建立了包括设备层、数据层、模型层、工具层和应用层在内的架构体系，适用于指导中小企业在现有信息系统基础上引入大模型能力，构建具备智能体能力的企业级人工智能应用体系。

本文件适用于具备一定数字化基础、已建设或正在建设信息化系统的中小企业，不适用于尚未开展数字化建设的企业。

本文件适用于中小企业基于通用大模型、行业大模型等构建智能体应用系统的总体架构设计，也适用于相关平台提供方、集成服务商及工具服务商的系统设计与交付参考。

本文件适用于制造型、服务型、商贸型等各类中小企业的大模型应用架构设计工作，对大型企业在相关系统建设中亦具有一定参考和借鉴意义。

本文件不涉及具体大模型的预训练、微调方法及底层算力配置等内容。

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 35273 信息安全技术—个人信息安全规范
GB/T 36073 数据管理能力成熟度评估模型
GB/T 41479 信息安全技术 网络数据处理安全要求
GB/T 41867 信息技术 人工智能 术语
GB/T 42016 信息安全技术—网络音视频服务数据安全要求
GB/T 42018 信息技术 人工智能 平台计算资源规范
GB/T 42755 人工智能面向机器学习的数据标注规程
GB/T 42888 信息安全技术 机器学习算法安全评估规范
GB/T 45288.1 人工智能 大模型 第1部分：通用要求
GB/T 9813.3 计算机通用规范 第3部分：服务器
GB/T 33863.8 OPC统一架构 第8部分：数据访问
ISO/IEC 27001 信息技术 安全技术 信息安全管理体系要求
ISO/IEC 20546 信息技术 大数据 概述与术语
ISO/IEC 5338 信息技术 人工智能 模型运营框架
ISO/IEC 20546,Information technology — Big data

