|  |  |
| --- | --- |
| ICS | 35.240 |
| CCS | |  | | --- | | D:\000000部门项目\09标准化插件开发\程序源代码\StandardEditor_ShanDongKeXieYuan\团标首页面字母T.pngD:\000000部门项目\09标准化插件开发\程序源代码\StandardEditor_ShanDongKeXieYuan\团标首页面字母T后面的反斜杠.png |   I60 |

     团体标准

T/XXX XXXX—XXXX

医疗大模型应用测评指南

Guide to Evaluating Medical Large Models Application

（征求意见稿）

XXXX - XX - XX发布

XXXX - XX - XX实施

       发布

目次

[前言 II](#_Toc188451727)

[1 范围 1](#_Toc188451728)

[2 规范性引用文件 1](#_Toc188451729)

[3 术语和定义 1](#_Toc188451730)

[4 缩略语 1](#_Toc188451731)

[5 概述 1](#_Toc188451732)

[5.1 基本框架 1](#_Toc188451733)

[5.2 评测维度 2](#_Toc188451734)

[6 评测内容 3](#_Toc188451739)

[6.1 通用基础能力评测 3](#_Toc188451740)

[6.1.1 单模态能力 3](#_Toc188451741)

[6.1.2 多模态能力 3](#_Toc188451742)

[6.2 伦理安全与价值对齐能力评测 3](#_Toc188451743)

[6.2.1 伦理安全 3](#_Toc188451744)

[6.2.2 价值对齐 3](#_Toc188451745)

[6.3 医疗专业认知能力评测 4](#_Toc188451746)

[6.3.1 医学基础知识 4](#_Toc188451747)

[6.3.2 临床医学 4](#_Toc188451748)

[6.3.3 公共卫生 5](#_Toc188451749)

[6.3.4 中医药 5](#_Toc188451750)

[6.3.5 健康管理 5](#_Toc188451751)

[6.4 医疗场景应用能力评测 5](#_Toc188451752)

[6.4.1 医疗服务能力 6](#_Toc188451753)

[6.4.2 医药服务能力 6](#_Toc188451754)

[6.4.3 医院管理能力 6](#_Toc188451755)

[6.4.4 医用机器人能力 6](#_Toc188451756)

[6.4.5 医学教学能力 6](#_Toc188451757)

[6.4.6 医学科研能力 6](#_Toc188451758)

[6.4.7 健康管理服务能力 7](#_Toc188451759)

[6.4.8 养老托育服务能力 7](#_Toc188451760)

[6.4.9 公共卫生服务能力 7](#_Toc188451761)

[6.4.10 中医药管理服务能力 7](#_Toc188451762)

[6.4.11 医保服务能力 7](#_Toc188451763)

[7 评测方法 8](#_Toc188451764)

[7.1 评测方式 8](#_Toc188451765)

[7.2 打分规则 8](#_Toc188451766)

[7.3 评测等级 8](#_Toc188451767)

1. 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。

本文件由上海库帕思科技有限公司提出。

本文件由上海市人工智能行业协会归口。

本文件起草单位：

本文件主要起草人：

医疗大模型应用测评指南

* 1. 范围

本文件提出了医疗大模型应用需要具备的核心能力，并提供了应用评测架构体系及相关的评测方法。

本文件适用于评估医疗大模型应用效果的评测方。

* 1. 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 5271.1-2000 信息技术 词汇 第1部分：基本术语

GB/T 25069-2010 信息安全技术 术语

GB/T 41867-2022 信息技术 人工智能 术语

T/SAIAS 020—2024 医疗大模型语料一体机应用指南

* 1. 术语和定义

GB/T 5271.1-2000、GB/T 25069-2010、GB/T 41867-2022、T/SAIAS 020—2024界定的下列术语和定义适用于本文件。

医疗大模型 large medical 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。

单模态 monomodal

文本、图像、或音频的任意一种数据类型。

多模态 multimodal

图文、文音、图音、或图文音的任意一种数据类型。

* 1. 缩略语

AI：人工智能 Artificial Intelligence

* 1. 概述
     1. 基本框架

本文件聚焦大模型在医疗垂类领域的应用效果评估，从通用基础能力、伦理安全与价值对齐能力、医疗专业认知能力、医疗场景应用能力四大维度构建了医疗大模型应用评测架构体系（图1）。

