

ICS 35.240

CCS 160

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIAS 020—2024

医疗大模型语料一体机应用指南

Application Guide of Medical Large Model Corpus Integrated Machine

2024 - 11 - 16 发布

2024 - 11 - 18 实施

上海市人工智能行业协会 发布

目 次

前 言 3

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 系统框架 2

6 医疗应用区要求 4

 6.1 一类应用 4

 6.2 二类应用 4

 6.3 三类应用 5

7 模型区要求 6

 7.1 基础模型要求 6

 7.2 垂类模型要求 6

8 智能体区要求 7

 8.1 智能体引擎要求 7

 8.2 智能体开发要求 7

9 医疗语料区要求 7

 9.1 一般性要求 7

 9.2 语料库要求 8

 9.3 语料工具链要求 10

10 算力区要求 11

 10.1 AI 加速计算资源配置要求 11

 10.2 通用计算资源配置要求 12

11 一体机操作系统要求 12

12 部署环境要求 13

参考文献 14

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。
本文件由上海市人工智能行业协会提出并归口。

本文件起草单位:上海库帕思科技有限公司、工业互联网创新中心(上海)有限公司、上海信投数字科技有限公司、中国信息通信研究院、上海市卫生和健康发展研究中心、上海市疾病预防控制中心、上海市先导产业促进中心、上海市人工智能行业协会、上海交通大学医学院附属瑞金医院、复旦大学附属中山医院、复旦大学附属华山医院、复旦大学附属儿科医院、上海交通大学医学院附属新华医院、海军军医大学第一附属医院、复旦大学附属浦东医院、万达信息股份有限公司、上海商汤善萃医疗科技有限公司、上海天数智芯半导体有限公司。

本文件起草人:山栋明、陈昕、郑忠斌、黄海清、方金武、赵艳、夏天、蒋璐伊、钱宇波、顾建英、王艺、李济宇、张晓波、冯天宜、朱立峰、张纪阳、钟俊浩、黄虹、孙新、金钢、施家樑、卢恒、刘红、李光亚、张宏亮、赵春昊、江龙平、蒋龙泉、刘捷宸、李贤、张敬谊、汪智、尤伟、李雨容、应嘉玮、武雅文、蔡剑飞、宋薇珊、柏志安、李绍欣、潘佳伟、钱琨、王治勋、张琪、程思杰、刘从进、王蓓蕾、章文杰、胡铭姗、胡从坤、刘硕鹏、李震洋、胡珂溢、倪雯静、郎巳良。

医疗大模型语料一体机应用指南

1 范围

本文件规定了医疗大模型语料一体机的能力要求，包括系统框架、医疗应用区、模型区、智能体区、医疗语料区、算力区、一体机操作系统等要求。

本文件适用于医疗大模型语料一体机的研发、配置、维护、采购等工作。

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 41867-2022 信息技术 人工智能 术语

GB/T 9813.3-2017 计算机通用规范 第3部分：服务器

T/SAIAS 015—2024 语料库建设导则

3 术语和定义

GB/T 41867-2022界定的术语和定义适用于本文件。

3.1

医疗大模型语料一体机 `medical large model corpus integrated machine`

专门为医疗领域设计的高性能计算设备，以一体机服务器形式存在。

3.2

医疗大模型 `medical large model`

基于深度学习等AI技术，针对医疗领域特定问题（如疾病诊断、药物研发等）训练而成的大型神经网络模型，具备处理和分析复杂医疗数据，提供精准医疗决策支持等能力。

3.3

医疗语料库 `medical corpus`

医疗领域相关文本、图片、音频、视频等数据的集合，可用于训练和优化医疗大模型，以提升其在医疗领域的应用效果。

3.4

临床决策支持 `clinical decision support`

基于大量患者数据，使用医疗大模型，为医生提供诊断和治疗建议的智能代理，旨在辅助医生作出更加精准和科学的医疗决策。

3.5

算力动态分配 `dynamic computing power allocation`

根据医疗大模型的实时需求和系统负载情况，动态调整算力资源分配的智能化策略，以确保系统在高负载情况下仍能稳定运行，并可优先处理关键任务。

4 缩略语

下列缩略语适用于本文件。

AEs 医疗不良事件(Adverse Events)

AI 人工智能(Artificial Intelligence)

API 应用程序编程接口(Application Programming Interface)