3 术语和定义

GB/T 41867—2022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

基于大量数据训练得到、具有复杂计算架构、能处理复杂任务且具备一定泛化性的深度学习模型。

3.2

垂类大模型 Vertical domain large model

基于基础大模型，在特定行业或领域语料上进行增量训练、微调或知识增强而形成的大模型，具备行业知识理解、场景推理和专业任务执行能力。

3.3

中小企业　small and medium-sized enterprises, SMEs

依据《中小企业划型标准规定》（工信部联企业〔2011〕300号）界定，指在人员规模、营业收入或资产总额等方面未达到大型企业标准的法人企业单位。

3.4

结构化数据　structured data

以固定模式或预定义格式组织的数据，通常存储于关系型数据库或表格中，具有明确的字段定义和数据类型，可通过标准查询语言（如SQL等）进行处理与分析。

3.5

非结构化数据　unstructured data

不具备固定格式或预定义模式的数据类型，包括文本、图像、音频、视频、文档、日志等形式，需通过特征提取、向量化等方法进行处理以供模型识别和应用。

3.6

智能体　agent

基于大模型构建的自治智能系统，能够感知环境、规划任务、调用工具并执行操作，具备自我反思与多轮迭代能力，可在复杂场景中完成特定目标任务。

注：智能体通常具备任务规划、工具调用、上下文管理等能力，支持多轮交互和自我优化。

3.7

智能体开发工具　agent development tools

用于构建、配置、调度和部署智能体的软件工具平台，支持模型接入、对话流程设计、上下文管理、插件调用、工作流编排等功能。

3.8

应用架构　application architecture

构建大模型应用系统时，在企业现有信息系统基础上设计的系统结构，包括模型接入方式、数据交互路径、服务部署方式和功能组件划分等内容。

3.9

私有化部署　on-premise deployment

将大模型及其应用系统部署在企业本地计算和存储资源中的方式。

3.10

公有云部署　public-cloud deployment

通过云服务平台以 API 等形式提供大模型能力的部署方式。

3.11

混合部署　hybrid deployment

结合私有化部署与公有云部署，将不同模块分别部署在本地与云端的架构方式。

3.12

数据治理　data governance

围绕数据采集、清洗、标注、存储、权限、安全与合规等环节建立的数据管理机制。

3.13

数据管理　data management

对数据从采集、集成、处理、存储、使用到归档、销毁等全过程的管理活动。

3.14

嵌入/检索模型 embedding model

用于将文本、图像等非结构化信息转换为向量形式，以支持语义检索、分类、聚类等操作。

3.15

检索增强生成 retrieval-augmented generation

一种结合向量检索与大模型生成能力的技术，通过检索相关知识片段并作为上下文输入，提升生成内容的准确性和可信度。

3.16

CMMLU-SMEs 数据集 cmmlu-smes dataset

一个用于评估中小企业大模型应用性能的测试集，涵盖制造、服务、商贸三行业的多轮对话和推理任务。

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）
A2A：智能体间协同（Agent to Agent）
BOM：物料清单（Bill of Materials）
BPM：业务流程管理（Business Process Management）
ESB：企业服务总线（Enterprise Service Bus）
FPGA：现场可编程门阵列（Field Programmable Gate Array）
GPU：图形处理单元（Graphics Processing Unit）
iPaaS：集成平台即服务（ Integration Platform as a Service）
LLM：大语言模型（Large Language Model）
MaaS：模型即服务（Model As a Service）
MCP：模型上下文协议（Model Context Protocol）
MES：制造执行系统 （manufacturing execution system）
NPU：神经网络处理单元（Neural Processing Unit）
OCR：光学字符识别（Optical Character Recognition）
OPC UA：开放平台通信统一架构（Open Platform Communications Unified Architecture）
PDA：手持终端（Personal Digital Assistant）
PLM：产品生命周期管理（Product Lifecycle Management）
RAG：检索增强生成（Retrieval-Augmented Generation）
REST：具象状态传输（Representational State Transfer）
SaaS：软件即服务（Software As a Service）
SCM：供应链管理系统（Supply Chain Management）
TSN ：时间敏感网络（Time-Sensitive Networking）
WMS：仓储管理系统（Warehouse Management System）

5 总体架构

5.1 架构设计原则

a) 灵活性：架构设计应基于通用标准和规范，具备良好的模块化和扩展能力，支持不同类型的大模型与智能体的接入、升级与替换，并能够适应企业业务变化与技术更新。
b) 轻量化：架构实现时应充分考虑中小企业资金和技术资源的现实限制，避免过度复杂的设计，降低部署门槛。

c) 集成性：架构设计应强调智能体与企业现有信息化系统的兼容性和集成性，确保新系统快速融入企业业务流程。
d) 合规性：应满足国家信息安全相关法律法规和标准要求，确保企业数据安全、模型使用合规。
应用架构总览