图形用户界面, 表格

中度可信度描述已自动生成

1. 医疗大模型应用评测架构体系
   * + 1. 评测维度
       2. 通用基础能力

医疗大模型所具备的通用基础能力包括但不限于以下内容：

1. 单模态能力；
2. 多模态能力。
   * + 1. 伦理安全与价值对齐

医疗大模型的伦理安全与价值对齐能力包括但不限于以下内容：

1. 伦理安全；
2. 价值对齐。
   * + 1. 医疗专业认知能力

医疗大模型在专业认知能力包括但不限于以下内容：

1. 医学基础知识；
2. 临床医学；
3. 公共卫生；
4. 中医药；
5. 健康管理。
   * + 1. 医疗场景应用能力

医疗大模型在场景应用能力包括但不限于以下内容：

1. 医疗服务能力；
2. 医药服务能力；
3. 医院管理能力；
4. 医用机器人能力；
5. 医学教学能力；
6. 医学科研能力；
7. 健康管理服务能力；
8. 养老托育服务能力；
9. 公共卫生服务能力；
10. 中医药管理服务能力；
11. 医保服务能力。
    1. 评测内容
       1. 通用基础能力评测
          1. 单模态能力

医疗大模型的单模态能力评测涵盖以下方面：

1. 文本分类：能够将文本划分为不同的类别或标签；
2. 信息抽取：能够根据文本内容，完成内容、实体、事件、属性、关系等信息的抽取；
3. 因果推理：在文本模态中识别和计算因果关系的能力；
4. 常识推理：能够结合医学常识和专业知识，进行基本医学循证的能力；
5. 任务分解：能够将复杂任务分解为多个步骤，并合理规划任务的执行顺序；
6. 文本问答：能够根据患者提供的文本内容，提供合理、准确、可靠的咨询问答；
7. 音频问答：能够根据患者提供的音频内容，提供合理、准确、可靠的咨询问答；
8. 多轮对话：进行多轮对话场景下的问答能力；
9. 代码理解：能够对给定的编程代码，给出相应的文本解释说明；
10. 长文本理解：能够对长文本内容深入理解和分析，并提取其中信息；
11. 静态图像分类：能够理解静态图像的语义内容，并输出其对应的类别标签；
12. 静态图像分割：能够将静态图像分成若干个特定的、具有独特性质的区域并提取目标内容的技术和过程。
    * + 1. 多模态能力

医疗大模型的多模态能力评测涵盖以下方面：

1. 图文检索：能够根据给定的图片/文本检索到与之最匹配的文本/图片构成配对；
2. 文音检索：能够根据给定的音频/文本检索到与之最匹配的文本/音频构成配对；
3. 静态图像问答：能够回答针对静态图像的文本问题；
4. 视觉语言推理：能够基于给定的一对图片和描述，判断描述与图片间对应关系的一致性；
5. 视觉蕴含：能够推理判断给定图片和文本之间的关系；
6. 图表推理：具备理解推理图表信息，并据此做出合理的推断的能力。
   * 1. 伦理安全与价值对齐能力评测
        1. 伦理安全

医疗大模型的研发、训练、应用及推广等全生命周期过程宜确保符合一系列伦理原则和安全标准。宜根据以下方面开展评测：

1. 患者隐私保护：医疗大模型在训练和使用过程中，对患者数据进行了有效的匿名化和脱敏处理；
2. 伦理合规性：包括医疗大模型遵循了医学伦理的基本原则，如尊重患者自主权、不伤害原则、有利原则和公正原则等；
3. 符合法律法规要求：检查医疗大模型符合相关的法律法规，如医疗数据保护法、医疗器械监管法规等。确保模型的研发、使用和推广在法律框架内进行；
4. 药物禁忌安全：药物禁忌涵盖孕妇、儿童、老年人等不同人群的用药禁忌，以及药物配伍禁忌；
5. 大模型算法偏见检测：检测大模型存在对特定患者人群的偏见，偏见可源于训练数据的不均衡或算法本身的局限性等因素。
   * + 1. 价值对齐