ASR 自动语音识别(Automatic Speech Recognition)

GPU 图形处理单元(Graphics Processing Unit)

NPU 神经网络处理单元(Neural Processing Unit)

SAEs 严重不良事件(Severity Adverse Events)

TFlops 每秒万亿次浮点运算(Tera Floating-point Operations Per Second)

TPU 张量处理单元(Tensor Processing Unit)

5 系统框架

医疗大模型语料一体机架构图如图1所示,包括医疗应用区、医疗模型区、算力区、医疗语料区、智能体区以及医疗一体机操作系统。架构中各区应充分解耦,可根据使用需求配置各区内容,具体如下:



图1 医疗大模型语料一体机架构图

- a) 医疗应用区应包括但不限于诊疗辅助、医院管理、健康管理、养老托育服务、临床专病辅助决策、基层全科医生辅助决策、公共卫生服务、临床用药辅助、药物研发、中医中药辅助、医学影像辅助诊断、手术辅助规划、医学教学、医学科研等应用；
- b) 模型区宜包含基础模型、垂类模型；
- c) 算力区宜包括普通算力、智能算力、内存配置；
- d) 智能体区应包括集成规划引擎、工具引擎、执行引擎、记忆引擎；
- e) 医疗语料区应包括语料库和语料工具链；
- 1) 语料库宜包括医疗通识语料、医疗行业知识语料、医疗临床数据、向量数据库；

2) 语料工具链应包括语料清洗工具、语料标注工具。

f) 医疗一体机操作系统宜包括医疗模型管理、引擎调度、智能体管理、应用管理、语料管理、算力调度管理及集群管理。

6 医疗应用区要求

医疗应用区宜根据医疗应用复杂程度分为一类应用、二类应用以及三类应用。

6.1 一类应用

一类应用包括诊疗辅助、医院管理、健康管理、养老托育服务。

6.1.1 诊疗辅助

诊疗辅助要求如下：

- a) 应具备根据患者症状、体检结果和实验室数据提供初步诊断的能力；
- b) 应具备辅助医生制定个性化治疗方案的能力；
- c) 应具备实时更新医学知识，以提供最新的诊疗建议。

6.1.2 医院管理

医院管理要求如下：

- a) 应具备辅助优化医院资源分配，提高医疗服务效率的能力；
- b) 应具备根据医院运营数据辅助决策制定的能力；
- c) 应具备提升医院医疗质量质量管理的能力；
- d) 应具备提供改善患者就医流程建议的能力。

6.1.3 健康管理

健康管理要求如下：

- a) 应具备跟踪个人健康数据，提供健康风险评估的能力；
- b) 应具备根据个人健康状况提供定制化的健康建议和干预措施。

6.1.4 养老托育服务

养老托育要求如下：

- a) 应具备为老年人和儿童提供个性化护理计划的能力；
- b) 应具备根据老年人和儿童的健康状况数据，及时预警健康风险；
- c) 应具备提供心理支持和社交互动，增强老年人和儿童的生活质量。

6.2 二类应用

二类应用包括临床专病辅助决策、基层全科医生辅助决策、公共卫生服务、临床用药辅助、药物研发、中医中药辅助。

6.2.1 临床专病辅助决策

临床专病辅助决策要求如下：

- a) 应具备针对特定疾病提供深入的临床知识支持的能力；
- b) 应具备辅助医生进行复杂病例分析和治疗方案选择的能力。

6.2.2 基层全科医生辅助决策

基层全科医生辅助决策要求如下：

- a) 应具备提供全面的基层医疗服务知识的能力；
- b) 应具备辅助全科医生进行常见病和多发病的诊断和治疗建议。

6.2.3 公共卫生服务

公共卫生服务要求如下：

- a) 应具备根据公共卫生数据预测疾病流行趋势的能力；
- b) 应具备制定和实施公共卫生干预措施的能力；
- c) 应具备提供公共卫生教育和健康促进知识的能力。

6.2.4 临床用药辅助

临床用药辅助要求如下：

- a) 应具备药物信息查询和药物相互作用分析的能力；
- b) 应具备辅助医生进行药物剂量调整和用药指导的能力；
- c) 应具备根据患者用药反应及时调整治疗方案的能力。