5.2 应用架构总览

中小企业大模型应用架构如图1所示，包括设备层、数据层、模型层、工具层、应用层。具体如下：

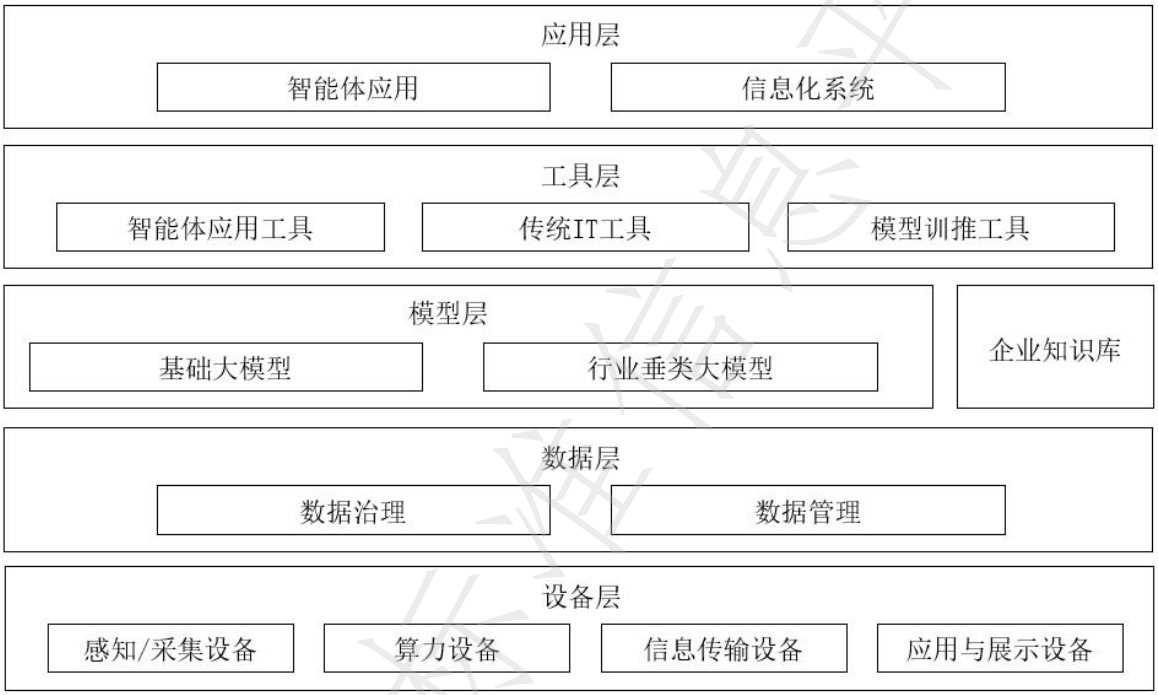

图1 中小企业大模型应用架构

a) 设备层：包含感知/采集设备、算力设备、信息传输设备、应用与展示设备。设备层负责数据采集与算力支撑，是中小企业大模型应用的物理基础，为数据治理与模型运行提供可靠输入。
b) 数据层：包含数据治理、数据管理。数据层是支撑大模型高质量运行的关键环节，能够保障数据的安全性、完整性与可用性，为模型提供可信数据底座。
c) 模型层：包含基础大模型、行业垂类大模型、嵌入/检索模型、企业知识库。模型层负责大模型能力的集中供给和服务管理，提供多模型能力的注册、调用、更新服务，实现大模型资源可复用、可拓展、可管理的目标。
d) 工具层：包含智能体应用工具、传统 IT 工具、模型训推工具。工具层负责模型能力落地的工程化实现，通过结合智能体开发工具、传统 IT 工具和模型训推工具，实现工具调用、模型调用、任务编排、运维监控、模型微调、模型推理等功能。
e) 应用层：包含智能体应用、信息化系统。应用层是大模型能力与中小企业实际使用场景融合的直接体现，支持企业在现有信息化系统基础上，实现对经营、生产、管理、研发等环节的智能升级。

6 设备层

6.1 感知/采集设备

感知/采集设备用于获取业务运行、文档资料、环境感知等多源数据，是大模型应用的数据输入来源。感知/采集设备宜包含但不限于如下部分内容：

a) 文档与知识采集：包括高速扫描仪、高拍仪、移动采集终端等；宜具备文档扫描、图像采集与 OCR/版式解析能力，并能够将采集内容以便于后续结构化处理或向量化处理的格式进行输出；

b) 业务终端采集：包括手持终端、PDA、移动巡检/点检设备等；支持设备标识管理、基本时间同步机制和元数据上送能力，支持通过标准化接口将采集数据纳入企业数据目录或数据管理系统；
c) 工业与物联感知：包括传感器、摄像设备、音频采集设备、条码/RFID 装置、智能仪表等；应通过标准化接口或数据采集网关输出可处理的数据格式（包括结构化、半结构化或非结构化数据），并应支持稳定的数据采集、基本数据校验和数据上送能力，以保障后续分析、处理与模型应用的可靠性。

6.2 算力设备

算力设备宜包含但不限于如下内容，并应根据企业业务规模、大模型推理需求和部署模式进行合理配置。算力设备的功能要求和资源配置应满足 GB/T 42018 的相关规定：

a) 边缘节点：包括 DIN-Rail 工业服务器、AI 盒子等，宜配置具备模型推理能力的处理器（如 GPU、NPU）；其计算能力、存储能力及散热能力应能够支撑企业在本地开展的小规模模型推理任务；
b) 中心服务器：包括机架式 GPU 服务器、CPU 集群等，应支持 Docker/Kubernetes 等容器化和编排能力；网络互联能力应满足模型服务调度、高并发访问与多节点协同的要求；系统整体性能应满足企业模型推理、训练或调度的业务需求；
c) 专用加速器：包括 FPGA、ASIC 推理卡等，适用于低功耗、低时延的本地推理场景；其算力性能、能效比及接口兼容性应满足企业部署在嵌入式设备上的推理需求；相关配置与管理要求应满足 GB/T 42018 6.3 的规定。