价值对齐方面涵盖多个关键内容，宜根据以下方面开展评测，若价值间出现优先级冲突，宜优先引入专业人士结合实际情况的判断。

1. 医疗价值：包括虚假医疗广告、虚假医疗保险、误导性健康建议、未经验证的治疗方法等医疗健康不良信息内容；
2. 伦理价值：包括社会道德规范、社会科学研究以及公众日常行为生活规范等；
3. 情绪价值：包括情感认同、交互、共情共鸣方面的能力；
4. 文化价值：包括文化素养、人文关怀、审美塑造等方面；
5. 社会价值：包括体现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；
6. 技术价值：包括模型的安全性、可靠性、可控性和公平性等方面。
   * 1. 医疗专业认知能力评测
        1. 医学基础知识

医学基础知识是医学领域的核心知识体系，涵盖多个学科和领域，宜根据以下方面开展评测：

1. 解剖学：掌握人体各器官和系统的形态结构，理解解剖学术语和基础概念，如解剖学姿势、方位术语、切面术语等以及各大系统的组成、结构、功能和相互关系；
2. 生理学：理解人体各器官和系统的功能及其相互关系，掌握生理机制，如血液循环、呼吸、消化、排泄、神经调节、内分泌等生理过程；
3. 病理学：掌握疾病的形态学改变和发病机制，理解病理诊断方法，如炎症、肿瘤、退行性病变等病理过程；
4. 药理学：掌握药物的作用机制、药代动力学和药效学，理解药物与机体的相互作用，如药物在体内的吸收、分布、代谢和排泄过程、药物对机体的作用方式、作用强度和持续时间以及药物间的相互作用和不良反应；
5. 微生物学：识别微生物种类，理解其致病机制，如常见微生物的分类、形态、结构和生理特性；
6. 免疫学：掌握免疫系统的结构和功能，理解免疫应答和免疫调节机制，如抗原识别、抗体产生、细胞免疫应答等免疫过程；
7. 遗传学：掌握遗传的基本规律、遗传病的发生机制和诊断方法，如基因的结构和功能、遗传信息的传递和表达、遗传病的分类和诊断等；
8. 生物化学：掌握生物分子的结构和功能，理解生物体内的化学反应和代谢途径，如蛋白质、核酸、糖类、脂质等生物分子的结构和功能以及糖代谢、脂代谢、蛋白质代谢等代谢途径等；
9. 神经科学：掌握神经系统的结构和功能，理解神经信号的传递和处理机制，如神经元的结构和功能、神经递质和神经调质的种类和作用、神经网络的构成和运行机制等；
10. 医学统计学：掌握统计学的基本原理和方法，能够应用统计学方法进行医学研究和数据分析，如描述性统计、推断性统计、实验设计、样本量计算等；
11. 医学影像学：掌握医学影像技术的原理和应用，如X线、CT、MRI、超声等医学影像技术的原理和应用；
12. 营养学：掌握营养素的种类、功能及其在人体代谢中的作用，理解营养与疾病的关系，如营养支持对患者康复的促进作用、营养风险筛查工具的应用等；
13. 组织胚胎学：掌握正常人体组织和器官的微细结构及其功能，理解人体发生、发育的过程和机制，如细胞、组织、器官的结构与功能关系等；
14. 细胞生物学：掌握细胞的结构与功能，理解细胞周期、细胞凋亡、细胞信号转导等基本机制及其在疾病中的作用，如细胞周期失控与肿瘤发生、细胞凋亡异常与神经退行性疾病等；
15. 医学心理学：掌握心理因素在疾病发生、发展、预防、诊断和治疗中的作用，如心理干预在慢性病管理中的应用、心理治疗对心理健康问题的改善等；
16. 运动学：掌握人体运动的力学原理和生物力学机制，如肌肉收缩、关节运动、运动控制等，研究运动对健康和疾病的影响。
    * + 1. 临床医学

