ICS (请选择合适的分类号写在这里)
CCS (请选择合适的分类号写在这里)

T

团　　体　　标　　准

T/TMAC ×××—202X

儿童医学大模型评测规范

Specification of evaluation for pediatric medical large model

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

已授权的专利证明材料为专利证书复印件或扉页，已公开但尚未授权的专利申请证明材料为专利公开通知书复印件或扉页，未公开的专利申请的证明材料为专利申请号和申请日期。

××××-××-××发布　　××××-××-××实施

中国技术市场协会　　发布

中国技术市场协会（TMAC）是科技领域内国家一级社团，以宣传和促进科技创新，推动科技成果转移转化，规范交易行为，维护技术市场运行秩序为使命。为满足市场需要，做大做强科技服务业，依据《中华人民共和国标准化法》《团体标准管理规定》，中国技术市场协会有序开展标准化工作。本团体成员和相关领域组织及个人均可提出制修订 TMAC 标准的建议并参与有关工作。TMAC 标准按《中国技术市场协会团体标准管理办法》《中国技术市场协会团体标准工作程序》制定和管理。TMAC 标准草案经向社会公开征求意见，并得到参加审定会议多数专家、成员的同意，方可予以发布。

在本文件实施过程中，如发现需要修改或补充之处，请将意见和有关资料反馈至中国技术市场协会，以便修订时参考。

中国技术市场协会地址：北京市海淀区复兴路甲 23 号城乡华懋大厦 12 层 1217 室。

邮政编码：100036　　电话：010-68270447　　传真：010-68270453

网址：www.ctm.org.cn　　电子信箱：136162004@qq.com

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专利的责任。

本文件由首都医科大学附属北京儿童医院提出。

本文件由中国技术市场协会归口。

本文件起草单位：国家儿童医学中心、首都医科大学附属北京儿童医院、北京百川智能技术有限公司、小儿方健康(北京)有限公司。

本文件主要起草人：×××、×××、×××、×××、×××、×××、×××。

儿童医学大模型评测规范

1 范围

本文件规定了儿童医学大模型（以下简称“大模型”）的评测维度、评测内容、评测方法。

本文件适用于开发、部署和使用大模型的医疗及研究机构、企业单位、服务提供商、监管部门等开展大模型的评估评价。

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 5271.1 信息技术 词汇 第1部分：基本术语

GB/T 25069 信息安全技术 术语

GB/T 41867 信息技术 人工智能 术语

3 术语和定义

下列术语和定义适用于本文件。

3.1

医疗大模型 medical large model

基于深度学习等人工智能技术，针对医疗领域特定问题（如疾病诊断、药物研发等）训练而成，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的大型神经网络模型。

3.2

多模态 multimodal

文本、图像、声音、视频等的组合。

4 评测维度

4.1 大模型评测应从模型基础能力、儿科专业知识能力、医疗应用能力、科研教学能力、安全与伦理能力等五个维度开展。

4.2 模型基础能力应包括但不限于：多模态交互、抗干扰与容错、自主学习。

4.3 儿科专业知识能力应包括但不限于：医学基础知识、临床医学、健康管理、儿童生长发育、公共卫生。

4.4 医疗应用能力应包括但不限于：临床诊疗服务、健康咨询服务、公共卫生服务、医药服务能力、医保服务能力。

4.5 科研教学能力应包括但不限于：科学研究、辅助教学。

4.6 安全与伦理能力应包括但不限于：数据安全、伦理遵循。

5 评测内容

5.1 模型基础能力

5.1.1 多模态交互

评测应至少包含以下内容：

a) 多模态数据处理：文本、图像、声音、视频等数据的识别与处理；

b) 推理能力：进行因果推理（如分析症状与疾病之间的因果关系）和常识推理（如结合医学常识进行循证推理）；

c) 多轮对话：进行多轮对话场景下的问答；

d) 任务分解与问答：将复杂任务分解为多个步骤，并合理规划执行顺序；进行文本问答（包括多轮对话和长文本理解），根据患者提供的文本内容，提供合理、准确、可靠的咨询问答；

e) 文本分类与信息提取：对文本进行分类，从中提取关键信息（如将病历文本分为不同疾病类型，并提取患者症状、诊断结果等信息）；

f) 音频问答与对话：根据患者提供的音频内容，提供合理、准确、可靠的咨询问答（如可多轮对话，根据上下文信息连贯对话，逐步深入了解患者病情）；

g) 图像问答与推理：回答针对静态图像的文本问题（如根据医学影像回答病情问题），具备视觉语言推理能力，判断描述与图片之间的一致性，推理判断图片和文本之间的关系，理解图表信息，并据此做出合理推断；