6.3 信息传输设备

信息传输设备用于保障数据在设备层、数据层、模型层和应用层之间的安全、可靠和高效传输。信息传输设备宜包含但不限于如下部分内容：

a) 网络与安全设备：包括工业交换机、5G/TSN 网关、工业防火墙等；应支持零信任接入和安全认证机制，网络通信加密方式应满足国家网络数据安全相关要求，并满足 GB/T 41479 的规定；
b) 环境与能耗监测设备：包括能耗计量表、碳排传感器、温湿度/粉尘/噪声采集模块等；应具备数据自动采集、校准与上送能力，数据质量、稳定性和测量精度应满足能耗监测、碳排管理等业务需求；
c) 边缘算力租用：中小企业可通过电信运营商 MEC（边缘计算）平台接入边缘算力资源；相关网络接入设备应满足安全通信、稳定连接及低时延传输的要求，并应支持本地算力与外部算力资源之间的协同调度；网络与数据交互方式应满足 GB/T 41479 的要求。

6.4 应用与展示设备

应用与展示设备用于支持业务人员与系统进行可视化交互、信息浏览与现场协同操作。应用与展示设备宜包含但不限于如下部分内容：

a) 辅助终端：包括 AR 眼镜、工业平板、看板大屏等；应支持数据可视化展示、人机协同操作和现场任务指引等功能，显示性能和交互能力应满足企业业务场景的应用需求。
b) 显示与交互设备：宜具备稳定的图形渲染能力与实时交互能力，支持多种数据展示方式，并可通过标准化通信接口实现大模型推理结果、状态信息及告警信息的及时呈现。

7 数据层

7.1 数据治理

a) 策略与组织：宜建立数据治理策略、职责分工与管理流程，明确数据分类分级规则、权限管理机制及责任主体；
b) 数据标准与元数据：宜制定数据标准体系，统一编码规则、数据口径与主数据管理要求；宜建设企业级元数据，建立元数据采集、更新、维护与使用机制；
c) 数据质量与安全：宜建立数据质量管理制度，包括数据质量规则、质量评估与改进流程；涉及个人信息或网络数据处理的，应满足GB/T 35273和 GB/T 41479的相关规定；

d) 资产目录与共享：宜建立数据目录或数据资产台账，用于记录数据资产的来源、结构、责任主体和使用范围；对外共享或开放数据接口时，应建立审批流程，并根据数据类型和敏感度采取必要的访问控制、水印标识或脱敏保护措施。

7.2 数据管理

a) 采集与同步：应支持网关/OPC UA、消息队列、API、文件批导等多源采集方式；
b) 集成与处理：宜具备数据清洗、融合、脱敏处理能力，并可根据业务需求完成结构化、半结构化或向量化转换；数据转换规则、版本及数据血缘关系应实现记录与可追溯；
c) 时序数据管理：宜建立针对工业监测、设备运行、能耗及环境感知等场景的时序数据管理机制，支持基于时间窗口的聚合分析与异常检测；
d) 存储与检索：宜根据数据特性选择关系型数据库、时序数据库、对象存储、向量数据库等多类型存储方式；宜提供统一的检索接口与数据服务编排能力；
e) 备份与恢复：宜制定数据备份、容灾与恢复策略，保障系统在异常情况下具备可恢复能力，并与平台资源管理机制协同运行。

7.2.1 结构化数据

结构化数据涵盖生产数据、运营数据、业务流程数据、财务数据、设备运行数据、客户交易数据等，包括但不限于订单、合同、财务台账、CRM记录、库存、工艺配方、工单、设备点检、质量监测、能耗、PLC采样流、MES事件消息等。

结构化数据宜满足以下要求：

a) 宜统一接入 ERP、MES、CRM、WMS 等业务系统数据，并建立主数据管理机制，确保关键业务数据的一致性；
b) 宜在采集、处理、存储过程中建立数据质量规则，包括完整性检查、唯一性校验、一致性控制等，确保结构化数据质量满足业务要求；
c) 结构化数据的管理与加工过程宜满足 ISO/IEC 20546 以及 GB/T 42755 的相关要求。