6.2.5 药物研发

药物研发要求如下：

- a) 应具备药物设计和筛选的辅助能力；
- b) 应具备药物效果预测和副作用评估的能力；
- c) 应具备临床试验设计和数据分析的能力。

6.2.6 中医中药辅助

中医中药辅助要求如下：

- a) 应具备中医理论学习和应用的能力；
- b) 应具备中药方剂配伍和调整建议的能力；
- c) 应具备中医诊断和治疗建议的能力。

6.3 三类应用

三类应用包括医学影像辅助诊断、手术辅助规划、医学教学、医学科研。

6.3.1 医学影像辅助诊断

医学影像辅助诊断要求如下：

- a) 应具备医学影像的辅助识别和分析能力；
- b) 应具备辅助医生进行疾病诊断和病情评估的能力。

6.3.2 手术辅助规划

手术辅助规划要求如下：

- a) 应具备辅助手术方案设计和风险评估的能力；
- b) 应具备手术过程中的实时辅助和建议能力；

c) 应具备制定术后恢复和并发症预防规划的能力。

6.3.3 医学教学

医学教学要求如下：

- a) 应具备提供医学教育资源和模拟训练的能力；
- b) 应具备辅助医学生和医生进行专业知识学习的能力；
- c) 应具备评估学习效果和提供个性化教学建议的能力。

6.3.4 医学科研

医学科研要求如下：

- a) 应具备医学文献检索和分析的能力；
- b) 应具备辅助科研设计和实验方法建议的能力。

7 模型区要求

模型区要求包括基础模型要求、垂类模型要求。

7.1 基础模型要求

基础模型要求包括基础能力要求、长文本处理能力要求、多模态数据处理要求、参数规模要求、性能要求。

7.1.1 基础能力要求

应具备自然语言理解、知识推理、文本生成、跨语言处理、跨模态理解、跨模态生成等方面能力，其能力宜通过行业公认的测评集测试。

7.1.2 长文本处理能力要求

应能处理8k以上的文本输入。

7.1.3 多模态数据处理能力要求

应能处理图像数据（X射线图像、CT图像、MRI图像、超声图像、病理切片图像等）、文本数据（电子病历、医生诊断报告、医学文献等）、实验室检查数据、基因数据、音频数据（医患沟通音频、心音、肺音等）、视频数据等多模态数据。

7.1.4 参数规模要求

模型参数规模要求如下：

- a) 一类应用宜使用不大于13B参数的稠密模型或使用混合专家模型激活总参数在13B及以下；
- b) 二类应用宜使用13B-30B参数的稠密模型或使用混合专家模型激活总参数在13B-30B；
- c) 三类应用宜使用30B-72B参数的稠密模型或使用混合专家模型激活总参数在30B-72B。

7.1.5 推理效率要求

模型推理效率要求如下：

- a) 应在处理文本长度不超过200token的条件下，响应时间不大于500ms；
- b) 文本生成速度应不小于10tokens/s。

7.2 垂类模型要求

垂类模型要求包括专业能力要求、精准性要求、知识库更新要求、价值对齐要求。

7.2.1 专业能力要求

应具备医疗知识问答、医学文本分类、医学信息抽取、医学自然语言推理、医学内容生成、医学多模态数据处理能力，能力宜通过行业公认的测评集测试。

7.2.2 医疗准确性要求

医疗准确性要求如下：

- a) 应具备精准的医疗场景问题解决能力；
- b) 应具备理解医疗行业特有术语和上下文的能力。

7.2.3 知识库更新要求

知识库更新要求如下：

- a) 应根据具体业务需求调整和优化；
- b) 应具备定期更新机制，以获取到最新的医学知识；
- c) 应具备实时监测机制，监测医学领域的最新动态，将获取的最新信息更新到知识库中；
- d) 应该具备持续学习能力，随行业发展不断更新知识库。

7.2.4 价值对齐要求

模型生成的内容应具备如下要求：

- a) 应具备医疗价值，不存在不良医疗信息内容；
- b) 应符合人类普遍认同的关于国家、民族、社会以及全人类层面的价值理念、处事原则和行为规范等；
- c) 应具备情感认同、交互、共情共鸣等方面的能力；
- d) 应具备医学领域的文化素养与人文关怀。