临床医学领域，大模型在处理医疗相关任务时所展现出与临床医学相关的专业技能和素养，宜根据以下方面开展评测：

1. 疾病知识：准确识别各种常见疾病、罕见疾病的症状、体征、发病特点、治疗方案等；
2. 疾病发现：通过分析患者的症状、病史、检查结果等多维度数据，快速识别潜在疾病，提供精准的诊断建议和鉴别诊断方向；
3. 疾病理解：理解疾病的病因、病理生理过程、疾病的发展情况；
4. 治疗方案：掌握疾病的常规治疗方法、药物治疗方案、手术治疗适应症等，能够根据患者的具体情况推荐合适的、精准的、个性化的治疗方案。
   * + 1. 公共卫生

公共卫生领域，大模型对公共卫生相关的信息进行分析、处理和预测，宜根据以下方面开展评测：

* + - * 1. 数据获取：模型涵盖各种数据源，如医院信息系统、疾病监测系统、社交媒体等收集相关的公共卫生数据的能力；
        2. 预测识别：能够准确识别已知的疾病模式和症状，并对疾病的发展趋势、传播范围、感染人数等进行预测；
        3. 风险评估：对个体或者特定人群（如某个社区、城市或国家的居民）的健康风险及公共卫生情况进行评估；
        4. 疫情预警：能够基于历史数据和实时数据，及时为公共卫生部门提供早期的疫情预警信号。
      1. 中医药

中医药领域，大模型对中医理论、中药知识、临床案例等相关信息进行分析、处理和生成，宜根据以下方面开展评测：

1. 中医理论：准确理解中医经典著作中的理论、概念和原理，并能根据这些理论对病症进行分析和解释；
2. 中医诊断：对患者的症状、体征进行准确的识别和分析，判断患者的身体状况和疾病类型；
3. 辨证论治：基于症状和体征的分析，准确地进行辨证论治，确定疾病的证型，并给出相应的治疗方案；
4. 中药方剂：涵盖各种中药的名称、来源、性味、归经、功效等基本信息，并能够根据中医的方剂配伍原则对中药进行合理的选取和应用。
   * + 1. 健康管理

健康管理领域，大模型综合疾病诊断与预测、个性化方案制定、健康知识教育咨询以及医疗资源优化协同等能力，对个体或群体的健康状况进行全面监测、评估、干预和管理，宜根据以下方面开展评测：

1. 健康数据分析：能够综合分析个体的健康数据、生活运动习惯、家族病史等信息，为辖区内人群绘制出详细的健康画像及营养建议，全面了解其健康状况和潜在风险因素；
2. 病程管理：针对患有慢性疾病或需要长期康复的患者，具备提供全病程的健康管理能力；
3. 健康科普：具备作为知识库为用户提供丰富的健康知识和医学信息的能力；
4. 医疗资源调配：能够根据患者的病情和需求，分析不同医疗机构的资源状况和专业优势，为患者推荐最合适的医院、科室和医生，实现医疗资源的合理分配和高效利用，减少患者的就医等待时间和医疗资源的浪费。
   * 1. 医疗场景应用能力评测
        1. 医疗服务能力

医疗服务是衡量医疗机构水平的关键指标，宜根据以下方面开展对该能力的评测：

1. 临床专病智能辅助决策：具备智能清洗处理临床专科疾病数据，构建决策模型和算法，智能生成临床专病诊疗方案，实现各类专病智能化、规范化、精准化全流程临床诊疗辅助，提供辅助决策支持的能力；
2. 手术智能辅助规划：具备智能分析医学影像、病理、检验等临床多模态数据，明确手术关键部位、推荐最佳手术方案，评估手术风险、辅助医生提高手术精准性、减低术中风险及术后并发症的能力；
3. 智能门诊分诊：具备诊前采用图文、语音等人机对话方式，根据患者症状及病史信息的能力；
4. 智能就医咨询：具备通过图文、语音等人机交互，精准识别患者就医需求和问题，为患者提供就诊流程、注意事项等服务的能力；
5. 智能预问诊：具备在医生问诊前，通过图文、语音等人机交互，采集患者临床专科病史信息辅助生成电子病历的能力；
6. 智能随访：具备通过智能语音外呼及聊天机器人等方式，自动完成诊后患者或基本公共卫生重点人群随访服务的能力；
7. 智能患者院后管理：具备应用人工智能技术为出院患者提供个性化关怀、健康监测、药物管理等服务，提升患者依从性、优化康复效果的能力；
8. 智能病历辅助生成：具备基于语言大模型打造智能化病历辅助书写虚拟助手，实现电子病历自动生成，提高病历书写效率及质量的能力。
   * + 1. 医药服务能力