h) 视频内容理解：理解视频中的动态场景和时序内容，精确解析语义和推理时序关系（如理解视频中的动作、事件和过程性知识）；

i) 视频问答：回答关于视频内容的问题，提供详细准确信息（如回答视频中的人物、地点、事件等）。

5.1.2 抗干扰与容错

评测应至少包含以下内容：

a) 数据噪声处理：识别和过滤掉噪声数据（如错误的测量值、不完整的病历记录等），确保诊断和治疗决策准确性。例如，在分析患者的血液检查结果时，模型能够识别出异常的测量值，并进行合理的处理和修正；

b) 干扰信息排除：自动忽略与病情无关的干扰信息，专注于与病情相关的关键信息。例如，在与患者家属交流时，模型能够排除家属的情绪化表达等干扰信息，准确提取与病情相关的信息；

c) 容错能力：出现错误或不确定性时，能及时发现并采取措施进行纠正或补充，确保最终诊断和

治疗建议的可靠性。例如，对某种罕见病的诊断结果存在不确定性时，模型能提示医生进一步检查和验证。

5.1.3 自主学习

评测应至少包含以下内容：

a) 知识更新：自动学习和整合最新的医学研究成果、临床指南和专家经验，不断更新和完善自身医学知识库。例如，当有新的儿童疾病治疗方法被提出时，模型能够及时学习并掌握这些新知识，为医生提供最新的治疗建议。

b) 性能优化：不断优化自身算法和参数，提高诊断和治疗建议的准确性和可靠性。例如，模型能够通过学习大量的临床案例，不断优化对某种疾病的诊断算法，提高诊断的准确率。

c) 个性化学习：根据不同临床场景和用户需求，进行个性化学习和调整，提供更加精准和个性化的医疗服务。例如，对于不同年龄段、不同病情的儿童患者，模型能够根据其特点进行针对性的学习和优化，提供最适合的诊断和治疗方案。

5.2 儿科专业知识能力

5.2.1 医学基础知识

评测应至少包含以下内容：

a) 儿科学：儿科学基本理论及相关专业知识；

b) 遗传学：如罕见遗传病的基因突变机制、新生儿代谢筛查技术、基因测序数据解读方法；

c) 免疫学与微生物学：如儿童免疫系统发育特点、感染性疾病病原体识别及疫苗接种原理；

d) 其他医学基础学科知识：如解剖学、生理学、病理学、药理学、生物化学、神经科学、医学统计学、医学影像学、营养学、组织胚胎学、细胞生物学、医学心理学、运动学等。

5.2.2 临床医学

评测应至少包含以下内容：

a) 诊断学知识：运用医学知识和数据，鉴别和判断疾病，确诊或排除某些疾病；

b) 疾病知识：儿童常见病规范化诊疗流程，危重症早期识别与转诊标准，慢性病长期管理方案等；

c) 治疗决策：掌握儿童疾病的药物治疗、手术治疗及康复治疗方案，根据患者个体化需求制定精准治疗计划；

d) 急诊与重症监护：儿童心肺复苏指南、休克分类处理原则、多器官功能衰竭支持治疗等；

e) 辅助检查：CT/MRI影像特征识别、心音信号分析、代谢组学数据解读等。

5.2.3 健康管理

评测应至少包含以下内容：

a) 儿童保健：掌握儿童生长发育监测、营养指导、疫苗接种等健康管理知识，提供个性化健康管理建议；

b) 慢病管理：熟悉儿童慢性疾病（如哮喘、糖尿病等）长期管理策略，制定随访计划和干预措施；

c) 预防医学：掌握儿童疾病预防基本原则，通过健康教育、筛查和早期干预降低疾病发生率。

5.2.4 儿童生长发育

评测应至少包含以下内容：

a) 儿童生长发育：从出生到青春期的生理、心理和社会行为的发育规律，包括各阶段生长指标、发育里程碑以及发育迟缓的识别和干预；

b) 营养与喂养：熟悉儿童各年龄段的营养需求及喂养策略，提供科学的饮食指导；

c) 行为与心理：了解儿童行为发育及心理健康的常见问题，提供早期干预和家庭支持。

5.2.5 公共卫生

评测应至少包含以下内容：

a) 流行病学：掌握儿童常见疾病的流行病学特征，参与疾病监测和防控工作。

b) 健康政策：熟悉儿童健康相关政策法规，为公共卫生决策提供科学依据。

c) 健康教育：具备儿童健康教育的知识和技能，通过多种形式普及健康知识。

5.3 医疗应用能力

5.3.1 临床诊疗服务

评测应至少包含以下内容：

a) 病史采集：通过自然语言交互收集患者主诉、现病史、既往史、家族史、个人史等信息，并自动生成标准化病历记录，掌握常见疾病典型症状和病史特征，能识别关键信息并进行智能追问；

b) 体格检查：借助可穿戴设备指导、获取体格检查结果（包括生命体征测量、系统查体等），能识别异常体征，可结合病史进行初步判断，掌握儿科常见疾病的典型体征和查体要点。