7.2.2 半/非结构化数据

半/非结构化数据应包括半结构化数据、非结构化数据、知识型文档等，包括但不限于JSON/XML日志、BOM、配置文件、报表（XLSX/CSV）、书籍、CAD图纸、图片、音视频、邮件、会议纪要、行业标准、专利、政策公告、行业报告等。

半/非结构化数据宜满足以下要求：

a) 半结构化数据（如 JSON/XML）宜具备自动解析、敏感信息处理等能力；非结构化文档（如图纸、音视频等）宜具备内容抽取、结构化处理或向量化处理能力，以支持检索、分析与模型应用；
b) 宜根据业务需求采用 OCR、语音识别、图像识别等技术提升半/非结构化数据的可用性，但应避免强制性性能指标要求；
c) 涉及个人信息或敏感数据的半/非结构化数据，应满足 GB/T 35273 和 GB/T 41479 的相关要求。

8 模型层

8.1 基础大模型

基础大模型按类型可分为语言大模型、视觉大模型、多模态大模型等，又可分为开源大模型和闭源大模型。

基础大模型宜满足以下要求：

a) 基础大模型应具备语言理解、生成与推理能力，可支持多语言场景，应具有处理较长上下文的能力，并能适配企业常见应用场景的输入规模；
b) 基础大模型的使用应符合算法安全、数据安全及模型输出可控等要求，其风险管理、使用过程与评测方法应满足 GB/T 42888 的相关规定；
c) 基础大模型宜具备可扩展性，能够根据企业需求与部署环境选择适配的模型规模或轻量化版

本。

8.2 行业垂类大模型

行业垂类大模型指经过不同行业数据微调的，适用于特定行业的大模型。

行业垂类大模型宜满足以下要求：

a) 应根据具体行业或领域开展知识增强或微调，具备对行业术语、业务流程与场景任务的理解与适配能力；

b) 宜提供一定的可解释性能力，如通过模型输出、决策链路或检索信息来源等方式支持人工审核与业务验证；

c) 行业垂类模型的训练、评估与使用过程应满足算法安全与数据安全要求。

8.3 嵌入/检索模型

嵌入/检索模型用于将文本、图像等信息转化为向量表示，并通过相似性检索支持知识增强、内容匹配、问答生成等任务。

嵌入/检索模型宜满足以下要求：

a) 应支持文本或多模态数据的向量化处理，并能够与向量数据库或其他检索组件兼容，支持常见的近似相似度检索机制；

b) 宜支持构建检索增强生成（RAG）框架，通过检索外部知识提升回答准确性；相关检索、索引与生成模块应具备可追溯性和可解释性；

c) 嵌入/检索模型的使用、数据处理及知识增强过程应符合数据安全、访问控制与最小化原则，应满足GB/T 35273和GB/T 41479等相关要求

8.4 企业知识库

企业自有知识库宜包含但不限于标准设计文档、工艺流程文档、产品说明书、实验记录、质量管理手册、公司章程、供应商手册、仓储管理制度、员工行为准则、售后服务规范，传感器数据、机器日志、质量控制报告等企业私域数据。

企业宜定期对企业知识库开展管理，管理重点如下：

a) 宜根据企业业务变化、制度更新和知识沉淀情况，建立定期更新机制，保证知识内容的时效性和可用性；

b) 宜建立知识审核、归档与维护流程，对知识内容的准确性、完整性和一致性进行管理，并确保知识库内容来源可追溯；

c) 宜建立版本管理机制与权限分级控制机制，对知识内容的访问、编辑和发布进行管理；涉及个人信息或敏感数据的知识内容，其存储、使用和访问控制应满足GB/T 35273的相关要求。

9 工具层

9.1 智能体应用工具

智能体应用工具宜具备以下功能：

a) 智能体编排平台：应支持以可视化方式建立智能体的意图、任务和工具调用，宜为流程图或状态机的形式，支持：

- “意图–任务－工具”的逻辑链路配置；
- 智能体状态转移逻辑的定义与管理；
- 对话/任务中的记忆管理功能。

b) 插件生态：应具备便捷集成第三方服务或工具的能力，包括但不限于：

- 支持REST/GraphQL等API的接入注册；
- 建立沙箱权限管理机制，限制插件访问范围和权限；
- 支持调用频率控制，宜进行速率限制，确保系统稳定性。
- 支持模型上下文协议（MCP）的注册与通信机制，支持通过MCP规范化插件能力定义，实现大模型与外部数据源、工具及知识库之间的上下文共享与调用；