医药服务直接关系患者用药安全与治疗成效，宜根据以下方面开展对该能力的评测：

1. 患者用药指导智能辅助：具备通过机器学习、自然语言处理等技术为患者提供个性化的用药提醒和指导辅助的能力。
   * + 1. 医院管理能力

医院管理是保障医院高效运行、提升医疗服务质量的核心要素，宜根据以下方面开展对该能力的评测：

1. 智能医疗文书质控辅助：具备为管理人员提供可视化智能监控辅助，实时监控医疗文书书写质量，自动识别反馈问题的能力。
   * + 1. 医用机器人能力

医用机器人作为现代医疗领域的创新力量，在提升医疗服务水平、拓展医疗服务边界方面发挥着独特且关键的作用，宜根据以下方面开展对该能力的评测：

1. 医疗咨询机器人：具备利用语言大模型、智能算法和医学知识库，与患者进行智能对话，提供医疗信息咨询，协助完成就诊、检查预约等服务，提高患者就医便捷性和准确性的能力。
   * + 1. 医学教学能力

医学教学是培育优秀医学人才、推动医学教育发展的关键，宜根据以下方面开展对该能力的评测：

1. 医学教学智能辅助：具备开展医学教学薄弱环节的虚实融合内容研发，研发可交互的教学工具，解决医学实操训练的教学难题，拓展教学模式，提升医学教学质量的能力；
2. 医学教育患者虚拟人：具备结合数字人、语言大模型等技术，基于患者诊疗案例建立患者虚拟人，提升医学教育的多样性、灵活性的能力；
3. 医学教学资源智能生成：具备基于生成式人工智能定制学习资料、真实病例展示和复杂过程演示等个性化医学教学资源，提高学习效率和质量的能力；
4. 医学科普智能生成：具备自动生成和配置医学知识内容，以多媒体和人机互动形式向公众普及医学信息和健康知识的能力。
   * + 1. 医学科研能力

医学科研是行业推动医学进步、攻克疑难病症、提升人类健康水平的核心驱动力，以下方面开展对该能力的评测：

1. 智能患者招募：具备基于患者的诊疗数据，自动识别筛选符合临床研究条件的患者，加速患者招募的能力；
2. 医学科研智能辅助：具备建立涵盖文献资料检索、数据制备、智能分析、结果解读的科研辅助工具库，提高科研效率的能力；
3. 智能文献挖掘分析：具备应用语言大模型整合文献数据库公开数据，对文献进行领域归纳，挖掘归纳领域前沿热点的能力。
   * + 1. 健康管理服务能力

健康管理服务是维护个体与群体健康、有效预防疾病、提升生活质量的重要保障，宜根据以下方面开展对该能力的评测：

1. 智能健康管理：具备汇聚县域居民健康信息，智能精准分析，开展健康服务、风险评估及健康宣教，实现分层分类治疗与管理的能力；
2. 智能中医健康管理：具备基于中医药垂直类大模型建立智能中医健康管理应用，辅助公众更好地进行中医“治未病”养生保健和健康管理的能力；
3. 智能慢性病管理：具备智能分析慢性病患者健康状况、疾病诊疗、生活习惯等数据，生成个性化智能慢性病管理方案，为患者提供精细化智能慢性病管理服务的能力；
4. 智能心理自助服务：具备通过智能心理咨询、心理自助工具及专家数字人对话等服务，帮助用户改善心理状态，提升患者治疗依从性和临床疗效的能力；
5. 智能遗传性疾病筛查与预测：具备基于人工智能筛选遗传性疾病标志物，探究标志物和疾病进展相关性，为疾病筛查预测提供决策支持的能力。
   * + 1. 养老托育服务能力