8 智能体区要求

智能体区要求包括应智能体引擎要求和智能体开发要求。

8.1 智能体引擎要求

智能体引擎要求如下：

- a) 应具备规划引擎、工具引擎、执行引擎、记忆引擎功能；
- b) 记忆引擎应支持向量数据库及外部数据访问。

8.2 智能体开发要求

应支持低代码或无代码开发智能体功能。

9 医疗语料区要求

医疗语料区要求包括一般性要求、语料库要求、语料工具链要求、向量数据库要求。

9.1 一般性要求

医疗语料区一般性要求包括：

- a) 应具备语料内容的独立性，可根据应用需求灵活配置各部分语料内容；
- b) 应按照至少每年、每半年、每月的更新周期对医疗通识语料、医疗行业知识语料、医疗临床数据进行更新；
- c) 应存在定期审核和更新机制，保证语料库的准确性和完整性；
- d) 应以知识块、知识对、三元组、知识图谱等方式存在；
- e) 应具有规范性、一致性、完整性、及时性、准确性、有效性和及时性；
- f) 应遵从数据安全管理的相关法律法规、标准以及合同约定等，并不得损害国家利益、社会公共利益和他人合法权益；
- g) 应具备正确的价值导向，符合科技伦理和社会道德。

9.2 语料库要求

医疗语料区语料库要求包括医疗通识语料要求、医疗行业知识语料要求、医疗临床数据要求和向量数据库要求。

9.2.1 医疗通识语料要求

医疗通识语料应包括医疗领域的相关知识，包括但不限于基础医学、临床医学、保健医学、康复医学、理疗学、麻醉学、内科学、外科学、妇产科学、儿科学、眼科学、耳鼻咽喉科学、口腔医学、皮肤病学、性医学、神经病学、精神病学、急诊医学、核医学、肿瘤学、护理学、预防医学与卫生学、药学、中医学、民族医学、中西医结合医学、生物学、生理学、遗传学、放射生物学、分子生物学、生态学、神经生物学、微生物学、病毒学、人类学、生物工程、心理学等。

医疗通识语料应包含但不限于如下部分内容：

- a) 书籍：包括国内外权威出版社获取的医疗健康领域教科书，及其他高质量生命健康相关书籍；
- b) 期刊论文：包括国内外知名期刊或者从权威数字出版平台获取的医疗健康领域的学术论文；
- c) 数据库：包括从医学数据库获取的临床试验、人类疾病、基因序列、蛋白质序列和功能信息、药物药理药效等数据；
- d) 知识库：包括从医学知识库中获取的生物、疾病和基因等维度之间的关联知识数据；
- e) 专利：包括从专利搜索引擎检索得到的医疗健康领域的专利正文；
- f) 循证医学知识库：通过整合和分析大量的医学研究数据和临床实践经验构建的循证医学结构化知识库；
- g) 临床指南：包括官方机构发布的关于某个疾病和专科领域内容有关的诊断、管理和治疗的决策标准文档；
- h) 药品数据库：包括药品的名称、成分、适应症、用法用量、不良反应、药物间相互作用、注意事项等信息的数据库；

- i) 医疗器械资料：如医疗设备的使用说明书、操作手册等，包括常见的诊断设备（如心电图机、血糖仪等）和治疗设备（如呼吸机、除颤仪等）的相关资料；
- j) 专家共识：包括由特定领域具有专业知识和经验的专家共同讨论、协商达成的统一意见或建议，一般发表于高水平的期刊论文中；
- k) 中医学典籍：包括中医领域的经典著作；
- l) 药材药方：包括经典中药书籍中记录的历代中医方剂，包括方剂的组成、用法、主治病症等；
- m) 相关法律法规：国家和地方关于医疗行业的政策、指南和标准等；
- n) 健康宣教材料：涵盖各类健康教育和预防知识，如疾病预防、健康生活方式等。

9.2.2 医疗行业知识语料要求

医疗行业知识语料应包含但不限于如下部分内容：

- a) 专科专病数据：结合医疗管理部门重点监测的单病种形成的专科专病数据；
- b) 医学数据库：包含了海量的医学文献资源的权威医学数据库；
- c) 医学知识图谱：医学知识图谱以图结构的形式表示医学领域的知识，包括疾病、症状、药物、治疗等实体以及它们之间的关系。