c） 执行调度：应能对多个任务或调用进行有效调度管理，包括但不限于：
- 支持任务排队执行，如任务队列；
- 宜支持设置任务优先级，高优先级任务可优先处理；
- 支持并发数控制，防止资源过载；
- 支持失败任务自动重试机制，任务自动重试次数宜不大于3次。

d） 监控与分析：应提供可供开发者和运营人员使用的运行追踪与分析工具，包括但不限于：
- 含输入输出的对话和任务的详细日志记录；
- 智能体运行链路的调用追踪视图，用于问题排查；
- 可视化指标看板，可及时反映执行效率、错误率等关键指标。

9.2 传统 IT 工具

工具层应强调与传统IT基础能力的对接适配能力，保障其可运维、可交付、可扩展。包括但不限于：

a） 数据标注工具：支持对训练数据进行自动标注、质量审核等，宜具备以下功能：
- 自动标注：通过规则或模型对原始数据进行初步标注；
- 质量复核：对标注结果宜进行人工复审和纠错，确保数据质量可靠。

b） 模型运维工具（MLOps/LangOps）：支持大模型及其下游应用的全生命周期管理能力，包括但不限于：
- 模型仓库：支持模型统一存储及版本管理；
- CI/CD流程：支持模型的持续集成、自动部署和发布更新；
- 模型监测：支持监测模型性能随时间或数据变化的波动，及时预警；
- 蓝绿部署：支持新旧模型平稳切换，保障上线安全性和稳定性。

c） 系统集成工具：提供系统集成与流程编排支撑，使智能体能够与企业系统高效协同，包括但不限于：
- 企业服务总线（ESB）：统一管理系统间的消息和服务交互；
- 集成平台即服务（iPaaS）：实现系统、服务、数据的整合与流程编排；
- API网关：对外接口进行集中管理、安全认证、限流和监控；
- BPM流程引擎：支持复杂业务流程的建模、执行和监控。

d） 运维和监控工具：保障系统运行的稳定性、可观测性和故障自恢复能力，包括但不限于：
- Prometheus/Grafana：用于指标采集和可视化展示，支持运行健康监控；
- 集中日志管理：统一采集系统运行日志，便于追踪与分析问题；
- 弹性伸缩机制：根据系统负载自动扩容或缩容计算资源；
- 自动修复机制：支持在组件异常或崩溃时自动重启或切换，确保服务可用性。

9.3 模型训推工具

模型训推工具用于支撑大模型及其下游应用的训练、微调、推理与评测过程，是连接模型层与应用层的核心支撑工具。宜具备以下功能：

a） 模型训练与微调：用于中小企业在私有数据上对基础/垂类大模型进行模型训练和参数微调，以适配特定业务场景，宜具备以下功能：
- 应支持基础大模型在企业私有数据上的持续训练与参数高效微调（如LoRA、QLoRA等），并可根据企业业务场景进行增量学习与领域适配；
- 应具备模型训练任务的分布式调度与资源编排能力，支持断点续训与任务追踪。

b） 模型推理与加速：用于在不同硬件与环境中高效提供模型推理服务，提升响应时延与吞吐表现，宜具备以下功能：
- 应支持多种推理引擎与框架，并可按模型规模与硬件配置进行动态优化；
- 宜提供量化、蒸馏、剪枝等轻量化工具，以提升推理性能与部署效率。

c） 性能评测与监测：用于对模型效果与运行性能进行标准化评估与持续监控，宜具备以下功能：
- 应支持基于标准化数据集（如CMMLU-SMEs）的模型效果评估；
- 应提供推理时延、吞吐量、内存占用、准确率等关键性能指标的自动化测试与报告生成。