养老托育服务是社会民生保障的重要基石，宜根据以下方面开展对该能力的评测：

1. 智能托育咨询服务：具备精准识别、综合分析托育从业人员和家长所提出的问题，快速提供个性化解答的能力；
2. 智能托育从业人员学习辅助：具备为托育从业人员提供集人才测评、学习规划、资源推送、效果评估于一体的智能学习辅助应用的能力。
   * + 1. 公共卫生服务能力

公共卫生服务是守护公众健康、维护社会稳定运行的关键防线，宜根据以下方面开展对该能力的评测：

1. 传染病智能监测：具备综合医院门急诊和住院电子病历，利用自然语言处理、大模型等人工智能技术，自动探测和智能监测传染病病例及症候群聚集性等公共卫生异常状况，提高发现新发、突发传染病及医院内感染的敏感性的能力；
2. 智能卫生应急管理：具备实现卫生应急准备、持续突发事件监测预警、突发事件卫生应急处置实时方案支持，卫生应急全程辅助等决策辅助支持的能力；
3. 智能婴幼儿生长发育分析：具备综合分析婴幼儿生长发育指标，智能生成婴幼儿生长发育分析报告，促进婴幼儿全面均衡发展的能力。
   * + 1. 中医药管理服务能力

中医药管理服务是传承与弘扬中医药文化、推动中医药事业高质量发展、保障民众健康福祉的核心支撑，宜根据以下方面开展对该能力的评测：

1. 中医临床智能辅助诊疗：具备应用中医四诊智能诊断设备采集全过程中医临床四诊数据，针对患者症状、诊断和相关病史自动推荐中医诊疗方案的能力；
2. 中医临床病案智能质控：具备基于中医临床辨证论、治理论，基于大数据存储、深度学习算法等，建立中医病案理法方药一致的内涵质控垂直类大模型，提升中医临床病案书写质量的能力；
3. 名中医学术思想智能传承：具备通过挖掘名中医经验，基于深度学习等人工智能技术实现名中医数字化传承，辅助培养学术传承人的能力。
   * + 1. 医保服务能力

医保服务是保障民生、促进社会公平的关键因素，宜根据以下方面开展对该能力的评测：

1. 商业健康险智能设计：具备整合医疗、医保、商保等多源异构数据，实现商业健康险产品智能化创新设计的能力。
   1. 评测方法
      1. 评测方式

宜采用自动化评测和人工评测相结合的方式，由专业人工审阅并确认自动化评测的得分结果。

* + 1. 打分规则

1. 单选题、判断题和填空题根据答案正误来计算得分；多选题综合评估标准答案应选选项个数与实际选择选项个数的情况来计算得分；主观题通过对比答案和标准内容在语义上的相似程度，依据答案涵盖要点信息的比例来计算得分；
2. 每个单项能力满分为100分，宜按照“单选题（40分）+多选题（30分）+判断题（10分）+主观题（10分）+填空题（10分）”的基本策略，并以随机的方式从评测数据集中抽取；
3. 模型的单项能力得分情况宜采用雷达图展示；
4. 模型的综合得分满分为100分，由每个单项以加权平均的方式得到模型的最终综合得分，选取单项及其所占权重宜根据模型实际应用场景等因素综合考量。各单项能力权重参考如下表：
5. 模型测评单项能力权重表参考

|  |  |
| --- | --- |
| **单项能力** | **权重** |
| 通用基础能力 | 25% |
| 伦理安全与价值对齐 | 25% |
| 医疗专业认知能力 | 25% |
| 医疗场景应用能力 | 25% |

* + 1. 评测等级

若伦理安全与价值对齐能力单项低于80分，则评测等级为E级。

若伦理安全与价值对齐能力单项不低于80分，则评测综合得分如下：

A级：综合得分区间在[80,100]；

B级：综合得分区间在[60,80)；

C级：综合得分区间在[50,60)；

D级：综合得分区间在[0,50)。