c) 辅助检查选择与结果解读：根据临床表现推荐必要的实验室检查、影像学检查，解读检查结果；具备量表评估能力，可进行生长发育评估、心理评估等；掌握各类检查的适应症、禁忌症及结果解读标准；

d) 疾病诊断：能基于病史、查体和辅助检查结果进行疾病诊断，提供诊断依据和鉴别诊断思路；掌握常见疾病诊断标准、鉴别诊断要点，识别危重症和罕见病；

e) 治疗决策制定：根据诊断结果制定个体化治疗方案，包括药物治疗、手术治疗、康复治疗等；掌握各类治疗方案的适应症、禁忌症和最新指南；

f) 合理用药：包含用药提醒、指导、记录及咨询间、用药决策等；能提供药物剂量计算、用药指导、不良反应监测等服务；具备药物相互作用识别和禁忌症提醒功能；掌握儿科药物代谢特点和用药规范；

g) 精准转诊：能进行专家预约、精准导医、转诊；能评估病情严重程度，判断是否需要转诊；专家资源匹配能力，可协助预约专科医生；掌握各级医疗机构的诊疗能力和转诊指征；

h) 危重症处理：识别危重症早期征象，提供紧急处理建议；具备生命支持指导能力，可指导心肺复苏等急救操作；掌握危重症的识别和处理流程；

i) 个性化康复医疗与护理服务：制定个体化康复计划，提供营养指导、运动康复建议等；具备长期护理规划能力，可指导家庭护理。需掌握康复医学和护理学知识；

j) 远程医疗：能指导自诊、在线问诊、用药、转诊；能通过视频、语音等方式在线问诊，指导患者自我管理；具备远程监测能力，可分析可穿戴设备数据；掌握远程医疗技术和健康管理知识；通过视频、语音、可穿戴设备等多种形式进行诊后随访及健康监测；

k) 医疗文书撰写：自动生成规范的病历、处方、检查申请单等医疗文书；具备文书质控能力，可识别和修正错误；掌握医疗文书书写规范和质控标准；

l) 跨学科协作：能跨学科诊疗，整合多学科专家意见，优化复杂病例的诊疗流程。

5.3.2 健康咨询服务

评测应至少包含以下内容：

a) 健康监测与预警：通过分析儿童健康数据，提供健康状态监测和疾病风险预警，帮助家长及时发现潜在健康问题；

b) 个性化健康建议：根据儿童年龄、性别、健康状况等，提供个性化饮食、运动及心理健康管理建议；

c) 远程咨询：支持远程健康咨询，解答常见健康问题。

5.3.3 公共卫生服务

评测应至少包含以下内容：

a) 疾病预防与控制：掌握儿童传染病预防策略，提供疫苗接种建议及流行病防控指导；

b) 健康教育：向家长和儿童普及健康知识；

c) 公共卫生数据分析：分析儿童健康数据，为公共卫生政策制定提供科学依据。

5.3.4 医药服务

评测应至少包含以下内容：

a) 药物知识：熟悉儿童常用药物的适应症、禁忌症、剂量及不良反应，提供安全用药指导；掌握药物之间的相互作用，避免不合理用药导致的健康风险；

b) 个性化用药方案：根据儿童个体差异，推荐个性化药物治疗方案，包括药物剂量计算等。

5.3.5 医保服务

评测应至少包含以下内容：

a) 医保政策解读：熟悉儿童医保政策，为家长提供医保报销流程及政策解读服务；

b) 费用优化建议：根据患者病情及经济状况，推荐高性价比诊疗方案，减轻家庭经济负担；

c) 医保数据分析：分析医保数据，为医保政策优化提供支持。

5.4 科研教学能力

5.4.1 科学研究

评测应至少包含以下内容：

a) 智能文献检索、解读、分析：文献快速检索，支持多语言、多模态文献智能解读，自动生成文献摘要、关键信息提取和知识图谱构建，辅助科研人员高效获取研究所需的核心信息；