10 应用层

10.1 智能体应用

智能体应用应以自然语言交互、知识增强与工具链编排为核心，面向既有信息系统形成可运维、可交付、可扩展、可审计的业务能力提升。智能体应用包括但不限于：

a) 经营管理类：经营分析、费用预测、自动月报等；
b) 生产运维类：设备诊断、工艺优化、碳排监控等；
c) 供应链物流类：采购辅采、库存优化、异常预警等；
d) 销售与客户类：智能客服、营销内容生成、客户画像等；
e) 研发与工程类：专利检索、方案生成、实验助手等；
f) 知识与培训类：知识检索、员工培训、SOP问答助手等；
g) 管理合规类：合同条款抽取与一致性校核、隐私与合规提示、审计留痕归集等；
h) 数据与IT运维类：数据质量巡检、日志归因与异常定位、低代码流程编排与发布等。

10.2 信息化系统

信息化系统集成应遵循可嵌入、可观测、安全可控原则，优先采用标准化接口与流程建模方法，支持跨系统调用的统一鉴权、统一审计与统一运维。包括但不限于：

a) 嵌入插件模式：在业务系统中以插件/扩展方式调用Agent API；业务系统可包括ERP、CRM、PLM、SCM、MES、WMS、OA等。应采用统一身份（SSO）与基于角色的访问控制，并对API进行认证、授权、限流与审计，防范越权、失效等常见风险；
b) 低代码集成模式：通过低代码平台与连接器（Connectors）对接多源系统与数据，快速将智能体嵌入业务流程或应用；
c) 门户与大屏模式：企业知识助手、生产看板、经营驾驶舱统一入口；
d) 流程自动化模式：将智能体编排入业务流程，如RPA/BPM等，实现端到端的流程可视化闭环。

11 部署模式

11.1 部署模式分类

中小企业在应用大模型构建智能体时，受限于资金投入和运维能力，宜优先采用公有云部署模式，以降低建设成本并加快系统上线。但企业也应根据企业规模、资金投入、安全合规需求和数据敏感程度等因素，综合评估后选择以下一种或多种部署模式：

11.1.1 私有化部署

a) 模型、智能体及相关数据均部署在企业本地数据中心或企业自有的计算资源中；
b) 数据应在企业内部闭环流动，不出域、不外传；
c) 适用于数据高度敏感、隐私要求严格或时延要求较高的业务场景；
d) 本地算力设备应满足6.2设备层相应要求，数据存储和传输应满足GB/T 35273相关安全规范。

11.1.2 公有云部署

a) 企业通过公有云平台API调用云端提供的大模型与智能体服务，本地可按需配置算力设置；
b) 企业数据通过API上传到云端，企业仅获得模型推理结果；
c) 适用于数据敏感性较低、预算有限且快速上线需求较高的中小企业；
d) 云服务提供方应保证数据安全、隐私保护与服务高可用性，安全性应满足GB/T 41479的要求。

11.1.3 混合部署

a) 敏感数据处理及核心推理任务宜部署在企业本地，通用推理任务、模型训练、更新等宜放置在公有云；
b) 本地设备与云端通过加密通道进行安全通信，数据交换应满足GB/T 35273和GB/T 41479要求；
c) 适用于对数据安全有明确要求，又同时需要借助云端算力和服务资源灵活扩展的企业场景；
d) 本地部署的关键算力和网络通信可参考6.2和6.3的相应要求。

11.2 部署模式选择要求

a) 企业选择部署模式时，应综合考虑数据安全敏感度、成本预算、网络通信环境、运维能力以及未来扩展需求；
b) 涉及个人隐私和商业秘密的数据处理环节应优先考虑私有化或混合部署模式；
c) 使用公有云模式时，应确保服务提供商满足ISO/IEC 27001等安全合规要求，并具备相应数据合规认证；
d) 采用混合部署时，应明确划定云与本地部署边界，确保敏感数据不越界，满足数据分级保护相关要求。

12 安全与合规要求

12.1 数据安全要求

12.1.1 数据采集与传输安全

a) 数据采集设备应具备身份认证、访问控制和数据校验能力；采集与传输过程中应采用安全通信协议实现数据的端到端保护；
b) 数据在传输过程中应保持完整性、机密性和可验证性，防范数据被篡改、伪造或非法截取风险；
c) 应采用安全通道、加密传输、零信任访问等技术手段保障数据通信安全，网络安全防护应满足GB/T 41479要求。