9.2.3 医疗临床数据要求

医疗临床数据应包含但不限于如下部分内容：

- a) 电子病历：包括针对个体的医院门诊、住院形成的电子病历，含诊断、实验室检验数据、影像检查及报告、治疗方案等内容；
- b) 医疗住院数据：包含入出院、转科记录、药品医嘱、住院非药品医嘱、住院费用、住院费用分类、住院发票、住院退药记录等业务的相关数据；
- c) 研发阶段临床试验数据：记录受试者在接受药物治疗期间发生的任何不良事件（AEs）、严重不良事件（SAEs）数据，以及评估药物达到预定疗效指标的情况、药物在体内吸收、分布、代谢和排泄过程等的的数据。

9.2.4 向量数据库要求

向量数据库提供语料向量形式存储，要求如下：

- a) 应支持多种向量类型，能够存储和处理多种类型的向量数据；
- b) 应具备向量索引机制，应支持快速的相似性检索和范围查询；
- c) 应提供查询功能；
- d) 应提供API接口，支持多种编程语言和框架；
- e) 应提供数据加密功能与访问控制机制；

- f) 应提供数据备份和恢复机制;
- g) 应支持不少于100万向量规模;
- h) 应满足不大于300ms返回速度。

9.3 语料工具链要求

语料工具链要求包括清洗工具要求和标注工具要求。

9.3.1 清洗工具要求

清洗工具要求包括文本数据清洗工具要求、图片数据清洗工具要求、音频数据清洗工具要求和视频数据清洗工具要求。

9.3.1.1 文本数据清洗工具要求

文本数据清洗工具功能要求如下:

- a) 应具备对医疗文档进行格式标准化处理的能力;
- b) 应具备获取医疗文档基本信息, 并进行标签记录的能力;
- c) 应具备识别医疗文档语言类型并进行标记的功能;
- d) 应具备清除转换后医疗文本中无关信息的能力;
- e) 应具备检测医疗文档敏感数据, 并确保其符合隐私保护标准的能力。

9.3.1.2 图像数据清洗工具要求

图像数据清洗工具功能要求如下:

- a) 应具备使用算法与模型抽取医疗图像文字信息, 或对图像进行医疗领域描述;
- b) 应具备使用深度学习模型对医疗图像进行分类, 以区分不同的疾病类型;
- c) 应具备使用算法与模型, 在医学图像中识别并分割出所需区域的功能;
- d) 应具备检测医疗图片敏感数据, 并确保其符合隐私保护标准的能力。

9.3.1.3 音频数据清洗工具要求

音频数据清洗工具功能要求如下:

- a) 应具备对医疗音频进行ASR处理, 并生成文本数据的能力;
- b) 应具备获取医疗音频基本信息, 并进行标签记录的能力;
- c) 应具备识别医疗音频语言类型并进行标记的功能;
- d) 应具备清除转换后医疗音频中无关信息的能力;
- e) 应具备检测医疗音频敏感数据, 并确保其符合隐私保护标准的能力。

9.3.1.4 视频数据清洗工具要求

视频数据清洗工具功能要求如下:

- a) 应具备自动获取医疗视频文件的帧率、时长等信息的功能;
- b) 应具备使用算法模型自动根据参数切分医疗视频镜头的功能;
- c) 应具备使用算法与模型抽取医疗视频文字信息, 并提取音频内容转换为文字信息的功能;
- d) 应支持自动抽取医疗视频中的关键帧, 或根据逻辑判定抽取关键帧的功能;
- e) 应支持自动识别医疗视频中的单帧水印, 并在相应的帧数上进行标记记录的功能;

f) 应具备检测医疗视频敏感数据，并确保其符合隐私保护标准的能力。

9.3.2 标注工具要求

标注工具要求包括文本数据标注工具要求、图像数据标注工具要求、音频数据标注工具要求和视频数据标注工具要求。

9.3.2.1 文本数据标注工具要求

关于文本数据标注工具的要求应参照标准“T/SAIAS 015—2024 第7.3.1节”执行。

9.3.2.2 图像数据标注工具要求

关于图像数据标注工具的要求应参照标准“T/SAIAS 015—2024 第7.3.2节”执行。

9.3.2.3 音频数据标注工具要求

音频数据标注工具用于根据指定的规范要求对音频数据进行标注、精确标注。

本文件规定了音频数据标注工具的最低功能要求，对可用的音频数据标注工具不作具体规定，本文件的使用者宜根据标注对象、标注要求和不同的数据集格式选择或开发合适的标注工具。