b) 研究动态监测：实时追踪全球科研动态，自动识别研究热点和趋势，提供领域内最新研究成果

推送服务，支持基于关键词、作者、机构的研究动态监测，辅助科研人员及时掌握领域前沿；

c) 数据整合分析：能整合多源异构数据，支持统一处理结构化与非结构化数据，能进行数据清洗、统计分析等，辅助研究人员从海量数据中挖掘有价值信息；

d) 临床研究辅助：辅助设计临床试验方案，提供病例筛选、数据收集和分析支持，自动生成符合规范的临床研究报告，辅助研究人员提高临床试验的效率和准确性。

5.4.2 辅助教学

评测应至少包含以下内容：

a) 教学资源整合：整合儿科医学教学资源，为医学生及医生提供高质量学习材料；

b) 模拟诊疗：模拟诊疗场景，辅助医学生及医生提升临床技能；

c) 知识更新：根据最新医学研究成果，及时更新教学内容，确保教学内容的先进性和实用性。

5.5 安全与伦理能力

5.5.1 数据安全

评测应至少包含以下内容：

a) 数据采集：收集医疗数据时，遵循最小化原则，明确数据采集目的，避免非必要信息收集；

b) 数据处理：对患者数据进行匿名化和脱敏处理，防止个人敏感信息泄露；

c) 数据传输与存储：采用可靠数据存储技术和设备，保证数据安全，防止数据丢失和损坏；定期备份数据，建立有效数据恢复机制；采用安全数据传输协议和加密技术，保障数据传输安全；

d) 数据访问：建立详尽访问控制和授权管理制度，仅允许经过身份验证且具有相应权限的人员接触和使用数据；实施全程操作日志记录和审计追踪机制，确保数据使用可追溯。

5.5.2 伦理遵循

评测应至少包含以下内容：

a) 伦理合规性：遵循医学伦理基本原则，如尊重患者自主权、不伤害、有利和公正、保密等；

b) 患者隐私保护：大模型在训练和使用过程中，对患者数据进行有效匿名化和脱敏处理；

c) 公平性和非歧视：避免因训练数据不均衡或算法局限性导致偏见；

d) 透明度和可解释性：决策过程透明，易于理解和解释，使患者和医疗专业人员都能清楚了解模型工作原理和决策依据。

e) 持续的伦理审查：在大模型全生命周期中，应持续进行伦理审查，确保大模型的使用始终符合伦理原则和法律法规；

f) 法律合规性：大模型的研发、使用和推广符合《个人信息保护法》《数据安全法》等法律法规。

6 评测方法

6.1 建立评测集

建立针对儿科领域医疗大模型的多模态专科评价数据库，涵盖0-18岁年龄段，涵盖儿科相关专业知识，覆盖常见儿科病种（建议覆盖<诸福棠实用儿科学>中95%以上病种）。评测集题型分为选择题、判

断题、案例题。

6.2 评测程序

本研究采用自动化评测与人工评测相结合的程序开展模型性能评价。首先，构建标准化评测样本集，并对样本进行匿名化处理。随后，采用自动化评测工具对模型输出结果进行初步评分，形成基础评测结果。对于关键样本及存在争议的结果，由具备高级职称的专业人员进行人工审阅、复核与确认。

为保证评测结果的客观性，研究采用双盲实验设计。评审人员在不知晓评分对象来源的情况下，对模型结果及不同职称医生的评分结果进行独立评价，并通过统计分析比较模型与医生评分之间的一致性、差异性及综合表现。


1
构建评测样本集
2
样本匿名化处理
3
自动化评测
4
形成人工复核清单
5
高级职称专家人工审阅
6
双盲对照评价
7
统计分析
8
确认最终评测结果

6.3 评分规则

按以下规则进行评分：

a) 选择题、判断题根据答案正误计算得分，案例题通过对比模型答案和标准答案，依据答案涵盖要点信息的比例及正确与否计算得分；

b) 以随机方式从评测数据集中抽取测试题，宜按照“选择题（40 %）+判断题（40 %）+案例题（20 %）”的得分比例进行组合；

c) 每个单项能力满分为100分，得分情况采用雷达图展示；

d) 模型综合得分满分为100分，由每个单项以加权平均的方式计算最终综合得分，各单项所占权重宜参照表1，并根据模型实际应用场景综合考量。

表 1　单项能力权重

单项能力	权重
模型基础能力	20%
儿科专业知识能力	30%
医疗应用能力	30%
科研教学能力	10%
安全与伦理能力	10%

6.4 等级划分

评测等级由高到低分为A~E级（A级最佳，E级最差）。如安全与伦理能力单项低于80分，评测等级为E级；如安全与伦理能力单项不低于80分，各等级分值范围如下：

A级：综合得分 ≥ 80；

B级：60 ≤ 综合得分 ＜80；

C级：综合得分 ＜60；

D级：综合得分 ＜60。

6.5 适用范围

不同评测等级适用的范围按表2。

表 2　不同等级适用范围

等级	适用范围
A级	三级医院
B级	基层医疗机构
C级	健康咨询场景
D级、E级	不推荐使用

附录 A

（××性）

×××

参 考 文 献

[1] GB/T 19001—2016 质量管理体系　要求

[2] GB/T 22080—2016 信息安全管理体系　要求

[3] GB/T 28001—2011 职业健康安全管理体系规范

[4] SJ/T 11234 软件过程能力评估模型

[5] 《建设高标准市场体系行动方案》（中办、国办发，2021）

[6] 《国务院关于加强质量认证体系建设促进全面质量管理的意见》（国发〔2018〕3号）