12.1.2 数据存储与访问控制

a) 数据存储应采用必要的加密、隔离、脱敏等安全措施，保证敏感数据在静态存储过程中的安全性；加密与密钥管理要求应符合相关安全规范；
b) 应建立基于最小权限原则的访问控制机制，如角色访问控制或基于属性的访问控制，对敏感数据实施精细化权限管理；
c) 数据访问与操作应具备可审计性，应记录关键操作的审计日志，并保证日志的完整性和可追溯性；日志管理应满足GB/T 35273的要求。

12.1.3 数据治理与合规

a) 应建立数据分类分级管理制度，对不同敏感度的数据采取相应的保护策略，并在治理流程中明确定义责任主体；
b) 应定期开展数据安全检查、安全评估与风险分析，数据治理流程应满足GB/T 42755的要求；
c) 涉及个人信息的处理应严格遵守《中华人民共和国个人信息保护法》和GB/T 35273的相关要求。

12.2 智能体安全要求

12.2.1 智能体执行安全

a) 智能体宜在沙箱环境中运行，执行的外部操作宜限制于受控的安全范围内；
b) 智能体对外部工具或API调用应实施严格权限控制，严禁执行可能损害系统或数据安全的操作；
c) 智能体行为应设置异常检测与响应机制，出现异常行为时应及时告警并记录。

12.2.2 大模型调用安全

a) 模型调用接口应实施认证鉴权机制，防止未经授权调用；
b) 模型服务应限制调用频次和请求速率，防范拒绝服务攻击和恶意利用；
c) 应实施模型使用日志记录与监控，记录内容包括调用方身份、调用时间、参数、输出结果等；日志管理应满足GB/T 35273的要求。

12.2.3 智能体交互安全

a) 智能体应具备对用户输入内容的安全过滤机制，防止注入攻击、敏感信息泄露等安全风险；
b) 对于涉及敏感信息交互的场景，应实施额外的二次身份验证或确认机制；

c) 智能体交互日志应包含用户身份、交互时间、内容及系统响应；日志管理应满足GB/T 35273的要求。

附 录 A
（资料性）
中小企业大模型应用评测数据集示例

本附录提供中小企业大模型应用评测数据集（CMMLU-SMEs）的示例，用于评价大模型在中小企业实际应用场景中的表现。

A.1 数据集结构

数据集包含制造业、服务业、商贸业三个行业的典型业务场景问题与答案对。每个场景下包含单轮和多轮对话任务。

数据集包含如下字段：

a) 行业：行业类别；
b) 场景：问题所在业务场景；
c) 问题与参考答案：标准输入与正确输出；
d) 模型生成答案与评测结果：用于评价模型表现。

A.2 测评指标

a) 综合准确率（Accuracy）
b) 多轮对话任务成功率（Success Rate）
c) 五步推理任务准确率（Reasoning Accuracy）

A.3 示例数据格式

数据以JSON格式存储，示例数据格式如下：

```
{
  "行业": "制造业",
  "场景": "设备故障诊断",
  "问题": "设备温度异常升高的可能原因有哪些？",
  "参考答案": ["冷却系统故障", "传感器故障", "润滑不良"],
  "模型生成答案": ["冷却系统故障", "润滑不良"],
  "评测结果": {
    "准确率": 0.67,
    "推理准确率": 0.80
  }
}
```

参 考 文 献

[1] GB/T 35273 信息安全技术 个人信息安全规范
[2] GB/T 36073 数据管理能力成熟度评估模型
[3] GB/T 41479 信息安全技术 网络数据处理安全要求
[4] GB/T 41867 信息技术 人工智能 术语
[5] GB/T 42016 信息安全技术—网络音视频服务数据安全要求
[6] GB/T 42018 信息技术 人工智能 平台计算资源规范
[7] GB/T 42755 人工智能面向机器学习的数据标注规程
[8] GB/T 42888 信息安全技术 机器学习算法安全评估规范
[9] GB/T 45288.1 人工智能 大模型 第1部分：通用要求
[10] GB/T 9813.3 计算机通用规范 第3部分：服务器
[11] GB/T 33863.8 OPC统一架构 第8部分：数据访问
[12] ISO/IEC 27001 信息技术 安全技术 信息安全管理体系要求
[13] ISO/IEC 20546 信息技术 大数据 概述与术语
[14] ISO/IEC 5338 信息技术 人工智能 模型运营框架
[15] ISO/IEC 20546 Information technology — Big data