音频数据标注工具应具有合理的页面布局，至少应同时呈现音频选区、工具区、标签区等三个区域：

- a) 音频选区：该区域用来显示音频和具体的打标签；
- b) 工具区：显示所有在标注过程中需要用到工具；
- c) 标签区（区分图层）：显示标注的内容。

音频数据标注工具还应支持医疗音频预标注，支持针对医疗音频的声音事件检测预标注、语音识别预标注（方言）、语音识别预标注（外语）、音频信号源分离预标注、语调模式识别预标注等功能。

9.3.2.4 视频数据标注工具要求

视频数据标注工具用于根据指定的规范要求对视频数据进行标注、精确标注。

本文件规定了视频数据标注工具的最低功能要求，对可用的视频数据标注工具不作具体规定，本文件的使用者宜根据标注对象、标注要求和不同的数据集格式选择或开发合适的标注工具。

视频数据标注工具应具有合理的页面布局，至少应同时呈现视频选区、工具区、标签区等三个区域：

- a) 视频选区：该区域用来显示视频和具体的打标签；
- b) 工具区：显示所有在标注过程中需要用到工具；
- c) 标签区（区分图层）：显示标注的内容。

视频数据标注工具还应支持医疗视频预标注，支持针对医疗视频的镜头分割预标注、关键帧提取预标注、行为识别预标注等功能。

10 算力区要求

一体机算力区要求分为智能算力配置要求、普通算力配置要求以及算力分配和调度要求。

10.1 AI 加速计算资源配置要求

AI加速计算资源提供AI加速能力，包括GPU、NPU、TPU等架构，其配置要求如下：

a) 使用稠密模型参数在13B及以下或使用混合专家模型激活总参数在13B及以下，并发请求数不大于5，FP16浮点总算力宜不低于280（TFlops/s），AI加速卡总内存宜不低于64GB，配置4张AI加速卡情况下，单卡FP16浮点总算力配置宜不低于 70（TFlops/s），单卡内存宜不低于16GB；

b) 使用稠密模型参数在13B以上30B及以下或使用混合专家模型激活总参数在13B以上30B及以下，并发请求数不大于10，FP16浮点总算力宜不低于640（TFlops/s），AI加速卡总内存宜不低于96GB，配置4张AI加速卡情况下，单卡FP16浮点总算力配置宜不低于 160（TFlops/s），单卡内存宜不低于24GB；

c) 使用稠密模型参数在30B以上72B及以下或使用混合专家模型激活总参数在30B以上72B及以下，并发请求数不大于15，FP16浮点总算力宜不低于1600（TFlops/s），AI加速卡总内存宜不低于192GB，配置8张AI加速卡情况下，单卡FP16浮点总算力配置宜不低于 200（TFlops/s），单卡内存宜不低于24GB。

10.2 通用计算资源配置要求

通用计算资源提供CPU算力，其配置要求如下：

- a) 当AI加速卡在4卡及以下时，CPU核心数不宜少于24，内存宜按照CPU核心数进行1:4适配；
- b) 当AI加速卡在8卡及以上时，CPU核心数宜不低于64，内存宜按照CPU核心数进行1:4适配。

表1 普通算力配置要求表

一体机卡数 配置要求	4卡及以下	8卡及以上
CPU核心数	宜不少于24	宜不少于64
内存	按照核心数进行1:4适配	

11 一体机操作系统要求

一体机操作系统应配置提供对整体资源管理和调度。具体参考要求如下：

- a) 应支持应用的安装部署、启用、停止、升级、卸载；
- b) 应支持智能体及引擎的安装、升级、启用、停用、卸载；
- c) 应支持模型、推理引擎和微调引擎的装载、加载、卸载；
- d) 应支持提示词、微调数据集、知识库数据集、测评数据集的存储、启用、更新、备份、删除功能；
- e) 应支持向量数据库的安装部署、启用、停止、升级、卸载、备份；
- f) 应支持对AI加速卡的资源分配、回收、抢占等调度功能；
- g) 应采用高可用性架构的算力分配和调度系统；
- h) 应集成容错机制；
- i) 应在AI加速卡单点故障发生时，对应资源隔离；
- j) 应配备监控系统，可实时跟踪算力资源的使用情况和系统性能指标；
- k) 应包含定期维护和更新机制。

12 部署环境要求

应符合GB/T 9813.3-2017中4.8的要求。

全国团体标准信息平台

参考文献

- [1] 《卫生健康行业人工智能应用场景参考指引》
-